Conclusion

The composite Prediction Reliability Score successfully integrates three independent uncertainty factors (applicability domain, deconjugation risk, and classification ambiguity) to identify molecules requiring experimental validation, but exhibits a strong negative correlation with predicted toxicity ($\rho = -0.75$, $p < 0.001$) rather than independence, indicating the score captures prediction confidence rather than being orthogonal to predicted hazard.

Methods

The analysis integrated three datasets using pandas: toxicity predictions (smiles2_toxicity_results.csv), applicability domain assessments (tbbpa_applicability_domain_report.csv), and deconjugation risk scores (deconjugation_risk_scores.csv). For each of 21 TBBPA metabolites, three component scores were calculated: (1) Domain_Penalty = 1 if out-of-domain, 0 otherwise; (2) Deconjugation_Penalty = log10(Deconjugation_Risk_Score) for conjugated molecules, 0 for non-conjugated; (3) Ambiguity_Penalty = 1 - 2*|P_max - 0.5|, where P_max is the highest probability among 26 classification endpoints. The composite Prediction Reliability Score was calculated as 2*Domain_Penalty* + *1*Deconjugation_Penalty + 1*Ambiguity_Penalty, with higher scores indicating less reliable predictions. Molecules were ranked by this score. The relationship between reliability score and predicted toxicity was assessed using Spearman correlation (scipy.stats.spearmanr) against both a composite Stress_Score (mean of 5 stress response endpoint probabilities) and the maximum endpoint probability. A final visualization was created showing molecules ranked by reliability score with domain status indicated. The weighting scheme (2:1:1) gave double weight to applicability domain violations.

Results

The Prediction Reliability Score successfully integrates three uncertainty factors across all 21 molecules, ranging from 0.011 (M17, most reliable) to 4.521 (M06, least reliable), representing a 42.9-fold difference. The top 5 molecules requiring experimental validation are: M06 (score=4.521, out-of-domain + high deconjugation risk [69.0] + ambiguity [P=0.659]), M07 (score=4.085, out-of-domain + high deconjugation risk [47.8] + ambiguity [P=0.797]), M05 (score=2.404, out-of-domain + ambiguity [P=0.798]), M02 (score=2.401, high deconjugation risk [80.6] + ambiguity [P=0.753]), and M11 (score=2.398, out-of-domain + moderate ambiguity [P=0.801]). Among the top 7 high-uncertainty molecules, 6/7 are out-of-domain, 3/7 have deconjugation risk, and 4/5 of the top 5 exhibit high ambiguity (penalty > 0.4). However, contrary to the hypothesis, the Prediction Reliability Score exhibits a strong negative correlation with predicted toxicity: Spearman $\rho = -0.740$ ($p = 0.0001$) vs. Stress_Score and $\rho = -0.753$ ($p = 0.0001$) vs. Max_Probability. This correlation is primarily driven by the Ambiguity_Penalty component ($\rho = -1.00$ with Max_Probability, $p < 0.0001$), while Domain_Penalty shows weak non-significant correlation ($\rho = -0.267$, $p = 0.242$) and Deconjugation_Penalty shows moderate negative correlation ($\rho = -0.626$, $p = 0.002$). High-uncertainty molecules (top 7) have mean Stress_Score = 0.055 and mean Max_Probability = 0.810, while low-uncertainty molecules (bottom 7) have mean Stress_Score = 0.299 and mean Max_Probability = 0.958, demonstrating that molecules with extreme (confident) predictions receive low reliability scores.

Challenges

The primary analytical challenge was the unexpected strong negative correlation between the Prediction Reliability Score and predicted toxicity, which contradicts the hypothesis that the score

would be "poorly correlated" with predicted hazard. This correlation arises from the mathematical structure of the Ambiguity_Penalty component (1 - 2*|P - 0.5|), which is by definition inversely related to prediction extremity. Molecules with high confidence predictions (P near 0 or 1) necessarily receive low ambiguity penalties, creating the observed negative correlation. This represents a conceptual tension: the score was intended to capture model uncertainty independent of predicted hazard level, but classification ambiguity is inherently a measure of prediction confidence. The weighting scheme (2:1:1) was selected to balance the three components but remains discretionary—alternative weightings would shift the relative contributions. Additionally, only 10/21 molecules have deconjugation risk scores (conjugated metabolites only), limiting the contribution of this component to the overall ranking. The interpretation required distinguishing between "prediction uncertainty" (epistemic uncertainty about model reliability) and "prediction confidence" (statistical uncertainty about classification boundaries), which are conceptually different but practically intertwined in the Ambiguity_Penalty metric.

Discussion

The Prediction Reliability Score successfully fulfills its primary objective of providing a single, actionable metric to prioritize molecules for experimental validation by integrating three independent risk factors for prediction unreliability. The top 5 molecules (M06, M07, M05, M02, M11) exhibit multiple concurrent reliability concerns: 4/5 are out-of-domain (exceeding Tox21 molecular weight thresholds), 3/5 have high deconjugation risk (metabolic instability could invalidate low predicted toxicity), and 4/5 have classification ambiguity (predictions near P=0.5-0.8 decision boundaries). However, the hypothesis that this score would be poorly correlated with predicted toxicity is not supported. Instead, the score exhibits a strong negative correlation ($\rho = -0.75$), indicating that molecules with LOW predicted toxicity and LOW prediction confidence receive HIGH unreliability scores. This reflects a fundamental insight: the Ambiguity_Penalty captures prediction confidence, not epistemic model uncertainty. Molecules with extreme predictions (high/low toxicity with P near 0 or 1) represent high-confidence classifications, whereas molecules near decision boundaries represent low-confidence classifications. This is scientifically appropriate—prediction confidence IS a form of reliability assessment. The negative correlation does not invalidate the score; rather, it reveals that uncertain predictions tend to occur for molecules with lower predicted toxicity (often conjugated metabolites), while high-toxicity molecules tend to have extreme, confident predictions. The score thus integrates two types of uncertainty: (1) epistemic uncertainty from applicability domain violations and deconjugation risk, and (2) statistical uncertainty from classification ambiguity. The practical utility is clear: molecules with high scores warrant experimental validation to verify predictions that are either extrapolated beyond training data, potentially unstable, or statistically ambiguous.

Proposed Next Hypotheses

1. Molecules with high Prediction Reliability Scores but experimentally confirmed toxicity would reveal systematic model failures, particularly for out-of-domain conjugated metabolites with high deconjugation risk (M06, M07), which could be used to define safety margins for future predictions.
2. Recalculating the Prediction Reliability Score using only epistemic uncertainty components (Domain_Penalty + Deconjugation_Penalty, excluding Ambiguity_Penalty) will produce a modified metric with weak or non-significant correlation to predicted toxicity ($\rho < 0.3$), better isolating model reliability from prediction confidence.

Artifacts

Artifact 1:

File name: prediction_reliability_scores.csv

Artifact description: Complete ranking table for all 21 TBBPA metabolites sorted by Prediction Reliability Score (descending), containing Molecule_ID, composite Prediction_Reliability_Score (range: 0.011-4.521), three component scores (Domain_Penalty, Deconjugation_Penalty, Ambiguity_Penalty), Max_Probability across endpoints, applicability domain status (Is_Out_Of_Domain), Deconjugation_Risk_Score, and Risk_Category. This table provides a single, sortable metric integrating applicability domain violations, metabolic instability risk, and classification ambiguity to prioritize molecules for experimental validation.