Few-shot Ablation Studies
LLMs for Drug-Drug Interaction Prediction

October 16, 2025

1 Empirical Design

We conducted an ablation analysis on the two best-performing models from RQ2 (GPT-4o and Claude 3.5 Sonnet) and Phi3.5 (2.7B) using the similarity-based few-shot strategy and considering SMILES only, genes only, species only, or pairwise combinations. The results are presented in Tables 1, 2, and 3.

Table 1: Ablation study results for Claude 3.5 Sonnet with similarity-based few-shot learning. Delta (Δ) represents the difference from the complete configuration (all sources).

Configuration	**Acc. (Δ)**	**Sens. (Δ)**	**F1 (Δ)**
SMILES only	0.8440 (+0.0064)	0.8606 (+0.0183)	0.8466 (+0.0082)
Genes only	0.8266 (−0.0110)	0.7963 (−0.0459)	0.8212 (−0.0172)
Organisms only	0.8541 (+0.0165)	0.8661 (+0.0239)	0.8558 (+0.0175)
SMILES + Genes	0.8156 (−0.0220)	0.7615 (−0.0807)	0.8050 (−0.0333)
SMILES + Organisms	0.8440 (+0.0064)	0.8147 (−0.0275)	0.8393 (+0.0010)
Genes + Organisms	0.8257 (−0.0119)	0.7688 (−0.0734)	0.8152 (−0.0232)

Table 2: Ablation study results for GPT-4o with similarity-based few-shot learning. Delta (Δ) represents the difference from the complete configuration (all sources).

Configuration	**Acc. (Δ)**	**Sens. (Δ)**	**F1 (Δ)**
SMILES only	0.7119 (−0.0798)	0.8862 (+0.0459)	0.7547 (−0.0467)
Genes only	0.7789 (−0.0128)	0.7303 (−0.1101)	0.7676 (−0.0338)
Organisms only	0.7853 (−0.0064)	0.8954 (+0.0550)	0.8066 (+0.0052)
SMILES + Genes	0.7670 (−0.0248)	0.7615 (−0.0789)	0.7657 (−0.0357)
SMILES + Organisms	0.7211 (−0.0706)	0.9028 (+0.0624)	0.7640 (−0.0374)
Genes + Organisms	0.7807 (−0.0110)	0.7725 (−0.0679)	0.7789 (−0.0225)

Table 3: Ablation study results for Phi-3.5 2.7B with similarity-based few-shot learning. Delta (Δ) represents the difference from the complete configuration (all sources).

Configuration	**Acc. (Δ)**	**Sens. (Δ)**	**F1 (Δ)**
SMILES only	0.5514 (−0.0275)	0.2330 (−0.6496)	0.3419 (−0.3351)
Genes only	0.5055 (−0.0734)	0.5266 (−0.3560)	0.5157 (−0.1613)
Organisms only	0.4899 (−0.0890)	0.9376 (+0.0550)	0.6477 (−0.0293)
SMILES + Genes	0.4954 (−0.0835)	0.9541 (+0.0715)	0.6541 (−0.0229)
SMILES + Organisms	0.5000 (−0.0789)	0.9596 (+0.0770)	0.6574 (−0.0196)
Genes + Organisms	0.4936 (−0.0853)	0.9743 (+0.0917)	0.6580 (−0.0190)

The ablation results reveal three patterns that challenge the direct applicability of traditional feature importance interpretation to LLM-based approaches. First, we observe non-monotonic performance relationships. For Claude 3.5 Sonnet, the organisms-only configuration achieves higher accuracy (0.8541) and F1-score (0.8558) than the complete multi-source configuration (0.8376 and 0.8384, respectively), with improvements of +1.65% and +1.75%. The SMILES-only configuration also slightly outperforms the baseline (+0.64% accuracy). Multi-source combinations like SMILES+Genes degrade performance by 2.20% in accuracy compared to the baseline. For Phi-3.5 2.7B, all single-source and pairwise configurations substantially underperform the complete baseline, with SMILES-only showing the most severe degradation (Δ accuracy: −2.75%, Δ F1: −33.51%), while organisms-only maintains high sensitivity (0.9376) at the cost of accuracy (Δ: −8.90%). This pattern cannot be explained by simple feature interactions or redundancy. In traditional machine learning with properly regularized models, adding features should either improve performance (if informative) or have a neutral impact (if redundant or if the model learns to ignore them). The degradation we observe when combining sources suggests that LLM performance depends on prompt structure and attention distribution rather than additive feature contributions.

Second, the three models exhibit divergent behaviors despite processing identical information. Claude 3.5 Sonnet shows the hierarchy: organisms (+1.65%) > SMILES (+0.64%) > baseline > genes (−1.10%), with multi-source combinations consistently underperforming. GPT-4o displays a reversed pattern where the complete configuration achieves optimal performance (F1: 0.8014), organisms-only nearly matches baseline ($\Delta = -0.64\%$ accuracy, +0.52% F1), while SMILES-only shows severe degradation ($\Delta = -7.98\%$ accuracy). Phi-3.5 2.7B presents a third distinct behavior where the complete configuration is essential for balanced performance, with all ablated versions showing substantial accuracy losses ranging from −2.75% (SMILES-only) to −8.90% (organisms-only), though organisms-only and pairwise combinations maintain high sensitivity (0.9376-0.9743) at the cost of precision. Critically, these divergent behaviors emerge from the same input data, the same task, and the same examples. In conventional feature importance analysis, the relative contribution of features should be determined by the data-generating process and task requirements, not by model architecture. A gene-drug interaction is either predictive of DDIs or not, regardless of the model used. The fact that Claude Sonnet 3.5 achieves better performance without genes, while GPT-4o requires them, and Phi-3.5 2.7B shows yet another pattern, suggests that ablations measure architecture-specific prompt processing artifacts rather than intrinsic information value for the DDI prediction task.

Third, certain multi-source combinations paradoxically underperform their constituent single sources. For Claude 3.5 Sonnet, SMILES+Genes (F1: 0.8050) performs 4.16% worse than SMILES alone (F1: 0.8466), and Genes+Organisms (F1: 0.8152) underperforms organisms alone (F1: 0.8558) by 4.06%. For GPT-4o, SMILES+Organisms (F1: 0.7640) degrades 4.26% compared to organisms alone (F1: 0.8066). For Phi-3.5 2.7B, this paradox does not manifest in the same way as all pairwise combinations show similar underperformance relative to the complete configuration, yet organisms-only achieves higher sensitivity (0.9376) than SMILES+Genes (0.9541) while maintaining better balance (F1: 0.6477 vs. organisms-only compared to the near-zero precision implied by the F1 decline in SMILES+Genes). This trend means that each ablated configuration induces the model to adopt a different reasoning strategy rather than simply removing one input to a fixed reasoning process. A SMILES-only prompt may activate pattern matching based on structural similarity learned during pre-training. In contrast, a genes-only prompt may trigger pathway-based reasoning, and the complete prompt may create competition between these strategies, leading to degraded performance for some architectures while proving essential for others like Phi-3.5 2.7B.