INTEGRATED ANALYSIS OF TBBPA METABOLITE TOXICITY PREDICTIONS

ANALYSIS OVERVIEW

This analysis integrated molecular descriptors and predicted toxicity endpoints for 21 TBBPA (tetrabromobisphenol A) transformation and metabolic products. The objectives were to:

1. Merge datasets and create comprehensive endpoint signatures
2. Identify correlations between molecular descriptors and toxicity probabilities
3. Generate priority lists for high-risk molecules

KEY FINDINGS

1. Dataset Integration and Risk Distribution

Successfully merged 21 molecules with:

- **26 classification toxicity endpoints** (89 total features from toxicity data)
- **51 molecular descriptors** (DFT-based and physicochemical properties)
- **Total: 140 features** per molecule

Risk distribution across all molecules (546 total predictions):

- **HIGH risk: 39 predictions (7.1%)** - Average 1.86 HIGH risks per molecule
- **MEDIUM risk: 31 predictions (5.7%)** - Average 1.48 MEDIUM risks per molecule
- **LOW risk: 476 predictions (87.2%)** - Average 22.67 LOW risks per molecule

Risk concentration pattern: 5 molecules (24% of the dataset) account for ALL 4 HIGH risk predictions each, supporting the hypothesis that a small subset accounts for the majority of high-risk predictions.

2. Molecular Descriptor-Toxicity Correlations

Strong monotonic correlations identified for key endpoints (Spearman ρ):

NR_ER (Nuclear Receptor - Estrogen Receptor):

- Strongest positive: XLogP ($\rho = 0.600$, $p = 0.0040$)
- Strongest negative: ODI_LUMO ($\rho = -0.682$, $p = 0.0007$)

NR_AR (Nuclear Receptor - Androgen Receptor):

- Strongest positive: Mol_Size_Short ($\rho = 0.866$, $p < 0.0001$)
- Strongest negative: ODI_HOMO ($\rho = -0.838$, $p < 0.0001$)

NR_PPAR_gamma (Nuclear Receptor - Peroxisome Proliferator-Activated Receptor):

- Strongest positive: LEA_Var ($\rho = 0.848$, $p < 0.0001$)
- Strongest negative: LEA_Ave ($\rho = -0.895$, $p < 0.0001$)

SR_MMP (Stress Response - Mitochondrial Membrane Potential):

- Strongest positive: LEA_Var ($\rho = 0.791$, $p < 0.0001$)
- Strongest negative: LEA_Ave ($\rho = -0.857$, $p < 0.0001$)

SR_ARE (Stress Response - Antioxidant Response Element):

- Strongest positive: LEA_Var ($\rho = 0.832$, $p < 0.0001$)
- Strongest negative: LEA_Ave ($\rho = -0.862$, $p < 0.0001$)

Key Pattern: Stress response endpoints (SR_MMP, SR_ARE) and NR_PPAR_gamma show remarkably consistent correlation patterns with surface electrostatic properties (LEA - Local Electron Affinity), indicating shared mechanistic drivers.

3. Priority Molecule Lists

LIST (i): OVERALL HAZARD PRIORITY (Ranked by HIGH risk count)

Rank 1: Molecule 4 - 4 HIGH risks

- SR_MMP: 0.9922, Eye_irritation: 0.9694, Respiratory_toxicity: 0.9366, CYP1A2: 0.8680

- Dibrominated trihydroxy TBBPA derivative

Rank 2: Molecule 12 - 4 HIGH risks

- SR_MMP: 0.9443, Respiratory_toxicity: 0.9094, Eye_irritation: 0.9047, CYP1A2: 0.8873
- Methoxy-hydroxy debrominated TBBPA coupling product

Rank 3: Molecule 13 - 4 HIGH risks

- SR_MMP: 0.9960, Eye_irritation: 0.9210, Respiratory_toxicity: 0.9021, SR_ARE: 0.8102
- TBBPA coupling product with pentabrominated phenolic ether

Rank 4: Molecule 17 - 4 HIGH risks

- Eye_irritation: 0.9947, Eye_corrosion: 0.9812, Respiratory_toxicity: 0.8839, CYP1A2: 0.8430
- Debrominated conjugated alkene derivative

Rank 5: Molecule 20 - 4 HIGH risks

- Eye_irritation: 0.9767, Respiratory_toxicity: 0.9499, Eye_corrosion: 0.9036, SR_MMP: 0.8313
- Tertiary alcohol TBBPA derivative

LIST (ii): ENDOCRINE PRIORITY (Ranked by max NR_* probability)

Rank 1: Molecule 13 - Max NR probability: 0.6271 (NR_ER)

- Top NR endpoints: NR_ER: 0.6271, NR_ER_LBD: 0.5587, NR_Aromatase: 0.2394

Rank 2: Molecule 4 - Max NR probability: 0.3859 (NR_AhR)

- Top NR endpoints: NR_AhR: 0.3859, NR_Aromatase: 0.2261, NR_ER: 0.1980

Rank 3: Molecule 12 - Max NR probability: 0.3138 (NR_AhR)

- Top NR endpoints: NR_AhR: 0.3138, NR_Aromatase: 0.1117, NR_ER: 0.0939

Rank 4: Molecule 21 - Max NR probability: 0.2734 (NR_AhR)

- Top NR endpoints: NR_AhR: 0.2734, NR_Aromatase: 0.0138, NR_PPAR_gamma: 0.0130

Rank 5: Molecule 17 - Max NR probability: 0.2284 (NR_PPAR_gamma)

- Top NR endpoints: NR_PPAR_gamma: 0.2284, NR_Aromatase: 0.0641, NR_AhR: 0.0484

Observation: No molecules classified as HIGH risk for key NR endpoints (NR_ER, NR_AR, NR_PPAR_gamma), indicating endocrine disruption probabilities are moderate but not reaching the HIGH threshold for this metabolite set.

LIST (iii): STRESS RESPONSE PRIORITY (Ranked by max SR_* probability)

Rank 1: Molecule 13 - Max SR probability: 0.9960 (SR_MMP)

- Top SR endpoints: SR_MMP: 0.9960, SR_ARE: 0.8102, SR_p53: 0.4398

Rank 2: Molecule 4 - Max SR probability: 0.9922 (SR_MMP)

- Top SR endpoints: SR_MMP: 0.9922, SR_ARE: 0.7658, SR_p53: 0.4195

Rank 3: Molecule 12 - Max SR probability: 0.9443 (SR_MMP)

- Top SR endpoints: SR_MMP: 0.9443, SR_ARE: 0.4170, SR_HSE: 0.2094

Rank 4: Molecule 9 - Max SR probability: 0.8963 (SR_MMP)

- Top SR endpoints: SR_MMP: 0.8963, SR_ARE: 0.3358, SR_HSE: 0.0997

Rank 5: Molecule 16 - Max SR probability: 0.8651 (SR_MMP)

- Top SR endpoints: SR_MMP: 0.8651, SR_ARE: 0.4613, SR_HSE: 0.1882

Critical Pattern: SR_MMP (mitochondrial membrane potential disruption) dominates stress response concerns. 6 molecules classified as HIGH risk for SR_MMP, making it the most prevalent HIGH-risk endpoint across the dataset.

4. Mechanistic Insights from Correlations

Surface electrostatic properties (LEA_Ave, LEA_Var) emerge as the most powerful predictors of toxicity:

- **LEA_Var** (local electron affinity variance) shows strong positive correlations with NR_PPAR_gamma ($\rho = 0.848$), SR_MMP ($\rho = 0.791$), and SR_ARE ($\rho = 0.832$)
- **LEA_Ave** (local electron affinity average) shows strong negative correlations with NR_PPAR_gamma ($\rho = -0.895$), SR_MMP ($\rho = -0.857$), and SR_ARE ($\rho = -0.862$)

This suggests that molecules with **heterogeneous electron-withdrawing surfaces** (high variance, low average LEA) are associated with increased stress response toxicity.

Molecular size and shape correlate with NR_AR activity:

- Mol_Size_Short ($\rho = 0.866$) and Weight ($\rho = 0.842$) show very strong positive correlations with NR_AR probability

Lipophilicity and density correlate with NR_ER activity:

- XLogP ($\rho = 0.600$) and Density ($\rho = 0.578$) positively associated with NR_ER

Frontier molecular orbital descriptors show negative correlations:

- ODI_LUMO negatively correlates with NR_ER ($\rho = -0.682$)
- ODI_HOMO negatively correlates with NR_AR ($\rho = -0.838$)

STATISTICAL SIGNIFICANCE

All reported top correlations for the five key endpoints achieved statistical significance ($p < 0.05$), with the strongest correlations ($|\rho| > 0.75$) achieving $p < 0.0001$. Given the small sample size (n=21), caution is warranted in generalizing these findings, but the effect sizes are substantial and biologically interpretable.

SUMMARY STATISTICS

- **High-risk molecule concentration**: 5 molecules (24%) account for 20/39 (51%) of all HIGH risk predictions
- **Dominant toxicity concern**: SR_MMP (6 HIGH risk molecules, 29% of dataset)
- **Secondary concerns**: Eye irritation (5 HIGH risk), Respiratory toxicity (7 HIGH risk)
- **Endocrine disruption**: Moderate probabilities, no HIGH classifications for key NR endpoints
- **Strongest descriptor-toxicity relationship**: LEA_Ave vs NR_PPAR_gamma ($\rho = -0.895$)

OUTPUTS GENERATED

1. **correlation_summary_key_endpoints.csv** - Top 5 positive and negative correlations for NR_ER, NR_AR, NR_PPAR_gamma, SR_MMP, SR_ARE (50 rows)
2. **endpoint_signatures_all_molecules.csv** - Risk level counts and HIGH risk endpoints for all 21 molecules
3. **priority_list_overall_hazard.csv** - Top 10 molecules ranked by HIGH risk count
4. **priority_list_endocrine.csv** - Top 10 molecules ranked by max NR endpoint probability
5. **priority_list_stress_response.csv** - Top 10 molecules ranked by max SR endpoint probability
6. **toxicity_analysis_summary.png** - Final summary figure with correlation patterns and priority molecule profiles

---

DISCRETIONARY ANALYTICAL DECISIONS

- **Correlation method**: Used Spearman rank correlation rather than Pearson correlation due to non-normal distribution of descriptors and small sample size (n=21), which makes non-

parametric methods more robust

- **Significance threshold**: Applied α = 0.05 for statistical significance reporting
- **Priority list length**: Reported Top 5 molecules for each priority dimension as specified in the objective, but saved Top 10 to files for completeness
- **Risk level classification**: Used the pre-defined HIGH/MEDIUM/LOW classifications from the original toxicity predictions rather than applying custom thresholds
- **Correlation summary depth**: Reported top 3 positive and 3 negative correlations in the visualization for clarity, but saved top 5 of each to CSV files as specified
- **Normalization in visualization**: Normalized HIGH count by dividing by 4 (observed maximum) to enable comparison with probability scores on 0-1 scale in Figure B
- **SMILES truncation**: Truncated SMILES strings to 50-80 characters in display outputs for readability while preserving full strings in saved files
- **Endpoint focus**: Prioritized the 5 key endpoints (NR_ER, NR_AR, NR_PPAR_gamma, SR_MMP, SR_ARE) for detailed correlation analysis as specified in the objective