**Background:**

We have a dataset which contains 16 compounds, and it belongs to the same class. The Primary Goal was to Develop the Machine learning based QSAR Model [1] to predict the IC50 values of novel compounds and to suggest new analog compounds by modifying the structure slightly and determine whether those changes improve biological activity. The Dataset was too small to for any type of machine learning model to get better result on test data later therefore to tackle such challenge we built top 5 compounds from the actual dataset and built the analogs through python libraries based on modifying groups like CH₃, OH, Cl, Br, etc. also called rule-based substitutions. Each analog was processed in the same way descriptors were computed, and pIC50 was predicted using the trained Elastic Net model.[4]

**Methodology:**

**Descrtors generation:**

Convert the IC50 values into the pIC50 and then We compute 1600+ Descriptors for parent compounds (16 compounds given at the start) through Modred [3]. Preprocessing and cleaning data is required for robust Model building, so we remove duplicates values there , constant due to same value across the column and Null values or zero value.  
**QSAR Model Training**

We built models on ElasticNet regression Algorithm to predict the activity of compounds. The model was trained solely on the parent compounds (original 16 molecules), and the performance was evaluated on a test split R² Score: High, indicating that the model could capture trends in the training data.RMSE and MAE Low, confirming that predicted pIC50 values were close to the actual experimental results.Accuracy (±0.5 pIC50): High confirming the model’s reliability for approximate potency estimation. This trained model was then used to predict the pIC50 of novel analogs.

R² Score: 0.7877, MAE: 0.1004, RMSE: 0.1173, Accuracy (±0.5 pIC50): 100.00%

So, we have almost 80% of learning rate model pick the hidden pattern in the data and learns.

**Feature Importance Analysis:**

For the structural Modification and Feature importance we used SHAP (SHapley Additive exPlanations) python based Library in order to understand which descriptors influenced activity the most and it shows the following descriptors envoled.

Top Influential Features:

1. PEOE\_VSA7 (charge distribution)
2. MaxAbsPartialCharge
3. MinEStateIndex
4. NHOHCount
5. BCUT2D\_MRLOW

**Analog Generations:**

In this project, new analog compounds were designed by applying a set of functional group replacement rules. Specifically, commonly found bioactive fragments such as –CH₃, –Cl, and –OH were substituted with chemically meaningful alternatives like –F, –NO₂, and –NH₂. The idea behind this strategy comes from the principles of Matched Molecular Pair Analysis (MMPA), which is widely used in medicinal chemistry to study how minor structural modifications affect biological activity [6]. While our method involved a simplified implementation using string replacements on SMILES to simulate chemical substitutions .it effectively allowed us to expand the chemical space for prediction. For future work or more precise transformations, tools like RDKit reaction SMARTS or dedicated matched-pair libraries could be integrated for chemically accurate analog design. Once generated, these analogs were evaluated using the trained QSAR model to estimate their biological activity.

**Analog vs Parent Comparison**

From the comparison we clearly see that model picks the small differences between analogs and parent. In-Silico QSAR model tells us original compounds has more accurate as due to the changing structure they cannot deviate as much.  
It correctly identified that small chemical modifications were unlikely to improve pIC50.  
This gives useful SAR guidance.[5]

Some of the analogs show better value than actual value like compound ID: FI-3-10

It has more pIC50 value than the actual value. Which is very good but still the change is very low. Hence, we can tell based on Model predictions that by modifying groups like CH₃, OH, Cl, Br, with one another cannot make big difference. i-e functional group swaps,("Cl", "F"), ("Br", "Cl"), ("CH3", "OH"), ("OH", "NH2"), ("NO2", "CN"), ("CH3", "NO2") here Clorine change their positions with the Florine and Bromine with Clorine and so on therefore our parent compounds was validated.

A graph with red and blue lines

AI-generated content may be incorrect.

Plot: the graph shows a positive linear relationship between the parent pIC50 values and the predicted pIC50 values of their analogs. The regression line suggests that higher activity in the parent compound generally corresponds to higher predicted activity in its analogs. However, the spread of the data points around the regression line and the confidence interval indicate the degree of variability and uncertainty in these predictions. The labels on the points allow for the identification of specific analogs and their predicted vs. parent pIC50 values

**Model Prediction and lead generation:**

The top 3 parent compounds still represent the most promising candidates.These compounds demonstrated the highest predicted pIC50 and were more stable in terms of model confidence. These can be considered final "lead compounds" from this in-silico study.

Table: shows different analogs leads with predicted activity value and actual activity value.

|  |  |  |
| --- | --- | --- |
| **Parent\_ID** | **Predicted\_pIC50** | **Actul\_pIC50** |
| FI-3-14 | 4.649018 | 4.935542011 |
| FI-3-9 | 4.916689 | 4.961379838 |
| FI-3-9 | 4.746819 | 4.961379838 |
| FI-3-10 | 4.715094 | 4.97428461609865 |

**Conclusions:**

This work employed machine learning and molecular descriptors to create a predictive QSAR model. Analog creation increased the chemical space but did not outperform the original compounds in terms of activity. The top three lead compounds discovered by this model can now undergo advanced in-silico screening or experimental validation for future medication development. Perform molecular docking on the top three leads to investigate how they interact with the biological target. Conduct molecular dynamics (MD) simulations to evaluate binding stability and conformational behavior.Perform toxicity and ADMET estimates to ensure drug-likeness and safety profiles. Because these processes are computationally inexpensive in comparison to lab effort, they represent a natural development from the current QSAR modeling phase.

**References**:

[1] Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., ... & Tropsha, A. (2014). QSAR modeling: where have you been? Where are you going to?. *Journal of Medicinal Chemistry*, 57(12), 4977-5010.

[2] Fuadah, Yunendah Nur, et al. "QSAR Classification Modeling Using Machine Learning with a Consensus-Based Approach for Multivariate Chemical Hazard End Points." *ACS omega* 9.51 (2024): 50796-50808.

[3] Dan, Yuanyuan, et al. "Predicting the Toxicity of Drug Molecules with Selecting Effective Descriptors Using a Binary Ant Colony Optimization (BACO) Feature Selection Approach." *Molecules* 30.7 (2025): 1548.

[4] Hans, Chris. "Elastic net regression modeling with the orthant normal prior." Journal of the American Statistical Association 106.496 (2011): 1383-1393.

[5] Verma, Rajeshwar P., and Corwin Hansch. "Camptothecins: a SAR/QSAR study." *Chemical reviews* 109.1 (2009): 213-235.

[6] Tyrchan, Christian, and Emma Evertsson. "Matched molecular pair analysis in short: algorithms, applications and limitations." Computational and structural biotechnology journal 15 (2017): 86-90.