GHINT: GPT-enhanced Hierarchical Interaction Network for Multimodal Clinical Trial Outcome Prediction

Sean Xu, Natalie Yang, Lexa Zhong

Carnegie Mellon University, 11-711 Advanced Natural Language Processing, Spring 25

Why It Matters?

High Cost & Long Timelines: Clinical trials often exceed $100 million and take over a decade to complete.
Smarter Resource Allocation: Early outcome forecasts help sponsors focus on the most promising compounds, optimize site selection, and streamline patient recruitment.
Enhanced Patient Safety: Predicting likely failures ahead of time reduces participant exposure to ineffective or unsafe interventions.
Deeper Protocol Insights: Leveraging advanced embeddings (e.g., LLM-based) on trial protocols uncovers nuanced design details that traditional models miss, boosting prediction accuracy and accelerating decision-making.

Baseline Model

Hierarchical Interaction Network (HINT)

HINT (Fu et al., 2022), a benchmark model for clinical trail outcome prediction, first encodes drug molecules, target diseases, and trial eligibility criteria into vector embeddings, then constructs a hierarchical interaction graph to capture their cross-modal relationships, and finally applies a dynamic, attentive GNN to predict trial success probabilities.

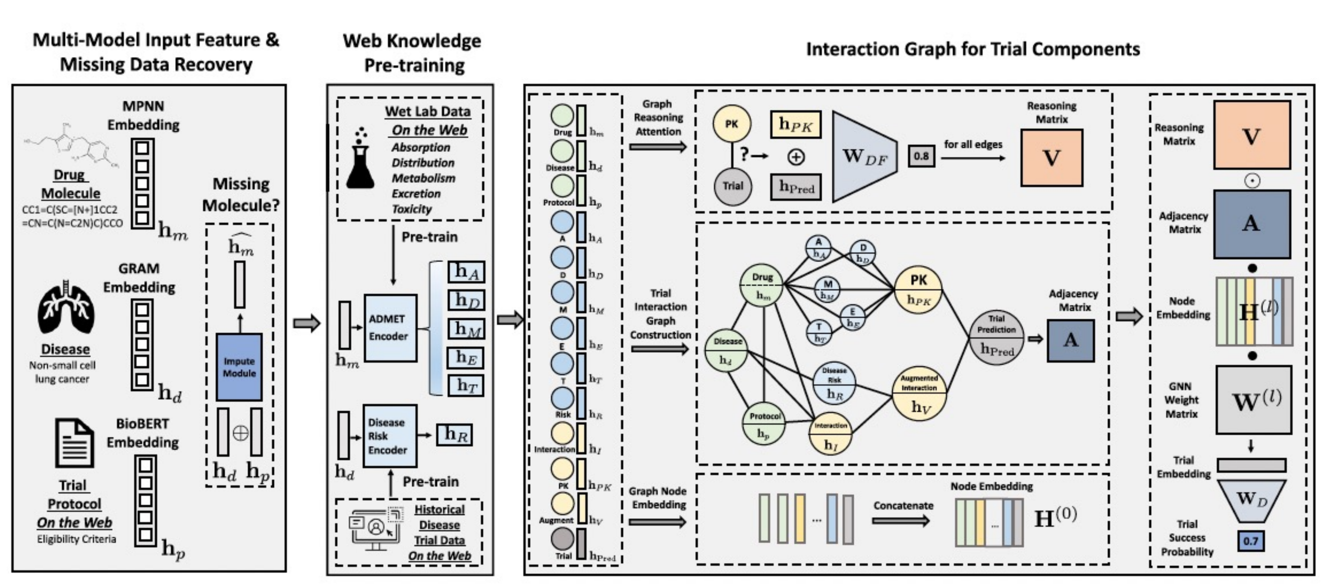

Figure 1: HINT Framework (credit to the original HINT paper)

	PR-AUC	F1	ROC-AUC
Phase I	*0.567 ± 0.010*	*0.665 ± 0.010*	*0.576 ± 0.008*
Phase II	*0.629 ± 0.009*	*0.620 ± 0.008*	*0.645 ± 0.006*
Phase III	*0.811 ± 0.007*	*0.847 ± 0.009*	*0.723 ± 0.006*

Table 1: HINT Results for Phase-level Outcome Predictions on Test Sets

What Can Be Improved?

Richer Data Inputs
- Go beyond eligibility criteria, SMILES strings and ICD-10 codes by adding additional protocol fields on ClinicalTrials.gov, e.g. outcome measures, allocation methods, intervention arms—so the model can leverage nuances of trial design.

Advanced Embeddings
- Experiment with promising text embedding models, such as `ClinicalBERT`, which is tailored for clinical text, and OpenAI's state-of-the-art `text-embedding-3-large`.

Methods

GPT-enhanced HINT (GHINT)

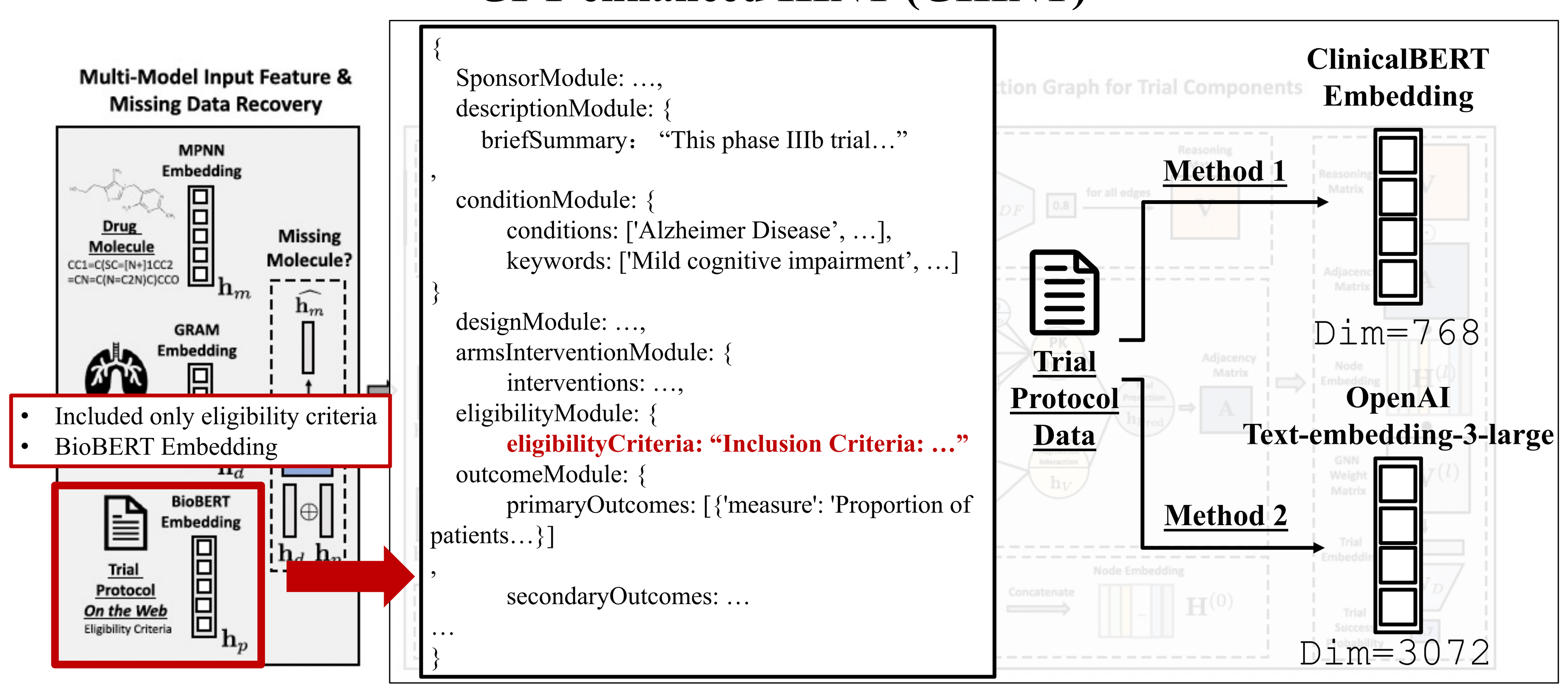

Figure 2: GHINT Framework

Additional Training Pipeline Enhancements
- **Optimizer**: Adam → AdamW
- **Activation**: ReLU → GELU
- Added **scheduler** to dynamically decay learning rates
- Streamlined **Dataset** and **DataLoader** for higher throughput

Ablation

We conducted ablation study using `TOP` phase II dataset: # train: 4,004; # valid: 445; # test: 1,653
It is the most critical to determine which **protocol embedding model**—`text-embedding-3-large` from OpenAI API or `ClinicalBERT`—works better. Besides, we experiment with different hyperparameters for **optimization** and **HINT Architecture**. Notably, we expect a larger **embedding output dimension**—into which different modalities are encoded—would boost the model performance.

Hyperparameter	Search Space
embeddings	[OpenAI, ClinicalBERT]
embedding_output_dim	[64, 128, 256]
n_highway	DiscreteUniform(2, 6)
mpnn_depth	DiscreteUniform(2, 10)
epoch	10
pre_training_epoch	DiscreteUniform(10, 30)
lr	Uniform(1e-4, 1e-3)
scheduler	[StepLR, ReduceLROnPlateau]
scheduler_gamma	[0.3, 0.5, 0.8]

Table 2: Search Space of Ablation Study

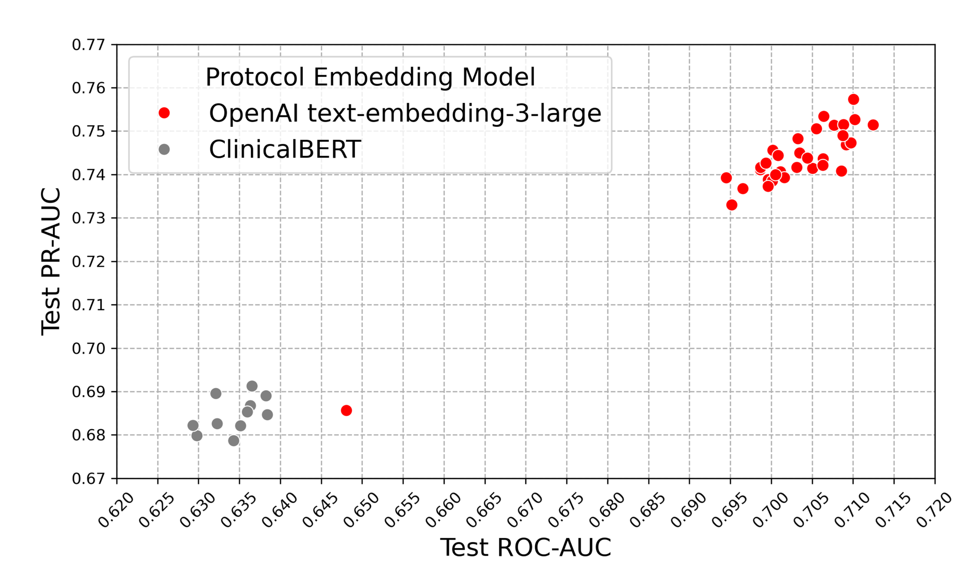

Figure 3: Performance of 44 Ablation Experiments

Hyperparameter	GHINT-v1	GHINT-v2
embeddings	OpenAI	OpenAI
embedding_output_dim	256	256
n_highway	6	5
mpnn_depth	4	3
epoch	10	10
pre_training_epoch	10	22
lr	2.45E-04	5.33E-04
scheduler	StepLR	StepLR
scheduler_gamma	0.5	0.5

Table 3: Hyperparameters of the 2 Best Models

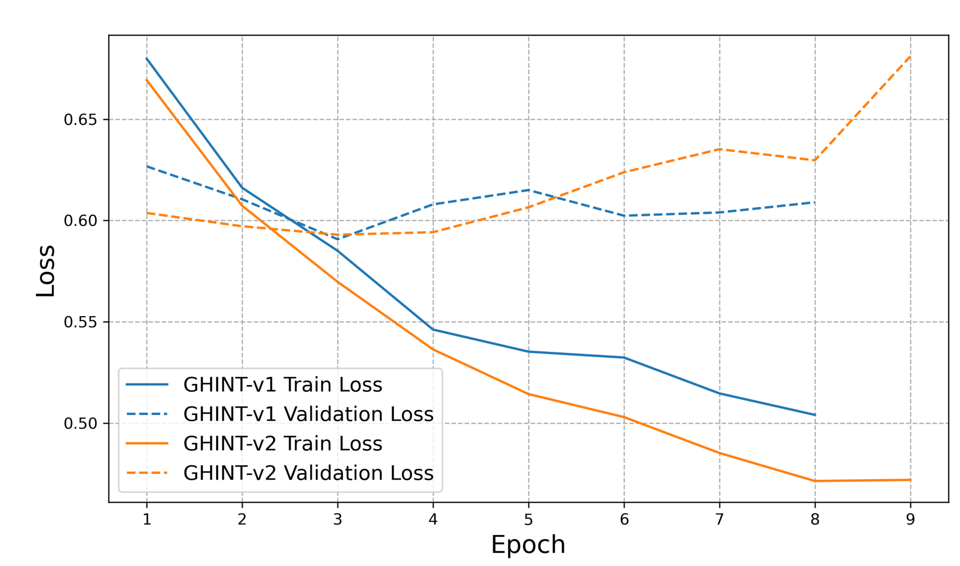

Figure 4: Loss Plot of the 2 Best Models

Results

Model	ROC-AUC	PR-AUC	F1
HINT	0.645	0.629	0.620
GHINT–v1	**0.712**	0.751	**0.714**
GHINT–v2	0.710	**0.757**	0.704
LIFTED	0.651	0.698	0.662

Table 4: Performance Comparison of Models for Predicting Phase II Clinical Trial Outcome

Model Performance
Through the ablation tests, two GPT text-embedding-enhanced models are significantly outperform the baseline HINT and LIFTED (Zheng et al., 2024) on the Phase II dataset.

Structure Supremacy
- GPT-based embeddings outperform BioBERT by capturing richer semantic and contextual information, which is crucial for understanding nuanced clinical trial text
- Unlike HINT's 50-dimension embeddings, our models use 256 dimensions, allowing for more expressive and informative representations.
- Beyond eligibility criteria, we also include sponsor, trial design, etc., as input features to capture more comprehensive trial context.

Discussion

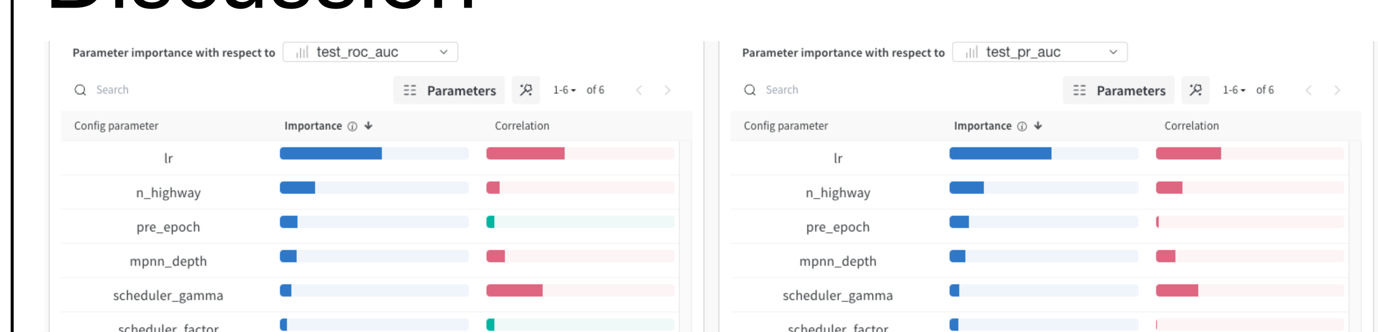

Figure 5: Hyperparameter Importance of Ablation Study

Hyperparameter Impact Assumptions
Our hyperparameter ablation study provides the following insights when using `text-embedding-3-large`:
- Learning Rate Sensitivity: A higher learning rate for AdamW tends to negatively impact performance, likely due to overshooting optimal minima.
- Highway Network Depth: The Highway layer serves as a learnable gating mechanism to control feature transformation across layers. However, increasing the number of Highway layers degrades performance, suggesting that added complexity may lead to overfitting or redundancy, particularly when the input embeddings are already semantically rich.
- Effect of Pretraining Epochs on ADMET encoder: Increasing the number of pretraining epochs on external web lab data improves performance. Longer pretraining provides the encoder with stronger domain-invariant representations, improving downstream generalization to prediction tasks

Bootstrap-Based Performance Evaluation
We propose to adopt bootstrap resampling like HINT to reduce evaluation variance, particularly important for small-sample clinical trial dataset.

Unified Dataset and Phase Integration
Unlike previous researches training models separately on clinical trials of different phases, we propose to train on the full dataset across all trial phases, allowing the model to capture cross-phase relationships, and enables learning from a more diverse training set.