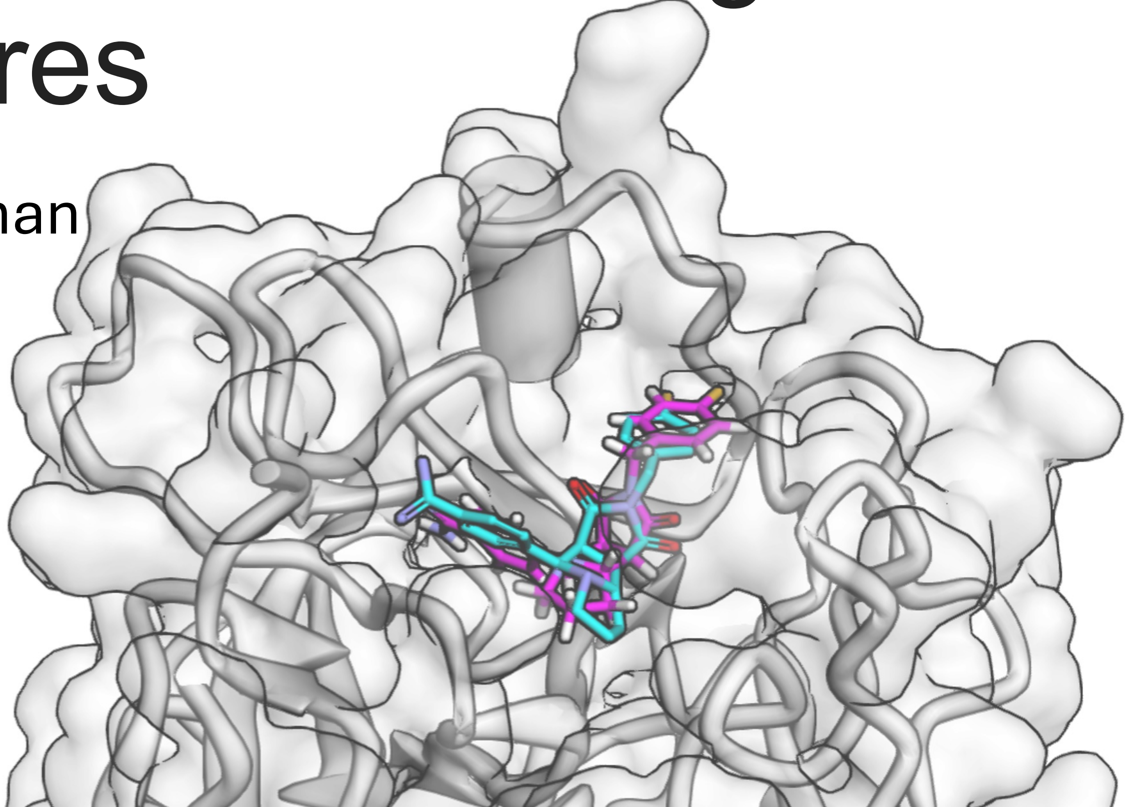
Molecular docking with Python in Jupyter Notebooks: Towards the development of accessible docking procedures

Lee Schoneman



Aims

- Overall Goal: develop an easy-to-use program that does not need to be closed to get materials or resources
  - Little command-line usage, highlighted areas for user input
  - Limit number of packages or programs needed to be downloaded by user
  - No assumption of knowledge, novice friendly
  - Flexible, allow users to choose what they want to execute

Breakdown of current notebooks in basil_dock

- Docking preparation
- Docking and preliminary analysis
- Data manipulation and collection
- Machine learning analysis

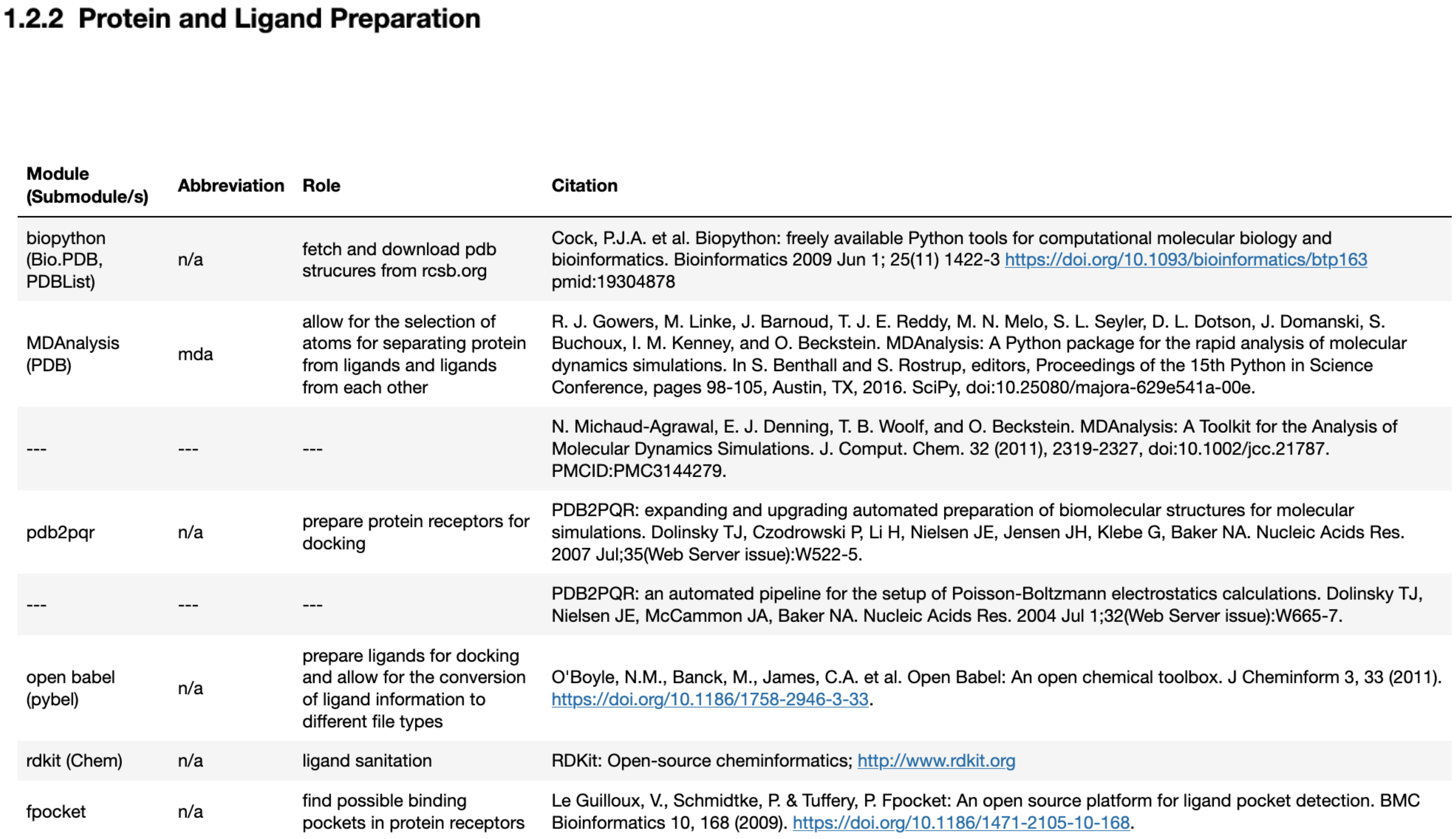
Template of Notebooks

- Purpose of the notebook/series
    - Target audience
    - Brief overview of the Jupyter notebook series
    - Stepwise summary for specific notebook being used
- Table of Libraries
    - Module/submodules used
    - Abbreviation used in code (if applicable)
    - Role
    - Citation
- Acknowledgements, if applicable

Header

Columns

Libraries used



1.2.2 Protein and Ligand Preparation

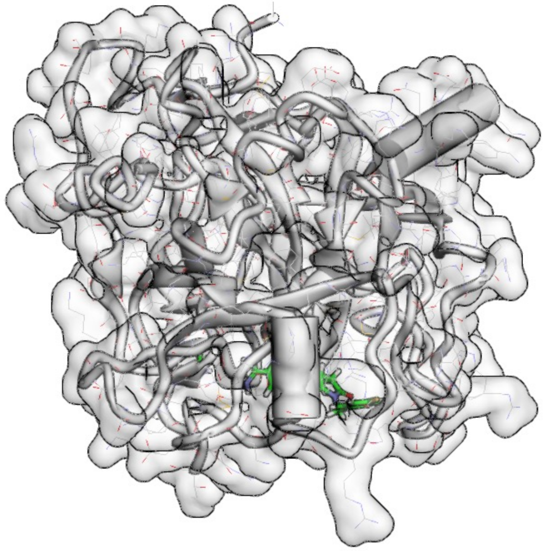
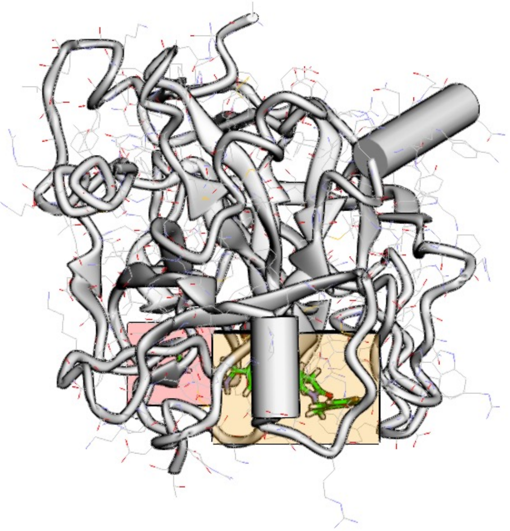
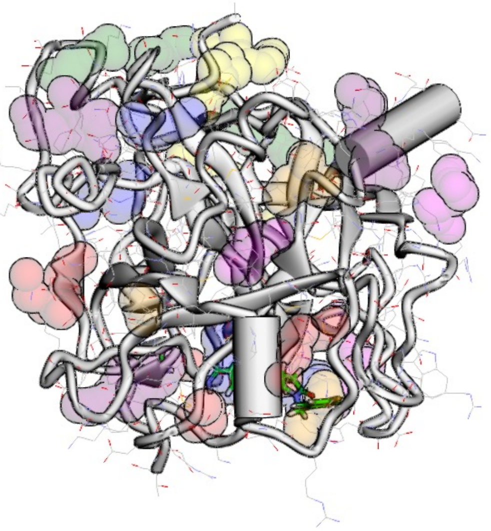
Module (Submodule/s)	Abbreviation	Role	Citation
biopython (Bio.PDB, PDBList)	n/a	fetch and download pdb strucures from rcsb.org	Cock, P.J.A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009 Jun 1; 25(11) 1422-3 https://doi.org/10.1093/bioinformatics/btp163 pmid:19304878
MDAnalysis (PDB)	mda	allow for the selection of atoms for separating protein from ligands and ligands from each other	R. J. Gowers, M. Linke, J. Barnoud, T. J. E. Reddy, M. N. Melo, S. L. Seyler, D. L. Dotson, J. Domanski, S. Buchoux, I. M. Kenney, and O. Beckstein. MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations. In S. Benthall and S. Rostrup, editors, Proceedings of the 15th Python in Science Conference, pages 98-105, Austin, TX, 2016. SciPy, doi:10.25080/majora-629e541a-00e.
---	---	---	N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein. MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. J. Comput. Chem. 32 (2011), 2319-2327, doi:10.1002/jcc.21787. PMCID:PMC3144279.
pdb2pqr	n/a	prepare protein receptors for docking	PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, Baker NA. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W522-5.
---	---	---	PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W665-7.
open babel (pybel)	n/a	prepare ligands for docking and allow for the conversion of ligand information to different file types	O'Boyle, N.M., Banck, M., James, C.A. et al. Open Babel: An open chemical toolbox. J Cheminform 3, 33 (2011). https://doi.org/10.1186/1758-2946-3-33.
rdkit (Chem)	n/a	ligand sanitation	RDKit: Open-source cheminformatics; http://www.rdkit.org
fpocket	n/a	find possible binding pockets in protein receptors	Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics 10, 168 (2009). https://doi.org/10.1186/1471-2105-10-168.

Notebook 1 – Docking Preparation

- Obtain PDB/MMCIF file from RCSB Protein Data Bank
  - Separate protein and ligands (if present) into separate files
- Import additional ligands if desired
  - RCSB PDB Chemical Component Dictionary
  - Local MOL2 file
  - SMILES strings
- Prepare and separate all ligands into individual MOL2 and PDBQT files
- Find possible binding pockets in protein
- Visualize protein and ligand/s

Notebook 1 output

- Protein receptor files: CIF, PDB, and PDBQT files for given receptor
- Ligand file/s: MOL2 and PDBQT files for each ligand utilized
- Protein pockets: CSV file containing information for each pocket

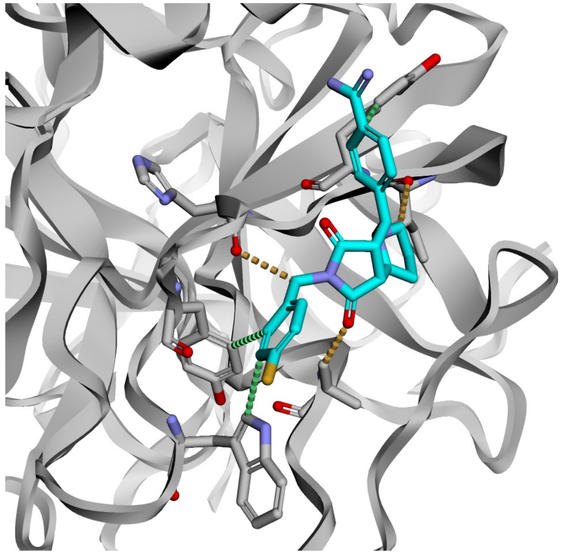




Notebook 2 – Docking and Preliminary Analysis

- Get docking box sizes from the docking-prep notebook
- Dock the ligand to protein using either VINA or SMINA
- Visualize different poses of ligands docked to the protein
- Visualize protein-ligand interactions of poses

Notebook 2 output

- VINA: PDBQT and SDF files for each ligand for a given pocket
- SMINA: SDF files for each ligand for a given pocket
- Docking information: CSV file containing conformation, pose, and position information for each ligand and location information for each pocket used in docking



Reference (FSN501): Magenta | Vina Pose (FSN501): Cyan
Pose: 3 | Score: -8.477

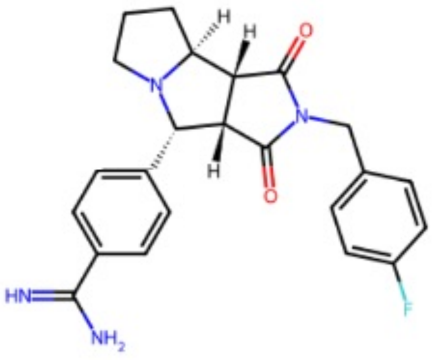
Notebook 3 – Data Manipulation and Collection

- Create derivatives for ligands by substituting/modifying functional groups on a canonical ligand
- Dock derivative/s to receptor
- Collect and store data
  - Score
  - Interaction type
  - Distance between interacting atoms from the ligand and protein
  - Functional group involved in interaction
- Visualize different poses of ligands docked to protein
- Visualize protein-ligand interactions of poses

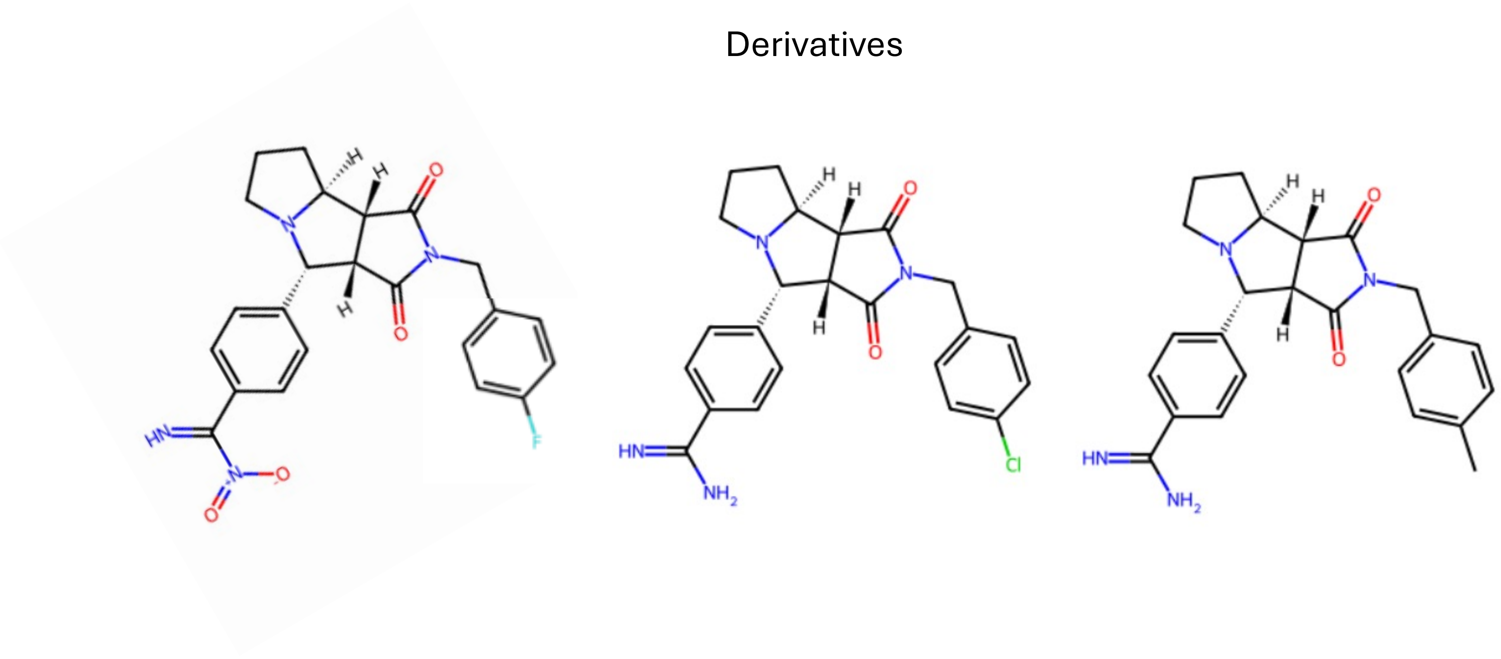
Notebook 3 output

- Same output as Notebook 2 but with related ligands

Original ligand



Derivatives



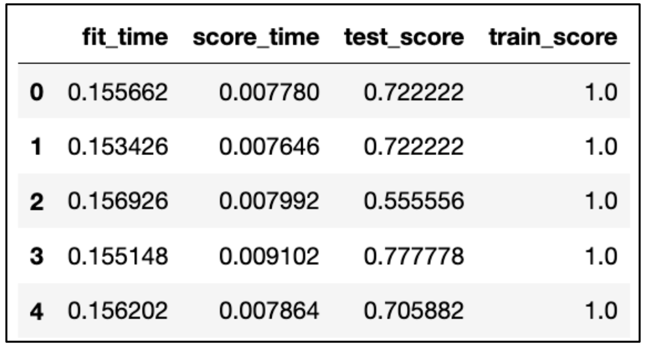
Notebook 4 – Machine Learning Analysis

- Determine the likelihood of a compound being orally bioactive using
  - Lipinski's Rule of Five
    - MW < 500
    - XLogP ≤ 5
    - HBD ≤ 5
    - HBA ≤ 10
  - Lipinski's Rule of Five with the Ghose filter
    - Meets Lipinski’s Rule of 5 criteria
    - 40 ≤ Molar refractivity ≤ 130
    - -0.4 ≤ XLogP ≤ 5.6
    - 20 ≤ Number of atoms ≤ 70
  - Veber’s Rule
    - Rotatable bonds ≤10
    - Polar surface area ≤140 square angstroms

Notebook 4 output

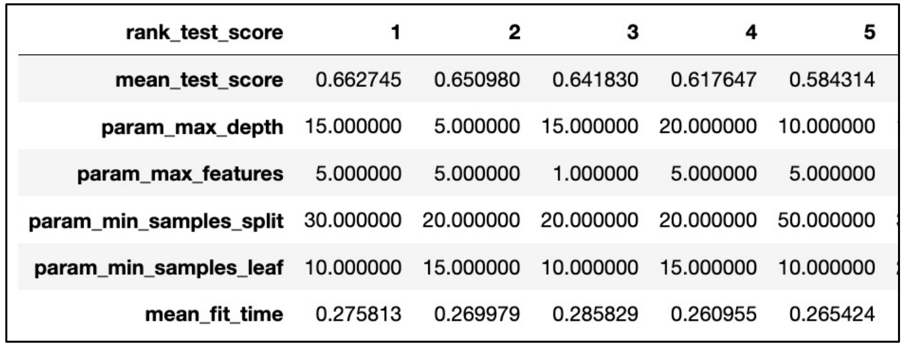
- Feature importance regarding oral bioactivity
- Predict oral bioactivity of experimental ligands

Initial train-test scoring



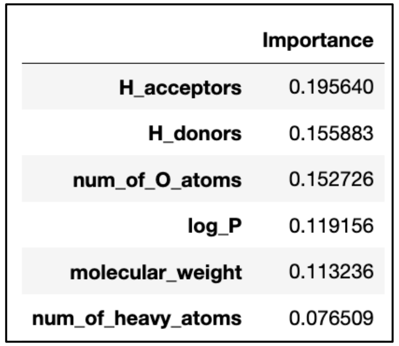
	fit_time	score_time	test_score	train_score
0	0.155662	0.007780	0.722222	1.0
1	0.153426	0.007646	0.722222	1.0
2	0.156926	0.007992	0.555556	1.0
3	0.155148	0.009102	0.777778	1.0
4	0.156202	0.007864	0.705882	1.0

Hyperparameter optimization



rank_test_score	1	2	3	4	5
mean_test_score	0.662745	0.650980	0.641830	0.617647	0.584314
param_max_depth	15.000000	5.000000	15.000000	20.000000	10.000000
param_max_features	5.000000	5.000000	1.000000	5.000000	5.000000
param_min_samples_split	30.000000	20.000000	20.000000	20.000000	50.000000
param_min_samples_leaf	10.000000	15.000000	10.000000	15.000000	10.000000
mean_fit_time	0.275813	0.269979	0.285829	0.260955	0.265424

Feature Importance



	Importance
H_acceptors	0.195640
H_donors	0.155883
num_of_O_atoms	0.152726
log_P	0.119156
molecular_weight	0.113236
num_of_heavy_atoms	0.076509

Looking Forward

- Impact of changing ligand functional groups on binding energy
- Teach users to explain these calculated differences
- Determine the importance of ligand functional groups and protein receptor residues in forming protein-ligand complex
- Consolidate all notebooks into a Jupyter book

Acknowledgements

The Rochester Institute of Technology

Angel Ruiz Moreno, developer of Jupyter Dock: Molecular Docking integrated in Jupyter Notebooks

Github: AngelRuizMoreno

Jessica Nash of MOLSSI, developer of iqb-2024 repository used in the IQB 2024 workshop - Python for Molecular Docking

Github: janash

Support for this project was provided by NSF 2142033.