Ligand Docking Tutorial

This tutorial will go over how to run a simple ligand docking experiment with RosettaLigand on the ROSIE server.

For this tutorial we will be docking oseltamivir (Tamiflu) to the influenza neuraminidase active site. The dosed form of Tamiflu are actually a prodrug, with the active form of the drug (oseltamivir carboxylate) being created through ester hydrolysis of the ethyl ester. As the carboxylate form is the form which actually binds to neuraminidase, we'll use that form during docking.

Initial set-up

For convenience, we'll be using a high-resolution structure of neuraminidase which was crystallized in the presence of oseltamivir carboxylate. This is "cheating" to some extent, as the structure of the sidechains and backbones are pre-optimized for binding of the desired small molecule. In a true experimental condition, we would likely be working with a protein crystallized in the apo state, or with a different ligand. (Or with an AlphaFold-predicted structural model.) The two approaches are often termed "self docking" and "cross docking" in the ligand docking literature.

1. Create a new working directory, and download the 2QWK structure as PDB Format.
    1. Go to https://rcsb.org and type '2qwk' in the search bar.
    2. Click on 'Download Files' on the right side of the page, then 'PDB Format'.
    3. Save the PDB file in your working directory as `2qwk.pdb`.
    (Note, the file may automatically download to your "Downloads" folder. If so, move the file into your working directory.)
2. Download the ligand structure as an SDF format.
    1. Go to https://rcsb.org and type 'G39' in the search bar.
    2. Click on 'Download Files' on the right side of the page, then 'Structure Data File (ideal sdf)'.
    3. Save the SDF file in your working directory as `G39_ideal.sdf`.
    (Note, the file may automatically download to your "Downloads" folder. If so, move the file into your working directory.)

Structure pre-processing

Protein pre-processing

Neuraminidase is biologically a tetramer, but the 2QWK structure only has a single chain in the asymmetric unit (and thus the PDB file). This is okay! The neuraminidase active site is contained entirely within one chain, and the docking will be minimally affected by the lack of other members of the tetramer. (Note this is not universally the case. For example, the active site of HIV protease sits at a dimeric interface, and requires both chains to be present.)

It is, however, worth cleaning up the structure of crystallization artifacts and other portions which are not needed for the docking. (Including the oseltamivir which is currently present!)

1. Open 2qwk.pdb with ChimeraX
2. Select the protein residues: `select protein`
   Not selected are two glycan chains, waters, the oseltamivir in the active site and two metal ions.
   While one of the metals is a structural ion within the neuraminidase chain, it will likely not not affect ligand docking, and it causes issues with some of the Rosetta steps. We'll omit it.
3. Save the selected residues as 2qwk_A.pdb (File->Save->File of Type->PDB->Save selected atoms only)

Refine the starting structure To avoid artifacts related to docking into a structure which isn't properly optimized within the Rosetta energy function, it's best to optimize the starting protein structure.

1. Go to https://rosie.rosettacommons.org and click the Relax application.
2. Upload the 2qwk_A.pdb structure on the submission page
   Note that the ROSIE display will show it as a tetramer, but that's only because the web page display code iss reading the header information in the PDB – Rosetta will only work with the single chain which is actually present in the PDB file.
3. Check the box labeled "tether backbone coordinates of the pdbs being relaxed to the coordinates in the xtal native". Don't tether sidechain atoms
4. Give the job a descriptive name and press "upload and queue job"

Once the job completes, go to the results page and click the download button next to "output/relaxed.pdb" to download the structure. Save the file as 2qwk_A_relaxed.pdb in your working directory.

Ligand pre-processing

1. Open G39_ideal.sdf in ChimeraX
   Note that ChimeraX will not show double bonds.

The wwPDB provides oseltamivir in the neutral form, whereas the physiological state will have it with a charged carboxylate and a protonated amine. To properly dock the compound, we need to convert the ligand to the appropriate (physiological) form.

Unfortunately, there isn't a way to perform this editing with ChimeraX at the moment. As such, we'll use a different approach.

1. Go back to the G39 page on rcsb.org
2. The "Isomeric SMILES" line contains a representation of the chemical structure of the molecule
   The SMILES format is way of representing chemical structures in text. It's somewhat straightforward to understand and modify if you understand the format.
3. Copy the Isomeric SMILES for oseltamivir onto the clipboard.
4. Go to the NIH NCI CADD Group Chemoinformatics Tools and User Services website at https://cactus.nci.nih.gov/index.html
5. Click on the "Online SMILES Translator" link
6. Paste the oseltamivir SMILES into the text box under Input Format
7. Edit the SMILES string to be the carboxylate form
   Convert `CCC(CC)O[C@@H]1C=C(C[C@@H]([C@H]1NC(=O)C)N)C(=O)O`
   to `CCC(CC)O[C@@H]1C=C(C[C@@H]([C@H]1NC(=O)C)[NH3+])C(=O)[O-]`
   (Changing the last 'N' and the last 'O')

8. In the output pane, select "SDF", "Kekule" and "3D"
9. Click the "Translate" button.
10. Click "Click here!" to download the structure.
11. Rename the downloaded file "oseltamivir.sdf" and move it to your working directory.
12. Open oseltamivir.sdf and compare it to G39_ideal.sdf

Docking location.

RosettaLigand is a local docking approach, and requires definition of a docking pocket. For the ROSIE server, this is specified in XYZ coordinates.

1. Open the original 2qwk.pdb file in ChimeraX.
2. Find the oseltamivir ligand in the binding pocket.

3. Ctrl-click an atom near the center of the ligand
4. Run the command `getcrd sel`
   The coordinates of the selected atom will be printed in the log
   (e.g. "Atom /A:800@C6 26.275 18.090 62.623")

RosettaLigand docking

We are now ready to dock the ligand using the ROSIE RosettaLigand application

1. Go to https://rosie.rosettacommons.org/ and click the "Ligand Docking" application
2. Open "[Ligand Docking Server Documentation]" in a new tab to read the documentation
3. Click "[Submit Ligand Docking task]"
4. Enter a short job description.
5. Upload 2qwk_A_relaxed.pdb as the Input PDB of the protein
6. Upload oseltamivir.sdf as the Input SDF of the ligand.
7. Click "Generate ligand conformers with the BCL" (200 conformers is fine)
8. Enter the X, Y, and Z coordinates from the "Docking Location" step.
9. Other parameters can be left as-is.

Analysis of docking results.

Interpreting ROSIE output

For RosettaLigand docking, ROSIE provides a preview of the 10 best structures by interface energy (predicted binding energy).

The plot displayed here is *not* a score-versus rmsd plot, but rather an interface energy versus total score (one including protein internal energy). This is because, in general, the starting ligand structure is not a good reference structure.

Below the graph is a score table. The main score to be interested in is the interface_delta score, which measures (in arbitrary units) how good Rosetta thinks the binding energy is. Other scores are primary the individual Rosetta terms of the complex, or just those of the protein-ligand interface (those prefixed with if_).

Download the top 10 lowest interface energy structures for further analysis.

Examining residue energy contributions.

The ROSIE energy breakdown app can also work with protein-ligand complexes.

1. Go to https://rosie.rosettacommons.org and click the "Energy Breakdown" application.
2. Upload the best structure from the RosettaLigand docking.
3. For ligand docking purposes, choose the "ligand" scorefunction, and upload the oseltamivir.sdf as a ligand parameter file (with ligand three letter code LG1)
4. Enter a descriptive job name and click the "Upload and queue job" button.

Once the job is done, you can examine the results.

1. Open `residue_energy_breakdown.tsv` with your spreadsheet program of choice.
2. Residue pairs should be listed with the partner which comes earlier in the PDB coming first. We can use that to select only those lines involving the ligand (which should be the last residue)
   1. Highlight and delete all the "onebody" terms at the top of the sheet.
   2. Sort by the `resid2` column

3. Highlight and delete all the lines (except the column label line) from the top of the sheet until the ligand 1X is listed in the `pdbid2` column
4. If successful, all the remaining lines should just be between residues in the protein and the docked ligand.

3. Sort the remaining lines (ascending) by the `total` column. More negative scores are better.

Residue-residue pairs with a highly negative score are contributing favorably to the binding interface (at least according to Rosetta). Residue pairs with a highly positive scores are potentially harming binding.

For a detailed description of what each energy term means, please see Alford et al. https://doi.org/10.1021/acs.jctc.7b00125 Briefly, here are the meanings of the non-zero terms you're likely to see with `ligand` scoring for protein/small molecule interactions:

- fa_atr - the attractive component of the Lennard Jones (van der Waals) interaction
- fa_rep - the repulsive component of the the Lennard Jones (van der Waals) interaction
- fa_elec - the Coulombic electrostatic interaction
- fa_sol - Implicit solvation burial terms (from Lazaridus & Karplus, with modifications).
- hbond_bb_sc - ligand-protein backbone hydrogen bonds
- hbond_sc - ligand-protein sidechain hydrogen bonds

(Other terms are either residue-internal energies, protein-protein interaction specific, or will otherwise not show up for protein/ligand interactions.)

Take a moment to look at the residue which are highly scored (either negatively or positively) in interaction with the ligand:

- According to the scoring, what sort of interactions are contributing to that score value?
- Looking at the docked structure in ChimeraX and visualizing those residues, does that interaction make sense?
- Can you determine which interactions are sidechain-based and which are backbone based?
- Would this analysis be helpful in determining how potential mutations affect binding?
- Could this analysis help guide improving the ligand?

Looking at the results with ChimeraX

ChimeraX has a tutorial about examining protein-ligand interfaces. https://www.cgl.ucsf.edu/chimerax/docs/user/tutorials/binding-sites.html

Work through the tutorial, but using the 2qwk structure and the structures from RosettaLigand docking.

Note: If the lowest energy structure from your RosettaLigand docking isn't close to the 2qwk, determine if any of the top ten structures are. If there is one, repeat the Energy Breakdown step with that structure, and compare it with the lowest energy structure – why does Rosetta think the low energy structure is better than the one which is closest to the crystallized structure?

Machine Learning docking approaches.

Using ML techniques to do ligand docking is an active area of research and a large amount of progress is being made.

One of the big developments recently is that recent structure prediction tools such as AlphaFold3, RoseTTAFold-AllAtom, Chai-1 and Boltz-1 have extended themselves to include prediction of non-protein residues, including arbitrary small molecule ligands.

Unfortunately, local installs of these tools are generally needed, as the AlphaFold3 server does not include support for arbitrary small molecules, and the other approaches are generally not available as a web-accessible server. The exeception is that Chai-1 is available (though with restrictions) on a web server at https://lab.chaidiscovery.com/ and Boltz-1 has a preliminary ColabFold integration (https://github.com/sokrypton/ColabFold).

An example Chai-1 prediction run is availible in files/chai1. You can load the resultant CIF files and compare them with the full 2qwk.pdb structure. (Note the matchmaker command of ChimeraX is useful for aligning one protein structure to another.) Keep in mind that these structures were generated from just the input sequence (fasta) and the ligand SMILES string. However, it is highly likely that 2qwk (and homologous structures) were in the training set, so the weights of Chai-1 were tuned such that it was able to recapitulate this structure, as such performance on this docking task is unlikely to be representative of how well it will do on an unknown task.

ML Ligand docking

There are also a number of ligand-docking specific ML models. The most well known of which is DiffDock, which uses a diffusion based approach to place a ligand in the protein binding site.

DiffDock has an official webserver on HuggingFace (a sever which is used heavily by machine learning people for storing models and datasets). https://huggingface.co/spaces/reginabarzilaygroup/DiffDock-Web

1. Go to https://huggingface.co/spaces/reginabarzilaygroup/DiffDock-Web
2. Upload 2qwk_A.pdb as the Input PDB
   Since DiffDock does not use the Rosetta energy function, you don't (necessarily) need to pre-relax the structure.
3. Upload oseltamivir.sdf as the Input Ligand (or alternatively, provide the SMILES string)
4. Leave Configuration blank.
5. Ignore the "examples" box and press the "Run DiffDock" button.

When the run completes, download the results (which is a zip file containing the input PDB and docked models in SDF format), and open the structures in ChimeraX along with the full 2qwk structure? When examining the structure, keep in mind that DiffDock was likely trained on 2qwk (and related structures), and as such its weights were tuned to do well on this task. Performance here is not necessarily representative of what would happen in cases where the results are unknown.