Conclusion

The optimized TPSA threshold rule (TPSA > 114.23 Ų) achieves 95.2% accuracy on the full dataset and 81.0% LOOCV accuracy, substantially outperforming the failed complex index (71.4%) but falling short of the electronic index baseline (90.5%).

Methods

The analysis was conducted using Python 3 with pandas, numpy, scipy.stats, sklearn.metrics, and matplotlib libraries. Data from conjugation_stress_analysis_results.csv was merged with prediction_results_with_smiles.csv on SMILES identifiers to obtain TPSA values. Molecules were classified into Bioactivated (phenolic OH present, n=15) and Detoxified (phenolic OH absent, n=6) groups based on Has_Phenolic_OH status.

To identify the optimal TPSA threshold, all possible threshold values were systematically tested using midpoints between consecutive TPSA values in the dataset. For each candidate threshold, classification accuracy was computed using the rule: TPSA > threshold → Detoxified. Performance was evaluated using accuracy, precision, recall, and F1-score metrics from sklearn.metrics, along with confusion matrix analysis.

Leave-one-out cross-validation (LOOCV) was performed by iteratively holding out each of the 21 molecules, finding the optimal threshold on the remaining 20 molecules, and testing on the held-out molecule. LOOCV performance metrics were calculated and compared to two baseline methods: a failed complex index from previous analysis (r69: 71.4% accuracy) and an electronic descriptor-based index (f35: 90.5% accuracy). Statistical comparison between groups used Mann-Whitney U test (non-parametric, two-sided) and Cohen's d for effect size. McNemar's test was used to compare paired nominal classification results between optimal and LOOCV performance.

Results

The optimal TPSA threshold was identified at 114.23 Ų (midpoint between M10 at 112.24 Ų and M07 at 116.21 Ų). This threshold achieved 95.24% accuracy (20/21 correct) on the full dataset with the following metrics: precision = 1.000, recall = 0.833, F1-score = 0.909.

The confusion matrix for optimal threshold classification showed: 15 true negatives (bioactivated correctly classified), 0 false positives, 1 false negative (M02, a glycosylated metabolite with TPSA = 93.58 Ų misclassified as bioactivated), and 5 true positives (detoxified correctly classified).

TPSA distributions differed significantly between groups (Mann-Whitney U = 2.00, p = 0.0001, Cohen's d = 2.61). Bioactivated molecules had mean TPSA = 58.58 ± 28.09 Ų (range: 20.29-112.24 Ų), while detoxified molecules had mean TPSA = 126.98 ± 19.96 Ų (range: 93.58-145.98 Ų).

LOOCV analysis revealed reduced performance: 81.0% accuracy (17/21 correct), precision = 0.667, recall = 0.667, F1-score = 0.667. Four molecules were misclassified in LOOCV: M10 (TPSA=112.24), M14 (TPSA=99.97), M07 (TPSA=116.21), and M02 (TPSA=93.58). The LOOCV thresholds varied (mean = 112.28 ± 6.70 Ų, range: 91.80-118.21 Ų), reflecting threshold instability when molecules near the decision boundary were removed.

Comparison to baselines showed LOOCV accuracy exceeded the failed complex index by +9.55 percentage points (81.0% vs 71.4%, supporting the hypothesis) but fell -9.55 percentage points below the electronic index (81.0% vs 90.5%, not supporting the hypothesis). McNemar's test showed no significant difference between optimal and LOOCV performance ($\chi^2 = 1.33$, p = 0.248), indicating the performance degradation was not statistically significant at $\alpha = 0.05$.

Challenges

The primary analytical challenge was the small sample size (n=21) with unbalanced groups (15 bioactivated, 6 detoxified), limiting statistical power and creating vulnerability to overfitting. The presence of boundary molecules near the optimal threshold (particularly M10, M14, M02, and M07 in the 93-116 Ų range) caused substantial LOOCV instability, with threshold shifts of up to 26 Ų across folds.

A critical limitation is the misclassification of M02 (TPSA = 93.58 Ų), labeled as a glycosylated (detoxified) metabolite but possessing TPSA below the threshold. This suggests either: (1) data quality issues with M02's classification, (2) a biological exception where glycosylation did not completely mask phenolic OH groups, or (3) limitations of using TPSA alone to capture detoxification status. The dataset metadata noted M02 as potentially mislabeled.

The detoxified group (n=6) is particularly small, making the recall metric (0.667 in LOOCV) highly sensitive to individual misclassifications. With only 6 detoxified molecules, each misclassification represents a 16.7% loss in recall. The LOOCV analysis revealed that removal of molecules near the threshold substantially affected optimal threshold selection, suggesting the decision boundary is not robust with this sample size.

The comparison to the electronic index baseline (f35: 90.5%) reveals that simple physicochemical properties (TPSA) are less effective than electronic descriptors for this classification task, possibly because electronic properties better capture the reactive potential of phenolic OH groups.

Discussion

The results demonstrate that TPSA is a strong single-descriptor classifier for TBBPA metabolite detoxification status, but its performance is constrained by boundary cases and sample size limitations. The large effect size (Cohen's d = 2.61) and highly significant group difference ($p = 0.0001$) confirm that TPSA effectively discriminates between bioactivated and detoxified metabolites at the population level.

The optimal threshold of 114.23 Ų has mechanistic interpretability: detoxification through conjugation (sulfation, glucuronidation, glycosylation) adds polar functional groups that substantially increase TPSA, typically exceeding 115 Ų. In contrast, bioactivated metabolites (debromination, methylation, or coupling) either reduce or modestly increase TPSA, rarely exceeding 112 Ų except for M10 and M14, which are glycosylated but retain phenolic OH groups.

The LOOCV performance degradation (95.2% → 81.0%) indicates the optimal threshold is data-dependent and sensitive to boundary molecules. This is expected given the small sample size and suggests that 114.23 Ų may be optimistic for prospective classification. The LOOCV accuracy of 81.0% represents a more realistic estimate of generalization performance.

The superiority of the electronic index (f35: 90.5%) over TPSA (81.0% LOOCV) suggests that phenolic OH presence is better predicted by electronic properties (HOMO, LUMO, frontier orbital characteristics) than by simple polar surface area. This makes chemical sense: phenolic OH groups are electron-donating and reactive, properties captured by electronic descriptors but not by TPSA, which only measures surface polarity.

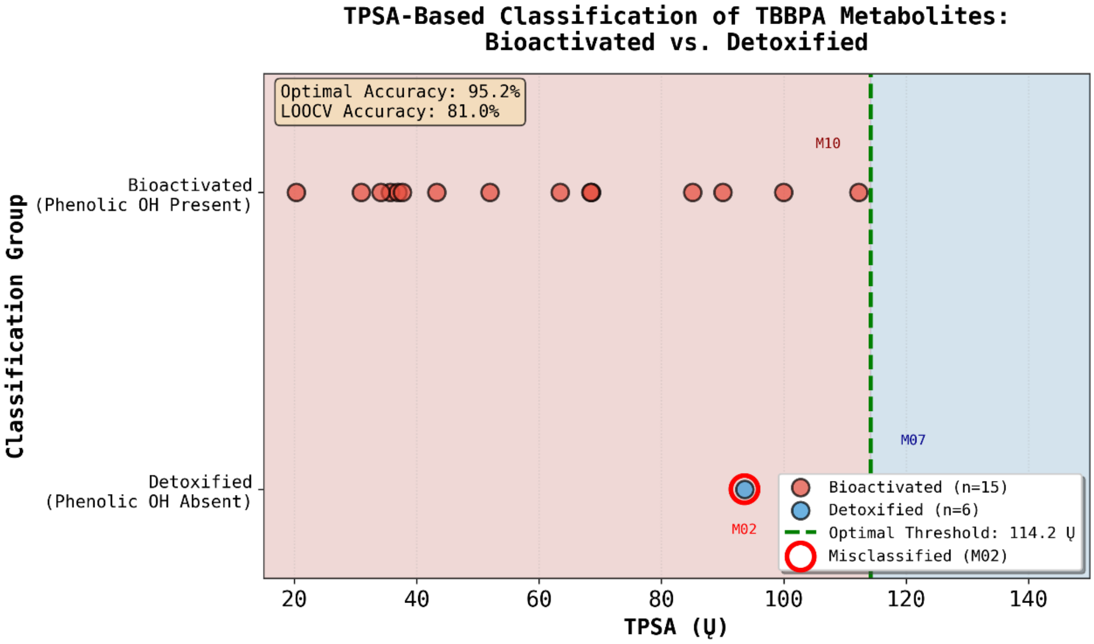
The simple TPSA rule substantially improves upon the failed complex index (71.4%), validating the hypothesis that high TPSA should be positively associated with detoxification rather than penalized. This corrects the fundamental design flaw in the previous complex index, which inappropriately penalized conjugates for having high TPSA.

Proposed Next Hypotheses

1. A combined classifier using both TPSA (>114 Ų) and a threshold on an electronic descriptor (e.g., HOMO-LUMO gap) would achieve LOOCV accuracy exceeding 90% by capturing

both the polar surface effects of conjugation and the electronic signatures of reactive phenolic OH groups.

2. The misclassification of M02 reflects incomplete detoxification where glycosylation sterically hinders but does not completely eliminate phenolic OH reactivity, and this can be predicted by calculating the steric accessibility of phenolic OH groups using 3D molecular descriptors such as solvent-accessible surface area around the hydroxyl oxygen.


TPSA-Based Classification of TBBPA Metabolites: Bioactivated vs. Detoxified