Task 6: AI-Based Hit Scoring / QSAR Prototype

Overview

The objective of Task 6 was to develop an AI-based hit scoring framework to predict compound bioactivity and prioritize candidates identified in Task 5. A QSAR pipeline was implemented using RDKit-derived molecular descriptors and Random Forest models to estimate pIC50 values and classify compounds as active or inactive. In addition to standard performance metrics, model interpretability and chemical space analysis were incorporated to assess reliability and guide hit prioritization.

Descriptor Generation and Dataset Preparation

Physicochemical descriptors were generated for each compound using RDKit, including molecular weight (MW), LogP, topological polar surface area (TPSA), hydrogen bond donors (HBD), hydrogen bond acceptors (HBA), and rotatable bonds. These features formed the input matrix for model training, while experimentally derived pIC50 values served as the regression target. Compounds were also labeled as active or inactive using a pIC50 threshold to enable classification analysis.

To obtain a realistic estimate of model generalization, scaffold-based splitting was applied using Murcko scaffolds, ensuring that structurally distinct chemotypes were separated between training and test sets.

Model Training and Performance Evaluation

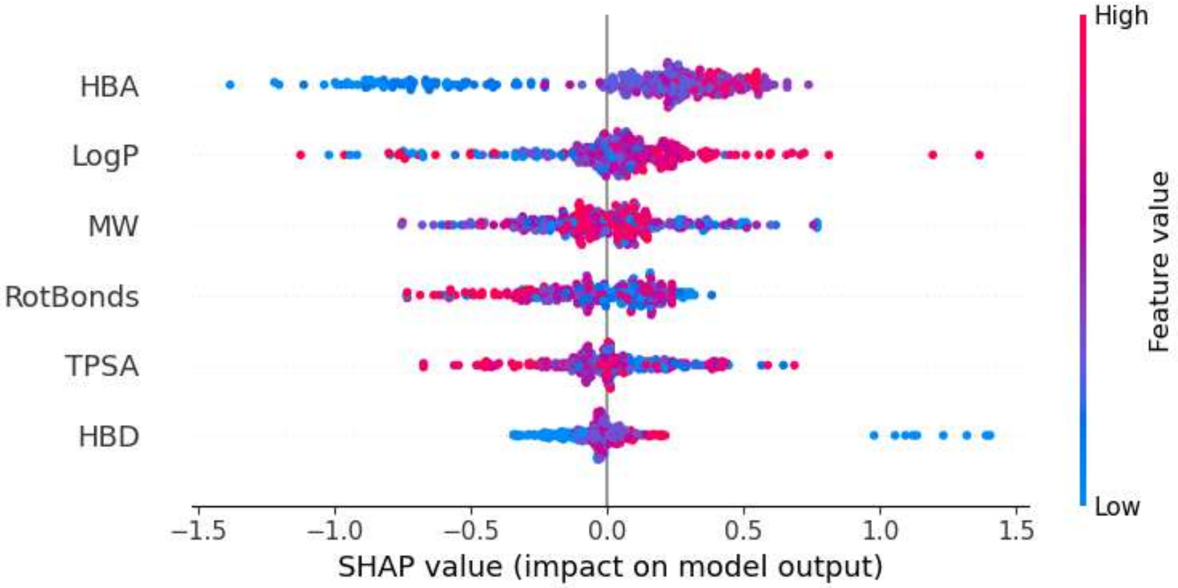
A Random Forest regressor was trained to predict pIC50 values, and a Random Forest classifier was used to distinguish active from inactive compounds. Model performance was evaluated using both regression and classification metrics. The regression model achieved an RMSE of approximately 1.07 pIC50 units with an R^2 of 0.32, indicating moderate predictive power typical of early-stage QSAR models trained on limited datasets. Classification performance was assessed using ROC analysis, which demonstrated discrimination above random baseline.

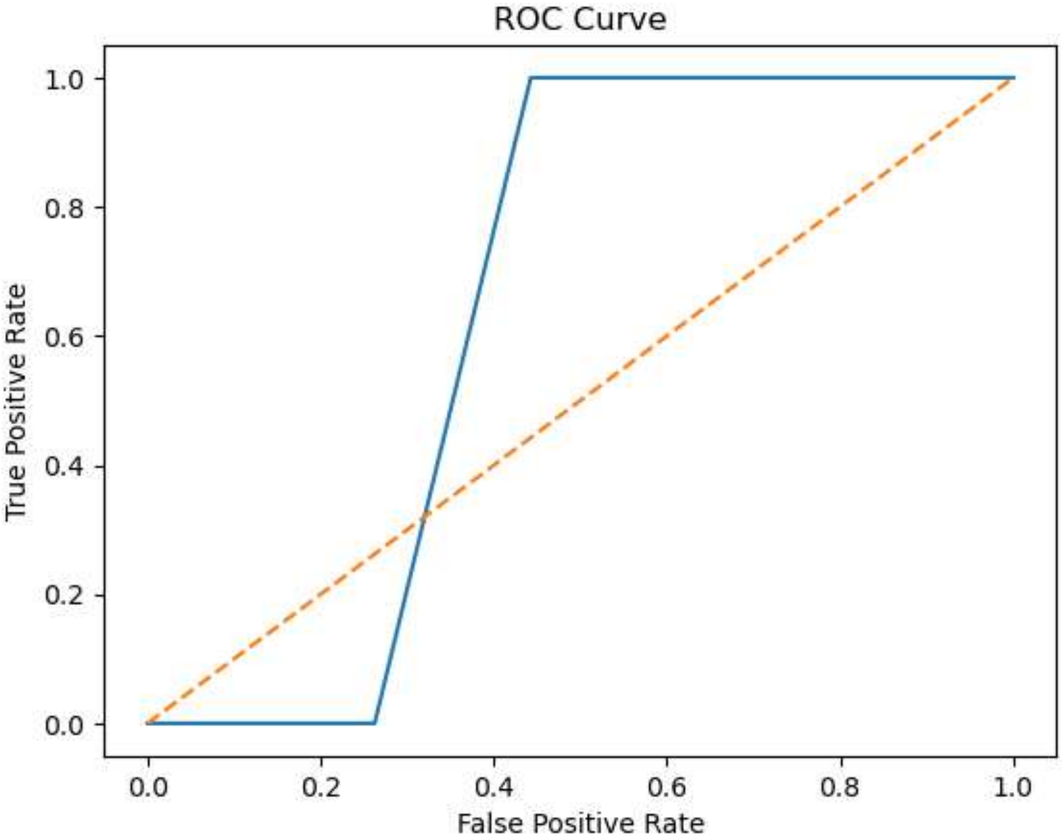
These results support the use of the model for compound prioritization rather than precise activity prediction.

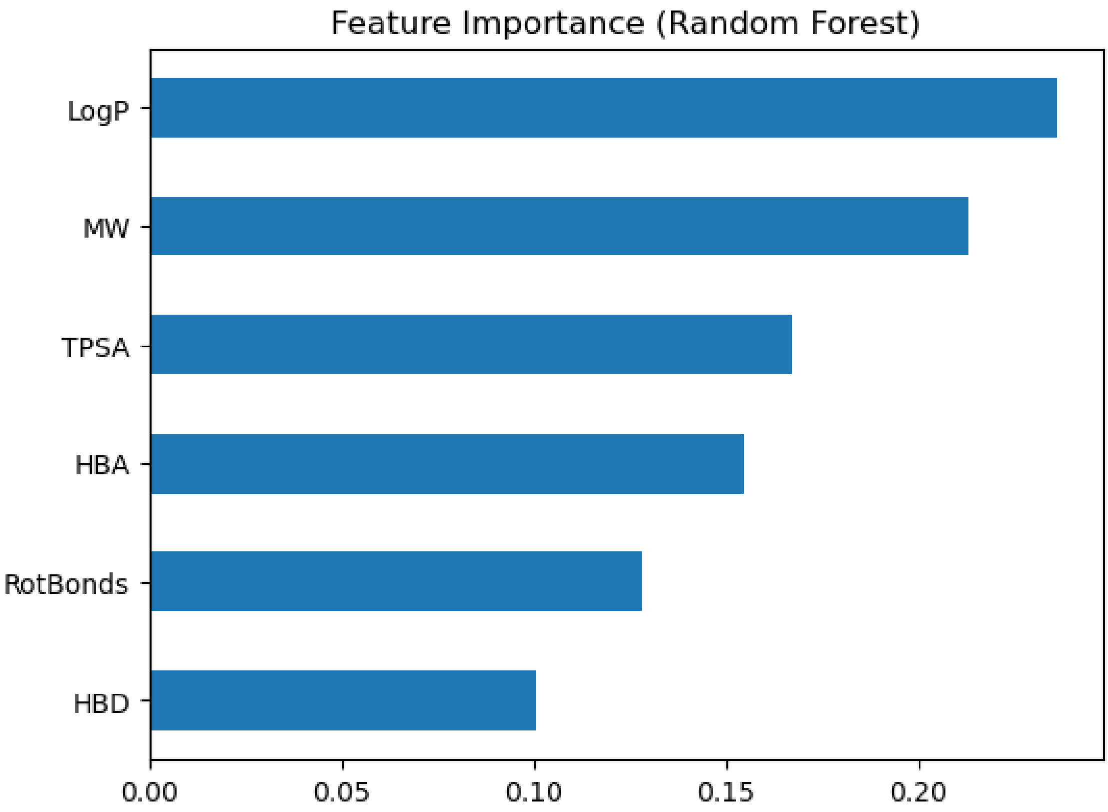
Explainability and Chemical Space Analysis

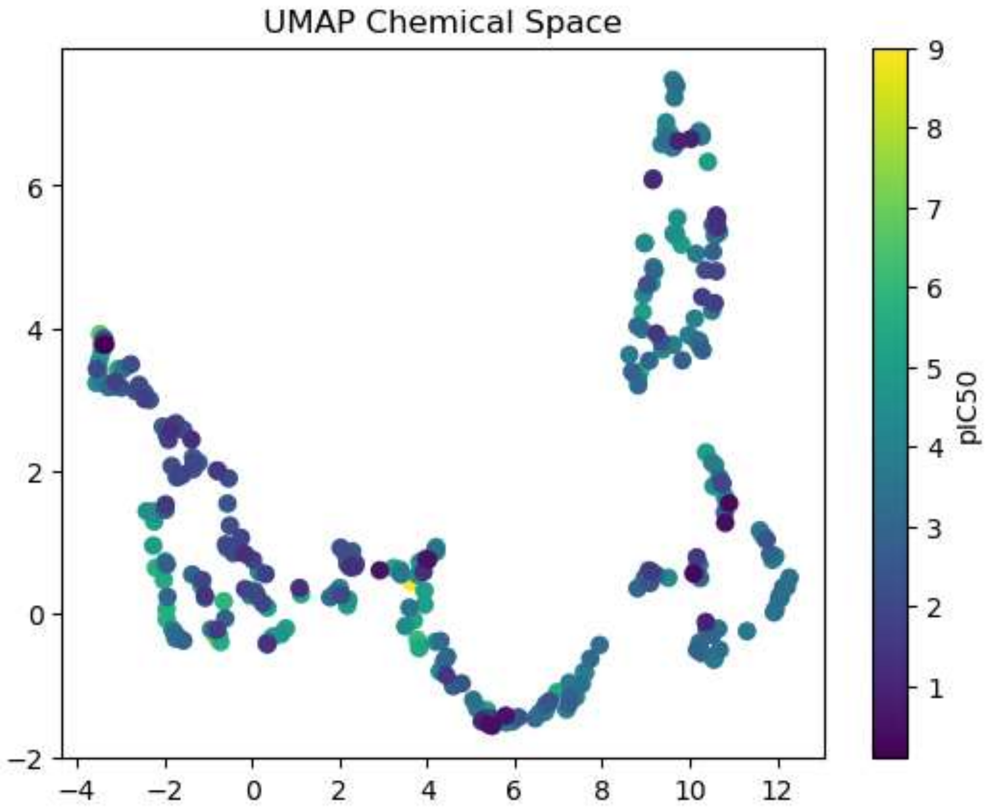
To interpret model predictions, SHAP (SHapley Additive exPlanations) analysis was applied. Hydrogen bond acceptors, LogP, and molecular weight emerged as the most influential features, followed by rotatable bonds and TPSA. Increased HBA and moderate lipophilicity were associated with higher predicted activity, while excessive molecular weight and flexibility negatively impacted predictions. These trends are chemically intuitive for DNA gyrase inhibitors and indicate that the model captures meaningful structure–property relationships.

Chemical diversity and structure–activity relationships were further explored using UMAP dimensionality reduction. The resulting chemical space projection revealed multiple distinct clusters, with high-activity compounds localized in specific regions. This suggests the presence of activity-enriched neighborhoods and supports the validity of the learned SAR patterns. An interactive UMAP visualization was generated using Plotly, enabling compound-level exploration of chemical space and facilitating identification of high-pIC50 clusters.


HBA
LogP
MW
RotBonds
TPSA
HBD
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
SHAP value (impact on model output)
High
Feature value
Low


ROC Curve
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate





AI-Based Hit Scoring and Reliability Assessment

The trained regression model was used to generate predicted pIC50 values for all compounds, producing an ML-ranked hit list to complement rule-based filtering from Task 5. This AI-based scoring layer provides an additional prioritization criterion for selecting promising candidates.

Several limitations affect model reliability, including small dataset size, heterogeneous ChEMBL assay conditions, and potential chemical series bias. Furthermore, the use of simple physicochemical descriptors restricts representation of detailed structural features. Consequently, the model is intended as a screening and prioritization tool rather than a definitive predictor of bioactivity. Despite these constraints, the integration of scaffold-aware validation, SHAP explainability, and chemical space visualization establishes a robust QSAR prototype suitable for early-stage hit ranking.

Outcome

Task 6 delivered an explainable, scaffold-aware QSAR workflow combining RDKit descriptors, Random Forest modeling, ROC and RMSE-based evaluation, SHAP feature interpretation, UMAP chemical space analysis, and ML-driven hit scoring. This framework provides a practical AI-based prioritization layer that complements physicochemical filtering and supports informed selection of candidate inhibitors for downstream validation.