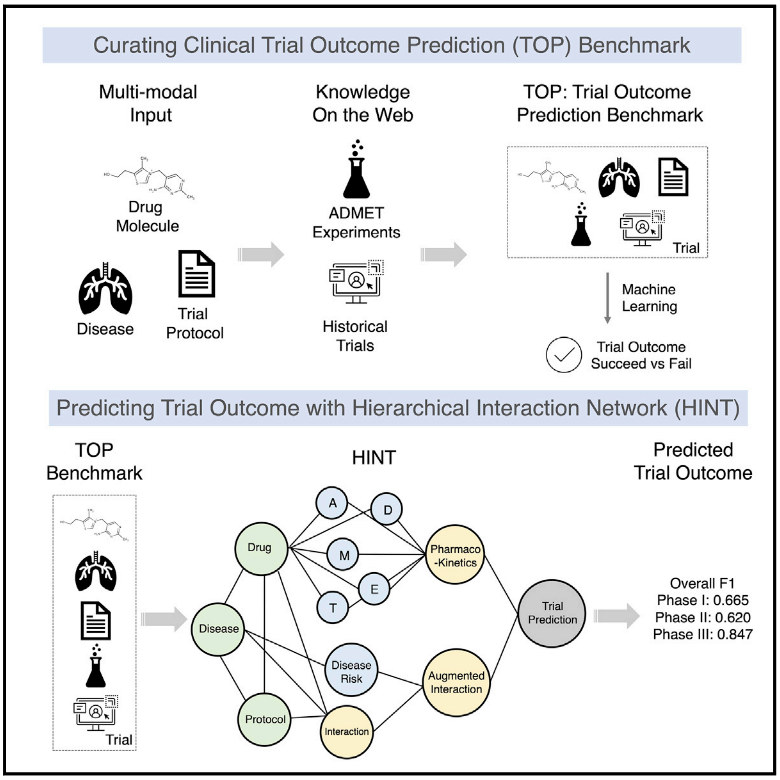
Article

Patterns

HINT: Hierarchical interaction network for clinical-trial-outcome predictions

Graphical abstract



Highlights

- We curate and release benchmark datasets for clinical-trial-outcome predictions
- Hierarchical interaction graph captures interactions among trial components
- HINT simulates the trial process and enables accurate outcome predictions

Authors

Tianfan Fu, Kexin Huang, Cao Xiao, Lucas M. Glass, Jimeng Sun

Correspondence

jimeng@illinois.edu

In brief

Fu. et al. propose a new method for predicting clinical trial outcomes. The proposed method represents multi-modal data and builds a hierarchical interaction graph to predict clinical trial outcomes. A benchmark dataset for clinical-trial-outcome predictions is created and shared. Extensive experiments indicate that the proposed method achieves high accuracy across different trial phases and multiple disease groups.

Fu et al., 2022, Patterns *3*, 100445
April 8, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.patter.2022.100445

CellPress

Patterns

CellPress
OPEN ACCESS

Article

HINT: Hierarchical interaction network for clinical-trial-outcome predictions

Tianfan Fu,[1] Kexin Huang,[2] Cao Xiao,[3] Lucas M. Glass,[4,5] and Jimeng Sun[6,7,*]

[1]Department of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[2]Department of Computer Science, Stanford University, Stanford, CA 94305, USA
[3]Amplitude, San Francisco, CA 94105, USA
[4]Analytics Center of Excellence, IQVIA, Cambridge, MA 02139, USA
[5]Department of Statistics, Temple University, Philadelphia, PA 19122, USA
[6]Computer Science Department and Carle's Illinois College of Medicine, University of Illinois at Urbana-Champaign, Urbana, IL 61820, USA
[7]Lead contact
*Correspondence: jimeng@illinois.edu

https://doi.org/10.1016/j.patter.2022.100445

THE BIGGER PICTURE Deep learning models have shown many successes in modeling biomedical data. Most existing models handle one type of data, while real-world data science applications often have multi-modal datasets with various characteristics and qualities along with domain-specific knowledge. This paper presents a new graph neural network method called the Hierarchical Interaction Network (HINT) that handles complex interaction patterns from multi-modal data for clinical-trial-outcome predictions. HINT offers a general data science methodology for handling complex interconnected datasets with underlying domain knowledge. In particular, HINT's ability to handle different types of input data (e.g., graphs, text, and categorical variables) and missing values can provide valuable algorithm design strategies for broader data science communities. Also, we curated and released a large and labeled benchmark dataset of 17,538 clinical trials for trial outcome predictions.

SUMMARY

Clinical trials are crucial for drug development but often face uncertain outcomes due to safety, efficacy, or patient-recruitment problems. We propose the Hierarchical Interaction Network (HINT) to predict clinical trial outcomes. First, HINT encodes multi-modal data (drug molecule, target disease, trial eligibility criteria) into embeddings. Then, HINT trains knowledge-embedding modules using drug pharmacokinetic and historical trial data. Finally, a hierarchical interaction graph connects all of the embeddings to capture their interactions and predict trial outcomes. HINT was trained and validated on 1,160 phase I trials, 4,449 phase II trials, and 3,436 phase III trials. It obtained 0.665, 0.620, and 0.847 F1 scores on separate test sets of 627 phase I, 1,653 phase II, and 1,140 phase III trials, respectively. HINT significantly outperforms the best baseline method on most metrics. The benchmark dataset and codes are released at https://github.com/futianfan/clinical-trial-outcome-prediction.

INTRODUCTION

A clinical trial is an indispensable step toward developing a new drug. Human participants are tested to respond to the drug (e.g., a drug molecule or drug combinations) for treating target diseases. The global clinical trial market has reached $44.3 billion in 2020 and is expected to grow to $69.3 billion by 2028.[1] The costs of conducting clinical trials are extremely expensive (up to hundreds of millions of dollars[2]), and the time to run a trial takes multiple years, with a low success probability.[3,4] Many factors, such as the inefficacy of the drug, drug safety issues, and poor trial protocol design, can fail a clinical trial.[5] Can we predict the trial outcome in an *in silico* manner? Here, the trial outcome refers to a binary success indicator whether the trial is completed to meet their primary endpoints. The vast amount of historical clinical trial data and massive knowledge bases about passed and failed drugs brings a new opportunity for using machine learning models to tackle the critical question: Can one predict the trial outcome before the trial starts?

Various public data sources can provide vital information for predicting the trial outcome. For example, the ClinicalTrials.gov database (publicly available at https://clinicaltrials.gov/) has 369,700 historical clinical trials, including important information about those trials. In addition, we utilize the standard medical codes of the diseases and their natural language descriptions through the National Institutes of Health website (publicly available at https://clinicaltables.nlm.nih.gov/). The DrugBank database (publicly available at https://www.drugbank.ca/) contains the biochemical description of many drugs, which allows for the computational modeling of drug molecules.[6]

Over the years, there have been early attempts to predict individual components in clinical trials to improve the trial results, including using electroencephalographic (EEG) measures to predict the effect of antidepressant treatments in improving depressive symptoms,[7] optimizing drug toxicity based on drug- and target-property features,[8] and leveraging phase II results to predict phase III trial results.[9] Recently, there has been interest in developing a general method for trial outcome predictions. As an initial attempt, Lo et al.[10] predicted drug approvals for 15 disease groups based on drug and clinical trial features using classical machine learning methods. Despite these initial efforts, several limitations impede the utility of existing trial outcome prediction models, including:

- Lack of benchmark data. Data science progress in any domain needs to be measured on large and accessible benchmark data. Such datasets in clinical trial domains are not available, which severely affects data science efforts on clinical-trial-related research. Such data are available in computer vision, e.g., ImageNet.[11]
- Limited task definition and study scope. Existing works either focus on predicting individual components of trials,[8,12–14] such as patient-trial matching, or only a subset of disease groups.[10] Although these works are helpful for part of the trial design, they do not predict the trial outcome for a broad set of target diseases. We are one of the first to study the general trial outcome prediction problem across different trial phases for many different diseases.
- Limited features used for prediction. Due to their limited task definition and study scope, existing works often only leverage restricted-disease-specific parts, which cannot be generalized for other diseases. These works also ignore the fact that a trial outcome is determined by various factors, including drug safety, treatment efficiency, and trial eligibility criteria. For example, the biomedical knowledge, such as drug molecule structures and historical trial data, for different diseases can be beneficial for modeling trial outcomes.
- Ignoring the complex relations among trial components. Due to the limited data and task scope, existing methods often simplify their predictions by limited input features and rely on classical classification methods (e.g., random forest) that are not explicitly designed for modeling the interaction of different trial components.[8,9,12,13,15] This simplified assumption impedes the prediction performance of the existing works.

Our approach

To provide accurate trial outcome predictions for all trials, we propose the Hierarchical Interaction Network (HINT). The HINT model is trained on a multi-modal dataset, including molecule information of the drugs, the target disease information, the trial eligibility criteria, and biomedical knowledge. HINT first encodes these multi-modal data into latent embedding vectors of the drug molecule, the target disease, and the trial risk, where an imputation module is designed to handle missing data. Next, we train a knowledge-embedding module from external knowledge on pharmacokinetic properties for improving drug embedding. We also train another trial-risk-embedding module using historical trial data for improving trial risk embedding. After that, we present an interaction graph module to connect all of the embeddings to capture various interaction effects from different trial components. Finally, HINT learns a dynamic attentive graph neural network to predict trial outcomes.

Contributions

Our main contributions are summarized as follows:

- Problem: We formally define a model framework for a general clinical-trial-outcome prediction task, which not only models various trial risks, including drug safety, treatment efficiency, and trial recruitment, but also models a wide range of drugs and indications (e.g., diseases). Our model framework can generalize over new trials given the drug molecule, target disease, and trial eligibility criteria (Problem formulation).
- Benchmark: To enable general clinical-trial-outcome predictions, we leverage a comprehensive set of datasets, including drugbank, disease codes, and clinical trial records to curate a Trial Outcome Prediction (TOP) dataset (Benchmark).
- Results: We evaluated HINT against state-of-the-art baselines using real-world data. HINT achieved 0.665, 0.620, and 0.847 F1 scores on phase I-, II-, and III-level predictions, respectively. In addition, HINT achieves statistically significant improvements compared with the best baseline method (Cross-Modal Psuedo-Siamese Network [COMPOSE]).[16] We also conduct an ablation study to evaluate the importance of key components of clinical trials to the prediction power and the effectiveness of the hierarchical formulation of a trial interaction graph. Lastly, we conduct a case study to show the potential real-world impact of HINT by successfully predicting some prominent trial outcomes (results and discussion).
- Method: We design a graph neural network method that explicitly simulates different clinical trial components and their interaction relations for predicting trial outcomes (Method).

Problem formulation

A clinical trial is designed to validate the safety and efficacy of a treatment set toward a target disease set on a patient group defined by the trial eligibility criteria.

Definition 1 (treatment set)

The treatment set includes one or multiple drug candidates, denoted by

$$\mathbb{M} = \{m_1, \ldots, m_{N_m}\}, \qquad \text{(Equation 1)}$$

where $m_1, \ldots, m_{N_m}$ are N_m drug molecules involved in this trial.

Note that we focus on clinical trials that aim at discovering new indications of drug candidates. The trials that do not involve drug molecules, such as surgery techniques and medical devices, are out of this scope and can be considered future work.

Definition 2 (target disease set)

Each trial targets one or more diseases. Suppose there are $N_d \geq 1$ diseases in a trial, we represent the target disease set as

$$\mathbb{D} = \{d_1, \ldots, d_{N_d}\}, \quad \text{(Equation 2)}$$

where $d_1, \ldots, d_{N_d}$ are N_d target diseases. We use d_i to represent the raw information associated with the disease including the disease name, description (text data), and the corresponding diagnosis code (e.g., International Classification of Diseases [ICD] codes[17]).

Each trial has eligibility criteria (in unstructured natural language) that describe criteria for enrolling patients, including participant characteristics such as age, gender, medical history, target disease conditions, and current health status.

Definition 3 (trial eligibility criteria)

The patient group is specified by the trial eligibility criteria. Formally, eligibility criteria consist of a set of inclusion and exclusion criteria for recruiting patients represented by the following:

$$\mathbb{C} = [\mathbf{c}_1^I, \ldots, \mathbf{c}_M^I, \mathbf{c}_1^E, \ldots, \mathbf{c}_N^E], \mathbf{c}_i^{I/E} \text{ is a sentence.} \quad \text{(Equation 3)}$$

M (N) is the number of inclusion (exclusion) criteria in the trial, and $\mathbf{c}_i^I$ ($\mathbf{c}_i^E$) denotes the i-th inclusion (exclusion) criterion. Each criterion **c** is a sentence in unstructured natural language.

Definition 4 (trial outcome)

Trial outcome is a binary label $y \in \{0, 1\}$, where $y = 1$ indicates the trial met their primary endpoints, while 0 means failing to meet with the primary endpoints.

Here, the primary endpoints are the statistical measures to indicate whether the drug candidate works or not. For example, for an antihypertensive drug trial, the primary endpoint can be the percentage of hypertension patients with controlled blood pressure (BP), e.g., systolic BP < 140 mm Hg.

Problem 1 (trial outcome prediction)

The predicted success probability is $\hat{y} \in [0, 1]$. The goal of HINT is to learn a deep neural network model f_θ for predicting the actual trial success status $\hat{y}$:

$$\hat{y} = f_\theta(\mathbb{M}, \mathbb{D}, \mathbb{C}), \quad \text{(Equation 4)}$$

where $\mathbb{M}, \mathbb{D}, \mathbb{C}$ are the treatment set, target disease set, and eligibility criteria, respectively. We focus on predicting the success for a particular phase of the trial. In general, there are three trial phases: phase I tests the toxicity and side effects of the drug, phase II determines the efficacy of the drug (i.e., if the drug works), and phase III focuses on the effectiveness of the drug (i.e., whether the drug is better than the current standard practice). The phase-level prediction determines whether a specific clinical trial study will successfully complete at the phase.

Benchmark

To standardize the clinical-trial-outcome predictions, we create a benchmark dataset for Trial Outcome Prediction named TOP, which incorporate rich data components about clinical trials, including drug, disease, and eligibility criteria. We first describe the data components and then report the processing steps to construct this benchmark dataset.

Benchmark dataset overview

TOP consists of 17,538 clinical trials with 13,880 small-molecule drugs and 5,335 diseases. Out of these trials, 9,999 (57.0%) succeeded (i.e., meeting primary endpoints) and 7,539 (43.0%) failed. For each clinical trial, we produce the following four data items: (1) drug molecule information including Simplified Molecular Input Line Entry System (SMILES) strings and molecular graphs for the drug candidates used in the trials; (2) disease information including ICD-10 codes (disease code), disease description, and disease hierarchy in terms of CCS codes (https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp); (3) trial eligibility criteria are in unstructured natural language and contain inclusion and exclusion criteria; and (4) trial outcome information includes a binary indicator of trial success (1) or failure (0), trial phase, start and end date, sponsor, and trial size (i.e., number of participants).

In addition to the primary clinical trial data, we also provide two auxiliary datasets. One is the pharmacokinetics (PK) dataset, which consists of wet lab experiment results for five important PK tasks (PK properties are absorption, distribution, metabolism, excretion, and toxicity) and the drug SMILES strings, provided in MoleculeNet (available at https://moleculenet.org/datasets-1). Another is the disease risk dataset, which is the historical success rate of the target disease and the disease descriptions, provided at ClinicalTrials.gov.

Tasks. Many tasks can be studied in terms of prediction using TOP. In this paper, we focus on trial primary outcome prediction as a binary classification. Future works can be done for more granular predictions on different types of outcomes such as patient enrollment and expected trial duration.

TOP *benchmark statistics*. For the PK auxiliary dataset, we have 640 drugs for absorption, 1,593 for distribution, 15,020 for metabolism, 15,982 for excretion, and 24,576 for toxicity. For the disease risk auxiliary dataset, we have 16,356 disease combinations and their success rate in the past. In Figure 1, we show the time distribution of some statistics. Table 1 shows statistics of the curated dataset.

Data curation process

We create the TOP benchmark for trial outcome predictions from multiple data sources, including the drug knowledge base, disease code (ICD-10 code), historical clinical trials,[17] and manually curated trial outcome labels. We split learning (training and validation)/test data on January 1, 2014. The earlier trials are used for training and validation, while the later trials are used for tests. We ensure that the completion dates of training/validation data are earlier than the start dates of the test data. The training/validation splits are random, with a ratio of 9:1.

Trial selection criteria. We apply a series of selection filters to ensure the selected trials have high-quality outcome labels as shown in Figure 2. We start with 369,700 raw clinical trial records from ClinicalTrials.gov. (1) We select only interventional trials, and 288.400 trials are left. (2) We select trials that test small-molecule drugs, and 143,400 trials are left. (3) We sample and label a set of 22,300 completed trials. (4) We further select trials with known drug molecule structures and available disease codes, and 17,600 trials are left in TOP. (5) After

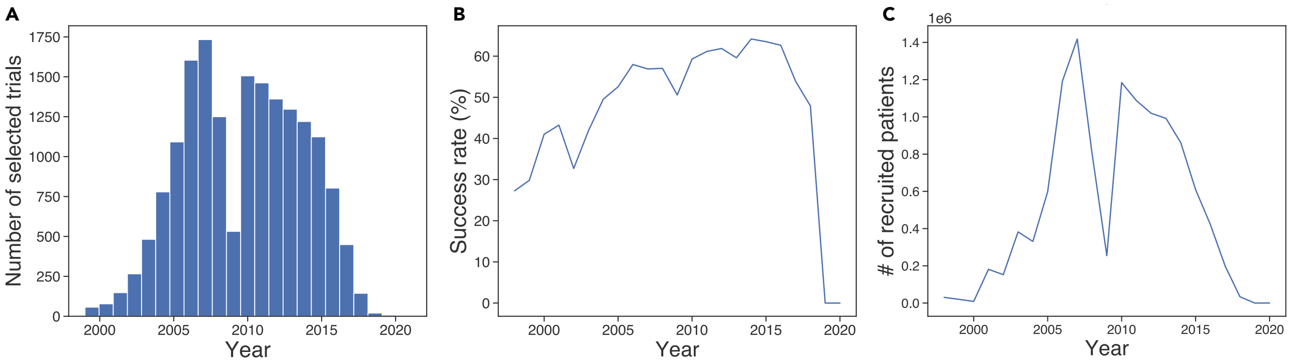


Figure 1. Time distribution

(A) A histogram of time distribution of trials' start dates. The fall in 2009–2010 is due to the financial crisis.[18] The number decreases after 2015 because it usually takes several years to complete a trial, and most of the trials that started after 2015 are not yet complete and thus have not been selected.

(B) The percetange of successful trials over the years. This also drops significantly after 2015 for the same reasons: most of the completed trials that started after 2015 were either stopped early or failed.

(C) The percetange of recruited patients over the years. The shape is similar to the trial number as in (A). Note that the trials that started before 1998 are limited in number and hence not shown in the above panels.

filtering by start/completion date (i.e., ensure train/validation and test have no time overlap), 12,500 trials are used in the experiments.

Each trial in ClinicalTrials.gov is an XML file, and we parse them to obtain all of the variables. In particular, for each trial, we obtain the NCT ID (i.e., identifiers to each clinical trial), disease names, drug molecules, brief titles and summaries, phases, and eligibility criteria.

Data processing and linking. Next, we describe how we process and link the parsed trial data to machine-learning-ready input and output formats:

- Drug molecule data are extracted from ClinicalTrials.gov and linked to the molecule structure (SMILES strings and the molecular graph structures) using DrugBank Database[6] (https://www.drugbank.com).
- Disease data are extracted from ClinicalTrials.gov and linked to ICD-10 codes and disease descriptions using clinicaltables.nlm.nih.gov and then to CCS codes via hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp.
- Trial eligibility criteria are extracted from ClinicalTrials.gov.
- Trial outcome labels (binary labels) are extracted through manual curation by an internal IQVIA team.
- Auxiliary drug PK data include five datasets across the main categories of PK. For absorption, we use the bioavailability dataset provided in Ma et al.'s supplemental information.[19] For distribution, we use the blood-brain-barrier experimental results provided in Adenot et al.'s study.[20]

Table 1. Data statistics

	Trials, #	Drugs, #	Diseases, #	Successes, #	Failures, #
All	17,538	13,880	5,335	9,999	7,539
All (filtered by start/completion date)	12,465	10,026	3,893	7,149	5,316
Neoplasm	4,246	2,456	2,008	1,752	2,494
Respiratory system	1,299	1,736	968	868	431
Digestive system	1,844	1,990	1,558	1,072	772
Nervous system	1,975	2037	1,369	1,171	804
Others	8,174	9,778	4,090	5,136	3,038
Start before 2000	179	144	193	43	136
Start between 2000–2004	1,753	1,092	317	771	982
Start between 2005–2009	6,211	2,358	1,267	3,472	2,739
Start between 2010–2014	6,846	3,277	1,613	4,185	2,661
Start between 2015–2021	2,549	2,987	1,710	1,528	1,021
Phase I	1,787	2,020	1,392	582/77/347	462/39/280
Phase II	6,102	5,610	2,824	1,925/196/918	2,079/249/735
Phase III	4,576	4,727	1,619	2,042/208/854	1,050/136/286

All clinical trial records were available at ClinicalTrials.gov on February 20, 2021. For phases I, II, and III, we show the #train/#validation/#test for both successes and failures. The train/validation and test are time-split by the date January 1, 2014, i.e., the start dates of the test set are after January 1, 2014, while the completion dates of the train/validation set are before January 1, 2014. We do not include the trials that started before and completed after the date in the training data. Train/validation sets are randomly split with a ratio of 9 to 1.

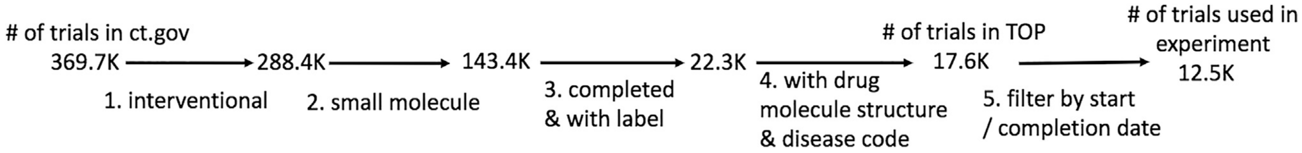


Figure 2. Workflow of data curation

For metabolism, we use the CYP2C19 experiment from Veith et al.'s[21] paper, which is hosted in the PubChem bioassay portal under AID 1851. For excretion, we use the clearance dataset from the eDrug3D database.[22] For toxicity, we use the ToxCast dataset[23] provided by MoleculeNet (https://moleculenet.org/datasets-1). We consider drugs nontoxic when they pass all toxicology assays. ADMET datasets are used to pretrain the drug encoder, as elaborated on in the supplemental information.

RESULTS

This section presents the performance comparison of different machine learning models on the benchmark TOP. It also shows the superior performance of the proposed method HINT on this benchmark.

Experimental setting

Evaluation settings

We consider phase-level evaluation, where we predict the outcome of a single-phase study. Since each phase has different goals (e.g., phase I is for safety, whereas phases II and III are for efficacy), we evaluate phases I, II, and III separately. Data statistics are shown in Table 1. All codes (including data collection and preprocessing, model construction, learning process, and evaluation) are publicly available at https://github.com/futianfan/clinical-trial-outcome-prediction.

Evaluation metrics

We use the following metrics to measure the performance of all methods.

- Precision-recall area under the curve (PR-AUC): PR curves summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds.
- F1: The F1 score is the harmonic mean of precision and recall.
- Area under the receiver operating characteristic curve (ROC-AUC): The ROC curve summarizes the trade-off between the true- and false-positive rates for a predictive model using different probability thresholds.
- p value: We report the results of hypothesis testing in terms of p values to showcase the statistical significance of our method over the best baseline results. If the p value is smaller than 0.05, we reject the null hypothesis and claim that our method significantly outperforms the best baseline method.

Baselines

We compare the proposed method HINT with several baselines, including conventional machine learning models and deep learning methods. We enhance their feature sets for all classical machine learning baselines to be the same as HINT. In particular, we include (1) 1,024-dimensional Morgan fingerprint features,[24] (2) GRAM embedding, where the GRAM model is pretrained using disease risk modules, and (3) Bidirectional Encoder Representations from Transformers (BERT) embedding of eligibility criteria. Then these three sets of features are concatenated as the input of all baselines. For deep learning baselines (DeepEnroll and COMPOSE), molecule encoders over input molecule graphs are added.

- Logistic regression (LR): LR was used[10,25] for trial outcome predictions.
- Random Forest (RF): Similar to LR, RF was used[10,25] for trial outcome predictions.
- XGBoost: An implementation of gradient-boosted decision trees designed for speed and performance. It was used in the context of individual patient trial outcome predictions in Rajpurkar et al. and Siah et al.[7,25]
- Adaptive boosting (AdaBoos) was used in Fan et al.[26] for individual Alzheimer's patients' trial result predictions.
- k Nearest Neighbor (kNN) + RF[10] combines statistical imputation techniques for handling missing data and standard classification methods. In the experiment, we chose the best-performing model reported[10] with kNN as the imputation technique and Random Forest as the classifier.
- Feedforward Neural Network (FFNN):[15] It uses the same feature as HINT. The feature vectors are fed into a three-layer feedforward neural network, where the hidden dimensions are 500 and 100, and the rectified linear unit (ReLU) function is used as an activation function in the hidden layer to provide nonlinearity.
- DeepEnroll[27] was originally designed for patient-trial matching, and it uses (1) a pre-trained BERT model[28] to encode eligibility criteria into sentence embedding, (2) a hierarchical embedding model for disease information, and (3) an alignment model to capture the eligibility criteria-disease interaction information. To adapt DeepEnroll for trial outcome predictions, the molecule embedding ($\mathbf{h}_m$) computed by the MPNN algorithm[29] over molecule graphs is concatenated to the output of the alignment model to make trial outcome predictions.
- COMPOSE[16] was also originally designed for patient-trial matching. It uses convolutional highway and memory networks to encode eligibility criteria and diseases, respectively, and an alignment model to model the interaction. COMPOSE incorporates the molecule information in the same way as DeepEnroll, as described above.

Disease subgroups

We also study the predictive performance of trial outcome predictions on different disease subgroups. These 4 groups take

Table 2. Empirical results of various approaches for phase-level-outcome predictions on test sets

Phase I trials			
# train: 1,044; # valid: 116; # test: 627; # patients/trial: 45			
Method	PR-AUC	F1	ROC-AUC
LR	0.500 ± 0.005	0.604 ± 0.005	0.520 ± 0.006
RF	0.518 ± 0.005	0.621 ± 0.005	0.525 ± 0.006
XGBoost	0.513 ± 0.06	0.621 ± 0.007	0.518 ± 0.006
AdaBoost	0.519 ± 0.005	0.622 ± 0.007	0.526 ± 0.006
kNN+RF[10]	0.531 ± 0.006	0.625 ± 0.007	0.538 ± 0.005
FFNN[15]	0.547 ± 0.010	0.634 ± 0.015	0.550 ± 0.010
DeepEnroll[27]	0.568 ± 0.007[†]	0.648 ± 0.011	0.575 ± 0.013
COMPOSE[16]	0.564 ± 0.007	0.658 ± 0.009	0.571 ± 0.011
HINT	0.567 ± 0.010	0.665 ± 0.010[†]	0.576 ± 0.008[†]
Phase II trials			
# train: 4,004; # valid: 445; # test: 1,653; # patients/trial: 183			
Method	PR-AUC	F1	ROC-AUC
LR	0.565 ± 0.005	0.555 ± 0.006	0.587 ± 0.009
RF	0.578 ± 0.008	0.563 ± 0.009	0.588 ± 0.009
XGBoost	0.586 ± 0.006	0.570 ± 0.009	0.600 ± 0.007
AdaBoost	0.586 ± 0.009	0.583 ± 0.008	0.603 ± 0.007
kNN+RF[10]	0.594 ± 0.008	0.590 ± 0.006	0.597 ± 0.008
FFNN[15]	0.604 ± 0.010	0.599 ± 0.012	0.611 ± 0.011
DeepEnroll[27]	0.600 ± 0.010	0.598 ± 0.007	0.625 ± 0.008
COMPOSE[16]	0.604 ± 0.007	0.597 ± 0.006	0.628 ± 0.009
HINT	0.629 ± 0.009[*,†]	0.620 ± 0.008[*,†]	0.645 ± 0.006[†]
Phase III trials			
# train: 3,092; # valid: 344; # test: 1,140; # patients/trial: 1,418			
Method	PR-AUC	F1	ROC-AUC
LR	0.687 ± 0.005	0.698 ± 0.005	0.650 ± 0.007
RF	0.692 ± 0.004	0.686 ± 0.010	0.663 ± 0.007
XGBoost	0.697 ± 0.007	0.696 ± 0.005	0.667 ± 0.005
AdaBoost	0.701 ± 0.005	0.695 ± 0.005	0.670 ± 0.004
kNN+RF[10]	0.707 ± 0.007	0.698 ± 0.008	0.678 ± 0.010
FFNN[15]	0.747 ± 0.011	0.748 ± 0.009	0.681 ± 0.008
DeepEnroll[27]	0.777 ± 0.008	0.786 ± 0.007	0.699 ± 0.008
COMPOSE[16]	0.782 ± 0.008	0.792 ± 0.007	0.700 ± 0.007
HINT	0.811 ± 0.007[*,†]	0.847 ± 0.009[*,†]	0.723 ± 0.006[*,†]

The mean and standard deviation of 30 independent runs (with different random seeds) are reported.

[*] Groups whose performances are significantly better than the best baseline (pass the t test, i.e., p value < 0.05).

[†] The best performing models on each metric for each trial phase.

up 19.4%, 8.4%, 12.7%, and 12.0% of all trials in the benchmark TOP, respectively.

- Neoplasm/tumor/cancer/oncology, e.g., cerebellar neoplasms, neuroectodermal tumors, breast cancer, and stomach neoplasms.
- Respiratory system diseases, e.g., tuberculosis, sinusitis, and tonsillitis.
- Digestive system diseases, e.g., cholera, esophageal disease, gastritis, and duodenitis.
- Nervous system diseases, e.g., meningitis, Parkinson's disease, and cerebral palsy.

Experiment (Exp) 1: Phase-level trial outcome predictions

First, we compare the performances of phase-level outcome predictions. For each phase, we train a separate model to make the prediction. We compare HINT with several baseline approaches, covering conventional machine learning models and deep-learning-based models. Among training data, we allocate 10% for examples as the validation set for model parameter tuning. The means and standard deviations of 30 independent runs (with different random seeds) are reported. We present the prediction performance in Table 2. We also report the relative success/failure proportion as a function of the predicted success probability on test sets in Figure 3, where the distributions of predictions for positive (success) and negative (fail) samples are significantly different. We have the following observations:

1. Deep-learning-based approaches including FFNN, DeepEnroll, COMPOSE, and HINT outperforms conventional machine learning approaches (LR, RF, XGBoost, AdaBoost, kNN+RF) significantly in outcome predictions for all three phases. This confirmed the benefit of deep learning models for clinical-trial-outcome predictions.
2. Among all deep learning methods, HINT performs best with F1 scores of 0.665 for phase I, 0.620 for phase II, and 0.847 for phase III. Compared with the strongest baseline (COMPOSE), HINT achieved 1.0%, 3.9%, and 7.0% relative improvement in terms of F1. The likely reason for this performance improvement is that HINT incorporates insightful multi-modal data embedding and finer-grained interactions between multi-modal data and trial components (i.e., nodes in interaction graphs).
3. When comparing the prediction performance across phases I, II, and III, we find that phase III achieves the highest accuracy for almost all methods, while phases I and II are more challenging with lower accuracy.

Exp 2: Evaluation on various disease groups

We evaluate HINT on different disease groups, including neoplasm (oncology/cancer/tumor), respiratory system diseases, digestive system diseases, and nervous system diseases. We also evaluate the prediction performance on high/low-prevalence disease trials (top/bottom 10% in prevalence), and the frequencies are counted in the training sets. Some examples of low-prevalence diseases are I25.700 (atherosclerosis of coronary artery bypass graft[s], unspecified, with unstable angina pectoris), Z62.6 (inappropriate [excessive] parental pressure), and E22.2 (syndrome of inappropriate secretion of antidiuretic hormone), and some examples of high-prevalence diseases are E11.44 (type 2 diabetes mellitus with diabetic amyotrophy), C05.2 (alignant neoplasm of uvula), and F20.3 (undifferentiated schizophrenia). We present the results in Table 3. We observe that the prediction of neoplasm-related trials achieves 0.604 PR-AUC, 0.585 F1 score, and 0.595 ROC-AUC, which are significantly lower than the other cohorts, showing that predicting the outcome for neoplasm is most challenging. The prediction of respiratory-system-disease-related trials achieves 0.860 PR-AUC, 0.867 F1 score, and 0.805 ROC-AUC, obtaining the highest

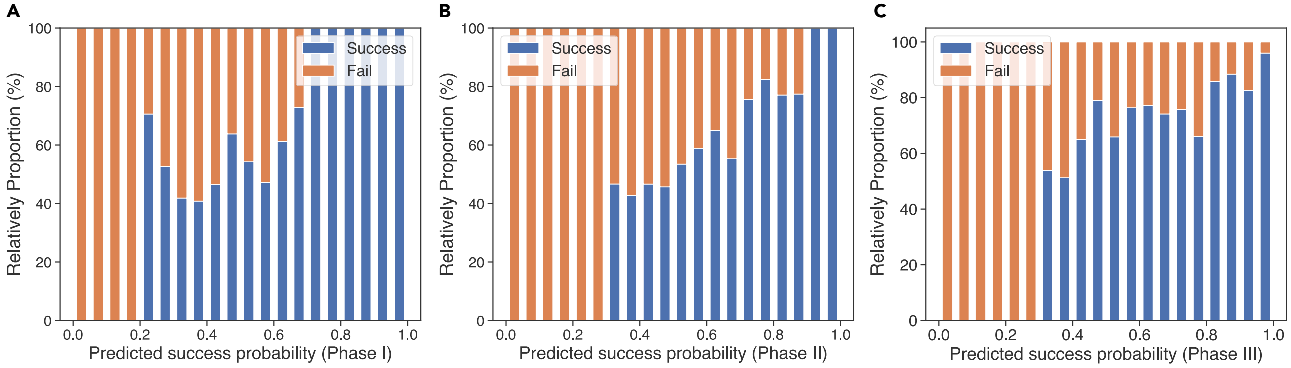


Figure 3. Predicted success probability versus relative success/failure proportion
Phase I, II, and III showing in (A), (B), and (C), respectively.

accuracy among all cohorts. Digestive-system-related trials also achieve a great prediction performance (0.750 PR-AUC, 0.786 F1 score, and 0.728 ROC-AUC). Also, we find that low-prevalence-disease-related trials still achieve a good prediction performance, validating the generalization of the proposed method.

Exp 3: Evaluation on missing data imputation

The complete data sample is $(\mathbb{M}, \mathbb{D}, \mathbb{C}, y)$, while missing data is $(\mathbb{D}, \mathbb{C}, y)$ (molecule information is missing). We use a data imputation module to estimate molecule embedding $\widehat{\mathbf{h}_m}$. To further validate the effectiveness of the data imputation module, we conducted experiments on phase-III-level predictions with missing molecules. Specifically, we fix the training set, where all data are complete, then vary the percentage of missing molecules from $\{0\%, 25\%, 50\%, 75\%\}$. We compare the performance of three methods: (1) HINT (with imputation), (2) HINT (without imputation), and (3) COMPOSE (the strongest baseline). Both (2) and (3) cannot leverage missing data; hence, the samples with missing molecule information are ignored in the training data. We conduct 30 independent runs (with different random seeds) and report the average F1 score and corresponding 95% confidence interval for all three methods on all scenarios and show the results in Figure 4. We find that all methods degrade with missing data, as expected. However, HINT with imputation outperforms the other methods, thanks to its capability to impute the missing molecule embeddings.

Exp 4: Case studies

Next, we provide qualitative examples of HINT applying to some recent trials in Table 4.

Examples of failed trials. One of the most promising drugs in 2019 was Entresto for heart failure, the leading cause of death in the United States. Entresto is sponsored by Novartis and was expected to have a 5-billion-dollar peak sale. However, in multi-country phase III trials with 4,822 patients enrolled, the result did not reduce death or meet any other endpoints. The trial took 5 years (2014–2019) and was estimated to cost $200 million dollars (we use the median per-patient cost multiplied by the number of patients to estimate the cost[30]). We feed the drug (Entresto), disease (heart failure), and their phase III eligibility criteria into HINT, and it predicts a low success probability of 0.476. This means that HINT could potentially have alerted the practitioners of the likely failure.

We also tested HINT on Fevipiprant, which was expected to be Novartis's blockbuster drug for asthma. The phase III trial of Fevipiprant took 4 years (2015–2019) and enrolled 894 patients, and it also incurred huge costs (an estimated 40 million dollars). Unfortunately, the primary endpoint was not met, and Fevipiprant was retired. We feed the drug (Fevipiprant), disease (asthma), and eligibility criteria into HINT, and it predicts a 0.352 success probability, which is low.

Similarly, for a recent phase II study on the effect of Pembrolizumab and Epacadostat on non-small cell lung cancer by Incyte and Merck, HINT correctly predicts the failure of the trial.

Error analysis

We also observe some cases that HINT predicts wrongly. For example, HINT generates a wrong prediction in a recent phase III study of Ustekinumab on a rare disease called lupus erythematosus (4th row in Table 4). We did the error analysis in a phase-III-level prediction, which contains 1,140 test points, where 874 points are predicted correctly. To better investigate HINT's mechanism, we divide the test data points into two disjoint groups: (1)

Table 3. Phase-III-level predictions for disease groups

Cohorts	In test set, %	PR-AUC	F1	ROC-AUC
Neoplasm	19.4	0.604 ± 0.015	0.585 ± 0.10	0.595 ± 0.016
Respiratory	8.4	0.860 ± 0.028	0.867 ± 0.023	0.805 ± 0.029
Digestive	12.7	0.849 ± 0.024	0.858 ± 0.028	0.813 ± 0.025
Nervous system	12.0	0.750 ± 0.023	0.786 ± 0.035	0.728 ± 0.025
High-prevalence diseases	10	0.845 ± 0.019	0.852 ± 0.021	0.748 ± 0.011
Low-prevalence diseases	10	0.783 ± 0.017	0.803 ± 0.012	0.702 ± 0.021
All diseases	100	0.811 ± 0.007	0.847 ± 0.009	0.723 ± 0.006

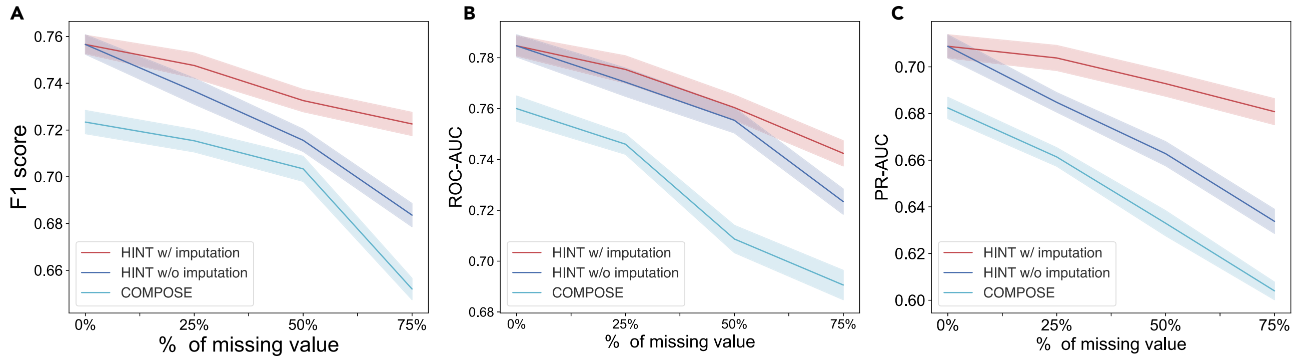


Figure 4. Empirical studies on missing molecule imputation on phase-III-level predictions

Evaluated on F1, ROC-AUC, and PR-AUC showing in (A), (B), and (C), respectively. For each data point, we conduct 30 independent runs with different random seeds, evaluate their mean and standard deviation, and plot the 95% confidence interval.

predicted correctly by HINT (874 data points) and (2) predicted wrongly by HINT (266 data points). For both groups, we evaluate whether the drugs and disease codes in the test set are already seen in the training set. More specifically, group 1 represents the drug/disease code appearing in the training set, whereas group 0 means not.

Regarding the drugs, for group 1, the mean and standard deviation of the indicator variable are 0.348 and 0.476, respectively; that is, 34.8% of the drugs appear in the training set. For group 2, the mean and standard deviation of the indicator variable are 0.345 and 0.475, respectively. In this case, groups 1 and 2 do not differ much.

Regarding disease codes, for group 1, the mean and standard deviation of the indicator variable are 0.928 and 0.156, respectively, which means that 92.8% of disease codes appear in the training set; for group 2, the mean and standard deviation are 0.814 and 0.302, respectively. Groups 1 and 2 differ significantly on these statistics. We conclude that the trials with unseen disease codes (not in the training set) are more likely to be mispredicted by HINT, which is more likely to happen with rare diseases. This is also consistent with the example of lupus erythematosus in Table 3.

Examples of successful trials

HINT can also predict the success probability of trials accurately, reassuring drug developers about the prospect of treatment. For example, HINT predicts several recent huge trial successes:

- Sitagliptin on diabetes by Merck 2017 received a 0.742 success probability.
- Etanercept for rheumatoid arthritis by Amgen 2019 received a 0.673 success probability.
- Afibercept for glaucoma by Bayer 2020 acquired a 0.854 success probability.
- Naltrexone for depression by University of Pittsburgh 2020 received a 0.747 success probability.
- cTACE Doxorubicin for liver cancer by Yale University 2020 received a 0.583 success probability.
- Phosphate supplement and vitamin D for X-linked hypophosphatemia by Ultragenyx 2020 received a 0.556 success probability.

In addition, we provide an interpretability analysis in Figure S1 and an ablation study in Table S2.

DISCUSSION

Trial outcome predictions

Existing works often focus on predicting individual patient outcomes in a trial instead of a general prediction about the overall trial success. They usually leverage expert-crafted features. For example, Wu et al.[31] leveraged support vector machines (SVMs) to predict the status of genetic lesions based on cancer clinical trial documents. Rajpurkar et al.[7] used gradient-boosted decision trees (GBDTs[32]) to predict the improvement in symptom scores based on the treatment symptom score and EEG measures for depressive symptoms with an antidepressant treatment. Hong et al.[8] focused on predicting clinical drug toxicity according to drug-property and target-property features and used an ensemble classifier of weighted least squares support vector regression. Note that these models are not tackling the same task as us. They are predicting at the patient level, whereas HINT focuses on the trial level. More relevant to us, Qi et al.[9] designed a Residual Semi-Recurrent Neural Network (RS-RNN) to predict phase III trial results based on phase II results. In contrast, the task of HINT is to predict for all clinical trial phases. Lo et al.[10] explored various imputation techniques and a series of conventional machine learning models (e.g., logistic regression, random forest, SVM) to predict the drug approval within 15 disease groups. Siah et al.[25] evaluated various conventional machine learning models for clinical-trial-outcome predictions. However, they did not leverage rich trial features, e.g., drug molecules and trial eligibility criteria, whereas HINT takes into account the multimodal data sources.

A related statistical practice during trial design is power analysis for sample-size estimation. Power/sample-size estimation is used to determine how many patients to recruit for a given trial in order to answer the research question in the study. More specifically, the sample size can be estimated given the treatment effect size between groups, the desirable power, and the statistical significance level. However, a strong assumption about the treatment effect has to be made to perform such power analysis

Table 4. Case studies: Prediction versus actual outcomes

Indication/disease	Drug	Sponsor	Year	Outcome	Prediction
Heart failure	Entresto	Novartis	2019	fail	0.476
Asthma	Fevipiprant	Novartis	2019	fail	0.352
Lung cancer	pembrolizumab and epacadostat	Incyte	2020	fail	0.498
Lupus erythematosus	ustekinumab	Janssen	2019	fail	0.567
Diabetes	sitagliptin	Merck	2017	success	0.742
Rheumatoid arthritis	etanercept	Amgen	2019	success	0.673
Neovascular glaucoma	aflibercept	Bayer	2020	success	0.854
Depression	naltrexone	U. Pitts.	2020	success	0.747
Liver cancer	cTACE doxorubicin	Yale U.	2020	success	0.583
X-linked hypophosphatemia	phosphate supplement and vitamin D	Ultragenyx	2020	success	0.556

Prediction is the HINT's predicted success probability. Low probability means the trial is likely to fail, and high probability means the trial is likely to succeed.

before trials. Moreover, power analysis does not utilize the rich trial information, unlike HINT. In comparison, HINT directly utilizes the information from the molecule structures, disease indication, and trial eligibility criteria to model the trial success probability. Finally, power analysis has to make assumptions about the treatment effect and variability of that effect, which can be tricky to estimate before the trial starts.

Trial representation learning

Recently, deep learning has been leveraged to learn representations from clinical trial data to support downstream tasks such as patient retrieval[16,27] and enrollment.[33] For example, Doctor2-Vec[33] learns hierarchical clinical trial embedding, where the unstructured trial descriptions were embedded using BERT.[28] DeepEnroll[27] and COMPOSE[16] leverage the pretrained BERT[28] model to encode clinical trial eligibility criteria. While these works optimize the representation learning for a single component in a trial, HINT models a diverse set of trial components such as molecule, disease, eligibility criteria, PK, and disease risk information and fuses them through an interaction graph neural network.

Limitations and future works

The current paper has several limitations. We plan to study them in our future works.

Supporting more trial types

HINT can handle interventional trials involving small molecules. Other trial types such as medical devices and biologics trials are not covered by the current model due to molecule encoding. From the method perspective, we can replace the molecule encoder with another encoder, such as protein sequence encoders for biologics. However, the challenge is in the limited training data of biologics. Nevertheless, supporting other trial types can be a future extension of the current work.

Supporting rare diseases

Like any machine learning model, HINT requires sufficient training data to train accurate predictive models. However, low-prevalence diseases, especially rare diseases, are difficult to handle due to the lack of sufficient historical trials as training data for HINT.

Enhancement of model interpretability

HINT is a graph neural network model that integrates comprehensive data sources to predict trial outcomes. Due to the complex interaction patterns, it can be difficult to explain those predictions. We provide an example to illustrate how to understand all intermediate predictions and provide some explanation in the supplemental material. However, we also recognize that the model interpretability should be further studied in future work.

Trial outcome labels

HINT assumes a simple binary label of success or not. However, there might be more granular classes of the trial outcomes, especially for the failed trials. It will be more useful if the model can classify the trial to more specific failure reasons. However, reliably creating such granular trial labels can be quite challenging. Significantly, the reason for a trial's failure might not be well documented or understood. Trial publications are often disproportionally skewed toward successful trials, although detailed explanations of failed trials can benefit future trials.

Look-ahead bias

Some subtle look-ahead biases might be inherent in the pretrained embedding modules, e.g., ADMET and trial risk encoders. This is because the data used for pretraining are not point in time, i.e., they might not be available in 2014. This means that future information (e.g., ADMET properties for a new molecule just published in 2020) could be indirectly used to create embeddings for drug molecular features. To address this challenge, a prospective study can be conducted moving forward to validate the HINT model for future trials.

Conclusion

In this paper, we create a machine learning benchmark for trial outcome predictions. We design a graph-neural-network-based method HINT to leverage multi-sourced data and incorporate multiple factors in a hierarchical interaction graph for predicting trial outcomes. Also, HINT can handle missing data via an imputation module. Empirical studies indicate that HINT outperforms multiple baseline methods in several prediction metrics on phase-level trial outcome predictions. Future works include expanding to other trial types beyond intervention trials of small molecules and expanding the binary trial outcome labels.

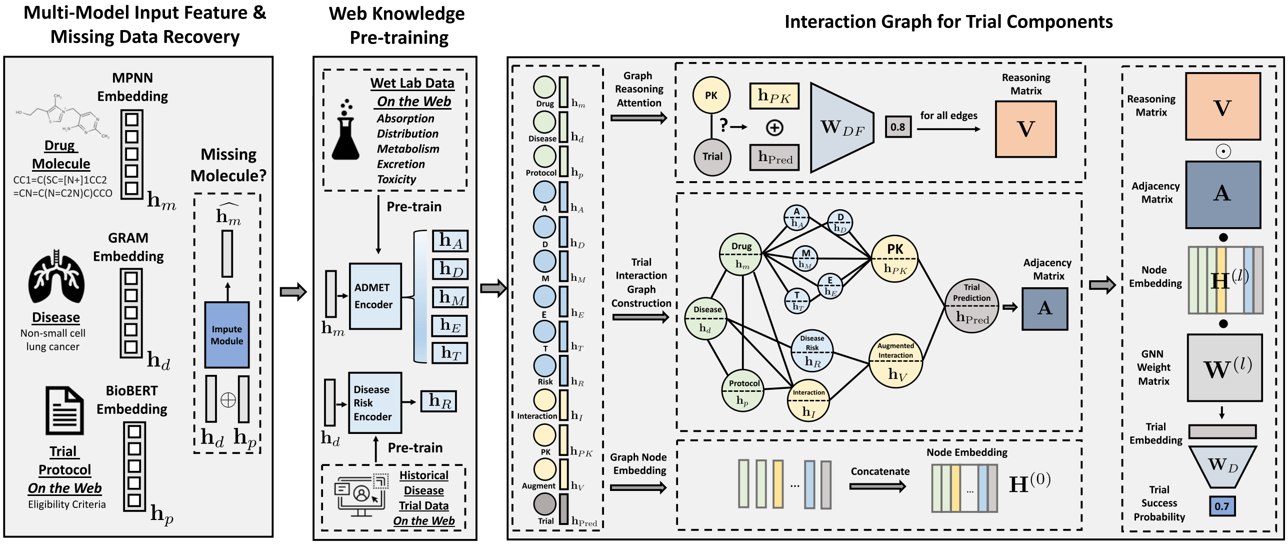


Figure 5. HINT **framework**

HINT is an end-to-end neural network pipeline with the following components: drug molecule embedding $\mathbf{h}_m$, disease embedding $\mathbf{h}_d$, and trial eligibility criteria embedding $\mathbf{h}_p$. Before constructing an interaction graph using these components, HINT pretrain some embeddings (blue nodes) using external knowledge about drug properties and disease risks. Then, we construct an interaction graph to characterize interactions between various trial components. Trial embeddings are learned based on the interaction graph to capture both trial components and their interactions. Based on the learned representation and the dynamic attentive graph neural network (Equation 13), we make trial outcome predictions.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for code and data should be directed to and will be fulfilled by the lead contact, Jimeng Sun (jimeng@illinois.edu).

Material availability

This study did not generate any physical materials.

Data and code availability

The benchmark datasets and codes (including data collection and preprocessing, model construction, learning process, and evaluation), referred as the Works, are publicly available for noncommercial use only at https://github.com/futianfan/clinical-trial-outcome-prediction. The dataset sources are publicly available and processed by the authors. The trial outcome labels are provided by IQVIA.

Method

HINT includes (1) the construction of the trial interaction graph and (2) the predicted trial outcome using the dynamic attentive graph neural network on the interaction graph. The details of the method are available in the supplemental information, where Table S1 list important mathematical notations. Here, we provide the high-level components in the method.

(1) Trial interaction graph construction: We construct hierarchical interaction graph $\mathcal{G}$ to connect all input data sources and important factors affecting clinical trial outcomes. The interaction graph $\mathcal{G}$ is constructed in a way to reflect the real-world trial development process, and it consists of four tiers of nodes that are connected between tiers:

(1.1) Input nodes (colored green in Figure 5) include drugs, diseases, and eligibility criteria with node features of input embedding $\mathbf{h}_m$, $\mathbf{h}_d$, $\mathbf{h}_p \in \mathbb{R}^d$. Formally, we represent (1) molecular graphs $\mathbb{M} = \{m_1, \cdots, m_{N_m}\}$, (2) disease codes $\mathbb{D} = \{d_1, \cdots, d_{N_d}\}$ (Equation 2), and (3) eligibility criteria (Equation 3) as follows:

$$\text{Molecule Embedding} \quad \mathbf{h}_m = \frac{1}{N_m}\sum_{j=1}^{N_m} f_m(m_j), \mathbf{h}_m \in \mathbb{R}^d, \qquad \text{(Equation 5)}$$

where $f_m(\cdot)$ is the molecule embedding function and can be a Morgan fingerprint,[24] SMILES encoder,[34] graph message passing neural network (MPNN),[29,35] or graph neural network.[36] We average all molecules' embeddings to get the drug embedding.

$$\text{Disease Embedding} \quad \mathbf{h}_d = \frac{1}{N_d}\sum_{i=1}^{N_d} \text{GRAM}(d_i), \mathbf{h}_d \in \mathbb{R}^d, \qquad \text{(Equation 6)}$$

where $\text{GRAM}(d_i)$ represent an embedding of disease d_i using the graph-based attention model GRAM.[37]

$$\text{Eligibility Criteria Embedding} \quad \mathbf{h}_p = f_p(\mathbb{C}), \mathbf{h}_p \in \mathbb{R}^d, \qquad \text{(Equation 7)}$$

where we apply Clinical-BERT,[38,39] which is a domain-specific version of BERT,[28] to embed each sentence of eligibility criteria and aggregate them into an embedding vector.

(1.2) External knowledge nodes (colored blue in Figure 5) include (1) ADMET embeddings $\mathbf{h}_* = \mathcal{X}_*(\mathbf{h}_m) \in \mathbb{R}^d$, where $* \in \{A, D, M, E, T\}$ (i.e., $\mathbf{h}_A$, $\mathbf{h}_D$, $\mathbf{h}_M$, $\mathbf{h}_E$, $\mathbf{h}_T$), and (2) disease risk embedding $\mathbf{h}_R = \mathcal{R}(\mathbf{h}_d)$. They are pretrained on external knowledge.

(1.3) Aggregation nodes are colored yellow in Figure 5.

The PK nodes gather all information of the five ADMET properties:

$$\text{Pharmacokinetics (PK)} \quad \mathbf{h}_{PK} = \mathcal{K}(\mathbf{h}_A, \mathbf{h}_D, \mathbf{h}_M, \mathbf{h}_E, \mathbf{h}_T), \mathbf{h}_{PK} \in \mathbb{R}^d. \qquad \text{(Equation 8)}$$

Then, an interaction node models the interaction among the drug molecule, disease, and eligibility criteria:

$$\text{Interaction} \quad \mathbf{h}_I = \mathcal{I}(\mathbf{h}_m, \mathbf{h}_d, \mathbf{h}_p), \mathbf{h}_I \in \mathbb{R}^d. \qquad \text{(Equation 9)}$$

In addition, we have an augmented interaction model to combine (1) the trial risk of the target disease $\mathbf{h}_R$ and (2) the interaction among the disease, molecule, and eligibility criteria $\mathbf{h}_I$.

$$\text{Augmented Interaction} \quad \mathbf{h}_V = \mathcal{V}(\mathbf{h}_R, \mathbf{h}_I), \mathbf{h}_V \in \mathbb{R}^d. \qquad \text{(Equation 10)}$$

(1.4) Prediction node (colored gray in Figure 5) summarizes the PK and the augmented interaction to obtain the final prediction:

$$\text{Trial Prediction} \quad \mathbf{h}_{\text{pred}} = \mathcal{P}(\mathbf{h}_{PK}, \mathbf{h}_V), \mathbf{h}_{\text{pred}} \in \mathbb{R}^d. \qquad \text{(Equation 11)}$$

(2) Dynamic attentive graph neural network: The trial embeddings provide initial representations of different trial components and their interactions via a graph. To further enhance the predictions, we design a dynamic attentive graph neural network to leverage this interaction graph to model the influential trial components and help improve predictions.

Mathematically, the interaction graph $\mathcal{G}$ is the input graph where nodes are trial components and edges are the relations among these trial components. We denote $\mathbf{A} \in \{0, 1\}^{K \times K}$ as the adjacency matrix of $\mathcal{G}$. The node embeddings $\mathbf{H}^{(0)} \in \mathbb{R}^{K \times d}$ are initialized to

$$\mathbf{H}^{(0)} = [\mathbf{h}_d, \mathbf{h}_m, \mathbf{h}_p, \mathbf{h}_A, \mathbf{h}_D, \mathbf{h}_M, \mathbf{h}_E, \mathbf{h}_T, \mathbf{h}_R, \mathbf{h}_{PK}, \mathbf{h}_I, \mathbf{h}_V, \mathbf{h}_{\text{pred}}]^\top \in \mathbb{R}^{K \times d}, \qquad \text{(Equation 12)}$$

$K = |\mathcal{G}|$ is the number of nodes in graph $\mathcal{G}$. $K = 13$ in this paper. We further enhance the node embeddings using a graph convolutional network (GCN).[40]

$$\mathbf{H}^{(l)} = \text{RELU}\left(\mathbf{B}^{(l)} + (\mathbf{V} \odot \mathbf{A})\left(\mathbf{H}^{(l-1)}\mathbf{W}^{(l)}\right)\right), l = 1, \cdots, L, \mathbf{H}^{(l)} \in \mathbb{R}^{K \times d}, \qquad \text{(Equation 13)}$$

where $\mathbf{B} \in \mathbb{R}^{K \times d}$ is a bias parameter, $\mathbf{W}^{(l)} \in \mathbb{R}^{d \times d}$ is the weight matrix in the l-th layer to transform the embedding, L is depth of GCN, and $\odot$ is the element-wise multiplication.

Different from conventional GCNs,[40] we introduce a learnable layer-independent attentive matrix $\mathbf{V} \in \mathbb{R}_+^{K \times K}$. $\mathbf{V}_{i,j}$, the (i, j)-th entry of $\mathbf{V}$, measures the importance of the edge that connects the i-th and j-th nodes in $\mathcal{G}$. We evaluate $\mathbf{V}_{i,j}$ based on the i-th and j-th nodes' embeddings in $\mathbf{H}^{(0)}$, which are denoted $\mathbf{h}^i, \mathbf{h}^j \in \mathbb{R}^d$ ($\mathbf{h}^i \in \mathbb{R}^d$ is transpose of the i-th row of $\mathbf{H}^{(0)} \in \mathbb{R}^{K \times d}$ in Equation 12):

$$\mathbf{V}_{i,j} = g_2\left(\text{CONCAT}\left(\mathbf{h}^i, \mathbf{h}^j\right)\right), i, j \in \{1, \cdots, K\}, \mathbf{V}_{i,j} \in \mathbb{R}_+, \qquad \text{(Equation 14)}$$

where $g_2(\cdot)$ is a two-layer fully connected neural network with ReLU and sigmoid activation functions in the hidden and output layers, respectively. Note that the attentive matrix $\mathbf{V}$ is element-wise multiplied to the adjacency matrix $\mathbf{A}$ (Equation 13) so that message of the edge with higher prediction scores would give a higher weight to propagate.

Training: The target is binary label $y \in \{0, 1\}$, and $y = 1$ indicates the trial succeeds, while 0 means it fails. After GNN message passing, we obtain an updated representation for each trial component. We then use the last-layer (L-th layer) representation on the trial prediction node to generate the trial outcome prediction, $\hat{y} = \text{Sigmoid}(\text{FC}(\mathbf{h}_{\text{pred}}^L))$, where $\hat{y} \in [0, 1]$, L is the depth of the GCN. We use one-layer fully connected networks with sigmoid activation functions. Then, binary cross-entropy loss is used to guide the model training:

$$\mathcal{L}_{\text{classify}} = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}). \qquad \text{(Equation 15)}$$

HINT is trained in an end-to-end manner.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100445.

ACKNOWLEDGMENTS

This work was supported by IQVIA, NSF award SCH-2014438, PPoSS 2028839, IIS-1838042, NIH award R01 1R01NS107291-01, and OSF Healthcare.

AUTHOR CONTRIBUTIONS

Conceptualization, T.F., K.H., C.X., L.M.G., and J.S.; investigation, T.F., K.H., C.X., L.M.G., and J.S.; experiments, T.F. and K.H.; writing – original draft, T.F., K.H., C.X., L.M.G., and J.S.; writing – review & editing, T.F., K.H., C.X., L.M.G., and J.S.; funding acquisition, J.S.; resources, J.S.; supervision, J.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 31, 2021
Revised: November 29, 2021
Accepted: January 14, 2022
Published: February 3, 2022

REFERENCES

1. Research and Markets Ltd. Clinical trials market size, share & trends analysis report by phase (phase i, phase II, phase III, phase IV), study design (interventional, observational, expanded access), indication, region, and segment forecasts, 2021-2028. https://www.researchandmarkets.com/reports/4396385/clinical-trials-market-size-share-and-trends.
2. Martin, L., Hutchens, M., Hawkins, C., and Radnov, A. (2017). How much do clinical trials cost. Nat. Rev. Drug Discov. *16*, 381–382.
3. Peto, R. (1978). Clinical trial methodology. Nature *272*, 15–16.
4. Ledford, H. (2011). 4 ways to fix the clinical trial: clinical trials are crumbling under modern economic and scientific pressures. nature looks at ways they might be saved. Nature *477*, 526–529.
5. Friedman, L.M., Furberg, C.D., DeMets, D.L., Reboussin, D.M., and Granger, C.B. (2015). Fundamentals of Clinical Trials (Springer).
6. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). Drugbank 5.0: a major update to the drugbank database for 2018. Nucleic Acids Res. *46*, D1074–D1082.
7. Rajpurkar, P., Yang, J., Dass, N., Vale, V., Keller, A.S., Irvin, J., Taylor, Z., Basu, S., Ng, A., and Williams, L.M. (2020). Evaluation of a machine learning model based on pretreatment symptoms and electroencephalographic features to predict outcomes of antidepressant treatment in adults with depression: a prespecified secondary analysis of a randomized clinical trial. JAMA Netw. Open *3*, e206653.
8. Hong, Z.Y., Shim, J., Son, W.C., and Hwang, C. (2020). Predicting successes and failures of clinical trials with an ensemble LS-SVR. medRxiv. https://doi.org/10.1101/2020.02.05.20020636.
9. Qi, Y., and Tang, Q. (2019). Predicting phase 3 clinical trial results by modeling phase 2 clinical trial subject level data using deep learning. Proc. Machine Learn. Res. *106*, 288–303.
10. Lo, A.W., Siah, K.W., and Wong, C.H. (2019). Machine learning with statistical imputation for predicting drug approvals. Harv. Data Sci. Rev. *1*, 7.
11. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. *25*, 1097–1105.
12. Gayvert, K.M., Madhukar, N.S., and Elemento, O. (2016). A data-driven approach to predicting successes and failures of clinical trials. Cell Chem. Biol. *23*, 1294–1301.
13. Artemov, A.V., Putin, E., Vanhaelen, Q., Aliper, A., Ozerov, I.V., and Zhavoronkov, A. (2016). Integrated deep learned transcriptomic and structure-based predictor of clinical trials outcomes. BioRxiv, 095653. https://doi.org/10.1101/095653.
14. Dong, J., Wang, N.-N., Yao, Z.-J., Zhang, L., Cheng, Y., Ouyang, D., Lu, A.-P., and Cao, D.-S. (2018). Admetlab: a platform for systematic admet evaluation based on a comprehensively collected admet database. J. Cheminformatics *10*, 1–11.
15. Tranchevent, L.-C., Azuaje, F., and Rajapakse, J.C. (2019). A deep neural network approach to predicting clinical outcomes of neuroblastoma patients. BMC Med. Genomics *12*, 1–11.
16. Gao, J., Xiao, C., Glass, L.M., and Sun, J. (2020). COMPOSE: cross-modal pseudo-siamese network for patient trial matching. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 803–812.
17. Anker, S.D., Morley, J.E., and von Haehling, S. (2016). Welcome to the Icd-10 Code for Sarcopenia, *volume 7* (Wiley Online Library), pp. 512–514.

18. Martin-Moreno, J.M., Alfonso-Sanchez, J.L., Harris, M., and Lopez-Valcarcel, B.G. (2010). The effects of the financial crisis on primary prevention of cancer. Eur. J. Cancer *46*, 2525–2533.

19. Ma, C.-Y., Yang, S.-Y., Zhang, H., Xiang, M.-L., Huang, Q., and Wei, Y.-Q. (2008). Prediction models of human plasma protein binding rate and oral bioavailability derived by using ga–cg–svm method. J. Pharm. Biomed. Anal. *47*, 677–682.

20. Adenot, M., and Lahana, R. (2004). Blood-brain barrier permeation models: discriminating between potential cns and non-cns drugs including p-glycoprotein substrates. J. Chem. Inf. Computer Sci. *44*, 239–248.

21. Veith, H., Southall, N., Huang, R., James, T., Fayne, D., Artemenko, N., Shen, M., Inglese, J., Austin, C.P., Lloyd, D.G., et al. (2009). Comprehensive characterization of cytochrome p450 isozyme selectivity across chemical libraries. Nat. Biotechnol. *27*, 1050–1055.

22. Pihan, E., Colliandre, L., Guichou, J.-F., and Douguet, D. (2012). e-drug3d: 3d structure collections dedicated to drug repurposing and fragment-based drug design. Bioinformatics *28*, 1540–1541.

23. Richard, A.M., Judson, R.S., Houck, K.A., Grulke, C.M., Volarath, P., Thillainadarajah, I., Yang, C., Rathman, J., Martin, M.T., Wambaugh, J.F., et al. (2016). Toxcast chemical landscape: paving the road to 21st century toxicology. Chem. Res. Toxicol. *29*, 1225–1251.

24. Cereto-Massagué, A., Ojeda, M.J., Valls, C., Mulero, M., Garcia-Vallvé, S., and Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. Methods *71*, 58–63.

25. Siah, K.W., Kelley, N., Ballerstedt, S., Holzhauer, B., Lyu, T., Mettler, D., Sun, S., Wandel, S., Zhong, Y., Zhou, B., et al. (2021). Predicting drug approvals: the novartis data science and artificial intelligence challenge. Patterns *2*, 100312.

26. Fan, Z., Xu, F., Li, C., and Yao, L. (2020). Application of KPCA and adaboost algorithm in classification of functional magnetic resonance imaging of alzheimer's disease. Neural Comput. Appl. 1–10.

27. Zhang, X., Xiao, C., Glass, L.M., and Sun, J. (2020). Deepenroll: patient-trial matching with deep embedding and entailment prediction. In Proceedings of The Web Conference 2020, pp. 1029–1037.

28. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)), pp. 4171–4186.

29. Huang, K., Fu, T., Glass, L.M., Zitnik, M., Xiao, C., and Sun, J. (2020). DeepPurpose: a deep learning library for drug–target interaction prediction. Bioinformatics *36*, 5545–5547.

30. Moore, T.J., Zhang, H., Anderson, G., and Alexander, G.C. (2018). Estimated costs of pivotal trials for novel therapeutic agents approved by the us food and drug administration. JAMA Intern. Med. *178*.

31. Wu, Y., Levy, M.A., Micheel, C.M., Yeh, P., Tang, B., Cantrell, M.J., Cooreman, S.M., and Xu, H. (2012). Identifying the status of genetic lesions in cancer clinical trial documents using machine learning. BMC genomics *13*, 1–9.

32. Ye, J., Chow, J.-H., Chen, J., and Zheng, Z. (2009). Stochastic gradient boosted distributed decision trees. In Proceedings of the 18th ACM conference on Information and knowledge management, pp. 2061–2064.

33. Biswal, S., Xiao, C., Glass, L.M., Milkovits, E., and Sun, J. (2020). Doctor2vec: dynamic doctor representation learning for clinical trial recruitment. Proc. AAAI Conf. Artif. Intelligence *34*, 557–564.

34. Wang, S., Guo, Y., Wang, Y., Sun, H., and Huang, J. (2019). Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics, pp. 429–436.

35. Jin, W., Barzilay, R., and Jaakkola, T. (2018). Junction tree variational autoencoder for molecular graph generation. In International conference on machine learning (PMLR), pp. 2323–2332.

36. Fu, T., Xiao, C., Li, X., Glass, L.M., and Sun, J. (2021). Mimosa: multi-constraint molecule sampling for molecule optimization. In Proceedings of the AAAI Conference on Artificial Intelligence, *35*, pp. 125–133.

37. Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., and Sun, J. (2017). GRAM: graph-based attention model for healthcare representation learning. In Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 787–795.

38. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., Redmond, W.A., and McDermott, M.B.A. (2019). Publicly available clinical bert embeddings. In NAACL HLT 2019, p. 72.

39. Huang, K., Altosaar, J., and Ranganath, R. (2019). ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission (CHIL Workshop).

40. Kipf, T.N., and Welling, M. (2017). Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations (ICLR).