Knowledge-Based Systems 338 (2026) 115486

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

LLMs For drug-Drug interaction prediction using textual drug descriptors

Gabriele De Vito [a,*], Filomena Ferrucci [a], Athanasios Angelakis [b,c,d]

[a] Department of Computer Science, University of Salerno, Via Giovanni Paolo II 132, Salerno, 84084 Fisciano, Italy
[b] Department of Epidemiology and Data Science, Amsterdam University Medical Center, Amsterdam, Netherlands
[c] Digital Health; Methodology, Amsterdam Public Health Research Institute, Amsterdam, Netherlands
[d] Data Science Center, University of Amsterdam, Amsterdam, Netherlands

ARTICLE INFO

Keywords:
Large language models (LLMs)
Drug-drug interactions (DDIs)
Zero-shot learning
Few-shot learning
Fine-tuning.

ABSTRACT

As treatment plans involve more medications, anticipating and preventing drug-drug interactions (DDIs) becomes increasingly important. Such interactions can result in harmful side effects and may reduce therapy effectiveness. Currently, most computational approaches for DDI prediction rely heavily on complex feature engineering and require chemical information to be structured in specific formats to enable accurate detection of potential interactions. This study presents the first investigation of the application of Large Language Models (LLMs) for DDI prediction using drug characteristics expressed solely in free-text form. Specifically, we use SMILES notations, target organisms, and gene associations as inputs in purpose-designed prompts, allowing LLMs to learn the underlying relationships among these descriptors and accordingly predict possible DDIs. We evaluated the performance of 18 distinct LLMs under zero-shot, few-shot, and fine-tuning settings on the DrugBank dataset (version 5.1.12) to identify the most effective paradigm. We then assessed the generalizability of the fine-tuned models on 13 external DDI datasets against well-known machine learning baselines. The results demonstrated that, while zero-shot and few-shot paradigms showed only modest utility, fine-tuned models achieved superior sensitivity while maintaining competitive accuracy and F1-score compared to baselines. Notably, despite its small size, the Phi-3.5 2.7B model attained a sensitivity of 0.978 and an accuracy of 0.919. These findings suggest that computational efficiency and task-specific adaptation are more important than model size in order to capture the complex patterns inherent in drug interactions, and outline a more accessible paradigm for DDI prediction that can be integrated into clinical decision support systems.

1. Introduction

Drug-drug interactions (DDIs) are a serious issue for clinicians as they can affect the desired response in patients taking multiple medications at the same time, and produce adverse effects or inadequate levels of treatment effectiveness [1]. The consequences of these interactions range from adverse drug reactions to reduced therapeutic efficacy and, in the worst case, can even be life-threatening [1,2]. This problem is getting worse as polypharmacy becomes more common [3], given the rise in the population's average age.

All this is confirmed by recent studies [2,4]. On the one hand, the percentage of older adults taking more than one drug is increasing and varies from 40 to 50% in Western countries (United States, Ireland, Sweden) to over 80% in some Asian regions (South Korea, Taiwan) [4]. On the other hand, DDIs cost U.S. healthcare about $5.4B annually [2]. Therefore, there is a clear need to predict potential DDIs before clinical use for patient safety and for efficient drug development.

The detailed mechanisms underlying DDIs are complex and various molecular pathways may be involved, such as competition for drug-metabolizing enzymes (e.g., cytochrome P450), alterations in drug transport proteins, changes in drug absorption due to pH variations, and interference with receptor binding. Perhaps more importantly, drug administration order can significantly influence interaction outcomes, making it an inherently directional problem [5–7]. For example, the effect of drug A on the metabolism of drug B may differ from the effect of drug B on the metabolism of drug A.

Traditional approaches to identifying DDIs rely on experimental methods, including in vitro and in vivo studies [8]. However, given the large number of possible drug combinations, these approaches are time-consuming, expensive, and often impractical. In addition, direct testing of potentially harmful interactions raises significant ethical concerns due to the risk of adverse effects on human subjects [9].

The advances in computational biology and the increasing availability of molecular data have stimulated the development of in silico

* Corresponding author.
E-mail addresses: gadevito@unisa.it (G. De Vito), fferrucci@unisa.it (F. Ferrucci), a.angelakis@amsterdamumc.nl (A. Angelakis).

https://doi.org/10.1016/j.knosys.2026.115486
Received 5 May 2025; Received in revised form 6 January 2026; Accepted 2 February 2026
Available online 4 February 2026
0950-7051/© 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

methods for DDI prediction, in order to complement traditional experimental approaches and improve interpretability and clinical validation [10]. Current machine learning-based prediction methods [11–14], although their success, often require integration of heterogeneous data sources and complex preprocessing [1].

Large Language Models (LLMs) have recently emerged as powerful tools in understanding and generating human-like text [15,16], demonstrating a remarkable capacity to identify hidden patterns in textual data [17]. They have also shown promise in various pharmaceutical applications [18–22], but their potential for direct DDI prediction remains largely underexplored.

This study presents the first investigation of LLMs for DDI prediction, leveraging their ability to process multiple drug information simultaneously (SMILES notation, target organisms, and gene interactions) as text. We explore a complementary approach to established methods (e.g., graph neural network methods), introducing a fundamentally different paradigm that does not require complex feature engineering or structured chemical representations: We enable LLMs to directly interpret and reason about drug interactions using purely textual representations. To this aim, we design specific prompts to instruct LLMs to use the multiple textual drug information to predict DDIs. We evaluate this approach using zero-shot, few-shot, and fine-tuning techniques across 18 different LLMs, ranging from state-of-the-art models to smaller, more efficient variants.

Seven research questions drive the evaluation:

RQ1 *How effectively do LLMs predict DDIs in a zero-shot setting?*
This research question evaluates LLMs' ability to leverage SMILES notation, target organisms, and gene interactions to predict drug-drug interactions without task-specific training.

RQ2 *How effectively do LLMs predict DDIs in a few-shot setting?*
This research question explores whether providing example interactions improves LLMs' DDI prediction performance through in-context learning.

RQ3 *What is the impact of fine-tuning on LLMs' DDI prediction performance?*
This question investigates how fine-tuning affects LLMs' ability to process multiple drug representations for DDI prediction, with particular attention to the relationship between model size and performance.

RQ4 *How do fine-tuned LLMs compare to existing approaches?*
This question assesses the performance of fine-tuned LLMs against baselines - L2-regularized logistic regression and deep learning - across multiple external datasets, evaluating prediction effectiveness and generalizability.

RQ5 *Are fine-tuned LLM predictions robust to dataset construction strategies?*
This research question is related to the dataset construction strategy (i.e., negatives generated from non-interacting drug pairs and balanced datasets), and explores whether it introduced label noise, class imbalance artifacts, or whether fine-tuned LLMs' predictions relied on entity memorization rather than learned interaction patterns.

RQ6 *Are fine-tuned LLM predictions robust to computational and prompt perturbations?*
This research question dives into the stability and sensitivity of models' responses with respect to computational and prompt perturbations.

RQ7 *What error patterns and learned interaction characteristics emerge in fine-tuned LLMs for DDI prediction?*
This question aims to characterize error patterns and learned interactions in fine-tuned models by analyzing prediction outcomes against molecular complexity and examining confidence distributions to distinguish genuine patterns from systematic biases.

The analyses reveal several key findings. First, while LLMs demonstrate limited effectiveness in zero-shot DDI prediction (average sensitivity of 0.5463), and few-shot approaches show improved but variable performance (with some models reaching an accuracy up to 0.8376), fine-tuning significantly improve their performance. Surprisingly, smaller models such as Phi-3.5 (2.7B parameters) achieve the best sensitivity of 0.978, significantly outperforming the L2-regularized logistic regression [13] and MSDAFL deep learning model [23] baselines and larger LLMs, while maintaining competitive accuracy (0.919) and F1-score (0.906). This performance pattern is consistent across 13 external validation datasets, suggesting that model size is not the determining factor for DDI prediction tasks. Robustness analyses further confirmed that these models learned authentic pharmacological patterns rather than dataset artifacts, maintaining stable performance under label noise (>0.85 sensitivity at 10% corruption), extreme class imbalance (weighted F1≥0.95 at 1:52 ratio), and prompt perturbations, with 99.9-100% stability across computational configurations.

The main contributions of the paper are the following:

- We present the first investigation of LLMs for DDI prediction, evaluating 18 models ranging from 1.5B to over 250B parameters.
- We introduce a text input representation that combines SMILES notation, target organisms, and gene interactions, enabling LLMs to leverage multiple aspects of drug information simultaneously.
- We demonstrate a performance progression across prediction approaches, from limited zero-shot performance to improved few-shot learning, with fine-tuned smaller models ultimately achieving superior sensitivity (0.978) compared to larger LLMs and traditional machine learning baselines, while maintaining competitive accuracy (0.919) and F1-scores (0.906).
- We validate the findings through extensive analyses across 13 external datasets, confirming the robustness and generalizability of the fine-tuned LLM-based DDI prediction method.
- We evaluate fine-tuned LLM's robustness across three dimensions –dataset construction strategy, computational stability, and error and learned interaction patterns– demonstrating that models learn pharmacological patterns rather than dataset artifacts or entity memorization, with 99.9-100% stability and consistent performance under realistic clinical scenarios (1:52 class imbalance).
- We provide datasets (excluding the DrugBank dataset, version 5.1.12), source code, results and supplementary material, as a contribution to the research community in our online repository [24].

Structure of the paper. The remainder of this paper is organized as follows. Section 2 reviews related work in DDI prediction. Section 3 describes the datasets used and data preparation process. Section 4 illustrates our method, including model selection, while the empirical evaluation design is described in Section 5. Section 6 discusses the results. Finally, Section 7 presents the implications of the findings and potential future directions, while Section 8 reports the conclusions.

2. Related work

In this section, we review existing approaches for DDI identification and prediction. We then conclude with recent applications of LLMs in drug discovery and the research gap our work addresses.

2.1. Machine learning-based prediction methods

Various computational methods for DDI prediction have been proposed, which mainly explored structured molecular data and involved extensive feature engineering.

Early approaches, like those by Ferdousi et al. [11] and Kastrin et al. [12], were based on the idea that drugs similar in properties may interact [25] and therefore used several drug similarity metrics combined with traditional ML algorithms to identify DDIs.

Techniques based on matrix factorization approached DDI prediction from a different angle by decomposing interaction matrices to help discover latent patterns [26,27]. Rohani et al. used Integrated Similarity-Constrained Matrix Factorization (ISCMF) [26] to integrate multiple similarity types. Shtar et al. [27] framed DDI prediction as a link prediction problem using Adjacency Matrix Factorization (AMF) with factor propagation over graph nodes.

Traditional machine-learning strategies have also been explored to improve DDI prediction. Mei et al. [13] used drug target profiles (i.e., genes) with L2-regularized logistic regression, achieving competitive results and demonstrating that simpler approaches could be effective in this field.

Graph neural networks (GNNs) marked a significant paradigm shift, facilitating DDI predictions by allowing for the direct modeling of molecular structure and network interactions. Using unified architectures, the initial GNN techniques, including GCN-BMP [28] and DeepDrug [29], combined structural graph information with sequential SMILES representations. However, these methods failed to capture the intricate relationships within interaction networks sufficiently and instead focused mainly on individual drug characteristics. Wang et al. [30] attempted to address this limitation by using multi-view learning, which simultaneously processes both molecular structures and DDI network topology through contrastive learning. Despite the fact that this technique enabled more comprehensive interaction modeling, the researchers recognized the need for a more granular understanding of interaction.

Nyamabo et al. [31] pioneered substructure-substructure interaction modeling using co-attention mechanisms (SSI-DDI) – a concept that has been successively advanced by Li et al. [32] with dual-view learning, which couples intra-drug and inter-drug message passing at the atomic level. Zhao et al. [33] tackled the generalization challenge by incorporating contrastive self-supervised learning with deep mix-hop architectures to capture high-order dependencies.

Complementary to these molecule-centric models, Mohamed et al. [34] surveyed and empirically evaluated knowledge-graph embedding methods on biomedical knowledge graphs, showing that low-rank embeddings can accurately predict drug-target links and polypharmacy side-effects (i.e., DDIs).

Current state-of-the-art methods have tried to address the challenge of combining diverse drug information sources. Chao et al. introduced MSDAFL [23], a dual attention framework that captures both intra-drug substructure patterns and inter-drug substructure interactions through self-attention and interactive attention mechanisms. On the other hand, Su et al. proposed KGCNN [35], which advances spatial modeling by incorporating a spatial-aware capsule aggregator that captures relationships between neighboring molecules in biomedical knowledge graphs.

Although these methods achieved good performance, they have similar limitations: they require complex feature engineering, structured chemical representations (such as graphs or matrices), and extensive preprocessing of heterogeneous data sources. Perhaps most importantly, no current model can process multiple drug characteristics (action type, property, pathway, etc.) simultaneously without time-consuming, sophisticated preprocessing or integration frameworks.

2.2. LLMs in drug discovery

Recently, LLMs have been successfully used in drug discovery applications [20,21], especially in drug-target interaction (DTI) prediction. Ahmed et al. developed DTI-LM [20], an innovative framework for DTI prediction that processes protein amino acid sequences and drug SMILES format using LLMs. The authors found that while LLMs can identify similarities and interactions between proteins, they are unable to represent the similarities between drugs adequately. Oniani et al. [21] proposed a method that utilized both protein language models (ESM-2) and chemical language models (ChemBERTa) for predicting DTIs, further corroborating the observations of Ahmed et al. on the limitations of chemical language models and the importance of larger datasets. Recent developments have also introduced specialized biomedical LLMs, such as BioClinical ModernBERT [36] for long-context biomedical text processing, MedGemma [37] for medical applications, Me-LLama [38] for foundation medical models, and OpenBioLLMs [39] for healthcare and life sciences. These domain-specific models show promise for biomedical tasks, but their application to DDI prediction has not been systematically evaluated. To the best of our knowledge, only a few studies have employed general-purpose LLMs [22] in the context of DDI prediction. Xu et al. presented DDI-GPT [22], a model that integrates LLMs with knowledge graphs for DDI prediction. DDI-GPT provides interpretable predictions through gene importance scoring and network analyses. However, it relies on a curated knowledge graph and specific feature engineering techniques (sentence-tree injection and visibility masks), which limit its scalability compared to LLMs that operate directly on textual drug information. Although these studies indicate a growing opportunity for LLMs in drug discovery, the direct application of LLMs for DDI prediction remains an emerging field, presenting opportunities for further research.

2.3. Research gap

The review highlights several limitations of DDI prediction approaches. First, machine learning methods need complex feature engineering and face challenges in integrating multiple data sources seamlessly [1]. Second, in order to attain optimal performance, deep learning techniques require large-scale training data, meticulous architecture design, and intensive hyperparameter tuning [1]. In addition, as we have seen in the section above, LLM applications in this context have primarily focused on specific tasks or aspects, such as DTIs [20,21] or knowledge graph integration [22]. Therefore, the potential of LLMs to process multiple textual drug characteristics (e.g., SMILES notation, target organisms, and gene interactions) for DDI prediction remains unexplored. Our goal is to bridge this gap by determining whether (and how) LLMs can efficiently process only textual drug information to forecast DDIs as a complementary approach to established methods (e.g., graph neural network methods) without the need for intricate feature engineering or organized chemical representations. We aim at exploring various adaptation techniques in order to pinpoint the advantages and disadvantages of LLMs in identifying DDIs and, if applicable, establish a more accessible approach in this field.

3. Datasets and data processing

This section describes the datasets and data processing steps implemented to achieve this objective. Data and source codes are publicly accessible in our online repository [24].

3.1. Datasets

We used two primary data sources to assess LLM performance. The first one is DrugBank [40], a database containing 16,581 drugs, 3921 unique target genes, and 1,420,072 known drug-drug interactions. This rich dataset was the primary source for zero-shot, few-shot, and fine-tuning analyses, and we also used it to generate negative examples (non-interacting drug pairs) for all datasets.

The second data source is the DDI repository provided by Ayvaz et al. [41], which we employed for external validation, as it aggregates drug interactions from 14 distinct sources from clinical knowledge bases, annotated corpora, healthcare systems, reference resources, and clinical guidelines.

The initial number of drugs and DDIs for each dataset is reported in Table 1.

3.2. Data processing

We prepared the datasets for the LLM empirical analyses following a three-step data processing pipeline (see Fig. 1).

Table 1
Drugs and initial DDI for datasets.

Category	Dataset	Drugs	DDI
Primary Dataset	DrugBank	16,581	1,420,072
Clinical knowledge bases	CredibleMeds	63	83
	HEP	557	11,194
	HIV	556	19,198
Annotated corpora	DDI Corpus 2011	244	334
	DDI Corpus 2013	410	787
	NLM Corpus	131	238
	PK DDI Corpus	85	146
Healthcare systems	OSCAR EMR	227	10,325
	WorldVista	378	13,693
Reference resources	French DDI Referrals	854	62,047
	KEGG	1033	52,104
	NDF-RT	425	1,876
Clinical guidelines	ONC High Priority DDI List	123	193
	ONC Non-Interruptive DDI List	187	2,101

Table 2
Datasets with positive and negatives drug pairs.

Dataset	Positive	Negative	Total
DrugBank	1,035,150	1,035,150	2,070,300
CredibleMeds	5	5	10
HEP	1271	1271	2542
HIV	5651	5651	11,302
DDI Corpus 2011	32	32	64
DDI Corpus 2013	74	74	148
NLM Corpus	9	9	18
PK DDI Corpus	2	2	4
OSCAR EMR	2052	2052	4104
WorldVista	1513	1513	3,026
French DDI Referrals	4297	4297	8,594
KEGG	6631	6631	13,262
NDF-RT	119	119	238
ONC High Priority DDI List	0	0	0
ONC Non-Interruptive DDI List	291	291	582
LLMs training set	500	500	1,000
LLMs validation set	545	545	1,090

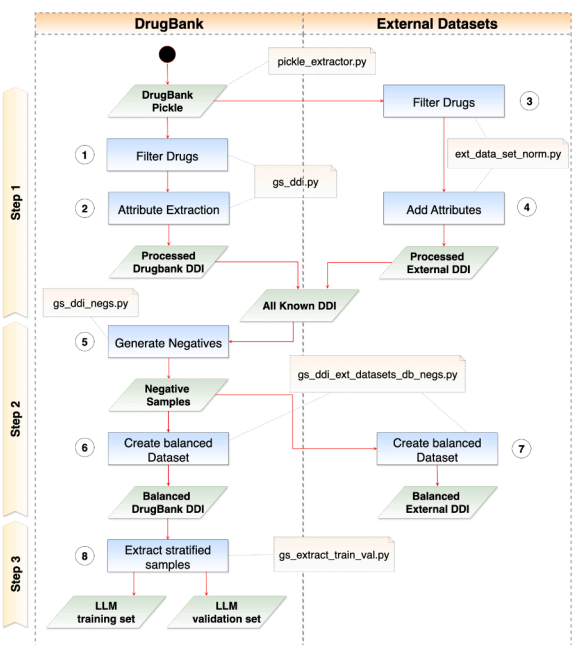

Fig. 1. Data preprocessing pipeline.

Following the principle that "two drugs potentially interact when a drug alters the other drug's therapeutic effects through targeted genes or signaling pathways´´ [13], we focus on drugs targeting at least one gene.

In the first step, we filtered the DrugBank database to include only drugs that were approved or experimental and excluded withdrawn or illicit drugs. Furthermore, we selected only drug pairs where both drugs targeted at least one gene. For each drug pair, we extracted DrugBank IDs, SMILES notations, target organisms, and two binary vectors (length 3,921) representing gene targets, where each position corresponded to a gene in DrugBank's lexicographically ordered gene list. At the end of this process, we collected 1,035,150 positive DDIs. We then processed the 14 datasets from Ayvaz et al. [41], filtering out drugs not present in the processed DrugBank dataset and removing any interactions already present in the DrugBank dataset. We enriched the remaining drug pairs with the necessary features from DrugBank, such as SMILES, target organisms, and gene vectors. We obtained 13 usable datasets, due to the fact that the ONC High Priority DDI dataset contained no valid interactions after filtering, totaling 21,947 additional unique DDIs. In the second step, we combined all known interactions from DrugBank and external datasets to generate 1,057,097 negative examples. For each candidate negative pair (A,B), we verified that neither (A,B) nor (B,A) existed as known interactions in any dataset, thereby ensuring proper handling of DDI directionality and preventing potential 'backward leakage' where reversed versions of the same drug pair could appear in different splits.

We used 1,035,150 negative examples to balance the processed DrugBank dataset and the remaining 21,947 examples to balance the external datasets to have an equal number of positive and negative examples. This approach ensured an equal number of positive and negative examples in all datasets, in accordance with accepted DDI prediction techniques [13,31] and established procedures for addressing class imbalance in machine learning [42,43].

To preliminary assess the negative sampling strategy, we analyzed structural and biological features (SMILES lengths and genes counts for single drugs and pairs) on 43,894 drug pairs from external datasets. All drug pairs are unique by design in their directional form (A,B), ensuring independence at the pair level. Using 10,000 permutation tests of pair labels with Bonferroni correction ($\alpha_{adj} = 0.0083$) [44,45], we found that while median differences were significant for five out of six features, all effect sizes were small-to-medium (Cohen's d = 0.129-0.354), implying substantial overlap between classes. All this indicates that models cannot trivially separate classes using simple thresholds but must learn complex feature interactions (see details in Supplementary Material S.R3.2.1).

In the last step, we randomly selected 1000 examples from the DrugBank balanced dataset for training and 1090 for validation using stratified sampling, as supported by recent research on LLMs that demonstrated that performance plateaus beyond the 1,000-example threshold [46–49]. Finally, we performed a systematic verification to confirm the absence of backward leakage across all data splits (training, validation, and external test sets). The analysis revealed zero instances where training contains (A,B) while validation or any external test set contains (B,A), and vice versa, thereby ensuring proper handling of DDI directionality across the entire experimental setup.

The data preprocessing pipeline and the processed datasets (with the exception of the DrugBank dataset, which cannot be publicly distributed) are available in our online repository [24] for reproducibility. The final distribution of positive and negative drug pairs for each dataset is illustrated in Table 2.

4. Methods

This section describes the methodological approaches used to adapt LLMs for DDI prediction. More in detail, we implemented three increasingly sophisticated adaptation approaches using the preprocessed

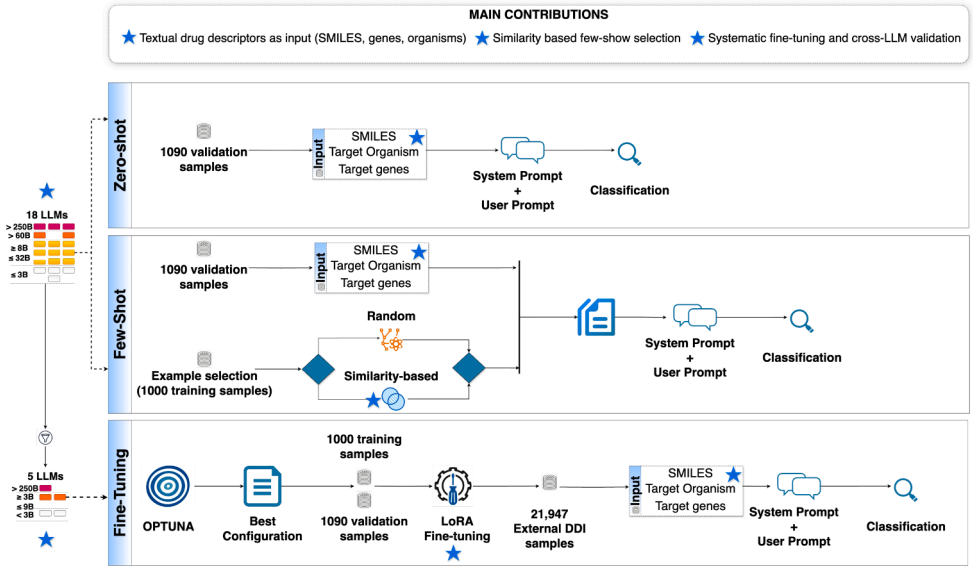

Fig. 2. Methods overview showing the three-tier LLM approach for DDI prediction: Zero-shot, few-shot, and fine-tuning workflows with drug pair input processing and classification output.

datasets described in Section 3.2 (see Fig. 2): (1) zero-shot learning leveraging models' inherent knowledge, (2) few-shot learning with exemplar drug interactions, and (3) fine-tuning for task-specific optimization. In the following, we outline the model selection criteria and then describe the implementation of each adaptation approach.

4.1. Selected LLMs

We evaluated 18 different LLMs, ranging from 1.5B to over 250B parameters, which can be categorized into four groups:

- Proprietary models: GPT-4o [50], Claude 2.1 [51], and Gemini 1.5 Pro-[52], representing state-of-the-art commercial LLMs accessed through their respective APIs.
- Open-weights large models: Llama 3.3-70B [53] and Qwen2 72B [54], which are publicly available models with architectures and parameter count comparable to commercial solutions.
- Open-weights middle-range models: Granite 3.1 8B [55], Llama 3.1 8B [53], Gemma 2 9B [56], Falcon 3 10B [57], Mistral-Nemo 12B [58], Qwen 2.5 14B [54], Gemma2 27B [56], Aya-expanse-32b [59], and Qwen 2.5 32B [54], representing a balance between computational efficiency and model capacity.
- Open-weights efficient models: Phi-3.5 2.7B [60], Qwen2.5 3B [54], Llama3.2 3B [53], and DeepSeek R1 distilled Qwen 1.5B [61], representing recent advances in efficient model architectures.

There were several considerations that guided our selection. First, we aimed for model diversity across different scales, including both large-scale architectures and smaller models (1.5B-7B parameters) to investigate the relationship between model size and performance and assess practical deployment scenarios. Second, we took into account LLM accessibility, selecting both proprietary and open-weight models in order to evaluate the feasibility of DDI prediction across different deployment scenarios. Finally, we included the latest models, such as Phi-3, Gemma 2, and DeepSeek R1, to evaluate cutting-edge architectures. We accessed the proprietary models by using their respective API; whereas, for local deployment of open-weights models, we utilized LM Studio [62]. LM Studio provides a unified interface for model deployment and inference, and supports various model architectures and configurations. This platform enabled us to maintain consistent experimental conditions across all open-weight models, offering an OpenAI-compatible REST API.

4.2. Zero-shot approach

The first adaptation approach we implemented is zero-shot learning, which aims to assess LLMs' inherent ability to predict DDIs without task-specific training and without providing examples of interactions. For prompt design, we paid attention to three key aspects: i) providing drug information in a structured format, ii) clearly specifying the binary classification task, and iii) implementing explicit output constraints.

To this aim, we conceived the prompt structure reported in Fig. 3. On the one hand, the system prompt emphasizes the importance of drug administration order and constrains the model's output to a binary classification. On the other hand, the user prompt explicitly includes all available drug characteristics: drug names, molecular structures (SMILES notation, such as "CC(C)(C)OC=O´´), target organisms, and gene interactions (i.e., Glucosamine-targeted gene "IFNG´´). This information is inserted directly as formatted text strings, i.e., with SMILES notation "as is´´, genes as comma-separated names, and organisms as text labels. LLM tokenizers process it as ordinary sequences of characters or words, without any intrinsic chemical understanding; this means that the model treats it as textual patterns, and any domain knowledge emerges from patterns learned during pretraining, where such notations co-occurred with discussions of drug interactions.

Although the instructions requested only the classification term, some models may provide additional explanations to support their classification. For these cases, we used a pattern-matching approach that searched the response text for one of the classification keywords ("interaction´´ or "no interaction´´), giving priority to exact matches that followed the "CLASSIFICATION:´´ marker when present.

4.3. Few-shot approach

The second adaptation approach is few-shot learning, which allows teaching the LLM with respect to the DDI prediction task through as few as 2–5 examples of input-output pairs. We implemented and compared two strategies for example selection, namely random and similarity-based, in order to assess whether the inclusion of curated examples improves LLM performance in DDI prediction. For both approaches, we used a balanced set of 10 examples, comprising 50% positive and 50% negative interactions, extracted from the training dataset (see Section 3.2), to maintain an equal representation of both interaction classes. We limited examples to 10 due to context length constraints in

System prompt

You are an expert in drug-drug interaction. Given two drugs, where the order of administration counts, the genes and organisms targeted by the two drugs and the SMILES formulas of the two drugs, classify whether their administration causes 'interaction' or 'no interaction.' Answer only with the classification ('interaction' or 'no interaction'), nothing else.

User prompt: Zero-shot setting

Drug1: drug1
SMILES for drug1: smiles1
Organism targeted by drug1: org1
Genes targeted by drug1: genes1
Drug2: drug2
SMILES for drug2: smiles2
Organism targeted by drug2: org2
Genes targeted by drug2: genes2
CLASSIFICATION:

Fig. 3. Zero-shot prompt template.

some LLMs (i.e., a 8192 token limit) while maintaining sufficient information for effective inference.

In the first strategy, we used random selection, where examples were chosen arbitrarily from the training set. While this approach introduces models to the general task format and diverse interaction patterns, it does not ensure that examples have similarities to the query pair.

We developed a second strategy using an embedding-based similarity search to test whether contextually relevant examples would yield better performance. We computed a composite similarity measure for each test drug pair based on six embedding vectors: SMILES notation, target organisms, and target genes for each drug in the pair.

We employed OpenAI's text-embedding-3-small model (released January 2024) to compute embeddings for similarity matching. This model supports up to 8192 tokens per input using the cl100k_base encoding scheme and generates 1,536-dimensional embeddings. More specifically, we calculated embeddings for six different components: SMILES notation, target organisms, and gene interactions for each drug.

The composite similarity score S between a query pair (d_i, d_j) and a training example pair (d_k, d_l) was calculated as:

$$S((d_i, d_j), (d_k, d_l)) = \frac{1}{6} \sum_{c \in C} \text{cosim}(e_c(d_i, d_j), e_c(d_k, d_l)) \quad (1)$$

where $C = \{s_1, s_2, o_1, o_2, g_1, g_2\}$ represents the six components (SMILES, organisms, genes for each drug), e_c denotes the embedding function for component c, and $\text{cosim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}$ is the cosine similarity.

We then identified the top matches while retaining the balance of positive and negative examples. This equally weighted approach allows the drug characteristics to have the same influence on similarity assessment without manual feature engineering efforts.

We used the same system prompt as in the zero-shot setting (see Section 4.2) and augmented the user prompt with the selected examples as shown in Fig. 4. In order to extract classifications from LLM outputs, we exploited the pattern-matching strategy described in Section 4.2.

To ensure that our approach did not require any truncation of drug information, we analyzed the token distribution across all components in the training dataset (i.e., the dataset used for examples selection is the training set as reported in Section 3.2).

As shown in Table 3, SMILES notations averaged 45.65 ± 46.45 and 45.12 ± 47.14 tokens for Drug 1 and Drug 2, respectively, gene lists averaged 18.10 ± 37.86 and 16.25 ± 29.63 tokens, and organism names averaged approximately three tokens. The total token count across all six components (SMILES, genes, and organisms for both drugs) averaged 125.66 ± 80.09 tokens, with a maximum of 657 tokens observed in the most complex drug pair. Even at the 99th percentile (510.12 tokens), all inputs remained well within the model's capacity, requiring no truncation and ensuring complete information preservation.

User prompt: Few-shot settings

INSTRUCTIONS
Analyze the provided drug information and consider the following examples to answer correctly:
** Examples **

Example 1:
Drug1: example_drug1_name
SMILES for drug1: example_smiles1
Organism targeted by drug1: example_org1
Genes targeted by drug1: example_genes1
Drug2: example_drug2_name
SMILES for drug2: example_smiles2
Organism targeted by drug2: example_org2
Genes targeted by drug2: example_genes2
CLASSIFICATION: interaction [or no interaction]
...

Example 10:
...

CURRENT TASK INPUT DATA
Drug1: {drug1}
SMILES for drug1: {smiles1}
Organism targeted by drug1: {org1}
Genes targeted by drug1: {genes1}
Drug2: {drug2}
SMILES for drug2: {smiles2}
Organism targeted by drug2: {org2}
Genes targeted by drug2: {genes2}
CLASSIFICATION:

Fig. 4. Few-shot prompt template.

4.4. Fine-tuning approach

As the final level in the adaptation approach, we implemented a fine-tuning technique. Fine-tuning helps specialize LLMs for specific tasks by adjusting their parameters through supervised learning on task-specific data. In our context, we fine-tuned selected models to optimize their DDI prediction capabilities using the same textual drug information and prompt structure described in Section 4.2. To implement this approach, we leveraged the LLM training and validation sets described in Section 3.2. Specifically, we used 1000 balanced training examples (500 positive interactions, 500 negative interactions) and 1090 validation examples (545 positive, 545 negative), selected via stratified sampling from the processed DrugBank dataset. We created JSONL files containing conversational sequences structured with system prompts, user prompts (as presented in Section 4.2), and assistant responses (either "interaction" or "no interaction"). The instruction format, training set composition, and validation set composition were kept identical for both proprietary and the open-weight models fine-tuned with LoRA, enabling direct performance comparison across all models.

Among proprietary models, we chose GPT-4o as the state-of-the-art representative for fine-tuning analysis. Indeed, Google and Anthropic do not offer fine-tuning capabilities for Gemini 1.5 Pro-and Claude Sonnet 3.5. For fine-tuning GPT-4o, we utilized OpenAI's API and set the following hyperparameter values, as suggested by the OpenAI documentation [63]: 3 epochs, batch size 3, and learning rate multiplier 0.3.

Table 3
Token distribution statistics by component for few-shot similarity matching in the training set.

Component	Count	Min	Max	Mean	Median	Std	P25	P75	P95	P99
SMILES 1	945	2	591	45.65	38.0	46.45	25.0	51.0	95.0	175.9
SMILES 2	938	2	610	45.12	37.0	47.14	24.0	52.0	106.3	195.5
Organism 1	992	1	53	2.99	1.0	5.38	1.0	1.0	11.0	29.0
Organism 2	989	1	53	2.91	1.0	5.17	1.0	1.0	11.0	26.0
Genes 1	1000	1	465	18.10	5.0	37.86	3.0	14.0	87.0	162.0
Genes 2	1000	1	275	16.25	6.0	29.63	3.0	14.0	77.0	156.0
Total Token per pair	1000	6	657	125.66	108.0	80.09	78.0	150.0	249.20	510.12

Table 4
Hyperparameter search space.

Hyperparameter	Type	Values/Range	Notes
Learning Rate	Categorical	{1e-5, 5e-5, 1e-4, 2e-4, 3e-4, 5e-4}	
Batch Size	Categorical	{8}	{4} for Gemma2 9B
LoRA Layers	Categorical	{16, 24, 28}	7 projections per layer
LoRA Rank	Categorical	{16, 32}	
LoRA Alpha Scaling	Categorical	{1, 2}	1: $\alpha = \text{rank}$; 2: $\alpha = 2\times\text{rank}$
LoRA Dropout	Float	[0.0, 0.03]	Step: 0.01
LoRA Scale	Float	[4.0, 4.3]	Step: 0.1

We chose four representatives for open-weights models: Phi-3.5 2.7B, Qwen2.5 3B, Deepseek R1 Distilled Qwen 1.5B, and Gemma2 9B. Our interest in these LLMs was motivated by exploring how smaller models performed as compared to state-of-the-art models like GPT-4o, as well as some practical considerations about computational resources and deployment options.

To determine the fine-tuning strategy, we considered multiple alternatives. In preliminary experiments with Phi-3.5 2.7B, full fine-tuning consistently led to catastrophic forgetting – i.e., a phenomenon where a neural network loses previously acquired knowledge when learning new information – regardless of learning rate or number of epochs. The model converged to degenerate solutions, predicting only one class with high confidence while ignoring the other. Reducing trainable parameters by fine-tuning only a subset of layers improved discrimination but resulted in suboptimal performance. For detailed results, see Appendix A. We also considered QLoRA [64] but opted against it since it trades accuracy for memory efficiency through quantization, and our focus was on maximizing prediction performance.

Finally, we employed Low-Rank Adaptation (LoRA) [65] for fine-tuning open-weight LLMs. LoRA enables efficient fine-tuning while preserving the model's general capabilities, adapting pre-trained weights $\mathbf{W}_0$ by adding trainable low-rank matrices:

$$\mathbf{W} = \mathbf{W}_0 + \frac{\alpha}{r}\mathbf{B}\mathbf{A} \tag{2}$$

where $\mathbf{B}$ and $\mathbf{A}$ are low-rank matrices with rank r much smaller than the original weight dimensions, and α is a scaling factor.

To optimize the fine-tuning parameters, we implemented a hyperparameter search using Optuna [66], aiming to minimize validation loss while avoiding overfitting. We employed a two-phase Optuna-based optimization process. In Phase 1 (Aggressive Screening), we executed 1000 trials per model and employed a *SuccessiveHalvingPruner* for early elimination of unpromising configurations. Each trial consisted of 30 iterations at maximum, with validation every five steps (2 batches), which typically yields 6–10 candidates per model ranked by best validation loss.

In Phase 2 (Intensive Validation), the best configurations were rerun without pruning for 300 iterations with validation every 50 steps (25 batches). The best configuration per model was selected based on minimum validation loss. Because Phase 1 quickly removes subpar configurations (the majority of trials were pruned after 1–2 validations at iterations 5–10) and Phase 2 firmly validates the best candidates, this strategy is effective.

The search space was identical across models (see Table 4), except for Gemma2 9B, which used a batch size of 4 due to memory constraints (others used 8). LoRA adapters were applied to seven projection layers, namely self_attn.q_proj, self_attn.k_proj, self_attn.v_proj, self_attn.o_proj, mlp.gate_proj, mlp.up_proj, and mlp.down_proj.

We used the Adam optimizer [67] with a cosine decay learning scheduler across 1000 trials. We monitored validation loss and stopped training when it ceased to improve for two consecutive validation iterations. During fine-tuning, model adapters were saved every 100 iterations and merged with the base model layers at the end of training. For the final evaluation, we picked the checkpoint with the lowest validation loss for each model.

The computational details regarding the fine-tuning processes are reported in Supplementary Section S.R6.3.1. To corroborate that performance plateaus around 1000 training examples with fine-tuned LLMs, we conducted an additional analysis using Phi-3.5 2.7B with 2000 training samples (twice our original size). The results, reported in the Supplementary material (see S.R6.3.2), confirmed the plateau and supported our initial design choice.

5. Empirical evaluation

In this section, we outline the empirical evaluation design used to assess LLMs' performance in DDI prediction, their predictive capabilities, robustness, and interpretability. All analyses were performed using a MacBook Pro-M3 Max, equipped with 96GB of RAM, 14 cores, and an integrated GPU with 30 cores.

5.1. Metrics

We assessed the performance of the zero-shot, few-shot, and fine-tuning methods using various complementary metrics, namely sensitivity, precision, F1-score, and accuracy. 'Sensitivity' measures the model's ability to correctly identify positive interactions, which is particularly crucial in drug safety applications. 'Precision´ and 'F1-score´ respectively quantify the proportion of true positive predictions and the harmonic mean of precision and sensitivity. We also calculated the 'accuracy´ to measure general performance across both classes. Since the datasets are fully balanced regarding the 'target´ (i.e., the same number of positive and negative data instances), 'accuracy´ is a reliable performance metric. We did not use ROC-AUC since the LLMs do not output probabilities.

5.2. Research method for RQ_1

To address RQ1, we systematically evaluated the 18 LLMs described in Section 4.1 through the zero-shot approach. Each model processed the same 1090 balanced drug pairs from the validation dataset (see Section 3.2) using the zero-shot prompt structure detailed in Section 4.2. We set the temperature to 0.0 for all evaluations and models to mitigate non-deterministic outputs, and accessed proprietary models through their respective APIs, and open-weight models through LM Studio's REST API. Each model generated a binary classification ("interaction´´ or "no interaction´´) for each drug pair, which we compared against

ground truth labels. We then calculated the four metrics defined in Section 5.1 (namely, Accuracy, Precision, Sensitivity, and F1-score) in order to establish a baseline for LLMs' intrinsic DDI prediction capabilities as a comparison with the other adaptation approaches.

5.3. Research method for RQ_2

To answer RQ2, we applied and assessed the two few-shot strategies described in Section 4.3: random selection and similarity-based selection of examples. We followed the procedure reported in Section 5.2 for both strategies, using the same inference configurations, LLMs, validation dataset, and performance metrics, to maintain consistent conditions and enable direct comparison between zero-shot and few-shot approaches. To further validate the robustness of few-shot approaches, we conducted perturbation analysis (see Supplementary Material S.R6.2.1) and ablation studies examining individual contributions of drug textual information (see Supplementary Material S.R6.2.2).

5.4. Research method for RQ_3

RQ3 aims to assess the impact of fine-tuning on the capabilities of LLMs in DDI prediction. We used the five fine-tuned models described in Section 4.4 (i.e., GPT-4o, Phi-3.5 2.7B, Qwen 2.5 3B, DeepSeek R1 Distilled Qwen 1.5B, and Gemma2 9B), picking the resulting fine-tuned version of GPT-4o and the best fine-tuned versions for the open-weight models in terms of validation loss. Using the same experimental setup and metrics as RQ1 and RQ2, we evaluated each fine-tuned model on the validation set, computing accuracy, sensitivity, and F1-score. We compared these results against the top-performing LLMs identified in RQ1 (zero-shot, see Section 5.2) and RQ2 (few-shot, see Section 5.3).

To further validate the fine-tuning gains without potential optimistic bias from checkpoint selection on the validation set, we conducted an independent evaluation on the two largest external datasets: KEGG (N = 13,262 pairs) and HIV (N = 11,302 pairs). These datasets were never used for training, example selection, or checkpoint selection. We compared all five fine-tuned models against the two best-performing LLMs from RQ1 and RQ2 (ranked by F1-score values) using accuracy, sensitivity, and F1-score metrics, providing an unbiased assessment of fine-tuning effectiveness on completely unseen data.

5.5. Research methods for RQ_4

RQ4 aims to assess the generalizability of fine-tuned LLMs (i.e., GPT-4o, Phi-3.5 2.7B, Qwen 2.5 3B, DeepSeek R1 Distilled Qwen 1.5B, and Gemma2 9B) and evaluate their performance against baselines. We selected the L2-regularized logistic regression model [13] and the MSDAFL deep-learning model [23] as baselines. The baseline comparisons served two purposes: first, to reproduce the original results reported in the published works [13,23], and second, to provide a direct performance comparison between our LLM-based approaches and established state-of-the-art methods on the same datasets. The L2-regularized logistic regression model was selected due to the impressive performance metrics, achieving an 'accuracy' of 0.9479, 'sensitivity' of 0.9556, 'specificity' of 0.948, and 'ROC-AUC' of 0.9884, the authors claimed in their original work. First of all, we created new training and validation sets (95% and 5%, respectively) from the balanced DrugBank dataset described in Section 3.2. Then we tuned the model using a stratified 10-fold cross validation and exploring the regularization parameter C in the range $[2^{-16}, 2^{16}]$, as specified in the original paper [13].

On the other hand, we chose MSDAFL as the second baseline because the authors compared the performance to ten leading methods (such as GCN-BMP, DeepDrug, and SSI-DDI) and demonstrated superior performance across multiple metrics in their 2024 study. We utilized the authors' replication package and followed their methodology to train the model. Specifically, we extracted 21,000 examples from the balanced DrugBank dataset described in Section 3.2 using stratified sampling for creating the training set, and used it to train the model for 300 epochs with a learning rate schedule of 0.001 for the first 150 epochs and 0.0001 for the subsequent 150 epochs. As stated in the original work [23], we used the Xavier initialization [68] to initialize the model and the Adam optimizer [67] to optimize it, and we also kept the original hyperparameter settings, namely: a batch size of 512, a node embedding dimension of 128, 60 representative vectors (M), and five layers in the Graph Isomorphism Network [69] structure.

In order to assess the fine-tuned LLMs and trained baselines, we evaluated each of them on the 13 external balanced datasets detailed in Section 3.2 and used the metrics described in Section 5.1. For LLMs, we adopted the same setup as in RQ1 and RQ2. We finally compared the LLM's performance against the baselines employing accuracy, sensitivity, and F1-score metrics.

5.6. Research methods for RQ_5

The absence of complete negative interaction databases and the scarcity of positive examples in clinical practice are two inherent difficulties in DDI prediction. We created negative samples from drug pairs with no known interactions according to established deep learning [42,43,70] and DDI prediction techniques [13,31], and we balanced the class distribution to prevent models from trivially optimizing for the majority class. Nevertheless, this approach raises three possible challenges that need to be empirical validated, namely 1) artificial 1:1 balancing may inflate metrics compared to real-world prevalence (where positives are rare); 2) sampled negatives may include unknown positives that have not yet been documented, introducing label noise; 3) fine-tuned models may rely on learning interaction patterns instead of memorizing drug names from training exposure.

RQ5 specifically examines the robustness of the dataset construction strategy. We conducted four analyses to answer this research question.

First, we carried out a prospective validation using an updated version of DrugBank (version 5.1.14 February 2025), released one year after our initial data collection. The new DrugBank dataset comprises 2,855,848 drug interactions, which is more than twice the number of interactions we extracted from the previous version (1,035,150) after applying the same processing procedure described in Section 3.2. We extracted all negative drug pairs from the training (N = 500) and validation (N = 545) sets, and 13 external datasets (N = 21,947 negative pairs) – described in Section 3.2 – and checked whether any had been subsequently confirmed as interacting in the newer version of DrugBank. In order to ascertain whether the fine-tuned models accurately predicted the interaction in spite of the sampled negative label, we examined the initial predictions for each prospectively validated interaction, and quantified the false negative rate (i.e., the proportion of sampled negative pairs subsequently confirmed as positive interactions).

Second, we conducted a sensitivity analysis by randomly selecting 1%, 5%, and 10% of negative samples and flipping their labels to positive in the external test datasets, and evaluating all fine-tuned models across these new datasets using the metrics reported in Section 5.1 to assess the fine-tuned LLMs' performance degradation.

Third, we employed an imbalanced dataset with a 1:52 ratio, matching the natural distribution found in the 5.1.14 version of DrugBank, where positive interactions account for roughly 1.9% of all potential drug pairs. We created this imbalanced dataset, preserving the diversity and representativeness of the original external datasets. Specifically, we applied stratified subsampling to select 400 positive examples and 20,800 negative examples from the external datasets, ensuring that each source dataset contributed proportionally to its representation in the original data. For instance, if dataset X contained 30% of all available positive examples, it contributed approximately 30% of the sampled positives. This approach resulted in a final test set of 21,200 examples (400 positives and 20,800 negatives). Finally, we computed the accuracy, weighted F1-score, and sensitivity (for both positives and

negatives) metrics for the fine-tuned LLMs and baselines (L2-regularized logistic regression model and the MSDAFL deep-learning model) and compared the results.

Finally, to test whether models relied on drug name memorization rather than learned interaction patterns, we stratified each evaluation set (validation and external datasets) into three disjoint groups based on training exposure, taking into account that pair-level duplication was strictly prevented across all splits (see Section 3.2): NO ENTITY OVERLAP (neither drug seen during training), ONE ENTITY OVERLAP (exactly one drug seen), and TWO ENTITY OVERLAP (both drugs seen individually in training).

For each model and dataset, we performed three pairwise comparisons using the disjoint groups (NO vs. ONE, NO vs. TWO, ONE vs. TWO). For proportion-based metrics, namely accuracy, sensitivity, precision, and error rate = (False Positives + False Negatives)/N, we applied two-sided tests for independent proportions using Fisher's exact test when any cell count was < 5 or total $n < 30$, otherwise the pooled Z-test. We computed Wilson 95% confidence intervals for group-specific proportions and Cohen's h for effect sizes, interpreted using standard conventions (small $|h| \approx 0.2$, medium $|h| \approx 0.5$, large $|h| \approx 0.8$).

For F1-score, we used stratified nonparametric bootstrap with 10,000 resamples per group, computing percentile-based 95% confidence intervals for group-specific F1 and for pairwise differences ΔF1. Two-sided p-values were derived from the proportion of bootstrap iterations where ΔF1 crossed zero.

Multiplicity was controlled using the Holm step-down procedure applied separately to each model and dataset combination. For each combination, the 15 raw p-values (12 from proportion metric pairwise tests + 3 from F1 pairwise bootstrap tests) were ranked from smallest to largest and sequentially tested against decreasing thresholds $\alpha/(m-k+1)$, where $m = 15$ is the number of tests per combination and k is the rank.

A test was declared significant only if its Holm-adjusted p-value ≤ 0.05. Systematic performance advantages for higher entity overlap indicate memorization effects; conversely, the absence of such patterns supports that models learned generalizable interaction patterns independent of entity familiarity.

5.7. Research methods for RQ_6

To answer RQ6, we conducted three sets of analyses to assess the stability and robustness of fine-tuned LLMs' predictions.

To investigate the cross-run stability of fine-tuned LLMs' responses, we adopted the same setup and performance metrics used in RQ1 and RQ2, and computed agreement percentages across five runs along with means, standard deviations, and 95% confidence intervals for all performance metrics (accuracy, sensitivity, F1-score) on the external datasets. The two-sided 95% confidence intervals was calculated as

$$CI = \mu \pm t_{0.975,n-1}, \frac{s}{\sqrt{n}}, \tag{3}$$

where μ and s are the sample mean and standard deviation over n repeats (default $n = 5$), and $t_{0.975,n-1}$ is the Student's t-quantile.

Given that the LLMs' predictions were carried out in parallel batches (see source code in the online repository), we conducted two additional analyses, each comprising five computational configurations. More in detail, we assessed LLMs' inference using five different seeds (42, 123, 456, 789, 1024) at a fixed batch size of 15 and five different batch sizes (1, 4, 16, 32, 64) at a fixed seed (123) in order to investigate computational non-determinism. Indeed, although temperature = 0 minimizes sampling variance by design, these sweeps allow us to verify robustness to floating-point precision, GPU scheduling, and batching effects that can introduce small variations in practice. For each group of five configurations (seed sweep and batch size sweep), we evaluated the fine-tuned models on the validation set, computing accuracy, sensitivity, and F1-score for each of the five runs. We then calculated the agreement percentage (i.e., the proportion of identical predictions across the five runs within each group) along with 95% confidence intervals for all metrics across the five runs of each configuration group.

Regarding the fine-tuned models' sensitivity to prompt hit perturbations, we tested robustness to semantically prompt changes for the two best fine-tuned models by F1 score resulting from RQ4, using four variants (P1-P4): P1 reorders drug fields to genes→organisms→SMILES, testing dependence on positional encoding; P2 paraphrases the instruction line while preserving semantic content ("classify whether their administration causes´´ → "determine if administering these drugs results in´´), testing lexical sensitivity; P3 switches to a list format with Drug A/Drug B labels, testing structural robustness; P4 uses a compact horizontal format with pipe-separated fields, testing spacing/delimiters and reduced texts. Full prompts are provided in our online repository. We then evaluated each variant (P1-P4) on the KEGG dataset using accuracy, sensitivity, and F1-score. For each variant, we calculated the agreement percentage with the original prompt predictions and the absolute difference (Δ) between each metric and the corresponding baseline value obtained with the original prompt used in RQ4.

5.8. Research methods for RQ_7

RQ7 focuses on characterizing the error and learned interaction patterns in the fine-tuned LLMs. To this aim, we proceeded through two phases. We first performed a qualitative error analysis across the 13 external datasets in the top-performing fine-tuned models resulting from RQ4 in terms of sensitivity and F1-score values. To this aim, we stratified predictions into true positives, true negatives, false positives, and false negatives categories and analyzed three correlates: (1) molecular complexity (combined SMILES string length), (2) pharmacological complexity (combined number of target genes), and (3) drug-level patterns (drugs misclassified across multiple datasets). For each dataset and prediction category, descriptive statistics (mean ± SD) were calculated, and patterns were reported when they were consistent across at least 50% of the datasets.

We then analyzed the prediction confidence of the fine-tuned version of GPT-4o across all the validation and external datasets. The confidence score is calculated via log probabilities as

$$confidence = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log p_i\right) \tag{4}$$

Using the Mann-Whitney U test with a one-sided alternative hypothesis (TP > FP), we compared the confidence distributions of true positives and false positives, calculating p-values and rank-biserial correlation as measures of effect size. $r = 1 - \frac{2U_{\min}}{n_1 \cdot n_2}$ is the formula for the rank-biserial correlation r. The smaller of the two Mann-Whitney U statistics, $U_{\min}$, has the following interpretation thresholds: negligible ($|r| < 0.1$), small ($0.1 \leq |r| < 0.3$), medium ($0.3 \leq |r| < 0.5$), and large ($|r| \geq 0.5$). A significantly higher confidence level for true positives ($p < 0.001$, positive effect size) suggests that learned interaction patterns, not indiscriminate positive predictions, are the cause of the model's high sensitivity.

We focused only on GPT-4o because, up to date, LM Studio does not return log probabilities for predicted tokens in its API. The analysis aims to verify that high sensitivity reflects genuine discriminative ability rather than uniform positive bias.

6. Results and discussion

This section presents and discusses the findings around the seven research questions.

6.1. RQ1: Zero-shot learning analysis

The zero-shot evaluation (see Fig. 5 and Table B.18 for details) revealed several interesting patterns across different models, sizes, and architectures. Proprietary models (i.e., GPT-4o, Claude 3.5 Sonnet, and

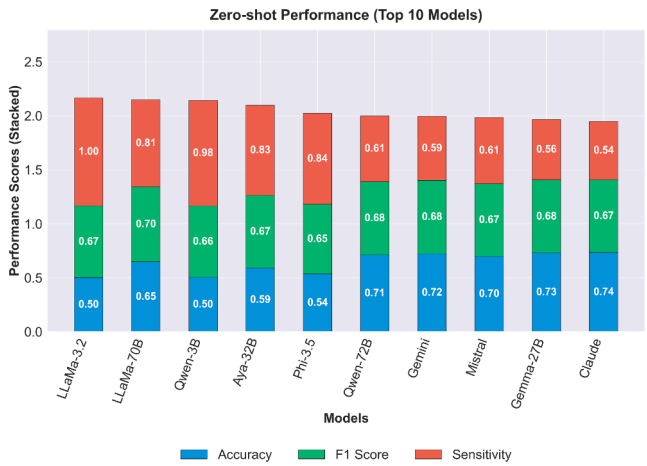

Fig. 5. Zero-shot performance of the top 10 LLMs on DDI prediction: Validation dataset.

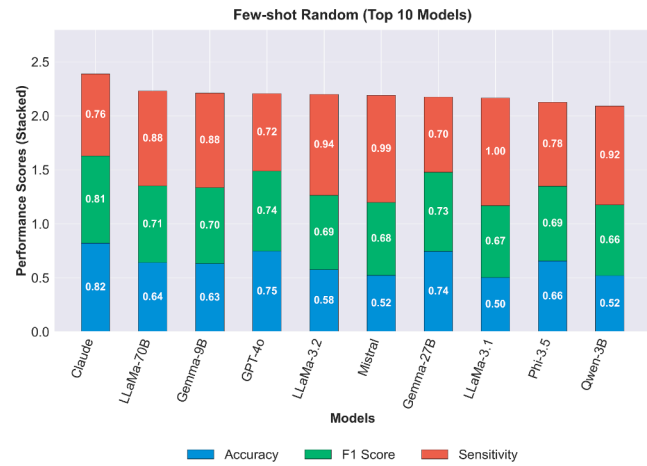

Fig. 6. Random few-shot performance of the top 10 LLMs on DDI prediction: Validation dataset.

Table 5
Zero-shot performance of reasoning-enhanced LLMs on DDI prediction.

# Params	Model	Acc.	Prec.	Sens.	F1
>250B	OpenAI O3	**0.7009**	0.8789	0.4661	0.6091
	Anthropic Opus-4	0.6881	**0.8839**	0.4330	0.5813
4B	Phi-4 4B reasoning	0.4853	0.4646	0.1927	0.2724
	Qwen3 4B thinking	0.5569	0.8780	0.1321	0.2297
14B	Nemotron 14B	0.5138	0.9412	0.0294	0.0569

Gemini 1.5 Pro) generally demonstrated superior performance, with Claude 3.5 Sonnet achieving the highest 'accuracy' (0.7358) and 'precision' (0.8859) among all models; nonetheless, even these state-of-the-art LLMs achieved relatively modest 'sensitivity' scores (0.5413-0.5927), which indicates that they have limitations in identifying positive interactions.

The results also showed variance in DDI prediction performance across different open-weight models, although model size did not consistently correlate with performance. On the one hand, smaller models (2-3B parameters) exhibited high sensitivity (e.g., Llama-3.2 3B: 0.9982), despite having lower overall accuracy, which suggests a bias toward positive predictions rather than true discriminative ability. On the other hand, mid-range LLMs, such as Gemma2 27B (accuracy: 0.7303) and Mistral-Nemo 12B (accuracy: 0.6991), performed similarly to proprietary models, whereas larger models like Qwen 2.5 72B (accuracy: 0.7119) did not significantly outperform their smaller counterparts. The findings indicate that although LLMs can use their pre-trained knowledge for DDI prediction, their zero-shot performance is limited, especially in terms of sensitivity.

To find out if sophisticated reasoning capabilities could overcome the drawbacks of conventional LLMs, we ran additional experiments using more recent reasoning-enhanced models. To this aim, we selected five additional LLMs, such as OpenAI O3, Anthropic Opus-4, Qwen3 4B thinking, Phi-4 4B reasoning, and NVIDIA Nemotron 14B.

Table 5 reports the results of these experiments. With an accuracy of 0.701 and an F1-score of 0.609, O3 performed the best, but it still fell short of the best-performing conventional LLMs, such as Claude 3.5 Sonnet. Notably, O3 showed a high precision (0.879) but a low sensitivity (0.466), which may suggest a conservative approach to positive predictions. Anthropic Opus-4 exhibited similar patterns with slightly lower performance (accuracy: 0.688, Sensitivity: 0.433, F1: 0.581), while the smaller reasoning models showed concerning limitations. Phi-4 4B reasoning, Qwen3 4B thinking, and Nemotron 14B all demonstrated poor sensitivity scores (0.193, 0.132, and 0.029, respectively), indicating severe difficulty in identifying positive drug interactions despite their advanced reasoning capabilities. These findings imply that reasoning-enhanced models do not address the issues of DDI prediction in zero-shot settings, even though they can offer structured analytical processes.

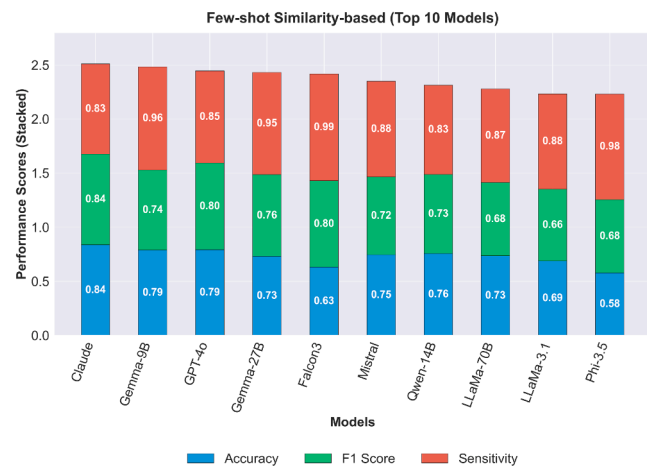

Fig. 7. Similarity-based few-shot performance of the top 10 LLMs on DDI prediction: Validation dataset.

All this leads us to answer RQ1:

> **RQ_1 Zero-shot Learning Analysis**: LLMs show limited effectiveness in zero-shot DDI prediction, with even the best models achieving only moderate accuracy and modest sensitivity. The variable performance across different model sizes suggests that pre-trained knowledge alone is insufficient for reliable DDI prediction.

6.2. RQ2: Few-shot learning analysis

Figs. 6 and 7 (see Tables B.19 and B.20 for details) report the results of random and similarity-based few-shot strategies. Both approaches demonstrated some effectiveness, as they improved LLM performance compared to the zero-shot setting (see Section 6.1), although there is still room for improvement.

In general, the similarity-based strategy outperformed the random one across most models, confirming our hypothesis that providing contextually relevant examples enhances the ability of LLMs to make accurate predictions. This improvement was particularly evident in proprietary and several mid-sized models, with GPT-4o showing substantial

Table 6
Optimal LoRA parameters for fine-tuned models obtained through hyper-parameter search.

Model	Epochs	Layers	LR	Rank	Alpha	Scale	Dropout
Phi-3.5 2.7B	3	16	2e-4	16	16	4.0	0.0
Qwen 2.5 3B	4	16	2e-4	16	16	4.0	0.0
DeepSeek R1 1.5B	4	28	2e-4	32	32	4.0	0.01
Gemma2 9B	5	16	1e-5	16	16	4.0	0.1

gains (F1: 0.7403→0.8014) and Claude 3.5 Sonnet maintaining its leadership (F1: 0.8384).

Distinctive performance patterns emerged across model families. Llama models (3.1 8B, 3.2 3B, and 3.3 70B) consistently exhibited extremely high sensitivity (>97%) but lower precision, indicating a tendency to "overpredict´´ positive interactions. In contrast, Gemini 1.5 Pro-and Qwen models showed the opposite pattern, suggesting a more conservative approach to identifying DDIs.

As with the zero-shot approach, model size did not consistently predict the performance benefits of few-shot learning. Although proprietary models performed better overall, a few mid-sized models – Falcon 3 10B (F1: 0.7975), Aya-Expanse 32B (F1: 0.7625), Gemma2 27B (F1: 0.7556), and Qwen 2.5 72B (F1: 0.7629) – performed exceptionally well with similarity-based selection. Surprisingly, the largest open models did not consistently outperform their smaller counterparts. Llama-3.3 70B achieved only a 0.6789 F1 score with similarity-based selection, underperforming compared to several 8-14B models.

All this suggests that for highly specialized tasks such as identifying DDIs, it is more important to have a model specifically optimized for that domain than a model that is simply larger but generic.

These results lead us to answer RQ2 as follows:

RQ_2 Few-shot Learning Analysis: LLMs show improvement with few-shot learning for DDI prediction compared to the zero-shot setting, with similarity-based selection outperforming random selection. Despite this approach narrowing the gap between open-weight and proprietary models, even the best performers, like Claude 3.5 Sonnet (F1 = 0.8384), achieve results that remain sub-optimal for reliable pharmaceutical applications, indicating DDI prediction is challenging even with in-context examples.

6.3. RQ3: Impact of fine-tuning

Table 6 reports the best parameters found for each fine-tuned model along with their corresponding training epochs. The computational details regarding the fine-tuning processes are reported in the Supplementary material (S.R6.3.1).

Fig. 8 (see Table B.21 for details) presents the performance of LLMs after fine-tuning.

Fine-tuning significantly improved the performance of all models on the validation set. Significant gains were seen in smaller models, with Qwen 2.5 3B demonstrating strong performance (accuracy: 0.878, sensitivity: 0.969) and Phi-3.5 2.7B performing on par with GPT-4o (accuracy: 0.913 vs. 0.926) and even outperforming it in sensitivity (0.960 vs. 0.930). It is interesting to note that DeepSeek R1 Distilled Qwen 1.5B (i.e., the smallest model in our study) produced competitive results (accuracy: 0.895, sensitivity: 0.919) after fine-tuning, indicating that model size has no bearing on DDI prediction performance.

Based on the results from Sections 6.1 and 6.2, we selected Claude Sonnet 3.5 and GPT-4o as the best-performing models from the few-shot analysis for comparison with fine-tuned models on independent data. The comparison in terms of F1-scores on the KEGG and HIV external datasets is reported in Table 7.

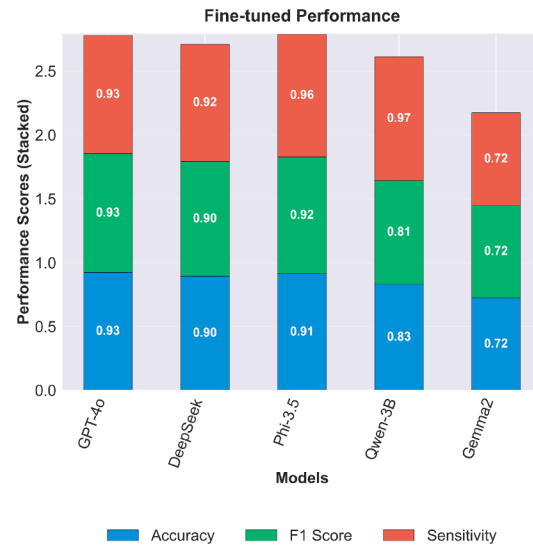

Fig. 8. Fine-tuning results of different LLMs on DDI prediction: Validation dataset.

Table 7
Performance comparison on independent external datasets (KEGG and HIV) of the fine-tuned models against the best two few-shot LLMs (Claude Sonnet 3.5 and GPT-4o).

Model	KEGG			HIV		
	Sens.	Acc.	F1	Sens.	Acc.	F1
Claude Sonnet 3.5 (similarity few-shot)	0.833	0.792	0.800	0.810	0.795	0.798
GPT-4o (similarity few-shot)	0.718	0.756	0.746	0.789	0.775	0.778
DeepSeek R1 1.5B (fine-tuned)	0.956	0.910	0.914	0.971	0.915	0.919
Phi-3.5 2.7B (fine-tuned)	**0.990**	0.922	0.927	**1.000**	**0.927**	0.932
Qwen2.5 3B (fine-tuned)	0.985	0.895	0.903	**1.000**	0.898	0.907
Gemma2 9B (fine-tuned)	0.714	0.824	0.803	0.765	0.846	0.833
GPT-4o (fine-tuned)	0.962	**0.937**	**0.939**	0.982	0.948	**0.950**

The results confirm the advantage of fine-tuning over few-shot learning. On the KEGG dataset, fine-tuned models achieved higher performance across all metrics: sensitivity (0.714-0.990 vs. 0.718-0.833), accuracy (0.824-0.937 vs. 0.756-0.792), and F1-score (0.803-0.939 vs. 0.746-0.800). Similarly, on the HIV dataset, fine-tuned models outperformed few-shot approaches across all metrics: sensitivity (0.879-0.990 vs. 0.789-0.819), accuracy (0.898-0.927 vs. 0.775-0.795), and F1-score (0.833-0.950 vs. 0.778-0.798). Notably, even the smallest fine-tuned model (DeepSeek R1 Distilled Qwen 1.5B) substantially outperformed both few-shot models on these independent datasets (accuracy: 0.910-0.971, sensitivity: 0.915-0.956, F1: 0.914-0.919).

Given the results and the above analysis, we can answer RQ3 as follows:

RQ_3 Impact of Fine-tuning: Fine-tuning dramatically improves LLMs' DDI prediction capabilities, with even small models achieving performance comparable to or exceeding larger models. All this suggests that model size is less critical than task-specific adaptation for effective DDI prediction.

6.4. RQ4: Comparative analysis

Several noteworthy findings emerge after comparing replicated machine learning models to LLM-based approaches across external datasets. Although slightly worse than but comparable to the results in the original paper (accuracy: 0.948, sensitivity: 0.948), the replication of the L2-regularized logistic regression model performed competitively

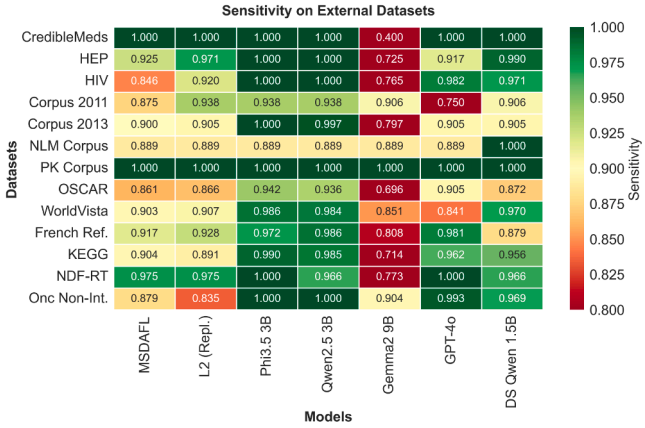

Fig. 9. Sensitivity comparison across external datasets.

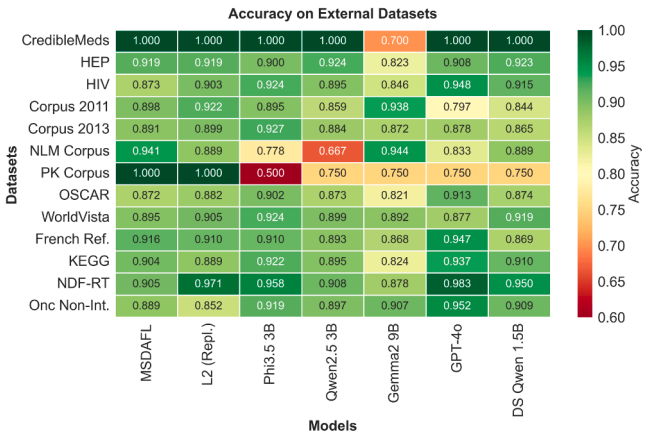

Fig. 10. Accuracy comparison across external datasets.

Table 8
Sensitivity robustness under label noise. Models maintain sensitivity > 0.85 on external datasets even at 10% noise, far exceeding the empirically observed 0.01% mislabeling rate.

Model	0% Noise	1% Noise	5% Noise	10% Noise
GPT-4o	0.933	0.928 (-0.005)	0.904 (-0.029)	0.877 (-0.056)
Phi-3.5 2.7B	0.978	0.966 (-0.013)	0.950 (-0.028)	0.920 (-0.058)
DeepSeek 1.5B	0.953	0.974 (+0.021)	0.950 (-0.002)	0.920 (-0.032)
Qwen 2.5 3B	0.975	0.970 (-0.005)	0.949 (-0.026)	0.922 (-0.053)
Gemma 2 9B	0.787	0.783 (-0.004)	0.763 (-0.024)	0.737 (-0.050)

Green : improvement; Yellow : <1% loss; Orange : 1–3% loss; Red : <6% loss.

(accuracy: 0.918, sensitivity: 0.925, F1-score: 0.919). Similarly, the MSDAFL model implementation showed good results (accuracy: 0.916, sensitivity: 0.913, F1-score: 0.918), but it fell short of the performance ranges reported in the original paper (accuracy: 0.953-0.987, F1-score: 0.900-0.993).

When examining the fine-tuned LLM performance, Phi-3.5 2.7B consistently demonstrated superior sensitivity (0.978 average) across the 13 external datasets, outperforming traditional machine-learning approaches and other LLMs (see Fig. 9 and Table B.22 for details). This model achieved perfect sensitivity (1.000) in 7 of the 13 datasets, including CredibleMeds, HEP, HIV, PK Corpus, and NDF-RT. Qwen 2.5 3B followed closely with an average sensitivity of 0.975, while DeepSeek Qwen 1.5B achieved 0.953, surpassing traditional machine learning models.

Interestingly, the traditional machine learning approaches showed slightly better overall performance when examining accuracy and F1-

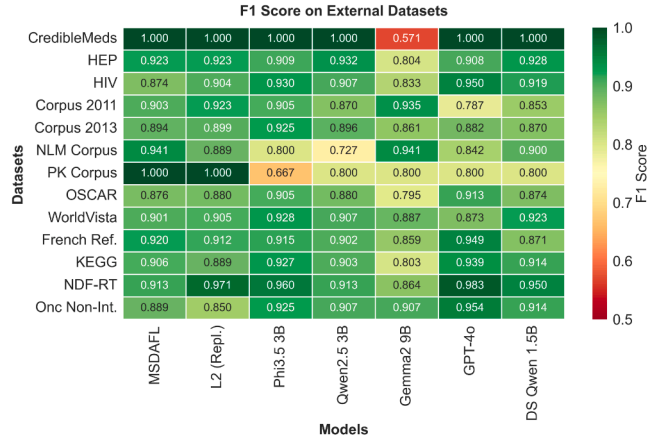

Fig. 11. F1 comparison across external datasets.

score (for details, see Figs. 10 and 11, and Tables B.23 and B.24). The L2-regularized model had the highest F1-score (0.919) and average accuracy (0.918), with MSDAFL coming in second (accuracy: 0.916, F1-score: 0.918). With the best average accuracy (0.902) and F1-score (0.906) among LLMs, GPT-4o excelled on the HIV, OSCAR, French DDI Referrals, KEGG, NDF-RT, and ONC Non-Interruptive datasets. These results suggest that, although fine-tuned LLMs are excellent at detecting true positive interactions (high sensitivity), they may generate more false positives than conventional machine learning techniques. Even so, Phi-3.5 2.7B's remarkable sensitivity (0.978) is a big plus in DDI prediction scenarios where it is usually riskier to miss a possible interaction (false negative) than to mistakenly flag a non-interaction (false positive).

Given the results, we can answer RQ4 as follows:

RQ_4 Comparative Analysis: Fine-tuned LLMs demonstrate performance comparable to the L2-regularized logistic regression and MSDAFL models on the external datasets, while not requiring complex feature engineering, structured chemical preprocessing, or large-scale training data. Notably, smaller models like Phi-3.5 2.7B and Qwen 2.5 3B achieve superior sensitivity scores while maintaining competitive accuracy and F1-scores.

6.5. RQ5: Dataset construction strategy validation

Prospective validation. The prospective validation revealed that of the 22,992 negative pairs, only 2 (0.01%) were reclassified as positive interactions (i.e., [Pentostatin, Iptacopan] and [Zafirlukast, Rifampin]) in the new DrugBank dataset (version 5.1.14), providing evidence that our negative sampling strategy introduced minimal label noise. Importantly, both newly confirmed interactions were correctly predicted as positive by our fine-tuned models (Phi-3.5 (2.7B), Qwen 2.5 3B, DeepSeek R1 Distilled Qwen 1.5B, and GPT-4o), despite being labeled as negatives during the datasets processing, excluding Gemma 2 9B that correctly predicted [Zafirlukast, Rifampin] but incorrectly classified [Pentostatin, Iptacopan]. This result suggests that the fine-tuned models learned robust interaction patterns that generalize beyond the training labels, further validating the approach.

Sensitivity analysis to label noise. The sensitivity analysis of the fine-tuned LLMs to label noise followed the same direction.

As we can note in Table 8, models' performance degraded gracefully across all noise levels. On external datasets at 10% label corruption–representing a conservative upper bound far exceeding our empirically observed 0.01% mislabeling rate–Phi-3.5 maintained sensitivity of

Table 9
Metrics on 1:52 imbalanced test set. Sensitivity is reported for positives (Pos) and negatives (Neg).

Model	Accuracy	Weighted F1	Sensitivity (Pos)	Sensitivity (Neg)
Gemma2 9B	**0.96**	**0.97**	0.74	**0.96**
GPT-4o	**0.96**	**0.97**	0.97	**0.96**
Deepseek 1.5B	0.92	0.94	0.95	0.92
Phi-3.5 2.7B	0.94	0.95	**0.99**	0.94
Qwen2.5 3B	0.86	0.91	**0.99**	0.85
L2-regularized (baseline)	0.92	0.95	0.90	0.92
MSDAFL (baseline)	0.91	0.92	0.91	0.91

0.920 (5.8% reduction), accuracy of 0.854 (2.7% reduction), and F1 of 0.878 (2.2% reduction). GPT-4o showed comparable robustness (sensitivity: 0.877, accuracy: 0.873, F1: 0.882). Notably, sensitivity remained above 0.85 for all top-performing models even under 10% corruption, suggesting our findings are robust to moderate label noise. Interestingly, the smallest model (DeepSeek R1 1.5B) exhibited improved performance at 1–5% noise levels on validation data (sensitivity: +3.2%, F1: +1.5% at 1%), suggesting that minor label perturbations may provide implicit regularization benefits for compact architectures. This effect diminished at 10% noise, where performance degraded as expected. These results additionally indicate that our dataset construction strategy operates well within the robustness margins of the models, with actual label noise being two orders of magnitude lower than the tested 10% threshold.

Evaluation on Naturally Imbalanced Dataset. The robust discriminative capabilities (see Table 9) of the fine-tuned LLMs were further supported by the models' evaluation under unbalanced class distribution.

Gemma2 9B and GPT-4o both achieved the highest weighted F1 score (0.97) and accuracy (0.96), outperforming traditional baselines (L2-regularized: 0.95, MSDAFL: 0.92). Smaller models also demonstrated excellent performance, with Phi-3.5 2.7B obtaining a weighted F1 of 0.95, DeepSeek 1.5B reaching 0.94, and Qwen2.5 3B attaining 0.91. Most fine-tuned LLMs had high sensitivity for positive interactions despite extreme class imbalance, with Phi-3.5 2.7B and Qwen2.5 3B attaining a sensitivity of 0.99, in contrast to the baselines' sensitivity of 0.90-0.91. The consistency of superior or comparable performance across both balanced and imbalanced test sets corroborates that the results observed in our analyses reflect genuine pattern recognition capabilities rather than artifacts of dataset construction.

Drug entity overlap and memorization effects. Finally, results from the entity-overlap analysis revealed no specific memorization patterns. On the validation set (1,090 examples: 47.0% no overlap, 34.8% one entity, 18.3% two entities), mean F1-scores were similar across groups and no pairwise differences reached significance (minimum adjusted $p = 0.065$ across all models) after Holm correction. The effect sizes (median $|\text{Cohen's } h| < 0.3$ for significant proportion tests; typical $|\Delta F1| < 0.10$) did not consistently favor higher entity overlap, and only isolated comparisons were significant across external datasets. In small sample size datasets, such as CredibleMeds, there were no detectable differences, while in DDI 2011, a few isolated cases showed significance (e.g., DeepSeek R1 Distilled Qwen 1.5B: $\Delta F1 \approx -0.188$, 95% CI $[-0.360, -0.062]$, Holm-adjusted $p = 0.003$), but directions were mixed and not monotonically aligned with greater overlap. There was no discernible benefit for entity familiarity in the performance on the other larger external datasets (e.g., DDI 2013, KEGG, HIV), which showed similar results across overlap groups. These results provide no evidence that entity overlap affected performance and disprove memorization as the underlying mechanism.

Additional results and details can be found in Supplementary Sections S.R6.4.4 (overlap analysis), S.R6.4.5 (prospective validation), S.R6.4.6 (label noise), and S.R6.4.7 (imbalanced datasets). The above-mentioned results lead us to answer RQ5 as follows:

RQ_5 Dataset Construction Strategy Validation: The dataset construction strategy is valid, introducing minimal label noise (0.01%). Fine-tuned LLMs maintain robust performance under label corruption (sensitivity >0.85 at 10% noise) and extreme class imbalance (weighted F1 = 0.97 at 1:52 ratio), with no evidence of entity memorization. Models learned generalizable pharmacological patterns rather than dataset artifacts.

Table 10
Stability results for fine-tuned LLMs across external datasets.

Metric	Models				
	Phi3.5 3b	Qwen 2.5 3B	Gemma2 9B	GPT-4o +250B	DeepSeek Qwen 1.5B
Accuracy	0.881	0.863	0.851	**0.902**	0.894
SD inter-runs	0.000	0.00023	0.00002	0.000	0.000
% agreement	100.00%	99.94%	99.99%	100%	100%
CI	–	[0.863–0.863]	[0.851–0.851]	–	–
Sensitivity	**0.978**	0.975	0.787	0.933	0.953
SD inter-runs	0.000	0.0004	0.00005	0.000	0.000
% agreement	100.00%	99.94%	99.99%	100%	100%
CI	–	[0.974–0.976]	[0.787–0.787]	–	–
F1-score	0.900	0.879	0.835	**0.906**	0.901
SD inter-runs	0.000	0.00022	0.00003	0.000	0.000
% agreement	100.00%	99.94%	99.99%	100%	100%
CI	–	[0.879–0.879]	[0.835–0.835]	–	–

6.6. RQ6: Stability and perturbation robustness assessment

The results of the two stability evaluations were convergent. On the one hand, agreement across all five runs was 100% for GPT-4o, Phi-3.5 2.7B, and DeepSeek R1 Distilled Qwen 1.5B with zero standard deviation across all metrics. Qwen 2.5 3B and Gemma2 9B achieved 99.94% and 99.99% agreement, respectively, confirming the high stability of fine-tuned models' responses (see Table 10).

Narrow 95% confidence intervals were derived from the minimal variance observed: Qwen 2.5 3B had CIs of [0.863–0.863] for accuracy, [0.879–0.879] for F1-score, and [0.974–0.976] for sensitivity; Gemma2 9B had CIs of [0.851–0.851] for accuracy, [0.835–0.835] for F1-score, and [0.787–0.787] for sensitivity. The near-zero width of these intervals (except for Gemma2's sensitivity) reflects the extremely high consistency across runs. Specifically, in a single iteration out of five, Qwen 2.5 3B showed different classifications for four drug pairs (one from each of the datasets DDI Corpus 2013, HEP, KEGG, and OSCAR), while Gemma2 9B misclassified three pairs (one from each of the three datasets French DDI Referrals, HEP, and OSCAR).

On the other hand, agreement across the ten computational configurations (as described in Section 5.7) was 100% for all models with zero standard deviation (see Tables 11 and 12). The 100% agreement reflects the deterministic nature of greedy decoding (temperature = 0) combined with fixed seeds, which eliminates sampling stochasticity-this level of stability is both expected and necessary for clinical deployment.

Regarding prompt perturbation robustness, the results (see Table 13) showed that the best-performing models from RQ4 (see Section 6.4), namely GPT-4o and Phi-3.5 2.7B, remained highly stable. Specifically, GPT-4o achieved accuracy ranging from 0.929 to 0.934 and F1 from 0.933 to 0.935, while Phi-3.5 2.7B achieved accuracy from 0.909 to 0.920 and F1 from 0.916 to 0.926. Averages and standard deviations were small, and the performance pattern reported in the paper held consistently: GPT-4o preserved the highest accuracy and F1 on KEGG, while Phi-3.5 maintained very high sensitivity (average 0.993) with competitive F1.

Table 11
Fine-tuning results of different LLMs on DDI prediction using five different random seeds (42, 123, 456, 789, 1024) at a fixed batch size of 15.

Model	Acc.	Prec.	Sens.	F1
GPT-4o	**0.926**	0.922	0.930	**0.926**
DeepSeek R1 Distilled Qwen1.5B	0.895	0.877	0.919	0.898
Phi-3.5 2.7B	0.913	0.878	0.960	0.917
Qwen 2.5 3B	0.878	0.820	**0.969**	0.888
Gemma2 9B	0.832	**0.923**	0.725	0.812

Note: Mean values over 5 independent runs. All models achieved perfect stability (SD = 0, 100% agreement across runs). 95% CI collapses to point estimates due to zero variance.

Table 12
Fine-tuning results of different LLMs on DDI prediction using five different batch sizes (1, 4, 16, 32, 64) at a fixed seed (123).

Model	Acc.	Prec.	Sens.	F1
GPT-4o	**0.926**	0.922	0.930	**0.926**
DeepSeek R1 Distilled Qwen1.5B	0.895	0.877	0.919	0.898
Phi-3.5 2.7B	0.913	0.878	0.960	0.917
Qwen 2.5 3B	0.878	0.820	**0.969**	0.888
Gemma2 9B	0.832	**0.923**	0.725	0.812

Note: Mean values over 5 independent runs. All models achieved perfect stability (SD = 0, 100% agreement across runs). 95% CI collapses to point estimates due to zero variance.

Table 13
Prompt perturbation robustness on KEGG for fine-tuned GPT-4o and Phi-3.5 2.7B. Each perturbation variant is evaluated on the KEGG test set.

Model	Perturbation	Accuracy	Sensitivity	F1
GPT-4o	P1	0.9340	0.9548	0.9354
	P2	0.9296	0.9787	0.9329
	P3	0.9315	0.9508	0.9328
	P4	0.9323	0.9609	0.9342
	AVG	**0.9320**	**0.9610**	**0.9340**
	SD	**0.0020**	**0.0110**	**0.0010**
Phi-3.5 2.7B	P1	0.9204	0.9922	0.9258
	P2	0.9204	0.9928	0.9257
	P3	0.9184	0.9911	0.9239
	P4	0.9085	0.9950	0.9158
	AVG	**0.9170**	**0.9930**	**0.9230**
	SD	**0.0050**	**0.0010**	**0.0040**

Additional results are provided in Supplementary sections S.R6.3.3 (validation set) and S.R6.4.3 (external datasets) for hit perturbations, and S.R6.4.2 for ablation studies.

Given the results, we can answer RQ6 as follows:

RQ_6 Stability and Perturbation Robustness Assessment: Fine-tuned LLMs demonstrate high stability (≥ 99.9% agreement across runs, 100% across computational configurations) and robustness to prompt perturbations (± 1% metric variation). These results confirm stable and reliable predictions suitable for clinical deployment.

6.7. RQ7: Error and learned patterns characterization analysis

Error Patterns Characterization. The qualitative error analysis across external datasets in the top-performing fine-tuned models, Phi-3.5 2.7B and GPT-4o, identified no systematic patterns.

Phi-3.5 2.7B showed a strong ability to identify true DDIs but tended to overestimate interactions. False positives were associated with higher molecular complexity (i.e., SMILES string length) in 50% of datasets (e.g., KEGG: 125.1±104 SMILES characters for false positives vs. 114.2±71 for false negatives) and with a higher number of target genes in 60% of cases (e.g., HIV: 8.9±17.0 no. of target genes for false positives vs. 4.9±9.6 for false negatives). Some drugs were systematically misclassified as interacting in certain datasets (e.g., capecitabine, insulin degludec, and fexofenadine in the KEGG dataset), although the pattern is not generalizable.

GPT-4o exhibited a more conservative behavior, but with the opposite bias, i.e., difficulty identifying true interactions. False negatives were linked to higher SMILES complexity in 60% of datasets (e.g., KEGG: 197.7±227.8 for false negatives vs. 124.7±106.9 for false positives), while no consistent patterns emerged for the number of target genes. Only four drugs showed recurring errors (i.e., metformin, thiamine, pseudoephedrine, eslicarbazepine acetate) in more than one dataset.

Despite opposite biases (overestimation vs. underestimation), the models shared 18 problematic drugs (e.g., methazolamide, pseudoephedrine) that were misclassified by both LLMs.

This aspect, which is independent of the model architecture, may suggest that certain drugs have features (e.g., structural or pharmacological) that current LLMs cannot capture, likely tied to complex mechanisms of action or atypical physicochemical properties that are difficult to encode linguistically.

Learned Patterns Characterization. The confidence analysis for GPT-4o fine-tuned revealed interesting findings.

True positive (TP) predictions consistently exhibited significantly higher confidence (mean: 0.986, SD: 0.06) than false positive (FP) predictions (mean: 0.904, SD: 0.17) across external datasets (Mann-Whitney U tests, $p < 0.001$ for all comparisons, mean effect size $r = +0.50$, large effect).

For example, on the validation set, TP confidence (0.98) significantly exceeded FP confidence (0.87; $p < 0.001$, $r = +0.32$), with consistent patterns observed across HIV ($\Delta = 0.10$), KEGG ($\Delta = 0.09$), and French DDI ($\Delta = 0.08$) datasets.

This calibration gap shows that GPT-4o assigns high confidence selectively to genuine interactions rather than indiscriminately to all positive predictions, indicating learned discrimination rather than systematic positive bias.

We provided the detailed results in Supplementary Sections S.R6.4.1 (error analysis) and S.R6.4.8 (confidence analysis).

The results reported and discussed above lead us to answer RQ7 as follows:

RQ_7 Error and Learned Patterns Characterization Analysis: Fine-tuned LLMs show opposite error patterns: Phi-3.5 2.7B overestimates interactions (false positives linked to molecular complexity), while GPT-4o underestimates them (false negatives associated with SMILES complexity). Both share 18 problematic drugs, suggesting intrinsic encoding limitations. GPT-4o demonstrates learned discrimination with significantly higher confidence for true positives (0.986) versus false positives (0.904, $p < 0.001$), indicating authentic DDI pattern recognition.

7. Implications, limitations and future work

This section discusses the practical implications of the findings and outlines limitations and future research directions.

7.1. Practical implications

The findings have several implications for both the deployment of DDI prediction systems and their clinical applications. First of all, the

excellent performance of smaller models (2–3 billion parameters) represents an important breakthrough in accessibility, because these models can run efficiently in various healthcare settings without requiring a heavy computational load or specialized infrastructure. The ability to deploy them locally also addresses important privacy concerns, since sensitive medical data can be processed on-premises rather than via cloud APIs, and their lightweight nature makes it easier to integrate them into existing healthcare information systems. Second, the high sensitivity achieved by the models in identifying DDIs across multiple datasets – an ability that is particularly valuable in cases of polypharmacy where multiple drugs are administered at the same time – makes them suitable as screening tools in clinical practice. Third, the models' ability to process multiple types of drug information simultaneously could offer opportunities to uncover emerging patterns and hidden correlations that are unexplored to date.

Furthermore, they could support clinical decision-making at various levels:

- Pre-prescription screening to identify potential interactions before medication is prescribed
- Support for pharmacists in medication review processes
- Aid in clinical research for identifying potential drug combinations for investigation
- Real-time decision support in emergency medicine where rapid assessment of drug interactions is crucial

The combination of all these aspects implies that, given their computational efficiency, smaller LLMs could be used as practical tools in everyday clinical workflows to improve patient safety.

7.2. Future directions

Our study demonstrated the potential of LLMs for DDI prediction, but it also outlined several research directions that warrant further investigation. The strong performance of smaller models suggests exploring even more efficient architectures and biomedical LLMs designed explicitly for processing molecular and biological data; for example, by developing specialized pre-training strategies that incorporate domain-specific knowledge about DDIs and biological pathways. A second interesting direction is including additional drug-related information into the LLM prompt, which could enable more precise predictions regarding the timing and severity of potential interactions. Indeed, although our approach combines several types of drug characteristics, future work could explore other information, such as pharmacokinetic properties, metabolic pathways, and the temporal aspects of drug administration. Additionally, future work could explore more explicit representations of interaction directionality, such as incorporating directional markers or comparing predictions for reversed drug pairs (A,B) versus (B,A) to validate the LLMs' understanding of asymmetric interactions and strengthen the modeling of directional drug-drug relationships. A third possible direction to investigate is the interpretability of model predictions, which would help build trust in these tools and provide valuable suggestions to healthcare practitioners. Future research could develop approaches to explain why some drug combinations were flagged as dangerous, for example, by exploring the attention weights or generating importance scores to help delineate the relative importance of each model input in prediction.

Finally, to improve the usefulness of LLMs in managing novel medications and rare interactions, future studies could investigate methods for continuously updating models' knowledge base as soon as new information on drug interactions becomes available.

7.3. Limitations

Our study faced a few potential limitations, which we systematically addressed through our methodology. Regarding experimental design, the selection of fine-tuning parameters could have influenced model performance; we mitigated this through comprehensive hyperparameter optimization using Optuna. The potential bias from random training/validation splits was addressed by validating the findings across 13 external datasets. Following the methodology in [13], we used balanced datasets to prevent learning bias. Another limitation relates to the negative sampling strategy we used, i.e., random generation of negative drug pairs, which carries the theoretical risk of mislabeling future true positives as negatives. To mitigate this issue, we generated negatives exclusively from drug pairs absent in our extensive interaction database (over 1.4 million known interactions from 14 sources) and constrained selection to drugs targeting at least one gene (according to established DDI prediction principles [13,31]). To empirically assess this risk, we conducted three complementary analyses: prospective validation, label noise evaluation, and prediction confidence analysis (see Section 6.4). The analyses combined with consistent performance across 13 heterogeneous external datasets and high sensitivity scores (0.978 average for Phi-3.5), indicate that our negative sampling strategy introduces negligible label noise. Models maintain robust performance even under substantial artificial label corruption (sensitivity >0.85 at 10% corruption), and show no evidence of memorization artifacts in entity-overlap analysis. Any remaining undetected interactions would cause our metrics to underestimate true model capability–a conservative bias acceptable for safety-critical applications. Concerning generalizability, while the fine-tuning analysis included only five selected models (one proprietary and four open-weights), we ensured diversity by choosing models with different architectures and sizes (ranging from 1.5B to over 250B parameters). The robustness of our approach was further validated across 13 diverse external datasets representing various domains and complexity levels. To ensure the reliability of the results, we performed five repeated classifications for fine-tuned models (RQ3 and RQ4), which demonstrated consistent performance with no variability in the outcomes. Beyond methodological issues, LLM architectures present ethical and regulatory obstacles to real-world clinical implementation. LLMs' black-box nature makes clinical decision-making more difficult to interpret, and obtaining regulatory approval as medical decision support tools (such as FDA approval for medical devices) would necessitate thorough clinical validation studies and adherence to healthcare regulations. Given these factors, it would be more appropriate to use LLM-based systems for DDI prediction as screening tools under clinical supervision as opposed to independent decision-makers.

8. Conclusion

This study presents the first investigation of LLMs for DDI prediction, introducing a novel approach that leverages these models' ability to process multiple drug characteristics (i.e., SMILES, target organisms, and genes) simultaneously as text input through carefully designed prompts. While the findings across 13 external datasets demonstrates that LLMs show limited effectiveness in zero-shot settings and improvement but variable performance in few-shot settings, fine-tuned models achieved remarkable performance, with smaller models like Phi-3.5 2.7B performing comparably to or better than larger models and state-of-the-art machine-learning approaches. All this suggests that model size is less critical than task-specific adaptation for effective performance and provides evidence that LLMs can be powerful tools for DDI prediction. These findings has several practical implications, making DDI prediction more accessible for clinical applications without requiring extensive computational resources. Future work could develop more efficient LLM architectures, incorporate additional drug-related information into the LLM's prompt, and improve model interpretability for clinical applications.

CRediT authorship contribution statement

Gabriele De Vito: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation,

Formal analysis, Data curation, Conceptualization; **Filomena Ferrucci:** Writing – review & editing, Validation, Methodology, Investigation, Formal analysis; **Athanasios Angelakis:** Writing – review & editing, Validation, Supervision.

Data availability

We have shared the link to our data and code in the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been partially supported by the Italian PNRR MUR project PE0000013-FAIR.

Appendix A. Experiments with full fine-tuning

Tables A.15 and A.14 report the best results (average and class) achieved training 50% of the trainable parameters and using different fine-tuning configurations: learning rate (1e-5, 5e-6, 2e-5) and training duration (1, 1.5, 2, 3 epochs).

Table A.14
Full fine-tuning results for Phi-3.5 2.7B with half trainable parameters.

Model	Acc.	Prec.	Sens.	F1
Phi-3.5 2.7B (16 layers)	0.871	0.847	0.905	0.875

Table A.15
Detailed classification metrics for Phi-3.5 2.7B full fine-tuning with half trainable parameters.

Class	Precision	Recall	F1-score
No Interaction (0)	0.90	0.84	0.87
Interaction (1)	0.85	0.90	0.87

Tables A.16 and A.17 report the best results (average and class) achieved training four layers and using different fine-tuning configurations: learning rate (1e-5, 5e-6, 2e-5) and training duration (1, 1.5, 2, 3 epochs).

Table A.16
Full fine-tuning results for Phi-3.5 2.7B with 4 trainable layers.

Model Configuration	Acc.	Prec.	Sens.	F1
Phi-3.5 2.7B (4 layers)	0.691	0.647	0.840	0.731

Table A.17
Detailed classification metrics for Phi-3.5 2.7B full fine-tuning with 4 trainable layers.

Class	Precision	Recall	F1-score
No Interaction (0)	0.77	0.54	0.64
Interaction (1)	0.65	0.84	0.73

Appendix B. Additional results

Table B.18
Zero-shot performance of different LLMs on DDI prediction.

# Params	Model	Acc.	Prec.	Sens.	F1
>250B	Claude 3.5 Sonnet	**0.7358**	0.8859	0.5413	0.6720
	Gemini 1.5 Pro-	0.7220	0.7995	0.5927	0.6807
	GPT-4o	0.6459	0.6710	0.5725	0.6178
< 2B	Deepseek Qwen1.5B	0.4807	0.4677	0.2789	0.3494
	Llama-3.2 3B	0.5009	0.5005	**0.9982**	0.6667
2-3B	Qwen 2.5 3B	0.5037	0.5019	0.9761	0.6629
	Phi-3.5 2.7B	0.5358	0.5221	0.8440	0.6452
	Granite 3.1 8B	0.4734	0.3535	0.0642	0.1087
8-9B	Llama-3.1 8B	0.5294	0.5656	0.2532	0.3498
	Gemma2 9B	0.6376	0.8178	0.3541	0.4942
	Falcon 3 10B	0.5853	0.8394	0.2110	0.3372
10-14B	Mistral-Nemo 12B	0.6991	0.7395	0.6147	0.6713
	Qwen 2.5 14B	0.6569	**0.9212**	0.3431	0.5000
	Gemma2 27B	0.7303	0.8476	0.5615	0.6755
27-32B	Aya-Expanse 32B	0.5917	0.5617	0.8349	0.6716
	Qwen 2.5 32B	0.5982	0.6754	0.3780	0.4847
70-72B	Llama-3.3 70B	0.6477	0.6123	0.8055	**0.6957**
	Qwen 2.5 72B	0.7119	0.7667	0.6092	0.6789

Table B.19
Random few-shot performance of different LLMs on DDI prediction.

# Params	Model	Acc.	Prec.	Sens.	F1
	Claude 3.5 Sonnet	**0.8211**	**0.8661**	0.7596	**0.8094**
>250B	Gemini 1.5 Pro-	0.7394	0.8595	0.5725	0.6872
	GPT-4o	0.7477	0.7626	0.7193	0.7403
<2B	Deepseek Qwen1.5B	0.5156	0.5209	0.3890	0.4454
	Llama-3.2 3B	0.5761	0.5443	0.9358	0.6883
2-3B	Qwen 2.5 3B	0.5202	0.5113	0.9156	0.6561
	Phi-3.5 2.7B	0.6560	0.6254	0.7780	0.6934
	Granite 3.1 8B	0.6248	0.6104	0.6899	0.6477
8-9B	Llama-3.1 8B	0.5028	0.5014	**0.9963**	0.6671
	Gemma2 9B	0.6321	0.5889	0.8752	0.7041
	Falcon 3 10B	0.7128	0.7900	0.5798	0.6688
10-14B	Mistral-Nemo 12B	0.5229	0.5118	0.9927	0.6754
	Qwen 2.5 14B	0.6807	0.8530	0.4367	0.5777
	Gemma2 27B	0.7450	0.7708	0.6972	0.7322
27-32B	Aya-Expanse 32B	0.6771	0.7468	0.5358	0.6239
	Qwen 2.5 32B	0.6963	0.8365	0.4881	0.6165
70-72B	Llama-3.3 70B	0.6413	0.5958	0.8789	0.7102
	Qwen 2.5 72B	0.7018	0.7209	0.6587	0.6884

Table B.20
Similarity-based few-shot performance of different LLMs on DDI prediction.

# Params	Model	Acc.	Prec.	Sens.	F1
	Claude 3.5 Sonnet	**0.8376**	0.8345	0.8422	**0.8384**
>250B	Gemini 1.5 Pro-	0.7257	**0.8534**	0.5450	0.6652
	GPT-4o	0.7917	0.7659	0.8404	0.8014
< 2B	Deepseek Qwen1.5B	0.5046	0.5052	0.4422	0.4716
	Llama-3.2 3B	0.5431	0.5231	0.9780	0.6816
2-3B	Qwen 2.5 3B	0.6394	0.6271	0.6881	0.6562
	Phi-3.5 2.7B	0.5789	0.5491	0.8826	0.6770
	Granite 3.1 8B	0.6138	0.5675	0.9560	0.7122
8-9B	Llama-3.1 8B	0.5009	0.5005	**0.9890**	0.6646
	Gemma2 9B	0.6881	0.6350	0.8844	0.7393
	Falcon 3 10B	0.7899	0.7696	0.8275	0.7975
10-14B	Mistral-Nemo 12B	0.6321	0.5811	0.9468	0.7202
	Qwen 2.5 14B	0.7459	0.7780	0.6881	0.7303
	Gemma2 27B	0.7578	0.7626	0.7486	0.7556
27-32B	Aya-Expanse 32B	0.7303	0.6811	0.8661	0.7625
	Qwen 2.5 32B	0.7046	0.8159	0.5284	0.6414
70-72B	Llama-3.3 70B	0.5330	0.5173	0.9872	0.6789
	Qwen 2.5 72B	0.7349	0.6899	0.8532	0.7629

Table B.21
Fine-tuning results of different LLMs on DDI prediction.

Model	Acc.	Prec.	Sens.	F1
GPT-4o	**0.926**	0.922	0.930	**0.926**
Deepseek Qwen1.5B	0.895	0.877	0.919	0.898
Phi-3.5 2.7B	0.913	0.878	0.960	0.917
Qwen 2.5 3B	0.878	0.820	**0.969**	0.888
Gemma2 9B	0.832	**0.923**	0.725	0.812

Table B.22
Sensitivity comparison across external datasets.

Dataset	Sensitivity						
	Msdafl (Repl.)	L2 (Repl.)	Phi3.5 3b	Qwen 2.5 3B	Gemma2 9B	GPT-4o +250B	Deepseek Qwen 1.5B
CredibleMeds	**1.000**	**1.000**	**1.000**	**1.000**	**0.400**	**1.000**	**1.000**
HEP	0.925	0.971	**1.000**	**1.000**	0.725	0.917	0.990
HIV	0.846	0.920	**1.000**	**1.000**	0.765	0.982	0.971
Corpus 2011	0.875	**0.938**	**0.938**	**0.938**	0.906	0.750	0.906
Corpus 2013	0.900	0.905	**1.000**	0.997	0.797	0.905	0.905
NLM Corpus	0.889	0.889	0.889	0.889	0.889	0.889	**1.000**
PK Corpus	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
OSCAR	0.861	0.866	**0.942**	0.936	0.696	0.905	0.872
WorldVista	0.903	0.907	**0.986**	0.984	0.851	0.841	0.970
French Ref.	0.917	0.928	0.972	**0.986**	0.808	0.981	0.879
KEGG	0.904	0.891	**0.990**	0.985	0.714	0.962	0.956
NDF-RT	0.975	0.975	**1.000**	0.966	0.773	**1.000**	0.966
Onc Non-Int.	0.879	0.835	**1.000**	**1.000**	0.904	0.993	0.969
AVG	0.913	0.925	**0.978**	0.975	0.787	0.933	0.953
SD	–	–	0.035	0.035	0.146	0.075	0.046

Table B.23
Accuracy comparison across external datasets.

Dataset	Accuracy						
	Msdafl (Repl.)	L2 (Repl.)	Phi3.5 3b	Qwen 2.5 3B	Gemma2 9B	GPT-4o +250B	DeepSeek Qwen 1.5B
CredibleMeds	**1.000**	**1.000**	**1.000**	0.900	0.700	**1.000**	**1.000**
HEP	0.919	0.919	**0.924**	0.895	0.823	0.908	0.923
HIV	0.873	0.903	0.927	0.898	0.846	**0.948**	0.915
Corpus 2011	0.898	0.922	0.875	0.859	**0.938**	0.797	0.844
Corpus 2013	0.891	0.899	**0.919**	0.884	0.872	0.878	0.865
NLM Corpus	0.941	0.889	0.778	0.667	**0.944**	0.833	0.889
PK Corpus	**1.000**	**1.000**	0.500	0.750	0.750	0.750	0.750
OSCAR	0.872	0.882	0.902	0.873	0.821	**0.913**	0.874
WorldVista	0.895	0.905	**0.924**	0.899	0.892	0.877	0.919
French Ref.	0.916	0.910	0.910	0.893	0.868	**0.947**	0.869
KEGG	0.904	0.889	0.922	0.895	0.824	**0.937**	0.910
NDF-RT	0.905	0.971	0.958	0.908	0.878	**0.983**	0.950
Onc Non-Int.	0.889	0.852	0.919	0.897	0.907	**0.952**	0.909
AVG	0.916	**0.918**	0.881	0.863	0.851	0.902	0.894
SD	–	–	0.125	0.072	0.069	0.073	0.059

Appendix C. Supplementary material

Supplementary analyses and results are provided in our online repository [24], as follows:

- **S.R3.2.1 Drug Pairs distribution analysis**, analyzes structural and biological features on 43,894 unique drug-drug pairs (21,947 positive interactions and 21,947 negatives) compiled from 13 external datasets in order to validate the negative sample strategy and check how different negatives and positives are.
- **S.R6.2.1 Few-shot perturbation** assesses the sensitivity of LLMs in few-shot settings to prompt and examples perturbations.
- **S.R6.2.2 Few-shot ablation studies** reports the ablation studies for LLMs in few-shot settings, considering SMILES only, genes only, species only, or pairwise combinations in prompts.
- **S.R6.3.1 Fine-tuning computational info** reports the computational details of the fine-tuning process.
- **S.R6.3.2 Fine-tuning Phi3.5 on 2000 samples** empirical validatates the choice of 1000 samples for fine-tuning LLMs.
- **S.R6.3.3 Fine-tuned LLM hit perturbations - Validation set** assesses the robustness of fine-tuned LLMs against hit perturbations on the validation set.
- **S.R6.4.1 Error Analysis GPT-4o vs PHI3.5 2.7B** reports a qualitative error analysis across external datasets in the top-performing fine-tuned models, Phi-3.5 2.7B and GPT-4o.

Table B.24
F1 comparison across external datasets.

Dataset	F1-score						
	Msdafl (Repl.)	L2 (Repl.)	Phi3.5 3b	Qwen 2.5 3B	Gemma2 9B	GPT-4o +250B	DeepSeek Qwen 1.5B
CredibleMeds	**1.000**	**1.000**	**1.000**	0.909	0.571	**1.000**	**1.000**
HEP	0.923	0.923	**0.930**	0.905	0.804	0.908	0.928
HIV	0.874	0.904	0.932	0.907	0.833	**0.950**	0.919
Corpus 2011	0.903	0.923	0.882	0.870	**0.935**	0.787	0.853
Corpus 2013	0.894	0.899	**0.925**	0.896	0.861	0.882	0.870
NLM Corpus	**0.941**	0.889	0.800	0.727	**0.941**	0.842	0.900
PK Corpus	**1.000**	**1.000**	0.667	0.800	0.800	0.800	0.800
OSCAR	0.876	0.880	0.905	0.880	0.795	**0.913**	0.874
WorldVista	0.901	0.905	**0.928**	0.907	0.887	0.873	0.923
French Ref.	0.920	0.912	0.915	0.902	0.859	**0.949**	0.871
KEGG	0.906	0.889	0.927	0.903	0.803	**0.939**	0.914
NDF-RT	0.913	0.971	0.960	0.913	0.864	**0.983**	0.950
Onc Non-Int.	0.889	0.850	0.925	0.907	0.907	**0.954**	0.914
AVG	0.918	**0.919**	0.900	0.879	0.835	0.906	0.901
SD	–	–	0.083	0.055	0.094	0.067	0.049

- **S.R6.4.2 Fine-tuned LLMs Ablation studies** conducts an ablation study for the fine-tuned version of Phi-3.5 2.7B using SMILES-only inputs, and on the fine-tuned version of GPT-4o using organisms-only inputs.
- **S.R6.4.3 Fine-tuned LLMs hit perturbations - External Datasets** assesses the robustness of fine-tuned LLMs against hit perturbations on the external datasets.
- **S.R6.4.4 Fine-tuned LLMs Overlap analysis** provides the results of the drug-entity overlap analysis between training and evaluation sets.
- **S.R6.4.5 Fine-tuned LLMs Prospective Analysis** validates the negatives sample strategy analyzing prospectively, verifying whether any negative sample had been subsequently confirmed as interacting using an updated version of DrugBank (version 5.1.14 February 2025), released one year after our initial data collection.
- **S.R6.4.6 Fine-tuned LLMs Label Noise** reports a sensitivity analysis of the fine-tuned models by introducing 1%, 5%, and 10% random label flips for negative samples in the external datasets.
- **S.R6.4.7 Fine-tuned LLMs Imbalanced Datasets results** provides the detailed results for the evaluation of the fine-tuned LLMs' performance on naturally imbalanced dataset.
- **S.R6.4.8 Fine-tuned GPT-4o Confidence Analysis** analyzes the prediction confidence of the fine-tuned version of GPT-4o across all the validation and external datasets, in terms of log probabilities of the responses' tokens.

References

[1] X. Lin, L. Dai, Y. Zhou, Z.-G. Yu, W. Zhang, J.-Y. Shi, D.-S. Cao, L. Zeng, H. Chen, B. Song, et al., Comprehensive evaluation of deep and graph learning on drug–drug interactions prediction, Brief. Bioinf. 24 (4) (2023) bbad235.

[2] R.E. David, K.G. Ferrara, J. Schrecker, D. Paculdo, S. Johnson, R. Bentley-Lewis, R. Heltsley, J.W. Peabody, Impact of medication nonadherence and drug-drug interaction testing on the management of primary care patients with polypharmacy: a randomized controlled trial, BMC Med. 22 (1) (2024) 540.

[3] M. Delara, L. Murray, B. Jafari, A. Bahji, Z. Goodarzi, J. Kirkham, M. Chowdhury, D.P. Seitz, Prevalence and factors associated with polypharmacy: a systematic review and meta-analysis, BMC Geriatr. 22 (1) (2022) 601.

[4] H. Lee, Y.-H. Baek, J.H. Kim, T.-C. Liao, W.C.Y. Lau, K.K.C. Man, X. Qin, S. Wood, J. Ilomäki, J.S. Bell, et al., Trends of polypharmacy among older people in asia, australia and the united kingdom: a multinational population-based study, Age Ageing 52 (2) (2023) afad014.

[5] S. Mukhopadhyay, P. Zhang, K.K.W. To, Y. Liu, C. Bai, S.S.Y. Leung, Sequential treatment effects on phageantibiotic synergistic application against multi-drug-resistant Acinetobacter baumannii, Int. J. Antimicrob. Agents. 62 (5) (2023) 106951.

[6] C.H. Chung, S. Chandrasekaran, A flux-based machine learning model to simulate the impact of pathogen metabolic heterogeneity on drug interactions, PNAS nexus 1 (3) (2022) pgac132.

[7] B. Yang, H. Liang, J. Xu, Y. Liu, S. Ma, Y. Li, C. Wang, Multi-drug sequential release systems: construction and application for synergistic tumor treatment, Int. J. Pharm. 670 (2025) 125156.

[8] D. Ramsden, E.S. Perloff, A. Whitcher-Johnstone, T. Ho, R. Patel, K.D. Kozminski, C.L. Fullenwider, J.G. Zhang, Predictive in vitro-in vivo extrapolation for time dependent inhibition of CYP1A2, CYP2C8, CYP2C9, CYP2C19, and CYP2D6 using pooled human hepatocytes, human liver microsomes, and a simple mechanistic static model, Drug Metab. Dispos. 50 (2) (2022) 114–127.

[9] L. van Kempen, M.C. de Vries, E.J.P. de Koning, N. de Graeff, Who should be included in first-in-human trials? a systematic review of reasons, J Transl. Med. 23 (1) (2025) 649.

[10] Y. Zhao, J. Yin, L. Zhang, Y. Zhang, X. Chen, Drug–drug interaction prediction: databases, web servers and computational models, Brief. Bioinf. 25 (1) (2024) bbad445.

[11] R. Ferdousi, R. Safdari, Y. Omidi, Computational prediction of drug-drug interactions based on drugs functional similarities, J. Biomed. Inform. 70 (2017) 54–64.

[12] A. Kastrin, P. Ferk, B. Leskošek, Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning, PLoS ONE 13 (5) (2018) e0196865.

[13] S. Mei, K. Zhang, A machine learning framework for predicting drug–drug interactions, Sci. Rep. 11 (1) (2021) 17619.

[14] H. Yu, S. Zhao, J. Shi, Stnn-ddi: a substructure-aware tensor neural network to predict drug–drug interactions, Brief. Bioinf. 23 (4) (2022) bbac209.

[15] E. Kasneci, K. Sessler, et al., ChatGPT for good? on opportunities and challenges of large language models for education, Learn Individ. Differ. 103 (2023) 102274.

[16] W.X. Zhao, K. Zhou, J. Li, et al., A survey of large language models, (2023). arXiv:2303.18223

[17] G. De Vito, Assessing healthcare software built using IoT and LLM technologies, in: Proceedings of the 28Th International Conference on Evaluation and Assessment in Software Engineering, 2024, pp. 476–481.

[18] A.J. Thirunavukarasu, D.S.J. Ting, K. Elangovan, L. Gutierrez, T.F. Tan, D.S.W. Ting, Large language models in medicine, Nat. Med. 29 (8) (2023) 1930–1940.

[19] G. De Vito, F. Ferrucci, A. Angelakis, HELIOT: LLM-Based CDSS for adverse drug reaction management, Knowl. Based. Syst. (2025) 114184.

[20] K.T. Ahmed, M.I. Ansari, W. Zhang, DTI-LM: language model powered drugtarget interaction prediction, Bioin. 40 (9) (2024) btae533.

[21] D. Oniani, J. Hilsman, C. Zang, J. Wang, L. Cai, J. Zawala, Y. Wang, Emerging opportunities of using large language models for translation between drug molecules and indications, Sci. Rep. 14 (1) (2024) 10738.

[22] C. Xu, O. Elemento, K.C. Bulusu, H. Pan, DDI-GPT: explainable prediction of drug-drug interactions using large language models enhanced with knowledge graphs, (2024)

[23] C. Hou, G. Duan, C. Yan, MSDAFL: molecular substructure-based dual attention feature learning framework for predicting drugdrug interactions, Bioin. 40 (10) (2024) btae596.

[24] G. De Vito, A. Angelakis, LLM DDI GitHub repository, (2025), https://github.com/gadevito/LLMDDI.

[25] S. Vilar, R. Harpaz, E. Uriarte, L. Santana, R. Rabadan, C. Friedman, Drug-drug interaction through molecular structure similarity analysis, J. Am. Med. Inform. Assoc. 19 (6) (2012) 1066–1074. https://doi.org/10.1136/amiajnl-2012-000935

[26] N. Rohani, C. Eslahchi, A. Katanforoush, ISCMF: integrated similarity-constrained matrix factorization for drugdrug interaction prediction, Network Model. Anal. Health Inform. Bioinform. 9 (2020) 1–8.

[27] G. Shtar, L. Rokach, B. Shapira, Detecting drug-drug interactions using artificial neural networks and classic graph similarity measures, PLoS ONE 14 (8) (2019) e0219796.

[28] X. Chen, X. Liu, J. Wu, GCN-BMP: investigating graph representation learning for DDI prediction task, Meth. 179 (2020) 47–54.

[29] Q. Yin, R. Fan, X. Cao, Q. Liu, R. Jiang, W. Zeng, Deepdrug: a general graph-based deep learning framework for drug-drug interactions and drug-target interactions prediction, Quant. Biol. 11 (3) (2023) 260–274.

[30] Y. Wang, Y. Min, X. Chen, J. Wu, Multi-view graph contrastive representation learning for drug-drug interaction prediction, in: Proceedings of the Web Conference 2021, 2021, pp. 2921–2933.

[31] A.K. Nyamabo, H. Yu, J.-Y. Shi, SSIDDI: substructuresubstructure interactions for drugdrug interaction prediction, Brief. Bioinf. 22 (6) (2021) bbab133.

[32] Z. Li, S. Zhu, B. Shao, X. Zeng, T. Wang, T.-Y. Liu, DSN-DDI: an accurate and generalized framework for drugdrug interaction prediction by dual-view representation learning, Brief. Bioinf. 24 (1) (2023) bbac597.

[33] C. Zhao, S. Liu, F. Huang, S. Liu, W. Zhang, CSGNN: contrastive self-supervised graph neural network for molecular interaction prediction, in: Ijcai, 2021, pp. 3756–3763.

[34] S.K. Mohamed, A. Nounu, V. Nováček, Biological applications of knowledge graph embedding models, Brief. Bioinf. 22 (2) (2021) 1679–1693.

[35] X. Su, B. Zhao, G. Li, J. Zhang, P. Hu, Z. You, L. Hu, Knowledge graph neural network with spatial-Aware capsule for drug-Drug interaction prediction, IEEE J. Biomed Health Inform. 29 (3) (2025) 1771–1781. https://doi.org/10.1109/JBHI.2024.3419015

[36] T. Sounack, J. Davis, B. Durieux, A. Chaffin, T.J. Pollard, E. Lehman, A.E.W. Johnson, M. McDermott, T. Naumann, C. Lindvall, Bioclinical ModernBERT: a state-of-the-art long-context encoder for biomedical and clinical NLP, (2025). arXiv:2506.10896

[37] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, S. Xu, F. Jamil, C. Hughes, C. Lau, et al., Medgemma technical report, (2025). arXiv:2507.05201

[38] Q. Xie, Q. Chen, A. Chen, C. Peng, Y. Hu, F. Lin, X. Peng, J. Huang, J. Zhang, V. Keloth, et al., Me-llama: foundation large language models for medical applications, Res. square (2024) rs–3.

[39] P. Ankit, OpenBioLLMs: advancing OpenSource large language models for healthcare and life sciences. Hugging Face repository, 2025, accessed on 14 February 2025, https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B.

[40] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, et al., Drugbank 5.0: a major update to the drugbank database for 2018, Nucleic Acids Res. 46 (D1) (2018) D1074–D1082.

[41] S. Ayvaz, J. Horn, O. Hassanzadeh, Q. Zhu, J. Stan, N.P. Tatonetti, S. Vilar, M. Brochhausen, M. Samwald, M. Rastegar-Mojarad, et al., Toward a complete dataset of drug–drug interaction information from publicly available sources, J. Biomed. Inform. 55 (2015) 206–217.

[42] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Transp. Knowl Data Eng. 21 (9) (2009) 1263–1284.

[43] M. Buda, A. Maki, M.A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, Neural Netw. 106 (2018) 249–259.

[44] P. Good, Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, Springer Science & Business Media, 2013.

[45] M.D. Ernst, Permutation methods: a basis for exact inference, Statist. Sci. (2004) 676–685.

[46] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, et al., Lima: less is more for alignment, Adv Neural Inf Proc. Syst. 36 (2023) 55006–55021.

[47] M. Oliver, G. Wang, Crafting efficient fine-tuning strategies for large language models, (2024). arXiv:2407.13906

[48] X. Zhao, X. Leng, L. Wang, N. Wang, Research on fine-tuning optimization strategies for large language models in tabular data processing, Biomim. 9 (11) (2024) 708.

[49] S. Barnett, Z. Brannelly, S. Kurniawan, S. Wong, Fine-tuning or fine-failing? debunking performance myths in large language models, (2024). arXiv:2406.11201

[50] OpenAI, GPT-4 technical report,(2023). arXiv:2303.08774

[51] Anthropic, Claude 3.5 Sonnet, 2024, [Online; accessed 15 November 2024], https://claude.ai/.

[52] Google, Gemini 1.5 Pro, 2024, [Online; accessed 15 November 2024], https://ai.google.dev/.

[53] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., The llama 3 herd of models, (2024). arXiv:2407.21783

[54] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, et al., Qwen2. 5 technical report, (2024). arXiv:2412.15115

[55] M. Mishra, M. Stallone, G. Zhang, Y. Shen, A. Prasad, A.M. Soria, M. Merler, P. Selvam, S. Surendran, S. Singh, et al., Granite code models: a family of open foundation models for code intelligence, (2024). arXiv:2405.04324

[56] G. Team, M. Riviere, S. Pathak, P.G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al., Gemma 2: improving open language models at a practical size, (2024). arXiv:2408.00118

[57] F.-L. Team, The falcon 3 family of open models, 2024, (2024). https://huggingface.co/blog/falcon3.

[58] M. A. Team, Mistral nemo, 2024, [Online; accessed 15 November 2024]. https://mistral.ai/en/news/mistral-nemo.

[59] J. Dang, S. Singh, D. D'souza, A. Ahmadian, A. Salamanca, M. Smith, A. Peppin, S. Hong, M. Govindassamy, T. Zhao, et al., Aya expanse: combining research breakthroughs for a new multilingual frontier, (2024). arXiv:2412.04261

[60] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A.A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, et al., Phi-3 technical report: a highly capable language model locally on your phone, (2024). arXiv:2404.14219

[61] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al., Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning, (2025). arXiv:2501.12948

[62] I. Element Labs, LM Studio website, (2024), [Online; accessed 15 November 2024], https://lmstudio.ai/.

[63] OpenAI, GPT models fine-tuning, (2024). [Online; accessed 15 November 2024], https://platform.openai.com/docs/guides/fine-tuning.

[64] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: efficient finetuning of quantized llms, Adv Neural Inf Proc. Syst. 36 (2023) 10088–10115.

[65] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: low-rank adaptation of large language models, ICLR 1 (2) (2022) 3.

[66] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: a next-generation hyperparameter optimization framework, in: Proceedings of the 25Th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2623–2631.

[67] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, (2014). arXiv:1412.6980

[68] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[69] K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks?, (2018). arXiv:1810.00826

[70] J.M. Johnson, T.M. Khoshgoftaar, Survey on deep learning with class imbalance, J. Big Data 6 (1) (2019) 1–54.