In silico ADMET, virtual screening and machine learning study of EGFR inhibitors for the treatment of type 2 diabetes and pancreatic cancer

Ricardo Romero*

September 5, 2023

Departamento de Ciencias Naturales, UAM Unidad Cuajimalpa, Avenida Vasco de Quiroga 4871, Col. Santa Fe Cuajimalpa, Alcaldía Cuajimalpa de Morelos, C.P. 05348, Ciudad de México.

Abstract

A differential gene expression analysis was performed in two studies of type 2 diabetes on pancreatic samples (Langerhans cells). In both cases the Epidermal Growth Factor Receptor (EGFR) was identified as the top hub gene in the protein-protein interaction networks of the upregulated genes. Functional enrichment analysis revealed that EGFR is also involved in pancreatic cancer signaling pathways. A virtual screening was conducted using EGFR inhibitors from the ZINC20 database. The results were filtered based on binding energies as well as ligands' absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties.

To further filter the results, two machine learning models were implemented and trained on a different set of ligands obtained from the ChEMBL database, one for regression and another for classification, to respectively predict pIC50 values and bioactivity class, resulting in six finalist compounds with viable characteristics, in terms of bioactivity and drug likeness, to be developed as EGFR inhibitors. Approved drugs afatinib, erlotinib, and almonertinib were used as controls, and in all cases the finalist compounds obtained better scores than the control drugs

*rromero@cua.uam.mx

1 Introduction

Type 2 diabetes (T2D) is a complex and chronic metabolic disorder, caused by both genetic and environmental factors[1, 2, 3]. Its main characteristic is hyperglycemia, which results from impaired insulin secretion, insulin resistance, or both. It is estimated that over 400 million people worldwide are affected by T2D, and its prevalence continues to rise. Other characteristics include obesity, dyslipidemia, hypertension, and inflammation, which have been linked to the development of other complications such as cardiovascular disease, neuropathy, nephropathy and retinopathy.

Several studies have shown that T2D is associated with reduced mass and/or reduced number of β-cells in the pancreatic islets[4, 5], which are clusters of endocrine cells located within the pancreas[6]. These cells are responsible for secreting hormones such as insulin, glucagon, and somatostatin, among others, as well as altered functionality of these cells[7], leading to insufficient insulin release and eventually hyperglycemia. Several genetic variants[8, 9, 10], identified through genome-wide association studies (GWAS), have been associated with the risk of developing T2D by influencing beta-cell function, mass, or response to various stimuli. Overall, the association between T2D and pancreatic islets highlights the importance of studying the underlying molecular mechanisms involved in the pathophysiology of disease onset and progression.

The management of T2D involves lifestyle changes such as diet and exercise, along with pharmacological prescriptions such as metformin, sulfonylureas, DPP4 inhibitors, etc., aiming to reduce blood sugar levels and improve insulin sensitivity[11, 12, 13]. However, these drugs do not stop or reverse the progression of the disease, focusing primarily on sugar control and symptom

management. This highlights the need for better theoretical and experimental results that could eventually lead to the development of improved drugs for its treatment.

On the other hand, the Epidermal Growth Factor Receptor (EGFR) is a transmembrane protein also known as ErbB1 or HER1, which plays a critical role in cellular signaling pathways. It belongs to the ErbB family of tyrosine kinase receptors and is expressed on the surface of various cells, including epithelial cells. Activation of EGFR initiates a cascade of intracellular events that lead to cell proliferation, survival, differentiation, and migration. The main features of the human EGFR include its extracellular domain, responsible for recognizing growth factors such as epidermal growth factor (EGF) and transforming growth factor-alpha (TGF-α). Upon binding to its ligand, EGFR undergoes conformational changes that result in dimerization and auto-phosphorylation of specific tyrosine residues within its cytoplasmic kinase domain. This phosphorylation activates downstream signaling pathways involved in the regulation of cell growth, tissue development, wound healing, and maintenance of epithelial tissues. Dysregulation or mutations in EGFR have been associated with various cancers[14], making it an important target for therapeutic interventions, such as targeted therapies and small molecule inhibitors. Although the relation between EGFR and diabetes is not conclusive, there are already studies pointing to their association, such as the proposal of EGFR as a therapeutic target for diabetic kidney disease and insulin resistance[15, 16].

Several FDA-approved drugs[17], specifically designed to inhibit EGFR signaling, have proven to be effective treatments for certain types of cancers, including non-small cell lung cancer (NSCLC), colorectal cancer, and head and neck squamous cell carcinomas. These include Gefitinib (Iressa), which is an EGFR tyrosine kinase inhibitor approved for the treatment of NSCLC. Erlotinib (Tarceva) is another EGFR tyrosine kinase inhibitor indicated for NSCLC and pancreatic cancer. Afatinib (Gilotrif) is an irreversible tyrosine kinase inhibitor targeting both EGFR and HER2, used in NSCLC treatment with specific mutations. Osimertinib (Tagrisso) is a drug approved for advanced or metastatic NSCLC patients who have developed resistance to other EGFR inhibitors due to a T790M mutation. Cetuximab (Erbitux) is a monoclonal antibody targeting EGFR, used in colorectal cancer, head and neck cancer, and some types of lung cancer. These targeted therapies can provide personalized treatment options by selectively blocking EGFR activity or its downstream signaling pathways, leading to improved clinical outcomes for patients with EGFR-driven malignancies.

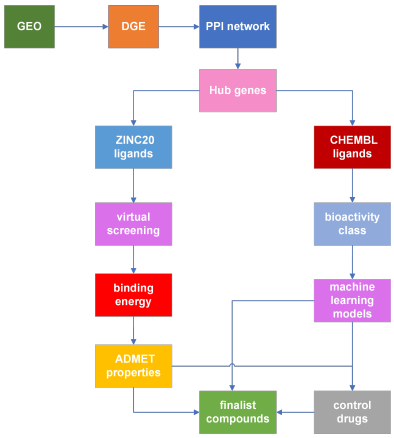


Figure 1: Flowchart of the procedure followed in the present study. The data is obtained from the Gene Expression Omnibus (GEO), and an analysis of differential gene expression (DGE) is performed. The protein-protein interaction (PPI) network is derived, and hub genes are identified. Ligands are obtained from ZINC20, and virtual screening is conducted. The results are filtered based on binding energy and ADMET properties. Simultaneously, machine learning models are trained to further filter compounds based on their bioactivity. The final compounds are compared with control drugs.

The objective of this study is to identify suitable compounds to be developed as drugs for a more effective treatment of type 2 diabetes, beyond the currently available drugs for the disease control. To achieve this, it is necessary to first identify, through differential expression analysis and interaction networks, the main hub genes associated with T2D based on pancreatic islet samples. Subsequently, a virtual screening using molecular docking and machine learning, as well as filtering of ADMET properties, will be performed to identify compounds with the best characteristics to proceed to the in vitro phase.

2 Methods and Materials

2.1 Data Acquisition and Analysis

Two suitable studies for the present analysis were obtained from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo) database, with accession numbers GSE25724 and GSE196797. These studies were selected based on the following criteria:

- The studies should focus on type 2 diabetes using pancreatic islet samples.
- The datasets should contain a clear distinction between T2D and non-T2D (control) samples.
- The differentially expressed genes obtained from the gene expression analysis should be sufficient to provide a protein-protein interaction (PPI) rich network.

Although the last requirement may seem somewhat arbitrary, its purpose will be clarified in the results section. The data from the GSE25724 study were analyzed online using the GEO2R tool for differential gene expression provided by NCBI for microarray data, comparing DT2 (6 samples) versus non-DT2 (7 samples). The data were normalized, and adjusted p-values using the Benjamini and Hochberg method, logarithmic transformation, and precise weights from the R limma package were employed. Box plots and volcano plots were generated with a significance level of 0.05 and a log2 fold change (log2FC) threshold of zero. The resulting data were downloaded, and genes were selected based on the criteria log2FC > 0 (since we were interested only in overexpressed genes) and adjusted p-value < 0.05. Subsequently, the genes were loaded into the STRING server (https://string-db.org) to obtain the PPI network, which was then imported into Cytoscape software[18]. The top ten hub genes, based on their degree of interaction, were identified using the CytoHubba plugin[19].

Data from GSE196797, which is a miRNA high throughput sequencing study, were not available for analysis using the GEO2R tool. Therefore, they were analyzed in R Studio using the DESeq2 package of R[20], available from the Bioconductor resource center[21]. Normalized gene count data with 2 DT2 samples versus 6 non-DT2 samples were used. Genes with a p-value < 0.05 and log2FC > 0 were selected from the differential expression analysis. The PPI network and top ten hub genes were obtained following the same procedure as the previous study. Both the raw datasets and the R script used are available in the supplementary materials.

2.2 Enrichment Analysis

The top ten hub genes from the GSE25724 study were selected for the enrichment analysis, as they provide the best interaction network from both studies, as shown in the results section. The genes were loaded into the Enrichr database[22, 23] and compared with the KEGG 2021 Human signaling pathways database[24].

To support the results of the hub genes, the GeneTrail 3.2 database[25] was used for transcriptomic analysis. The hub genes were loaded, and alternatively, the accession number of the study was directly provided to obtain the top ten enriched categories, in both cases. This analysis includes data from the WikiPathways 2021 Human database[26] and Gene Ontology[27, 28] (GO). From the miRNet database[29, 30], groups of miRNAs related to diabetes and to pancreatic neoplasms were obtained by analyzing all the miRNAs associated with the hub genes, using the miRNA disease database and a hypergeometric test.

2.3 Virtual screening procedure

The number one hub gene in both studies was the Epidermal Growth Factor Receptor (EGFR), so the corresponding protein was searched for in the Uniprot database[31]. From the filtered results for humans, entry with accession number P00533 was selected to retrieve the fasta sequence of the protein, which was then loaded into the online software protein BLAST from NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and compared with the Protein Data Bank[32] (PDB) database.

Entry with accession number 7SYD[33] was chosen, showing a coverage of 100% and an identity of 99.92%. The PDB structure corresponds to the complete extracellular module of EGFR bound to EGF in a juxtaposed conformation, obtained through electron microscopy (Figure 1). The structure was stripped off of its EGF ligands using USCF Chimera[34], energy was minimized using Swiss PDB-viewer 4.10[35] and further prepared for docking by assigning Gasteiger charges using Chimera again.

Ligands were downloaded from ZINC20 database[36]. They were selected based on biological samples and targeting the EGFR gene. Ligands available for purchase were used, initially totaling 550 compounds; however, after extracting them, converting sdf files, and performing energy minimization using Open Babel software[37] for both tasks, only 475 functional ligands remained. Energy minimization was performed with the Amber force field with 5000 steps. Virtual screening was carried out using Autodock Vina[38, 39] software, with coordinates centered at (158, 171.66, 204.89) and a grid size of

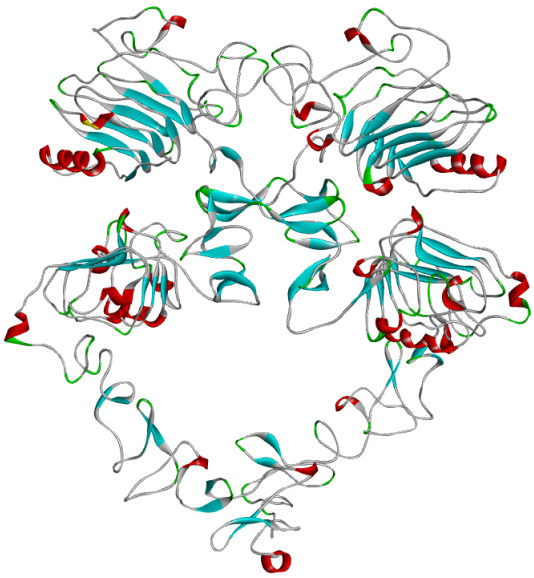


Figure 2: Structure of the Epidermal Growth Factor Receptor (EGFR) molecule, PDB 7SYD, used for molecular docking. The picture shows the structure of the complete extracellular module of EGFR with its ligand EGF removed, in a juxtaposed conformation. The 7SYD structure was obtained through electron microscopy.

(20, 20, 150) to cover the entire domain II of the EGFR structure, which is known to be the active site of the molecule. The energy range was set to 4 with an exhaustiveness value of 10. A cutoff value for energy was established at -8.5 kcal/mol, and only complexes with equal or lower values than this limit were considered or, equivalently, those with higher or equal absolute values. This resulted in 130 potential hits from the original set of 475 ligands.

2.4 ADMET Filtering

The results from the molecular docking procedure were loaded into the Swiss ADME database[40], and their properties were calculated and retrieved. The compounds underwent filtering based on complete compliance with the drug-likeness rules of Lipinski[41], Ghose[42], Veber[43], Egan[44], and Muegge[45]. Additionally, a solubility filter was applied, requiring that at most one condition among the three classes of solubility, ESOL[46], Ali[47], and Silicos-IT, be poorly soluble while the other two should be soluble or moderately soluble. The OSIRIS Property Explorer software (https://www.organic-chemistry.org/prog/peo/) was then utilized to search for toxicological risks, implementing an additional filter that required compounds to pass mutagenic, tumorigenic, irritant, and reproductive effective filters displayed in green. Values for solubility, drug-likeness score, and drug score produced by the software were also recorded. Calculation details for these parameters can be found on the program's website.

These filters reduced the number of hits to 17; they were supplemented with control drugs Afatinib, Almonertinib, Erlotinib along with Curcumin, Dimethylcurcumin, and Tirfostin-AG-835. Afatinib was obtained from the DGIdb gene-drug interaction database while the latter two came from GeneCards database as inhibitors of EGFR. These control drugs were chosen based on their superior characteristics among each generation of approved EGFR inhibitors[17]. Three-dimensional structures of both control drugs and added compounds were downloaded from PubChem (https://pubchem.ncbi.nlm.nih.gov/) and optimized for docking using Avogadro[48] with the Generalized Amber Force Field (GAFF) with 5000 steps.

To further validate the docking procedure and classify the results, docking was repeated ten times for each of the 23 final compounds, but this time with an exhaustiveness of 30 to compensate for the large search space, and average binding energy was calculated. Using these results along with those from OSIRIS software, a simple average score was obtained: Score = 1/4(solubility + drug-likeness + Drug score + |binding energy|). Solubility was left without taking absolute value because, although it is negative, less negativity indicates better solubility, unlike binding energy.

2.5 Machine learning bio-availability prediction

Two machine learning models were implemented, one to predict the pIC50 values of the final compounds using a regression model, and the other to predict their bio-availability class, whether active or non-active, independently of the pIC50 values, using a classification model. The idea is for these models to complement each other and provide more robust predictions.

Data were downloaded from the ChEMBL database[49] to train the models. A search was performed for compounds targeting human EGFR, and dataset with identifier CHEMBL203 was selected. Data including IC50 values, in nanomoles, were downloaded for these compounds. Originally, there were 16,715 compounds

in this set, but after cleaning the data by removing zero or missing IC50 values, as well as duplicate or missing canonical SMILES, 9,508 elements remained.

The compounds were then classified based on their IC50 values: active if IC50 $\leq$ 1000 nm; inactive if IC50 $\geq$ 10000 nm; and intermediate for all others. The IC50 values were normalized to a cut-off value of 10^9 nm and then converted into pIC50 values. Lipinski descriptors were calculated for these compounds and Mann-Whitney statistical test was applied to each of the four descriptors to verify if indeed active and inactive classes belong to different distributions. Using Python's rdkit package, Morgan fingerprints with radius 2 and 1024 bit vectors were obtained for these compounds.

To train the models, a Random Forest regressor and classifier were respectively used on the training set, which consisted of 80% of all available data (with the remaining 20% assigned to the test set), while model parameters were optimized through random search and cross-validatred. For the regression model, the objective variable (y) was the pIC50 value while independent variables (X) were Morgan fingerprints. For the classification model, pIC50 values were converted into binary code: 1 for pIC50 $\geq$ 6 (active) and 0 for the rest (non-active, including inactive and intermediate), which became the dependent variable, while independent variables remained the same. Once models were trained, they were validated on the testing set and then applied to the final set of 23 compounds, including controls, obtained after ADMET filtering. The Python code used is available in supplementary materials.

Figure 2 shows a flowchart depicting the steps followed in this study.

3 Results

The main results of the study include the identification and classification of six compounds with suitable characteristics to be developed as potential EGFR inhibitors. In descending order according to their overall score (defined in section 2.4) and using ZINC database identifiers, these compounds are: ZINC000003816310, ZINC000053683345, ZINC000147701032, ZINC000001494443, ZINC000103297550, and ZINC000020148984. Their properties are presented in Tables 1 and 2 along with the control drugs Afatinib, Almonertinib, and Erlotinib used for comparison, as well as the compounds Curcumin, Dimethylcurcumin, and Tirfostin-AG-835 obtained from drug databases.

The results of differential gene expression analysis as well as IPP networks and hub genes, for both studies used, are presented in Figures 3 to 6. In both cases, the EGFR hub gene had the highest degree of interaction. The molecular docking results are shown in Table 3 for the two highest-scoring compounds and for Almonertinib, which had the highest score among controls (the remaining results can be found in supplementary materials). Since docking was performed ten times per compound with an average calculated from these runs, complexes with the lowest standard deviation were selected to extract interactions in all cases. Figures 7 to 9 display these complexes.

For enrichment analysis, hub genes from GSE25724 study were utilized since its interaction network is superior among both studies (as observed in Figure 6). The results using Enrichr and GeneTrail are shown in Figure 10. In GeneTrail analysis specifically, regardless if directly uploading the hub genes or just providing the GEO accession number, resulted in identical top ten enriched categories, which indirectly verifies the validity of obtained hub genes. The most important finding from enrichment analysis is that EGFR is involved in pancreatic cancer signaling pathways. Among the set of hub genes, five of them - EGFR, ERBB2, RAC1, MAPK3, and SMAD3 - participate in pancreatic cancer signaling pathways. The analysis also provided sets of miRNAs associated with diabetes mellitus and pancreatic neoplasms (available in supplementary materials). Figure 11 depicts the miRNA interaction network.

Table 4 presents the results of Mann-Whitney test for active and inactive compounds obtained from ChEMBL and classified as explained in section 3.5, based on Lipinski descriptors: molecular weight (MW), octanol-water partition coefficient (LogP), number of hydrogen bond donors (#H donors), and number of hydrogen bond acceptors (#H acceptors). In all cases, it can be concluded that the compounds belong to different distributions; however, both classes occupy the same chemical space as shown in Figure 12.

Regarding machine learning models, the regression model achieved an R^2 score of 0.73 with a MSE of 0.61, while the classification model achieved an accuracy rate of 85% for classifying non-active compounds (inactive and intermediate) and an accuracy rate of 89% for classifying active compounds, with an average precision rate of 87% and an AUC value of 0.86. Corresponding graphs

are displayed in Figure 13.

4 Discussion and concluding remarks

An analysis of the data in Table 1 shows that compound ZINC000003816310, which obtained the highest score, has a pIC50 value of 5.94 according to the regression model and was classified as non-active by the classification model. Therefore, it is safe to assume that this compound lies on the borderline between active and intermediate compounds. Given its favorable characteristics, it would be the top candidate for further development as a lead compound. The remaining five compounds were all classified as active by both models, and their priority for development would follow the same order presented here based on their scores. It is worth noting that all six proposed compounds achieved significantly better scores compared to control drugs, which exhibited a rather poor performance. As for curcumin, which appears as a ligand in several studies related to cancer, our results show that its bio-activity is far from being within the active range to be seriously considered a viable compound, despite having a good overall score. The same holds true for tyrphostin-AG-835 and dimethylcurcumin showed even worse bio-activity.

Another noteworthy result is the association of EGFR gene with diabetes; which, of course, would necessarily require experimental verification, but that nonetheless provide support to studies already showing such association, as the ones mentioned in the introduction. There is also the involvement of EGFR in pancreatic cancer signaling pathway, as showed in the enrichment analysis. In summary, the findings give more weight to the association between EGFR and type 2 diabetes, and we can conclude that our objective of identifying potential drug candidates for its treatment has been achieved, narrowing down viable hits for in vitro studies. Moreover, since EGFR is a well-known oncogene, the obtained results can also serve as a basis for eventually obtaining improved drugs for diseases requiring its inhibition.

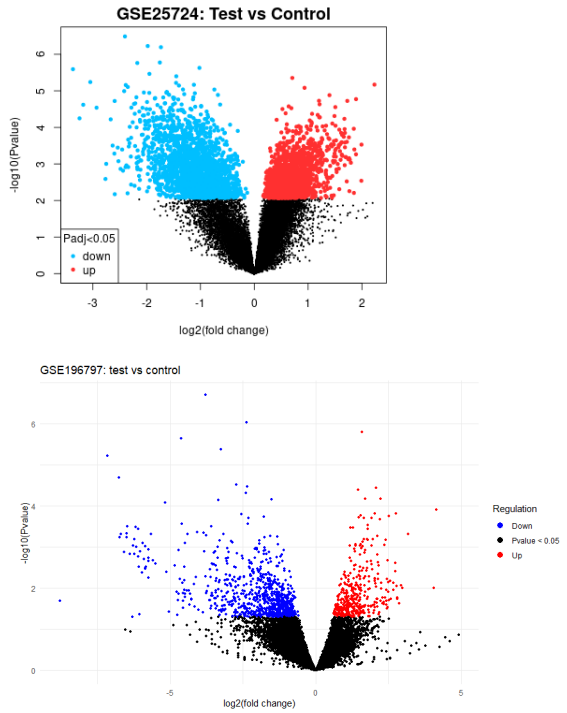


Figure 3: Volcano graphs of the differentially expressed genes in both studies

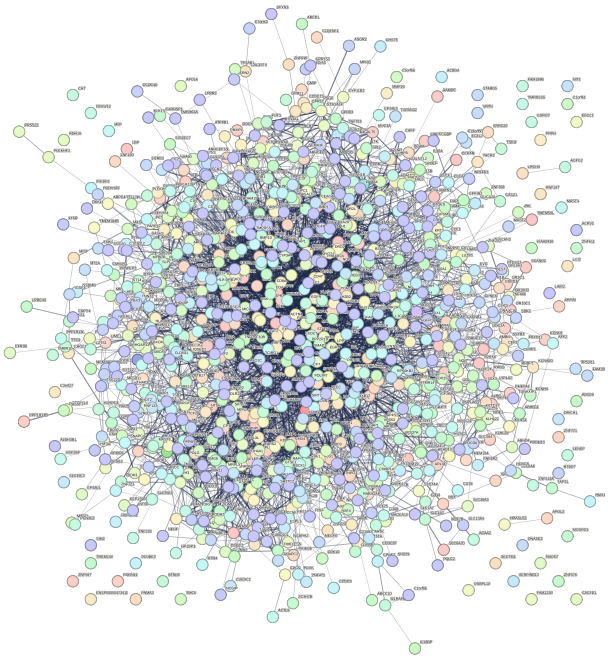


Figure 4: Protein-Protein Interaction (PPI) network for the over expressed genes in the GSE25724 study. It consists of 998 nodes and 5036 edges, with an average node degree of 10.1 and an average local clustering coefficient of 0.338. The expected number of edges is 4699, with a PPI enrichment p-value of 5.99e-07.

Compound	Canonical SMILES	Average binding energy	SD	Solubility	Druglikeness	Drug score	Final score	pIC50	Class	Classification
ZINC000003816310	CN1CCN(CC1)c1ccc2c(c1)nH]c(n2)c1c(=O)]nH]c2c(c1N)c(F)ccc2	-9.69	0.25	-2.39	5.77	0.85	3.48	5.94	Intermediate	Non-active
ZINC000053683345	COc1cc2c(cc1OCC1CCN(CC1)C)nH]cn/c/2=N/c1ccc(cc1F)Br	-8.86	0.02	-5.19	3.87	0.52	2.01	6.31	Active	Active
Dimethylcurcumin	COc1nccc(n1)c1c(ncn1)C@@H]1CC]C@@H]([CC1)O)c1ccc(cc1)F	-8.02	0.24	-4.3	3.68	0.62	2.00	4.97	Inactive	Non-active
ZINC000147701032	COc1nccc(n1)c1c(ncn1)C@@H]1CC]C@@H]([CC1)O)c1ccc(cc1)F	-9.19	0.53	-3.36	0.89	0.66	1.85	7.08	Active	Active
Tyrphostin-AG-835	N#C/C(=C\c1ccc(c(c1)O)O)/C(=O)NC@H](c1ccccc1)C	-8.68	0.01	-3.07	0.73	0.73	1.77	5.43	Intermediate	Non-active
ZINC000001494443	CN1CCC(CC1)Nc1ncc2c(n1)c(ncn2)Nc1ccc(c(c1)Cl)F	-8.75	0.40	-4.68	2.23	0.63	1.73	8.35	Active	Active
ZINC000103297550	Oc1cccc(c1)Nc1ccnc2c1c1ccccc1[nH]2	-8.81	0.01	-4.97	0.56	0.54	1.23	6.27	Active	Active
Curcumin	COc1cc(C=CC(=O)CC(=O)C=Cc2ccc(c(c2)OC)O)ccc1O	-9.12	0.08	-3.62	-3.95	0.4	0.49	5.02	Intermediate	Non-active
ZINC000020148984	COc1cccc(c1)c1cn(c2c1c(N)ncn2)c1ccccc1	-8.98	0.02	-7.98	-0.35	0.33	0.24	6.20	Active	Active
Almonertinib	C=CC(=O)Nc1cc(Nc2nccc(n2)c2c=n(c3c2cccc3)C2CC2)c(cc1N(CCN(C)C)C)OC	-8.60	0.33	-4.46	-4.09	0.08	0.03	9.17	Active	Active
Afatinib	CN(CC=CC(=O)Nc1cc2c(ncnc2cc1OC1COCC1)Nc1ccc(c(c1)Cl)F)C	-9.19	0.83	-5.48	-4.11	0.24	-0.04	8.62	Active	Active
Erlotinib	COCCOc1cc2c(ncnc2cc1OCCOC)Nc1cccc(c1)C#C	-7.84	0.32	-3.53	-6.73	0.38	-0.51	6.42	Active	Non-active

Table 1: Compounds properties. The Average column shows the average binding energy from the ten molecular docking runs, in kcal/mol. SD is the associated standard deviation. Solubility, Druglikeness, and Drug Score are the scores obtained from OSIRIS Property Explorer. The final score is calculated using the formula 1/4(solubility + druglikeness + drug score + |binding energy|). The pIC50 column contains the numerical values predicted by the machine learning regression model, and Class represents the corresponding classification explained in section 2.5 (after conversion to IC50 values). The Classification column displays the results from the machine learning classifier model.

Compounds	MW	TPSA	BBB	Lipinski	Ghose	Veber	Egan	Muegge	PAINS	Brenk	Leadlikeness
ZINC000003816310	392.43	94.04	No	0	0	0	0	0	0	0	1
ZINC000053683345	475.35	62.74	Yes	0	0	0	0	0	0	0	2
Dimethylcurcumin	396.43	74.22	Yes	0	0	0	0	0	0	3	3
ZINC000147701032	368.40	73.06	Yes	0	0	0	0	0	0	0	1
Tyrphostin-AG-835	308.33	95.35	No	0	0	0	0	0	1	3	0
ZINC000001494443	387.84	78.86	Yes	0	0	0	0	0	0	0	1
ZINC000103297550	257.30	60.94	Yes	0	0	0	0	0	0	0	1
Curcumin	368.38	93.06	No	0	0	0	0	0	0	2	2
ZINC000020148984	316.36	65.96	Yes	0	0	0	0	0	0	0	1
Almonertinib	525.64	87.55	No	1	3	1	0	0	0	1	3
Afatinib	485.94	88.61	No	0	1	0	0	0	0	1	3
Erlotinib	393.44	74.73	Yes	0	0	0	0	0	0	1	2

Table 2: Druglikeness properties. Molecular weight (MW) in g/mol, topological surface area (TPSA) in $Å^2$, brain blood barrier permeability (BBB). The Lipinski, Ghose, Veber, Egan and Muegge columns indicate the number of violations of the respective rules. The columns PAINS (Pan Assays Interference Structures), Brenk and Leadlikeness indicate the number of alerts of the corresponding conditions.

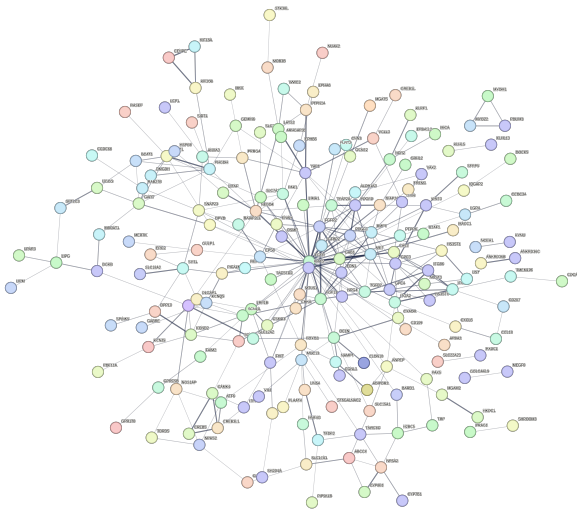
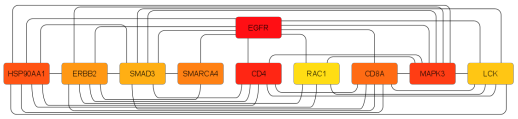
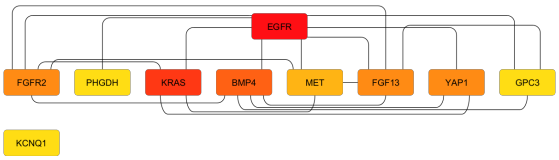


Figure 5: PPI network for up-regulated genes in study GSE196797. It consists of 298 nodes and 314 edges, with an average node degree of 2.11 and an average local clustering coefficient of 0.314. The expected number of edges is 229 with a PPI enrichment p-value of 6.37e-08.

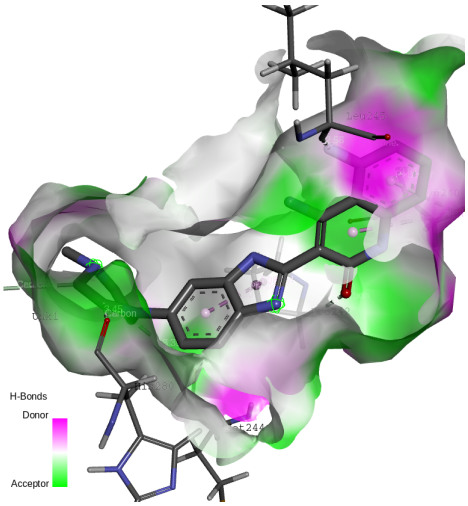


(a) Top ten hub genes from GSE25724

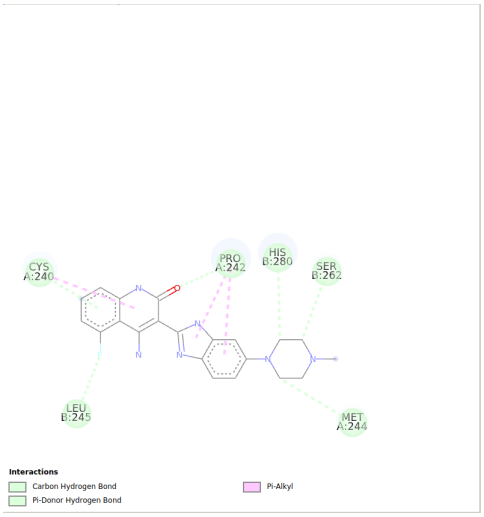


(b) Top ten hub genes from GSE196797

Figure 6: The top ten hub genes with the highest degree of interaction in the PPI network for both studies. Color code for the degree of interaction: darker red for the highest, lighter yellow for the lowest.

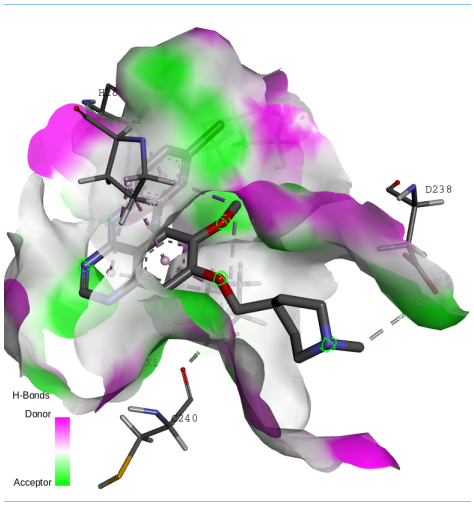


(a) Three-dimensional diagram based on the characteristics of donor-receptor hydrogen bonds between amino acid residues.

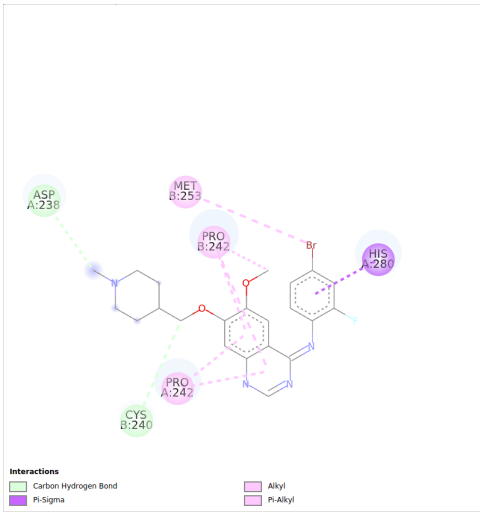


(b) Two-dimensional diagram showing various interactions between the ligand and the protein.

Figure 7: Molecular interactions for the complex EGFR-ZINC000003816310

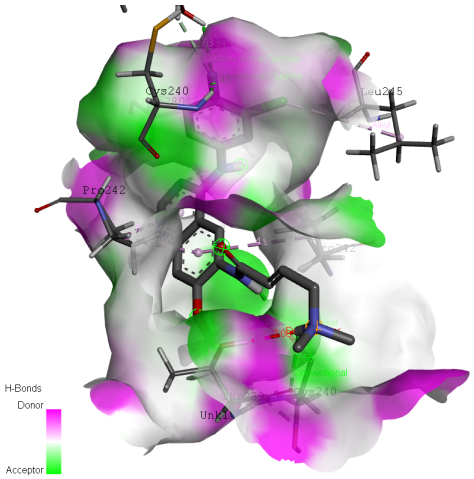


(a) Three-dimensional diagram based on the characteristics of donor-receptor hydrogen bonds between amino acid residues.

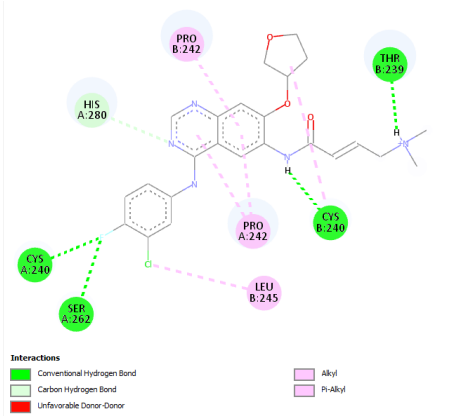


(b) Two-dimensional diagram showing various interactions between the ligand and the protein

Figure 8: Molecular interactions for the complex EGFR-ZINC000053683345

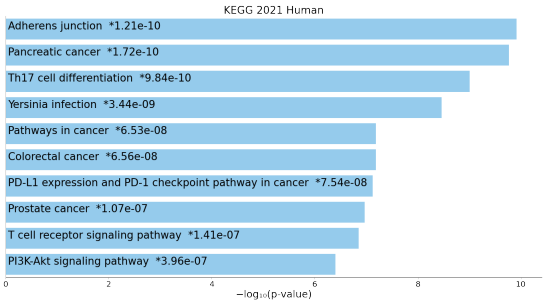


(a) Three-dimensional diagram based on the characteristics of donor-receptor hydrogen bonds between amino acid residues.

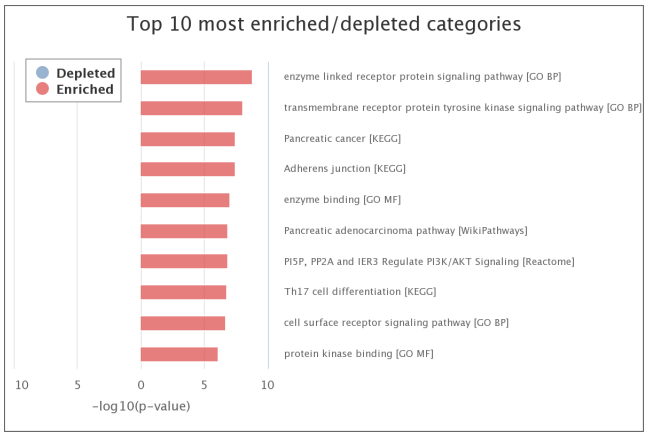


(b) Two-dimensional diagram showing various interactions between the ligand and the protein

Figure 9: Molecular interactions for the complex EGFR-Almonertinib



(a) Results from enrichment analysis against Kegg 2021 database



(b) Results from GeneTrail for the ten most enriched classes

Figure 10: Results from enrichment analysis of the ten hub genes from GSE25724

Compound	Binding energy (Kcal/mol)	Donnor	Distance	Acceptor	Categy	Interaction type
ZINC000003816310	-9.73	UNL1:C	2.99	B:SER262:OG	Hydrogen Bond	Carbon Hydrogen Bond
		UNL1:C	3.45	B:HIS280:O	Hydrogen Bond	Carbon Hydrogen Bond
		UNL1:C	3.39	A:MET244:O	Hydrogen Bond	Carbon Hydrogen Bond
		UNL1	4.50	A:PRO242	Hydrophobic	Pi-Alkyl
		UNL1	4.08	A:PRO242	Hydrophobic	Pi-Alkyl
		A:PRO242:HA	2.36	UNL1:O	Hydrogen Bond	Carbon Hydrogen Bond
		UNL1	5.20	A:CYS240	Hydrophobic	Pi-Alkyl
		UNL1	4.92	A:CYS240	Hydrophobic	Pi-Alkyl
		B:LEU245:HA	2.58	UNL1:F	Hydrogen Bond; Halogen	Carbon Hydrogen Bond; Halogen (Fluorine)
ZINC000053683345	-8.86	UNL1	4.95	A:PRO242	Hydrophobic	Pi-Alkyl
		UNL1	4.25	A:PRO242	Hydrophobic	Pi-Alkyl
		UNL1	4.01	B:PRO242	Hydrophobic	Pi-Alkyl
		UNL1	5.30	B:PRO242	Hydrophobic	Pi-Alkyl
		UNL1:C	3.69	UNL1	Hydrophobic	Pi-Sigma
		A:HIS280:HB2	2.75	UNL1	Hydrophobic	Pi-Sigma
		UNL1:BR	5.38	B:MET253	Hydrophobic	Alkyl
		UNL1:C	4.03	B:PRO242	Hydrophobic	Alkyl
		UNL1:C	3.64	B:CYS240:O	Hydrogen Bond	Carbon Hydrogen Bond
		UNL1:C	3.64	A:ASP238:OD2	Hydrogen Bond	Carbon Hydrogen Bond
Almonertinib	-8.65	UNL1:HN	2.56	B:THR239:O	Hydrogen Bond	Conventional Hydrogen Bond
		UNL1:H	2.46	B:CYS240:O	Hydrogen Bond	Conventional Hydrogen Bond
		B:CYS240	5.25	UNL1	Hydrophobic	Alkyl
		B:THR239:HG1	2.08	UNL1:HN	Unfavorable r	Unfavorable Donor-Donor
		UNL1	4.91	B:PRO242	Hydrophobic	Pi-Alkyl
		UNL1	4.21	A:PRO242	Hydrophobic	Pi-Alkyl
		UNL1	4.97	A:PRO242	Hydrophobic	Pi-Alkyl
		UNL1:CL	5.03	B:LEU245	Hydrophobic	Alkyl
		A:CYS240:HN	2.44	UNL1:F	Hydrogen Bond;Halogen	Conventional Hydrogen Bond;Halogen (Fluorine)
		A:SER262:HB2	2.63	UNL1:F	Hydrogen Bond	Carbon Hydrogen Bond
		A:SER262:HG	2.71	UNL1:F	Hydrogen Bond;Halogen	Conventional Hydrogen Bond;Halogen (Fluorine)

Table 3: Binding energy and type of molecular interaction for the top finalist compounds and the control almonertinib. The distance is reported in $Å^2$, and UNL1 refers to the corresponding ligand in each case.

Descriptor	Statistic	P value	Result
Molecular weight	8924972.0	1.38e-192	Different distributions (reject H0)
LogP	7761898.5	4.09e-64	Different distributions (reject H0)
#H donors	6906548.0	1.48e-15	Different distributions (reject H0)
#H acceptors	8671850.5	7.33e-162	Different distributions (reject H0)

Table 4: Mann-Whitney test results($\alpha = 0.05$) for active and inactive compounds, from ChEMBL, for their Lipinskidescriptors.

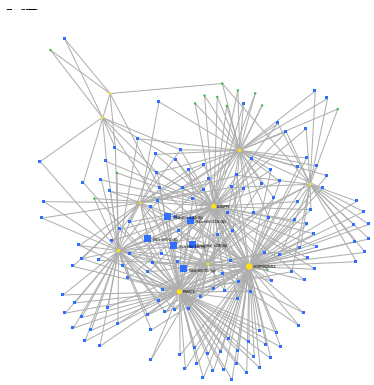


Figure 11: miRNAs (in blue) interaction network obtained from miRNet using the hub genes (in yellow) from the GSE25724 study.

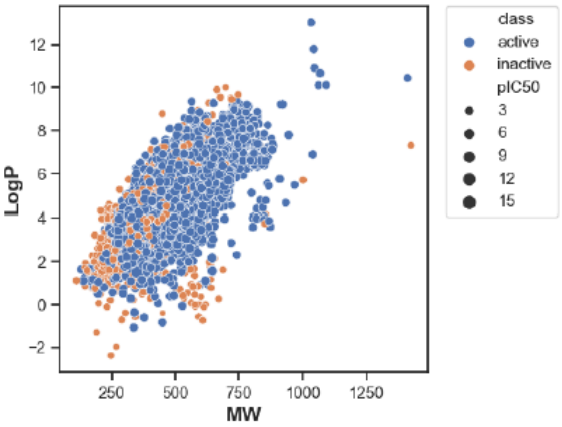
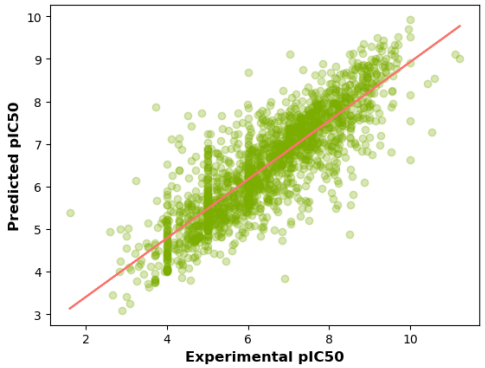
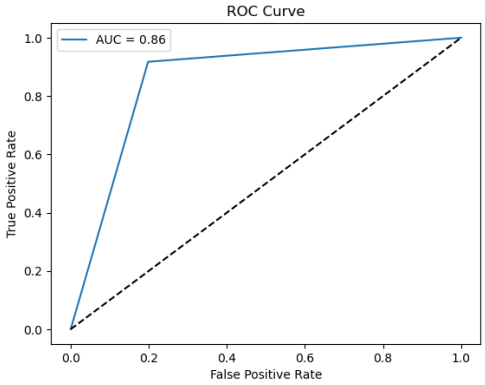


Figure 12: Scatter plot of molecular weight (MW) vs octanol-water partition coefficient (LogP) for active and inactive compounds obtained from ChEMBL. The plot demonstrates that both classes occupy the same chemical space despite originating from different distributions based on their Lipinski descriptors.



(a) Scatter plot of predicted vs actual results for pIC50 values obtained using the random forest regression model.



(b) ROC curve and area under the curve (AUC) for the random forest classification model.

Figure 13: Performance plots for both machine learning models.

References

[1] DeFronzo RA, Ferrannini E, Groop L, Henry RR, Herman WH, Holst JJ, et al. Type 2 diabetes mellitus. Nature Reviews Disease Primers. 2015 Jul;1(1):15019.

[2] Galicia-Garcia U, Benito-Vicente A, Jebari S, Larrea-Sebal A, Siddiqi H, Uribe KB, et al. Pathophysiology of Type 2 Diabetes Mellitus. Int J Mol Sci. 2020 Aug;21(17).

[3] Goyal R, Jialal I. Type 2 diabetes - statpearls - NCBI bookshelf. Treasure Island (FL): StatPearls Publishing; 2023. Available from: https://www.ncbi.nlm.nih.gov/books/NBK513253.

[4] Hu F, Qiu X, Bu S. Pancreatic islet dysfunction in type 2 diabetes mellitus. Archives of Physiology and Biochemistry. 2020;126(3):235-41. PMID: 30293453. Available from: https://doi.org/10.1080/13813455.2018.1510967.

[5] Marchetti P, Del Guerra S, Marselli L, Lupi R, Masini M, Pollera M, et al. Pancreatic Islets from Type 2 Diabetic Patients Have Functional Defects and Increased Apoptosis That Are Ameliorated by Metformin. The Journal of Clinical Endocrinology and Metabolism. 2004 11;89(11):5535-41.

[6] Da Silva Xavier G. The Cells of the Islets of Langerhans. Journal of Clinical Medicine. 2018;7(3). Available from: https://www.mdpi.com/2077-0383/7/3/54.

[7] Böni-Schnetzler M, Meier DT. Islet inflammation in type 2 diabetes. Seminars in Immunopathology. 2019 Jul;41(4):501-13. Available from: https://doi.org/10.1007/s00281-019-00745-4.

[8] Goodarzi MO, Palmer ND, Cui J, Guo X, Chen YDI, Taylor KD, et al. Classification of Type 2 Diabetes Genetic Variants and a Novel Genetic Risk Score Association With Insulin Clearance. The Journal of Clinical Endocrinology 'I&' Metabolism. 2019 11;105(4):1251-60. Available from: https://doi.org/10.1210/clinem/dgz198.

[9] Shoily SS, Ahsan T, Fatema K, Sajib AA. Common genetic variants and pathways in diabetes and associated complications and vulnerability of populations with different ethnic origins. Scientific Reports. 2021 Apr;11(1):7504. Available from: https://doi.org/10.1038/s41598-021-86801-2.

[10] ElHajj Chehadeh S, Sayed NS, Abdelsamad HS, Almahmeed W, Khandoker AH, Jelinek HF, et al. Genetic Variants and Their Associations to Type 2 Diabetes Mellitus Complications in the United Arab Emirates. Frontiers in Endocrinology. 2022;12. Available from: https://www.frontiersin.org/articles/10.3389/fendo.2021.751885.

[11] Davies MJ, Aroda VR, Collins BS, Gabbay RA, Green J, Maruthur NM, et al. Management of Hyperglycemia in Type 2 Diabetes, 2022. A Consensus Report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes Care. 2022 09;45(11):2753-86. Available from: https://doi.org/10.2337/dci22-0034.

[12] Artasensi A, Pedretti A, Vistoli G, Fumagalli L. Type 2 Diabetes Mellitus: A Review of Multi-Target Drugs. Molecules. 2020;25(8). Available from: https://www.mdpi.com/1420-3049/25/8/1987.

[13] Reed J, Bain S, Kanamarlapudi V. A Review of Current Trends with Type 2 Diabetes Epidemiology, Aetiology, Pathogenesis, Treatments and Future Perspectives. Diabetes Metab Syndr Obes. 2021 Jun;14:3567-602.

[14] Yarden Y, Pines G. The ERBB network: at last, cancer therapy meets systems biology. Nature Reviews Cancer. 2012 Aug;12(8):553-63. Available from: https://doi.org/10.1038/nrc3309.

[15] Li Z, Li Y, Overstreet JM, Chung S, Niu A, Fan X, et al. Inhibition of Epidermal Growth Factor Receptor Activation Is Associated With Improved Diabetic Nephropathy and Insulin Resistance in Type 2 Diabetes. Diabetes. 2018 06;67(9):1847-57. Available from: https://doi.org/10.2337/db17-1513.

[16] Sheng L, Bayliss G, Zhuang S. Epidermal Growth Factor Receptor: A Potential Therapeutic Target for Diabetic Kidney Disease. Frontiers in Pharmacology. 2021;11. Available from: https://www.frontiersin.org/articles/10.3389/fphar.2020.598910.

[17] Abourehab MAS, Alqahtani AM, Youssif BGM, Gouda AM. Globally Approved EGFR Inhibitors: Insights into Their Syntheses, Target Kinases, Biological Activities, Receptor Interactions, and Metabolism. Molecules. 2021;26(21). Available from: https://www.mdpi.com/1420-3049/26/21/6677.

[18] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Research. 2003;13(11):2498-504. Available from: http://genome.cshlp.org/content/13/11/2498.full.html.

[19] Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Systems Biology. 2014 Dec;8(4):S11. Available from: https://doi.org/10.1186/1752-0509-8-S4-S11.

[20] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014 Dec;15(12):550. Available from: https://doi.org/10.1186/s13059-014-0550-8.

[21] Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods. 2015 Feb;12(2):115-21. Available from: https://doi.org/10.1038/nmeth.3252.

[22] Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013 Apr;14:128.

[23] Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016 May;44(W1):W90-7.

[24] Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Research. 2020 10;49(D1):D545-51. Available from: https://doi.org/10.1093/nar/gkaa970.

[25] Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, et al. GeneTrail – advanced gene set enrichment analysis. Nucleic acids research. 2007;35(suppl_2):W186-92.

[26] Martens M, Ammar A, Riutta A, Waagmeester A, Slenter D, Hanspers K, et al. WikiPathways: connecting communities. Nucleic Acids Research. 2020 11;49(D1):D613-21. Available from: https://doi.org/10.1093/nar/gkaa1024.

[27] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nature Genetics. 2000 May;25(1):25-9. Available from: https://doi.org/10.1038/75556.

[28] Consortium TGO, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023 03;224(1):iyad031. Available from: https://doi.org/10.1093/genetics/iyad031.

[29] Chang L, Xia J. In: Song Q, Tao Z, editors. MicroRNA Regulatory Network Analysis Using miRNet 2.0. New York, NY: Springer US; 2023. p. 185-204. Available from: https://doi.org/10.1007/978-1-0716-2815-7_14.

[30] Chang L, Zhou G, Soufan O, Xia J. miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology. Nucleic Acids Research. 2020 06;48(W1):W244-51. Available from: https://doi.org/10.1093/nar/gkaa467.

[31] Consortium TU. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research. 2022 11;51(D1):D523-31. Available from: https://doi.org/10.1093/nar/gkac1052.

[32] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research. 2000 01;28(1):235-42. Available from: https://doi.org/10.1093/nar/28.1.235.

[33] Huang Y, Ognjenovic J, Karandur D, Miller K, Merk A, Subramaniam S, et al. A molecular mechanism for the generation of ligand-dependent differential outputs by the epidermal growth factor receptor. eLife. 2021 nov;10:e73218. Available from: https://doi.org/10.7554/eLife.73218.

[34] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera, A visualization system for exploratory research and analysis. Journal of Computational Chemistry. 2004;25(13):1605-12. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20084.

[35] Guex N, Peitsch MC. SWISS-MODEL and the Swiss-Pdb Viewer: An environment for comparative protein modeling. ELECTROPHORESIS. 1997;18(15):2714-23. Available from: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/elps.1150181505.

[36] Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, et al. ZINC20, A Free Ultralarge-Scale Chemical Database for Ligand Discovery. Journal of Chemical Information and Modeling. 2020;60(12):6065-73. PMID: 33118813. Available from: https://doi.org/10.1021/acs.jcim.0c00675.

[37] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. Journal of Cheminformatics. 2011 Oct;3(1):33. Available from: https://doi.org/10.1186/1758-2946-3-33.

[38] Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry. 2010;31(2):455-61. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21334.

[39] Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. Journal of Chemical Information and Modeling. 2021 Aug;61(8):3891-8. Available from: https://doi.org/10.1021/acs.jcim.1c00203.

[40] Daina A, Michielin O, Zoete V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Scientific Reports. 2017 Mar;7(1):42717. Available from: https://doi.org/10.1038/srep42717.

[41] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3-25.1. Advanced Drug Delivery Reviews. 2001;46(1):3-26. Special issue dedicated to Dr. Eric Tomlinson, Advanced Drug Delivery Reviews, A Selection of the Most Highly Cited Articles, 1991-1998. Available from: https://www.sciencedirect.com/science/article/pii/S0169409X00001290.

[42] Ghose AK, Viswanadhan VN, Wendoloski JJ. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. Journal of Combinatorial Chemistry. 1999;1(1):55-68. PMID: 10746014. Available from: https://doi.org/10.1021/cc9800071.

[43] Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. Journal of Medicinal Chemistry. 2002;45(12):2615-23. PMID: 12036371. Available from: https://doi.org/10.1021/jm020017n.

[44] Egan WJ, Merz KM, Baldwin JJ. Prediction of Drug Absorption Using Multivariate Statistics. Journal of Medicinal Chemistry. 2000;43(21):3867-77. PMID: 11052792. Available from: https://doi.org/10.1021/jm000292e.

[45] Muegge I, Heald SL, Brittelli D. Simple Selection Criteria for Drug-like Chemical Matter. Journal of Medicinal Chemistry. 2001;44(12):1841-6. PMID: 11384230. Available from: https://doi.org/10.1021/jm015507e.

[46] Delaney JS. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. Journal of Chemical Information and Computer Sciences. 2004 May;44(3):1000-5. Available from: https://doi.org/10.1021/ci034243x.

[47] Ali J, Camilleri P, Brown MB, Hutt AJ, Kirton SB. Revisiting the General Solubil-

ity Equation: In Silico Prediction of Aqueous Solubility Incorporating the Effect of Topographical Polar Surface Area. Journal of Chemical Information and Modeling. 2012 Feb;52(2):420-8. Available from: https://doi.org/10.1021/ci200387c.

[48] Hanwell MD, Curtis DE, Lonie DC, Vandermeersch T, Zurek E, Hutchison GR. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. Journal of Cheminformatics. 2012 Aug;4(1):17. Available from: https://doi.org/10.1186/1758-2946-4-17.

[49] Mendez D, Gaulton A, Bento AP, Chambers J, DeÂ Veij M, FÃ©lix E, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Research. 2018 11;47(D1):D930-40. Available from: https://doi.org/10.1093/nar/gky1075.