

MUDI: A Multimodal Biomedical Dataset for Understanding Pharmacodynamic Drug-Drug Interactions

Tung-Lam Ngo*

22028092@vnu.edu.vn

VNU University of Engineering and
Technology (VNU-UET)

Ba-Hoang Tran*

21020631@vnu.edu.vn

VNU University of Engineering and
Technology (VNU-UET)

Duy-Cat Can*[†]

duy-cat.can@chuv.ch

Lausanne University Hospital (CHUV)
University of Lausanne (UNIL)

Trung-Hieu Do

dotrunglehieu05220161@daihocyhanoi.edu.vn

Hanoi Medical University
National Geriatric Hospital, Hanoi

Oliver Y. Chén

olivery.chen@chuv.ch

Lausanne University Hospital (CHUV)
and University of Lausanne (UNIL)

Hoang-Quynh Le[‡]

lhquynh@vnu.edu.vn

VNU University of Engineering and
Technology (VNU-UET)

Abstract

Understanding the interaction between different drugs (drug-drug interaction or DDI) is critical for ensuring patient safety and optimizing therapeutic outcomes. Existing DDI datasets primarily focus on textual information, overlooking multimodal data that reflect complex drug mechanisms. In this paper, we (1) introduce MUDI, a large-scale Multimodal biomedical dataset for Understanding pharmacodynamic Drug-drug Interactions, and (2) benchmark learning methods to study it. In brief, MUDI provides a comprehensive multimodal representation of drugs by combining pharmacological text, chemical formulas, molecular structure graphs, and images across 310,532 annotated drug pairs labeled as Synergism, Antagonism, or New Effect. Crucially, to effectively evaluate machine-learning based generalization, MUDI consists of unseen drug pairs in the test set. We evaluate benchmark models using both late fusion voting and intermediate fusion strategies. All data, annotations, evaluation scripts, and baselines are released under an open research license.

CCS Concepts

• **Information systems** → **Multimedia databases**; **Information extraction**; • **Computing methodologies** → **Machine learning approaches**; *Supervised learning by classification*; *Neural networks*; • **Applied computing** → **Life and medical sciences**; *Bioinformatics*; *Health informatics*.

Keywords

Multimodal Dataset, Drug-Drug Interaction, Pharmacodynamics, Biomedical Data Mining, Multimodal Fusion

1 Introduction

Polypharmacy, the concurrent use of multiple medications, is common in treating complex diseases or comorbid conditions, especially in elderly patients. It can lead to drug-drug interactions (DDIs), where one drug alters the effect of another, potentially reducing efficacy or causing adverse events [19]. A promising approach is to

predict DDIs proactively when assigning drugs, improving patient safety and treatment outcomes while reducing healthcare costs. To do so effectively requires two key components: a suitable predictive method and a well-targeted dataset for training and evaluation.

Designing such a dataset must reflect the inherently multimodal nature of DDIs, which arise from diverse pharmacological foundations, including chemical properties, pharmacological description, and molecular structure [1]. Yet most existing datasets focus narrowly on one of these modalities, typically textual data [24, 34, 36], limiting models' ability to capture complex biochemical interactions. Some recent multimodal DDI studies address this, but often rely on fragmented or non-standardized sources, with limited modality coverage or label inconsistency. For example, the state-of-the-art datasets, such as DDInter [31], simply merge pharmacokinetic and pharmacodynamic interactions into a single label set and ignore interaction directionality. These lines of evidence suggest a need for a comprehensive, well-curated multimodal dataset.

To address these gaps, we introduce **MUDI** – a **M**ultimodal **B**iomedical **D**ataset for **U**nderstanding **P**armacodynamic **D**rug-**D**rug **I**nteractions. MUDI is a large-scale, richly annotated collection of paired drugs, integrating multiple data modalities: drug descriptions, molecular structure graphs, molecular structure images, and chemical formulas. Unlike existing datasets, MUDI provides directed labels (e.g., Synergism, Antagonism) and undirected labels (e.g., New Effect). Notably, MUDI's test set contains a substantial portion of interactions involving unseen drugs to assess model generalization. Moreover, we provide multimodal baselines with intermediate and late fusion to benchmark learning from heterogeneous biomedical inputs. Finally, we release the full MUDI dataset, annotation guidelines, baseline implementations, and benchmarking results publicly under an open license.

We summarize our key contributions as follows:

- We construct and release **MUDI**, a multimodal pharmacodynamic DDI dataset, curated based on clinical knowledge, containing 310,532 annotated drug pairs.
- We provide benchmark results using **multimodal baselines** for predicting DDIs from heterogeneous biomedical inputs.
- We publicly release the dataset, annotations, code, and evaluation pipelines to support reproducible multimodal research.

*Shared first authors.

[†]Also with VNU University of Engineering and Technology (VNU-UET).

[‡]Corresponding authors.

2 Related Work

DDI Prediction Datasets. Numerous datasets have been developed for DDI prediction, varying in scope and annotation strategies. DrugBank [30] is the most comprehensive and widely used. However, its DDI information is presented as free-text descriptions rather than structured labels or relations. The associated drug metadata is also not systematically organized into distinct modalities, limiting its applicability for multimodal machine learning. Several DDI-focused datasets have been constructed based on DrugBank and supplementary sources such as FAERS [27], MEDLINE [28], and other biomedical databases. These datasets are typically single-modal, providing DDI information using either text or SMILES strings. Among them, some describe interactions without assigning labels (e.g., LIDDI [3], TDC [14], BioSNAP [37]), while others provide binary or inconsistently defined label sets without a systematic scheme, and often omit relation directionality (e.g., SemEval 2013 [4], HODDI [29], TWOSIDES [26], Mendeley DDI dataset [32]).

The DDInter dataset [31] represents a recent state-of-the-art with emphasis on multimodality. However, it merges pharmacokinetic and pharmacodynamic interactions into a single label set, ignoring their distinct roles and causal links. Its assumption of symmetric interactions overlooks real-world directionality, where one drug may affect another without reciprocal effects. Finally, molecular graphs are not provided, missing key spatial and topological cues critical for modeling molecular structure.

DDI Multimodal Approaches. Recently, there has been a growing trend towards multimodal approaches for drug-drug interaction prediction. Most studies aim to incorporate, expand, and diversify drug-related multimodal information from various data sources to improve model performance. However, the richness and availability of these modalities remain limited. Some studies extract additional aspects from textual descriptions, such as targets and pathways [6, 9], yet this remains multi-aspect features derived from a single modality. Others incorporate knowledge graphs [2, 25], producing text-graph hybrids rather than fully multimodal models. Notably, some researchers have utilized SMILES strings from DrugBank to capture molecular structural information [1, 20] or chemical substructures [6]. More advanced methods, such as 3DGT-DDI [12], leverage 3D molecular graph structures obtained from PubChem to better represent molecular geometry. Nevertheless, current multimodal DDI research still relies on fragmented sources, narrow modality coverage, and non-standardized preprocessing.

How MUDI Complements Existing Resources. MUDI serves as a standardized and reproducible resource for multimodal DDI prediction, addressing key limitations of prior datasets (see Table 4). MUDI is specifically curated for pharmacodynamic interactions, with biologically grounded labels and interaction directionality. Unlike prior fragmented datasets, MUDI integrates six structured modalities, supporting standardized evaluation and downstream pharmacological integration. Finally, MUDI emphasizes **accessibility and reproducibility**, releasing the full dataset, annotation guidelines, and baseline implementations under an open license to foster transparent, accelerated research in multimodal biomedical AI.

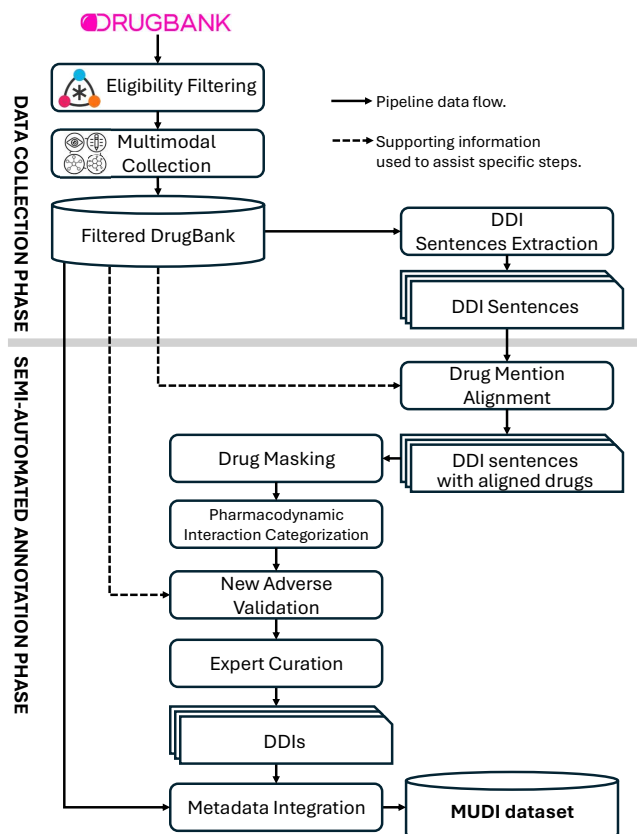


Figure 1: Overview of MUDI dataset construction pipeline.

3 Dataset Construction

MUDI is constructed from a pharmacodynamic perspective on drug-drug interactions, targeting clinically meaningful effects resulting from drug co-administration. Each interaction is categorized into one of three abstract-level pharmacodynamic labels:

- **Synergism** (directed relationship): The co-administration of two drugs results in an enhanced effect of one of the drugs. This enhancement may involve therapeutic efficacy, but it can also manifest as increased toxicity or adverse effects.
- **Antagonism** (directed relationship): The concurrent use of two drugs reduces or neutralizes the effect of one of the drugs. This reduction may pertain to therapeutic benefit, but may also involve diminished toxicity or side effects.
- **New Effect** (undirected relationship): The combination of two drugs leads to a new effect – either adverse or therapeutic – that is not associated with either drug when used individually.

These labels are based on established pharmacological theory [8, 23]. Drug pairs that do not fall into any of the above categories are considered to exhibit **no or unclear interaction**, indicating a lack of evidence or insufficient characterization of their pharmacodynamic relationship.

Figure 1 provides an overview of the MUDI dataset construction pipeline, illustrating both the data collection and semi-automated annotation phases along with their respective steps.

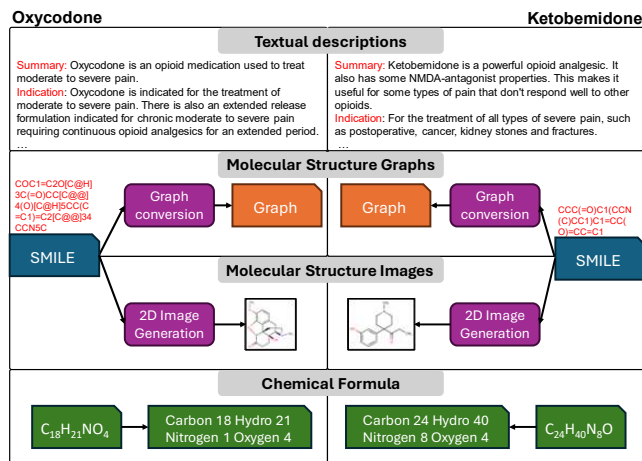


Figure 2: Multimodal examples.

3.1 Data Collection Phase

Eligibility Filtering: We begin with a comprehensive list of drugs from DrugBank (version 5.1.12) [30]. This list is refined using Hetionet version 1.0 [13], which includes only drugs approved for human use and excludes experimental, toxic, or veterinary compounds. In addition, Hetionet prioritizes drugs with clear therapeutic indications and biomedical connectivity, improving clinical relevance and downstream integration.

Multimodal Collection: For each eligible drug, we extract four data modalities from DrugBank. Drugs missing any modality are excluded. Four types of modalities are extracted or generated:

- **Textual modality:** Includes drug name and descriptions (summary, indications, pharmacodynamics, mechanism of action, and metabolism). Note that the drug-drug interaction fields in the metadata were excluded from the textual modality.
- **Molecular structure graph:** SMILES (simplified molecular input line entry system) strings are processed with RDKit [17] into graphs where atoms are nodes and bonds are edges.
- **Molecular structure image:** The processed SMILES were also used to generate standardized 2D chemical structure images rendered at a fixed resolution.
- **Chemical formula:** Formulas are converted into normalized sequences with explicit atomic counts (e.g., 'C₁₀H₁₃N₅O₄' becomes 'Carbon 10 Hydrogen 13 Nitrogen 5 Oxygen 4').

Figure 2 illustrates an example of the multimodal representations for two drugs, *Oxycodone* and *Ketobemidone*. The result of this step is referred to as **Filtered DrugBank** and is retained for a later metadata integration step.

DDI Sentences Extraction: As mentioned above, the drug-drug interaction fields present in the DrugBank metadata are excluded from the textual modality and handled separately. In this step, we extract and separate this field into individual sentences that explicitly describe the interaction relationships between drug pairs. These extracted sentences serve as the basis for downstream annotation.

3.2 Semi-automated Annotation Phase

It is called the semi-automated annotation phase, as the process is primarily automated but involves human validation: linguistic annotators corrected masking errors, and medical experts curated the final labels.

Drug Mention Alignment: The input to this step is a DDI sentence describing an interaction between a known pair of drugs. Since DrugBank does not provide explicit tags linking drug names to their textual mentions within the sentence, we perform automatic alignment to identify the specific text spans referring to the two interacting drugs. To achieve this, we apply flexible matching against all known synonyms of each drug.

Drug Masking: Once drug mentions are aligned, their corresponding text spans are replaced with placeholders [DRUG1] and [DRUG2]. This masking abstracts away lexical differences across sentences and facilitates more consistent downstream classification. As a result, the original set of 244,921 DDI descriptions was reduced to 287 distinct masked sentence templates. However, due to limitations in the automated drug mention alignment step, some sentences contained incorrect masking – such as partial drug names or missed detections when names overlapped (e.g., Promazine and Acepromazine). These sentences were manually reviewed and corrected by three linguistics students, resulting in 241 finalized sentence types. This simple manual curation achieved near-perfect inter-annotator agreement. See Appendix A.1 for masking details.

Pharmacodynamic Interaction Categorization: To assign pharmacodynamic labels, each masked sentence was automatically categorized into one of three predefined interaction types using a set of lexical rule-based heuristics. These heuristics were derived from recurring linguistic templates and semantic patterns observed in the DrugBank descriptions. For example, expressions indicating increased concentration or efficacy were mapped to Synergism, while patterns suggesting reduced activity were assigned to Antagonism. Sentences that could not be confidently categorized were provisionally labeled as New Effect. Full details of the heuristic rules and example templates are provided in Appendix A.2.

New Effect Validation: This step validates whether a New Effect truly refers to an adverse or therapeutic effect that is not present in either drug profiles. We compare biomedical effect terms extracted from the DDI description sentences with the metadata of both drugs using stemming and synonym expansion techniques [18, 22]. If the effect term is absent from both drug profiles, the sentence is retained as a New Effect; otherwise, it is reclassified as Synergism.

Expert Curation: To ensure biomedical accuracy, two physicians were involved in this expert curation phase. They review all 241 annotations and make corrections as needed based on the actual context and drug profiles. In cases of disagreement, the two experts discussed to reach a final consensus.

Metadata Integration: For each drug pair, we combine all collected modalities of two drugs, along with their interaction label, to form a complete data instance. The resulting dataset is then split into training and test sets, with careful consideration to ensure that the test set contains novel examples for robust evaluation.

Table 1: Distribution of pharmacodynamic interaction labels in MUDI.

Label	Direction type	Train (#)	Test (#)	Total (#)
Synergism	Directed (uni)	94,128	38,421	132,549
	Directed (bi)	92,140	32,580	124,720
Antagonism	Directed (uni)	27,320	11,914	39,234
New Effect	Undirected	7,527	6,502	14,029
Total		221,115	89,417	310,532

Table 2: Generalization coverage of test drug pairs based on training set exposure.

Drug Pair Type	Count	Proportion
Both drugs seen in training	4,099	4.58 %
One drug seen in training	78,822	88.15 %
Neither drug seen in training	6,496	7.27 %

3.3 Data Statistics and Analysis

The MUDI dataset contains 1,295 unique drugs annotated with four modality types: structured text, SMILES sequences, molecular structure images, and graphs. From these drugs, we generate 310,532 labeled drug-drug interaction instances in three pharmacodynamic classes: Synergism, Antagonism, and New Effect (see Section 3.2).

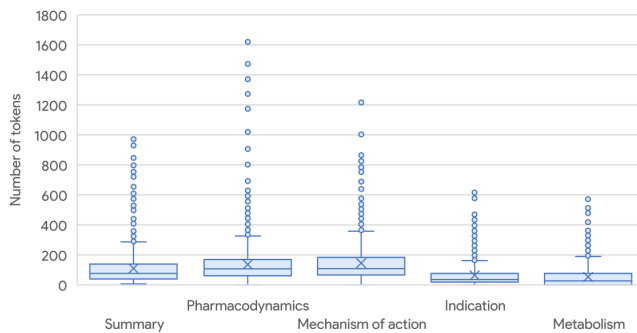
3.3.1 Label Distribution. Table 1 summarizes the label distribution across train and test sets. The majority (82.85%) of samples are Synergism, followed by Antagonism (12.63%) and New Effect (4.52%). It reflects real-world pharmacological distribution, where synergistic effects are more frequently reported, while new effect interactions remain rare [11]. This natural imbalance encourages the design of robust classification strategies.

In addition to label variety, MUDI introduces **direction-aware annotations** – a key novelty over prior datasets. While Antagonism is strictly uni-directional and New Effect is undirected, some Synergism instances are bi-directional, allowing nuanced modeling of asymmetric and reciprocal effects.

3.3.2 Drug Pair Coverage and Generalization. To assess the model’s capacity for generalization, we explicitly construct a test set containing drug combinations that are underrepresented or entirely unseen during training. Table 2 categorizes test drug pairs based on the presence of their constituent drugs in the training set.

The majority of test interactions (88.15%) involve drug pairs where only one drug is familiar to the model, creating a semi-unseen generalization scenario. Only 4.58% of test pairs contain two seen drugs, while 7.27% involve entirely unseen drugs, enabling rigorous evaluation of zero-shot prediction. This careful partitioning strategy enables the assessment of generalization capabilities, reflecting practical challenges in biomedical applications where new drug combinations are continuously emerging [33].

3.3.3 Modality-Level Data Profiling. Each drug entity in MUDI is represented through multiple modalities, each offering unique biological or structural information. We analyze their properties to provide insights for modality-specific and fusion-based models.

**Figure 3: Token length distributions for five pharmacological text fields in MUDI.****Table 3: Statistics of molecular graphs derived from SMILES representations.**

Statistic	Mean	Std. Dev.
Atoms per graph	27.06	15.00
Bonds per graph	27.79	16.25
Average node degree	2.10	0.22

Pharmacological Text Descriptions. Token length distributions across five fields are shown in Figure 3. *Pharmacodynamics*, *Summary*, and *Mechanism of Action* exhibit heavy-tailed lengths, often exceeding 1,000 tokens. In contrast, *Indication* and *Metabolism* are concise and consistent, often comprising under 100 tokens. These variations call for adaptable text encoding strategies that can accommodate both long-form and short-form biomedical content.

Molecular Graph Properties. Table 3 reports statistics for SMILES-derived graphs. The average molecular graph in MUDI contains approximately 27 atoms and 28 bonds, with a node degree of 2.1, reflecting typical chemical sparsity. However, the relatively high standard deviations suggest notable diversity in graph sizes and connectivity patterns across the dataset. These properties motivate the use of graph neural networks that are robust to varying graph sizes and capable of capturing both local connectivity and long-range interactions. Additionally, models must handle sparsity efficiently to avoid overfitting on smaller or simpler structures while preserving signals from more complex compounds.

Chemical Structure Images. Each drug’s SMILES string is also rendered into a 1000×800 resolution image. These standardized 2D representations support consistent training for vision-based models like CNNs and transformers.

3.3.4 Comparison with Existing DDI Resources. Table 4 compares MUDI with leading DDI datasets in terms of size, modality coverage, and semantic detail. MUDI is the only dataset that integrates **rich multimodal features** per drug and uses **directed pharmacodynamic interaction labels**. Unlike most existing datasets, which are limited to single-modality inputs or binary classification tasks, MUDI supports multimodal learning and fine-grained reasoning over clinically meaningful interaction categories.

This combination of scale, multimodality, and semantic depth positions MUDI as a unique benchmark for evaluating robust, multimodal approaches to pharmacodynamic DDI prediction.

Table 4: Comparison of our MUDI dataset with existing DDI-related datasets and resources.

Dataset	Scope	Data size	Drugs	Number of categories [§]	Multi-modal	Textual description	SMILES	Chemical formula	Molecular structure image	Molecular structure graph	Directed relations
DrugBank [30]	Drug information	1,420,072*	16,581 [†]	–	–	✓	✓	✓	(✓)	–	–
BioSNAP [37]	Drug interaction	48,514	1,514	–	–	–	–	–	–	–	–
LIDDI [3]	Side effect	103,774	345	–	–	–	–	–	–	–	–
TDC [14]	DDI description	191,808	1,706	–	–	–	✓	–	–	–	–
DDInter [31]	Mechanism and risk	236,834	1,833	10	✓	✓	✓	✓	(✓)	–	–
HODDI [29]	Side effect	109,744 [‡]	2,506	1	–	–	–	–	–	–	–
Mendeley DDI [32]	DDI description	222,696	1,868	20	–	–	–	–	–	–	–
DDI-2013 [4]	Semantic DDI	5,021	1,913	4	–	✓	–	–	–	–	–
TWOSIDES [26]	Side effect	63,473	645	1	–	–	✓	–	–	–	–
Our MUDI	Pharmacodynamics	310,532	1,295	3	✓	✓	✓	✓	✓	✓	✓

*: DDI-related sentence. [†]: Only 4,532 drugs are involved in drug-drug interactions. [‡]: Negative samples are counted in the total. [§]: Number of positive class labels; 1 implies binary classification.

(✓): Data exists but was not included in the provided download.

3.4 Licensing, Access, and Ethics

Open Release and Licensing. To support transparency and reproducibility, MUDI is released under the Creative Commons Attribution-NonCommercial 4.0 License (CC BY-NC 4.0), which allows sharing and adaptation with attribution, but prohibits commercial use. The dataset, preprocessing scripts, and baseline models are available on [Zenodo](#)¹ and [GitHub](#)², along with documentation and full download instructions (Appendix B).

Ethical Considerations. MUDI is built from publicly available biomedical sources (specifically, DrugBank [30]) and contains no human subject data or identifiable information, thus requiring no ethical review. We encourage responsible use aligned with licensing terms and recommend expert consultation for clinical applications. Responsible usage guidelines are provided in Appendix C.

4 Baseline Experiments

4.1 Baseline Models

To benchmark MUDI and support future research, we implement a suite of baseline models covering both single-modality and multi-modal fusion strategies. Each model takes a pair of drugs as input and predicts one of three interaction types. To train models effectively, we apply negative sampling by including 200,000 additional *No Interaction* pairs as negative examples. Implementation details and equations are provided in Appendix D.

Single-Modality Baselines. We train six modality-specific models: (i) name, (ii) description (merged from summary, pharmacodynamics, mechanism, metabolism, and indication), (iii) formula, (iv) SMILES (text), (v) molecular structure graph, and (vi) image. For **text-based** inputs (name, description, SMILES, formula), we fine-tune BioMedBERT [10], using the final [CLS] token as the drug embedding. **Graph** inputs are processed using a graph convolutional network [16] with max pooling. **Image** inputs are encoded using a Vision Transformer [7]. Each single-modality model concatenates two drug embeddings and feeds them to a shallow classifier.

Late Fusion Baseline. We construct a late fusion model that aggregates predictions from six single-modality classifiers. The final prediction is computed via majority voting across the six classifiers, with tie-breaking handled in a fixed modality priority order. This baseline reflects ensembling without joint feature modeling.

Intermediate Fusion Baseline. We construct an intermediate fusion baseline by concatenating the modality-specific embeddings into an interleaved sequence. The resulting fused vector is passed through a two-layer MLP classifier. This captures cross-modal interactions while preserving pairwise structure.

4.2 Experimental Setup

Evaluation Metrics. We report Precision, Recall, and F1 scores for each of the three positive classes, along with Micro and Macro averages. Negative examples (*No Interaction*) are included during training but excluded from positive-class metric computation, following common biomedical evaluation practice [21].

Evaluation Settings. We evaluate under two settings: (i) **direction-aware** matching, where (DRUG1, DRUG2) \neq (DRUG2, DRUG1), and (ii) **direction-agnostic** matching, where order is ignored. Detailed evaluation scripts and protocols are provided in Appendix E.

Implementation Environment and Hyper-parameters. All models are implemented in PyTorch 3.10 and trained on NVIDIA T4 GPUs. The training environment, model-specific hyper-parameters, and hardware details are fully documented in Appendix F.

4.3 Results

Table 5 presents the performance of our multimodal baselines under two evaluation settings: direction-aware and direction-agnostic matching. Across all metrics, the results confirm the inherent difficulty of MUDI, particularly in predicting less frequent labels such as *New Effect*. In the direction-aware setting, F1 scores are generally lower, with *New Effect* reaching only 27.84% under intermediate fusion and 15.84% under late fusion. Notably, the exclusion of directionality during evaluation substantially improves performance. For example, the micro-averaged F1 of intermediate fusion rises from 52.74% (direction-aware) to 66.69% (direction-agnostic). This underscores the challenge of modeling subtle interactions and label imbalance.

Among the fusion strategies, intermediate fusion consistently outperforms late fusion across both settings. Its superior performance (up to 69.76% F1 for Synergism and 66.69% overall micro-F1) demonstrates the advantage of integrating multimodal features before final decision layers. In contrast, late fusion appears less effective, likely due to the weak predictive capacity of individual unimodal branches in isolation.

¹<https://zenodo.org/records/15544551>

²<https://github.com/hoangbros03/MUDI>

Table 5: Multimodal baseline results on MUDI dataset.

		Precision	Recall	F1
Intermediate Fusion				
Direction-aware matching	Synergism	54.79	54.49	54.64
	Antagonism	55.39	47.88	51.36
	New Effect	61.68	17.97	27.84
	Micro-Averaged	55.04	50.61	52.74
	Macro-Averaged	57.29	40.11	47.18
Direction-agnostic matching	Synergism	78.98	62.47	69.76
	Antagonism	71.61	45.32	55.51
	New Effect	78.65	35.59	49.00
	Micro-Averaged	77.91	58.29	66.69
	Macro-Averaged	76.41	47.79	58.80
Late Fusion				
Direction-aware matching	Synergism	38.32	70.82	49.73
	Antagonism	48.12	36.12	41.26
	New Effect	62.79	9.06	15.84
	Micro-Averaged	39.25	60.32	47.56
	Macro-Averaged	49.74	38.67	43.51
Direction-agnostic matching	Synergism	62.14	76.44	68.55
	Antagonism	65.86	36.12	46.65
	New Effect	76.97	17.99	29.16
	Micro-Averaged	62.62	66.85	64.67
	Macro-Averaged	68.32	43.52	53.17

To assess the contribution of each modality, we report single-modality results in Appendix G.1, Table 6. Graph-based features perform best, achieving 65.44% micro-F1 under direction-agnostic matching. Textual fields (e.g., name and description) and images also show competitive results. However, SMILES and chemical formula underperform, indicating the need for more expressive encoders.

Together, these findings highlight (1) the need for robust multimodal integration to capture complementary biomedical signals, and (2) the utility of MUDI in benchmarking model generalization and cross-modal reasoning under realistic constraints.

5 Discussion

Potential Applications. The MUDI dataset supports several promising research directions within multimodal biomedical machine learning. First, it serves as a robust resource for developing models aimed at **early detection of pharmacodynamic interactions**, which can help researchers better understand complex drug behaviors. MUDI can also be leveraged in developing **DDI-aware recommendation systems** to assist researchers and pharmacists in preclinical compound screening and pharmaceutical development, as well as preliminary **interaction risk screening** for early-stage drug discovery. Additionally, MUDI enables the creation of **educational tools** for training healthcare professionals and pharmacologists on multimodal biomedical data interpretation. Its inclusion of unseen drug pairs further supports research into **generalizable prediction methods** for novel or investigational compounds, thereby promoting innovation in zero-shot biomedical learning. Nevertheless, given the dataset’s limitations in clinical validation and potential label noise, models trained on MUDI should not be directly applied to clinical or regulatory decisions (see Appendix C). Therefore, appropriate usage involves foundational research, educational training, or benchmarking within controlled experimental conditions, rather than immediate clinical implementation.

Insights. The baseline results provide valuable insights into multimodal modeling on MUDI. First, intermediate fusion, which integrates multimodal information earlier within the model, consistently outperforms late fusion strategies across all evaluation settings, underscoring the importance of early interactions between modalities. Second, performance significantly improves when directional constraints are relaxed during evaluation, particularly for the *Synergism* class, indicating that directional ambiguity is an important factor affecting model predictions. Lastly, the single-modality analysis (Appendix G.1) highlights the strong predictive capability of molecular graph representations relative to other modalities, suggesting the particular utility of structural information in predicting pharmacodynamic interactions.

Limitations. Despite careful curation, MUDI presents some limitations. A key constraint is the absence of **gold-standard validation** from clinical trials or wet-lab experiments. Interaction labels are derived from structured drug information and, although carefully annotated, may not fully reflect the complexity of **real-world clinical practice**. Additionally, reliance on textual descriptions introduces potential label noise, as **linguistic ambiguities or inconsistencies** in drug metadata can lead to occasional mislabeling [5].

Future Work. Several directions exist to further extend MUDI’s utility and scientific impact. Future work could expand the dataset with separate **pharmacokinetic** labels (e.g., absorption, metabolism) to support **multi-label** interaction modeling and disentangle the dynamics between pharmacokinetics and pharmacodynamics. Another direction is integrating **biomedical knowledge graphs** [35] to enrich drug representations, support structured reasoning, and enhance both performance and explainability.

6 Conclusion

In this paper, we (1) introduce MUDI, a large-scale, multimodal biomedical dataset for predicting pharmacodynamic interactions between drug-drug interactions (DDIs) and (2) benchmark competitive machine learning to predict DDI using MUDI. MUDI addresses key gaps in existing resources by integrating pharmacological text, molecular graphs, and chemical structure images, providing a comprehensive multimodal data source for investigating complex interactions between medications. Additionally, MUDI includes previously unseen drugs in the test set, thereby supporting robust evaluation of machine learning models’ generalization capabilities beyond known drug pairs.

The strengths of MUDI lie in its scale, multimodal richness, real-world clinical relevance, benchmarking results, and (free) availability. By covering diverse modalities and focusing specifically on pharmacodynamic effects, MUDI enables the development of potentially more accurate, generalizable, and targeted DDI prediction models. Further, MUDI’s structured annotation guidelines, open-access licensing, and reproducibility protocols enhance its utility for research, clinics, and potentially, drug discovery.

Exploring MUDI, future work may gain further insights into complex interactions between drugs. Future work may also use MUDI to improve medication assignment when treating comorbidities, and, potentially, based on selected multimodal features predictive DDI effects, improve drug development.

References

- [1] Masaki Asada, Makoto Miwa, and Yutaka Sasaki. 2021. Using drug descriptions and molecular structures for drug–drug interaction extraction from literature. *Bioinformatics* 37, 12 (2021), 1739–1746.
- [2] Masaki Asada, Makoto Miwa, and Yutaka Sasaki. 2023. Integrating heterogeneous knowledge graphs into drug–drug interaction extraction from the literature. *Bioinformatics* 39, 1 (2023), btac754.
- [3] Juan M Banda, Tobias Kuhn, Nigam H Shah, and Michel Dumontier. 2015. Provenance-centered dataset of drug–drug interactions. In *The Semantic Web-ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part II 14*. Springer, 293–300.
- [4] Jari Björne, Suwisa Kaewphan, and Tapio Salakoski. 2013. UTurku: Drug named entity recognition and drug–drug interaction extraction using SVM classification and domain knowledge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. 651–659.
- [5] Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 32, suppl_1 (2004), D267–D270.
- [6] Yifan Deng, Xinran Xu, Yang Qiu, Jingbo Xia, Wen Zhang, and Shichao Liu. 2020. A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics* 36, 15 (2020), 4316–4322.
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.
- [8] Norig Ellison. 2002. Goodman & Gilman’s the pharmacological basis of therapeutics. *Anesthesia & Analgesia* 94, 5 (2002), 1377.
- [9] Yanglan Gan, Wenxiao Liu, Guangwei Xu, Cairong Yan, and Guobing Zou. 2023. DMFDDI: deep multimodal fusion for drug–drug interaction prediction. *Briefings in Bioinformatics* 24, 6 (2023), bbad397.
- [10] Yu Gu, Robert Tinn, Hao Cheng, Matthew Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* 3, 1 (2021), 1–23.
- [11] Rave Harpaz, William DuMouchel, Nigam H Shah, David Madigan, Patrick Ryan, and Carol Friedman. 2012. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology & Therapeutics* 91, 6 (2012), 1010–1021.
- [12] Haohuai He, Guanxing Chen, and Calvin Yu-Chian Chen. 2022. 3DGT-DDI: 3D graph and text based neural network for drug–drug interaction prediction. *Briefings in bioinformatics* 23, 3 (2022), bbac134.
- [13] Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *elife* 6 (2017), e26726.
- [14] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. 2021. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548* (2021).
- [15] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)* (2015). <https://arxiv.org/abs/1412.6980>
- [16] Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)* (2017).
- [17] Gregory Landrum. 2013. RDKit: Open-source cheminformatics. <http://www.rdkit.org>. Accessed: 2025-04-19.
- [18] Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Proceedings of the 5th annual international conference on Systems documentation* (1986), 24–26.
- [19] Bradley M McQuade and Andrea Campbell. 2021. Drug prescribing: drug-drug interactions. *FP essentials* 508 (2021), 25–32.
- [20] Ishani Mondal. 2020. BERTChem-DDI: Improved Drug-Drug Interaction Prediction from text using Chemical Structure Information. In *Proceedings of Knowledgeable NLP: the First Workshop on Integrating Structured Knowledge and Neural Networks for NLP*. 27–32.
- [21] Yifan Peng, Anthony Rios, Ramakanth Kavuluru, and Zhiyong Lu. 2018. Extracting chemical-protein relations with ensembles of SVM and deep learning models. In *Proceedings of the BioCreative VI Workshop*. 155–157.
- [22] Martin F Porter. 1980. An algorithm for suffix stripping. *Program* 14, 3 (1980), 130–137.
- [23] Humphrey P Rang, Maureen M Dale, James M Ritter, Rod J Flower, and Graeme Henderson. 2011. *Rang & Dale’s pharmacology*. Elsevier Health Sciences.
- [24] Sunil Kumar Sahu and Ashish Anand. 2018. Drug-drug interaction extraction from biomedical texts using long short-term memory network. *Journal of biomedical informatics* 86 (2018), 15–24.
- [25] Yiyang Shi, Mingxiu He, Junheng Chen, Fangfang Han, and Yongming Cai. 2024. SubGE-DDI: A new prediction model for drug-drug interaction established through biomedical texts and drug-pairs knowledge subgraph enhancement. *PLOS Computational Biology* 20, 4 (2024), e1011989.
- [26] Nicholas P Tatonetti, Phyllis P Ye, Roxana Daneshjou, and Russ B Altman. 2012. Data-driven prediction of drug effects and interactions. *Science translational medicine* 4, 125 (2012), 125ra31–125ra31.
- [27] U.S. Food and Drug Administration. 2024. FDA Adverse Event Reporting System (FAERS) Database. <https://www.fda.gov/drugs/drug-approvals-and-databases/fda-adverse-event-reporting-system-faers-database>. Accessed: 2025-05-29.
- [28] U.S. National Library of Medicine. 2024. MEDLINE Database. https://www.nlm.nih.gov/medline/medline_home.html. Accessed: 2025-05-29.
- [29] Zhaoying Wang, Yingdan Shi, Xiang Liu, Can Chen, Jun Wen, and Ren Wang. 2025. HODDI: A Dataset of High-Order Drug-Drug Interactions for Computational Pharmacovigilance. *arXiv preprint arXiv:2502.06274* (2025).
- [30] David S Wishart, Yannick D Feunang, An C Guo, Elaine J Lo, Ana Marcu, Jason R Grant, Timothy Sajed, Daniel Johnson, Cecilia Li, Naina Sayeeda, et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research* 46, D1 (2018), D1074–D1082.
- [31] Guoli Xiong, Zhijiang Yang, Jiakai Yi, Ningning Wang, Lei Wang, Huimin Zhu, Chengkun Wu, Aiping Lu, Xiang Chen, Shao Liu, et al. 2022. DDInter: an online drug–drug interaction database towards improving clinical decision-making and patient safety. *Nucleic acids research* 50, D1 (2022), D1200–D1207.
- [32] Hui Yu. 2020. Data of multiple-type drug-drug interactions. [doi:10.17632/md5czfsfnd.1](https://doi.org/10.17632/md5czfsfnd.1)
- [33] Tianlin Zhang, Jiaxu Leng, and Ying Liu. 2020. Deep learning for drug–drug interaction extraction from the literature: a review. *Briefings in bioinformatics* 21, 5 (2020), 1609–1627.
- [34] Zhehuan Zhao, Zhihao Yang, Lijing Luo, Hongfei Lin, and Jian Wang. 2016. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics* 32, 22 (2016), 3444–3453.
- [35] Jie Zhou, Cong Fu, Jianzhong Zhao, Weilin Lv, Sheng Zha, Qing Li, and Jiawei Han. 2021. A comprehensive survey on knowledge graphs: Representation, acquisition and applications. *IEEE Transactions on Neural Networks and Learning Systems* 32, 5 (2021), 2241–2263.
- [36] Yuchen Zhu, Lei Li, Hang Lu, Aoying Zhou, and Xiaoyong Qin. 2020. Extracting drug-drug interactions from texts with BioBERT and multiple entity-aware attentions. *Journal of biomedical informatics* 106 (2020), 103451.
- [37] Marinka Zitnik, Rok Sosić, Sagar Maheshwari, and Jure Leskovec. 2018. BioSNAP Datasets: Stanford Biomedical Network Dataset Collection. <http://snap.stanford.edu/biodata>.

A Annotation Guidelines

This section describes the manual annotation protocol used to label drug-drug interaction (DDI) pairs in the MUDI dataset. Our objective is to create a standardized, clinically meaningful categorization of pharmacodynamic interactions into three classes: Synergism, Antagonism, and New Effect.

A.1 Drug Name Masking Policy

To ensure consistent pattern recognition and reduce annotator bias toward specific drug identities, all drug mentions in the original DrugBank interaction descriptions are replaced with abstract placeholders before annotation. This masking step is essential for allowing models and human annotators to focus on the nature of the pharmacodynamic interaction rather than the specific lexical forms of drug names.

Standardized Placeholders. The two interacting drugs in each sentence are masked using [DRUG1] and [DRUG2]. Any additional drug names appearing in the same sentence are replaced with [DRUGOTHER]. This abstraction is applied to both brand names and generic names, as well as chemical synonyms.

Drug Mention Alignment. Since DrugBank does not provide token-level alignment between drug entities and their textual positions, we employ a flexible string-matching algorithm to detect mentions of each drug and its synonyms. This process uses:

- Canonical names and aliases from DrugBank metadata.
- Case-insensitive matching.
- Partial overlap resolution (to disambiguate e.g., “Promazine” vs. “Acepromazine”).

Manual Review and Correction. In some cases, automatic matching resulted in incomplete or incorrect spans – especially when drugs shared lexical substrings or when spacing/punctuation was irregular. To address this, a team of three linguistics students manually reviewed and corrected all unique masked sentence templates. The original corpus of 244,921 raw DDI descriptions was thereby reduced to 287 distinct masked templates, from which 241 final sentence types were retained after quality filtering. Inter-annotator agreement during this manual review phase was near-perfect, with disagreements resolved by discussion and unanimous consensus.

Impact on Annotation Quality. This masking strategy eliminates the risk of model or annotator bias due to prior familiarity with drug names or brand-specific expectations. By enforcing a uniform abstracted representation across the dataset, we enable more consistent labeling of pharmacodynamic effects and allow models to generalize beyond known drug pairs.

A.2 Labeling Rules

Interaction descriptions in MUDI are categorized into one of three pharmacodynamic classes based on a set of carefully constructed lexical heuristics. These heuristics are grounded in recurring sentence structures observed in DrugBank and designed to reflect pharmacological theory [8, 23].

Synergism. Labeled when [DRUG1] enhances the pharmacological effect, bioavailability, or systemic concentration of [DRUG2].

Typical patterns include increased absorption, inhibited excretion, or elevated therapeutic efficacy.

Rule Templates for Synergism:

- [DRUG1] can cause an increase in the absorption of [DRUG2] resulting in an increased serum concentration and potentially a worsening of adverse effects.
- [DRUG1] may decrease the excretion rate of [DRUG2] which could result in a higher serum level.
- [DRUG1] may increase the [activities names] activities of [DRUG2].
- The bioavailability of [DRUG1] can be increased when combined with [DRUG2].
- The excretion of [DRUG1] can be decreased when combined with [DRUG2].
- The metabolism of [DRUG1] can be decreased when combined with [DRUG2].
- The protein binding of [DRUG1] can be decreased when combined with [DRUG2].
- The serum concentration of [DRUG1] can be increased when it is combined with [DRUG2].
- The serum concentration of [metabolite name], an active metabolite of [DRUG1], can be increased when used in combination with [DRUG2].
- The therapeutic efficacy of [DRUG1] can be increased when used in combination with [DRUG2].

Antagonism. An interaction is labeled as Antagonism if the description suggests that [DRUG1] inhibits, reduces, or interferes with the pharmacodynamic action of [DRUG2], including diminished absorption, faster metabolism, or decreased efficacy.

Rule Templates for Antagonism:

- [DRUG1] can cause a decrease in the absorption of [DRUG2] resulting in a reduced serum concentration and potentially a decrease in efficacy.
- [DRUG1] may decrease effectiveness of [DRUG2] as a diagnostic agent.
- [DRUG1] may decrease the [activities names] activities of [DRUG2].
- [DRUG1] may increase the excretion rate of [DRUG2] which could result in a lower serum level and potentially a reduction in efficacy.
- The absorption of [DRUG1] can be decreased when combined with [DRUG2].
- The bioavailability of [DRUG1] can be decreased when combined with [DRUG2].
- The excretion of [DRUG1] can be increased when combined with [DRUG2].
- The metabolism of [DRUG1] can be increased when combined with [DRUG2].
- The risk or severity of [adverse effects] can be decreased when [DRUG1] is combined with [DRUG2].
- The serum concentration of [DRUG1] can be decreased when it is combined with [DRUG2].
- The serum concentration of [metabolite name], an active metabolite of [DRUG1], can be decreased when used in combination with [DRUG2].

- The therapeutic efficacy of [DRUG1] can be decreased when used in combination with [DRUG2].

New Effect. An interaction is labeled as New Effect when the interaction leads to a novel adverse effect not known to be associated with either [DRUG1] or [DRUG2] independently. Initially, an interaction assigned the New Effect label when the sentence contains biomedical event terms (typically adverse effects) that cannot be clearly attributed to either Synergism or Antagonism patterns. These are typically masked sentences that do not indicate enhancement or suppression but instead describe the emergence of a distinct pharmacological effect.

To validate whether the reported effect is indeed novel to the drug combination, we perform a comparison between the extracted biomedical term and the known side effect profiles of both drugs. This process includes:

- (1) Automatically identifying the biomedical effect term in the masked sentence.
- (2) Matching the term against the adverse event metadata of each drug using stemming and synonym expansion [18, 22].
- (3) Assigning the New Effect label only if the effect is absent from both drug profiles.
- (4) Reassigning the sentence to Synergism if the effect is already associated with one of the drugs.

All candidate New Effect annotations are subsequently reviewed by domain experts to ensure biomedical validity. This step ensures that no pharmacologically implausible or redundant labels remain in the final dataset.

No or Unclear Interaction. Interaction descriptions that do not exhibit clear pharmacodynamic effects – such as therapeutic enhancement, attenuation, or novel adverse outcomes – are not assigned a positive label. This includes cases with vague, incomplete, or purely pharmacokinetic information lacking clinical consequence. In line with established pharmacological theory [8, 23], such drug pairs are considered to exhibit *no or unclear interaction*, indicating insufficient evidence to support classification into one of the defined pharmacodynamic categories.

A.3 Annotation Quality Control

To ensure the biomedical validity and consistency of interaction labels in MUDI, we implemented a two-phase expert curation protocol. After automated annotation using lexical heuristics (Section 3.2), each of the 241 distinct masked interaction templates was reviewed by two physicians with domain expertise in clinical pharmacology and drug safety.

Each expert independently examined the interaction context, checked consistency with known drug properties, and validated the correctness of the assigned label with respect to both pharmacodynamic semantics and biomedical relevance. In cases where the experts initially disagreed, they conducted a focused discussion to reach a consensus. This adjudication phase ensured the removal of residual annotation noise introduced by the automated pipeline, particularly in borderline or semantically ambiguous cases.

This dual-review and consensus-based process ensures that MUDI maintains clinically credible labels and supports reliable downstream model development.

B Dataset Access and Organization

This appendix provides practical information for obtaining and using the MUDI dataset, including access instructions, licensing terms, and directory structure.

B.1 Access and Licensing

The MUDI dataset is derived from DrugBank, a publicly accessible biomedical database. According to DrugBank’s data usage policy³, academic and non-commercial use of DrugBank content is permitted under its custom license for research purposes. In full compliance with this policy, MUDI builds on openly accessible DrugBank fields and restricts redistribution to non-commercial academic use only.

We release the MUDI dataset under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)⁴. This license enables researchers to copy, distribute, and adapt the dataset for academic purposes, provided that:

- Proper credit is given to both the MUDI project and the original DrugBank resource.
- Any derived works or models are clearly marked as adaptations.
- No commercial use is made without explicit written permission.

Users may download the dataset, documentation, preprocessing scripts, and baseline code from the following links:

- **Zenodo:** <https://zenodo.org/records/15544551> – the dataset archive with DOI for stable citation.
- **GitHub:** <https://github.com/hoangbros03/MUDI> – the code-base repository for preprocessing, baseline models, and future updates.

A permanent DOI link⁵ is assigned via Zenodo to ensure stable referencing and citation.

By downloading and using MUDI, users agree to comply with the license terms and responsible usage guidelines outlined in Appendix C.

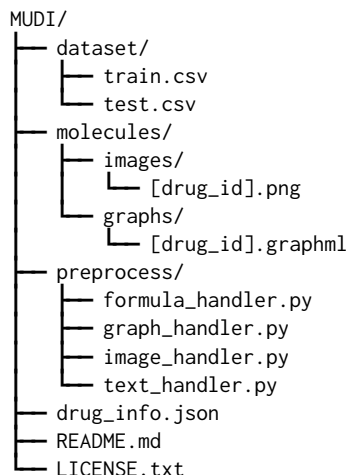
B.2 Dataset Structure

The MUDI dataset is organized into a clear and modular directory layout to facilitate ease of use, reproducibility, and multimodal experimentation:

³<https://go.drugbank.com/legal>

⁴<https://creativecommons.org/licenses/by-nc/4.0/>

⁵<https://doi.org/10.5281/zenodo.15544551>



dataset/train.csv and test.csv. Each row in these CSV files corresponds to a labeled drug-drug interaction (DDI) instance and contains the following fields:

- DRUG1: Unique identifier for the first drug.
- Interaction: One of the pharmacodynamic classes (Synergism, Antagonism, or New Effect).
- DRUG2: Unique identifier for the second drug.

All drug identifiers are keys in `drug_info.json` for retrieving textual, structural, and molecular representations.

molecules/images/. Each PNG image represents the 2D chemical structure of a drug generated from its SMILES string using RDKit. All images are:

- Named as `[drug_id].png`.
- Stored at a standardized resolution of 1000×800 pixels.
- Ready for direct use with image encoders such as Vision Transformer.

molecules/graphs/. Each file is a GraphML-encoded molecular graph where:

- Nodes represent atoms.
- Edges represent bonds (e.g., single, double, aromatic).

The files follow the standard GraphML format, with recommended compatibility via the NetworkX Python library.

drug_info.json. This JSON file consolidates metadata for each drug. Each entry contains:

- name: Human-readable drug name, e.g., “Amitriptyline”.
- description: A dictionary containing pharmacological text fields used for textual modeling:
 - summary: A concise overview of the drug’s identity, primary use, and general characteristics.
 - indication: Approved medical conditions or diseases that the drug is prescribed to treat.
 - metabolism: Description of the drug’s metabolic pathway, including hepatic enzymes involved (e.g., CYP450 family).
 - pharmacodynamics: Explanation of the biological effects, mechanism of drug action, and dose-response relationships.

- moa (mechanism of action): Detailed molecular-level explanation of how the drug achieves its intended effect, such as receptor binding or enzyme inhibition.

- formula: The molecular formula representing the elemental composition of the drug (e.g., C20H25N3O).
- smiles: The canonical SMILES (Simplified Molecular Input Line Entry System) string encoding the molecular structure in a compact, text-based format.

These fields are used to build modality-specific representations for textual, formula-based, and structural modeling.

B.3 File Standards and Preprocessing Code

File Formats.

- All interaction data is UTF-8 encoded CSV.
- Molecular graphs follow the GraphML standard.
- Chemical structure diagrams are stored as PNG images.
- Metadata is provided as structured JSON.

Preprocessing Scripts. The `preprocess/` directory includes modular Python scripts for converting raw inputs into model-ready formats:

- `text_handler.py`: Create a JSON object holding textual information, including name, description, SMILES, and formula.
- `formula_handler.py`: Improve the representation of molecule elements within chemical formulas before they are passed to the text handler.
- `image_handler.py`: Load the images from the dataset and convert them into tensors.
- `graph_handler.py`: Load and create graph objects.

Key dependencies include RDKit, networkx, and transformers.

B.4 Getting Started

To quickly begin using MUDI:

- Refer to `README.md` for installation, tutorials, and citation information.
- Use the `drug_info.json` file to retrieve all relevant metadata for each drug.
- Apply the preprocessing scripts to regenerate modality-specific features as needed.
- Evaluate models using the provided train/test splits and compute metrics such as precision, recall, micro-F1, and macro-F1.

The provided setup is fully reproducible and extensible for future multimodal biomedical research.

C Responsible Usage and Ethical Guidelines

This appendix details the ethical foundations, responsible usage requirements, and recommended best practices for working with the MUDI dataset. These principles aim to promote transparency, safety, and compliance in biomedical AI research.

C.1 Data Provenance and Privacy

The MUDI dataset is built entirely from non-sensitive, publicly accessible data obtained from the DrugBank database [30], a reputable biomedical resource. All data sources are governed by DrugBank’s

academic use policy⁶, which permits reuse for non-commercial research.

Importantly, MUDI does not contain any protected health information (PHI), patient-level records, or personally identifiable information (PII). No data originates from clinical trials, electronic medical records, or real-world hospital systems. As such, the dataset does not fall under the scope of human subjects research and does not require ethical approval from an institutional review board (IRB). It is also exempt from compliance obligations under HIPAA, GDPR, or related data privacy regulations.

C.2 Intended Use

MUDI is intended exclusively for academic research, education, and non-commercial purposes. Acceptable use cases include, but are not limited to:

- Development, benchmarking, and publication of multimodal learning algorithms for biomedical knowledge discovery.
- Research in drug-drug interaction (DDI) prediction, representation learning, cross-modal retrieval, and zero-shot biomedical reasoning.
- Classroom use in university-level courses or technical workshops on machine learning, drug discovery, or bioinformatics.

Any commercial use of the dataset is prohibited under the terms of the CC BY-NC 4.0 license without explicit written permission from the authors.

C.3 Known Limitations and Usage Caveats

Despite thorough curation and validation, MUDI remains a research-focused dataset and carries certain limitations:

- **No Clinical Validation:** Pharmacodynamic interaction labels are generated through lexical rules and expert curation, but not independently verified in wet-lab or clinical settings. The dataset is not intended for clinical use or decision support.
- **Potential Label Ambiguity:** The source descriptions from DrugBank are natural language statements, which may contain implicit or ambiguous interaction signals. While the annotation pipeline includes validation steps, some residual label noise is inevitable.
- **Pharmacodynamic Scope Only:** MUDI exclusively targets pharmacodynamic interactions. It does not cover pharmacokinetic DDIs such as those involving absorption, distribution, metabolism, or excretion (ADME) pathways.
- **Bias Toward Common Drugs:** Interaction labels are inherently more complete for well-studied drugs. This may bias model performance toward drugs with richer metadata and documented histories.

Researchers should exercise caution when interpreting results for clinical decision support or downstream biomedical applications.

C.4 Responsible Research Practices

We encourage users of MUDI to adopt the following practices to uphold ethical standards and maximize the scientific value of their work:

- Clearly cite the MUDI dataset and its associated publication in all derivative research.
- Disclose all modeling assumptions, training data subsets, and evaluation procedures to support reproducibility.
- Publicly release code and model checkpoints when possible, subject to the same licensing terms.
- Transparently communicate dataset limitations, especially when proposing real-world or clinical applications.
- Avoid deploying or advertising models trained on MUDI for direct use in patient care without formal clinical validation and regulatory approval.

D Baseline Model Configurations

This appendix provides detailed descriptions of the baseline models used to benchmark the MUDI dataset, including architecture choices, modality-specific preprocessing, and mathematical formulations.

D.1 Single-Modality Baselines

D.1.1 Text-only Baseline (BioMedBERT). The text-only model uses BioMedBERT [10] to encode concatenated pharmacological fields: summary, indication, mechanism of action, pharmacodynamics, and metabolism. The fields are joined into a single input sequence, separated by special tokens.

Given an input sequence \mathbf{x}^{text} , the model computes hidden representations \mathbf{h}_i for each token:

$$\mathbf{h}_i = \mathcal{E}_{\text{BioMedBERT}}(\mathbf{x}^{\text{text}})_i,$$

where i indexes the tokens.

We extract the embedding corresponding to the [CLS] token, $\mathbf{h}_{[\text{CLS}]}$, and apply a linear classification layer:

$$\hat{\mathbf{y}}_{\text{text}} = \text{softmax}(\mathbf{h}_{[\text{CLS}]} \mathbf{W}_t + \mathbf{b}_t),$$

where $\mathbf{W}_t \in \mathbb{R}^{d \times C}$ and $\mathbf{b}_t \in \mathbb{R}^C$ are learnable parameters, d is the embedding dimension, C is the number of output classes, and $\hat{\mathbf{y}}_{\text{text}} \in \mathbb{R}^C$ is the predicted class distribution.

D.1.2 Graph-only Baseline (GCN). The graph-only model uses a two-layer Graph Convolutional Network (GCN) [16] to process the molecular structure graphs generated from SMILES strings.

Each molecule graph is represented as an adjacency matrix \mathbf{A} and a feature matrix \mathbf{X} containing atom features. The GCN updates node features as:

$$\begin{aligned} \mathbf{H}^{(1)} &= \sigma \left(\tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{X} \mathbf{W}^{(0)} \right), \\ \mathbf{H}^{(2)} &= \sigma \left(\tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{H}^{(1)} \mathbf{W}^{(1)} \right), \end{aligned}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ is the adjacency matrix with self-loops added, $\tilde{\mathbf{D}}$ is the corresponding degree matrix, and σ denotes the ReLU activation function. The learnable parameters $\mathbf{W}^{(0)} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{hidden}}}$ and $\mathbf{W}^{(1)} \in \mathbb{R}^{d_{\text{hidden}} \times d_{\text{out}}}$ are weight matrices for the first and second GCN layers, respectively, where d_{in} is the input node feature dimension, d_{hidden} is the hidden dimension, and d_{out} is the output node feature dimension.

After the second GCN layer, we obtain node-level embeddings $\mathbf{H}^{(2)} = [\mathbf{h}_1^{(2)}, \mathbf{h}_2^{(2)}, \dots, \mathbf{h}_n^{(2)}]$, where $\mathbf{h}_i^{(2)} \in \mathbb{R}^d$ is the feature vector

⁶<https://go.drugbank.com/legal>

of the i -th node and n is the number of nodes in the molecular graph.

We apply global max pooling across nodes to produce a graph-level embedding:

$$\mathbf{z}_{\text{graph}} = \frac{1}{n} \sum_{i=1}^n \mathbf{h}_i^{(2)}.$$

The pooled graph representation $\mathbf{z}_{\text{graph}}$ is then passed through a linear classifier:

$$\hat{\mathbf{y}}_{\text{graph}} = \text{softmax}(\mathbf{z}_{\text{graph}} \mathbf{W}_g + \mathbf{b}_g),$$

where $\mathbf{W}_g \in \mathbb{R}^{d \times C}$ and $\mathbf{b}_g \in \mathbb{R}^C$ are the classification weights and bias, and C is the number of interaction classes.

D.1.3 Image-only Baseline (ViT). The image-only model employs a Vision Transformer (ViT) [7] to encode 2D chemical structure images.

Given an input image $\mathbf{x}^{\text{img}} \in \mathbb{R}^{H \times W \times 3}$, where H and W denote the height and width, the image is divided into N non-overlapping patches, each of size $P \times P$ pixels.

Each patch is flattened into a vector and linearly projected into a d -dimensional embedding space via a learnable matrix $\mathbf{W}_p \in \mathbb{R}^{(P^2 \times 3) \times d}$:

$$\mathbf{z}_i = \mathbf{x}_i^{\text{patch}} \mathbf{W}_p + \mathbf{b}_p, \quad \forall i = 1, \dots, N$$

where $\mathbf{x}_i^{\text{patch}}$ is the flattened pixel vector of the i -th patch.

The sequence of patch embeddings $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N\}$ is prepended with a learnable [CLS] token embedding $\mathbf{z}_{[\text{CLS}]} \in \mathbb{R}^d$, and positional encodings are added to preserve spatial information.

The resulting sequence is input into a standard Transformer encoder:

$$\mathbf{H} = \mathcal{E}_{\text{ViT}}([\mathbf{z}_{[\text{CLS}]}, \mathbf{z}_1, \dots, \mathbf{z}_N]),$$

where \mathcal{E}_{ViT} denotes the stack of transformer layers.

We extract the output corresponding to the [CLS] token, denoted as $\mathbf{h}_{[\text{CLS}]} \in \mathbb{R}^d$, and apply a linear classifier:

$$\hat{\mathbf{y}}_{\text{image}} = \text{softmax}(\mathbf{h}_{[\text{CLS}]} \mathbf{W}_v + \mathbf{b}_v),$$

where $\mathbf{W}_v \in \mathbb{R}^{d \times C}$ and $\mathbf{b}_v \in \mathbb{R}^C$ are learnable parameters, d is the hidden dimension, and C is the number of output classes.

D.2 Multimodal Baselines

D.2.1 Late Fusion Baseline. The late fusion baseline combines predictions from six independent single-modality classifiers, each trained on a distinct representation of drug information. Let \mathcal{M} denote the set of modalities:

$$\mathcal{M} = \{\text{name, description, SMILES, formula, graph, image}\}.$$

Each modality is processed by a dedicated model:

- **Name, Description, SMILES, Formula:** Each field is input into a separate BioMedBERT encoder to produce four independent textual predictions. For formula, the chemical formula is translated into a sequence of full element names (e.g., C20H25N3O \rightarrow carbon 20 hydrogen 25 nitrogen 3 oxygen) to align with the input of language models.
- **Graph:** The molecular structure graph is encoded using a two-layer GCN.

- **Image:** The 2D chemical structure is processed by a Vision Transformer.

Given a drug pair, each model $m \in \mathcal{M}$ produces a predicted label $\hat{y}_m \in C$, where $C = \{\text{Synergism, Antagonism, New Effect}\}$ is the set of pharmacodynamic interaction classes.

The final prediction \hat{y}_{late} is obtained through majority voting across modalities:

$$\hat{y}_{\text{late}} = \arg \max_{c \in C} \sum_{m \in \mathcal{M}} \mathbb{I}(\hat{y}_m = c),$$

where $\mathbb{I}(\cdot)$ is the indicator function.

In the case of a tie (i.e., multiple classes receiving equal votes), we apply a deterministic rule that prioritizes modalities based on their average F1 performance on the MUDI dataset, in the following order: graph \rightarrow name \rightarrow image \rightarrow SMILES \rightarrow formula \rightarrow description. This ordering is based on the empirical performance of the individual models on the development set.

D.2.2 Intermediate Fusion Baseline. The intermediate fusion baseline constructs a joint representation by integrating six modality-specific embeddings for each drug in the input pair. For a given drug pair (d_1, d_2) , we extract embeddings from the following modalities: name, description, SMILES (text), formula, molecular graph, and chemical image.

Each modality-specific encoder independently processes both drugs:

$$\begin{aligned} \mathbf{z}_m^{(1)} &= \mathcal{E}_m(d_1), \\ \mathbf{z}_m^{(2)} &= \mathcal{E}_m(d_2), \end{aligned} \quad \text{for each modality } m \in \mathcal{M},$$

where \mathcal{M} is the set of modalities, and $\mathbf{z}_m^{(i)} \in \mathbb{R}^{d_m}$ denotes the embedding of drug d_i in modality m .

We concatenate the two drug embeddings for each modality:

$$\tilde{\mathbf{z}}_m = [\mathbf{z}_m^{(1)}; \mathbf{z}_m^{(2)}] \in \mathbb{R}^{2d_m},$$

and subsequently form the full multimodal representation by concatenating across all modalities:

$$\mathbf{z}_{\text{fused}} = [\tilde{\mathbf{z}}_{\text{name}}; \tilde{\mathbf{z}}_{\text{desc}}; \tilde{\mathbf{z}}_{\text{smiles}}; \tilde{\mathbf{z}}_{\text{formula}}; \tilde{\mathbf{z}}_{\text{graph}}; \tilde{\mathbf{z}}_{\text{image}}] \in \mathbb{R}^{d_{\text{fused}}},$$

where $d_{\text{fused}} = 2(d_n + d_d + d_s + d_f + d_g + d_i)$.

This joint embedding is passed through a two-layer multilayer perceptron (MLP) with ReLU activation:

$$\begin{aligned} \mathbf{h}_1 &= \sigma(\mathbf{z}_{\text{fused}} \mathbf{W}_1 + \mathbf{b}_1), \\ \hat{\mathbf{y}}_{\text{inter}} &= \text{softmax}(\mathbf{h}_1 \mathbf{W}_2 + \mathbf{b}_2), \end{aligned}$$

where:

- $\mathbf{W}_1 \in \mathbb{R}^{d_{\text{fused}} \times d_{\text{hidden}}}$ and $\mathbf{b}_1 \in \mathbb{R}^{d_{\text{hidden}}}$ are parameters of the first MLP layer,
- $\mathbf{W}_2 \in \mathbb{R}^{d_{\text{hidden}} \times C}$ and $\mathbf{b}_2 \in \mathbb{R}^C$ are parameters of the classification head,
- C is the number of pharmacodynamic interaction classes.

This fusion strategy enables the model to learn pairwise dependencies and cross-modal interactions between drugs in a unified and expressive representation space.

D.3 Classification Task Definition

The central task in MUDI is formulated as a multi-class classification problem over pharmacodynamic drug-drug interactions. Given a pair of drugs (d_1, d_2) , the goal is to predict a single interaction label $y \in C$ based on their multimodal features. The label set C includes four possible classes:

- **Synergism** – drug d_1 enhances the effect of d_2 .
- **Antagonism** – drug d_1 reduces or nullifies the effect of d_2 .
- **New Effect** – the combination produces a novel outcome not present in individual use.
- **No Interaction** – no significant pharmacodynamic interaction is known or observed.

Each sample is annotated with one of the four mutually exclusive labels, with directional semantics included for Synergism and Antagonism. Specifically, (d_1, d_2) and (d_2, d_1) may correspond to different labels or directions unless symmetry is explicitly annotated (e.g., in New Effect cases).

To align with clinical interest in detecting meaningful drug interactions, our evaluation protocol concentrates on the three **positive interaction classes** – Synergism, Antagonism, and New Effect. The No Interaction label is used during training to simulate realistic class imbalance and improve discrimination, but is excluded from performance metric computation during test-time evaluation, following best practices in biomedical literature [21].

D.4 Prediction Thresholds

All models produce a probability distribution over the four interaction classes via a softmax output layer. Final class predictions are made using a maximum likelihood decision rule:

$$\hat{y} = \arg \max_{c \in C} p(c | \mathbf{x}),$$

where $C = \{\text{Synergism}, \text{Antagonism}, \text{New Effect}, \text{No Interaction}\}$, and $p(c | \mathbf{x})$ is the predicted probability for class c given multimodal input \mathbf{x} .

No additional confidence thresholding is applied. This choice ensures fair and consistent comparison across models, particularly under class imbalance conditions.

D.5 Reproducibility

To promote transparency and facilitate fair comparison, we standardize all experimental procedures as follows:

- **Randomness control:** All experiments are conducted with fixed random seeds for PyTorch, NumPy, and system-level generators to ensure consistent results across runs.
- **Dataset splits:** We use the same predefined training and test sets for all baseline models and fusion strategies.
- **Evaluation consistency:** All models are evaluated using a unified set of metrics and evaluation scripts, ensuring consistent treatment of prediction outputs under both direction-aware and direction-agnostic settings.

The full evaluation pipeline, including scoring functions and matching logic, is publicly available in the official repository (see Appendix B). This setup enables full replication of our results and supports future benchmarking efforts on the MUDI dataset.

E Evaluation Protocols and Settings

E.1 Evaluation Metrics

To evaluate model performance on clinically meaningful interaction types, we report standard classification metrics computed over the three positive classes: Synergism, Antagonism, and New Effect.

Precision (P). For each class, Precision is the proportion of correctly predicted instances among all instances assigned to that class:

$$\text{Precision}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c},$$

where TP_c and FP_c denote true positives and false positives for class c .

Recall (R). Recall is the proportion of correctly predicted instances among all actual instances of the class:

$$\text{Recall}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c},$$

where FN_c is the number of false negatives for class c .

F1 Score (F1). The F1 score is the harmonic mean of Precision and Recall:

$$\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}.$$

Micro-averaged Metrics. Micro-averaging aggregates true positives, false positives, and false negatives across all positive classes before computing Precision, Recall, and F1:

$$\text{Precision}_{\text{micro}} = \frac{\sum_c \text{TP}_c}{\sum_c (\text{TP}_c + \text{FP}_c)}, \quad \text{Recall}_{\text{micro}} = \frac{\sum_c \text{TP}_c}{\sum_c (\text{TP}_c + \text{FN}_c)},$$

$$\text{Micro-F1} = \frac{2 \cdot \text{Precision}_{\text{micro}} \cdot \text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}}.$$

Macro-averaged Metrics. Macro-averaging computes the unweighted mean of the per-class metrics. We first compute macro-averaged Precision and Recall:

$$\text{Precision}_{\text{macro}} = \frac{1}{3} \sum_{c=1}^3 \text{Precision}_c, \quad \text{Recall}_{\text{macro}} = \frac{1}{3} \sum_{c=1}^3 \text{Recall}_c.$$

Then, we define the Macro-F1 score as the harmonic mean of these macro-averaged values:

$$\text{Macro-F1} = \frac{2 \cdot \text{Precision}_{\text{macro}} \cdot \text{Recall}_{\text{macro}}}{\text{Precision}_{\text{macro}} + \text{Recall}_{\text{macro}}}.$$

Treatment of Negative Class (No Interaction). Although the negative class is included during training to enhance model calibration and decision boundaries, it is excluded from all evaluation metrics. This decision reflects established practice in biomedical relation extraction [21], which emphasizes performance on clinically actionable positive interactions.

E.2 Evaluation Settings

We evaluate model performance under two distinct matching settings to reflect different use scenarios:

- **Direction-aware Matching.** Drug pairs are treated as ordered tuples; that is, (DRUG1, DRUG2) and (DRUG2, DRUG1) are considered distinct. This setting requires models to not only

detect the correct interaction type but also capture the directionality – i.e., which drug initiates or modulates the effect.

- **Direction-agnostic Matching.** Drug pairs are treated as unordered sets. A prediction is considered correct if it matches the ground-truth interaction type for either (DRUG1, DRUG2) or its reversed pair (DRUG2, DRUG1). This relaxed setting reflects clinical cases where directionality is either symmetric or not explicitly defined.

Unless otherwise noted, all results reported in the main text follow the stricter direction-aware setting, which better aligns with real-world pharmacodynamic modeling.

F Training Environment and Hyperparameter Configurations

This appendix details the computational environment, software stack, and hyperparameter configurations used for training and evaluating all baseline models.

F.1 Hardware and Software Environment

Experiments are conducted on a Linux server with the following specifications:

- CPU: Intel(R) Xeon(R) CPU (2.2 GHz, 2 cores)
- GPU: 2× NVIDIA T4 GPUs (16GB VRAM each)
- RAM: 32GB DDR4 Memory
- Storage: 128GB NVMe SSD

The software environment is standardized as follows:

- Operating System: Ubuntu 22.04 LTS
- Python: 3.10
- PyTorch: 2.0.1
- CUDA: 11.8
- Transformers Library (HuggingFace): 4.31
- DGL (Deep Graph Library): 1.1.1
- scikit-learn: 1.2.2
- RDKit: 2022.09.5
- Additional packages: NumPy 1.24, SciPy 1.10, Matplotlib 3.7

F.2 General Training Settings

Unless otherwise specified, the following settings are shared across all baseline models:

- Optimizer: Adam [15]
- Initial learning rate: 5×10^{-5}
- Batch size: 32
- Learning rate scheduler: linear decay with warm-up (10% of total steps)
- Weight decay: 1×10^{-2}
- Dropout rate: 0.1 (applied after embeddings and in MLPs)
- Number of epochs: 100
- Gradient clipping: maximum norm of 1.0

Early stopping is applied based on validation loss with a patience of 5 epochs.

F.3 Model-Specific Hyper-parameters

F.3.1 Text-only Baseline (BioMedBERT).

- Pretrained checkpoint: ‘BioMedBERT-Base (uncased)’

- Maximum sequence length: 512 tokens
- Hidden size: 768
- Number of transformer layers: 12
- Number of attention heads: 12
- Fine-tuned end-to-end on MUDI dataset

F.3.2 Graph-only Baseline (GCN).

- Number of GCN layers: 2
- Hidden dimension (d_{hidden}): 768
- Input features: 37 atom features (one-hot encoded)
- Activation: ReLU
- Readout: global max pooling

F.3.3 Image-only Baseline (ViT).

- Pretrained checkpoint: ‘ViT-B/16’
- Image size: 1000×800 pixels
- Patch size: 16×16
- Hidden size: 768
- Number of transformer layers: 12
- Number of attention heads: 12
- MLP head dimension: 3072
- Fine-tuned end-to-end on MUDI dataset

F.3.4 Late Fusion Baseline.

- No additional training is performed.
- Predictions are aggregated from six single-modality models: Name, Description, SMILES, Formula, Graph, and Image.
- Tie-breaking priority: graph → name → image → SMILES → formula → description.

F.3.5 Intermediate Fusion Baseline.

- Embedding dimension for each modality: $d = 768$.
- Concatenated embedding dimension: $d_{\text{fusion}} = 6 \times 768 = 4608$.
- Fusion MLP: Two fully connected layers.
- Hidden dimension: 1024.
- Activation: ReLU.
- Dropout: 0.1 after each layer.
- Output: 4-way softmax classification.

F.4 Training Time

On average, training our baseline model takes:

- **Single-modality model:** 3.5–4 hours per modality (BioMedBERT, GCN, ViT, etc.).
- **Intermediate Fusion:** 4–4.5 hours, including preloading all modality-specific encoders and training the fusion MLP.
- **Late Fusion:** No additional training time, as predictions are directly aggregated from pretrained single-modality models.

All timing estimates are based on dual-GPU training using two NVIDIA T4 GPUs with 16GB memory on each GPU.

G Additional Results

G.1 Single-Modality Analysis

Table 6 presents the classification performance of each individual modality on the MUDI dataset under both direction-aware and direction-agnostic matching. Among all modalities, the molecular **graph**-based model performs best, achieving a Micro-F1 of 65.44% and Macro-F1 of 57.36% in the direction-agnostic setting.

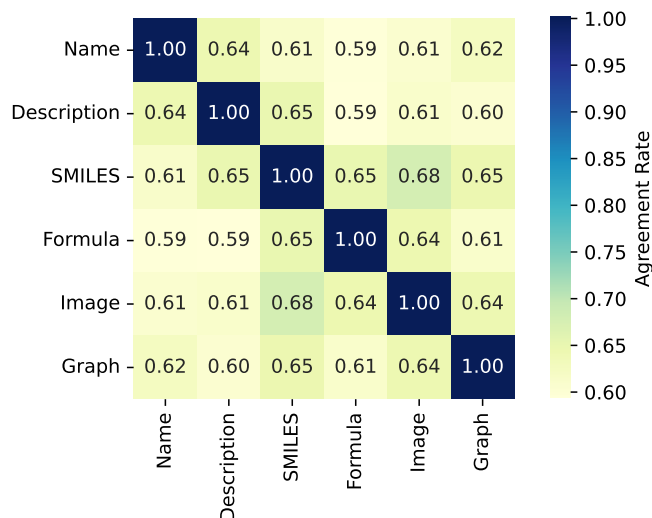


Figure 4: Agreement rate between different modalities.

This demonstrates the effectiveness of topological molecular representations in capturing pharmacodynamic interactions. The **name**-based model also yields surprisingly strong results, suggesting that drug identity alone encodes useful priors, especially when paired drugs have known interaction profiles.

In contrast, modalities like **SMILES**, **formula**, and **description** show limited standalone performance, particularly for the *New Effect* class. This result is consistent with the difficulty of extracting discriminative features from raw strings or sparse chemical formulas, and the noisiness of unstructured textual fields. Direction-aware results are consistently lower across all modalities, highlighting the increased challenge when models must account for interaction asymmetry.

To further understand how different modalities contribute to predictions, we visualize their agreement in Figure 4. The heatmap shows that while all modality pairs exhibit moderate correlation (typically between 0.59 and 0.68), **SMILES** and **image** achieve the highest agreement at 0.68, likely due to shared molecular-level information. This finding motivates future work on modality selection, weighted ensembling, or modality-specific gating to optimize fusion strategies.

G.2 Additional Ablation Studies

To further understand the contribution of modalities, we conduct several ablation experiments, each removing a modality to exhibit its impact on performance. Figure 5 illustrates the reductions in macro and micro F1 scores for Intermediate Fusion. Additionally, Figure 6 shows the decrease in the macro-averaged F1 metric across two fusion strategies.

The ablation study results in Figure 5 reveal a significant impact from Molecular Graph and demonstrate its enormous impact, as the scores are reduced by around 9%. Conversely, the performance when excluding Image and SMILE channels slightly improves the micro-averaged F1 but declines the macro-averaged F1, indicating

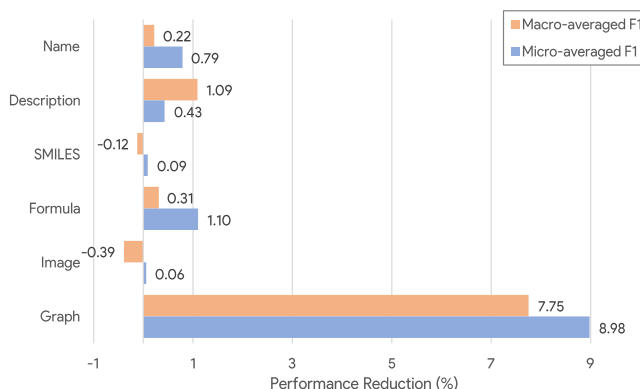


Figure 5: Macro and Micro F1 reduction with Intermediate Fusion.

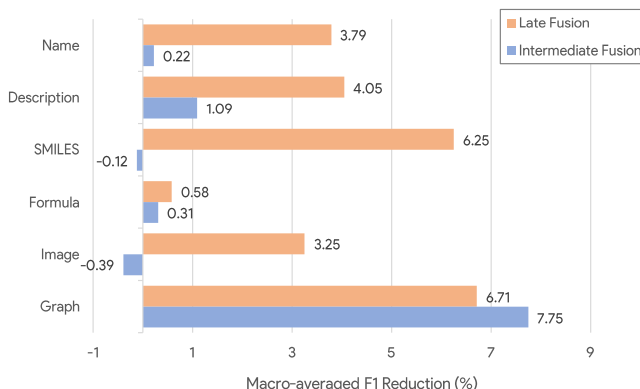


Figure 6: Macro F1 reduction between two fusion strategies.

potential conflict when adding these input sources. All other modalities contribute positively to prediction accuracy, however their individual impact is comparatively small.

As shown in Figure 6, Late Fusion generally shows a greater performance reduction than Intermediate Fusion when a specific modality is ablated. While removing SMILES in Intermediate Fusion shows a minor increase, its exclusion in Late Fusion results in a substantial 6.25% reduction, which underscores the inherent importance of this modality. This highlights the need for advanced fusion methods that can effectively integrate heterogeneous information from diverse modalities, as current baseline models may not fully leverage their complementary contributions.

Received 30 May 2025; revised xx xxx 2025; accepted xx xxx 2025

Table 6: Results of Single-Modality baselines on our MUDI dataset.

Modality	Metric	Direction-aware Matching					Direction-agnostic Matching				
		Synergism	Antagonism	New Effect	Macro-averaged	Micro-averaged	Synergism	Antagonism	New Effect	Macro-averaged	Micro-averaged
Name	Precision	38.73	41.13	33.84	37.90	38.81	62.92	59.21	53.74	58.62	62.25
	Recall	60.57	36.03	18.31	38.30	53.53	68.21	36.03	35.1	46.45	61.18
	F1	47.25	38.41	23.76	38.10	44.99	65.46	44.8	42.46	51.83	61.71
Description	Precision	41.86	0.00	0.00	13.95	41.86	65.78	0.00	0.00	21.93	65.78
	Recall	65.33	0.00	0.00	21.78	50.03	71.92	0.00	0.00	23.97	56.31
	F1	51.02	0.00	0.00	17.01	45.58	68.72	0.00	0.00	22.90	60.68
SMILES	Precision	32.35	30.64	0.00	21.00	32.17	54.32	47.95	0.00	34.09	53.58
	Recall	62.27	36.76	0.00	33.01	53.33	68.93	36.76	0.00	35.23	60.24
	F1	42.58	33.42	0.00	25.67	40.13	60.76	41.62	0.00	34.65	56.72
Formula	Precision	29.75	21.11	24.1	24.99	28.56	50.68	35.78	36.94	41.13	48.51
	Recall	56.55	31.57	8.07	32.06	48.98	63.99	31.57	15.84	37.13	56.22
	F1	38.99	25.3	12.09	28.09	36.08	56.56	33.54	22.17	39.03	52.08
Image	Precision	32.37	28.77	29.54	30.23	31.86	54.45	45.15	47.94	49.18	53.05
	Recall	54.82	37.94	12.76	35.17	48.96	62.09	37.94	25.01	41.68	56.25
	F1	40.71	32.72	17.82	32.51	38.6	58.02	41.23	32.87	45.12	54.6
Graph	Precision	53.09	61.27	52.30	55.55	54.05	77.23	74.56	71.24	74.34	76.69
	Recall	53.96	46.09	17.79	39.28	49.91	61.25	43.88	34.97	46.70	57.07
	F1	53.52	52.61	26.55	46.02	51.90	68.32	55.25	46.92	57.36	65.44