Fine-tuned LLMs Abaltion Studies
LLMs for Drug-Drug Interaction Prediction

October 16, 2025

1 Empirical Design

To further investigate whether few-shot patterns analyzed in "S.R6.2.2 Few-shot ablation studies" persist under task-specific optimization, we conducted targeted fine-tuning experiments on two architectures using their best-performing single-source configurations from the few-shot ablations. Specifically, we fine-tuned Phi-3.5 2.7B using SMILES-only inputs and GPT-4o using organisms-only inputs, employing the same methodology and training procedures described in Section 4.4 in the main manuscript. These configurations were selected because organisms-only achieved near-baseline performance for GPT-4o in few-shot evaluation (Δ accuracy: -0.64%, Δ F1: $+0.52\%$), while SMILES represents the structural foundation of DDI prediction and showed the smallest degradation for Phi-3.5 2.7B among single-source configurations (Δ accuracy: -2.75% compared to -7.34% for genes and -8.90% for organisms).

Table 1: Fine-tuning results for Phi-3.5 2.7B (SMILES-only) and GPT-4o (organisms-only) compared to complete multi-source baselines across validation and external datasets.

	Phi-3.5 2.7B (SMILES-only)			**GPT-4o (Organisms-only)**		
Dataset	**Acc.**	**Sens.**	**F1**	**Acc.**	**Sens.**	**F1**
Validation	0.690	1.000	0.763	0.888	0.987	0.898
CredibleMeds	0.800	1.000	0.833	0.800	1.000	0.833
HEP	0.700	1.000	0.769	0.878	0.998	0.891
HIV	0.696	1.000	0.767	0.881	0.998	0.893
Corpus 2011	0.609	1.000	0.719	0.703	0.750	0.716
Corpus 2013	0.682	1.000	0.759	0.858	0.973	0.873
NLM Corpus	0.611	1.000	0.720	0.833	0.889	0.842
PK Corpus	0.500	1.000	0.667	0.500	1.000	0.667
OSCAR	0.705	1.000	0.772	0.881	0.984	0.892
WorldVista	0.710	1.000	0.775	0.825	0.870	0.833
French Ref.	0.700	1.000	0.769	0.885	0.999	0.897
KEGG	0.696	1.000	0.767	0.880	0.989	0.892
NDF-RT	0.693	1.000	0.765	0.908	1.000	0.915
ONC Non-Int.	0.682	1.000	0.759	0.878	1.000	0.891
AVG (external)	**0.676**	**1.000**	**0.757**	**0.824**	**0.958**	**0.849**

Table 2: Performance deltas between single-source fine-tuned models and complete multi-source baselines.

	Δ Phi-3.5 2.7B			**Δ GPT-4o**		
Dataset	**Acc.**	**Sens.**	**F1**	**Acc.**	**Sens.**	**F1**
Validation	-0.223	$+0.040$	-0.153	-0.038	$+0.057$	-0.028
CredibleMeds	-0.200	0.000	-0.167	-0.200	0.000	-0.167
HEP	-0.224	0.000	-0.160	-0.029	$+0.081$	-0.017
HIV	-0.230	0.000	-0.165	-0.067	$+0.016$	-0.056
Corpus 2011	-0.266	$+0.063$	-0.163	-0.094	0.000	-0.070
Corpus 2013	-0.236	0.000	-0.166	-0.020	$+0.068$	-0.009
NLM Corpus	-0.167	$+0.111$	-0.080	0.000	0.000	0.000
PK Corpus	0.000	0.000	0.000	-0.250	0.000	-0.133
OSCAR	-0.197	$+0.058$	-0.133	-0.033	$+0.079$	-0.021
WorldVista	-0.214	$+0.014$	-0.153	-0.052	$+0.029$	-0.040
French Ref.	-0.210	$+0.028$	-0.146	-0.061	$+0.018$	-0.051
KEGG	-0.226	$+0.010$	-0.160	-0.057	$+0.027$	-0.047
NDF-RT	-0.265	0.000	-0.194	-0.076	0.000	-0.068
ONC Non-Int.	-0.237	0.000	-0.166	-0.074	$+0.007$	-0.063
AVG (external)	-0.206	$+0.022$	-0.143	-0.078	$+0.025$	-0.057

The fine-tuning results (Tables 1 and 2) show that single-source models, despite task-specific optimization, consistently underperform their multi-source counterparts across nearly all metrics and datasets. For Phi-3.5 2.7B trained on SMILES-only inputs, validation accuracy drops 22.3 percentage points (0.690 vs. 0.913), while external dataset performance averages 67.6% accuracy and 75.7% F1-score compared to the complete model's 88.1% and 90.6%, respectively. The SMILES-only model achieves perfect sensitivity (1.000) across all datasets, indicating systematic overprediction of interactions rather than discriminative learning.

GPT-4o trained on organisms-only inputs shows more moderate degradation, achieving 88.8% validation accuracy (Δ: -3.8%) and maintaining competitive sensitivity (0.987, Δ: $+5.7\%$). Across external datasets, the organisms-only model averages 82.4% accuracy (Δ: -7.8%) and 84.9% F1 (Δ: -5.7%).

These patterns reveal that LLM ablations do not measure what traditional ablation studies aim to quantify. When we remove a source and reprompt the model, we are not isolating that source's contribution to a unified reasoning process. Instead, we are comparing different tasks with different optimal strategies. SMILES-only predictions rely on structural heuristics that may succeed for certain interaction types (e.g., competitive binding) but systematically fail for mechanism-based DDIs. Genes-only predictions leverage pathway overlap patterns that capture enzyme-mediated interactions but miss transport or absorption effects. Organisms-only predictions may exploit biological context but lack molecular specificity. Each configuration represents a different inductive bias rather than a decomposition of the same model's decision-making.

Given these limitations, we support our multi-source approach with evidence from three independent sources, acknowledging that none definitively proves necessity. First, our fine-tuned models achieve consistent performance across thirteen heterogeneous external datasets, with Phi-3.5 (2.7B) reaching an average sensitivity of 0.978 when trained on complete multi-source inputs. The ablation experiments demonstrate that single-source fine-tuning systematically fails. The cross-domain robustness across clinical knowledge bases, annotated corpora, and healthcare systems spanning different interaction types suggests the multi-source representation captures diverse DDI mechanisms that cannot be learned from any single information modality. Second, our error analysis (Section 6.4) reveals source-specific failure modes wherein Phi-3.5 (2.7B) false positives correlate with higher molecular complexity (mean SMILES length 125.1 vs. 114.2 characters) and more target genes (8.9 vs. 4.9), while GPT-4o false negatives associate with structural atypicity (mean SMILES length 197.7 characters) independent of gene information. These complementary error patterns imply that different sources provide discriminative power for different interaction subtypes. Third, the performance progression from zero-shot (average sensitivity 0.5463) through few-shot (Sonnet F1: 0.8384) to fine-tuning (Phi-3.5 F1: 0.936 with complete inputs vs. 0.757 with SMILES-only) demonstrates that effective integration of multiple sources requires task-specific adaptation. Zero-shot models have access to all information but cannot leverage it effectively. Few-shot models show variable benefits depending on prompt structure and architecture-specific biases. Fine-tuned single-source models achieve high sensitivity through systematic overprediction rather than true discriminative learning. Only fine-tuned multi-source models achieve consistent gains, meaning that learning to combine sources is non-trivial and not reducible to simple pattern matching on individual sources.

Last but not least, beyond empirical validation, our multi-source approach is motivated by the mechanistic diversity of DDIs documented in pharmacology. Different interaction types require different information modalities:

- Structural interactions (competitive enzyme binding, transporter competition) depend on molecular similarity captured by SMILES
- Metabolic interactions (enzyme induction/inhibition) depend on shared gene targets
- Species-specific effects (variations in P450 isoforms) depend on the organism context

A model limited to any single source would be theoretically incomplete for the task, as different DDI mechanisms require different information modalities. Structural features (SMILES) are essential for predicting competitive interactions at binding sites and transport proteins [1, 2], while gene target profiles capture gene-mediated (e.g., cytochrome P450-mediated) metabolic interactions [3]. Prior methods have demonstrated that gene-only approaches [3] excel at enzyme-mediated DDIs but miss structural interactions, whereas structure-only methods [1, 2] capture competitive binding but overlook pathway-based effects. Our multi-source design integrates these complementary information types to address the mechanistic diversity of DDIs.

The non-standard behavior we observed suggests that future work on LLM-based biomedical prediction should focus on developing evaluation frameworks designed specifically for contextual reasoning systems. Traditional ablation studies, developed for models with explicit feature representations and additive contributions, do not translate directly to systems where performance depends on how information is presented, tokenized, and attended to within complex prompt structures.

References

[1] Santiago Vilar et al. "Drug—drug interaction through molecular structure similarity analysis". In: *Journal of the American Medical Informatics Association* 19.6 (Nov. 2012), pp. 1066–1074. ISSN: 1067-5027. DOI: 10.1136/amiajnl-2012-000935.

[2] Xin Chen, Xien Liu, and Ji Wu. "GCN-BMP: investigating graph representation learning for DDI prediction task". In: *Methods* 179 (2020), pp. 47–54.

[3] Suyu Mei and Kun Zhang. "A machine learning framework for predicting drug–drug interactions". In: *Scientific Reports* 11.1 (2021), p. 17619.