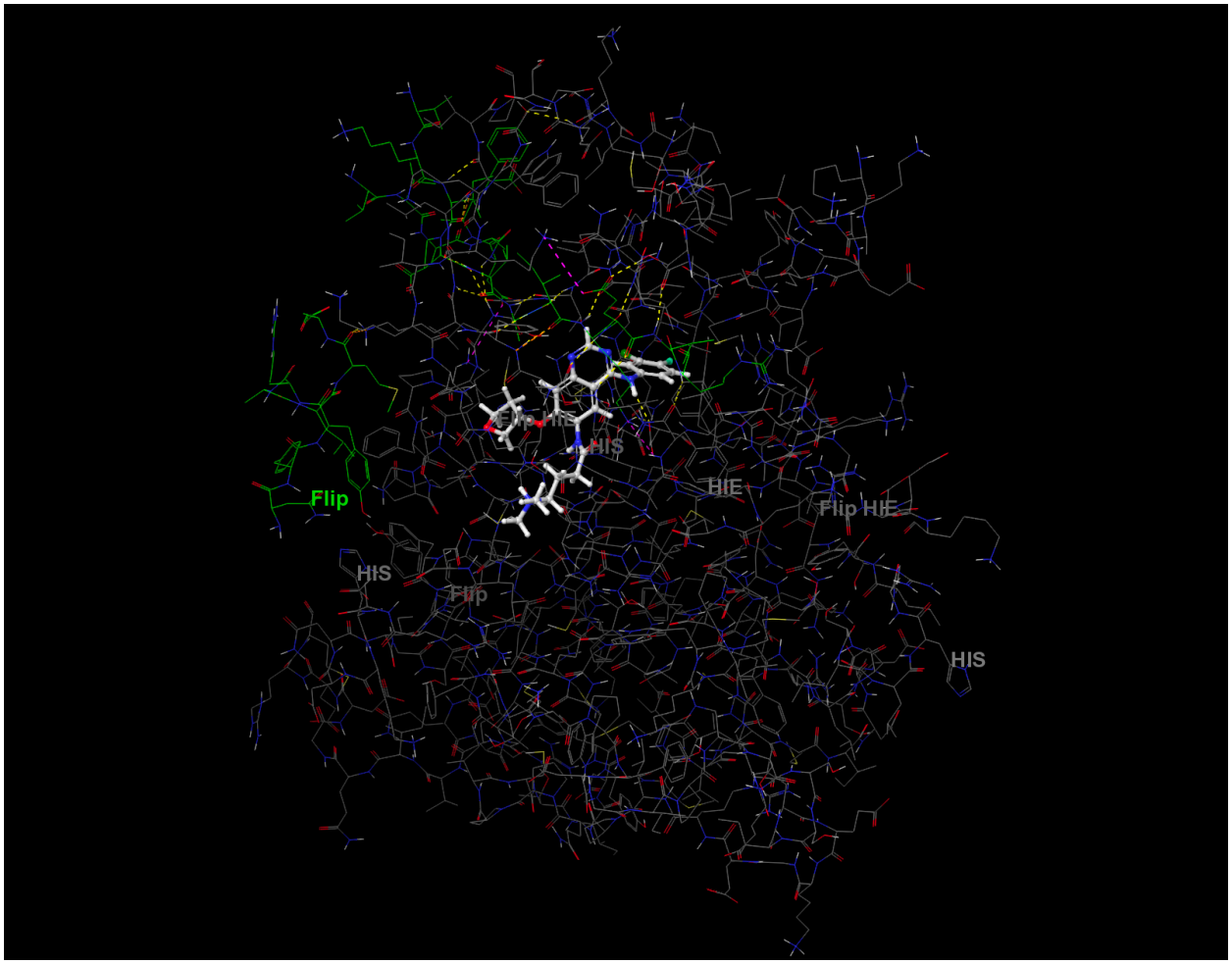
SDU

Final Project - FA509



Vishnu Mothukuri

Objective

This project aims to study the inhibition of EGFR, epidermal growth factor receptor while employing molecular docking using Glide and BioLuminate. It involves studying the receptor-ligand interactions while comparing the binding affinities of various FDA-approved EGFR inhibitors currently prescribed in the market.

Introduction

Epidermal Growth Factor (EGF) is a key player in cellular functions, steering processes like growth, development, and tissue repair. Its counterpart, the Epidermal Growth Factor Receptor (EGFR), serves as a central figure with tyrosine kinase capabilities, awaiting specific signals for activation. EGF binds to EGFR. The initiation of EGFR activation involves the formation of receptor dimers, where pairs of receptors collaborate. This dimerization event leads to the autophosphorylation of tyrosine residues located in the carboxyl-terminus of EGFR by the kinase domain. Notably, instances of EGFR overexpression or genetic mutations can result in uncontrolled cell proliferation, leading to cancer.

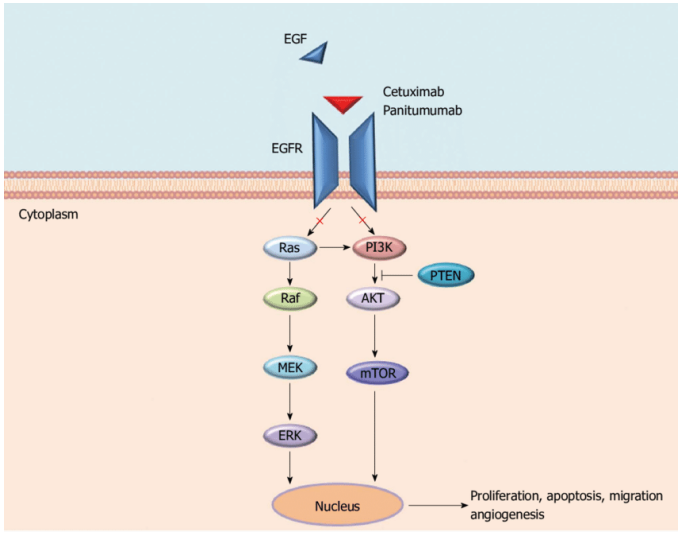


Figure 1 EGFR in the cell membrane

Glide Docking

Glide is a widely used software for molecular docking, a process crucial in computational drug design. Molecular docking simulates the interaction between a ligand and a target protein, predicting how these molecules fit together and the strength of their interaction. Glide generates a grid around the binding site of the protein. This grid represents the area where the docking process will take place. The grid includes information about the electrostatic and van der Waals properties, which influence ligand binding. Ligands are flexibly docked into the grid. Glide allows ligands to adopt different conformations and orientations within the binding site. Glide generates conformations internally during the docking process; this procedure is known as "flexible docking." Conformation generation is limited to variation around acyclic torsion bonds, sampling low-energy ring conformations, and generation of pyramidalizations at certain trigonal nitrogen

centers. Rigid docking allows the existing ligand structure to be translated and rigidly rotated relative to the receptor but skips the conformation generation step. In this project, flexible docking was used to dock the ligands.

Glide uses several scoring functions to evaluate the potential of a ligand to bind effectively to the target protein. Standard Precision (SP) docking is designed to be faster and less computationally demanding. SP mode uses a slightly simplified scoring function and a less rigorous conformational search algorithm than XP.
Extra Precision (XP) docking is more computationally intensive and aims for higher accuracy in predicting ligand binding. It is used when a more precise understanding of ligand-protein interactions is needed. It includes additional terms in the scoring function, such as coulomb energy, van der Waals energy, binding energy, and penalties for ligand desolvation and strain, to predict binding affinities more accurately. It is typically used for a more focused set of compounds after initial screening, as it rules out false positives and is valuable for lead optimization.

Epik uses Hammett and Taft methodology to predict ionization states and their energetic penalties. Epik can also predict different tautomeric forms and calculate energetic penalties for every ligand state it predicts. The ligands have been prepared using Epik for ionization and tautomerization, and Epik penalties for adopting higher-energy states are added to the docking score. Glide calculates a compound's Docking Score by adding the Epik state penalty to the compound's GlideScore.

Method

Receptor preparation

The file containing the structure of EGFR in complex with hydrazone, a potent dual inhibitor, was downloaded and loaded from PDB. It was then prepared using the Protein Preparation Workflow. The receptor was preprocessed, and missing side chains were filled in. After this, the hydrogen bonding network within the structure was optimized. The hydrogen bonding network is optimized by reorienting hydroxyl and thiol groups, water molecules, amide groups of asparagine (Asn) and glutamine (Gln), and the imidazole ring in histidine (His) and predicting protonation states of histidine, aspartic acid (Asp) and glutamic acid (Glu) and tautomeric states of histidine. PROPKA was used for specifying the protonation states. It predicts the pKa values of ionizable groups in the protein. Following this, a restrained minimization is performed on the structure to delete water molecules. All atoms are minimized, and only the waters near het groups are kept. The default OPLS4 force field was used for the minimization.
After this, the grid for the ligand interaction was generated with a scaling factor of 1.00 and a partial charge cutoff of 0.25.

Ligand preparation

The three ligands AQ4 (Erlotinib), IRE (Gefitinib), and 0WN (Afatinib) are tyrosine kinase inhibitors that bind to the tyrosine kinase domain in the epidermal growth factor receptor and stop the activity of the EGFR. They were downloaded from PDB and prepared in LigPrep at pH 7.00 +/- 1.00 with the OPLS4 force field. The IRE ligand was docked in both the protonated and non-protonated forms and showed better binding in the protonated form.
These prepared ligands were docked first with SP and then with XP, with a maximum of 5 poses and a scaling factor of 1.00.

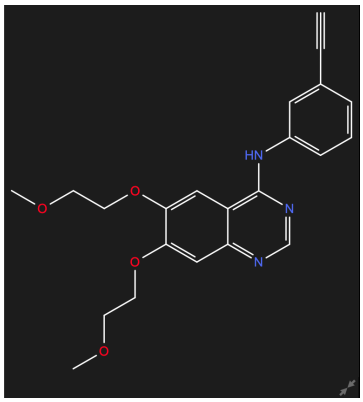


Figure 2 AQ4

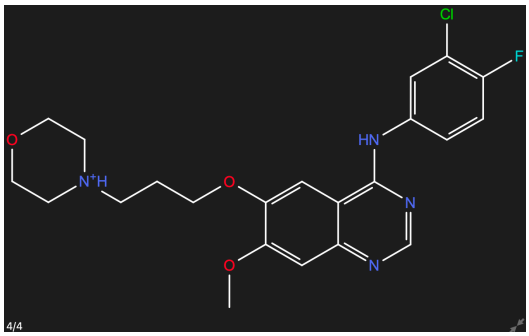


Figure 3 IRE

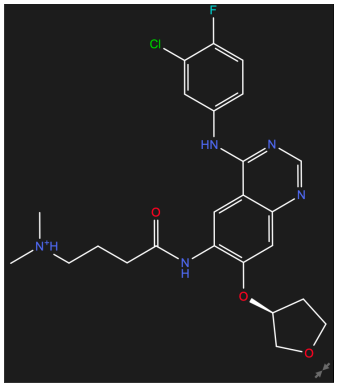


Figure 4 0WN

Results

Table 1 shows the results from docking AQ4 in SP followed by XP.

Sample - AQ4	Scoring	State Penalty	Force Field	Docking score	Glide Gscore
1	SP	0.0000	S-OPLS	-7.655	-7.655
2	SP	0.0000	S-OPLS	-7.649	-7.649
3	SP	0.0000	S-OPLS	-7.473	-7.473
4	SP	0.0000	S-OPLS	-7.076	-7.076
5	SP	0.0000	S-OPLS	-6.630	-6.630

6	XP	0.0000	S-OPLS	-10.302	-10.302
7	XP	0.0000	S-OPLS	-8.927	-8.927
8	XP	0.0000	S-OPLS	-8.353	-8.353
9	XP	0.0000	S-OPLS	-8.135	-8.135
10	XP	0.0000	S-OPLS	-7.806	-7.806

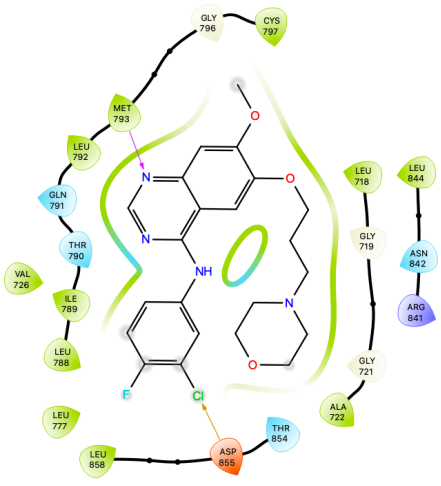
Table 2 shows the results from docking IRE in SP followed by XP in the protonated and non-protonated forms.

Sample-IRE	Scoring	Protonation	State Penalty	Docking score	Glide Gscore
1	SP	No	0.1963	-9.412	-9.608
2	SP	Yes	0.7500	-9.055	-9.805
3	SP	No	0.1963	-9.031	-9.228
4	SP	Yes	0.7500	-8.935	-9.685
5	SP	Yes	0.7500	-8.706	-9.456
6	SP	Yes	0.7500	-8.696	-9.446
7	SP	No	0.1963	-8.336	-8.532
8	SP	No	0.1963	-7.818	-8.014
9	SP	No	0.1963	-5.646	-5.842
10	XP	No	0.1963	-9.638	-9.835
11	XP	No	0.1963	-9.475	-9.672
12	XP	Yes	0.7500	-9.086	-9.836
13	XP	Yes	0.7500	-8.887	-9.637
14	XP	Yes	0.7500	-8.886	-9.636
15	XP	Yes	0.7500	-8.475	-9.225
16	XP	Yes	0.7500	-8.428	-9.178
17	XP	No	0.1963	-8.394	-8.590
18	XP	No	0.1963	-8.032	-8.228

Table 3 shows the results from docking 0WN in SP followed by XP.

Sample - 0WN	Scoring	State Penalty	Force Field	Docking score	Glide Gscore
1	SP	0.0030	S-OPLS	-8.022	-8.025
2	SP	0.0030	S-OPLS	-7.465	-7.468
3	SP	0.0030	S-OPLS	-7.411	-7.414
4	SP	0.0030	S-OPLS	-7.245	-7.248
5	SP	0.0030	S-OPLS	-7.108	-7.111
6	XP	0.0030	S-OPLS	-9.484	-9.487
7	XP	0.0030	S-OPLS	-7.421	-7.424
8	XP	0.0030	S-OPLS	-6.655	-6.658
9	XP	0.0030	S-OPLS	-6.272	-6.275
10	XP	0.0030	S-OPLS	-6.187	-6.190

Figures 5 and Figure 6 show the Ligand Interaction Diagrams of the deprotonated pose (left) and the protonated pose (right) of IRE (Gefitinib).



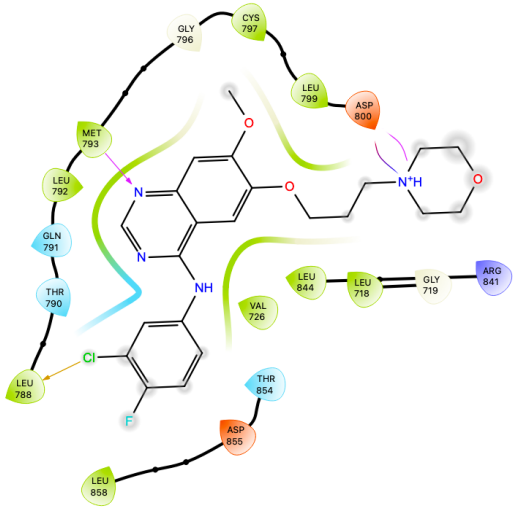


Figure 7 shows the Ligand Interaction Diagram for the best docking score by AQ4 (Erlotinib).

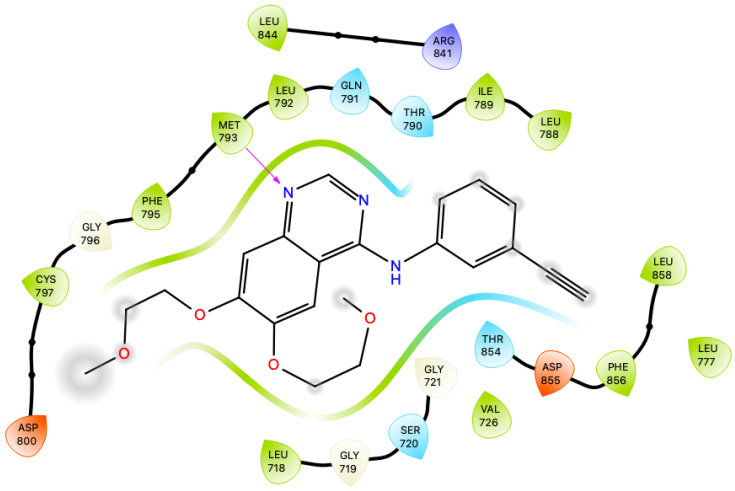
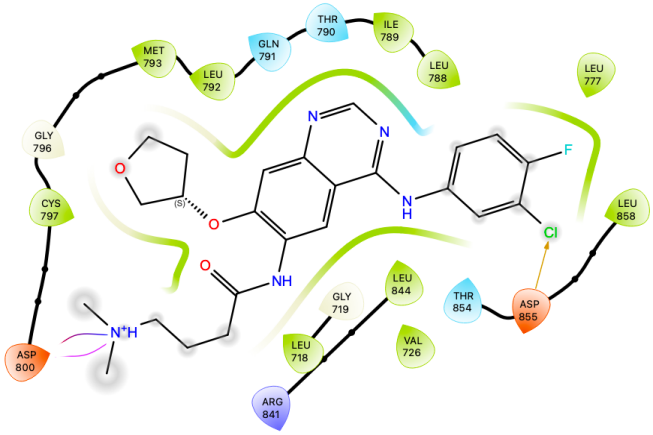


Figure 8 shows the Ligand Interaction Diagram for the best docking score by 0WN (Afatinib).



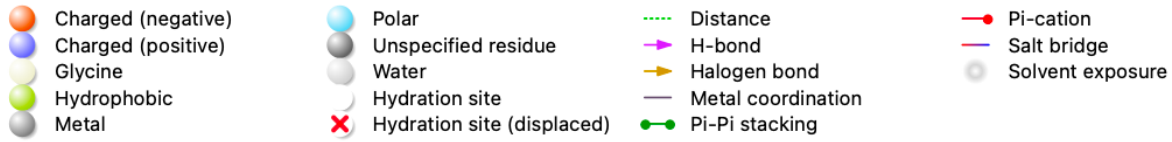


Figure 9 LID Legend

Discussion

According to literature review, first-generation tyrosine kinase inhibitors of EGFR like Gefitinib and Erlotinib bind to the ATP-binding pocket of EGFR. Amino acid Met 793 lies in the “hinge” region of the kinase.

As we can observe from the LID of Gefitinib, a hydrogen bond is formed with Met 793. It also has a fluorine group in the para position that extends towards Leu 858. The chloro group is surrounded by the sidechains of residues Leu 788 and Thr 790. These results go hand in hand with the literature. The protonated N also forms a salt bridge with the negatively charged amino acid Asp 800. In the case of Erlotinib, a hydrogen bond to the Met 793 amino acid backbone is observed. Afatinib’s LID shows us the formation of a salt bridge of the protonated N with Asp 800.

It is worth looking at the different docking scores obtained using the SP and XP scoring functions. As expected, XP provides us with a more precise and, in most cases, higher docking score. Docking with Erlotinib (AQ4) gave the highest glide score, indicating that this ligand has the best binding affinity. In the case of Gefitinib (IRE), the protonation states made a difference in the glide score when compared with SP. However, when XP was used for the comparison, the protonated ligand had only a marginally better overall binding affinity.

Conclusion

Through exploring the different tyrosine kinase inhibitors of EGFR, this project compares the binding affinity of different ligands using Glide. This process can be helpful to predict and aid the drug development process. We can see that Extra Precision docking was the most efficient in estimating the binding affinities of these drugs.

EGFR is a critical target in many types of cancers, including lung, breast, and colorectal cancers. This project's focus on identifying effective inhibitors of EGFR can lead to the development of more targeted therapies for these cancers, leading to personalized treatment. The analysis of FDA-approved inhibitors could also reveal new uses for existing drugs, a process known as drug repurposing. By examining how different inhibitors interact with the EGFR, the study can be further explored to provide insights into the mechanisms of drug resistance. This is particularly relevant for cancers where resistance to EGFR inhibitors is common, helping develop next-generation inhibitors.

References

- Yun, C. H., Boggon, T. J., Li, Y., Woo, M. S., Greulich, H., Meyerson, M., & Eck, M. J. (2007). Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell*, *11*(3), 217–227. https://doi.org/10.1016/j.ccr.2006.12.017
- Amelia T, Kartasasmita RE, Ohwada T, Tjahjono DH. Structural Insight and Development of EGFR Tyrosine Kinase Inhibitors. *Molecules*. 2022; 27(3):819. https://doi.org/10.3390/molecules27030819
- Glide User Manual Copyright © 2015 Schrödinger, LLC. All rights reserved.

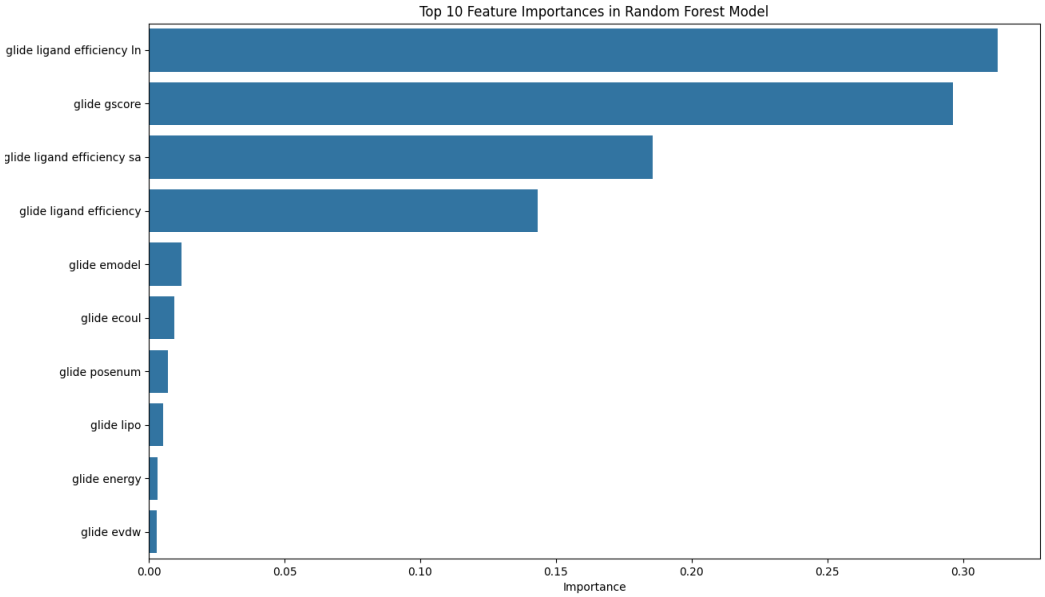
SDU

Integrating Machine Learning with Molecular Docking

Drug development was something that sparked an interest in me. Coming from a computational biology background, I wanted to further explore this project's scope by employing machine learning. A Random Forest Regressor, a machine-learning model used for regression tasks, was created. Its goal was to predict the docking score of a ligand in relation to a specific protein target, in this case, EGFR, before performing the actual docking using Glide.

Feature preparation

A Random Forest is an ensemble learning method. It constructs multiple decision trees during training and outputs the average prediction of the individual trees. This method is effective for regression (predicting a continuous value) and reduces the risk of overfitting, which is common in single decision trees. The model requires input in the form of features, which are measurable properties or characteristics of the ligands. Here, the features I inputted into the model were the various properties and descriptors of ligands. The model's mean squared error (MSE) is approximately 0.018, and the R^2 score is around 0.979, indicating a good fit for the test data. Training data (80%) is used to train the machine learning model. The model learns to make predictions based on this data. Test data (20%) is used to evaluate the model's performance. This set is not used during training and tests how well the model generalizes to new, unseen data.



Training the model

The training data, which includes known ligands with their corresponding docking scores and various molecular descriptors, is used to 'teach' the model the relationship between these features and the docking scores. The model learns from the training data. It finds patterns that correlate the features of the ligands (like molecular weight, charge, etc.) with their docking scores. This learning phase involves adjusting the model parameters to minimize the difference between the predicted and actual docking scores in the training data.

Prediction

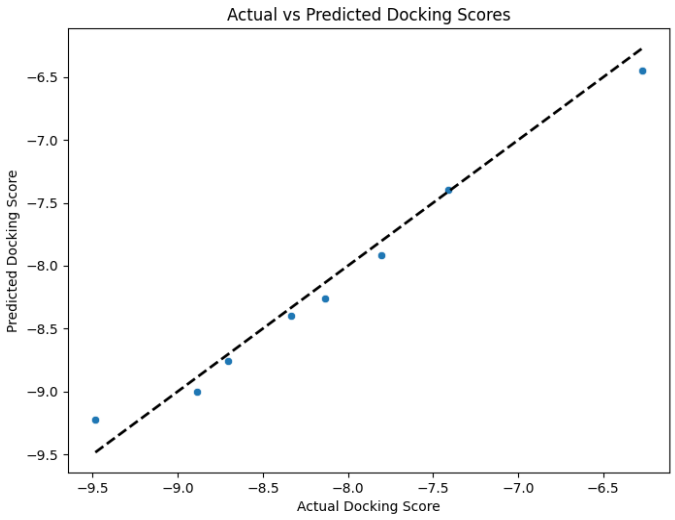
Once trained, the model can predict the docking score for a new ligand. This prediction is based on the patterns it learned during training. When a new ligand is introduced, its features are fed into the model, which predicts its docking score based on what it has learned.

Before predicting the binding affinity of a ligand, new ligand data must be preprocessed to match the format of the training data. This includes handling missing values, ensuring the same features are present, and arranging them in the same order. If the new data has missing values, they are filled with the mean values from the training set to maintain consistency.

Output

The output is the predicted docking score for the ligand, which estimates how well a ligand might bind to the target protein. This score is then used to assess the potential efficacy of the ligand as a drug candidate, guiding further experimental validation.

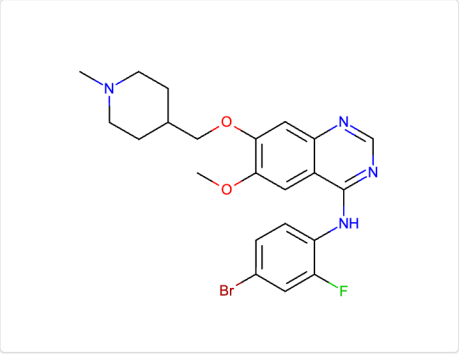
The scatter plot illustrates the relationship between the actual and predicted docking scores for the test set. The closer the points are to the dashed diagonal line, the more accurate the predictions. The plot shows that most predictions are quite close to the diagonal line, indicating that the Random Forest model performs well in predicting the docking scores for these ligands.



Ligand	Actual Docking Score from Glide	Predicted Docking Score
0WN - Afatinib	-9.484	-9.22487
0WN - Afatinib	-6.272	-6.44623
IRE - Gefitinib	-8.706	-8.75862
IRE - Gefitinib	-8.886	-9.00105
AQ4 - Erlotinib	-8.135	-8.26377
AQ4 - Erlotinib	-7.806	-7.91576

To further test the model, ligand Vandetanib or ZD6 was used for the docking.
First, the ligand was imported, and all the necessary features were extracted. These features were then imported into the machine learning model to predict the binding affinity of ZD6 with EGFR.

The predicted docking score for the new ligand (ZD6) using the trained Random Forest model was approximately -8.03.



ZD6

To test the accuracy of this, I manually performed docking using Glide as it was done for the previous ligands. These are the results from the Glide docking using both SP and XP.

Sample - ZD6	Scoring	State penalty	Force Field	Docking Score	Glide GScore
1	SP	0.0000	S-OPLS	-7.289	-7.289
2	SP	0.0000	S-OPLS	-6.915	-6.915
3	SP	0.0000	S-OPLS	-6.782	-6.782

Sample - ZD6	Scoring	State penalty	Force Field	Docking Score	Glide GScore
4	SP	0.0000	S-OPLS	-6.749	-6.749
5	SP	0.0000	S-OPLS	-6.626	-6.626
6	XP	0.0000	S-OPLS	-8.712	-8.712
7	XP	0.0000	S-OPLS	-8.582	-8.582
8	XP	0.0000	S-OPLS	-8.146	-8.146
9	XP	0.0000	S-OPLS	-8.062	-8.062
10	XP	0.0000	S-OPLS	-4.439	-4.439

Discussion

The predictions of the model were fairly accurate and provide an idea of the binding affinities of ligands based on their properties. One of the reasons why the prediction of ZD6 binding affinity was not as accurate as the rest is due to conformation generation. As the maximum number of poses was set to 5, Glide employed multiple poses to find the pose with the highest binding affinity. The data from all these poses was not made available to the machine learning model; only the data from the initial ligand preparation was given. Whereas in the case of the previous ligands, it had access to data on all poses of the respective ligands.
While in this project, I only used three ligands, a larger dataset with more ligands, receptors, etc., can increase the model's accuracy.

Conclusion

The integration of Glide docking results with machine learning models can further enhance the predictive capabilities and efficiency of the drug discovery process. A machine learning model can assist in identifying potential drug candidates based on their predicted binding affinities. This makes the overall drug discovery process more efficient while screening numerous ligands or lead compounds.