Practical Session 4: Virtual Screening Based on Molecular Similarity and QSAR Modeling to Predict Activity Against EGFR (Epithelial Growth Factor Receptor) Protein Kinase

Student: Léopold Sossey Alaoui

I. Introduction : bibliographic review of EGFR (Epithelial Growth Factor Receptor) protein kinase

The epidermal growth factor receptor (EGFR) is a growth factor receptor that induces cell differentiation and proliferation upon activation through the binding of one of its ligands. The receptor is located at the cell surface, where the binding of a ligand activates a tyrosine kinase in the intracellular region of the receptor. This tyrosine kinase phosphorylates a number of intracellular substrates that activates pathways leading to cell growth, DNA synthesis and the expression of oncogenes such as fos and jun. EGFR is thought to be involved in the development of cancer, as the EGFR gene is often amplified, and/or mutated in cancer cells. [1]

Therefore, its signaling pathway is one of the most important pathways that regulate growth, survival, proliferation, and differentiation in mammalian cells. Reflecting this importance, it is one of the best-investigated signaling systems, both experimentally and computationally, and several computational models have been developed for dynamic analysis. [2]

Thousands of very effective EGFR inhibitors have been developed in the last decade. The known inhibitors originated from a very diverse chemical space but--without exception--all of them act at the Adenosine TriPhosphate (ATP) binding site of the enzyme. [3] Thanks to those known inhibitors we performed a virtual screening to select the most similar molecules with the best inhibitor activities and then built a Quantitative Structure-Activity Relationship (QSAR) model to predict activity against the EGFR protein kinase.

II. Materials and methods

a. Virtual Screening Based on Similarity Principles

i. Dataset

We used a dataset of 6,213 molecules reported as inhibitors of the EGFR protein kinase. Each entry contained three descriptors:

- molecule_chembl_id, the unique identifier from the ChEMBL database,
- SMILES, the structural representation of the molecule,
- pI50, the biological activity endpoint.

Gefitinib (ChEMBL ID: CHEMBL939), a clinically validated EGFR inhibitor, was selected as the reference compound for similarity-based virtual screening.

ii. Fingerprint Generation and Similarity Computation

Question 1: What are MACCS and Morgan fingerprints, and how do they differ ?

MACCS fingerprints are 166 bit structural keys based on predefined SMARTS patterns. Each bit corresponds to the presence or absence of a specific chemical substructure. They provide a fixed, interpretable representation of molecular fragments.

Morgan fingerprints, by contrast, are circular fingerprints generated by iteratively hashing atom-centered environments up to a defined radius. They capture local topological neighborhoods and produce a more flexible, data-driven representation of molecular structure.

Thus, MACCS fingerprints rely on predefined structural motifs, whereas Morgan fingerprints encode topological environments through iterative hashing.

Question 2: What is the difference between the Tanimoto and Dice similarity coefficients ?

Tanimoto coefficient allows the comparison of two molecules, based on their fingerprints Sab = $C\ /\ (A + B - C)$ A and B are the bit number of the considered fingerprint equal to 1, respectively in the two considered molecules and C the bits equal to 1 in both fingerprints. Sab is the similarity rate between two molecules, between 0 and 1. At 1, molecules are identical.

Dice is similar to Tanimoto, but the intersection is counted twice. By consequence, it is more sensible than Tanimoto on similarity.

iii. Virtual Screening Procedure

Fingerprints were computed using the RDKit library. For each molecule, similarity to Gefitinib was calculated using both Tanimoto and Dice coefficients, applied to both MACCS and Morgan fingerprints.

This resulted in four screening combinations:

1. MACCS + Tanimoto
2. MACCS + Dice
3. Morgan + Tanimoto
4. Morgan + Dice

The molecules retained by the best virtual screening combination, meaning the combination that retains the most similar molecules with the best inhibition activities were saved in a CSV file (similarity greater than 0.5). The filtered dataset was exported as a CSV file for QSAR modeling.

b. QSAR model creation with the filter molecular dataset

i. QSAR Modeling

Question 3: What are the main steps required to build a QSAR model ?

QSAR modeling typically involves the following steps:

1) Compilation of a dataset of chemical compounds.
2) Collection of experimental activity data for the target property.
3) Calculation of molecular descriptors.
4) Selection or filtering of descriptors to reduce redundancy.
5) Construction of a statistical or machine‑learning model relating descriptors to activity.
6) Validation of the model using internal or external validation strategies.
7) Use of the validated model to predict the activity of new compounds.

ii. QSAR Workflow Implementation

QSAR modeling was performed using the KNIME analytics platform. The workflow included:

- calculation of molecular descriptors for the screened molecules,
- normalize the data,
- removal of highly correlated descriptors,
- construction of a linear regression model,
- validation using leave‑one‑out cross‑validation (LOOCV),
- prediction of pIC50 values for the filtered dataset.

This workflow enabled the identification of key structural features associated with EGFR inhibition and allowed the prediction of biological activity for new candidate molecules using the validated linear QSAR model.

III. Results and discussion

a. Virtual Screening result

After computing similarity scores for each fingerprint metric combination, the distributions of similarity values were plotted to compare the behavior of the four virtual screening methods.

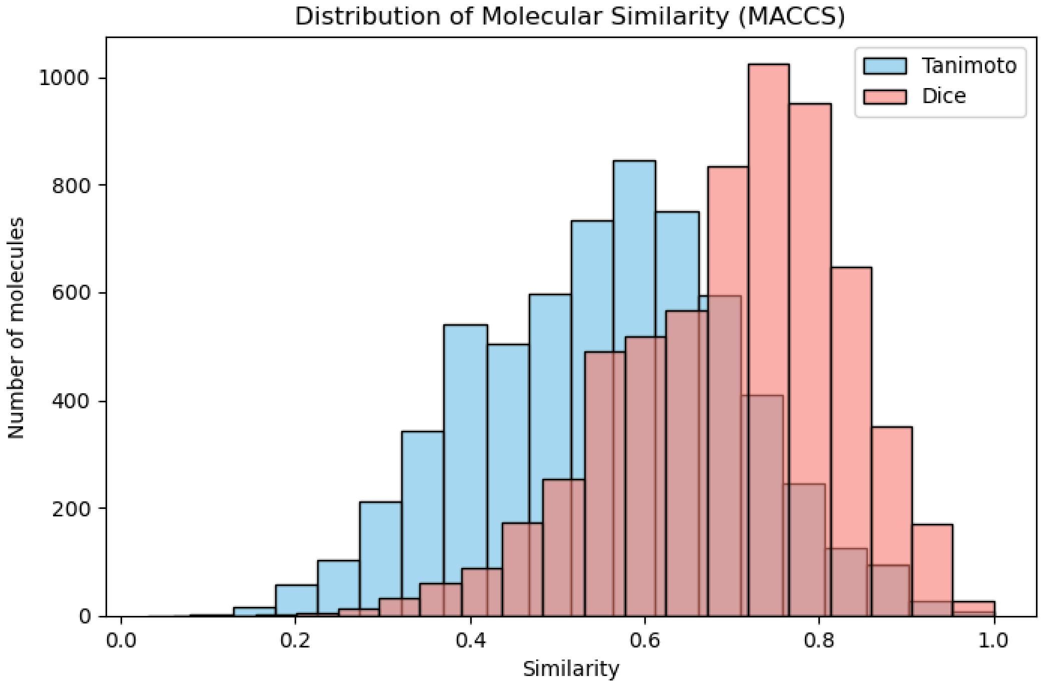


Figure 1: histogram of Tanimoto and Dice similarity based on MACCS fingerprint

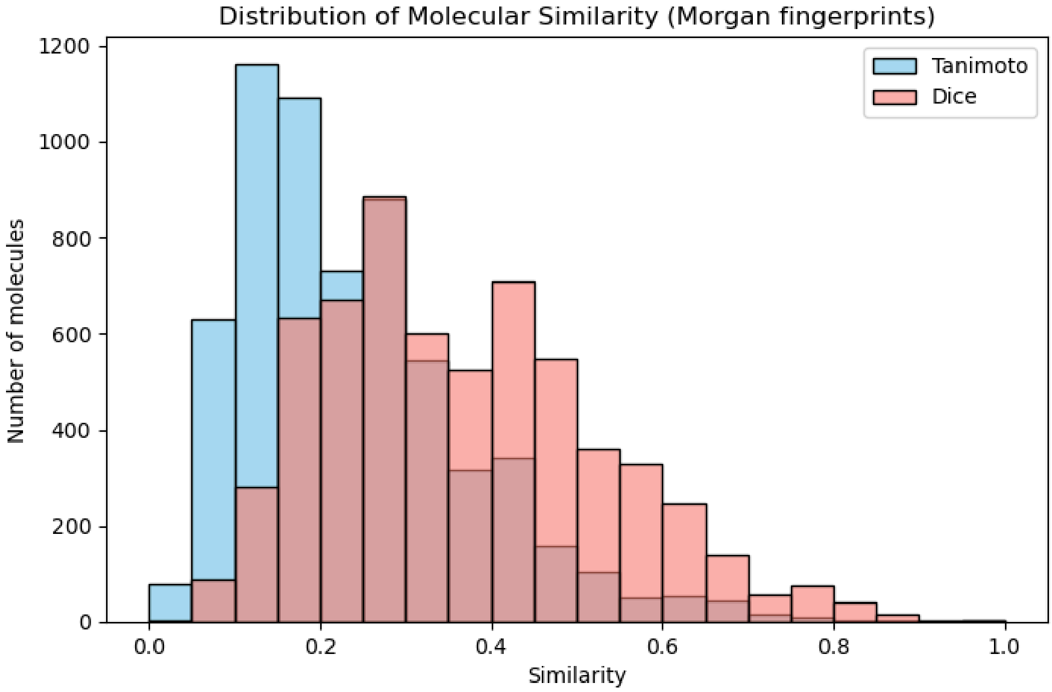


Figure 2: histogram of Tanimoto and Dice similarity based on Morgan fingerprint

The histograms illustrate the distribution of molecular similarity scores obtained with the two fingerprint types. The x-axis represents similarity values (0.0–1.0), and the y-axis indicates the number of molecules. Tanimoto similarity is shown in blue, and Dice similarity in red.

When similarity is computed using MACCS fingerprints, a large proportion of molecules exhibit similarity values above 0.5. This indicates that MACCS fingerprints are less restrictive, producing generally higher similarity scores.

In contrast, similarity values computed using Morgan fingerprints are mostly below 0.5, demonstrating that Morgan fingerprints are more discriminative and better at distinguishing structural differences between molecules.

From these distributions, the number of molecules retained with a similarity threshold of 0.5 follows the trend:

MACCS + Dice > MACCS + Tanimoto > Morgan + Dice > Morgan + Tanimoto

However, selecting the best virtual screening method requires not only maximizing the number of retained molecules but also prioritizing those with the highest biological activity. For this reason, the pIC50 values of the selected molecules (similarity > 0.5) were compared across the four methods.

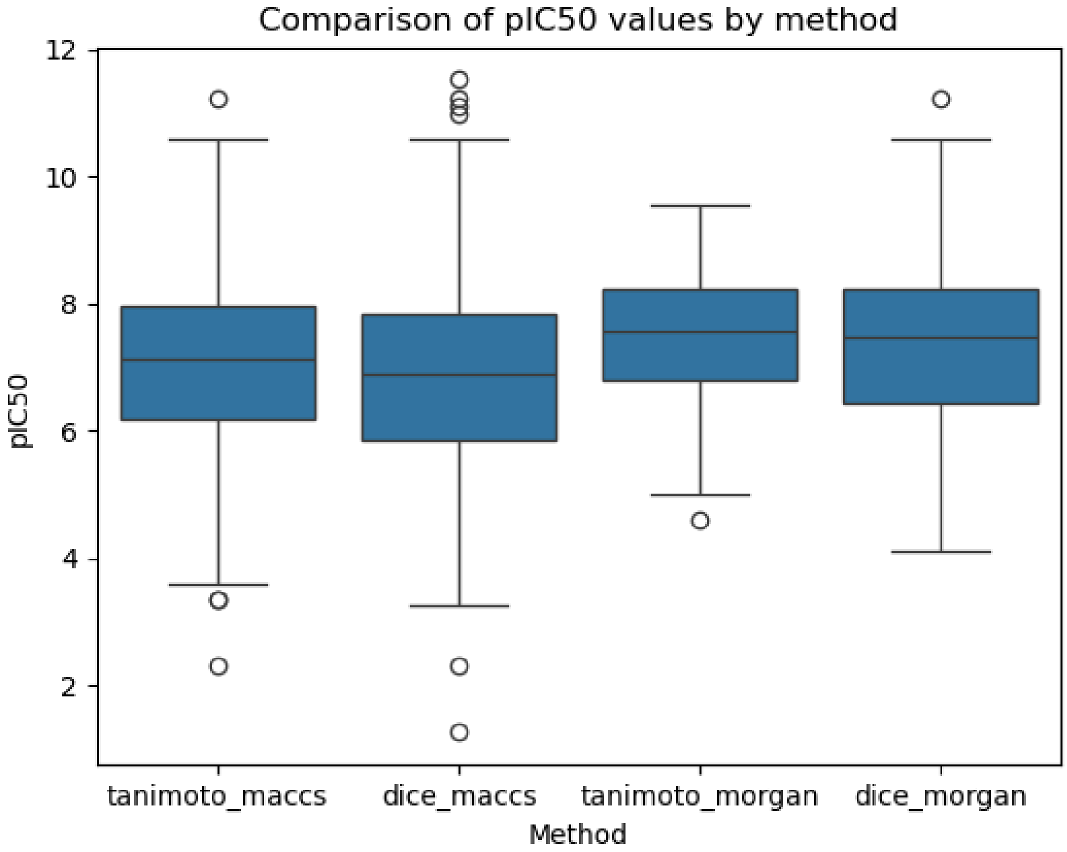


Figure 3: histogram of Tanimoto and Dice similarity based on Morgan fingerprint

The boxplot shows that the Tanimoto + Morgan combination yields the highest median pIC50, indicating that it preferentially selects molecules with stronger inhibitory activity against EGFR. Although the Dice + Morgan method shows a slightly lower median pIC50, it retains a much larger number of molecules above the similarity threshold.

For this reason, the Dice + Morgan based method was retained for the subsequent QSAR modeling steps.

Due to the strong imbalance in sample sizes and the presence of repeated pIC50 values, formal statistical comparisons (ANOVA or Kruskal–Wallis) were not considered reliable and were therefore not performed.

c. QSAR model result:

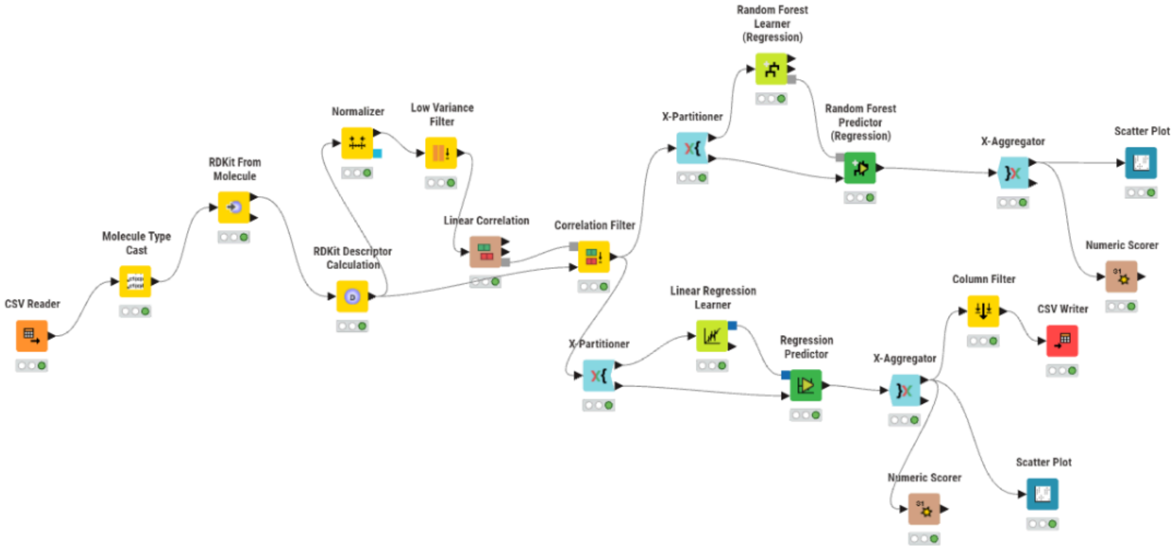


Figure 4 : KNIME workflow for QSAR modeling

The figure presents the complete KNIME workflow used to build and evaluate QSAR models based on the molecular descriptors of the screened compounds. It illustrates each step of the cheminformatics pipeline. The workflow starts with descriptor normalization and the removal of highly correlated descriptors using a correlation threshold of 0.925. Two predictive models are then trained: a linear regression model and a random forest model.

Both models are evaluated using the leave-one-out cross-validation (LOOCV) method, meaning the model is trained on all molecules except one, which is kept aside for prediction. This process is repeated for every molecule in the dataset.

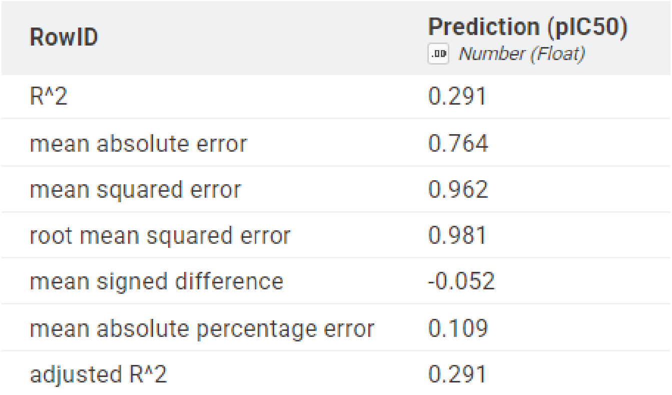
Questions:

What is the equation associated with your QSAR model here?

After exporting the prediction table obtained with the X-Aggregator node in CSV format, the linear regression equation computed by R is:

```
pIC50 =  2.9307 +  0.7232 * SlogP + 0.0261 * TPSA + 1.9229 * NumLipinskiHBA +
-0.3047 * NumLipinskiHBD + 0.2186 * NumRotatableBonds + -0.2043 * NumHBA +
-0.6618 * NumAmideBonds + -1.9624 * NumHeteroAtoms + 0.1797 *
NumStereocenters + -0.1429 * NumUnspecifiedStereocenters + 0.8247 * NumRings
+ -0.2345 * NumAromaticRings + -1.0094 * NumSaturatedRings + NA *
NumAliphaticRings + 0.2072 * NumAromaticHeterocycles + 1.143 *
NumSaturatedHeterocycles + 0.1403 * NumAliphaticHeterocycles + NA *
NumAromaticCarbocycles + NA * NumSaturatedCarbocycles + -3.0976 *
FractionCSP3 + 1.1817 * Chi1v + 0.5667 * Chi4v + 0.5102 * Chi4n + 1.3646 *
HallKierAlpha + -1.2344 * kappa2 + 0.2238 * slogp_VSA1 + 0.2389 * slogp_VSA2
+ 0.2138 * slogp_VSA3 + 0.1898 * slogp_VSA4 + 0.18 * slogp_VSA5 + 0.1959 *
slogp_VSA6 + 0.256 * slogp_VSA7 + 0.1206 * slogp_VSA8 + NA * slogp_VSA9 +
0.0813 * slogp_VSA10 + 0.0901 * slogp_VSA11 + 0.2597 * slogp_VSA12 + 0.0564 *
smr_VSA1 + -0.3187 * smr_VSA2 + -0.1455 * smr_VSA3 + -0.1598 * smr_VSA4 +
-0.1605 * smr_VSA5 + -0.195 * smr_VSA6 + -0.0894 * smr_VSA7 + NA * smr_VSA8 +
-0.013 * smr_VSA9 + NA * smr_VSA10 + 0.2327 * peoe_VSA1 + 0.1581 * peoe_VSA2
+ 0.199 * peoe_VSA3 + 0.1115 * peoe_VSA4 + 0.0997 * peoe_VSA5 + 0.0854 *
peoe_VSA6 + 0.0859 * peoe_VSA7 + 0.0728 * peoe_VSA8 + 0.0703 * peoe_VSA9 +
0.0558 * peoe_VSA10 + 0.0631 * peoe_VSA11 + 0.08 * peoe_VSA12 + -0.0063 *
peoe_VSA13 + NA * peoe_VSA14 + 2.3616 * MQN2 + -1.3973 * MQN3 + -3.1688 *
MQN4 + -3.8755 * MQN5 + -2.3494 * MQN6 + -3.6878 * MQN7 + 0.7716 * MQN8 +
0.959 * MQN9 + -0.4033 * MQN10 + NA * MQN11 + 1.5 * MQN13 + 1.5491 * MQN14 +
2.1477 * MQN15 + 0.0815 * MQN16 + NA * MQN18 + 0.6351 * MQN24 + 0.0463 *
MQN26 + -2.483 * MQN27 + -5.5115 * MQN28 + -8.3118 * MQN29 + -0.7893 * MQN30
+ -3.7972 * MQN31 + -7.0479 * MQN32 + 2.5188 * MQN33 + 0.5448 * MQN34 +
-0.2565 * MQN35 + -0.2631 * MQN36 + -1.3736 * MQN37 + NA * MQN38 + -0.2262 *
MQN39 + NA * MQN40 + 0.8793 * MQN41
```

Here is the model’s performance:



RowID	Prediction (pIC50) .00 Number (Float)
R^2	0.291
mean absolute error	0.764
mean squared error	0.962
root mean squared error	0.981
mean signed difference	-0.052
mean absolute percentage error	0.109
adjusted R^2	0.291

figure 5: result of the linear regression

The linear QSAR model built in this study shows weak predictive performance, with an R² of 0.291. This means the model only explains about 29% of the variation in pIC50 values, which is quite low. For a model to be acceptable, it usually needs to explain at least more than half of the variance. This low R² suggests that the relationship between the descriptors and pIC50 is probably not linear, and that a simple linear model is not enough to capture the complexity of the data.

What insights can you draw from analyzing the scatter plot and correlation matrix?

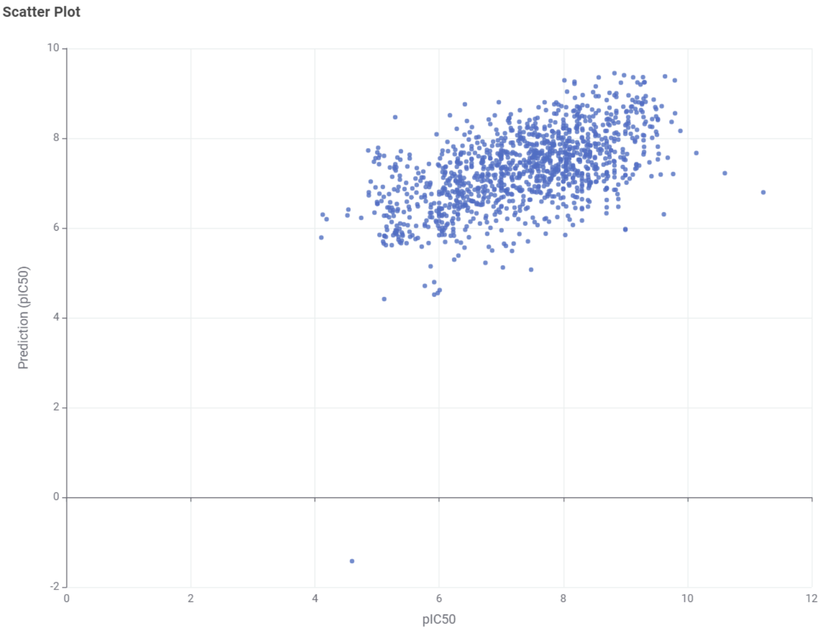
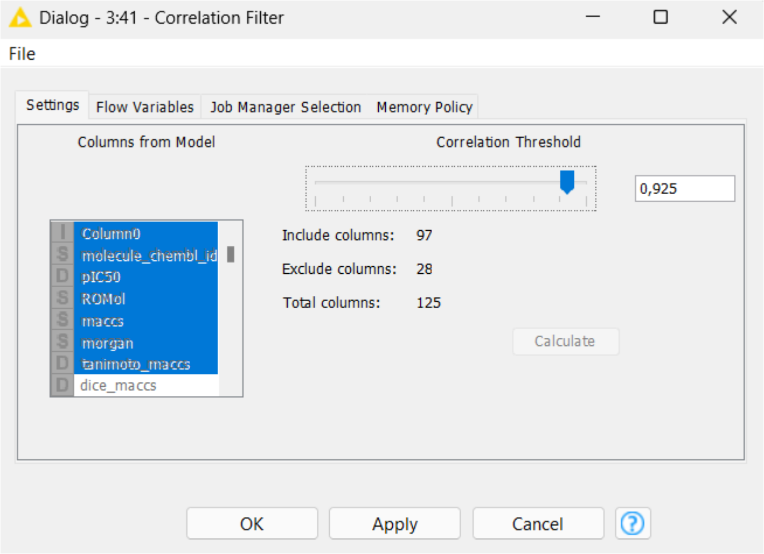


figure 6: Scatter plot Prediction(pIC50) ~ pIC50

Insights from the Scatter Plot:

Most data points are clustered along a diagonal trend from lower-left to upper-right, indicating that the model captures a general relationship between molecular descriptors and activity. The predicted values tend to be compressed between 6 and 10, while observed values span a wider range (4 to 12). This suggests the model may underpredict high activity and overpredict low activity. Additionally, there is one clear outlier with a negative predicted pIC50, indicating that the model struggles with atypical compounds.



Columns from Model

I Column0
S molecule_chembl_id
D pIC50
S ROMol
S maccs
S morgan
D tanimoto_maccs
D dice_maccs

Correlation Threshold: 0,925

Include columns: 97

Exclude columns: 28

Total columns: 125

figure 7: Insights from correlation matrix

Insights from the Correlation Matrix:

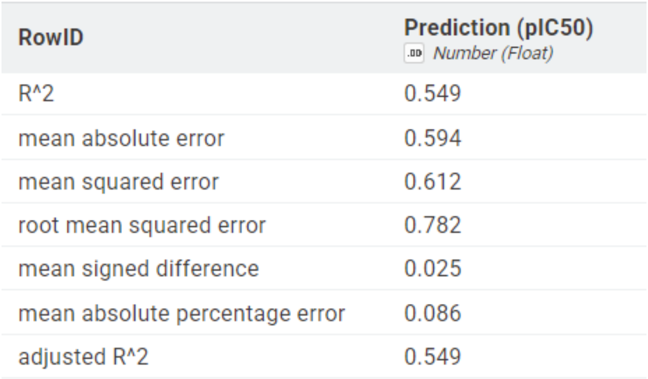
After applying the correlation filter with a threshold of 0.925, 28 descriptors were removed out of 125 total, leaving 97 for modeling. This means that some descriptors were highly correlated with each other and didn't bring new information. Removing them helps reduce redundancy and avoids problems like overfitting or misleading results. The remaining descriptors are more independent and should improve the stability and reliability of the model.

What can you say about the molecules identified as "outliers"?

Outliers are molecules whose predicted pIC50 values differ significantly from their observed values. In the scatter plot, they appear as points far from the diagonal line, indicating poor prediction accuracy. These molecules may have unusual chemical features not well represented in the training set, making them hard to model.

What solutions would you suggest to improve this model ?

Several strategies can be implemented to improve the predictive performance of the QSAR model. The current linear regression explains only a small portion of the variance ($R^2 = 0.291$), indicating that the relationship between molecular descriptors and pIC50 is likely non linear and more complex than what a linear model can capture. To address this, a non linear approach using a Random Forest model was tested, which is better at learning non linear patterns. The performance improved significantly: the R² increased from 0.291 to 0.549, meaning the model explains more than half of the variance in pIC50 values, which means it captures a real and meaningful relationship between the descriptors and the activity.



RowID	Prediction (pIC50) Number (Float)
R^2	0.549
mean absolute error	0.594
mean squared error	0.612
root mean squared error	0.782
mean signed difference	0.025
mean absolute percentage error	0.086
adjusted R^2	0.549

figure 8: Result for random forest model

Conclusion

The goal of this study was to virtually screen molecules similar to known EGFR protein kinase inhibitors and then use these selected compounds to train a model capable of predicting whether a molecule is likely to inhibit the kinase. Although the Morgan + Tanimoto method produced the highest median pIC50 among the selected molecules, the Morgan + Dice combination offered the best balance between activity enrichment and dataset size, and was therefore chosen for QSAR modeling.

A linear regression model was first trained, but its performance was poor, explaining only 29% of the variance. In contrast, the Random Forest model improved the predictive power to an R² of 0.549, which begins to be acceptable for QSAR studies. However, this performance remains far below the values typically reported in scientific publications, where well-optimized models often reach R² values above 0.9.

Overall, this work highlights the importance of selecting appropriate fingerprints and applying proper descriptor filtering.

Bibliography

[1] Voldborg BR, Damstrup L, Spang-Thomsen M, Poulsen HS. Epidermal growth factor receptor (EGFR) and EGFR mutations, function and possible role in clinical trials. Ann Oncol. 1997 Dec;8(12):1197-206. doi: 10.1023/a:1008209720526. PMID: 9496384.

[2] Oda K, Matsuoka Y, Funahashi A, Kitano H. A comprehensive pathway map of epidermal growth factor receptor signaling. Mol Syst Biol. 2005;1:2005.0010. doi: 10.1038/msb4100014. Epub 2005 May 25. PMID: 16729045; PMCID: PMC1681468.

[3] Szántai-Kis C, Kövesdi I, Eros D, Bánhegyi P, Ullrich A, Kéri G, Orfi L. Prediction oriented QSAR modelling of EGFR inhibition. Curr Med Chem. 2006;13(3):277-87. doi: 10.2174/092986706775476098. PMID: 16475937.