WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Identifying Barley Phytochemicals as Potential Inhibitors of EGFR in Triple-Negative Breast Cancer using Molecular Docking and QSAR

Muhammad Abdur Rehman [a]

[a] *Atta-ur-Rahman School of Applied Biosciences, National University of Sciences and Technology (NUST), Islamabad, Pakistan*

DOI:https://doi.org/10.55248/gengpi.5.1024.2817

ABSTRACT

Herbal medicine has been used for centuries to treat a variety of diseases predominantly in the eastern world. Although significant potential exists in the therapeutic activities of phytochemicals, more research is required to develop plant-based drugs. We aim to identify barley (*Hordeum vulgare*) phytochemicals with the highest binding scores against overexpressed EGFR in aggressive breast cancer subtypes, such as triple-negative breast cancer through molecular docking. Then, we develop a simple reusable QSAR model to predict the binding affinity of phytochemicals. Due to the small size of the dataset, our QSAR model showed moderate predictive power. The phytochemicals with the highest binding affinity were all flavonoids (orientin, saponarin, proanthocyanidin, and cianidanol). Flavonoids are already being studied to explore their anti-cancer properties, which further confirms our findings.

Keywords: Computational drug discovery, *Hordeum vulgare*, barley, breast cancer, human epidermal growth factor and receptor, EGFR, molecular docking, machine learning, quantitative structure-activity relationship (QSAR)

1. Introduction:

1.1 Hordeum vulgare:

Belonging to the family of Poaceae, barley (*Hordeum vulgare*) is one of the first cultivated cereal grains grown widely around the world (12% of all cereal crops) (Cassman, 1999). Barley can show resilience to ecological strains such as drought and extreme cold, and possess a faster maturity rate as compared to other crops making it cost-effective to grow (Ahmed Sallam, 2019) . In addition to a rich nutritional profile, barley contains a huge therapeutic and pharmaceutical potential due to the presence of phytochemicals, which are known to have anticarcinogenic effects and prevent the onset of chronic illnesses such as high blood pressure, heart disease, and colon cancer (Kuldeep Singh, 2023). Barley has been used as herbal medicine in traditional Chinese medicine for gastrointestinal issues and complications related to skin, blood, liver, etc., since 2000 BCE (Yu Zhou, 2022). It is also a major component of the Indian Ayurveda system of medicine.

1.2 Breast cancer:

Breast cancer is the most common cause of cancer-related deaths among women across the world. Accounting for 36% of all cancers, breast cancer has shown the highest mortality rate of 315309 in Asia reported in 2022, according to the International Agency for Research on Cancer (Beata Smolarz, 2022) . Breast cancer can be categorized into subtypes, based on the molecular biomarkers, such as hormone receptor-positive, HER2-positive, and triple-negative breast cancer (TNBC). Human epidermal growth factor and receptor (EGFR, also known as ErbB1 and HER1) is a tyrosine kinase transmembrane receptor responsible for regulating cellular processes such as cell proliferation, differentiation, and survival. Dysfunction or dysregulation of EGFR, associated with tumor growth and metastasis, is found in aggressive cancer subtypes, particularly TNBC and inflammatory breast cancer (IBC) (Hiroko Masuda, 2012) . Metaplastic breast cancer, which is a specific expression of TNBC shows 25% of EGFR gene amplifications resulting in overexpression of the receptor (Khawla Al-Kuraya, 2004).

1.3 Computer-aided drug design:

Since breast cancer is a heterogeneous disease, treating it is a complex phenomenon as each subtype requires specific treatment and therapies. This demands the development of novel drugs and therapies against specific cancer biomarkers which include the integration of advanced technology in the drug development process. Computer-aided drug discovery employs software, algorithms, and computational techniques to screen large libraries of compounds and identify potential drug candidates for further testing. Two computational approaches that we will use in this research are explained below.

1.3.1 Molecular docking:

Molecular docking is an important part of the computational drug discovery process. This approach is a structure-based method employed to identify the most reliable molecular interactions based on binding affinity and conformation of the molecules (Jiyu Fan, 2019). Molecular docking techniques are being widely used to identify optimal interactions between potential drug candidates and the target molecules. Various software are employed to perform molecular docking such as MolDock, AutoDock Vina, PyRx, etc.

1.3.2 Quantitative structure-activity relationship (QSAR):

Machine learning-based algorithms have a diverse role in virtual screening, target discovery, lead compound discovery, and protein-ligand interactions in drug discovery (Lauv Patel, 2020). A quantitative structure-activity relationship (QSAR) model utilizes one or more machine learning algorithms to create a mathematical or statistical relationship between the input and output variables of a dataset and make predictions on the unknown dataset.

1.4 Research workflow:

In this research, we perform molecular docking of phytochemicals found in barley with human EGFR and visualize the ligand-protein interactions. Afterward, we use the binding affinity data obtained from molecular docking and create a reusable QSAR pipeline to predict the binding affinities of compounds. We finally check the accuracy of our model and conclude our research by identifying top phytochemicals that show high binding affinity with the target to be further considered in the drug discovery process against breast cancer subtypes that contain dysfunctional EGFR.

2. Steps in molecular docking:

The steps used in performing and visualizing molecular docking of barley phytochemicals with human epidermal growth factor and receptor (EGFR) are listed below.

2.1 Phytochemical data retrieval:

A list of 50 phytochemicals present in various parts of barley (fruit, leaf, and seed) was generated with the help of Indian Medicinal Plants, Phytochemistry and Therapeutics (IMPPAT), which is a curated database of 17967 phytochemicals discovered in Indian medicinal plants till date (Mohanraj, 2018) (R.P. Vivek-Ananth, 2023). These phytochemicals will be used as ligands for the target protein in further steps.

2.2 Ligand structure retrieval:

PubChem is the largest database for chemical information on 119M compounds [accessed on: 30th September 2024] (Kim, 2023). This database was used to download 3D conformer structures of all 50 compounds in SDF format.

2.3 Protein structure retrieval:

Protein Data Bank (PDB) is a database containing 3D structures of biomolecules such as proteins, nucleic acids, etc. Using this database, we obtained “Crystal Structure of the Complex of Human Epidermal Growth Factor and Receptor Extracellular Domains (1IVO),” the target protein 3D structure as shown in Figure 1 (Ogiso, 2023).

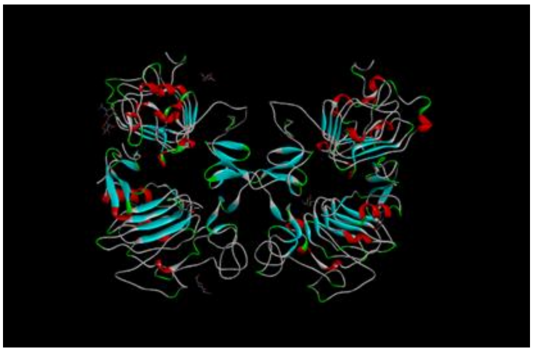

Figure 1: Three-dimensional structure of the complex of human epidermal growth factor and receptor extracellular domains (1IVO) visualized through BIOVIA Discovery Studio

2.4 Protein preparation:

The target protein was prepared using BIOVIA Discovery Studio, which provides a graphical interface for protein modeling and structure manipulation. During the preparatory steps, all the water molecules and heteroatoms, atoms that do not belong to the protein primary chain, were removed from the protein until we were left with chain A, protein groups, and ligand groups as shown in Figure 2. After the addition of polar hydrogens, the file was saved as a PDB file.

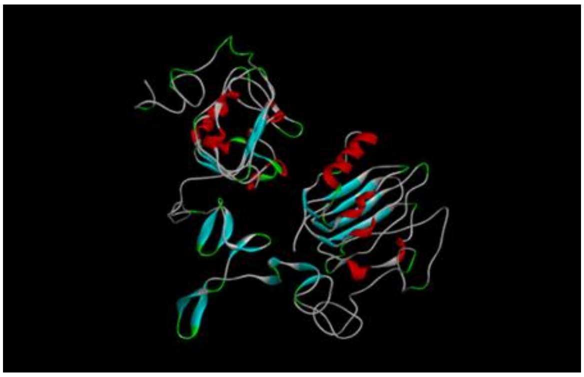

Figure 2: Prepared target protein structure visualized with BIOVIA Discovery Studio

2.5 Performing molecular docking:

For molecular docking, we required software that has a graphical interface and can run multiple compounds against the target simultaneously. PyRx is a virtual screening software that helps computational drug discovery scientists screen multiple compounds together against a target. It also provides a user interface, instead of a command prompt (Dallakyan S, 2015) . The prepared protein molecule was loaded in PyRx to be converted to a macromolecule in PDBQT format. Ligand molecules were inserted using the Open Babel function one by one. Energies of all the ligands were minimized and the ligands were converted to PDBQT format as well. The grid was adjusted to cover the whole protein so the active sites are identified by the software itself. This could yield new binding sites for the ligands. The ligands were divided into subsets of 20 and 30, and molecular docking was performed twice against the same receptor with both subsets. The exhaustiveness score was kept at 8 so the best of 9 conformations is selected. The molecular docking results of 50 compounds against the human epidermal growth factor and receptor (EGFR) were obtained and saved in a CSV file.

2.6 Visualization:

The best conformation model of the top 4 compounds with the highest binding affinity was visualized using BIOVIA Discovery Studio. This was done to identify and analyze chemical interactions between ligands and the active sites of the protein.

3. Quantitative Structure-Activity Relationship (QSAR) Model:

For the QSAR, a linear regression model was employed to predict the binding scores of compounds using the molecular docking results obtained through molecular docking. The choice of linear regression model was because our dataset was small (50 compounds with corresponding binding affinities), so we needed a simple machine learning algorithm. A linear regression model is a simple statistical model that aims to detect the best-fit straight line between independent X and dependent Y variables. The model can be represented as,

$$Y = a + bX$$

where,

- Y = dependent value
- X = independent value
- b = regression parameter (slope of the line)
- a = intercept (Smita Rath, 2020)

For our model, the original dataset contained compound PubChem CID, name, canonical smiles obtained from the ChEMBL database, and binding affinity scores obtained through previous steps of docking.

3.1 Calculation of Lipinski descriptors:

We needed to develop numerical values for compounds so that the model finds a mathematical relationship between the structure (X) and binding affinity of compounds (Y). For all the compounds in the dataset, Lipinski descriptors were calculated using the canonical smiles as input for each entry. Lipinski Rule of Five (RO5) is a set of rules that evaluates the drug-likeness of compounds. According to this rule, the molecular weight of compounds should be less than 500 Daltons, LogP (Octanol-water partition coefficient) should be less than 5, the number of hydrogen bond donors should be less than or equal to 5, and number of hydrogen bond acceptors should be less than or equal to 10 (Christopher A. Lipinski, 1997).

3.2 Data splitting:

The data was split into X and Y variables with a ratio of 80:20. The independent variable (X) contained the molecular descriptor information, while the dependent variable (Y) was composed of the binding affinity of compounds with the target protein EGFR.

3.3 Model initialization and training:

The linear regressor was initialized and trained on the training set, X_train and Y_train. Predictions were made on the test set, X_test. The performance of the model was evaluated using R-squared and mean-squared error.

3.4 Cross-validation:

To check the validity and generalizability of our model, 5-fold cross-validation was done for the dataset. As a result, the model was trained on 4 parts of the data and tested on the remaining 1 part for as long as the possible combinations of the dataset were achieved. The performance metrics for the cross-validation test were noted as mean R-squared score and standard deviation of R-squared score.

The QSAR model can be accessed at: *QSAR to predict binding affinities of compounds*

4. Molecular docking analysis:

Eastern medicine presents a pool of untapped compounds to be developed and used as plant-based drugs. In Table 1, we present our findings on the molecular docking (binding affinity) scores of phytochemicals present in barley, with human epidermal growth factor and receptor (EGFR) that is overexpressed in aggressive cancer subtypes, such as triple-negative breast cancer (TNBC).

Ligand	**Name of the Ligand**	**Binding affinity (kcal/mol)**
5281675_E=415.84	Orientin	-8.3
441381_E=620.94	Saponarin	-8.3
108065_E=551.24	Proanthocyanidin	-8.2
9064_E=204.84	Cianidanol	-8.2
72276_E=230.94	(-)-Epicatechin	-8.2
146798_E=440.35	Procyanidin B3	-8.1
66868_E=974.16	Porphyrin	-8.1
5280441_E=409.01	Vitexin	-8
5280794_E=546.19	Stigmasterol	-8
5280445_E=242.10	Luteolin	-8
442584_E=739.63	Carlinoside	-8
5280666_E=324.01	Chrysoeriol	-7.8
135398658_E=289.04	Folic acid	-7.7
222284_E=590.88	Beta-Sitosterol	-7.6
5379265_E=424.28	5,7-Dihydroxy-3',4',5'-trimethoxyflavone	-7.6
6466_E=827.33	Gibberellic acid	-7.1
493570_E=317.94	Riboflavin	-7

439242_E=688.30	Raffinose	-6.9
16754_E=645.57	Glaucine	-6.9
5281328_E=588.80	Fucosterol	-6.9
10153_E=596.79	Corydine	-6.8
5281416_E=104.17	Esculetin	-6.7
5280691_E=167.68	p-Coumaroylagmatine	-6.6
6857447_E=255.66	beta-Tocopherol	-6.3
5988_E=487.38	Sucrose	-6.1
689043_E=98.60	Caffeic acid	-6.1
185872_E=300.96	Epiheterodendrin	-6
5793_E=184.96	D-Glucose	-5.9
472107_E=317.61	(1H-Indol-3-yl)methanamine	-5.9
6890_E=352.17	Gramine	-5.8
45934203_E=199.96	Sodium pangamate	-5.8
11032594_E=278.77	Osmaronin	-5.8
134774_E=183.95	N-(3-Carboxy-2,3-dihydroxypropyl)-4-((carboxymethyl)amino)threonine	-5.8
54670067_E=200.65	Ascorbic acid	-5.6
171548_E=328.21	Biotin	-5.6
10742_E=110.08	Syringic acid	-5.6
11067153_E=674.11	Mugineic acid	-5.5
5280896_E=381.66	Abscisic acid	-5.5
5610_E=97.63	Tyramine	-5.4
68313_E=134.23	Hordenine	-5.3
6613_E=133.81	Pantothenic acid	-5.3
9727_E=104.89	N-Methyltyramine	-5.2
439944_E=163.76	2-Carboxy-D-arabinitol	-5.1
938_E=58.73	Nicotinic acid	-5.1
445639_E=80.35	Oleic acid	-5
5280934_E=142.21	Linolenic acid	-5
4650_E=31.30	1,4-Benzoquinone	-4.7
31268_E=212.35	Pyrrolidine	-3.7
305_E=103.14	Choline	-3.5
3776_E=27.00	Isopropyl alcohol	-3.2

Table 1: Binding affinity scores in kcal/mol obtained through molecular docking of 50 phytochemicals present in barley against human EGFR. “E” represents the energy of the ligands with specific PubChem CID.

A broad range of binding affinities was observed among the 50 phytochemicals present in the dataset. Compared with others in the list, compounds with the highest binding affinity were orientin, saponarin, proanthocyanidin, and cianidanol, with binding affinity of -8.3, -8.3, -8.2, and -8.2 respectively. These compounds show strong binding affinity through molecular and chemical interactions with the target, suggesting a higher likelihood of potential inhibitory activity.

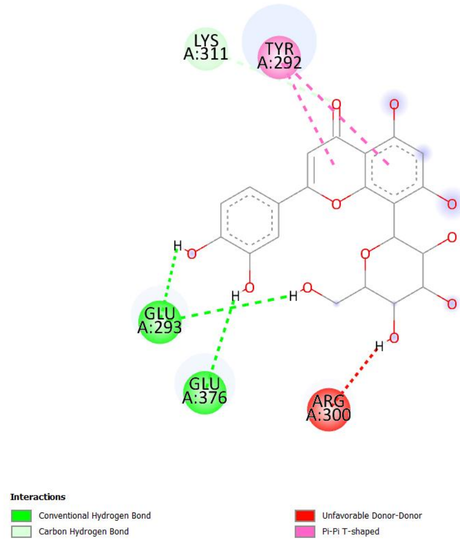

Figure 3

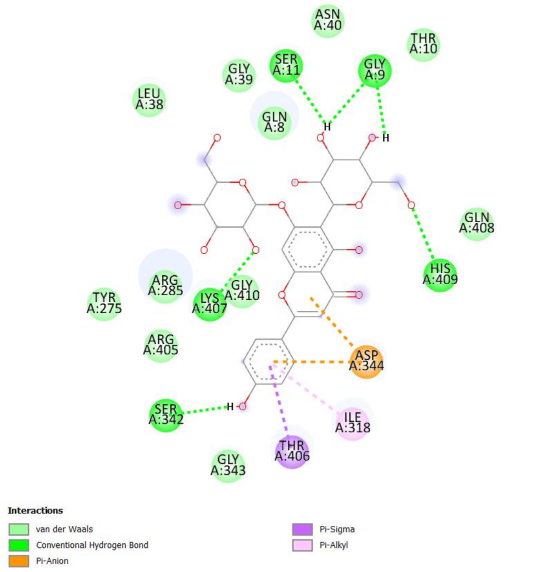

Figure 4

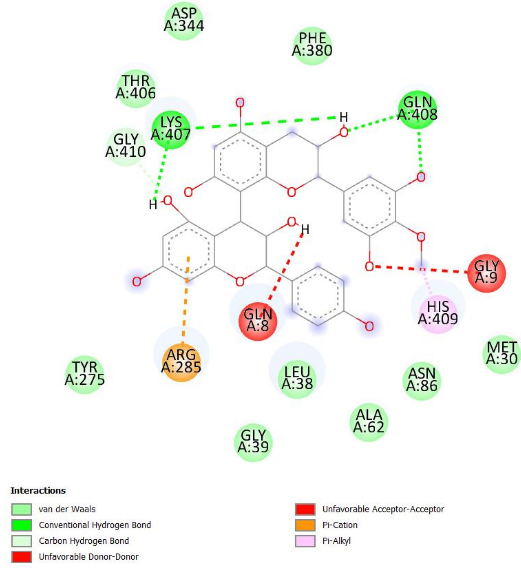

Figure 5

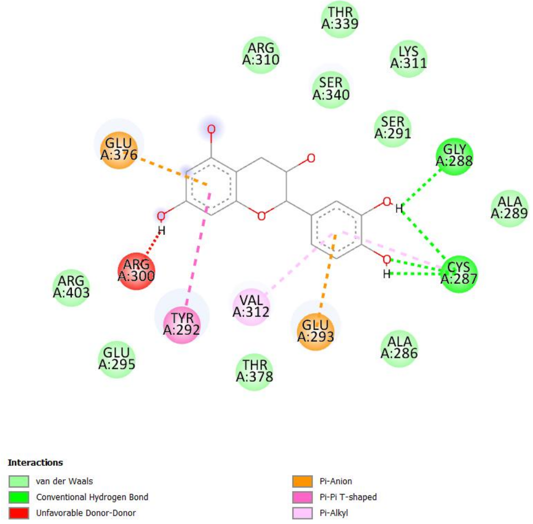

Figure 6

Figure 3-6: Visualization of two-dimensional chemical interactions between orientin, saponarin, proanthocyanidin, and cianidanol with active sites of EGFR (1IVO), respectively.

4.1 Analyzing chemical interactions between EGFR (1IVO) and ligands with the highest binding affinities:

All the ligands with the highest binding affinity scores belong to the class of flavonoid compounds. Two-dimensional interactions of ligands with the highest binding affinities have been shown in Figure 3-6.

Interaction with orientin:

The interactions of orientin (a flavonoid compound) with active sites of EGFR (1IVO) can be seen in Figure 3. Orientin binds with EGFR through two glutamic acid residues, GLU-293 and GLU-376 by favorable hydrogen bonds which are responsible for positioning the ligand within the active sites of the protein (Renxiao Wang, 2003). The Pi-Pi T-shaped interaction between TYR-92 and aromatic rings of orientin enhances the hydrophobic binding and contributes positively to the interaction. An unfavorable bond with ARG-300 is also observed, which could be due to steric repulsions and might cause a little hindrance in the binding of the molecules.

Interaction with saponarin:

Figure 4 represents a good binding of another flavonoid, saponarin with EGFR (1IVO) through multiple hydrogen bonds with serine, glycine, histidine, and lysine residues (SER-342, SER-11, GLY-9, HIS-409, LYS-407), and pi-sigma bond with threonine (THR-406) that stabilizes the aromatic rings in the binding pocket of protein, and no unfavorable bonds. Since hydrogen bonds promote the stability of the ligand-protein interactions, multiple hydrogen bonds with no hindrance-causing bonds increase the chances of higher inhibitory activity and stability of saponarin.

Interaction with proanthocyanidin:

Figure 5 shows the binding of proanthocyanidin (flavonoid polyphenolic compound) with EGFR (1IVO) active sites containing glutamine and lysine residues (GLN-408 and LYS-407) through hydrogen bonds and 2 unfavorable donor-donor and acceptor-acceptor bonds with glutamine and glycine residues (GLN-8 and GLY-9). The figure also shows multiple van der Waals interactions between proanthocyanidin and EGFR. Although weaker than hydrogen bonds, van der Waals interactions promote hydrophobic contact and make sure the ligand is well accommodated in the active site of the protein.

Interaction with cianidanol:

Figure 6 represents cianidanol’s chemical interaction with EGFR (1IVO). Cianidanol (catechin) is a polyphenolic flavonoid compound that binds with EGFR through hydrogen bonds with GLY-288, and CYS-287 residues on active sites of the protein. This interaction also contains multiple van der Waals interactions and Pi-Pi T-shaped bonds with tyrosine (TYR-292) residue. The interaction also includes two Pi-anion bonds between glutamine residues that are favorable non-covalent bonds. With multiple stabilizing van der Waals interactions, hydrogen bonds, and Pi bonds, cianidanol could be an effective inhibitory drug for EGFR (1IVO) inhibition.

5. Scoring the QSAR model:

The QSAR model was built to produce a reusable pipeline to predict the binding affinity scores of compounds. The model evaluation was done through performance metrics, such as mean squared error (MSE), which is an average square of the error between actual and predicted values, and R-squared (R^2) error which represents the amount of variance in the dependent variable (binding affinity) explained by independent variable (descriptors). A model is a better fit if its MSE is low – less difference between actual and predicted values. The value of MSE for our model was 0.47 indicating that the error between the predicted and actual values is low but not negligible. The R-squared (R^2) value was 0.61, meaning that 61% of the variance is explained by the model. This is a reasonable value, but not too strong to prove the reliability of the model. The model demonstrates moderate predictive power, but in the context of QSAR, R^2 values greater than 0.60 are considered acceptable (Seema KESAR, 2019). The cross-validation results were obtained as a mean R^2 score equal to 0.51, and a standard deviation of R^2 scores of 0.17. The standard deviation score indicates the difference between the values of the five subsets of the data. This explains that the model does not hold a strong predictive power, but rather a moderate one. It is important to note that due to the small dataset, features with low variance were not removed. Removing low variance features in large datasets could considerably increase the efficiency of the model.

6. Conclusion:

We successfully performed molecular docking of our phytochemical library from barley (*Hordeum vulgare*) with human epidermal growth factor and receptor (EGFR). Flavonoids (orientin, saponarin, proanthocyanidin, and cianidanol) have shown the highest binding affinity with the receptor, and we present our study to drug discovery scientists and researchers to further test these compounds for their inhibitory activity against EGFR in triple-negative breast cancer (TNBC) or other related cancers. The QSAR model showed moderate predictive ability due to the small dataset and the presence of low variance features that could be handled with larger datasets.

7. Future Suggestions:

If required, the active sites of the protein can be predicted by CASTp or any other software that will help identify the key binding region in the protein structure. For the QSAR model, a larger dataset with more relevant descriptors could train the model better and enhance its predictive power.

References:

1. Ahmed Sallam, A. M. (2019). Drought Stress Tolerance in Wheat and Barley: Advances in Physiology, Breeding and Genetics Research. International Journal of Molecular Sciences. doi:doi:10.3390/ijms20133137

2. Beata Smolarz, A. Z. (2022). Breast Cancer—Epidemiology, Classification, Pathogenesis and Treatment (Review of Literature). Cancers. doi:10.3390/cancers14102569

3. Cassman, K. G. (1999). Ecological intensification of cereal production systems: Yield. Proceedings of the National Academy of Sciences. doi:10.1073/pnas.96.11.5952

4. Christopher A. Lipinski, F. L. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews. doi: https://doi.org/10.1016/S0169-409X(96)00423-1

5. Dallakyan S, O. A. (2015). Small-Molecule Library Screening by Docking with PyRx. Methods Molecular Biology. doi:10.1007/978-1-4939-2269-7_19

6. Hiroko Masuda, D. Z. (2012). Role of Epidermal Growth Factor Receptor in Breast Cancer. Breast cancer research and treatment. doi: 10.1007/s10549-012-2289-9

7. Jiyu Fan, A. F. (2019). Progress in molecular docking. Quantitative Biology, 7, 83-89. doi:https://doi.org/10.1007/s40484-019-0172-y

8. Khawla Al-Kuraya, P. S. (2004). Prognostic Relevance of Gene Amplifications and Coamplifications in Breast Cancer. Cancer Research, 64(23). doi:https://doi.org/10.1158/0008-5472.CAN-04-1945

9. Kim, S. C. (2023). PubChem 2023 update. Nucleic Acids Res. doi:https://doi.org/10.1093/nar/gkac956

10. Kuldeep Singh, J. K. (2023). Pharmacological and therapeutic potential of Hordeum vulgare. Pharmacological Research - Modern Chinese Medicine, 8. doi:https://doi.org/10.1016/j.prmcm.2023.100300.

11. Lauv Patel, T. S. (2020). Machine Learning Methods in Drug Discovery. Molecules. doi:https://doi.org/10.3390/molecules25225277

12. Mohanraj, K. K.-A. (2018). IMPPAT: A curated database of Indian Medicinal Plants, Phytochemistry And Therapeutics. Sci Rep, 8. doi:https://doi.org/10.1038/s41598-018-22631-z

13. Ogiso, H. I. (2023). Crystal Structure of the Complex of Human Epidermal Growth Factor and Receptor Extracellular Domains. doi:https://doi.org/10.1016/S0092-8674(02)00963-7

14. R.P. Vivek-Ananth, K. M. (2023). IMPPAT 2.0: An Enhanced and Expanded Phytochemical Atlas of Indian Medicinal Plants. ACS Omega. doi:10.1021/acsomega.3c00156

15. Renxiao Wang, Y. L. (2003). Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem. doi:10.1021/jm0203783

16. Seema KESAR, S. K. (2019). Quantitative Structure–Activity Relationship Analysis of Selective Rho Kinase Inhibitors as Neuro-regenerator Agents. Turkish Journal of Pharmaceutical Sciences. doi: 10.4274/tjps.galenos.2018.70288

17. Smita Rath, A. T. (2020). Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(5), 1467-1474. doi:https://doi.org/10.1016/j.dsx.2020.07.045

18. Yu Zhou, Y. F. (2022). Assessment of soil quality for guided fertilization in 7 barley agro-ecological areas of China. Plos One. doi:https://doi.org/10.1371/journal.pone.0261638