ChemAI

ADMET PREDICTION

1. Absorption (HIA)

Model Goal:
The objective of this work is to develop a reliable, data-driven machine learning system capable of predicting human intestinal absorption by combining two key molecular determinants — permeability and aqueous solubility. By integrating large-scale experimental datasets with advanced cheminformatics features and optimized learning algorithms.

1. **Dataset Summary**:A large permeability dataset was extracted from ChEMBL, containing multiple assay formats (Papp, PCaco-2, permeability coefficient, etc.).

Loaded raw rows	35,437
After outlier filtering	29,145
Unique clean SMILES	21,074

- Final Dataset Structure
- 1 SMILES per molecule
- 7 RDKit physicochemical descriptors2048-bit Morgan fingerprints

2. Feature Engineering
RDKit descriptors (07):MolWt, MolLogP, TPSA, NumHDonors, NumHAcceptors, NumRotatableBonds, HeavyAtomCount,
Fingerprints:
Morgan Fingerprint (radius=2, 2048 bits)
Final X and y Shapes:
X: (21,074 molecules × 2057 features)
y: 21074

3. **Train/Test Split**
Train/Test Split
X_train: (16,859 × 2055)
X_test: (4,215 × 2055)
y_train: (16,859)
y_test: (4,215)

4. Model Development— Permeability Prediction (logPapp)
Optuna-Tuned LightGBM (Final Model)
Metric Value
- R^2: 0.6137
- MAE: 0.3608

Final chosen hyperparameters:
- "n_estimators": 939,
- "learning_rate": 0.06988,
- "num_leaves": 86,
- "max_depth": 12,
- "subsample": 0.7704,
- "colsample_bytree": 0.6071,
- "min_child_samples": 10

5. Model Development— Solubility Model (ESOL logS)
Dataset Summary
Raw ESOL dataset: (1128 × 10)
Valid molecules: 1128
Featurization success: 0 failures
Feature matrix: (1128 × 2057)
X_esol: (1128 × 2055)
y_esol: (1128,)

Train/Test Split
X_train_s: (902 × 2055)
X_test_s: (226 × 2055)

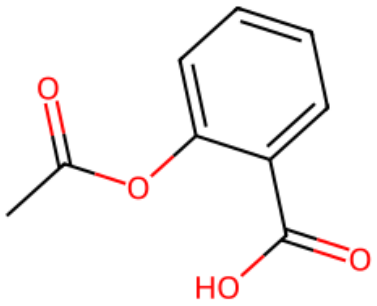
Model Performance (LightGBM)
R^2: 0.8982
MAE: 0.4808

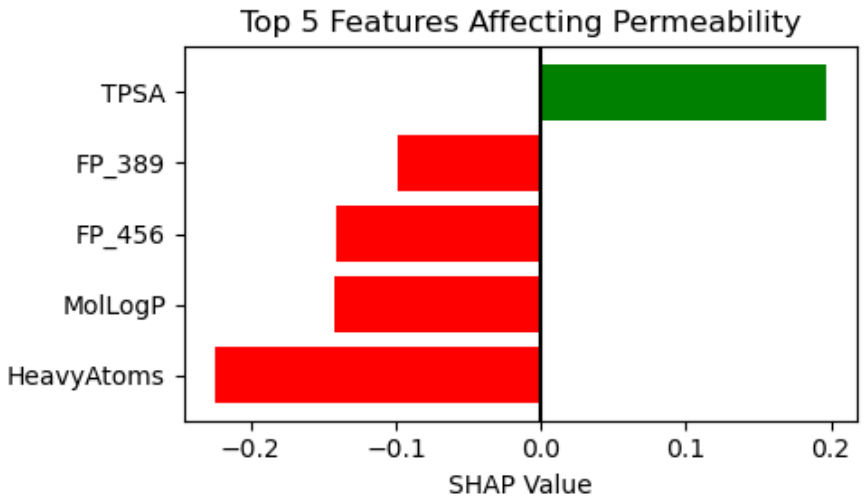
6. Combined Absorption (HIA) Scoring text
Permeability model → predicts logPapp and Solubility model → predicts logS
A combined absorption score (0–1) is computed by weighted scaling:
60% permeability
40% solubility
Thresholds were selected from experimental ADME ranges:
Score ≥ 0.5 → HIGH absorption
Score < 0.5 → LOW absorption
This approach provides a robust surrogate for HIA when experimental Caco-2 or PAMPA assays are unavailable.

Prediction:

🔹 Prediction Results

SMILES : CC(=O)Oc1ccccc1C(=O)O
Predicted logPapp : -5.432
Predicted logS : -1.474
Absorption score : 0.801
Absorption class : HIGH





🔹 Textual Explanation of Top Features:

- HeavyAtoms: decreases predicted permeability (SHAP = -0.225)
- TPSA: increases predicted permeability (SHAP = 0.197)
- MolLogP: decreases predicted permeability (SHAP = -0.142)
- FP_456: decreases predicted permeability (SHAP = -0.142)
- FP_389: decreases predicted permeability (SHAP = -0.099)

ChemAI

ADMET PREDICTION

2.Volume of Distribution (Vd)

Model Goal:
To build a robust, explainable machine-learning model that predicts the steady-state Volume of Distribution (Vd, L/kg) directly from molecular structure (SMILES).
This model supports ADMET profiling and early drug discovery by helping identify molecules with high or low tissue distribution.

1. **Dataset Summary**:The dataset was mined from ChEMBL, filtered and standardized to ensure high-quality Vd values:

Loaded raw rows	100,546
Filtered for Vd-related assays	32,423
Unique clean SMILES	13,583
Valid Vd (L/kg) retained	30,155
Final aggregated dataset	12,916

- Final Dataset Structure
- 1 SMILES per molecule
- Median Vd (L/kg) aggregated across assays
- 10 RDKit descriptors
- 2048-bit Morgan fingerprints

2. Feature Engineering
RDKit descriptors (10):MolWt, MolLogP, TPSA, NumHDonors, NumHAcceptors, NumRotatableBonds, FractionCSP3, HeavyAtomCount, RingCount,AromaticProportion
Fingerprints:
Morgan Fingerprint (radius=2, 2048 bits)
Final X and y Shapes:
X: (12,916 molecules × 2058 features)
y: log10(Vd) values

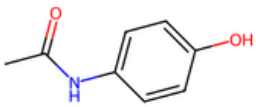
3. **Train-Test Splits**
After removing outliers using IQR filtering:
Final cleaned samples: 12,629
Train: 9,123
Validation: 1,611
Test: 1,895

4. Model Development
Optuna-Tuned LightGBM (Final Model)
Metric Value

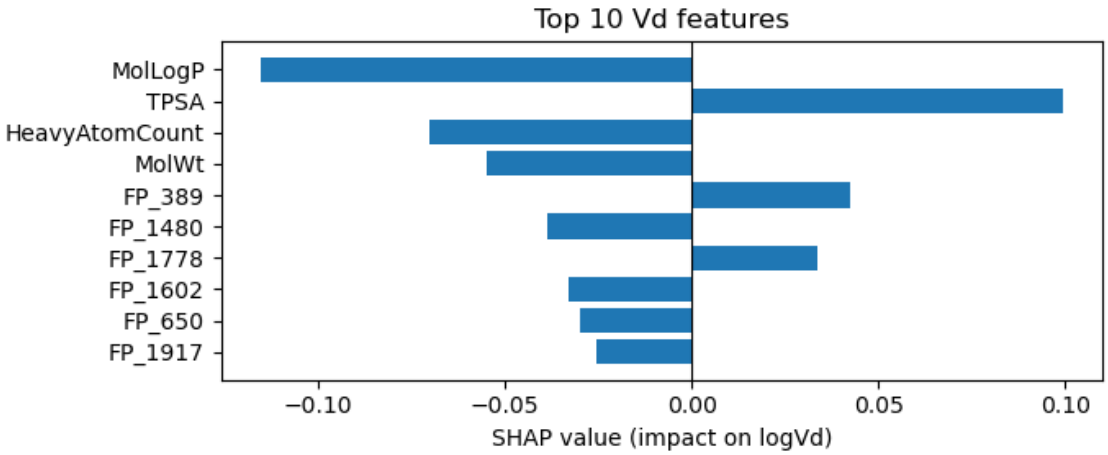
- R^2 0.4568
- MAE 0.3091
- RMSE 0.4063
- MSE 0.1651

Final chosen hyperparameters:
n_estimators: 681
max_depth: 10
num_leaves: 102
learning_rate: 0.0609
min_child_samples: 10
subsample: 0.809
colsample_bytree: 0.753
reg_alpha: 0.0900
reg_lambda: 0.6471

6.Prediction
SMILES: CC(=O)NC1=CC=C(C=C1)O
logVd: 0.0360
Vd (L/kg): 1.0865
=== TOP SHAP FEATURES ===



MolLogP	SHAP = -0.1153 (decreases)
TPSA	SHAP = +0.0993 (increases)
HeavyAtomCount	SHAP = -0.0702 (decreases)
MolWt	SHAP = -0.0550 (decreases)
FP_389	SHAP = +0.0426 (increases)
FP_1480	SHAP = -0.0388 (decreases)
FP_1778	SHAP = +0.0339 (increases)
FP_1602	SHAP = -0.0329 (decreases)
FP_650	SHAP = -0.0298 (decreases)



=== TEMPLATE SUMMARY ===

MolLogP (decreases, |SHAP|=0.12) | TPSA (increases, |SHAP|=0.10) | HeavyAtomCount (decreases, |SHAP|=0.07) | MolWt (decreases, |SHAP|=0.05) | FP_389 (increases, |SHAP|=0.04) | FP_1480 (decreases, |SHAP|=0.04) | FP_1778 (increases, |SHAP|=0.03) | FP_1602 (decreases, |SHAP|=0.03) | FP_650 (decreases, |SHAP|=0.03) | FP_1917 (decreases, |SHAP|=0.03)

ADMET PREDICTION

3.CYP450 Metabolism Prediction

Model Goal:
To develop accurate machine-learning models that can predict CYP450 inhibition (pIC_{50}) for the five major human enzymes, helping researchers identify metabolic liabilities, optimize lead compounds, and reduce experimental screening cost.

1. **Dataset Summary**:A large permeability dataset was extracted from ChEMBL, containing multiple assay formats (Papp, PCaco-2, permeability coefficient, etc.).

Total rows extracted from ChEMBL	57,050
After numeric IC_{50} filtering	50,215
After molar conversion	37,207
After removing invalid IC_{50}	37,194

pIC_{50} Value Distribution
Range: 0.024 → 12.878
Mean: 5.06
Final cleaned dataset: 37,194 rows
unique molecules:12,984

After grouping and aggregating (median pIC_{50} per SMILES–isoform):

CYP Isoform	Rows	Unique SMILES
CYP3A4	9,804	
CYP2D6	6,516	
CYP2C9	5,753	
CYP1A2	4,084	
CYP2C19	3,575	

2. Feature Engineering
RDKit descriptors (07):MolWt, MolLogP, TPSA, NumHDonors, NumHAcceptors, NumRotatableBonds, HeavyAtomCount,
Fingerprints:
Morgan Fingerprint (radius=2, 2048 bits)

- **3. Model Architecture**
- All CYP models use: LightGBM Regressor
- 80/20 train–test split
- Early stopping for convergence
- RMSE, MAE, and R^2 for evaluation
- Consistent fingerprint + descriptor feature set
- his provides strong performance, even across CYPs with differing data volumes.

4. Final Model Performance (Enzyme-wise)

CYP3A4 (n = 9,804)
R^2: 0.626
MAE: 0.355

CYP1A2 (n = 4,084)
R^2: 0.646
MAE: 0.366

CYP2C9 (n = 5,753)
R^2: 0.502
MAE: 0.333

CYP2C19 (n = 3,575)
R^2: 0.360
MAE: 0.344

CYP2D6 (n = 6,516)
R^2: 0.536
MAE: 0.334

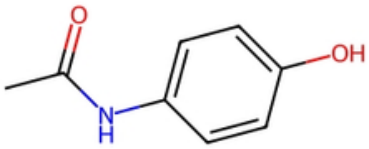
5. Interpretation

CYP3A4 and CYP1A2 showed the strongest predictive power, reflecting: larger dataset size, better assay consistency, well-defined substrate specificity

CYP2C19 exhibited lower R^2 due to: diverse chemistry, smaller dataset, higher experimental variability.
Despite this, the models provide robust quantitative inhibition predictions across all five isoforms.

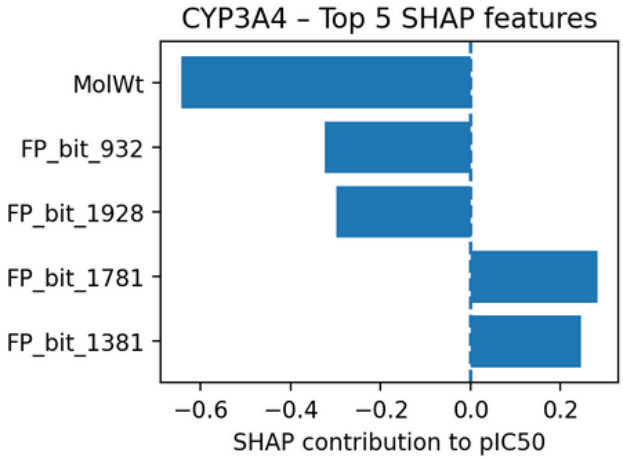
Prediction:

🔹 Prediction Results: paracetamol
SMILES : CC(=O)NC1=CC=C(C=C1)O

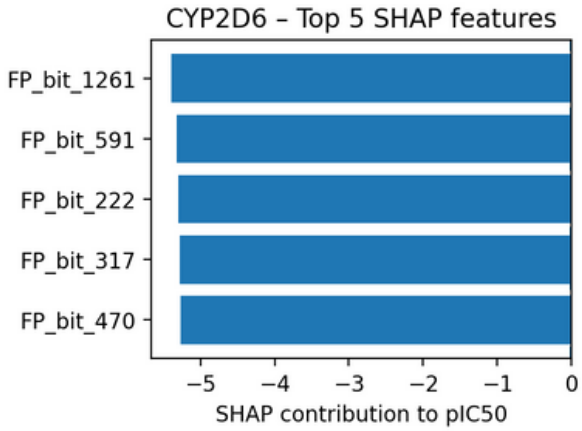


Enzyme	pIC50	pIC50_µm	Risk Class
CYP3A4	4.929	11.785626844383408	Weak inhibitor
CYP2D6	4.595	25.397958066525373	Weak inhibitor
CYP2C9	4.14	72.43239064362662	Weak inhibitor
CYP2C19	4.601	25.07772089013473	Weak inhibitor
CYP1A2	4.733	18.485217453399088	Weak inhibitor

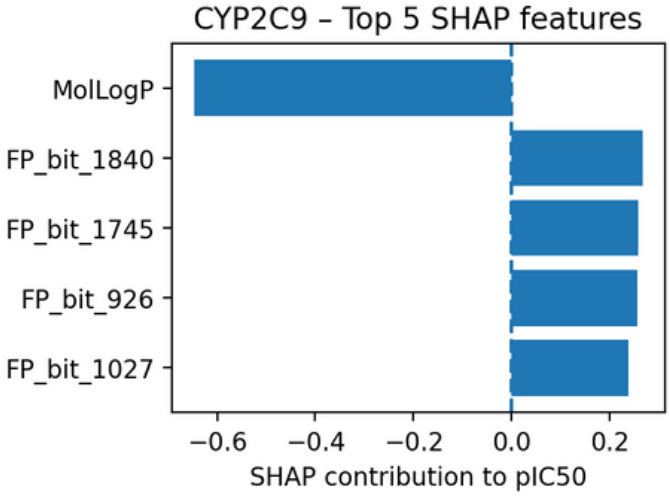
- Highest inhibition risk: CYP3A4 (Weak inhibitor), estimated IC50 ≈ 11.79 µM.



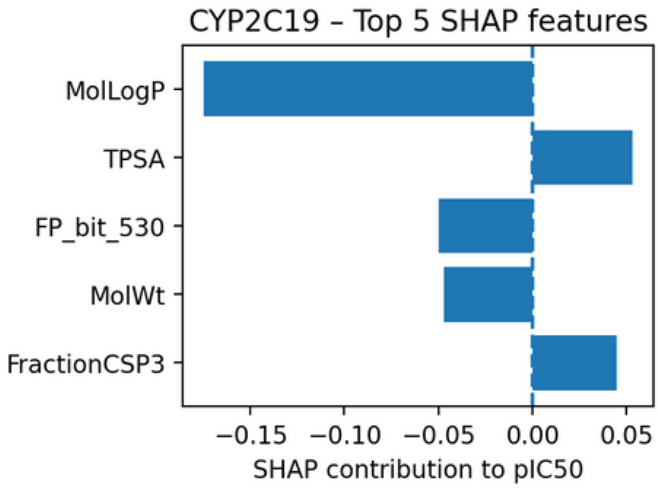
- MolWt decreases predicted CYP3A4 inhibition (SHAP = -0.643)
- FP_bit_932 decreases predicted CYP3A4 inhibition (SHAP = -0.323)
- FP_bit_1928 decreases predicted CYP3A4 inhibition (SHAP = -0.299)
- FP_bit_1781 increases predicted CYP3A4 inhibition (SHAP = 0.282)
- FP_bit_1381 increases predicted CYP3A4 inhibition (SHAP = 0.245)



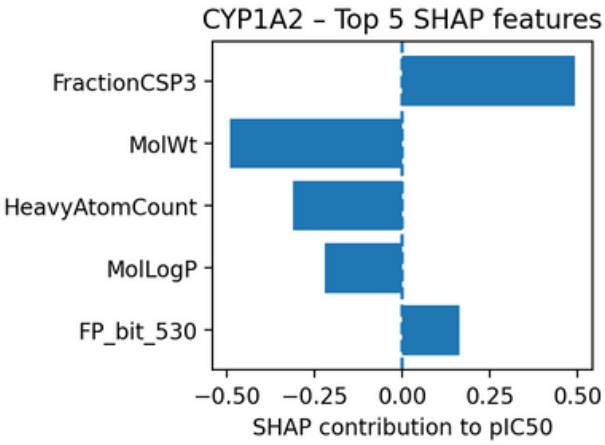
- FP_bit_1261 decreases predicted CYP2D6 inhibition (SHAP = -5.397)
- FP_bit_591 decreases predicted CYP2D6 inhibition (SHAP = -5.325)
- FP_bit_222 decreases predicted CYP2D6 inhibition (SHAP = -5.303)
- FP_bit_317 decreases predicted CYP2D6 inhibition (SHAP = -5.280)
- FP_bit_470 decreases predicted CYP2D6 inhibition (SHAP = -5.267)



- MolLogP decreases predicted CYP2C9 inhibition (SHAP = -0.647)
- FP_bit_1840 increases predicted CYP2C9 inhibition (SHAP = 0.267)
- FP_bit_1745 increases predicted CYP2C9 inhibition (SHAP = 0.258)
- FP_bit_926 increases predicted CYP2C9 inhibition (SHAP = 0.256)
- FP_bit_1027 increases predicted CYP2C9 inhibition (SHAP = 0.238)



- MolLogP decreases predicted CYP2C19 inhibition (SHAP = -0.175)
- TPSA increases predicted CYP2C19 inhibition (SHAP = 0.053)
- FP_bit_530 decreases predicted CYP2C19 inhibition (SHAP = -0.050)
- MolWt decreases predicted CYP2C19 inhibition (SHAP = -0.047)
- FractionCSP3 increases predicted CYP2C19 inhibition (SHAP = 0.045)



- FractionCSP3 increases predicted CYP1A2 inhibition (SHAP = 0.494)
- MolWt decreases predicted CYP1A2 inhibition (SHAP = -0.490)
- HeavyAtomCount decreases predicted CYP1A2 inhibition (SHAP = -0.310)
- MolLogP decreases predicted CYP1A2 inhibition (SHAP = -0.220)
- FP_bit_530 increases predicted CYP1A2 inhibition (SHAP = 0.164)

ChemAI

ADMET PREDICTION

4.Toxicity

Model Goal:
The objective of this project is to develop a unified, explainable AI-based toxicity prediction system capable of identifying potentially hazardous small molecules using a combination of LD_{50} (acute toxicity), Ames mutagenicity, and Tox21 toxicity endpoints. By integrating diverse datasets, structural alerts, and advanced machine-learning models, the system provides an early assessment of safety risks, enabling medicinal chemists to deprioritize toxic compounds and focus on safer, drug-like candidates.

1. **Dataset Summary**:The toxicity module combines three major datasets:

LD_{50} Acute Toxicity	4,743
Tox21 Toxicity Panel	7,588
Ames Mutagenicity	8,567

- Dataset Merging & Feature Engineering
- After cleaning and merging:
- Unified master dataset
- 18,255 total molecules
- After missing value cleanup: 17,729 molecules
- 1 SMILES per molecule
- 7 RDKit physicochemical descriptors2048-bit Morgan fingerprints

2. Feature Engineering
RDKit descriptors (07):MolWt, MolLogP, TPSA, NumHDonors, NumHAcceptors, NumRotatableBonds, HeavyAtomCount,
7 expert-defined alerts:
Nitrosamine, Aromatic amine, Epoxide, Nitroaromatic,Allylic halide,Alkylating group Michael acceptor

3. **Train/Test Split**
Train: 14,183 molecules
Test: 3,546 molecules
Class Distribution
Non-toxic: 79.1%
Toxic: 20.9%
Imbalance handled using scale_pos_weight = 3.78 in XGBoost.

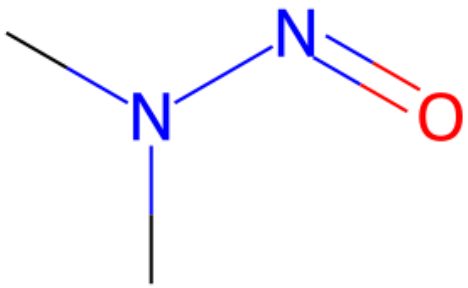
4. Model Development— XGBoost
The final model is based on XGBoost, chosen for its robustness with large, sparse, mixed-toxicity datasets.
XGBoost Configuration
300 estimators
max_depth = 8
learning_rate = 0.05
subsample = 0.8
colsample_bytree = 0.8
eval_metric = AUC
scale_pos_weight = 3.78
n_jobs = 4

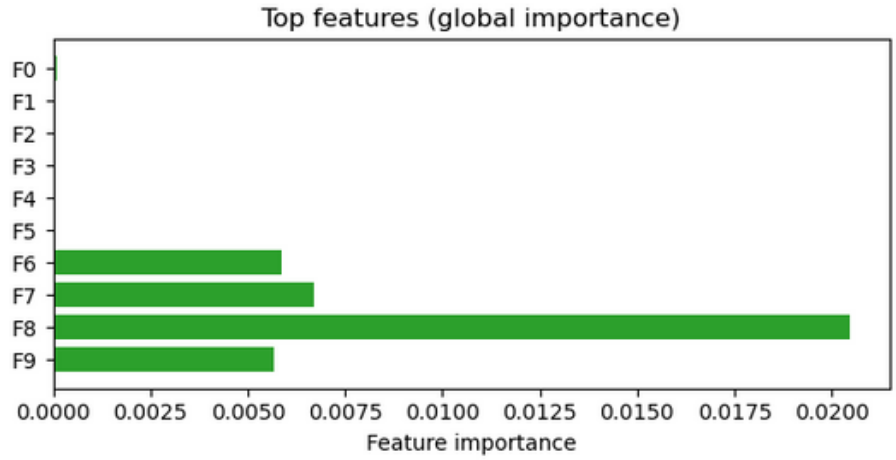
Model Performance
Optimized Threshold (0.68)
Metric Score
Accuracy 0.77
Precision 0.647
Recall 0.964
F1-score 0.775

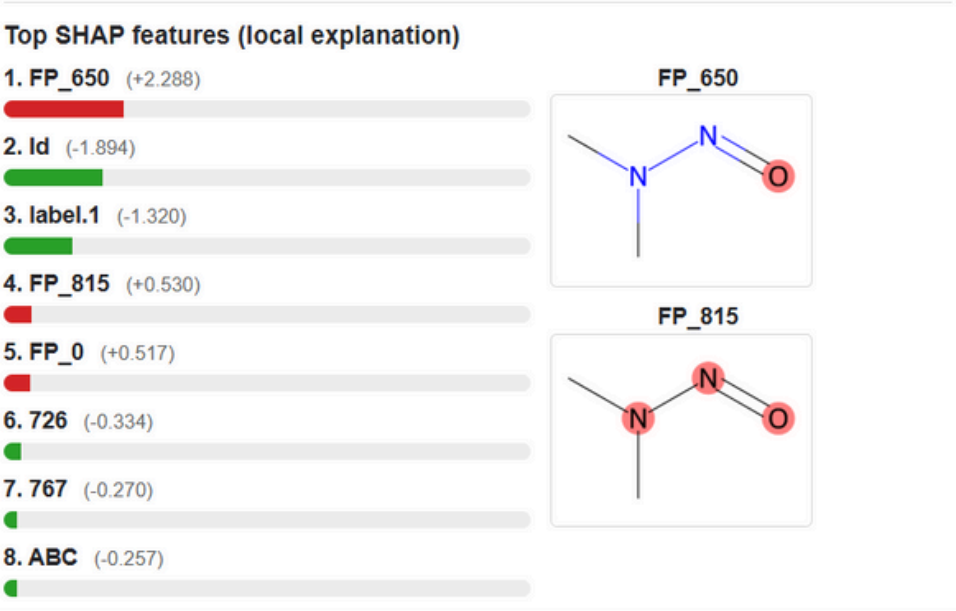
5. External Validation
External dataset: 5 molecules
Features aligned: 5 × 3,758
Model loaded successfully & predictions generated
Confirms the model generalizes to unseen chemistry

Prediction:

Name: N-Nitrosodimethylamine
SMILES: CN(C)N=O
Prediction: toxic
probability: 0.992639422416687







This molecule is predicted as Toxic (probability: 0.993).