Conclusion

I have created a comprehensive dataset description documenting two complementary CSV files containing 52 molecular descriptors and 89 toxicity predictions for 21 brominated compounds.

Methods

I conducted systematic exploration of both datasets using pandas for data manipulation, numpy for statistical analysis, and matplotlib/seaborn for visualization. The analysis included: (1) Loading and examining dataset dimensions and structure, (2) Statistical summarization with descriptive statistics and distribution analysis, (3) Data quality assessment including missing value detection and range validation, (4) Cross-dataset relationship analysis to confirm perfect matching between SMILES identifiers, (5) Detailed categorization of molecular descriptors and toxicity endpoints, (6) Skewness analysis to identify distribution characteristics, (7) Creation of summary visualizations showing molecular descriptor distributions and toxicity endpoint positivity rates, and (8) Generation of comprehensive documentation covering all aspects of the datasets.

Results

Dataset 1 (prediction_results_with_smiles.csv): 21 rows × 52 columns containing molecular descriptors including basic properties (AtomNum, Weight), electronic properties (HOMO: -0.319 to -0.306 eV, LUMO: 0.001 to 0.029 eV), physicochemical properties (XLogP: 3.42-6.97, TPSA: 20.29-145.98 Ų), and surface properties. No missing values detected. Dataset 2 (smiles2_toxicity_results.csv): 21 rows × 89 columns containing acute toxicity metrics (LD50: 106.49-4551.59 mg/kg), 26 toxicity probability scores (all bounded 0-1), and corresponding binary results and risk classifications. Key findings: 100% respiratory toxicity positivity, 52.4% eye irritation positivity, highly skewed acute toxicity distributions (skewness 2.77-3.02). Perfect dataset matching: all 21 compounds have identical SMILES identifiers between datasets, enabling complete merging to 141-column combined dataset.

Challenges

Small sample size (n=21) limits statistical power for predictive modeling and significance testing. High dimensionality (141 features vs 21 observations) creates risk of overfitting requiring careful feature selection. Many molecular descriptors show substantial positive skewness (up to 3.02) requiring transformation for parametric analyses. Extreme class imbalance in toxicity endpoints (many 0% or 100% positive rates) limits utility for classification modeling. Some endpoints like respiratory toxicity show no variation (all positive), preventing discriminative analysis.

Discussion

This dataset represents a valuable resource for structure-activity relationship (SAR) modeling of brominated compounds, combining quantum chemical descriptors with comprehensive toxicity predictions. The perfect dataset alignment enables integrated analysis of molecular properties and toxic effects. The high-quality data with no missing values facilitates reliable statistical analysis, though the small sample size necessitates cross-validation approaches. The diversity of toxicity endpoints (carcinogenicity, mutagenicity, cardiotoxicity, enzyme inhibition, receptor binding) enables multi-endpoint risk assessment. The substantial feature skewness is typical of molecular descriptor datasets and manageable with appropriate transformations.

Proposed Next Hypotheses

Molecular lipophilicity (XLogP) is positively correlated with bioconcentration factor (BCF) and negatively correlated with acute aquatic toxicity endpoints. Electronic properties (HOMO-LUMO gap) are predictive of respiratory toxicity probability scores and stress response pathway activation.

Artifacts

Artifact 1:

File name: DATASET_DESCRIPTION.txt

Artifact description: Comprehensive dataset documentation (15,880 characters) containing detailed descriptions of all 141 variables across both datasets, statistical summaries, data quality assessments, distribution characteristics, merging instructions, suggested analysis applications, environment configuration notes, and analytical challenges. Created through systematic exploration and statistical analysis of both CSV files.