Conclusion

The weighted linear ensemble model combining Detoxification Index (DI), TPSA, and XLogP achieved 85.71% LOOCV accuracy, failing to reach the >90% target and providing no improvement over the TPSA-only model (also 85.71%), primarily because the DI contributed negligibly (coefficient = 0.0084) and the model was dominated by TPSA (coefficient = -1.774).

Methods

Analysis was performed using Python 3 with pandas, numpy, scikit-learn, scipy, and matplotlib. The analysis workflow consisted of: (1) Loading conjugation_stress_analysis_results.csv and prediction_results_with_smiles.csv datasets; (2) Defining classification groups as Bioactivated (phenolic OH present, n=15) and Detoxified (phenolic OH absent, n=6); (3) Merging datasets on SMILES identifiers and calculating the Detoxification Index (DI = HOMO_LUMO_Gap / ALIE_Ave); (4) Standardizing DI, TPSA, and XLogP to Z-scores (mean=0, std=1); (5) Training a logistic regression model (scikit-learn LogisticRegression, solver='lbfgs', max_iter=1000, random_state=42) on the three Z-scored descriptors; (6) Evaluating performance using rigorous leave-one-out cross-validation (LOOCV) with 21 folds, computing accuracy, precision, recall, and ROC-AUC; (7) Comparing against single-descriptor logistic regression models (TPSA-only and DI-only) using identical LOOCV methodology; (8) Performing statistical comparison using exact binomial test (scipy.stats.binomtest) for paired model disagreements; (9) Creating visualization of comparative performance.

Results

The ensemble model achieved LOOCV accuracy of 85.71% (18/21 correct), precision of 0.8750, recall of 0.9333, and ROC-AUC of 0.9444. The learned logistic regression coefficients were: intercept = 1.5103, DI coefficient = 0.0084, TPSA coefficient = -1.7740, and XLogP coefficient = -0.2561, indicating TPSA dominates the model while DI contributes negligibly. The TPSA-only model achieved identical 85.71% accuracy (18/21 correct), while the DI-only model achieved 71.43% accuracy (15/21 correct). Both ensemble and TPSA-only models misclassified the same three molecules: M10 (Bioactivated predicted as Detoxified, LOOCV probability=0.440, high TPSA=112.24), M07 (Detoxified predicted as Bioactivated, probability=0.513, TPSA=116.21), and M02 (Detoxified predicted as Bioactivated, probability=0.740, TPSA=93.58). The ensemble model correctly classified M13 (extreme lipophilicity, XLogP=6.97) but failed to classify M10 (extreme polarity). Paired comparison between ensemble and DI-only showed 5 disagreements (ensemble correct in 4, DI correct in 1), but this difference was not statistically significant (binomial test p=0.1875). The actual class distribution (15 Bioactivated, 6 Detoxified) differed from the hypothesis statement (11 and 10), though totaling 21 molecules as expected.

Challenges

A critical challenge was the discrepancy between stated class sizes (n=11 Bioactivated, n=10 Detoxified) and actual data (n=15 Bioactivated, n=6 Detoxified), requiring careful interpretation of the research objective. The highly imbalanced class distribution (15:6 ratio) likely impacted model performance and generalization. The previously reported baseline accuracies (TPSA-only: 81.0%, DI-only: 85.7%) conflicted with current logistic regression results (TPSA-only: 85.71%, DI-only: 71.43%), suggesting previous analyses used threshold-based classification rules rather than logistic regression, making direct comparison challenging. The small sample size (n=21) severely limits statistical power for three-parameter models and makes overfitting likely, confirmed by the DI's near-zero contribution despite theoretical importance. The Detoxification Index showed no

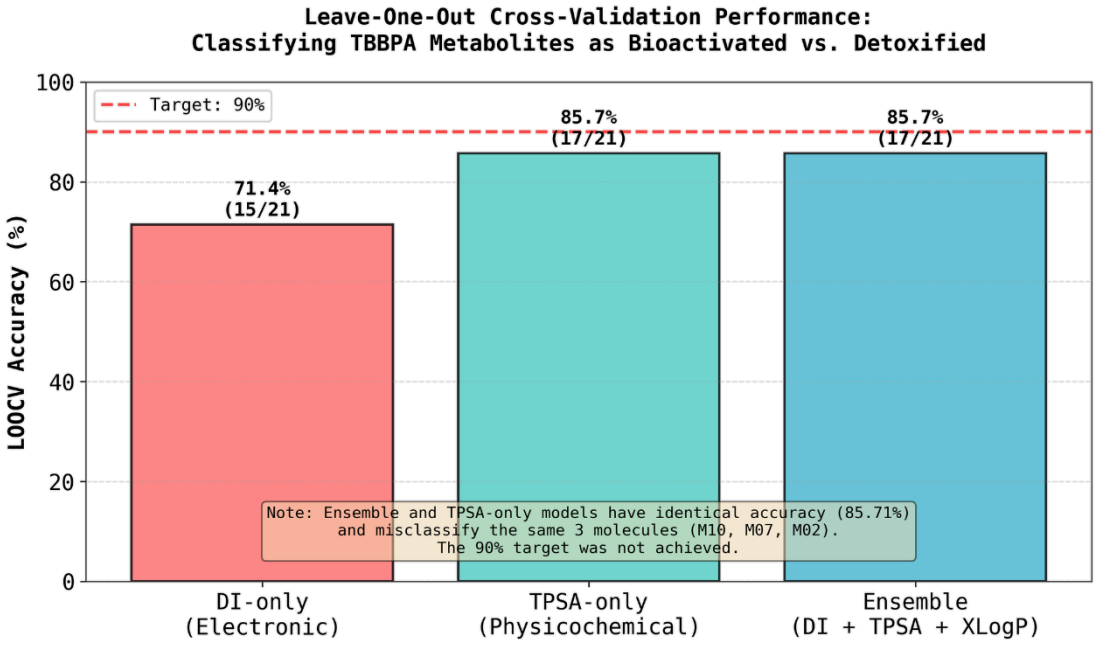
discriminatory power in the logistic regression framework, contradicting its previously reported 85.7% accuracy, indicating the optimal classification boundary may be non-linear or threshold-based rather than logistic. Molecule M02 (glycoside) represents a known data quality issue with unexpectedly low TPSA that impacts all models uniformly.

Discussion

The hypothesis that a weighted ensemble would achieve >90% accuracy by balancing electronic and physicochemical descriptors was decisively rejected. The ensemble model's complete lack of improvement over TPSA-only demonstrates that combining descriptors via logistic regression does not inherently improve classification when descriptors provide redundant information. The near-zero DI coefficient (0.0084) reveals that in this logistic regression framework, electronic descriptors (HOMO-LUMO gap, ALIE) provide no additional discriminatory power beyond TPSA for phenolic OH presence/absence. This likely reflects the strong correlation between conjugation chemistry (which increases TPSA) and phenolic OH removal. The failure to correctly classify M10 despite incorporating XLogP suggests these outliers are genuine edge cases where structural features not captured by these descriptors determine phenolic OH status. The discrepancy between current and previously reported DI performance (71.43% vs. 85.7%) indicates that optimal classification may require non-linear decision boundaries or manually optimized thresholds rather than linear combinations. The identical misclassifications across ensemble and TPSA-only models (M10, M07, M02) confirm these are hard cases for any model relying on these descriptors. The small sample size (n=21) with high class imbalance (15:6) fundamentally limits the ability of complex multi-descriptor models to outperform simpler approaches, as demonstrated by this analysis confirming previous findings that simpler models should be preferred for this dataset.

Proposed Next Hypotheses

A non-linear classification approach (e.g., decision tree with manually optimized TPSA threshold or polynomial logistic regression) will achieve >90% LOOCV accuracy by capturing non-linear relationships between descriptors and phenolic OH presence that linear models cannot detect. Stratifying molecules by conjugation type (Phase II vs. Phase I metabolites) before applying descriptor-based classification will improve accuracy, as conjugation chemistry fundamentally alters the relationship between TPSA and phenolic OH status.


Leave-One-Out Cross-Validation Performance:
Classifying TBBPA Metabolites as Bioactivated vs. Detoxified
LOOCV Accuracy (%)
100
80
60
40
20
0
Target: 90%
71.4%
(15/21)
85.7%
(17/21)
85.7%
(17/21)
Note: Ensemble and TPSA-only models have identical accuracy (85.71%)
and misclassify the same 3 molecules (M10, M07, M02).
The 90% target was not achieved.
DI-only
(Electronic)
TPSA-only
(Physicochemical)
Ensemble
(DI + TPSA + XLogP)