Analysis of structural and biological features on 43,894 unique drug–drug pairs
LLMs for Drug-Drug Interaction Prediction

October 16, 2025

1 Statistical methodology

To validate the negative sampling strategy, we analyzed structural and biological features on 43,894 unique drug–drug pairs (21,947 positive interactions and 21,947 negatives) compiled from 13 external datasets, covering 6,088 unique drugs. All (unordered) pairs are unique by design, ensuring independence at the pair level. Because features were highly right-skewed and drugs recur across pairs (mean 14.4 appearances per drug), we compared classes using permutation tests on the difference in medians, permuting pair labels 10,000 times. Despite drug-level correlations, validity is maintained under pair-level independence by permuting independent units (pairs) [2, 1]. We calculated effect sizes using Cohen's d with pooled standard deviations, and applied Bonferroni correction across the six features ($\alpha_{\text{adj}} = 0.0083$). The results are reported in Table 1. Median differences were significant for five out of six features; however, all effect sizes were small-to-medium (d=0.129–0.354) with large standard deviations (>means). All this means that negative samples are not trivially dissimilar from positive samples. The substantial overlap indicates that models cannot easily succeed by memorizing thresholds ("Drug 2 SMILES $>$ 60 $\rightarrow$ positive") or univariate rules, but must instead integrate multiple features and learn the complex, non-linear interaction between structural properties, target profiles, and interaction potential.

2 Results

2.1 Datasets Summary

Total samples loaded: 43,894
Valid samples processed: 43,894
Positive samples (interactions): 21,947 (50.0%)
Negative samples (no interactions): 21,947 (50.0%)

2.2 Drug Redundancy Analysis

Total drug instances (across all pairs): 87,788
Unique drugs: 6,090
Redundancy ratio: 14.42x
Average appearances per drug: 14.42

2.3 Permutation Test Analysis

Number of permutations: 10,000
Number of tests: 6
Original alpha: 0.05
Bonferroni-corrected alpha: 0.008333

Analyzing: SMILES Length (Drug 1)...

- Positive: mean=72.33, median=55.00, CV=102.5%
- Negative: mean=57.57, median=50.00, CV=93.9%
- Permutation p-value: 0.000000 ***
- Cohen's d: 0.228

Analyzing: SMILES Length (Drug 2)...

- Positive: mean=65.05, median=50.00, CV=110.9%
- Negative: mean=56.99, median=50.00, CV=88.5%
- Permutation p-value: 1.000000 ns
- Cohen's d: 0.129

Analyzing: Total SMILES Length...

- Positive: mean=137.37, median=112.00, CV=76.6%
- Negative: mean=114.56, median=102.00, CV=64.1%
- Permutation p-value: 0.000000 ***
- Cohen's d: 0.252

Analyzing: Target Genes (Drug 1)...

- Positive: mean=4.12, median=2.00, CV=174.0%
- Negative: mean=2.75, median=1.00, CV=268.5%
- Permutation p-value: 0.000000 ***
- Cohen's d: 0.188

Analyzing: Target Genes (Drug 2)...

- Positive: mean=5.25, median=2.00, CV=155.4%
- Negative: mean=2.80, median=1.00, CV=275.6%
- Permutation p-value: 0.000000 ***
- Cohen's d: 0.308

Analyzing: Total Target Genes...

- Positive: mean=9.37, median=6.00, CV=115.9%
- Negative: mean=5.56, median=3.00, CV=192.1%
- Permutation p-value: 0.000000 ***
- Cohen's d: 0.354

2.4 Results Summary

Table 1: Between-class contrasts (n. positives = n. negatives = 21,947): medians, p-values from permutation tests, significance (Bonferroni correction), and effect size.

Feature	Median (Pos)	Median (Neg)	Significance	Cohen's d
SMILES Length (Drug 1)	55	50	< 0.001 (***)	0.228
SMILES Length (Drug 2)	50	50	1.000 (ns)	0.129
Total SMILES Length	112	102	< 0.001 (***)	0.252
Target Genes (Drug 1)	2	1	< 0.001 (***)	0.188
Target Genes (Drug 2)	2	1	< 0.001 (***)	0.308
Total Target Genes	6	3	< 0.001 (***)	0.354

Notes: p-value from 10,000 permutations at the pairwise level on the difference of medians. Bonferroni correction on 6 tests: $\alpha_{\text{adj}} = 0.0083$; codes: *** $p < 0.0083$, ns = not significant. Cohen's d with pooled standard deviation; positive values indicate higher values in positives.