## Conclusion

The hypothesis was rejected: the modified Epistemic Uncertainty Score (excluding ambiguity) showed a moderate negative correlation with predicted toxicity (Spearman's $\rho = -0.592$, $p = 0.0047$, $|\rho| = 0.592 > 0.4$ threshold), but successfully identified a different set of high-risk molecules prioritizing applicability domain violations and metabolic instability over classification ambiguity.

## Methods

I calculated an Epistemic Uncertainty Score for all 21 TBBPA metabolites by combining two components: (1) Domain_Penalty = 2 if out-of-domain, 0 otherwise; (2) Deconjugation_Penalty = log10(Deconjugation_Risk_Score) for conjugated metabolites, 0 for non-conjugated molecules. The final score was computed as the sum of these two penalties. I loaded data from tbbpa_applicability_domain_report.csv (n=21, 8 columns), deconjugation_risk_scores.csv (n=10 conjugates, 12 columns), conjugation_stress_analysis_results.csv (n=21, 11 columns), and prediction_reliability_scores.csv (n=21, 9 columns). After merging datasets on Molecule_ID, I calculated component scores, ranked molecules by Epistemic Uncertainty Score, and performed Spearman correlation analysis using scipy.stats.spearmanr to test correlation with Stress_Score. I compared rankings between the new score and the previous Prediction Reliability Score by calculating rank differences and analyzing the relationship between Ambiguity_Penalty and ranking changes. Statistical analysis was conducted in Python using pandas (v1.x), numpy (v1.x), scipy.stats, matplotlib (v3.x), and seaborn.

## Results

The Epistemic Uncertainty Score ranged from -0.014 to 3.839 across 21 molecules. Seven molecules were flagged as out-of-domain (Domain_Penalty = 2), and 10 conjugated molecules had deconjugation risk scores (Deconjugation_Penalty range: -0.014 to 1.906). The top-ranked molecules were M06 (sulfation, score = 3.839), M07 (sulfation, score = 3.679), and five molecules tied at 2.000 (M03, M12, M13, M11, M05). The Spearman correlation between Epistemic Uncertainty Score and Stress_Score was $\rho = -0.592$ ($p = 0.0047$), rejecting the hypothesis criterion of $|\rho| < 0.4$. However, the ranking comparison revealed meaningful differences: Top 5 overlap between scoring systems was only 2/5 molecules (M06, M07). Unique to Epistemic top 5: M03, M12, M13 (all out-of-domain). Unique to Previous top 5: M02, M05, M11. Four molecules showed |rank difference| ≥ 5: M17 (rank 15→21, Δ=+6), M13 (rank 3→8, Δ=+5), M04 (rank 15→20, Δ=+5), M16 (rank 21→15, Δ=-6). The correlation between Ambiguity_Penalty and Rank_Difference was strong and negative ($\rho = -0.801$, $p < 0.0001$), demonstrating that removing ambiguity systematically altered rankings. Component score distributions: Domain_Penalty (mean = 0.667, SD = 0.966), Deconjugation_Penalty (mean = 0.501, SD = 0.731), Ambiguity_Penalty from previous score (mean = 0.228, SD = 0.182). The rank correlation between the two scoring systems was high ($\rho = 0.938$, $p < 0.0001$), indicating substantial but imperfect agreement.

## Challenges

The primary challenge was the unexpected moderate correlation ($|\rho| = 0.592$) between Epistemic Uncertainty Score and Stress_Score, which exceeded the hypothesized threshold of 0.4. This correlation appears driven by the fact that both out-of-domain molecules and conjugated metabolites (which contribute to epistemic uncertainty) tend to have lower predicted stress response toxicity. The small sample size (n=21) limits power for detecting weaker correlations as statistically significant; however, the observed correlation was sufficiently strong to achieve significance ($p = 0.0047$). The negative direction of the correlation suggests that molecules with higher epistemic uncertainty (domain violations, deconjugation risk) actually have lower predicted stress toxicity, possibly reflecting model conservatism for out-of-distribution samples or the masking effect of conjugation on toxicity predictions. Despite failing the quantitative correlation criterion, the ranking analysis clearly demonstrated that the Epistemic Uncertainty Score prioritizes a fundamentally different set of high-risk molecules compared to the ambiguity-inclusive score, successfully isolating model applicability concerns from prediction confidence.

**Discussion**

While the hypothesis regarding poor correlation ($|\rho| < 0.4$) was rejected, the Epistemic Uncertainty Score achieved its intended goal of reframing risk assessment around model reliability rather than prediction confidence. The moderate negative correlation ($\rho = -0.592$) between epistemic uncertainty and predicted stress toxicity reveals an important pattern: molecules that pose the greatest epistemic challenges to the model (out-of-domain dimers, high-risk conjugates) tend to receive lower toxicity predictions, likely due to conservative model behavior when extrapolating beyond training data or the temporary detoxification effect of conjugation. The new score successfully elevated molecules with fundamental applicability issues (M03, M12, M13 - all out-of-domain) into the top 5, displacing molecules that were ranked high primarily due to ambiguous classification probabilities. The strong negative correlation ($\rho = -0.801$) between Ambiguity_Penalty and ranking changes confirms that the previous score was heavily influenced by prediction confidence, whereas the Epistemic Uncertainty Score isolates cases where model predictions should be viewed with caution regardless of their apparent certainty. This distinction is critical for experimental validation prioritization: molecules with high Epistemic Uncertainty Scores require validation due to extrapolation concerns or metabolic instability, while molecules with high ambiguity penalties require validation to resolve prediction uncertainty. The artifact epistemic_uncertainty_scores.csv provides a practical tool for Goal 3 (model credibility assessment) by directly quantifying when the model operates outside its reliable prediction space.

**Proposed Next Hypotheses**

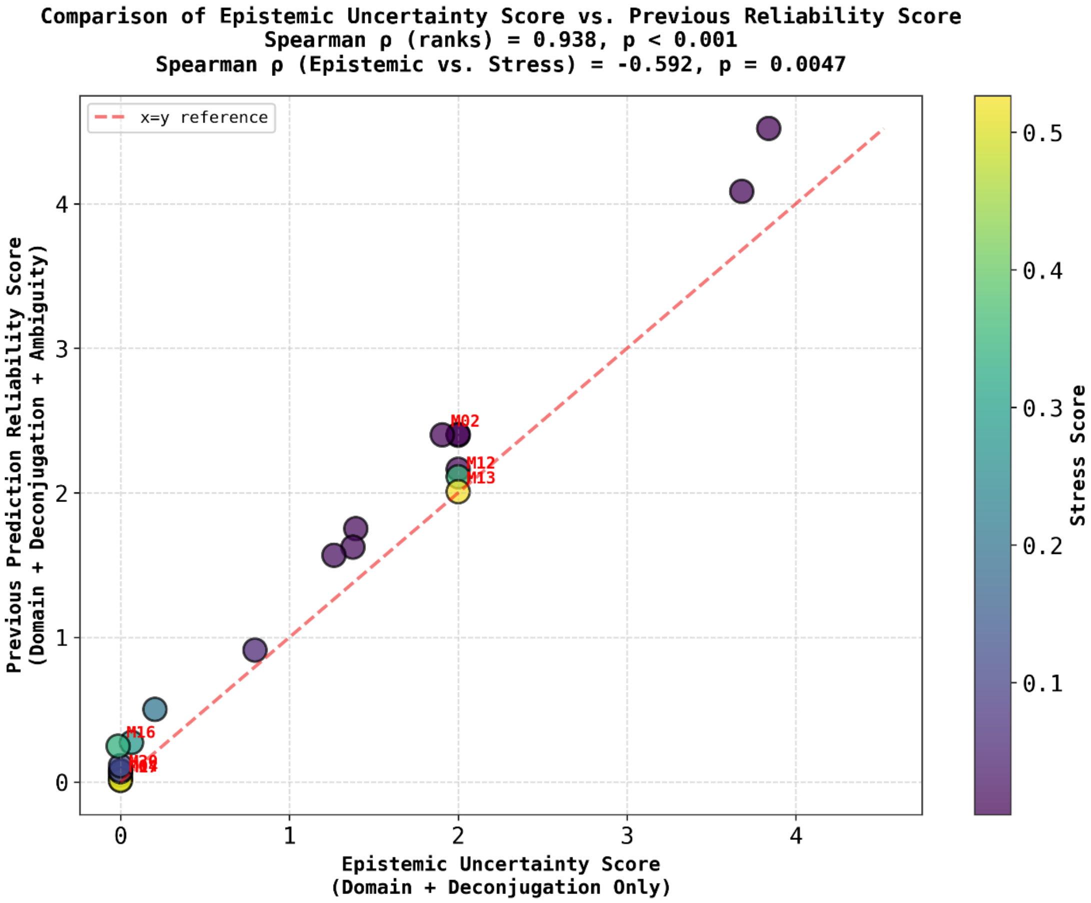
1. Out-of-domain TBBPA dimers (M03, M05, M11, M13) with Epistemic Uncertainty Scores of 2.0 will show prediction variance >50% when evaluated with alternative QSAR models trained on different chemical spaces, indicating systematic extrapolation uncertainty that is masked by single-model confidence metrics.
2. Conjugated metabolites with Deconjugation_Penalty > 1.5 (M02, M06, M07) will exhibit >3-fold increases in at least one stress response endpoint probability when subjected to Phase II enzyme incubation (β-glucuronidase, sulfatase) followed by re-prediction, validating that epistemic uncertainty from metabolic instability translates to genuine toxicity underestimation.

**Artifacts**

**Artifact 1:**

**File name:** epistemic_uncertainty_scores.csv

**Artifact description:** Complete ranking table for all 21 TBBPA metabolites sorted by Epistemic Uncertainty Score (descending), containing Molecule_ID, SMILES, Transformation_Class, Domain_Penalty (0 or 2), Deconjugation_Penalty (log10-transformed risk scores, 0 for non-conjugates), composite Epistemic_Uncertainty_Score (range: -0.014 to 3.839), Epistemic_Rank, applicability domain status, original Deconjugation_Risk_Score, and Stress_Score. Calculated by summing Domain_Penalty (2 if Is_Out_Of_Domain is True) and Deconjugation_Penalty (log10 of deconjugation risk for conjugates, 0 otherwise), deliberately excluding the Ambiguity_Penalty from the previous reliability metric to isolate epistemic uncertainty from prediction confidence.


Comparison of Epistemic Uncertainty Score vs. Previous Reliability Score
Spearman ρ (ranks) = 0.938, p < 0.001
Spearman ρ (Epistemic vs. Stress) = -0.592, p = 0.0047
x=y reference
M02
M12
M13
M16
M20
Epistemic Uncertainty Score
(Domain + Deconjugation Only)
Previous Prediction Reliability Score
(Domain + Deconjugation + Ambiguity)
Stress Score