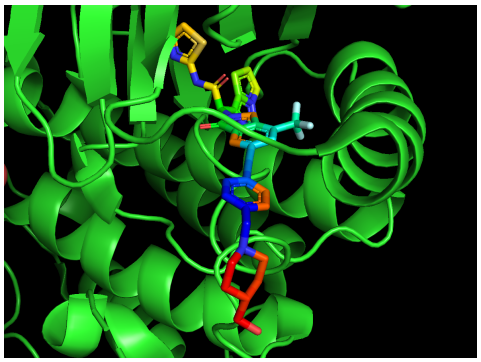
Open Source Software and Libraries for Data Analysis and Computer-Aided Drug Design

JIS KOCHUNIRAVATHU SAJI 7027109

REPORT

EGFR stands for epidermal growth factor receptor, a transmembrane receptor protein that is found in the tyrosine kinase (RTK) family.EGFRs are really important as any irregularities, mutations,overexpression etc can have serious implications and which inturn can be a hallmark of cancer. If EGFR is overexpressed it causes the cancer cells to divide more rapidly.so it is important to keep the levels down and we have EGFR inhibitors. Therefore it is important to find EGFR inhibitors .EGFR inhibitors work by inhibiting the downstream signaling pathways activated by EGFR receptors.The main example of EGFR in pathology are different types of cancer,inflammatory diseases,neurological disorders and cardiovascular diseases.There are mainly two types of inhibitors,small molecule inhibitors and monoclonal antibodies.The pdb id of the protein I have selected is ‘8A2A’.



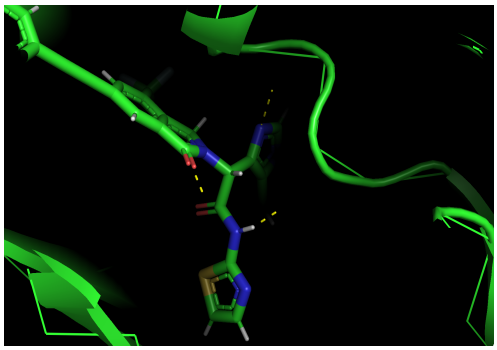
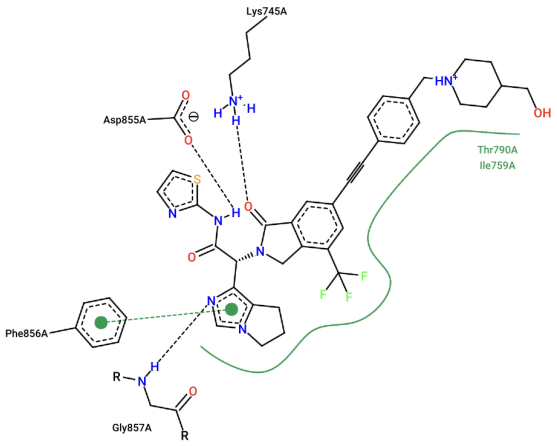


Table 1:Potential off targets from probis server

Ligand Name	PDB ID	Confidence
ADENOSINE-5'-DIPHOSPHATE	ADP	4.57
PHOSPHOAMINOPHOSPHONIC ACID-ADENYLATE ESTER	ANP	4.57
Gefitinib	IRE	4.57
PHOSPHOAMINOPHOSPHONIC ACID-ADENYLATE ESTER	ANP	4.57

[6,7-BIS(2-METHOXY-ETHOXY)QUINAZOLINE-4-YL]-(3-ETHYNYLPHENYL)AMINE	ANP	4.57



The structure chosen is '8A2A' and it was chosen because it is a relatively new protein(2022) and the resolution(1.43Ä) is pretty less and is also an Xray crystallized structure.When the structure was analyzed in protein plus,the number of hydrogen bond donors is 23,acceptors is 66 and the hydrophobic interactions are 64.

There are a total of 9 potential binding pockets in the structure. This information was provided by the DoGSiteScorer in the protein plus website by uploading the ligands and the structure. The best drug pocket according to the drug score is 'p_1' with 0.85 and the drug pocket where the ligand binds is 'p_0' with a score of 0.82. The volume of the 'p_0' pocket is 1302.53 and for 'p_1' it is 636.74.

The Pipeline(compound filtering)

The pipeline which was created used random forest , a machine learning approach , clustering using the butina algorithm ,creating a dissimilarity matrix and filtering using the lipinski rule of 5.

The first step in the pipeline is to sort the given list of molecules by their pIc50 values. The higher the pIc50 value the better is the drug. In the pipeline we sort the compounds into actives and inactives but putting a threshold of 6.3.

The next step involves breaking the active and inactive lists into fingerprints . We have Morgan and Maccs fingerprints available. It is observed that morgan fingerprints work better with my pipeline from the comparison of the five compounds for similarity by using both the methods.Two separate datasets were used to The two machine learning approaches we have used are k-fold and random forest classifier.The accuracy and AUC value of the model is 0.87 and 0.93 respectively. The accuracy,sensitivity,AUC and specificity of the random k-fold are given below.

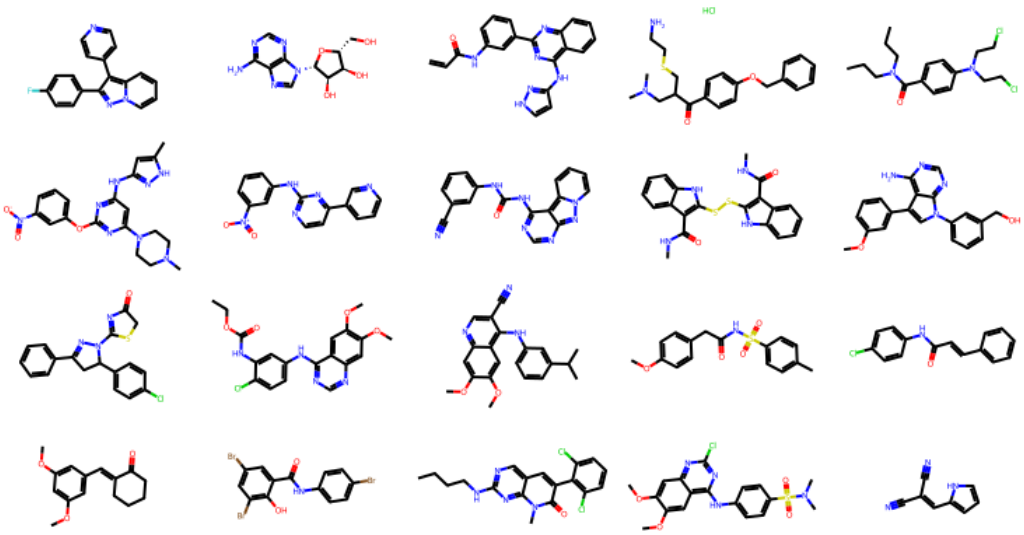
	Accuracy	AUC	Sensitivity	Specificity	Sum
0	0.863171	0.927089	0.050147	1.000000	2.840408
1	0.860435	0.916580	0.041298	0.997738	2.816051
2	0.856594	0.921437	0.044379	1.000000	2.822410
3	0.856594	0.916869	0.071006	0.995485	2.839955
4	0.860435	0.922406	0.035503	0.995485	2.813829

I have chosen to continue the pipeline with the random forest as the machine learning algorithm.The lipinski rule of 5 was applied to filter out the compounds.During this step the compounds was limited to 272 .Lipinski rule of 5 is important for the bioavailability of the drug.The rule only approves drugs which have a molecular weight less than 500,no more than 5 donor atoms and 10 acceptor atoms and has a log p less than 5.The next step is creating a dissimilarity matrix for the comparison. Tanimoto matrix was chosen because of its scale invariance,robustness to noise and computational efficiency. The matrix created was of shape 272*272. The next step in the pipeline is to cluster the compounds using butina clustering. Butina clustering clusters the structures according to their chemical similarity.A cutoff of 0.825 was used to cluster the compounds into 20 different clusters.Since we need different and diverse compounds , the first element in all the 20 clusters were taken as the final compounds

Table 2:

Methods for filtering	Total Compounds(1051)
forest ML model	403
Lipinski rule of 5	272
clustering(butina)	20
Docking	3

These are the selected compounds after screening.

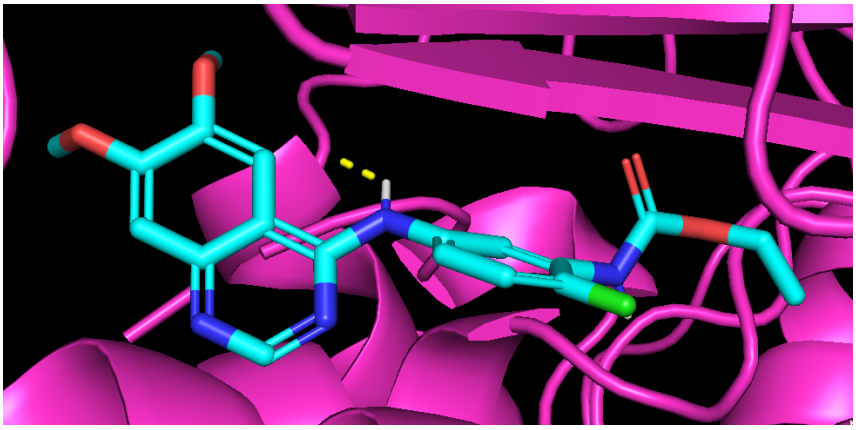


Docking pipeline

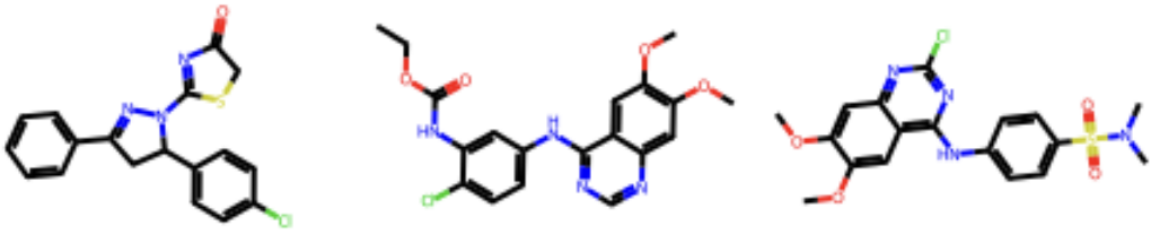
Using the help of pymol the ligand and protein was separated.Open babel was used to convert the molecules into pdbqt. Autodock vina was used to dock the protein with its ligand.

Table 3:This is the result of redocking using smina.

mode	affinity	RMSD
1	-14.6	0.00CD
2	-14.2	1.935
3	-12.0	14.511



Img1:This is the best pose of the selected protein(compund 2) in the compound(RMSD=2.183)



Img 2:Top 3 selected compound

Compound1:C1(=[C]C(=NN1C1=NC(=O)[C]S1)C1=[C][C]=[C][C]=[C]1)C1=[C][C]=C(Cl)[C]=[C]1

Compound2:N(C1=C2C(=N[C]=N1)[C]=C(C(=[C]2)O[C])O[C])C1=[C][C]=C(C(=[C]1)NC(=O)O[C][C])Cl

Compound3:C1(=C([C]=c2c(=[C]1)nc(Cl)nc2NC1=[C][C]=C([C]=[C]1)S(=O)(=O)N([C])[C])O[C])O[C]

The autodock vina program generated 3 conformations of the 20 compounds which equals to about 60 poses.The virtual screening took place in pymol and poseview by inspecting the ligands individually .The various interactions were also the main deciding factors in choosing the compounds.It s noted that structures which were similar to the eye had rmsd larger rmsd values.

Unwanted Compound properties

The potential off-targets were found using swiss target prediction and EGFR was excluded from the list of potential Off-targets

compound	Potential Off-targets	Uniprot ID	Target kinase
Compound 1	CDK2 LRRK2 PTGS1 PTGS2 GSK3B	P24941 Q5S007 P23219 P35354 P49841	Kinase Kinase Oxidoreductase Oxidoreductase Kinase
Compound 2	KDR SRC RET MAPK14 CDK1	P35968 P12931 P07949 Q16539 P06493	Kinase Kinase Kinase Kinase Other cytosolic protein
Compound 3	KDR BRAF TNNI3K PIK3R1 SRC	P35968 P15056 Q59H18 P27986 P12931	Kinase Kinase Kinase Enzyme Kinase

Table 4:The top 5 potential off- targets in the three compounds

To look at the ADMETox point of view we used Swiss ADME and eMolTox.When we ran the Swiss Adme, the bioavailability was 0.55 for all the three compounds ,which is ideal.Most of the other criterias were also fulfilled. The toxicity could not be validated as eMolTox was not able to find any models