

PRINCIPAL COMPONENT ANALYSIS AND QSAR MODELING OF SELECTED LIGAND AGAINST PROSTATE CANCER PROTEIN TARGET

Group:

MARADESA Adeleke	20724523
OGEDENGBE Ikeoluwa Ireoluwa	20724157

Abstract

In this research, we queried ChEMBL database and subjected some selected Ligands to virtual screening. We found four most stable Ligands with very active bioactivities on protein target. The Python RDKit was used to obtain values of molecular descriptors of those Ligands and QSAR models are built to predict inhibitory concentration (PIC50) and form of binding affinity of the Ligand using Random Forest (RF) regression and logistic regression, respectively. Principal component analysis (PCA) and t-SNE were used for descriptive analysis. Only two principal components explained 90% of the variation in the bioactivity data. From PCA and t-SNE, only Four Ligands were identified to be very active and stable, respectively. From RF- based QSAR, $R^2 = 0.802$, shows that about 80.2 % of factors affecting PIC50 are explained by molecular descriptors and logistic regression based QSAR show 100% classification accuracy. The result of molecular docking shows high binding affinity for M2 and M4 Ligands against prostate cancer protein receptor; these two Ligands are required for further therapeutic studies.

1 Introduction

Bioinformatic involves the application of mathematics, statistics, and computing in explaining biological information. This approach is based on employing computer-aided analyses, prediction of structural activities, homology modeling, protein-ligand docking, protein-protein docking, mutagenicity to explain the properties of ligand (drug-likeness), computational drug designing and predicting the inhibitory concentration of drug against different selected diseases [1]. In many areas of medicine, pharmacology and medicinal chemistry determining the binding affinity of drug against protein target is imperative in deciding the usability of the drug and the mode in which the drug can be administered. In bioinformatics, drug design helps to discover the potency of novel molecule. Many bioinformatics tools have been developed to study the properties, dissolvability and the bioactivities(such as binding affinity) of drug candidates and computer aided drug design

(CADD) has been widely used to study ligand-receptor interaction, drug-drug interaction in order to reduce cost and save time used in experimentation before arriving at new drug candidate[2].

In high throughput screening (HHT), the CADD has been praised for its high efficiency in reducing the number of ligands to be screened while maintaining the same level of lead molecule discovery. Some of the molecule will be predicted as being active while inactive ones are screened out. The processes involved in designing novel drug is very complex and challenging as a lot of time and money are normally wasted; this process is accelerated with the aid of CADD and other bioinformatic tools in order to develop new compound in quest for efficient therapeutic agents[3]

Also, when a novel drug candidate is identified, its bioactivities need to be properly understood and studied to determine the potency of the drug against certain protein-receptor, its dissolvability, and the mode in which it should be administered will be better revealed. In this research, prostate cancer protein receptor is used as the target. The binding affinity of some selected Ligand (drug candidate) is studied, and their molecular descriptors are investigated and QSAR modeling, which reveals the correlation between chemical structures and activities (Quantitative structure-activity relationship), is used to establish a model for predicting inhibitory concentration (IC₅₀) of the selected Ligands.

2 Data Collection and Methodology

The data used in this research work was extracted from ChEMBL. It is a database contains over 2.1 million bioactive molecules with drug-like properties. It comprises of chemical, genomic, and bioactive data which aids transformation of genomic informatic to stable novel drugs. We queried ChEMBL for some Ligands whose properties (drug-likeness) show high inhibitory activities on protein-receptor (prostate cancer). We extracted 29 different Ligands and calculated their molecular descriptors (Molecular Weight, LogP, Hydrogen Bond Donor (HBD), Hydrogen Bond Acceptor(HBA), and Number of Rotatable Bond) using Python RDKit. The data was saved to dataframe(.CSV). We used the data for QSAR modeling, PCA and t-SNE were used for descriptive analysis. Based on PCA, we identify important molecular descriptors, proportion of explained variation and the statistical significance of the descriptors. QSAR modeling is a mathematical model which established relations between molecular and inhibitory concentration (IC₅₀). Molecular docking was performed using those Ligand with lower predictive errors and the binding affinity of the selected Ligands with protein-receptor are investigated.

To perform molecular docking using the Ligands with the smallest MSE. The ligands will be used against the selected protein target where the binding affinity of the Ligands with the protein target can be obtained. The prostate cancer protein target was obtained by quarrying the protein data bank (PDB) database.

2 Results

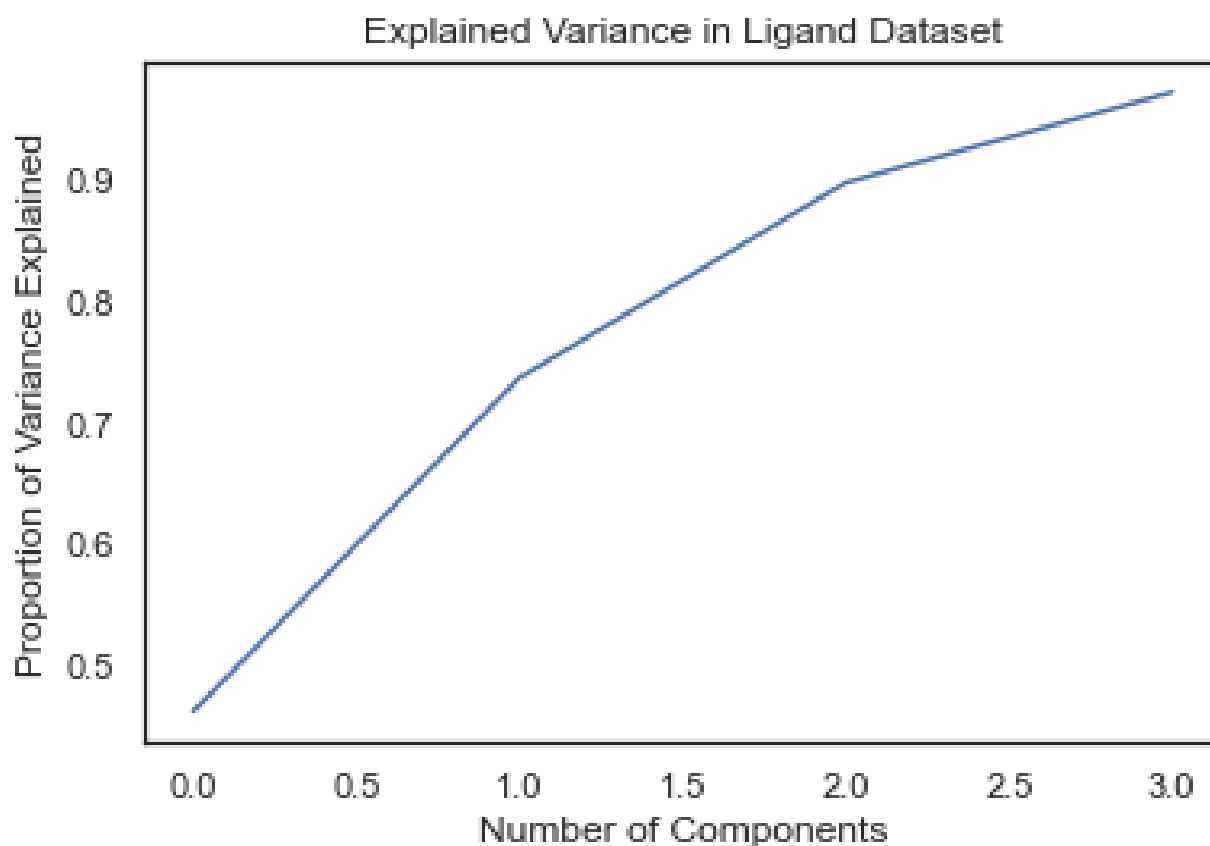


Fig 1: The Principal Component Analysis

From Fig 1 above, we deduce that the proportion of variance explained increases with components and only first two principal components have explained about 90% of the variability.

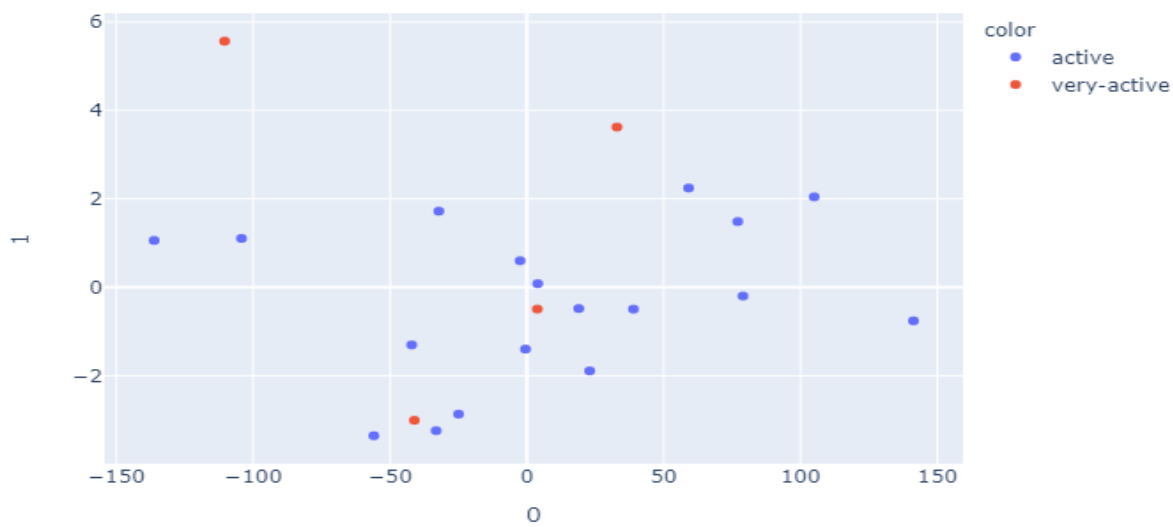


Fig 2: Bioactivities of the selected Ligands

From the Fig 1 above, we discovered that only four Ligands are very active and they are viewed as better drug candidate. This is explained by the first two principal components where over 90% of the variability are explained. The PC1 and PC2 explained 46.194196% and 73.716187% of the variation respectively.

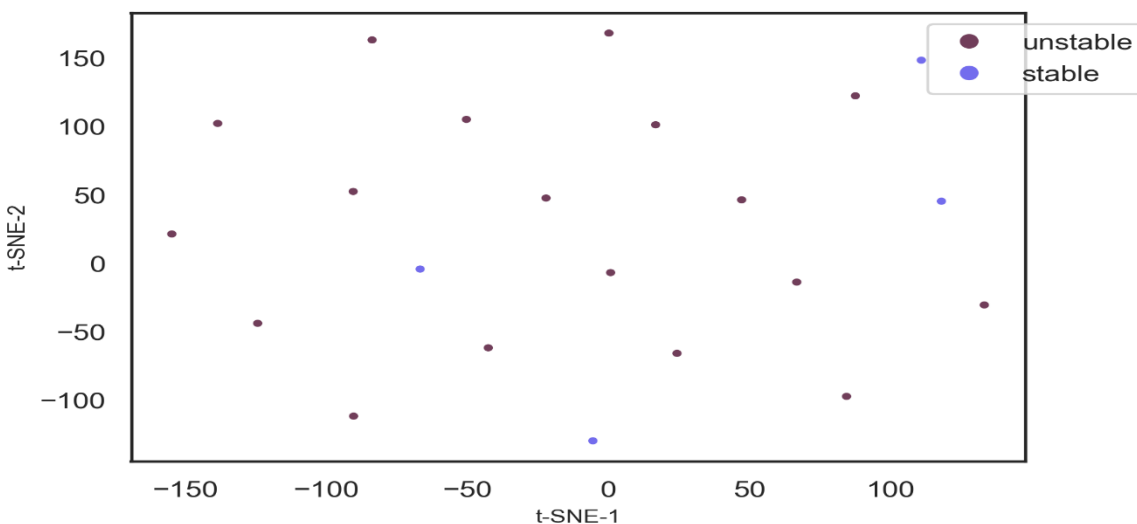


Fig 3: Bioactivities of the selected Ligands (t-SNE)

From the Fig 3 above, we discovered that only four Ligands are stable and can be viewed to be better drug candidate due to their stable form. This is explained by the first two components where over 90% of the variability are explained.

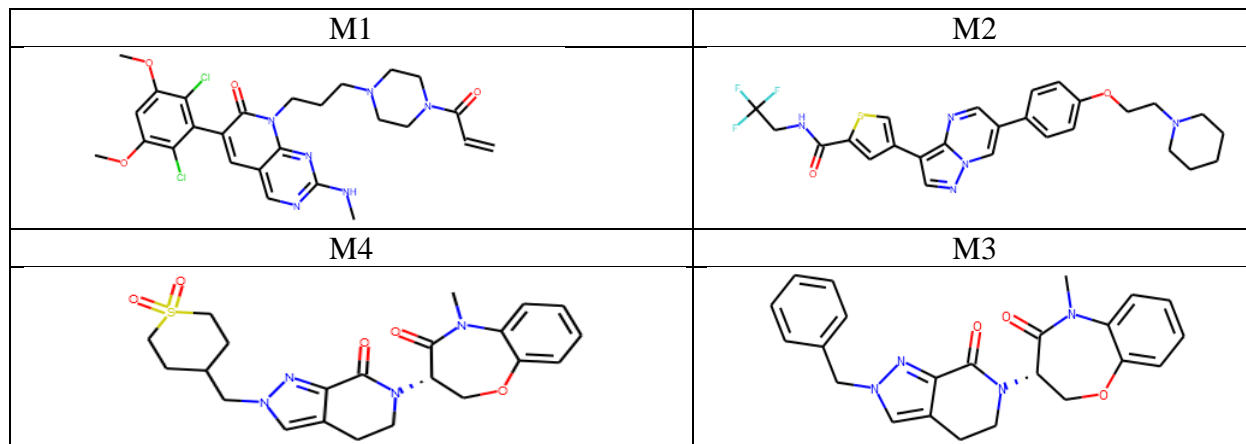


Fig 4: The Stable and Very-active Ligands

The Fig 4 above shows the Ligand screened out using MSE. All of those Ligands whose mean square errors are less than 0.05 are very active compound.

3 Quantitative Structure-activity relationship (QSAR) Modeling

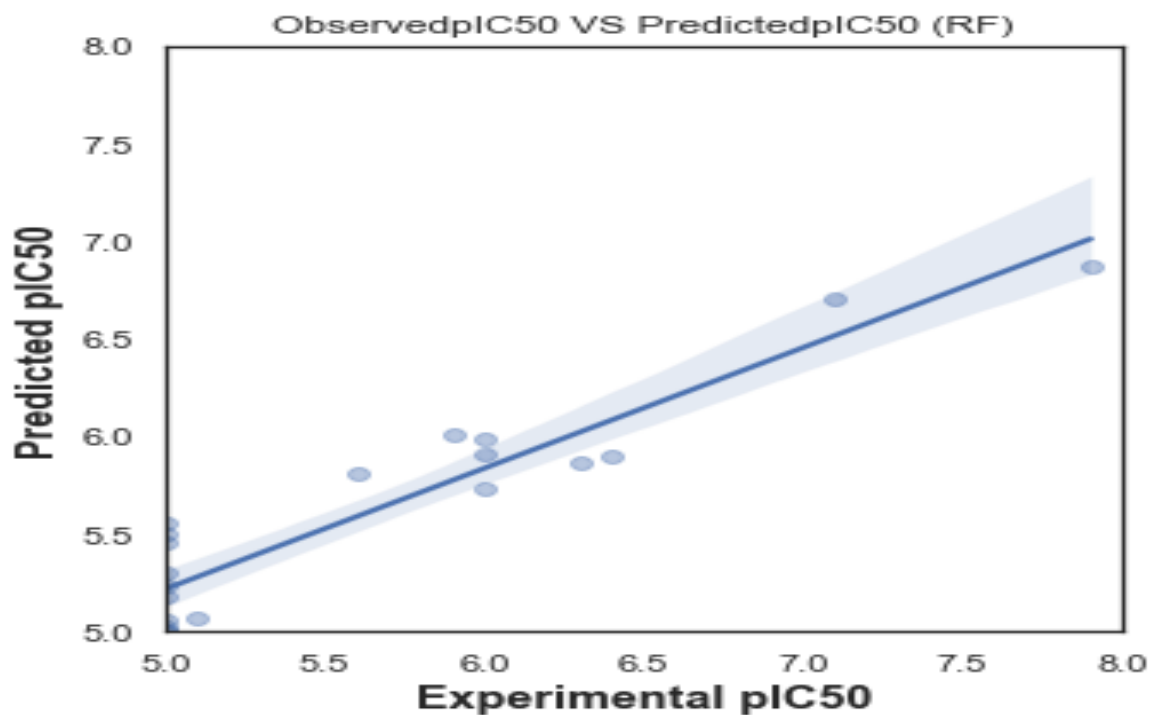


Fig 4: The Random forest Regression (Predicted IC50 VS experimental IC50)

The Fig 5 show predicted inhibitory concentration against the experimental; this shows the regression trendline of the model. The coefficient of determination $R^2 = 0.802$ shows that about 80.2% of the variation are explain. *i.e* about 80.2% of the factors affection the inhibitory concentration of the ligands can be explained by their molecular weight (MW), octanol water partition(LogP), hydrogen bond donor (HBD), hydrogen bond acceptor (HBA) and number of rotatable bond (RTB).

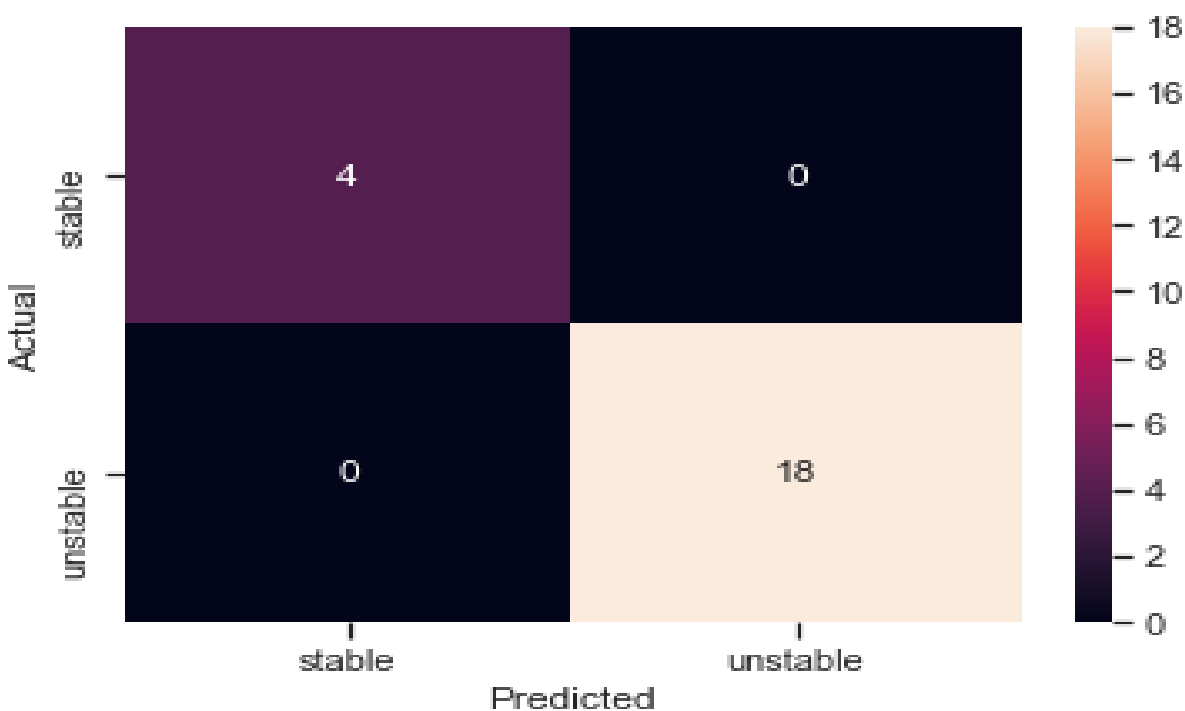
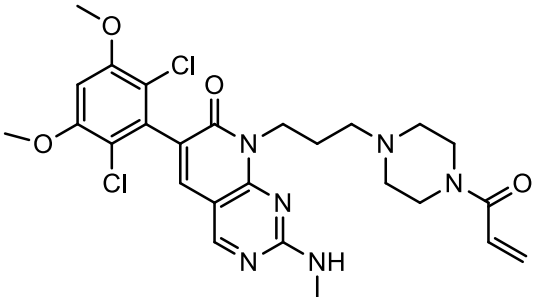
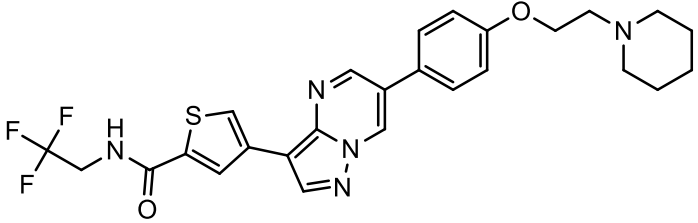
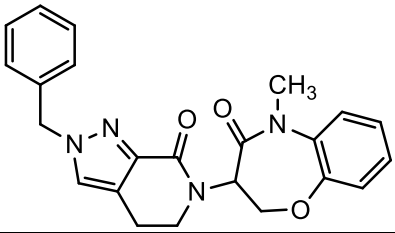
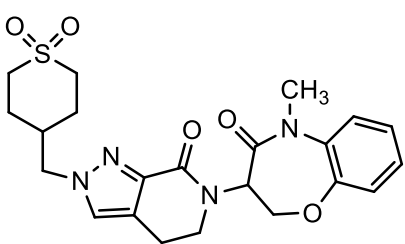


Fig 5: Confusion matrix

The Fig 6 above shows the predictive accuracy of Logistic regression. We deduce that the regression predicted the binding affinity accurately. The true positive (TP) is 4, true negative (TN) is 18, false positive (FP) and false negative (FN) are zero respective. Therefore, the predictive accuracy $y = \frac{TP+TN}{Total} = \frac{22}{22} = 100\%$. We say logistic regression provides excellent fit, and its predictive results are reliable in deciding the binding affinity of the Ligand. From this result, four Ligands are also predicted to be stable. The identified ligands will be used in molecular docking.

3 Molecular Docking of Ligand with Prostate cancer Receptor

Table 1: The Binding Affinity of the Ligands

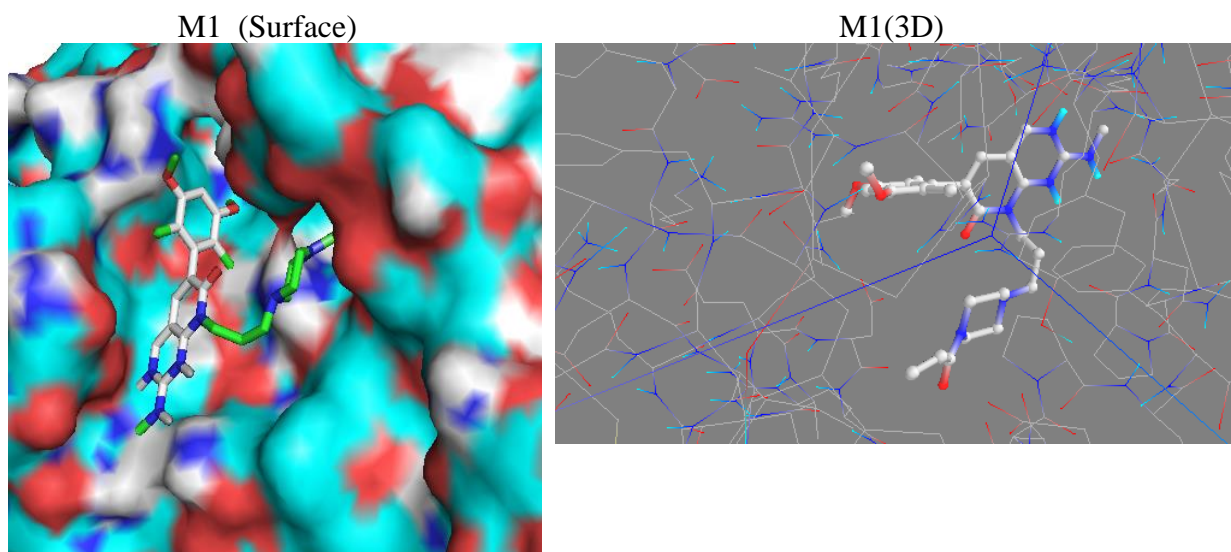
Ligand	Affinity (kcal/mol)	Hydrogen Bond Interaction
<p>M1</p> 	-7.5	TYR-425 ASP-424, Leu-448
<p>M2</p> 	-8.3	GLN-155, LEU-154, LEU-152
<p>M3</p> 	-7.6	TYR-367, ASP-424
<p>M4</p> 	-8.5	ASN2-224, PHE-265 PHE-277, ALA-164,

Docking Studies

In the quest for determining the biological activities of drug candidate with high inhibitory activities on androgen receptor (AR), ChEMBL was queried to select the Ligands whose inhibitory concentration shows drug-likeness. Four of the selected Ligands show better stability against androg

en receptor (AR) with the PDB (Protein Data bank) code 5KJ2. PYRX software was used for virtual screening. The binding affinity of the ligands to the binding pocket of this receptor was shown in the table 1 above. During docking nine different docking modes were generated and this is an indicative of different conformations of the Ligands at the active site of the protein Receptor(5KJ2). The lowest binding affinity values have been proved to be the best binding pocket of the Ligands.

we could observe that ligand encoded as M4 has the least affinity of -8.5kcal/mol followed by M2 (-8.3kcal/mol), M3 (-7.6kcal/mol) and M1 (-7.3kcal/mol). Therefore, Ligand M4 has higher potential to be used as a chemotherapeutic agent against prostate cancer. This can be further validated with clinical trials. Hydrogen bond interaction plays important role in structure-based drug design (SBDD). This also determines the energetic stability of ligands when bind to the active site of the protein receptor. Three hydrogen bonding was formed between M1 and 5KJ2 and this occurs between the residue Tyr-425, Asp-424 and Leu-448. Between M2 and 5KJ2; Gln-155, Leu-154 and Leu-152 were the residues forming hydrogen. We have Tyr-367, and Asp-424 responsible for the hydrogen bonding forming between M3 and 5KJ2 and between M4 and the receptor, we have Asn-224, Phe-265, Phe-277 and Ala-164 as the residues at the binding site. The docking result visualization (3D and surface representation) were also shown in Figure 6 below. The surface representation shows how each of the ligands fit to the active site of the protein. We could observe that the ligand M4 fit in to the binding pocket as compared to other ligands; hence, Ligand M4 has proved to be a better drug candidate for the treatment of prostate cancer. Clinical trials are strongly encouraged for Ligand M2 and M4.



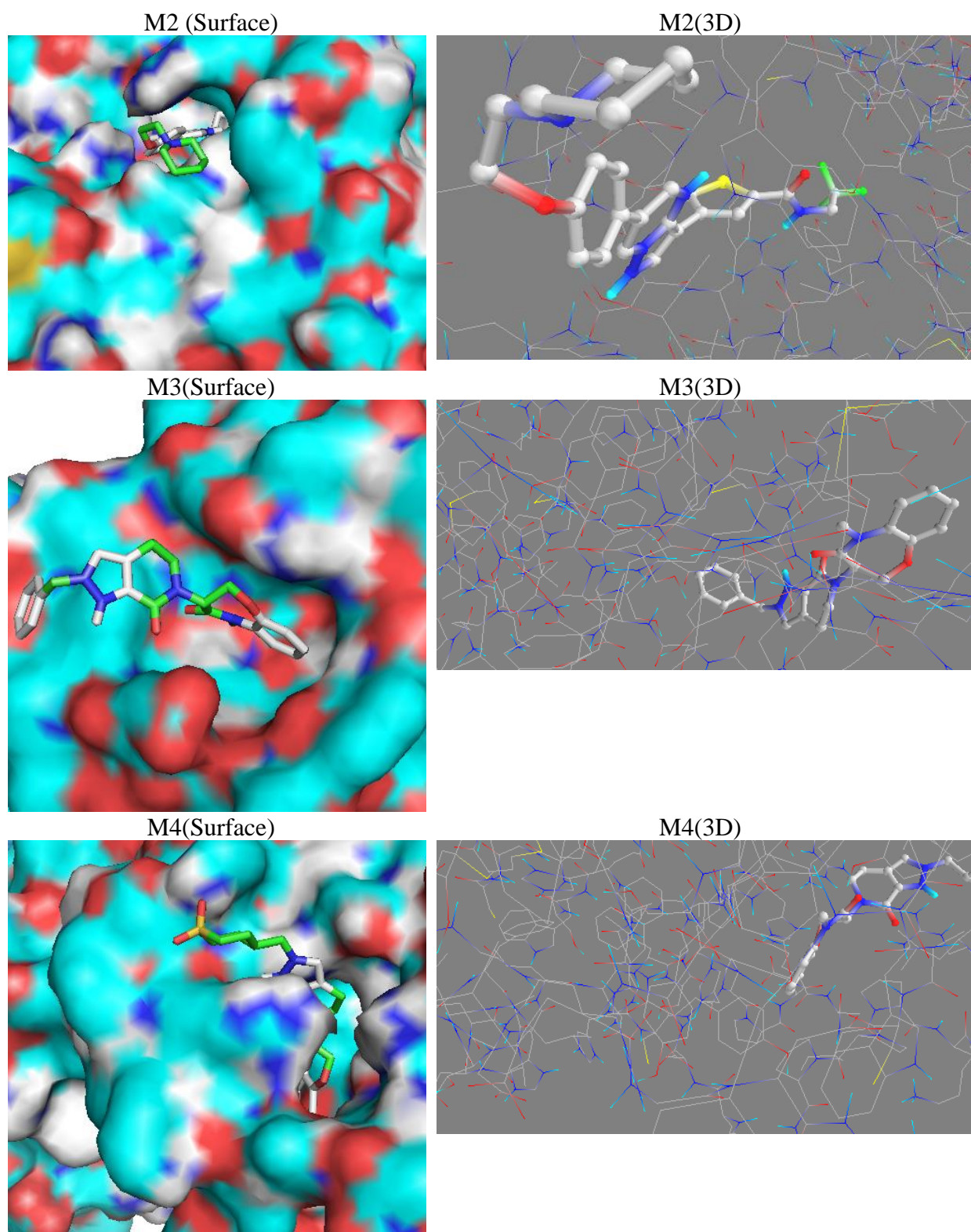


Fig 6: The Molecular docking with binding site

From Fig 6 above, we can conclude that there is effective binding in the interaction between Ligand and the receptor for M2 and M4. The Ligand M2 and M4 is therefore recommended for clinical trial and further therapeutic studies

4 Conclusion

Tumor metastasis has been area of focus in bioinformatics and medicinal chemistry for so long . In fact, many interventional therapies have been used to find better drug candidate for interventional therapy. This research has suggested two different Ligands as better drug candidate because of their high binding affinity and bioactivities on protein target (Prostate cancer). The results of this research work will be useful in interventional therapy.M2 and M4 are recommended for further studies.

INDIVIDUAL CONTRIBUTION

NAME	Problem Development	Coding Python	Docking (PYRX)	Report	Presentation
MARADESA Adeleke	✓	✓		✓	
OGEDENGBE Ikeoluwa Ireoluwa	✓		✓	✓	✓

Reference

- [1] Y. Dali, S. M. Abbasi, S. A. F. Khan, A. Larra, and R. Rasool, "Computational drug design and exploration of potent phytochemicals against cancer through in silico approaches," *Biomed. Lett.*, p. 6.
- [2] G. Sliwoski, S. Kothiwale, J. Meiler, and E. W. Lowe, "Computational Methods in Drug Discovery," *Pharmacol. Rev.*, vol. 66, no. 1, pp. 334–395, Jan. 2014, doi: 10.1124/pr.112.007336.
- [3] N. A. Meanwell *et al.*, "Inhibitors of HIV-1 Attachment: The Discovery and Development of Temsavir and its Prodrug Fostemsavir," *J. Med. Chem.*, vol. 61, no. 1, pp. 62–80, Jan. 2018, doi: 10.1021/acs.jmedchem.7b01337.

Appendix

Table 2: The Ligands used in Molecular Docking (the most active and stable binding affinity)

Ligands	IUPAC
M1	8-(3-(4-acryloylpiperazin-1-yl)propyl)-6-(2,6-dichloro-3,5-dimethoxyphenyl)-2-(methylamino)pyrido[2,3-d]pyrimidin-7(8H)-one
M2	4-(6-(4-(2-(piperidin-1-yl)ethoxy)phenyl)pyrazolo[1,5-a]pyrimidin-3-yl)-N-(2,2,2-trifluoroethyl)thiophene-2-carboxamide
M3	3-(2-benzyl-7-oxo-4,5-dihydro-2H-pyrazolo[3,4-c]pyridin-6(7H)-yl)-5-methyl-2,3-dihydrobenzo[b][1,4]oxazepin-4(5H)-one
M4	3-(2-((1,1-dioxidotetrahydro-2H-thiopyran-4-yl)methyl)-7-oxo-4,5-dihydro-2H-pyrazolo[3,4-c]pyridin-6(7H)-yl)-5-methyl-2,3-dihydrobenzo[b][1,4]oxazepin-4(5H)-one

Table 3: Virtual Screenings

ChEMBL_id	Canonical_Smiles	LogP	MolWt	NRTB	HB D	HB A
CHEMBL2205766	<chem>CC(C)(C)NS(=O)(=O)c1cncc(-c2ccn3nc(N)nc3c2)c1</chem>	1.45030	346.416	3.0	2.0	7.0
CHEMBL3741589	<chem>O=C(O)C1CN(Cc2ccc(OCc3ccc(Cl)c(Cl)c3)cc2)C1</chem>	4.08880	366.244	6.0	1.0	3.0
CHEMBL3745885	<chem>Cn1c(=O)c(S(=O)(=O)c2ccc(F)cc2F)cc2cnc(Nc3ccc4...</chem>	3.66440	467.457	4.0	2.0	7.0
CHEMBL3884319	<chem>CC1(C)C(=O)N([C@H]2CCc3c(O)cccc32)c2nc(Nc3ccc...</chem>	4.23750	386.455	3.0	2.0	5.0
CHEMBL3903725	<chem>Nc1ccc(-c2ccc3ncc4c(=O)[nH]c(=O)n(-c5cccc(C(F)...</chem>	3.28520	450.380	2.0	2.0	7.0
CHEMBL2006765	<chem>CCCN(C(=O)c1ccc(Nc2nc(NCC(F)(F)F)c3cc[nH]c3n2)cc1</chem>	3.81550	392.385	7.0	4.0	5.0
CHEMBL4078893	<chem>FC(F)(F)CNc1nc(Nc2ccc(N3CCOCC3)c(Cl)c2)nc2ccoc12</chem>	4.43060	427.814	5.0	2.0	7.0
CHEMBL4066664	<chem>FC(F)(F)CNc1nc(Nc2ccc3cn[nH]c3c2)nc2ccoc12</chem>	3.81690	348.288	4.0	3.0	6.0
CHEMBL4069365	<chem>CC1(C)OCC(=O)Nc2cc(Nc3nc(NCC(F)(F)F)c4occc4n3)...</chem>	4.14440	421.379	4.0	3.0	7.0
CHEMBL4062803	<chem>Cc1n[nH]c2cc(Nc3nc(NC4CC4)c4occc4n3)ccc12</chem>	3.72542	320.356	4.0	3.0	6.0
CHEMBL4068509	<chem>C=CC(=O)N1CCN(CCCn2c(=O)c(-c3c(Cl)c(OC)cc(OC)c...</chem>	3.54450	561.470	9.0	1.0	9.0
CHEMBL4063500	<chem>CN(C)C(=O)c1cc2cnc(Nc3ccc(NC(=O)CCCCCCC(=O)NO)...</chem>	4.77630	535.649	12.0	4.0	8.0

CHEMBL4276 946	CC(=O)n1nc(C)c(-c2ccc(Cl)c(Cl)c2)c1N	3.407 62	284.1 46	1.0	1.0	4.0
CHEMBL3889 78	CN[C@@H]1C[C@H]2O[C@@](C)([C@@H]1OC)n1 c3ccccc3...	4.354 00	466.5 41	2.0	2.0	6.0
CHEMBL4438 748	CN(C)c1ccc(C(=O)Nc2cccc(NC(=O)COc3ccc4c(=O)c co...	4.128 80	457.4 86	7.0	2.0	6.0
CHEMBL4569 508	O=C(NCC(F)(F)F)c1cc(-c2cnn3cc(- c4ccc(OCCN5CCCC...	5.281 60	529.5 88	8.0	1.0	7.0
CHEMBL4550 702	Cn1cc(-c2cnc3c(-c4csc(C(=O)NCC(F)(F)F)c4)cnn3c..	3.150 40	406.3 93	4.0	1.0	7.0
CHEMBL4568 087	Cn1cc(-c2cnc3c(- c4csc(C(=O)N[C@@H]5CCCC[C@@H]5...	2.858 00	421.5 30	4.0	2.0	8.0
CHEMBL4552 628	Cc1sc(C(=O)N[C@@H]2[C@H](N)CCCC2(F)F)cc1- c1cnn...	3.664 52	425.8 92	3.0	2.0	6.0
CHEMBL4549 667	CN1C(=O)[C@@H](N2CCc3c(nn(Cc4cccc4)c3Br)C 2=O)...	3.116 20	481.3 50	3.0	0.0	5.0
CHEMBL4088 216	CN1C(=O)[C@@H](N2CCc3cn(Cc4cccc4)nc3C2=O)COc2...	2.353 70	402.4 54	3.0	0.0	5.0
CHEMBL4097 778	CN1C(=O)[C@@H](N2CCc3cn(CC4CCS(=O)(=O)CC 4)nc3C...	1.130 20	458.5 40	3.0	0.0	7.0

bioactivities	Pic50	Predictedpic50	MSE	Binding_affinity
active	5.0	5.147	0.719104	unstable
active	5.0	5.053	0.719104	unstable
active	5.0	5.369	0.719104	unstable
active	5.0	5.118	0.719104	unstable
active	5.0	5.221	0.719104	unstable
very-active	6.3	5.848	0.204304	unstable
active	5.0	5.529	0.719104	unstable
active	5.0	5.001	0.719104	unstable
active	5.1	5.084	0.559504	unstable
active	5.0	5.013	0.719104	unstable
active	5.9	5.876	0.002704	stable
very-active	7.1	6.687	1.567504	unstable
active	5.0	5.118	0.719104	unstable
very-active	7.9	6.744	3.210704	unstable
active	5.0	5.325	0.719104	unstable
active	5.6	5.893	0.061504	unstable
active	4.6	4.843	1.557504	unstable
active	6.4	5.811	0.304704	unstable
active	4.9	4.941	0.898704	unstable
active	6.0	5.916	0.023104	stable
active	6.0	5.706	0.023104	stable
active	6.0	5.848	0.023104	stable