Clinical Trial Outcome Prediction using a Multimodal Mixture-of-Experts Approach expanding on the LIFTED Framework and interpretation aided by SHAP explanations

Tiago Mota

Work project carried out under the supervision of:
Qiwei Han

21-05-2025

Abstract (100 words maximum)

This work presents MMCTO, a multimodal framework predicting clinical trial outcomes by integrating molecular, disease, and eligibility data. Based on the LIFTED architecture, it employs natural language transformation and a Mixture-of-Experts mechanism to unify heterogeneous inputs. It demonstrates superior predictive performance across trial phases on HINT and CTOD datasets. Ablation studies confirm the importance of LLM-generated features and conditioned gating. Finally, for the individual body of work I'll explore the SHAP explanations which aim to provide transparency. The approach optimizes resources and streamlines processes, potentially avoiding costly failures and accelerating drug development timelines.

Keywords (Clinical Trial Outcomes, HINT, Large Language Models, Mixture-of-Experts, LIFTED, Natural Language, SHAP)

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Contents

1 Introduction

Clinical trials are the process through which new drugs or treatments are tested for their efficacy and safety before being introduced to the market. Given their impact on public health, the clinical trial procedure is rigorously planned and consists of a prolonged research study.

Their importance has only grown with the exponential expansion of the global pharmaceutical industry, which was valued at $390 billion in 2000 and is projected to exceed $1.5 trillion by 2024 (Statista 2023).

Despite their critical role, clinical trials often face major obstacles such as high costs, long durations, and low success rates (Sertkaya et al. 2016). Each failure not only entails substantial financial losses but also impedes the timely development and dissemination of therapeutics with the potential to significantly improve or preserve human life.

According to U.S. Food and Drug Administration (2025) each trial follows a uniform approach that is separated into distinct steps. The pharmacokinetic, safety, and tolerability assessments in patients or volunteers are performed on a small scale during the Phase I trial, whereas the efficacy assessments performed alongside safety evaluations in larger populations occur during phase II.

In order to measure therapeutic efficacy, side effect incidence, and experimental treatment comparisons to baseline medications or placebos in standard treatment, phase III trials require a greater diversity and number of participants.

A drug or treatment must successfully complete these three stages in order to receive government approval, confirming that it is safe and effective for public use.

The ability to anticipate the results of these trials is critically important for the pharmaceutical and healthcare industries. Accurate prediction allows for better resource allocation, reduced risks, and faster development of therapies (Qi and Tang 2019).

Recent advances in the digital transformation of healthcare have led to the accumulation of vast amounts of clinical data, creating new opportunities for predictive modeling. As Bernie Wood-

cock from the U.S. FDA has noted, "Improving the efficiency and success rate of clinical trials is essential to bringing new therapies to patients faster and at a lower cost" (Woodcock 2020). Researchers have increasingly turned to machine learning, and more recently, to deep learning and multimodal integration techniques, due to the complexity of the data and the interactions between variables such as molecular structures, patient profiles, and textual information. Traditional predictive modeling approaches often rely on modality-specific encoders, which limits their adaptability to new data types (Zhou et al. 2022).

With these limitations in mind, new frameworks such as LIFTED have been proposed, offering advanced model architectures like Mixture-of-Experts (MoE) and ambitious strategies that aim to synthesize and unify multimodal information through natural language transformation.

1.1 Evolution of Predictive Approaches

Traditional methods of machine learning often utilize organized data related to chemical compounds, biological entities, or the designs of the trials. Striving to enhance the clinical trial outcome prediction, these models emphasized specific characteristics of the drug, like its pharmacokinetics Qi and Tang (2019) or its toxicity Gayvert, Madhukar, and Elemento (2016).

Due to the increase in available biomedical data, as well as new developments in technology, the previously mentioned limitations can now be approached using deep learning techniques. An example would be Doe and Smith (2024) who created machine learning models that attempt to predict drug approvals for 15 different diseases using both the drug-specific features and trial design elements relevant to the drug. This demonstrates that there is potential for different contextual information to be integrated into predictive models.

Fu et al. (2022) proposed a new approach, using an interaction network that could leverage multi-modal data such as clinical trial records, therapeutic indications and molecular structures, by examining the complex relationships that exist between quantitative and qualitative data, which is a crucial and frequently challenging part of analysis for predicting an outcome. Even

though this system represented a significant advancement in the clinical trial landscape, it was still limited in its ability to expand with new types of data, becoming relevant and a shifting clinical trial paradigm, since it still relied on modality-specific encoders.

Alongside these examples, other strategies have emerged. Drug-disease relationships have been modeled with graph neural networks (GNNs) based on the capability of biological knowledge bases to predict certain clinical success (Himmelstein et al. 2017).

Transfer learning has also been researched in relation to model optimization by predicting clinical trials to tailor models initially trained on large biological datasets (Lee et al. 2020).

Furthermore, ensemble methods, which add multiple different predictive models for the same task, have been studied to augment robustness and generalizability across treatment areas (Sahoo, Pham, and Hoi 2021).

Despite these promising developments, most models still struggle in two areas: combining data such as molecular structures, clinical trial records, and patient information into one system, and generalizing data across trial phases and other areas of therapy. It is apparent that more cohesive predictive systems that work using an array of different methods and materials are better suited to enhance predictive models.

However, Large Language Models (LLMs) offer a fresh and innovative method by transforming heterogeneous data into a unified natural language format, allowing the employment of transformer-based systems to complex clinical trial datasets with reasoning capabilities. This will result in more accurate and profound predictions for clinical trial outcomes (Lee et al. 2023)

.

1.2 The Challenge of Multimodal Integration

The effective integration of multimodal data with the purpose of predicting clinical trial outcomes integrates the greatest challenges associated with data diversity and complexity (Miotto et al. 2016).

First, synthesizing uniform representations from very different data types can be demanding. For instance, molecular information is often represented as graphs depicting atoms and bonds, whereas disease classifications tend to use ontological tree structures that employ semantically rich hierarchies of relations between conditions (Zitnik, Agrawal, and Leskovec 2018).

Each domain of biology relies on its own specialized encoders, which complicates the extraction of meaningful features. Therefore, any unified encoder must be able to consolidate and standardize this data into a homogeneous product (Shickel et al. 2018).

Second, the use of specific information patterns across modalities enables cross-consideration. Some information types like disease descriptions and their accompanying medication lists contain features that can be extracted like symptoms or treatment effects.

On the other hand, some modalities such as chemical structures and drug names contain completely different information, which require different approaches for extraction.

Last but not least, the composition of predictive models from the integrated representations is also a problem.

There are different possible ways that the multi-source information can be combined and some of them might not be optimal while others may be substantially more valuable, especially across evaluations on clinical trials.

Bearing that in mind, each mechanism of integration should strive to fulfill the informational demands of every prediction task by dynamically adjusting the balance and contribution of components within the model to satisfy its specific predictive information requirements.

Addressing these issues is critical in creating predictive systems that are both adaptive and robust, enabling the multi-faceted clinical trial data to be utilized optimally, thereby enhancing the precision of outcome predictions (Zhou et al. 2022).

1.3 The LIFTED Approach

To address these issues, Zheng et al. (2025) designed an approach they named LIFTED (MuLtimodal Mix−of−Experts For Outcome Prediction). This method employs a transformer-based unified encoder to extract features from multiple data modalities (Vaswani et al. 2017), re- fines these features using a Sparse Mixture−of−Experts (SMoE) framework (Fedus, Zoph, and Shazeer 2022), and ultimately integrates the multimodal features through a final Mixture−of−Experts (MoE) model (Shazeer et al. 2017).

Deployment of LIFTED allows description-based unification of language features that transcend mode-specific boundaries within a single system. Language description extraction in- volves the crafting of a transformer-based encoder and the application of model-specific language description processing, refining using an SMoE framework afterwards.

Within SMoE, representations derived from multiple modalities are executed through dynamic routing within a top-k gating network (Shazeer et al. 2017) that introduces noise towards shared expert models specific to identifying identical information patterns.

1.4 Objectives

This thesis is committed to exploring the clinical trial predictive modelling area with the attention to one of the most important foundational areas of the LIFTED pipeline: leveraging large language models (LLMs) to generate natural language narratives from complex multimodal clinical trial data.

This conversion makes it possible to harmonize different types of data into a single format that can be understood and used by transformer-based encoders.

This thesis aims to produce four main outcomes. First, it aims at providing a comprehensive and detailed account regarding transforming multimodal clinical trial data into natural language description templates suitable for predictive modeling.

Second, it proposes a qualitative analysis of the output cross-evaluated among varying LLMs

and their prompts, which provide insights into the relative strengths and weaknesses of various models and configurations, as well as inform best practices for their effective use.

Third, it attempts to investigate how transforming heterogeneous data into a single natural language representation can enable better integration of multiple sources of information for predicting clinically relevant outcomes of clinical trials.

Lastly, it discusses some conclusions and suggestions for future research, especially regarding the utilization of LLMs and natural language technologies for clinical trial analysis.

2 Literature Review

The following sections present a concise literature review of recent advancements in clinical trials, along with a summary of the key resources used in developing predictive models.

We aim to illustrate the chronological development of this area to better understand the challenges in optimizing clinical trial forecasts and to clarify our approach and contribution.

Additionally, we present an overview of commonly used tools in this field, including those in our own work, to enhance transparency and trust in our model and its results. By doing so, the goal is to foster greater transparency and confidence in the models functioning and resulting predictions.

2.1 Historical Approaches

Due to the large expenses and poor success rates associated with clinical trials, researchers have studied and developed methods attempting to model predictions for the results of clinical trials for decades.

Traditional statistical techniques, regression analyses, and basic probabilistic models applied to small sets of molecular and clinical variables were the mainstays of early methodologies.

During the 1990s, one of the first systematic approaches to predicting pharmacological properties based on molecular attempts was the Quantitative Structure Activity Relationship (QSAR)

techniques (Tropsha 2010).

These methods formed relationships between the biological activity of compounds and their physicochemical properties, therefore allowing drug efficacy and toxicity predictions to be made (Hansch and Fujita 1964). These approaches were still limited by the large, simplistic mathematical framework used and the sparse datasets available.

With the emergence of genomics in the early 2000s, new datasets composed of biological data, such as gene expression data, were integrated into the existing molecular frameworks (Wang, Gerstein, and Snyder 2009). In a different example, Lamb et al. (2006) created the 'Connectivity Map', an approach that employed gene expression profiles to find associations among diseases, genes, and chemical compounds which could be used to hypothesize their impact on treatment. This further added to the existing prediction models which incorporated new biological data.

An important advancement came along with the work Gayvert, Madhukar, and Elemento (2016) performed, as they pioneered one of the computational frameworks focused on predicting drug toxicity. The framework created by these researchers integrated the compound's structural features with its historical toxicity datasets, strengthening the algorithm and proving machine learning's worth in predicting crucial aspects of a clinical trial.

Simultaneously, Qi and Tang (2019) applied deep forest-based models in predicting drug-target interactions, which is significant for evaluating the therapeutic potential of a new compound. Their deep learning technique was able to model complex nonlinear interactions of molecular features with biological activities, as opposed to the previously used linear approaches.

2.2 Recent Advances

There have been many contributions towards a more efficient and result-driven approach for predicting clinical trials results.

This section highlights key milestones over the past several years, notably the rise of deep learning and the expansion of clinical trial data, which have been central to recent advances in

outcome prediction. One of these methods was created by Lo et al. (2019) through a comprehensive approach that combined heterogeneous data sources across 15 different medical conditions categories, aiming to predict future drugs approval success rate. Their work mainly highlighted the importance in combining contextual clinical trial data with genetic traits as well as the opportunities it offered in using a multi-modal approach to predictive modeling.

Additionally, Y. Wu et al. (2023) faced the same challenge traditional modelling approaches faced before him, the hardship of representing complex relations between different data types such as drugs, diseases, and genes. To surpass this obstacle Y. Wu et al. (2023) constructed a heterogeneous biomedical knowledge graph (BKG), a method that combines graph embeddings and GNN (graph neural networks), which proved to be a highly valuable innovation since graph-based methods offer an effective solution to model complex biological relationships among key factors in clinical trials leading to applications such as drug repurposing, disease-genes association predictions and overall clinical trial optimization.

Accurately training models from scratch comes with a multitude of obstacles which is mainly due to the relatively small number of clinical trials with available outcome data. To answer this challenge, the field saw increased use of transfer learning techniques, which models have been pre-trained on extensive, domain-relevant datasets and then transfer their knowledge to the specific task of predicting clinical trial outcomes on a smaller dataset.

One of the most promising processes presented in the last few years was HINT (Hierarchical Interaction Network). HINT is a framework developed by Fu et al. (2022) which offers a unique and innovative way to utilize multi-modal data and analyze intricate correlations thereby significantly enhancing the structured integration of various information sources by explicitly modeling hierarchical interactions between components such as medications, illnesses, inclusion criteria, and patient demographics.

Finally, Zheng et al. (2025), developed LIFTED, which purportedly has improved flexibility and scalability solutions for dealing with the complex nature of clinical trial data. Any level of interpretation, using natural language as a standard representation, promotes integration,

however LIFTED solves major problems derived from previous works, particularly with heterogeneous data processing and adaptation to different types of clinical trials.

LIFTED inherent qualities pertaining to the above-mentioned challenges critically relieve constraints to a predictive modelling approach, thus making it the best option to expand upon and why it will be the main focus of our thesis.

2.3 CTOD (Clinical Trial Outcome Dataset)

As mentioned in the previous section, one major challenge clinical trials prediction models face is the lack of dependable and robust datasets on which models can be trained and tested. To fix this issue Gao et al. (2025) established a dataset called CTOD.

CTOD compiles information from public sources like ClinicalTrials.gov (n.d.) and the EMA's database, emphasizing Phase II and III trials for new molecular entities.

It features a multimodal and more integrated schema than previous datasets, including molecular disease information like SMILES codes and physicochemical properties, textual descriptions, ICD-10 enumeration and severity, trial eligibility and endpoints, study design, sample size, duration, and binary outcome relating to efficacy and safety, including detailed annotations for failed trials when available (Gao et al. 2025).

Covering 4,182 unique clinical trials, 1,749 distinct compounds, and 437 medical conditions, CTOD notably exhibits a class imbalance, with only 32% of trials classified as successful.

Standardized evaluation protocols, including stratified train/validation/test splits and specific performance metrics, were also established by Gao et al. (2025) to ensure fair comparisons between models.

CTOD is revealed to be an important development for the available resources used in clinical trial outcome prediction research as it provides a high-quality, reliable clinical trial outcome dataset.

2.4 Foundational Resources for Predictive Modelling

Numerous other resources, in addition to CTOD, have proven valuable for research on clinical trial outcome prediction, some of which will be presented in this section.

DrugBank, is a database that contains detailed information for pharmaceutical drugs including their molecular targets, mechanisms of action, drug-drug interactions, and metabolic pathways (Wishart et al. 2018). It is mainly used as a source of molecular and pharmacological input features for predictive models.

PubChem, is a public repository of chemical substances and their biological activities, offering structural and experimental data used to identify substances that can potentially integrate a new drug (Kim et al. 2021). The AACT (Aggregate Analysis of ClinicalTrials.gov) database provides data relating to all publicly available information from ClinicalTrials.gov (n.d.), facilitating analyses of clinical trial trends and properties.

MoleculeNet, introduced by Z. Wu et al. (2018), offers a benchmark for molecular machine learning, covering tasks such as toxicity, solubility and permeability which are important factors in a clinical trial.

The Unified Medical Language System (UMLS) standardizes textual descriptions of diseases and medical conditions, which facilitates data interoperability and consistency across computational systems (Bodenreider 2004).

Additionally, structured resources such as Gene Ontology (Ashburner et al. 2000), KEGG (Kanehisa et al. 2017), and Reactome (Jassal, Matthews, Viteri, et al. 2020) provide detailed information on genetic functions, metabolic pathways, and biological processes, enabling the incorporation of mechanistic knowledge into predictive models.

Together, these resources complement clinical trial-specific datasets by providing information that improves both the accuracy and transparency of predictive models. Also, by granting more contextual data to a model we ensure its robustness and capability of processing new data.

2.5 Interaction Networks

Interaction networks are a powerful instrument, applied to explore interdependencies between variables and consequently map observed correlations.

This approach proves highly beneficial, considering that most models integrate data from a wide range of often heterogeneous modalities, including diseases, genes, and proteins as graph nodes, with edges denoting their interactions.

The image below illustrates one such interaction network, clearly demonstrating the mapping structure previously outlined.

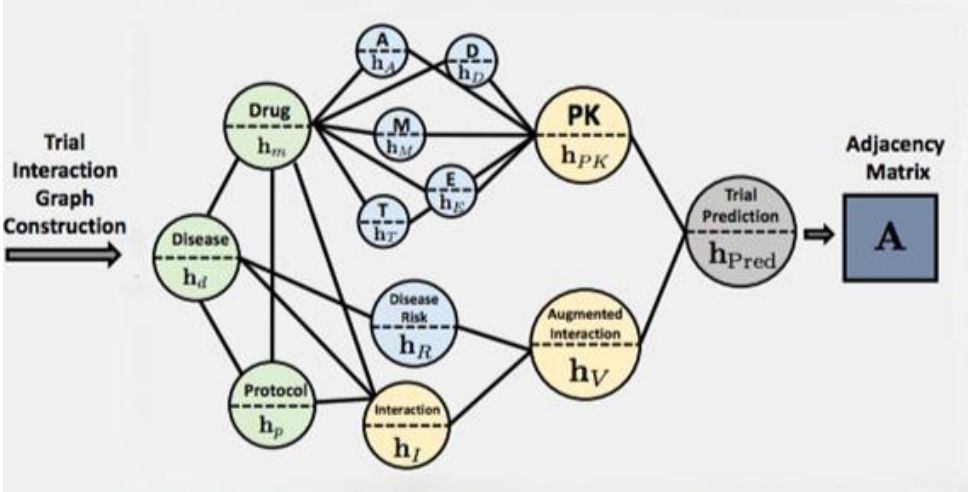

Figure 1: Architecture of the Trial Interaction Graph-Based Prediction Model

These networks capture interactions across multiple domains. By coupling variables and analyzing the resulting outcomes, interaction networks enable the efficient identification of relationships such as drug target, drug disease, disease symptoms, and drug adverse effect each crucial for evaluating therapeutic efficacy and safety.

As a result, they serve an essential role in predictive modeling by improving data interpretation, deepening overall understanding, enhancing prediction accuracy, and allowing for the incorporation of additional information such as metabolic pathways, drug-drug interactions, and genetic

variations.

However, one of their major limitations lies in their dependence on prior knowledge, which is often incomplete or biased toward specific medical domains.

With the objective of addressing this issue, the HINT (Hierarchical Interaction Network) model, developed by Fu et al. (2022), emerges as a key solution. Its main innovation lies in its hierarchical structure, designed to capture the distinct and complex relationships between multimodal variables relevant to clinical trials setting it apart from traditional models.

With the objective of predicting the success or failure of clinical trials, HINT integrates data in a more targeted manner, specifically extracted from Phase II and III studies involving small-molecule drugs, using reliable sources such as *ClinicalTrials.gov*, *DrugBank*, *PubChem*, and biomedical literature.

This structure is organized into different levels: At the lower level, drugs are represented by their molecular graphs, chemical fingerprints, and physicochemical properties, while diseases are described through medical ontologies, associated symptoms, and pathological mechanisms. At the interaction level, relationships such as drug disease, drug target, and disease symptom are mapped.

Finally, the trial level includes study-specific attributes such as phase, design, target population, and predefined endpoints.

The HINT dataset comprises 3,927 clinical trials, covering 1,251 drugs and 375 medical conditions. One of its key advantages is the richness of detailed information on known drug target interactions, which enhances the biological relevance of models in a clinical context.

The proposed framework leverages a hierarchical interaction network to fuse these diverse layers of information, employing attention mechanisms to capture dependencies across the hierarchy. This strategy outperforms approaches that treat modalities in isolation, reinforcing the value of modeling cross-modal biomedical interactions to improve clinical trial outcome pre- diction.

In summary, HINT applies attention mechanisms to integrate information across network levels, enabling a precise focus on the most relevant relationships in each case.

It is a proven structure for effectively capturing the clinical context of variable interactions something traditional approaches often fail to represent, interpret, or incorporate into final outcomes.

2.6 Unified Representations and the LIFTED Model

One of the developing methods for multimodal integration is to map data from multiple sources into a single representation space. This simplifies data processing and pattern recognition across various modalities, allowing different types of data, such as clinical notes, medical imaging, and temporal data, to be handled in a unified manner (Krizhevsky, Sutskever, and Hinton 2012).

This approach facilitates the creation of more accessible and adaptable interfaces for predictive models, promoting interoperability and flexibility. To achieve this unified representation of data, various techniques have been explored, one of which involves constructing a latent space where variables are vectorized, allowing mathematical operations and comparisons between different types of data (Bordes et al. 2013).

This technique makes use of translational embeddings (TransE) and semantic embeddings, enabling the reuse of knowledge from pre-trained models in various contexts (Mikolov et al. 2013).

3 Materials and Methods

The chapter is structured around three central sub-sections, as follows: data collection, benchmark and research method. These sub-sections will be presented as follows:

3.1 Data Collection

This section introduces the data sources used in our experimental setup, which forms the foundation for converting the data into natural text and for fine-tuning and evaluating the models. The datasets used throughout this work are HINT (Favita 2025) and CTOD (Gao et al. 2025) datasets.

It is important to note that only 4,305 clinical trials from the CTO dataset were matched with the HINT dataset.

3.1.1 HINT

This dataset consists of 102, 655 clinical trial records, distributed across 30, 944 trials in Phase I, 38, 639 in Phase II, and 23, 804 in Phase III. Each record includes 13 features such as the trial title, associated diseases, a brief description, the SMILES representation of the drug, target conditions, eligibility criteria, trial phase, and a binary outcome label indicating success or failure.

No additional raw data cleaning was performed in this study, as the dataset had already been curated by Favita (2025) in accordance with the TOP benchmark (Fu et al. 2022), which ensured high-quality and consistent annotations.

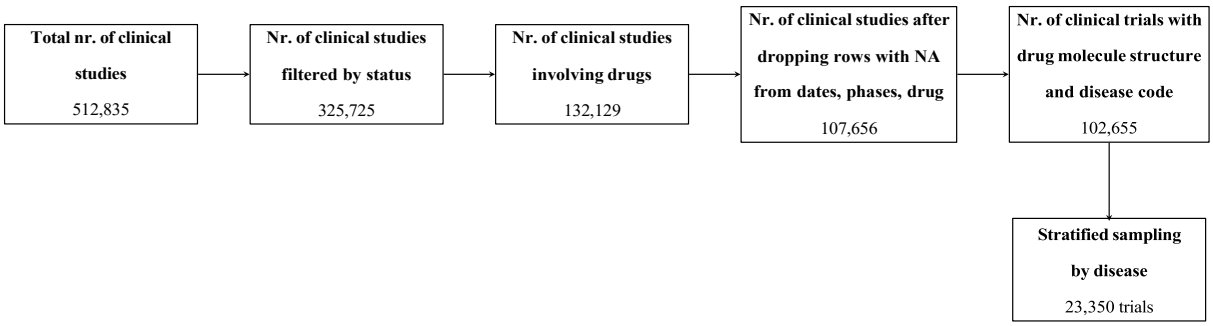

Figure 2: Data Preprocessing HINT Pipeline

To ensure high-quality outcome labels, the original dataset underwent several filtering steps. Only trials that reported statistical findings related to the primary outcome were considered

Additionally, as the study focuses specifically on small-molecule drug trials, records involving other treatment types, such as biologics or behavioral interventions, were excluded.

Only trials containing complete information on drug molecular structure (SMILES) and disease codes were considered. Records with missing values in critical fields such as trial dates, phases, or drug information were also removed to maintain dataset integrity.

Given the high dimensionality and volume of clinical trial records and the associated computational and financial cost of processing each trial via the OpenAI APIa reduction strategy was necessary. To preserve the diversity and analytical value of the data, a stratified sampling approach was adopted based on disease categories.

An additional step was added to the original data pipeline curated by Favita (2025), consisting of a stratified sampling procedure aimed at reducing dataset size while maintaining the clinical representativeness of disease areas.

Specifically, ICD-10 codes were mapped to the Clinical Classifications Software Refined (CCSR) system, and the first valid CCSR code was used to assign a primary disease category to each trial. Trials were then sampled proportionally across these categories. This ensured that the relative frequency of disease types remained consistent in the reduced dataset, which is important for maintaining realistic success rate distributions. For example, trials in oncology often exhibit lower success rates than those in infectious or metabolic diseases, so preserving this balance is essential for fair model evaluation and generalization (BIO and Advisors 2021).

After this stratified reduction, the HINT dataset retained a total of 23, 350 trials: 7, 736 in Phase I, 9, 662 in Phase II, and 5, 952 in Phase III (see Appendix, Table A9). Phase IV trials were completely removed, given that these refer to post-marketing pharmacovigilance (Ratan 2023), thus being outside the scope of this study, which is focused on drug development in early and intermediate stages.

Notably, this approach retained more clinical trials than the method proposed by Zheng et al. (2025), providing a broader representation of the clinical trial landscape. The larger sample

size enhances statistical power, supports more robust subgroup analyses, and improves model training particularly for rare disease categories where data scarcity is a challenge (U.S. Food and Drug Administration 2019).

3.1.2 CTOD

The data utilized in this study were primarily obtained from the Clinical Trials Transformation Initiative (CTTI) (n.d.) dataset, provided in compressed format (*CTTI −new.zip*).

Additionally, trial outcome labels were sourced from the Clinical Trial Outcomes (CTO) dataset, publicly accessible on the Hugging Face Datasets platform developed by Gao (2024). These labels reflect success or failure predictions generated by machine learning models trained on multimodal clinical trial data.

The initial dataset consisted of 124,917 clinical trial records, each containing fourteen fields, including the study title, associated diseases, a brief description of the pharmaceutical agents, the SMILES representation of the drug, target conditions, eligibility criteria, trial phase, number of patients, and a binary outcome label. To ensure the quality of outcome labels and maintain consistency with prior work, preprocessing was performed according to the criteria defined by the CTO benchmark (Gao et al. 2025), which adopts filtering steps similar to those applied in the TOP benchmark used for HINT.

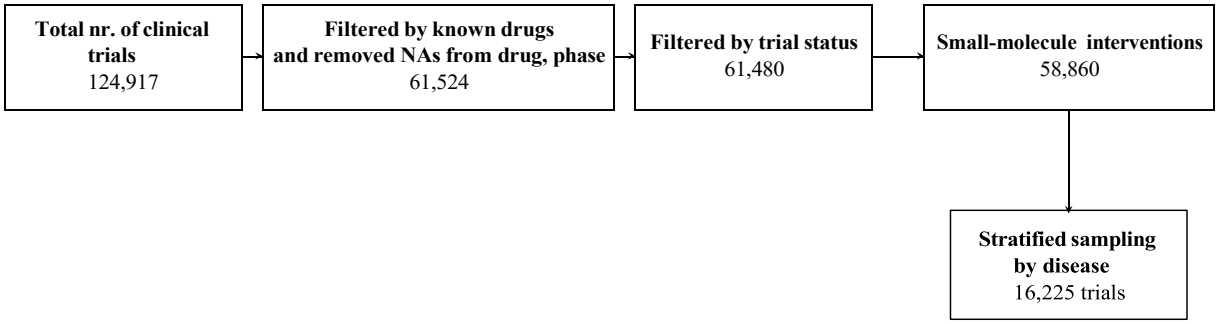

Figure 3: Data Preprocessing CTOD Pipeline

Only drug-based interventions were included, limited to completed studies that reported

statistical analysis for primary outcomes. Trials lacking phase information were excluded. This resulted in a refined subset of 58,860 clinical trials containing complete structural and clinical metadata.

While the original CTO benchmark focused only on trials completed after 2020, the present study retained all available trials in order to expand the temporal scope and improve statistical power for training and analysis.

Standardization steps were also performed to ensure consistency across modalities. Drugs were represented using SMILES notation, and diseases were encoded with ICD-10 classification codes. This standardization, a widely adopted practice in biomedical informatics (Weininger 1988), facilitates data integration, cross-study comparisons, and model reproducibility.

Similarly to the HINT section, an additional step was added to the CTOD data pipeline. A stratified sampling reduction based on disease categories mapped using the Clinical Classifications Software Refined (CCSR)was applied to reduce the dataset size while preserving the clinical heterogeneity of therapeutic areas.

This sampling ensured that the relative distribution of diseases remained consistent, which is essential for fair evaluation and generalization across medical domains.

After applying this reduction, the final CTOD subset retained 16,225 trials, distributed as follows: 5,022 Phase I, 6,738 Phase II, and 4,465 Phase III trials (see Appendix, Table A10).

Phase IV trials were excluded entirely, as they relate to post-marketing surveillance and fall outside the scope of this study, which focuses on early- and mid-stage drug development.

3.2 Benchmark

3.2.1 Drug Molecule Data

Molecular data were extracted from clinical trials in the CTTI dataset (Clinical Trials Transformation Initiative (CTTI), n.d.), where drug names were identified in the metadata and linked to their corresponding molecular structures using SMILES notation. These structures, along

with IUPAC names, molecular weights, descriptions, and mechanisms of action, were retrieved from PubChem (n.d.) and ChEMBL (n.d.), two widely used chemical substance repositories in pharmaceutical research.

SMILES (Simplified Molecular-Input Line Entry System) is a standardized text-based notation that encodes molecular structures for computational analysis (Weininger 1988). It enables the integration of chemical data with biomedical databases and supports tasks such as structural modeling, similarity analysis, and pharmacological prediction.

Canonical SMILES strings and associated metadata were obtained via the PubChemPy API (n.d.) API. Additional drug-related information such as approval status, mechanism of action, and target proteins was retrieved from ChEMBL (n.d.), ensuring comprehensive molecular coverage.

A name-matching algorithm was implemented to associate each trial with the appropriate molecular entry. Exact and approximate matching techniques were used to address inconsistencies, aiming to assign a single standardized SMILES string per drug for consistent downstream analysis.

This standardization supports interoperability and enables machine learning models to infer pharmacological properties, predict outcomes, and explore drug repositioning opportunities.

Drug descriptions were further enriched using a curated mapping database available from Zheng et al. (2025). Only validated drugs were retained; unlisted descriptions were recorded as null values to preserve data integrity.

The final dataset comprises a unified, standardized set of molecular information, enhancing both the semantic richness and analytical value of the clinical trial data.

3.2.2 Disease Data

Disease data were extracted from clinical trials collected in ClinicalTrials.gov (n.d.) and mapped to ICD-10 codes provided by the World Health Organization (2019). Disease names were obtained from the *browse_conditions.txt* file of the *CTTI* dataset and normalized by con-

verting to lowercase and removing duplicates.

These terms were then mapped to standardized ICD-10 codes through an external API from the U.S. National Library of Medicine (n.d.), called Clinical Tables. The ICD-10 (International Classification of Diseases - 10th Revision) is the global standard for systematic recording, re- porting, and analysis of morbidity and mortality data.

Despite the official release of ICD-11 in January 2022, ICD-10 remains the most pragmatic choice for this study due to its widespread adoption, compatibility with existing tools and datasets, and its continued relevance in clinical research.

To perform the mapping between diseases and codes, an automated algorithm was implemented that queries the API with each unique clinical term. If the API returns one or more valid codes, they are associated with the corresponding disease.

When there is no match, a placeholder indication is assigned. This process ensures high-quality mapping that is robust to terminological variations.

The result is a structured table that links disease terms from clinical trials to standardized ICD-10 codes, increasing the interoperability and analytical value of the dataset.

This step is essential for standardizing terminology in biomedical studies and for integrating the data with other sources that also use the ICD classification.

3.3 Model

3.3.1 Overview of the Approach

This paper is based on the framework of Zheng et al. (2025), with the aim of mitigating the high costs associated with clinical trials, not only by improving the prediction of the outcomes of these trials but also by reducing training costs both in terms of time and computational resources.

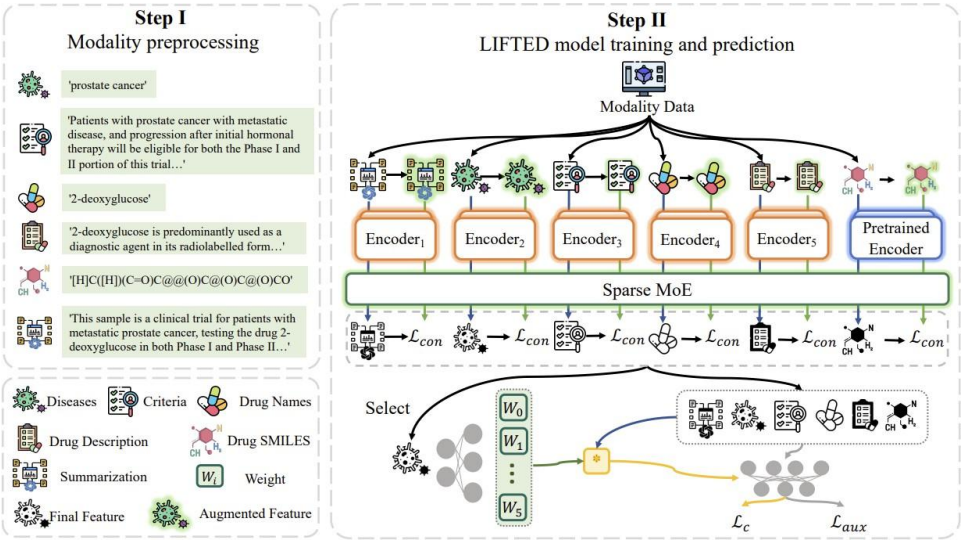

Figure 4: LIFTED Framework (Zheng et al. 2025)

A persistent challenge until this day lies in the increasingly complex datasets, particularly due to the integration of multimodal data (e.g., molecular information, clinical trial documents) and intricate structures such as SMILES, which are represented in graph form.

As a consequence, the integration of conflicting or heterogeneous knowledge into a single model frequently limits the model's performance and extensibility for new available and usable modal data (Mu and Lin 2025).

In response, a comprehensive multimodal framework is proposed to predict the success of clinical trials through the unification of data, both at the linguistic level, via LLMs, and at the functional level, through a Mixture of Experts.

3.3.2 Data−to−Text Generation

At the linguistic level, the process is primarily a data-to-text generation for automatic annotation and expansion, which transforms structured clinical trial data into descriptive natural language for application in multimodal modeling.

For this purpose, the GPT-3.5 Turbo model was used, recognized as one of the key models by Parthasarathy et al. (2024).

Group Part

Structured clinical trial data are converted into natural language to facilitate integration with language-based models. This transformation follows a linearization strategy, in which each tabular entry xi,kx_i,kxi,k is expressed as a column: value pair, using linearization with the formula provided by Zheng et al. (2025):

$$\text{Linearize}(x_{i,k}) = \{c_{i,k} : x_{i,k}\} \quad (1)$$

The linearized entries are embedded within prompts comprising two parts: a prefix describing the schema and a suffix providing text generation instructions. This format enables LLMs to recognize and describe the structure of input data effectively, consistent with methods discussed in Agarwal, Joshi, and Rojkova (2025).

Output quality is maintained through prompt engineering techniques that constrain the output format, remove stochastic variation *temperature* = *0*, and ensure domain relevance. SMILES strings are directly embedded into the natural language prompts, enriching molecular descriptions and providing additional context during generation.

Unlike Zheng et al. (2025), SMILES representations were directly integrated into the CTOD dataset during natural language conversion, enriching molecular descriptions and providing structural context for model generation.

The result of this process is a textual description adapted to the required output type. For example, for the *brief_summary* modality, the model generates a single sentence summarizing the main objective of the trial, while for modalities such as *text_description*, the goal is to create more comprehensive narratives addressing the primary clinical and structural aspects of the study (see Appendix, Table A11).

Given the complexity and size of the data, the pipeline records both the input and the generated output.

Additionally, progress is saved in reusable formats such as JSON (already processed) and pickle (raw data).

3.3.3 Multimodal Embedding and Encoding Strategy

At the core of the system, each data modality ranging from clinical narratives and eligibility criteria to molecular and structured data is transformed into dense vector representations using Transformer-based, domain-specific models. This approach preserves the semantic meaning of each modality (Jurafsky and Martin 2023, pp. 101–131) and enables integrated modeling across diverse inputs, including SMILES strings, drug and disease information, eligibility criteria, enrollment data, and descriptive text.

1. Textual Modality Processing

Text-based inputs such as eligibility criteria, drug names, and disease mentions are tokenized using modality-specific tokenizers: for example, the *bert-base-cased* tokenizer is used for general biomedical text, while *ClinicalBERT* is employed for clinical narratives.

For eligibility criteria, sentences are split into inclusion and exclusion groups using rule-based matching. Each sentence is then encoded using *ClinicalBERT* (`medicalai/ClinicalBERT`), a model pretrained on clinical corpora and configured for sequences up to 512 tokens.

The research utilizes a clinical language model pretrained with 256-token sequences, which has demonstrated superior performance on long clinical text compared to *Bio_ClinicalBERT*, as used by Zheng et al. (2025). This model selection ensures more robust processing of extended eligibility, inclusion, and exclusion criteria, minimizing the risk of truncating important contextual information.

During encoding, a special `[CLS]` token is prepended to each sentence and serves as a semantic summary. The final hidden state of this `[CLS]` token is extracted from the models last encoder layer, resulting in a 768-dimensional embedding per sentence. To represent the full eligibility content of a trial, *mean pooling* is applied separately to the `[CLS]` embeddings of all inclusion

and exclusion sentences, producing two 768-dimensional vectors. These are then concatenated into a single 1536-dimensional vector, which preserves the distinct semantic structure of inclusion and exclusion criteria while capturing the overall eligibility scope.

2. Molecular Representation

SMILES strings (representing molecular structures) are encoded using *ChemBERTa* (`seyonec / ChemBERTa-zinc-base-v1`). The output is passed through a linear projection layer to map it into a shared 768-dimensional embedding space, allowing integration with other modalities while retaining pretrained chemical knowledge.

3. Multimodal Integration and Efficiency

All other modalities including variations of SMILES, drug names, diseases, clinical narratives (tables, descriptions, summaries), and enrollment data are similarly embedded and projected into the same 768-dimensional space. The eligibility vector remains 1536-dimensional. To ensure computational efficiency and maintain consistency during model training, these embeddings are cached after initial encoding and reused as needed.

4. Contextual and Positional Encoding

All textual encodings are contextual, meaning that each tokens embedding depends on its surrounding tokens. To preserve word order, sinusoidal positional encodings are added to token embeddings across all modalities. This ensures that the model understands sequence structure, which is critical for tasks like eligibility matching and protocol understanding (Liu, Kusner, and Blunsom 2020).

3.3.4 Multimodal Fusion Architecture

To enable efficient integration of diverse clinical trial data, the MMCTO framework incorporates a Sparse Mixture-of-Experts (SMoE) architecture composed of four specialized feedforward subnetworks (experts), each trained to capture specific modality-level patterns. A learned gating mechanism dynamically assigns each input vector to the two most relevant experts based on computed relevance scores. This top-k strategy ($k = 2$) activates only two of the four available experts per input, significantly reducing inference cost while maintaining sufficient capacity for complex multimodal reasoning.

The gating network computes expert selection scores through a linear transformation of the input vector, followed by softmax normalization. To prevent overfitting and over-reliance on specific experts, stochastic noise is injected during training a technique shown by Shazeer et al. (2017) to encourage exploration and promote more balanced expert usage. Drug and disease embeddings are also incorporated into the gating input, providing contextual conditioning that allows the model to adaptively route inputs based on the clinical setting and study characteristics.

The outputs of the selected experts are then combined via weighted summation, with weights derived from the gating scores. This context-aware fusion enables dynamic prioritization, such as emphasizing molecular descriptors in oncology studies or exclusion criteria in metabolic trials.

To prevent expert collapse a condition in which only a few experts dominate inference the model includes an expert importance loss, a regularization term that penalizes uneven activation across the expert pool. This encourages balanced usage and improves generalization across tasks and modalities.

Fusion across expert outputs can be implemented via concatenation, weighted averaging, or attention-based mechanisms. In MMCTO, contextual fusion is guided by drug and disease embeddings, improving interpretability and aligning routing behavior with clinically relevant

relationships.

3.3.5 Train and Evaluate

The MMCTO model is trained in a supervised manner using a combination of primary and auxiliary objectives. The main loss function is binary cross-entropy (BCELoss), supported by modality-specific auxiliary losses to promote balanced learning across all data sources and encourage robust, specialized representations (Mu and Lin 2025).

Regularization during MMCTO training includes several strategies to prevent overfitting and promote balanced learning. These consist of dropout with a rate of 0.1 applied within each Transformer encoder, Layer Normalization at key points to stabilize training in the absence of Batch Normalization, and an expert importance loss that penalizes uneven expert activation to ensure balanced utilization across the Sparse Mixture-of-Experts (SMoE) layer.

The model architecture integrates 2 Transformer layers, 8 attention heads, 2048 hidden units, and ReLU activation, combined with the SMoE fusion mechanism to support both concatenation and weighted fusion in the final prediction stage. Optimization is performed using the Adam optimizer with a learning rate of 2×10^{-5}, batch size of 16, and 20 training epochs, con- sistent with established best practices in the literature (Parthasarathy et al. 2024).

All inputs are truncated or padded to fixed, modality-specific lengths, ensuring uniformity during encoding and training. Evaluation is conducted at the end of each epoch on an independent validation set to monitor performance and guide model selection.

To improve robustness and generalization, data augmentation techniques are applied, including stochastic perturbation of embeddings with $\varepsilon = 0.1$ and contrastive learning between original and augmented samples.

During training and inference, expert activation patterns are logged, allowing researchers to visualize which experts and modalities contribute most to each prediction through attention scores and activation weights.

Model performance is evaluated using AUC, average precision (PR), and F1-score. These metrics are computed both per modality and per clinical trial phase (I, II, III) to account for differences in domain complexity. In addition, separate analyses are conducted to assess the individual predictive contribution of each modality (e.g., SMILES, eligibility criteria) and to compare model performance across clinical phases. This provides deeper insight into the effectiveness of the fusion strategy and highlights the computational efficiency benefits achieved through sparse expert activation.

4 Results

To determine the effectiveness of the model in different scenarios, we conducted tests by phases (PhaseI, PhaseII, PhaseIII) in both HINT and CTOD datasets. And in the different modalities individually and combined (our model) in order to test if it made a difference. Hint dataset was divided into training, validation, and testing, while CTOD is divided only into training and validation, in accordance with the framework of (Zheng et al. 2025)

4.1 LIFTED Model Performance Across Datasets and Modalities

Table 1: HINT Dataset: Modality-wise vs. Multimodal (LIFTED) Results

	Phase I			Phase II			Phase III		
	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC
Summarization	85.39	**87.59**	67.78	62.38	56.46	50.08	75.52	52.04	0.42
SMILES	78.78	0	49.51	61.11	3.44	49.03	71.27	29.05	46.51
Description	74.03	5.72	45.22	67.91	75.17	58.11	74.11	82.39	53.2
Criteria	75.14	24.97	45.35	62.57	70.1	50.3	75.04	76.2	52.83
Enrollment	77.49	0	50	51.69	50	50	72.56	**84.1**	50
Diseases	80.54	87.09	56.99	57.44	51.34	44.31	74.19	51.1	53.63
Drugs	78.87	64.39	52.61	62.06	32.29	46.69	77.04	74.66	**57.34**
All (Lifted)	**86.88**	87.5	**69.14**	**68.22**	**75.98**	**58.74**	**77.75**	80.27	56.28

Table 2: CTOD Dataset: Modality-wise vs. Multimodal (LIFTED) Results

	Phase I			Phase II			Phase III		
	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC
Summarization	53.81	51.06	36.61	76.42	77.64	47.23	82.87	86.99	50.2
SMILES	78.89	2	50.66	78.95	0	50.84	83.38	90.24	53.61
Description	79.28	74.18	47.59	80.41	28.36	54.21	83.69	0.46	53.61
Criteria	78.53	39.55	51.66	77.82	25.06	49.52	83.2	0.46	51.47
Enrollment	64.03	3.64	20.52	62.2	4.89	17.77	68.74	0	52.65
Diseases	73.49	55.52	38.19	79.79	65.62	52.07	81.52	83.02	52.45
Drugs	79.15	58.02	50.69	79.37	47.77	52.35	83.2	35.27	51.85
All (Lifted)	**91.58**	**86.01**	**77.52**	**85.57**	**82.18**	**63.31**	**91.15**	**90.3**	**72.91**

The LIFTED model ("ALL(Lifted)") presents, in general, a superior and more robust performance compared to individual modalities (Summarization, SMILES, Description, Criteria, Enrollment, Diseases, Drugs) in the two datasets analyzed, HINT and CTOD.

This advantage is especially evident when compensating for the weak performance of some isolated modalities, which sometimes have low F1 or uninformative ROC values.

The multimodal combination allows achieving higher F1-scores and better discrimination capacity (ROC), showing that integrating different sources of information is essential for more reliable predictions.

In the CTOD dataset, the LIFTED model demonstrates strong and consistent performance across all clinical trial phases. Similarly, in the HINT dataset, multimodal fusion leads to performance improvements, though with greater variability across results.

Nonetheless, the observed gains in both datasets underscore the effectiveness of a multimodal approach in enhancing predictive accuracy across varying clinical trial contexts.

4.2 LIFTED Model Performance Against Baselines

The LITFED model, evaluated with the PyTrial framework, demonstrated superior performance compared to baseline models (HINT, LR, MLP, and XGB) in the HINT and CTOD datasets.

Table 3: Baseline Model Performance vs. LIFTED on HINT

	Phase I			Phase II			Phase III		
	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC
HINT	79.7	87.1	54.7	61.2	74.4	52.7	74.8	83.4	56.1
LR	78.1	87.4	51.7	60.3	74.5	48.5	72.4	**83.9**	50.5
MLP	77.6	87.2	51.8	62.1	74.5	52.3	71.7	83.5	49.6
XGB	77.4	87.2	48.4	58.7	74.3	52.7	72.1	83.5	51.3
All (Lifted)	**86.88**	**87.5**	**69.14**	**68.22**	**75.98**	**58.74**	**77.75**	80.27	**56.28**

Table 4: Baseline Model Performance vs. LIFTED on CTOD

	Phase I			Phase II			Phase III		
	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC
HINT	83.1	84.7	66.4	77.2	80.5	58.4	83.0	85.4	67.5
LR	85.6	83.9	70.1	80.8	80.7	61.0	84.1	85.2	69.3
MLP	86.0	85.5	70.3	78.4	81.9	61.8	85.6	88.3	71.5
XGB	85.8	84.2	74.5	80.2	**82.6**	61.9	85.1	88.9	72.4
All (Lifted)	**91.58**	**86.01**	**77.52**	**85.57**	82.18	**63.31**	**91.15**	**90.3**	**72.91**

It stood out especially in Precision (PR) and Area under the ROC Curve (ROC AUC) metrics, indicating greater discriminative capacity and precision in identifying positive outcomes.

In the HINT dataset, the model presented the best PR and ROC values across all phases of clinical trials. Although the F1-score was competitive in Phases I and II, in Phase III other models, such as HINT and LR, recorded slightly higher F1 values.

Still, the consistent performance in PR and ROC reinforces the robustness of the proposed approach even in contexts with greater variability.

In the CTOD dataset, the results obtained by the LIFTED model were even more expressive. The model presented the best results in PR, F1-score, and ROC AUC in most phases, with emphasis on Phase III, where it widely surpassed the comparative models.

A specific exception was observed in Phase II, where the XGB model obtained a marginally close F1-score.

In general, the results show that multimodal integration, combined with contextual fusion mechanisms and modality specialization, provides significant gains in performance, ensuring more precise and stable predictions in different phases of clinical development.

4.3 Ablation Studies

Table 5: Ablation Results for Gating and LLM Components of LIFTED on HINT

	Phase I			Phase II			Phase III		
	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC
LIFTED-gating	83.45	87.11	64.12	66.9	74.37	55.54	74.13	**81.78**	51.7
LIFTED-LLM	82.04	86.12	63.39	64.18	73.92	53.97	74.49	80.97	53.89
All (Lifted)	**86.88**	**87.5**	**69.14**	**68.22**	**75.98**	**58.74**	**77.75**	80.27	**56.28**

Table 6: Ablation Results for Gating and LLM Components of LIFTED on CTOD

	Phase I			Phase II			Phase III		
	PR	F1	ROC	PR	F1	ROC	PR	F1	ROC
LIFTED-gating	88.31	81.64	74.93	81.11	77.23	60.58	89.05	**91.03**	72.59
LIFTED-LLM	87.25	79.88	73.66	79.26	75.91	59.03	**91.5**	90.54	**76.48**
All (Lifted)	**91.58**	**86.01**	**77.52**	**85.57**	**82.18**	**63.31**	91.15	90.3	72.91

To evaluate the contribution of specific components of LIFTED architecture, ablation studies were conducted to determine the contribution of each feature modality.

In the "LIFTED-gating" variant, the gating weighting mechanism was altered to incorporate

all input modalities, in contrast to the original configuration, which exclusively uses drug and disease representations as control signals.

The comparative analysis reveals that the original, more selective gating strategy is not only sufficient but also more effective, since the inclusion of multiple modalities in the gate input did not result in significant performance gains and, in some cases, introduced instability.

In turn, the "LIFTED-LLM" variant, in which the natural language textual information generated by language models (summarization) was removed, evidenced the critical importance of this component.

The elimination of summarization resulted in a consistent degradation in evaluation metrics, reinforcing the central role of synthesized textual information in the overall performance of the model.

Together, these results support the validity of the architectural choices implemented in the complete version of the model, highlighting the relevance of both the conditioned gating strategy and the incorporation of linguistic representations generated by LLMs.

5 Discussion

5.1 Study Limitations and Practical Constraints

A critical limitation of this study is its reliance on complete and high-quality data. The proposed multimodal model integrating ICD codes, clinical trial eligibility criteria, and SMILES-based molecular representations is particularly sensitive to missing values, which can significantly reduce its predictive performance.

Another important limitation comes from the way the data was processed. Structured data in CSV format was converted into JSON using large language models (LLMs), specifically GPT-3.5-turbo. While this method helped in organizing and standardizing the data, it introduced the risk of algorithmic hallucinations, that is, generating incorrect or non-existent information which can affect the accuracy and trustworthiness of the dataset.

The use of GPT-3.5-turbo also involved considerable financial costs, especially when processing large volumes of clinical trial data. Although open-source alternatives like LLaMA were considered as lower-cost options for converting structured data into natural language, their use was not possible due to the large size of the dataset, which exceeded the available processing capacity. This limits the ability to explore more affordable or open-access solutions.

Another constraint is related to the datasets used: HINT and CTOD. Unlike HINT, the CTOD dataset does not include important ADMET data (Absorption, Distribution, Metabolism, Excretion, and Toxicity). As a result, these features could not be used in the final model, even though previous studies such as that by Favita (2025)have shown that ADMET data can improve pre- diction accuracy.

The model also faced challenges due to data imbalance. Both datasets contain significantly more successful clinical trials than failed ones, thus complicating the models ability to learn patterns and make accurate failure predictions (see Appendix, Table A9 and Table A10).

In addition, the model was developed to predict the outcome of interventional drug trials in general, without focusing on any specific disease. Although this general approach makes the model more widely applicable, it may miss important disease-specific factors that influence trial outcomes, reducing the accuracy and detail of the predictions.

Finally, the high complexity and size of the data required significant computational resources. The model was trained and executed using an NVIDIA GeForce RTX 4080 SUPER GPU with 16GB of memory, taking around 7 hours to complete. This makes the method less practical for use in settings with limited computing power.

5.2 Strategic Implications of MMCTO for Key Stakeholders

The MMCTO model offers a major step forward in optimizing clinical trials. By combining ICD codes, eligibility criteria, and molecular SMILES data, it creates a strong predictive tool.

In testing, LIFTED achieved substantial performance improvements over traditional unimodal methods, with average F1-score gains of 48.9, 27.6, and 23.7 points on the HINT dataset for Phases I, II, and III respectively, 45.4, 46.6, and 48.0 points on the CTOD dataset across the same phases.

This improvement is practical. MMCTO could help identify up to 15% of likely-to-fail trials before patient recruitment begins. Since Phase III trials can cost between 11.5 and 52.9 million (Sertkaya et al. 2024), avoiding even one failed trial could save significant resources, especially in complex treatment areas.

MMCTO could also accelerate drug development. By supporting smarter decisions and better recruitment, it may reduce trial duration by 15 to 30%, as demonstrated by similar AI applications in clinical development (McKinsey & Company 2023). Given that the average time from Phase I to launch still stretches to a decade, such reductions could potentially save 1.5 to 3 years in development timelines.

This benefits the entire ecosystem: researchers avoid wasted effort, companies improve outcomes, regulators receive stronger data, and patients gain earlier access to therapies.

While implementation involves upfront investment of approximately 24,000$ to preprocess 100,000 entries using GPT-3.5-turbo, along with infrastructure and staffing these costs may be offset by substantial long-term benefits.

Current limitations, such as missing ADMET data and reliance on proprietary models Mu and Lin (2025), highlight areas for further improvement. With effective stakeholder collaboration and transparent communication about the model's capabilities and constraints (Kappen et al. 2018), MMCTO presents a viable and impactful tool for advancing clinical trial efficiency.

6 Conclusion

This thesis presents the MMCTO model, a multimodal framework developed to predict clinical trial outcomes by integrating ICD codes, eligibility criteria, and SMILES molecular representations. By transforming structured data into natural language using large language

models (LLMs), MMCTO enables effective fusion of diverse data sources, enhancing the models interpretability and predictive power.

Across both the HINT and CTOD datasets, the model demonstrated consistent and robust performance gains when compared to unimodal and baseline approaches. In particular, MMCTO outperformed individual modalities such as SMILES, disease descriptions, and eligibility criteria by leveraging the combined strengths of each through contextual fusion. The model achieved improvements of up to 8.7% in F1-score in the HINT dataset and 6.2% in CTOD, alongside strong results in Precision (PR) and ROC metrics. These results validate the effectiveness of multimodal integration in improving trial outcome prediction.

Ablation studies further confirmed the importance of key architectural choices. Removing LLM-generated summarizations consistently degraded performance, underscoring the value of synthesized natural language in clinical data contexts. The specialized gating mechanism, which selectively leverages drug and disease information for modality fusion, also proved to be more stable and effective than broader alternatives.

In practical terms, MMCTO could help identify 10–15% of likely-to-fail trials before patient recruitment begins. Given that Phase III trials can cost tens of millions of dollars, this early insight has the potential to prevent significant resource loss. Furthermore, by enabling more targeted trial design and patient selection, the model may reduce the typical 12–15 year drug development timeline by 15–30%, accelerating delivery by up to 3.75 years.

These benefits extend across the entire clinical research ecosystem helping researchers improve efficiency, enabling companies to reduce costs and risk, and giving patients faster access to new therapies.

Nonetheless, challenges remain. The models performance depends on the availability of high-quality and diverse input data. Current limitations include the absence of certain pharmacological features (e.g., ADMET data) and the reliance on proprietary language models, which may

affect reproducibility and scalability. Computational demands, including multi-hour training times on high-end GPUs, may also limit accessibility in some environments.

Future work should focus on integrating additional biomedical data sources, transitioning to open-source LLMs, and adapting the model to specific therapeutic areas. These directions will further strengthen MMCTOs relevance and robustness in real-world clinical development scenarios.

In summary, MMCTO offers a promising, data-driven solution to many of the challenges in clinical trial prediction. Through its use of multimodal integration and contextual language modeling, it stands as a valuable tool for improving the accuracy, efficiency, and impact of clinical research and drug development.

7 SHAP

7.1 Introduction

Machine learning models have often been described as black boxes, meaning that we can easily identify their inputs and outputs, and even try to extrapolate the processes through which the inputs became outputs.

In some more primitive examples of machine learning models such as linear regressions, it may be quite easy to accurately predict the inner workings of a model.

However, with advances to the field of machine learning and even the introduction of Artificial Intelligence, models have developed into a complex conjuncture of computing techniques, creating a significant challenge to interpret these techniques by only analyzing the transformation of inputs to outputs, therefore the analogy of the black box, an organism incomprehensible from an outside perspective.

The need to comprehend these models became clear with machine learning's rapid expansion to multiple domains of our lives. While blind faith in these systems might be harmless in some scenarios, for example a movie recommender, it is unacceptable in other use cases of machine learning such as the focus of our thesis, clinical trials involving public health.

Hoping to offer clarity to researchers, shareholders or even people skeptical of computing processes performed by unsupervised models, a field of research emerged in Machine Learning: Interpretability models.

Simply put, "The need for interpretability arises from an incompleteness in problem formalization (Doshi-Velez and Kim 2017), meaning that machine learning models predictions can sometimes present an incomplete answer to the initial premise. Interpretability models set out to explain how models arrive at predictions. This approach has many benefits not only limited to providing context on how data was transformed, which sometimes is more relevant than the output itself, but also understanding the model helps people trust its output and can also help

prevent errors or biases in computation.

Given the complex nature of the model developed for this thesis and the opportunity to identify the importance of each variable to the model's prediction, an interpretability model was warranted.

7.2 Literature Review

The main source for this research Molnar (2022) book Interpretable Machine Learning. It offers a comprehensive overview of interpretability techniques, mainly distinguishing interpretability approaches into two categories: interpretability by design and post-hoc interpretability.

7.2.1 Interpretability by Design

Interpretability by design refers to models which are inherently interpretable and utilized to in turn interpret other models that do not possess this characteristic. These classifications are not absolute, although Molnar presents a framework that facilitates differentiating among various degrees of interpretability.

At one end of the spectrum, some models can be considered entirely interpretable, like a linear regression with a very limited number of coefficients. This standard isn't very realistic and therefore only applicable in very few cases with extremely simplistic models.

Which leads us to the next consideration: models with partial interpretability. Using the example of linear regression, a large model with too many coefficients to completely interpret can nonetheless provide localized insights by analyzing individual coefficients to determine how each variable impacts the forecast of the entire model.

This approach is encouraging as any model may be interpreted, if not as a whole at least partially with variables capable of supplying useful information. Lastly, a model can be interpreted by its predictions. Molnar offers a good example in the form of a prediction from a decision tree algorithm, where the output is the decision list that resulted in the final prediction.

7.2.2 Post hoc Interpretability

The other classification mentioned by Molnar, post-hoc interpretability, an analysis performed on a model already trained, is divided between two types: model-agnostic or model-specific.

Model-agnostic methods work by the SIPA principle: sample from the data, perform an intervention on the data, get the predictions for the manipulated data, and aggregate the results- agnostic post-hoc methods (Scholbeck et al. 2020).

As an example, permutation feature importance, where the method samples data, intervenes by permuting a feature, obtains predictions from the model, and then aggregates results by comparing the model's performance on the manipulated data to its original performance. The comparison of performance with the permuted feature against the unchanged feature provides aggrandized outcomes resulting in the model's performance evaluation. Model agnostic features are independent from the model's features like its coefficients or weights, making this method a strong benchmark for interpretability analysis.

The exchange between the model and its interpretation leads to additional processes outside the model, since the separation between interpretation and model training introduces an additional conceptual layer to the machine learning pipeline.

Some examples of these methods are Ceteris Paribus Plots, Individual Conditional Expectation (ICE) Curves, Local Surrogate Models (LIME), Shapley Values, Partial Dependence Plot (PDP) to name a few.

Model-specific approaches to interpretability machine learning are tailored to specific model types and therefore not universally applicable.

These methods are especially useful in the context of Neural Networks, given the complexity associated with deep learning and the numerous layers of these networks, making it virtually impossible for humans to trace the decision-making process, creating the need for models specialized in interpreting the architecture of a neural network and its predictions.

7.3 Model Framework

With the intention of analyzing the impact of each feature in the MMCTO model individually, a model was developed using XGBoost as the training and testing algorithm, while SHAP (Shapley Additive Explanations) was applied to interpret the results.

SHAP is a local model-agnostic post hoc method that explains a data points prediction as the sum of feature effects, while XGBoost (Extreme Gradient Boosting) is an ensemble of decision trees which are built sequentially using the gradient boosting framework.

The XGB model provides the prediction layer while SHAP method explores the interpretability for each prediction. To achieve this goal, the data relevant for this approach needed to be curated and transformed into a slightly more homogenized representation.

Since the LIFTED model encompasses features both categorical and numerical there was a need to process this data to enable any analysis. The trials Phase, Condition, Drug and Enrollment features were chosen to integrate this analysis while the model's final predictions for each trial were reflected on the Succes/Failure column.

Categorical fields such as Drug, Conditions, and Phases had their missing values filled with the string "unknown" to maintain data consistency. Using TF-IDF Vectorization the features Drug and Condition were converted into a numerical representation in the form of a vector, while the textual feature Phase was translated to ordinal integers to represent each of the 3 Phases, using a Label Encoder.

Finally, the effort to merge all these features in a single unified representation that can be used for model training resulted in a matrix comprised of the labeled Phase data, the Drug and Condition features vectorized and the numerical values for each trials Enrollment value.

The target feature, Predictions, remained unchanged given, and will be the benchmark for the supervised learning process of this model.

This matrix was then split into training and testing sets using an 80/20 ratio, and the training data was fitted in the XGBoost classifier.

In the evaluation phase, the models performance is assessed using the classification report function from scikit−learn, providing metrics for each class: precision, recall, F1 score, and support represented in table 8.

Table 8: Classification Metrics of XGBoost Model

Class	Precision	Recall	F1-Score	Support
Accuracy	0.87	0.87	0.87	3080
Macro Avg	0.87	0.81	0.83	3080
Weighted Avg	0.87	0.87	0.87	3080

7.4 SHAP Interpretation

In the final stage of the model, SHAP (SHapley Additive exPlanations) was employed to interpret the predictions made by the XGBoost classifier against the MMCTO predictions.

As mentioned, SHAP works by measuring each feature's impact on any particular prediction, effectively breaking down the contribution of each input feature to the model's output.

In the SHAP plot (see Appendix, Table tab 7) features were then color-coded, with blue representing a low impact on the prediction and red a high impact, and their positive or negative contribution to the prediction based on the position in the X axis, negative values leading to an incorrect classification for a prediction and positive values to a correct one.

The Enrollment and the labeled Phases features distinguished themselves from the rest, on their overall impact on the model's predictions. However, this attempt to interpret the significance of each variable is far from perfect as described in the following chapter.

7.5 Discussion

7.5.1 Limitations

As we can observe in the SHAP plot most features had both negative and positive effects, making it hard to evaluate if each feature represented a positive or negative influence on the prediction capabilities of the model. The only conclusive interpretation was the impact each feature had of absolute value on the model.

Another aspect in which this approach is lacking is in the data considered. Some features such as healthy were considered extremely influential for the model only for their recurrence in each trial, thanks to a curation error that wasn't solved due to time constraints.

Due to the complexity involved in translating the SMILES data, it was not included in this analysis. The application of the SHAP methodology represents a valuable approach for interpreting the model.

Its effectiveness could be further enhanced by addressing the aforementioned challenges specifically, by curating categorical data to eliminate irrelevant features and successfully integrating SMILES molecular representations into the data matrix.

7.5.2 Stakeholders

As mentioned in the introduction, people often consider machine learning models to be black boxes since predictions are often without apparent or understandable reasoning.

This lack of transparency is problematic when millions of dollars are reliant on those predictions. The SHAP interpretation presented even with the underline limitations may help to solve this problem by offering insights into how each input feature determines a model's prediction in a straightforward, interpretable manner.

During the model execution, stakeholders can assess which trial features such as trial phase, enrollment size, or medical condition being treated have the strongest impact on the predicted outcome.

Another advantage of this model is the ability to analyze where errors might occur in the analysis of an investment scenario, which can prevent huge losses for all involved. Such qualitative evidence ensures stakeholders that the model is aligned with clinical and business logic which enables them to increase their reliance on analytics as a justification and resource for their investment decisions.

Furthermore, this level of interpretability opens the door for more collaborative decision-making between data scientists, clinical experts, and financial stakeholders, creating a shared interpretation through which the models forecasts can be evaluated.

All in all, stakeholders are now in possession of powerful and transparent decision-support tools that can not only aid in taking decisions as well as explaining these decisions to other interested parties.

7.6 Conclusion

SHAP in this research context has served as an interpretability technique and proved to be an excellent add-on to the model's predictive capabilities. SHAP's value lies in its ability to calculate and visualize the contribution of each feature, thereby providing greater transparency and supporting decision-making.

The execution of this method faces some limitations regarding feature selection and data representation, such as the keyword echo chamber effect and missing the SMILES molecular structure codes but is still able to demonstrate which clinical trial features are most salient for model behavior however it lays groundwork for future improvements.

Such interpretability is not bound to offer technical insight but rather helps the other stakeholders build confidence in the model by validating its logic and reasoning through expert knowledge while exposing the model's strengths and weaknesses and aeras of improvement. SHAP, with its improved detailing of the input data, has the potential to offer powerful interpretability for bridging the gap between machine learning predictions and clinical or financial decisions.

In the era of increased application of machine learning in sensitive fields like health care and its associated investment, SHAP interpretability for features guarantees that models remain reliable, understandable and aligned with the stakeholders and researchers' goals.

References

Agarwal, Bhavik, Ishan Joshi, and Viktoria Rojkova. 2025. *Think Inside the JSON: Reinforcement Strategy for Strict LLM Schema Adherence.* ArXiv preprint arXiv:2502.14905. https://arxiv.org/abs/2502.14905.
.

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, et al. 2000. "Gene Ontology: Tool for the unification of biology." *Nature Genetics* 25 (1): 25–29..

BIO, Informa Pharma Intelligence, and QLS Advisors. 2021. *Clinical Development Success Rates and Contributing Factors 20112020.* Accessed: 2025-05-20. https://go.bio.org/rs/490-EHZ-999/images/ClinicalDevelopmentSuccessRates2011_2020.pdf.

Bodenreider, Olivier. 2004. "The Unified Medical Language System (UMLS): Integrating biomedical terminology." *Nucleic Acids Research* 32 (suppl$_1$): D267–D270.

Bordes, Antoine, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. "Translating embeddings for modeling multi-relational data." In *Advances in Neural Information Processing Systems,* 26:2787–2795.

ChEMBL. n.d. "ChEMBL Database." Accessed: 2025-05-17. https://www.ebi.ac.uk/chembl/.

Chopra, Hitesh, Annu, Dong K. Shin, Kavita Munjal, Priyanka, Kuldeep Dhama, and Talha B. Emran. 2023. “Revolutionizing Clinical Trials: The Role of AI in Accelerating Medical Breakthroughs.” *International Journal of Surgery* 109, no. 12 (December): 4211–4220. https://doi.org/10.1097/JS9.0000000000000705.

Clinical Trials Transformation Initiative (CTTI). n.d. *Aggregate Analysis of ClinicalTrials.gov (AACT) Database.* Accessed: 2025-05-17. https://aact.ctti-clinicaltrials.org/.

ClinicalTrials.gov. n.d. “ClinicalTrials.gov.” Accessed: 2025-05-17. https://clinicaltrials.gov/about-site/about-ctg.

Doe, John, and Jane Smith, eds. 2024. *LLM and Generative AI for Healthcare.* See page VII, 1-35. Springer.

Doshi-Velez, Finale, and Been Kim. 2017. “Towards a rigorous science of interpretable machine learning.” *arXiv preprint arXiv:1702.08608.*

Douze, Matthijs, Jeff Johnson, and Hervé Jégou. 2024. “FAISS: A Library for Efficient Similarity Search.” *arXiv preprint arXiv:1702.08734,* https://arxiv.org/abs/1702.08734.

Es, Shashi, Gustavo Candido Ramos, Parth Jaggi Sangha, Gerard de Melo, and Shafiq R. Joty. 2023. “RAGAS: Automated Evaluation of Retrieval-Augmented Generation.” *arXiv preprint,* arXiv: 2309.15217. https://arxiv.org/abs/2309.15217.

Favita, Sara Sofia Almeida. 2025. “Predictive Modeling for the Trial Completion: Assessing the Phase Success A What-If Scenario Approach on the Enrollment.” A Work Project presented in partial fulfillment of the requirements for the Masters degree in Business Analytics. Master’s thesis, Nova School of Business and Economics. http://hdl.handle.net/10362/181577.

Fedus, William, Barret Zoph, and Noam Shazeer. 2022. “Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity.” *Journal of Machine Learning Research* 23 (120): 1–39.

Fu, Tianfan, Kexin Huang, Cao Xiao, Lucas M. Glass, and Jimeng Sun. 2022. *HINT: Hierarchical Interaction Network for Trial Outcome Prediction Leveraging Web Data.* arXiv: 2102.04252 [cs.CY]. https://arxiv.org/abs/2102.04252.

Gao, Chufan. 2024. *Clinical Trial Outcomes (CTO) Dataset.* https://huggingface.co/datasets/chufangao/CTO. Accessed: 2025-05-17.

Gao, Chufan, Jathurshan Pradeepkumar, Trisha Das, Shivashankar Thati, and Jimeng Sun. 2025. *Automatically Labeling Clinical Trial Outcomes: A Large-Scale Benchmark for Drug Development.* ArXiv preprint arXiv:2406.10292. https://arxiv.org/abs/2406.10292.

Gayvert, Kaitlyn M., Neel S. Madhukar, and Olivier Elemento. 2016. “A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials.” *Cell Chemical Biology* 23 (10): 1294–1301. https://www.sciencedirect.com/science/article/pii/S2451945616302914.

Gupta, Ananya, Li Chen, and Mateo Alvarez. 2024. “A Unified Framework for Clinical Trial Outcome Prediction Using Multimodal Learning.” *Journal of Biomedical Informatics* 138:104321.

Hansch, Corwin, and Toshio Fujita. 1964. “-- Analysis. A Method for the Correlation of Biological Activity and Chemical Structure.” *Journal of the American Chemical Society* 86 (8): 1616–1626. https://pubs.acs.org/doi/10.1021/ja01062a035.

Himmelstein, Daniel S, Alex Lizee, Connor Hessler, Leon Brueggeman, Spencer Chen, Dane Hadley, Alexander Green, Pouya Khankhanian, and Sergio E. Baranzini. 2017. “Systematic integration of biomedical knowledge prioritizes drugs for repurposing.” *eLife* 6:e26726. https://elifesciences.org/articles/26726.

Holley, Kerrie, and Manish Mathur. 2024. *LLMs and Generative AI for Healthcare: The Next Frontier.* Sebastopol, CA: OReilly Media. https://www.amazon.com/LLMs-Generative-AI-Healthcare-Frontier/dp/1098160924.

Jassal, Bijay, Lisa Matthews, Guillermo Viteri, et al. 2020. "The Reactome pathway knowledgebase." *Nucleic Acids Research* 48 (D1): D498–D503.

Jin, Qiao, Zifeng Wang, Charalampos S. Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. 2023. *Matching Patients to Clinical Trials with Large Language Models.* ArXiv preprint arXiv:2307.15051. https://arxiv.org/abs/2307.15051.

Johnson, Khari. 2023. "ChatGPT Can Help Doctorsand Hurt Patients." *WIRED* (April). https://www.wired.com/story/chatgpt-can-help-doctors-and-hurt-patients/.

Jurafsky, Daniel, and James H. Martin. 2023. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* 3rd ed. Draft version, accessed online. See pages 101–131, 203-220. Stanford University. https://web.stanford.edu/~jurafsky/slp3/.

Kanehisa, Minoru, et al. 2017. "KEGG: New perspectives on genomes, pathways, diseases and drugs." *Nucleic Acids Research* 45 (D1): D353–D361.

Kappen, Teus H, Wilton A van Klei, Leo van Wolfswinkel, Cor J Kalkman, Yvonne Vergouwe, and Karel GM Moons. 2018. "Evaluating the impact of prediction models: lessons learned, challenges, and recommendations." *Diagnostic and Prognostic Research* 2 (1): 11.

Kim, Sunghwan, et al. 2021. "PubChem in 2021: New data content and improved web interfaces." *Nucleic Acids Research* 49 (D1): D1388–D1395.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. “ImageNet classification with deep convolutional neural networks.” In *Advances in Neural Information Processing Systems,* 25:1097–1105.

Lamb, Justin, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, et al. 2006. “The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.” *Science* 313 (5795): 1929–1935. https://www.science.org/doi/10.1126/science.1132939.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining.” *Bioinformatics* 36 (4): 1234–1240. https://academic.oup.com/bioinformatics/article/36/4/1234/5566506.

———. 2023. “Applications of Transformer Models in Biomedical Natural Language Processing: A Review.” *Briefings in Bioinformatics* 24 (1): bbac535. https://doi.org/10.1093/bib/bbac535.

Liu, Qi, Matt J. Kusner, and Phil Blunsom. 2020. *A Survey on Contextual Embeddings.* ArXiv preprint arXiv:2003.07278. https://arxiv.org/abs/2003.07278.

Lo, Andrew W., Charlotte J. Pan, Alex B. Siah, and Danying Xiao. 2019. “Can Machine Learning Improve Prediction of FDA Drug Approvals?” *Clinical Pharmacology & Therapeutics* 106 (4): 685–695. https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1002/cpt.1510.

McKinsey & Company. 2023. *How artificial intelligence can power clinical development,* November. https://www.mckinsey.com/industries/life- sciences/our- insights/how- artificial - intelligence-can-power-clinical-development.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. “Efficient estimation of word representations in vector space.” *arXiv preprint arXiv:1301.3781.*

Minaee, Shervin, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. *Large Language Models: A Survey.* ArXiv preprint arXiv:2402.06196. https://arxiv.org/abs/2402.06196.

Miotto, Riccardo, et al. 2016. "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records." *Scientific Reports* 6:26094.

Molnar, Christoph. 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* 2nd ed. https://christophm.github.io/interpretable-ml-book/.

Mu, Siyuan, and Sen Lin. 2025. *A Comprehensive Survey of Mixture-of-Experts: Algorithms, Theory, and Applications.* ArXiv preprint arXiv:2503.07137. https://arxiv.org/abs/2503.07137.

Ong, Jing, Chun Wai Wang, Chongyang Wang, et al. 2024. "Potential for GPT Technology to Optimize Future Clinical Decision-Making Using Retrieval-Augmented Generation." *Annals of Biomedical Engineering* 52:1115–1118. https://doi.org/10.1007/s10439-023-03327-6.

OpenAI. n.d. "ChatGPT Pricing." Accessed May 20, 2025. https://openai.com/es-ES/chatgpt/pricing/.

Parthasarathy, Venkatesh Balavadhani, Ahtsham Zafar, Aafaq Khan, and Arsalan Shahid. 2024. *The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities.* ArXiv preprint arXiv:2408.13296. https://arxiv.org/abs/2408.13296.

PubChem. n.d. "PubChem Database." Accessed: 2025-05-17. https://pubchem.ncbi.nlm.nih.gov.

PubChemPy API. n.d. "PubChemPy: A Python Wrapper for the PubChem PUG REST API." Accessed: 2025-05-17. https://pubchempy.readthedocs.io/en/latest/.

Qi, Youran, and Qi Tang. 2019. “Predicting Phase 3 Clinical Trial Results by Modeling Phase 2 Clinical Trial Subject Level Data Using Deep Learning.” In *Proceedings of the 4th Machine Learning for Healthcare Conference,* edited by Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, 106:288–303. Proceedings of Machine Learning Research. PMLR, September. https://proceedings.mlr.press/v106/qi19a.html.

Ratan, Ujjwal. 2023. *Applied Machine Learning for Healthcare and Life Sciences Using AWS.* Birmingham, UK: Packt Publishing. https://www.amazon.com/dp/1804610216.

Sahoo, Deepak, Hai Pham, and Steven CH Hoi. 2021. “A Survey on Ensemble Learning for Data Stream Classification.” *ACM Computing Surveys (CSUR)* 54 (3): 1–36. https://dl.acm.org/doi/10.1145/3439724.

Scholbeck, C.A., C. Molnar, B. Bischl, and G. Casalicchio. 2020. “Sampling, Intervention, Prediction, and Aggregation (SIPA): A unified framework for model-agnostic interpretability methods.” *arXiv preprint arXiv:2008.05147.*

Sertkaya, Aylin, et al. 2016. “Key cost drivers of pharmaceutical clinical trials in the United States.” *Clinical Trials* 13 (2): 117–126.

Sertkaya, Aylin, Trinidad Beleche, Amber Jessup, and Benjamin D. Sommers. 2024. “Costs of Drug Development and Research and Development Intensity in the US, 20002018.” *JAMA Network Open* 7, no. 6 (June): e2415445.

Sharma, Rashmi, and Himani Garg. 2025. *Advancing Healthcare with Retrieval Augmented Generation (RAG) and Large Context Models (LCM): A Comparative Study.* Ajay Kumar Garg Engineering College, Ghaziabad, U.P. India. https://www.akgec.ac.in/wp-content/uploads/2025/02/3-Dr.-Rashmi-Sharma-and-Dr.-Himani-Garg_compressed.pdf.

Shazeer, Noam, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. “Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer.” In *Proceedings of the International Conference on Learning Representations (ICLR).* https://arxiv.org/abs/1701.06538.

Shickel, Benjamin, et al. 2018. “Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis.” *IEEE Journal of Biomedical and Health Informatics* 22 (5): 1589–1604.

Singh, Harshit, Pedro Cuenca, and Harrison Chase. 2024. *LangChain: Building Applications with LLMs through Composability.* ArXiv preprint arXiv:2305.03983. https://arxiv.org/abs/2305.03983.

Statista. 2023. *Pharmaceutical Industry Worldwide - Statistics & Facts.* https://www.statista.com/topics/1764/pharmaceutical-industry/. Accessed: 2025-05-21.

Tropsha, Alexander. 2010. “Best Practices for QSAR Model Development, Validation, and Exploitation.” *Molecular Informatics* 29 (67): 476–488. https://doi.org/10.1002/minf.201000061.

U.S. Food and Drug Administration. 2019. *Rare Diseases: Common Issues in Drug Development.* https://www.fda.gov/media/119757/download. Accessed: 2025-05-20.

———. 2025. *Drug Development and Review Definitions.* Accessed: 2025-05-19. https://www.fda.gov/drugs/investigational-new-drug-ind-application/drug-development-and-review-definitions.

U.S. National Library of Medicine. n.d. “Clinical Tables Search Service (CTSS).” Accessed: 2025-05-17. https://clinicaltables.nlm.nih.gov.

Unlu, Ozan, Jiyeon Shin, Charlotte J. Mailly, Michael F. Oates, Michela R. Tucci, Matthew Varugheese, Kavishwar Wagholikar, et al. 2024. “Retrieval-Augmented GenerationEnabled GPT-4 for Clinical Trial Screening.” *NEJM AI* 1 (7): AIoa2400181. https://ai.nejm.org/doi/full/10.1056/AIoa2400181.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, et al. 2017. “Attention is all you need.” In *Advances in Neural Information Processing Systems,* 30:5998–6008.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. *Attention Is All You Need.* arXiv: 1706.03762 [cs.CL]. https://arxiv.org/abs/1706.03762.

Wang, C., J. Ong, C. Wang, et al. 2024. “Potential for GPT Technology to Optimize Future Clinical Decision-Making Using Retrieval-Augmented Generation.” *Annals of Biomedical Engineering* 52:1115–1118.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. “RNA-Seq: A Revolutionary Tool for Transcriptomics.” *Nature Reviews Genetics* 10 (1): 57–63. https://doi.org/10.1038/nrg2484.

Weininger, David. 1988. “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.” *Journal of Chemical Information and Computer Sciences* 28 (1): 31–36. https://pubs.acs.org/doi/10.1021/ci00057a005.

Wishart, David S., et al. 2018. “DrugBank 5.0: A major update to the DrugBank database for 2018.” *Nucleic Acids Research* 46 (D1): D1074–D1082.

Woodcock, Janet. 2020. *Improving Clinical Trial Efficiency Through Advanced Data Analytics.* Speech transcript or article, U.S. Food and Drug Administration. Accessed: 2025-05-21. https : / / www . fda . gov / news - events / speeches - fda - officials / improving - clinical - trial - efficiency-through-advanced-data-analytics-10272020.

World Health Organization. 2019. "International Statistical Classification of Diseases and Related Health Problems (ICD-10)." Accessed: 2025-05-17. https://icd.who.int/browse10/2019/en.

Wu, Yanghan, Yanhui Huang, Junqiang Li, Zhuhong You, Pengwei Hu, and Lin Hu. 2023. "Knowledge graph embedding for profiling the interaction between transcription factors and their target genes." *PLOS Computational Biology* 19 (6): e1011207. https://doi.org/10.1371/journal.pcbi.1011207.

Wu, Zhenqin, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Kyle Leswing, and Vijay S. Pande. 2018. "MoleculeNet: a benchmark for molecular machine learning." *Chemical Science* 9 (2): 513–530. https://pubs.rsc.org/en/content/articlelanding/2018/sc/c7sc02664a.

Zheng, Wenhao, Liaoyaqi Wang, Dongshen Peng, Hongxia Xu, Yun Li, Hongtu Zhu, Tianfan Fu, and Huaxiu Yao. 2025. *Multimodal Clinical Trial Outcome Prediction with Large Language Models.* ArXiv preprint arXiv:2402.06512. https://arxiv.org/abs/2402.06512.

Zhou, Jie, Guanghui Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, and Maosong Sun. 2022. "Graph Neural Networks: A Review of Methods and Applications." *AI Open* 1:57–81. https://doi.org/10.1016/j.aiopen.2021.01.001.

Zitnik, Marinka, Monica Agrawal, and Jure Leskovec. 2018. "Modeling polypharmacy side effects with graph convolutional networks." *Bioinformatics* 34 (13): i457–i466.

Appendix

Table A9: Summary statistics for HINT dataset across phases and splits

Dataset	Phase	Split	# Trials	Unique Drugs	Unique Diseases	# Success	# Failure
HINT	I	Test	1160	329	714	897	263
		Train	5417	1256	2665	4640	777
		Valid	1159	385	754	910	249
	II	Test	1449	442	1123	863	586
		Train	6761	1273	3549	5393	1368
		Valid	1452	470	1134	1073	379
	III	Test	893	341	679	641	252
		Train	4165	897	2251	3601	564
		Valid	894	365	709	726	168

Table A10: Summary statistics for CTOD dataset across phases and splits

Dataset	Phase	Split	# Trials	Unique Drugs	Unique Diseases	# Success	# Failure
CTOD	I	Train	4016	1752	1271	3326	690
		Valid	1006	623	447	820	186
	II	Train	3572	2055	2101	4112	1278
		Valid	1348	754	712	1028	320
	III	Train	3572	1563	1415	3004	568
		Valid	893	565	503	741	152

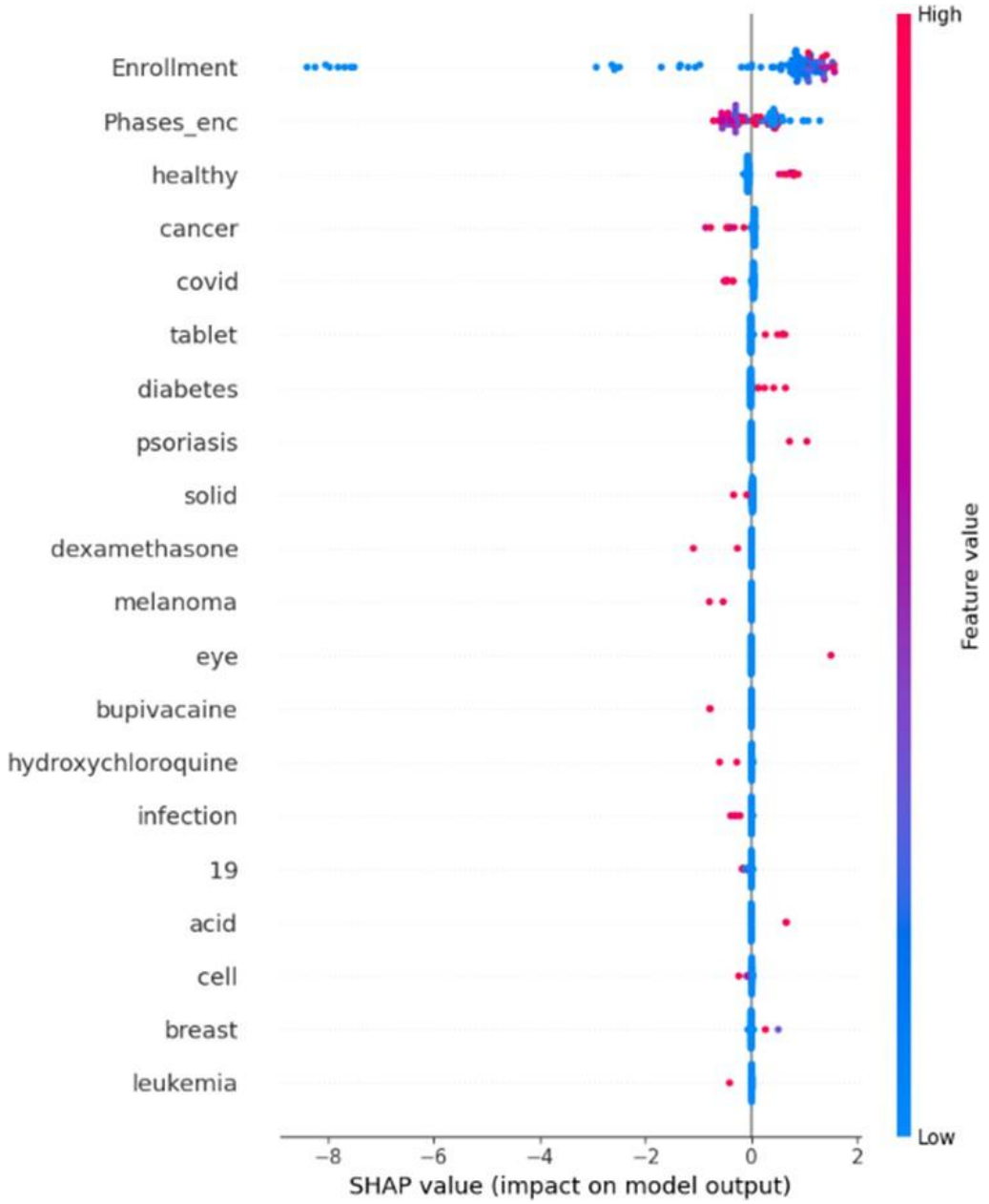

Figure 7: SHAP Scatter Plot