Improving Clinical Trial Outcome Prediction: Causal Machine Unlearning Approach and LLM-Augmented Dataset Enhancement

Teddy Ganea[1,2], Simon Pritchard[3,4], Jacob Rubenstein[1,5]

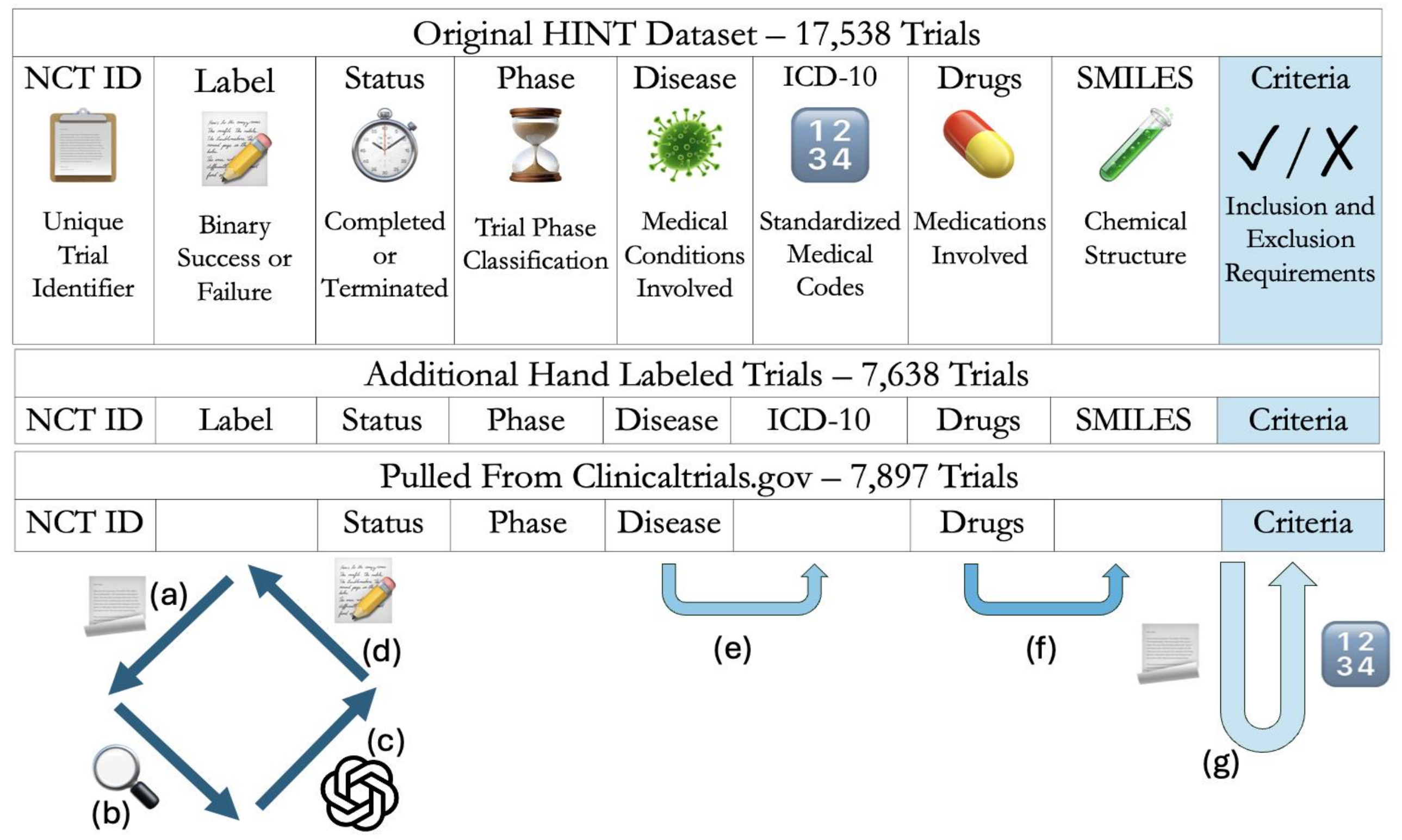
(tganea@stanford.edu, skpritch@stanford.edu, jacobrl@stanford.edu)

1. Department of Mathematics, 2. Department of Classics, 3. Department of Biology, 4. Department of Statistics, 5. Department of Public Policy; Stanford University

INTRODUCTION

Clinical trials are essential to pharmaceutical development, but **only 13.8% of all drug candidates eventually lead to approval, with an average of $2.6 billion dollars spent on each drug eventually brought to market.** A significant challenge is that labeling clinical trial outcomes requires specialized knowledge and is often ambiguous, making it difficult to create large high-quality datasets. **To address this, we expanded the HINT dataset by using LLMs to label additional trials, more than doubling the available training data while updating coverage to recent trials.** Additionally, clinical trial outcome prediction involves addressing spurious correlations that may lead models to focus on confounding factors rather than those that are relevant. **In our project, we extended the Hierarchical Interaction Network (HINT) architecture via dynamic UnLearning from Experience (dULE), a novel student-teacher causal unlearning approach designed to identify and mitigate spurious correlations.**

HINT DATASET AND AUGMENTATION



Original HINT Dataset – 17,538 Trials								
NCT ID	Label	Status	Phase	Disease	ICD-10	Drugs	SMILES	Criteria
Unique Trial Identifier	Binary Success or Failure	Completed or Terminated	Trial Phase Classification	Medical Conditions Involved	Standardized Medical Codes	Medications Involved	Chemical Structure	✓/X Inclusion and Exclusion Requirements
Additional Hand Labeled Trials – 7,638 Trials								
NCT ID	Label	Status	Phase	Disease	ICD-10	Drugs	SMILES	Criteria
Pulled From Clinicaltrials.gov – 7,897 Trials								
NCT ID		Status	Phase	Disease		Drugs		Criteria

Figure 1: (a) Created instructions for GPT-3.5. **(b)** Formatted relevant information. **(c)** Applied LLM to trials with 71% accuracy. **(d)** Incorporated binary success/failures into data. **(e)** Mapped diseases to ICD10 codes. **(f)** Pulled from drugbank API to convert drugs to their chemical structures. **(g)** Text to vectors, reduce dimensions.

LLM DATA LABELING

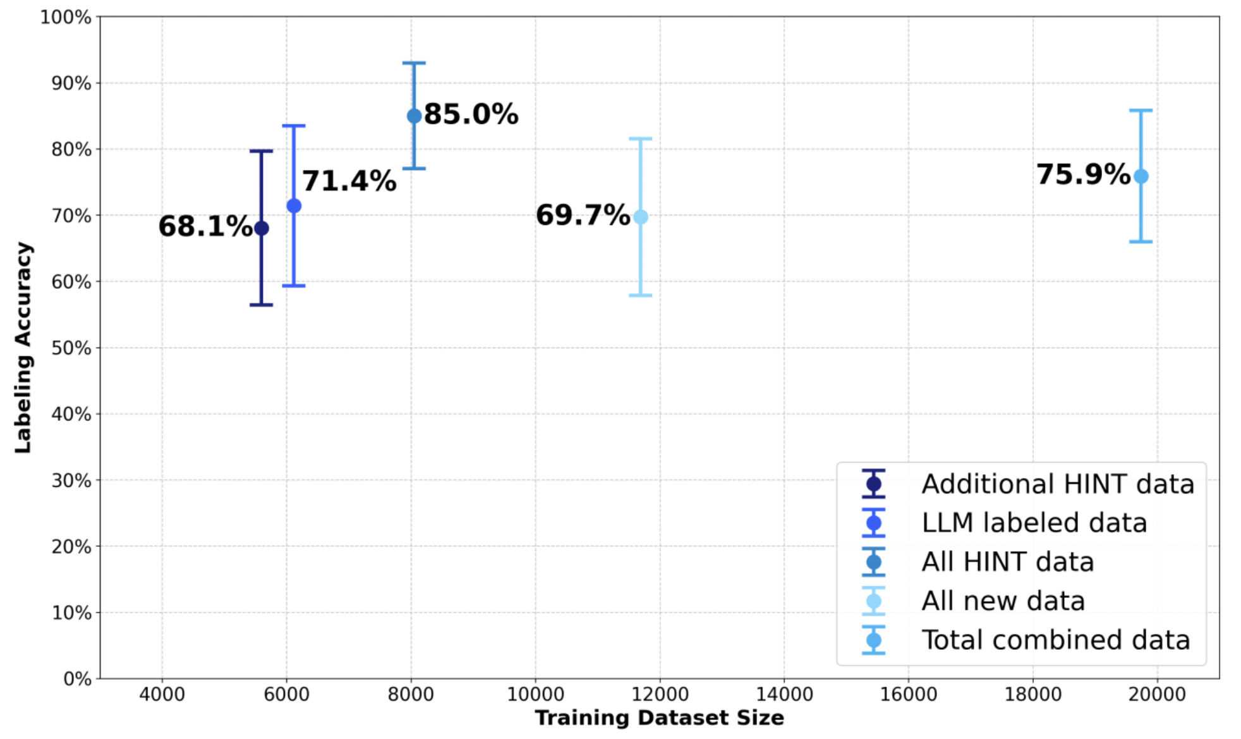


Figure 2: Our self-supervised LLM labeling vastly expands dataset size at the cost of significantly lower accuracy; 66% confidence intervals.

dULE-HINT MODEL ARCHITECTURE

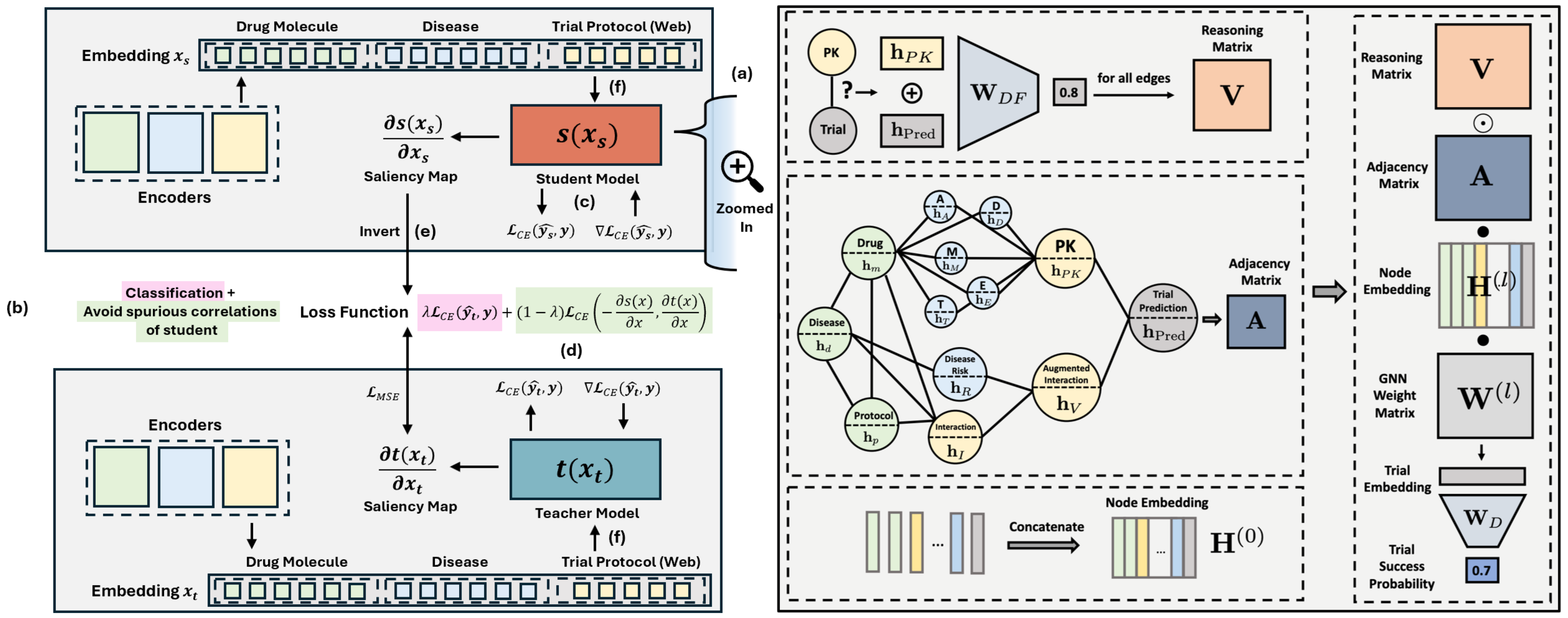


Figure 3: Building on HINT architecture, we implement **d**ynamic **U**n**L**earning from **E**xperience with a custom loss function in a "classroom" student-teacher environment. **(a)** Teacher and student models have identical architecture: embeddings X_s and X_t are processed through a one-layer GNN with self-attention and a causal adjacency matrix to predict trial success or failure. Both X_s and X_t use identical encoder structures, with one ADMET encoder pretrained from HINT. **(b)** Our "classroom" trains student and teacher in parallel. **(c)** Student uses cross-entropy loss backpropagated through both GNN and encoders. **(d)** Teacher loss combines CE and MSE between the teacher and the student's negative gradient. We innovate with dULE by dynamically adjusting lambda over epochs. **(e)** Using the negative student gradient penalizes the teacher from pursuing the same (potentially spurious) correlations. **(f)** We apply dULE to our graph models, computing gradients against the inputs (our encoding outputs X_s and X_t).

dULE-TEACHER TENDS TO OUTPERFORM STUDENT

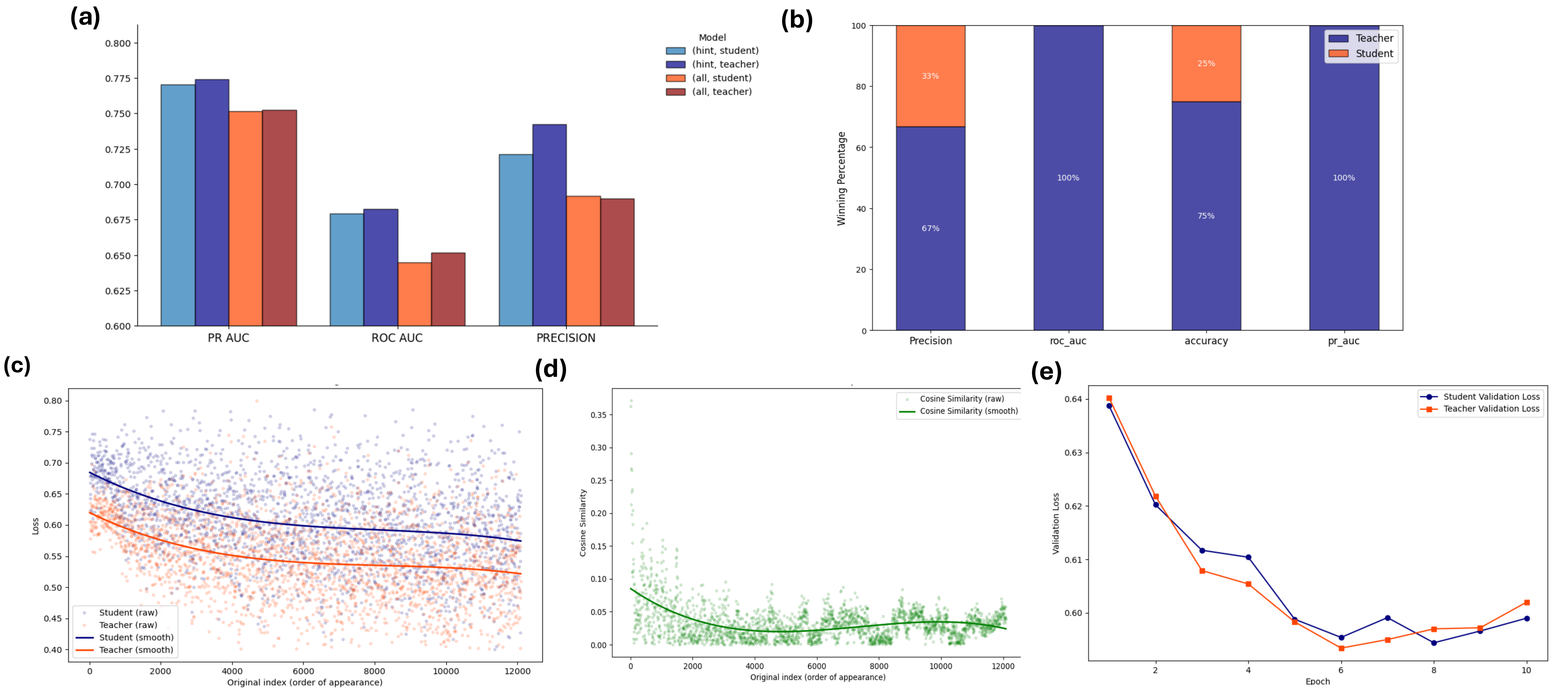


Figure 4: dULE-HINT model results and training behavior. **(a)** Respective performances of best model (cos_hint_0.9_0.0) on PR_AUC, ROC_AUC, and Precision. Teacher models tend to outperform student. **(b)** Teachers outperform pairwise students for superior performance within the top 50 models. **(c)** Teacher, student training loss across epochs. **(d)** Teacher-student gradient cosine similarities are low across epochs. **(e)** Teacher model achieves lower validation loss than student.

Robustness of Hyperparameter Tuning

Role	Dataset	ULE Loss	Lambda	d	Epoch	ROC-AUC	PR-AUC	Precision	Accuracy	F1
Student	HINT				5	0.682	0.768	0.739	0.645	0.699
Student	All				2	0.650	0.751	0.669	0.643	**0.749**
Teacher	HINT	mse	0.7	0.0	5	0.680	0.768	**0.751**	0.643	0.688
Teacher	HINT	mse	0.7	0.02	4	0.675	0.767	0.706	0.648	0.724
Teacher	HINT	mse	0.9	0.0	5	0.675	0.768	0.743	0.644	0.699
Teacher	HINT	mse	0.9	0.02	5	0.679	0.769	0.741	0.646	0.699
Teacher	HINT	cos	0.7	0.0	7	0.671	0.764	0.749	0.639	0.683
Teacher	HINT	cos	0.7	0.02	7	0.676	0.759	0.745	**0.651**	0.703
Teacher	HINT	cos	0.9	0.0	4	**0.683**	**0.773**	0.712	0.650	0.722
Teacher	HINT	cos	0.9	0.02	5	0.679	0.768	0.732	0.638	0.693
Teacher	All	mse	0.7	0.0	6	0.639	0.738	0.687	0.638	0.728
Teacher	All	mse	0.7	0.02	8	0.645	0.743	0.697	0.632	0.712
Teacher	All	mse	0.9	0.0	7	0.647	0.748	0.699	0.647	0.731
Teacher	All	mse	0.9	0.02	7	0.644	0.744	0.691	0.646	0.735
Teacher	All	cos	0.7	0.0	3	0.647	0.749	0.691	0.642	0.730
Teacher	All	cos	0.7	0.02	4	0.652	0.752	0.690	0.648	0.738
Teacher	All	cos	0.9	0.0	8	0.643	0.741	0.694	0.642	0.728
Teacher	All	cos	0.9	0.02	6	0.646	0.747	0.695	0.638	0.722

Table 1: Test statistics for our highest-performing models on "all" vs. "hint" data, and MSE vs. cosine-similarity based loss function, and various hyperparameter options. Performance is quite similar overall, with test accuracy, especially, being incredibly robust.

DISCUSSION

- Teacher models, especially with cosine-based loss functions, outperformed baseline while training nearly orthogonal to the student's gradient. It is notable this computer vision-inspired technique, with adaptations, succeeds with text data.
- More data didn't improve performance. While doubling size, we increased dataset heterogeneity and inaccuracy, making signals harder to learn. Greater prompt engineering, search access, and more advanced LLMs could avoid this accuracy penalty.
- Our only modest improvement on baseline suggests that clinical trial prediction is too complex a task for a one-layer graph neural network. State-of-the-art isn't an adequate starter architecture.
- Future work could involve modifying the loss function by taking a product of MSE and cosine-similarity, combining both a directionality and magnitude penalty to help our teacher model better explore the loss landscape.
- Additionally, we could ensemble with and/or compare to careful feature engineering via clustering algorithms fed into traditional ML algorithms like Random Forest.

REFERENCES

1. Tianfan Fu, Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Hint: Hierarchical interaction net-work for clinical-trial-outcome predictions. Patterns, 3(4):100445, Feb 2022.
2. Jeff Mitchell, Jesus Martınez del Rincon, and Niall McLaughlin. Unlearning from experience to avoid spurious correlations, 2024.

Role	Dataset	Loss Function	Lambda	d	Epoch	ROC-AUC	PR-AUC	Precision	Accuracy	F1
Student	HINT	[redacted]	[redacted]	[redacted]	5	0.682	0.768	0.739	0.645	0.699
Student	All	[redacted]	[redacted]	[redacted]	2	0.650	0.751	0.669	0.643	0.749
Teacher	HINT	mse	0.7	0	5	0.680	0.768	0.751	0.643	0.688
Teacher	HINT	mse	0.7	0.02	4	0.675	0.767	0.706	0.648	0.724
Teacher	HINT	mse	0.9	0	5	0.675	0.768	0.743	0.644	0.699
Teacher	HINT	mse	0.9	0.02	5	0.679	0.769	0.741	0.646	0.699
Teacher	HINT	cos	0.7	0	7	0.671	0.764	0.749	0.639	0.683
Teacher	HINT	cos	0.7	0.02	7	0.676	0.759	0.745	0.651	0.703
Teacher	HINT	cos	0.9	0	4	0.683	0.773	0.712	0.650	0.722
Teacher	HINT	cos	0.9	0.02	5	0.679	0.768	0.732	0.638	0.693
Teacher	All	mse	0.7	0	6	0.639	0.738	0.687	0.638	0.728
Teacher	All	mse	0.7	0.02	8	0.645	0.743	0.697	0.632	0.712
Teacher	All	mse	0.9	0	7	0.647	0.748	0.699	0.647	0.731
Teacher	All	mse	0.9	0.02	7	0.644	0.744	0.691	0.646	0.735
Teacher	All	cos	0.7	0	3	0.647	0.749	0.691	0.642	0.730
Teacher	All	cos	0.7	0.02	4	0.652	0.752	0.690	0.648	0.738
Teacher	All	cos	0.9	0	8	0.643	0.741	0.694	0.642	0.728
Teacher	All	cos	0.9	0.02	6	0.646	0.747	0.695	0.638	0.722

Improving Clinical Trial Outcome Prediction: Causal Machine Unlearning Approach and LLM Augmented Dataset Enhancement

Teddy Ganea[1,2], Simon Pritchard[3,4], Jacob Rubenstein[1,5]

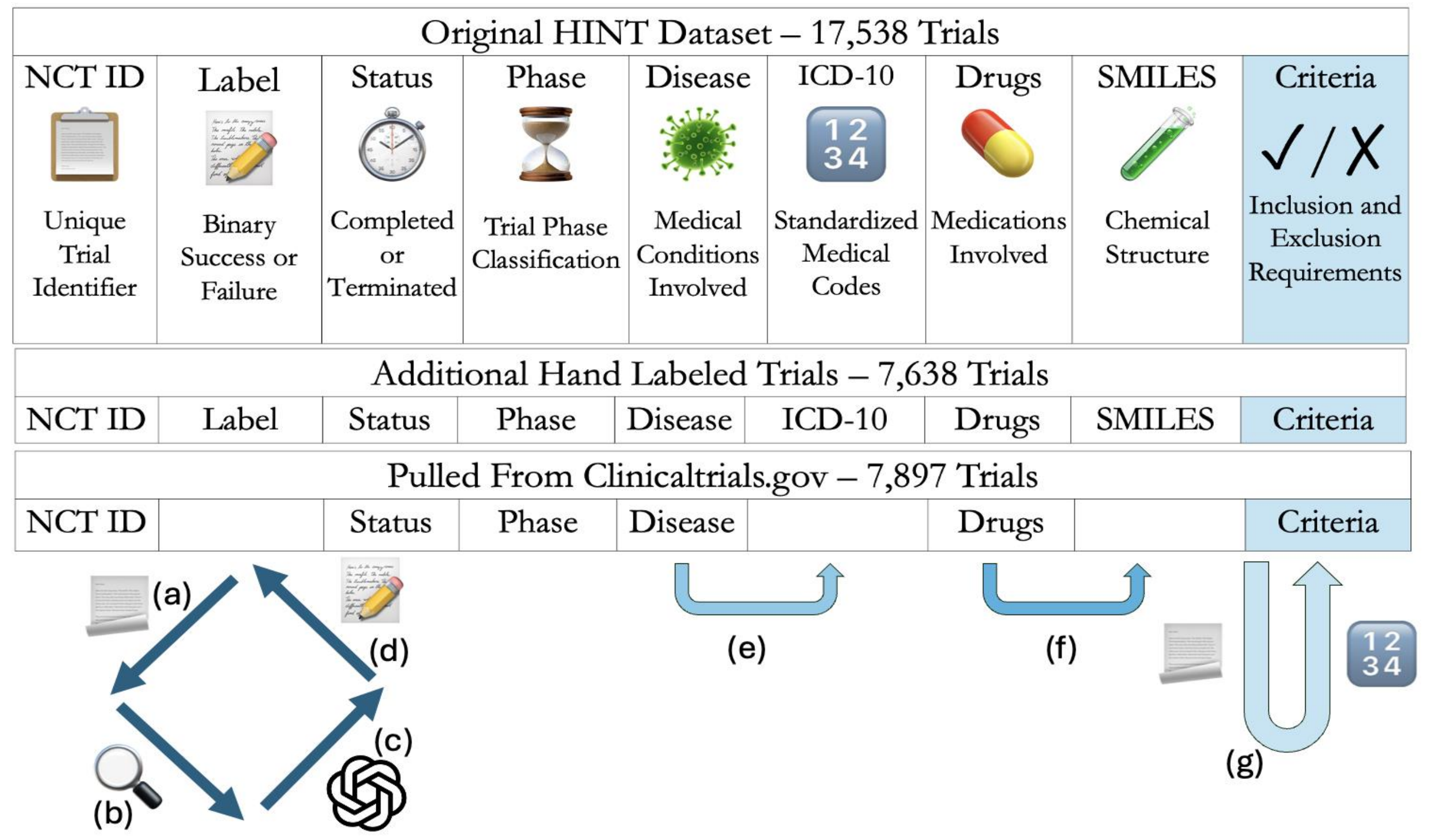
(tganea@stanford.edu, skpritch@stanford.edu, jacobr1@stanford.edu)

1. Department of Mathematics, 2. Department of Classics, 3. Department of Biology, 4. Department of Statistics, 5. Department of Policy; Stanford University

Abstract

Clinical trials are essential to pharmaceutical development, but **only 13.8% of all drug candidates eventually lead to approval, with an average of $2.6 billion dollars spent on each drug eventually brought to market.** Labeling clinical trial outcomes requires specialized knowledge and is often ambiguous, making it difficult to create large high-quality datasets. **To address this, we expanded the HINT dataset by using large language models to label additional trials, more than doubling the available training data while updating coverage to recent trials.** Additionally, a significant challenge in clinical trial outcome prediction involves addressing spurious correlations that may lead models to focus on confounding factors rather than those that are relevant. **In our project, we extended the Hierarchical Interaction Network (HINT) architecture via dynamic UnLearning from Experience (dULE), a novel student-teacher causal unlearning approach designed to identify and mitigate spurious correlations.**

Dataset + Augmentation



Original HINT Dataset – 17,538 Trials								
NCT ID	Label	Status	Phase	Disease	ICD-10	Drugs	SMILES	Criteria
Unique Trial Identifier	Binary Success or Failure	Completed or Terminated	Trial Phase Classification	Medical Conditions Involved	Standardized Medical Codes	Medications Involved	Chemical Structure	✓ / X Inclusion and Exclusion Requirements

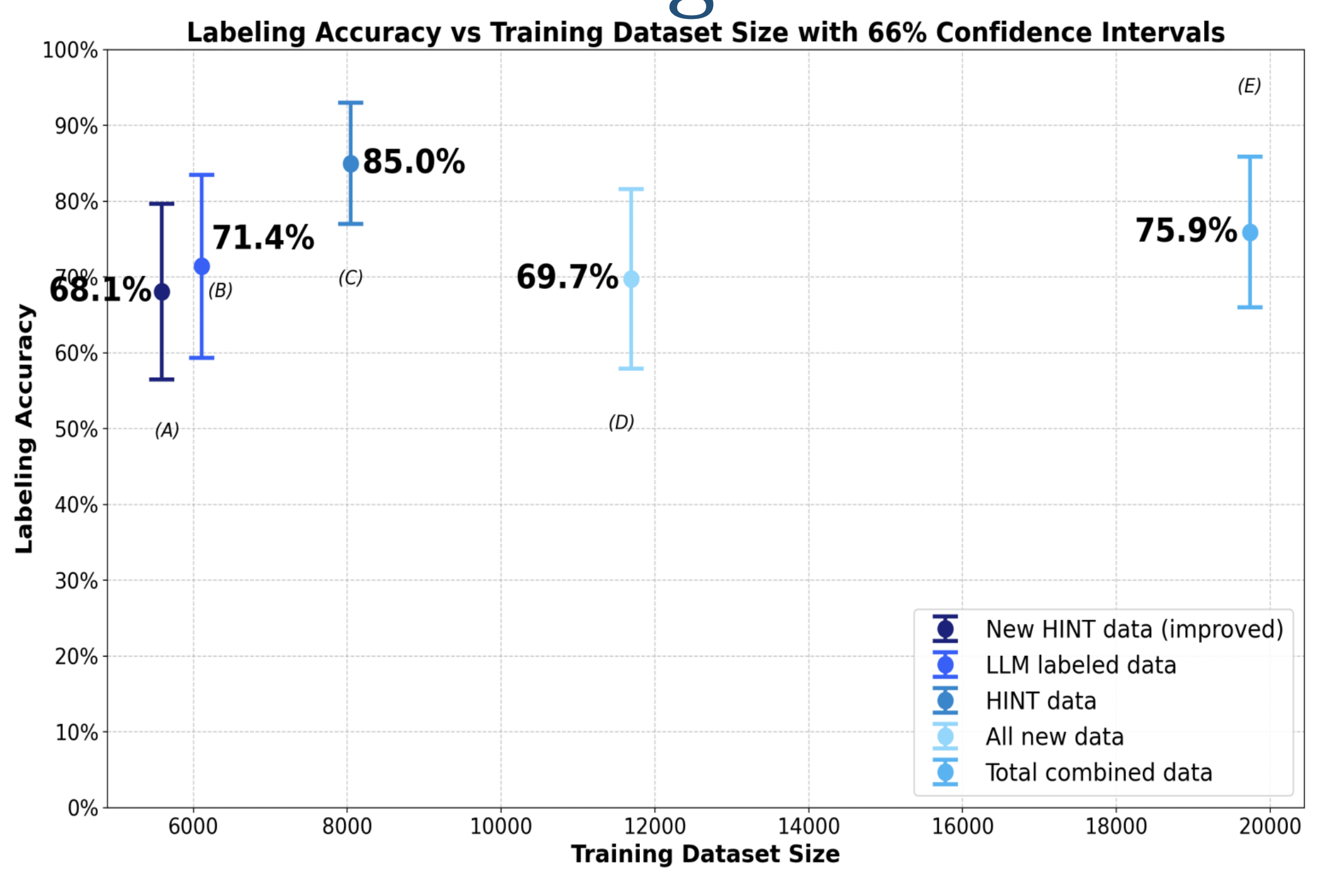
Additional Hand Labeled Trials – 7,638 Trials								
NCT ID	Label	Status	Phase	Disease	ICD-10	Drugs	SMILES	Criteria

Pulled From Clinicaltrials.gov – 7,897 Trials								
NCT ID		Status	Phase	Disease		Drugs		Criteria

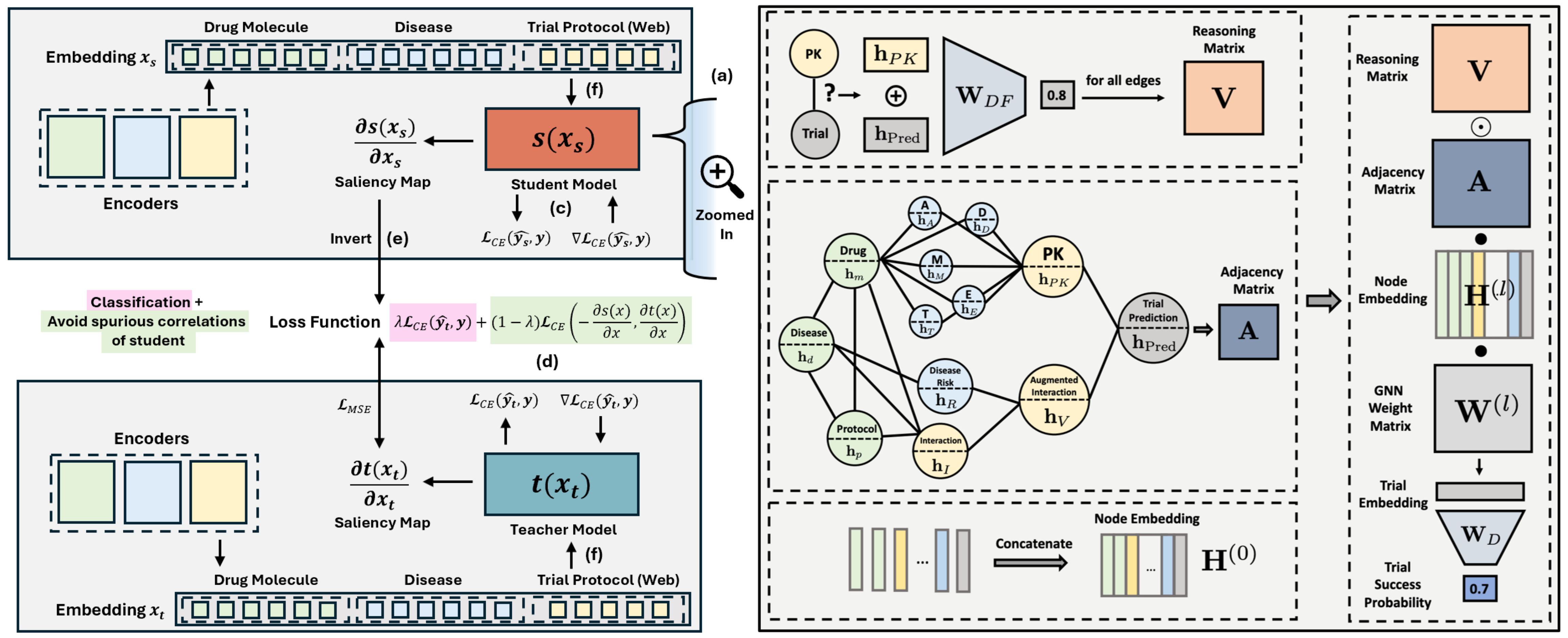
(a) Created trial-specific instructions for GPT-3.5; **(b)** Identified key data signals and refined prompts; **(c)** Applied LLM to trials with 71% accuracy; **(d)** Incorporated binary success/failures into data; **(e)** Created standardized mapping from diseases to ICD-10 codes; **(f)** Pulled from drugbank API to convert drugs to their chemical structures; **(g)** Text to vectors, reduce dimensions, maintain consistent features.

LLM Labeling

- GPT-3.5 classified trial outcomes as success/failure based on endpoints, significance, and termination reasons.
- Framework achieved 85% validation accuracy and 71% on manual test sets.
- Enabled efficient processing of 7,897 trials, trading slight accuracy reduction for larger, more diverse dataset.



Causal Unlearning Methodology



DULE-HINT Model Architecture: Building on HINT architecture, we implement dynamic UnLearning from Experience with a custom loss function in a "classroom" student-teacher environment.

(a) Teacher and student have identical architecture: embeddings X_s and X_t are processed through a one-layer GNN with self-attention and a causal adjacency matrix to predict trial success or failure. Both use identical encoder structures that process clinical trial protocol, drug, and disease data, with one ADMET encoder pretrained from HINT.
(b) Our "classroom" trains student and teacher in parallel using separate losses.
(c) The student uses binary cross-entropy loss backpropagated through both model and encoders.

(d) The teacher's loss combines BCE and MSE between the teacher and the student's negative gradient. This ULE approach avoids overpowering background correlations in the loss landscape. We innovate with dULE by dynamically adjusting lambda as student and teacher diverge.
(e) Using the negative student gradient penalizes the teacher from pursuing the same correlations.
(f) We apply dULE to our graph models, computing gradients against the inputs (our encoding outputs X_s and X_t).

Results #1

Results #2

Discussion

References

Tianfan Fu, Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Hint: Hierarchical interaction net-work for clinical-trial-outcome predictions. Patterns, 3(4):100445, Feb 2022.

Jeff Mitchell, Jesus Martınez del Rincon, and Niall McLaughlin. Unlearning from experience to avoid spurious correlations, 2024.

Improving Clinical Trial Outcome Prediction: Causal Machine Unlearning Approach and LLM Augmented Dataset Enhancement

Teddy Ganea[1,2], Simon Pritchard[3,4], Jacob Rubenstein[1,5]

(tganea@stanford.edu, skpritch@stanford.edu, jacobr1@stanford.edu)

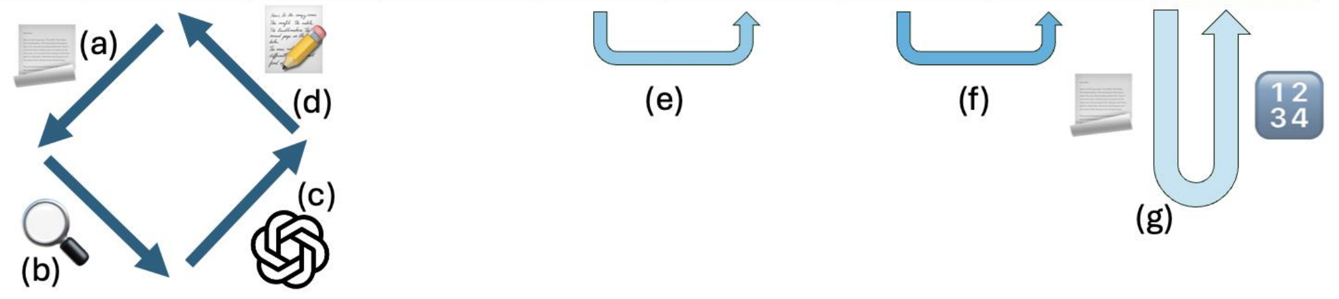
1. Department of Mathematics, 2. Department of Classics, 3. Department of Biology, 4. Department of Statistics, 5. Department of Policy; Stanford University

Abstract

Clinical trials are essential to pharmaceutical development, but **only 13.8% of all drug candidates eventually lead to approval, with an average of $2.6 billion dollars spent on each drug eventually brought to market.** Labeling clinical trial outcomes requires specialized knowledge and is often ambiguous, making it difficult to create large high-quality datasets. **To address this, we expanded the HINT dataset by using large language models to label additional trials, more than doubling the available training data while updating coverage to recent trials.** Additionally, a significant challenge in clinical trial outcome prediction involves addressing spurious correlations that may lead models to focus on confounding factors rather than those that are relevant. **In our project, we extended the Hierarchical Interaction Network (HINT) architecture via dynamic UnLearning from Experience (dULE), a novel student-teacher causal unlearning approach designed to identify and mitigate spurious correlations.**

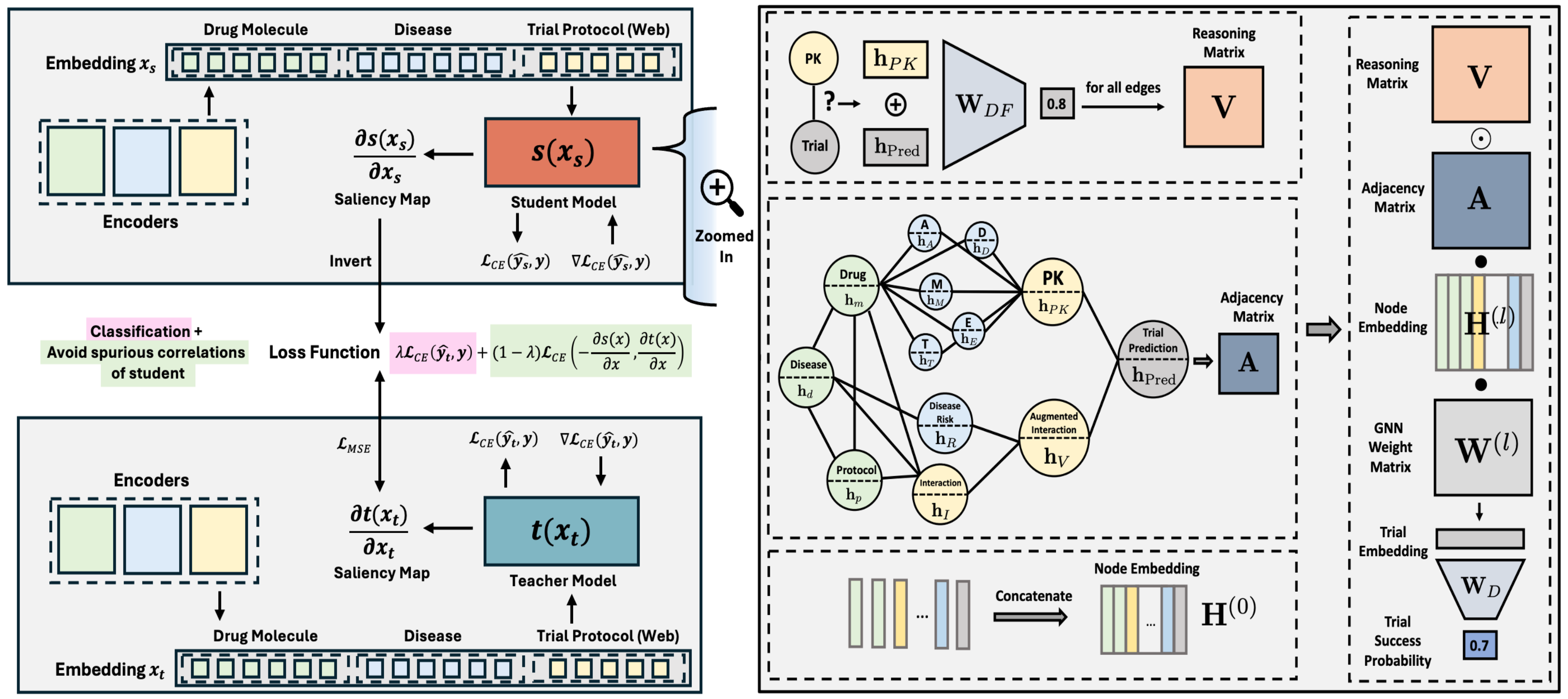
Dataset + Augmentation

Original HINT Dataset – 17,538 Trials								
NCT ID	Label	Status	Phase	Disease	ICD-10	Drugs	SMILES	Criteria
Unique Trial Identifier	Binary Success or Failure	Completed or Terminated	Trial Phase Classification	Medical Conditions Involved	Standardized Medical Codes	Medications Involved	Chemical Structure	✓ / X Inclusion and Exclusion Requirements
Additional Hand Labeled Trials – 7,638 Trials								
NCT ID	Label	Status	Phase	Disease	ICD-10	Drugs	SMILES	Criteria
Pulled From Clinicaltrials.gov – 7,897 Trials								
NCT ID		Status	Phase	Disease		Drugs		Criteria



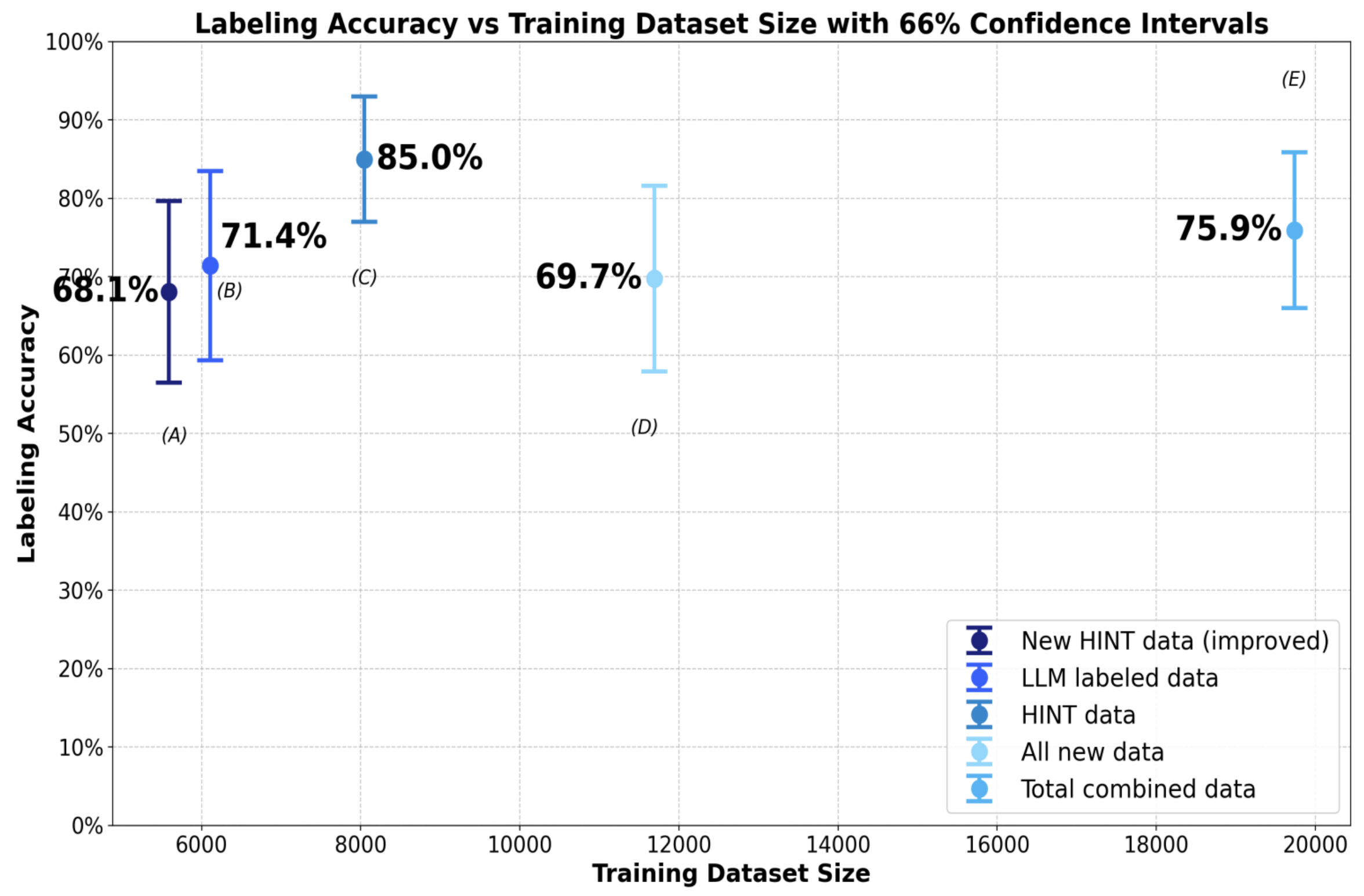
(a) Created trial-specific instructions for GPT-3.5; (b) Identified key data signals and refined prompts; (c) Applied classifier to trials with 71% accuracy; (d) Added binary success/failures back to data; (e) Created standardized mapping from diseases to ICD-10 codes; (f) Pulled from drugbank API to convert drugs to their chemical structures; (g) Text to vectors, reduce dimensions, maintain consistent features.

Causal Unlearning Methodology



DULE-HINT Model Architecture: Building on HINT architecture, we implement dynamic UnLearning from Experience with a custom loss function in a "classroom" student-teacher environment. (a) Teacher and student have identical architecture: embeddings X_s and X_t are processed through a one-layer GNN with self-attention and a causal adjacency matrix to predict trial success or failure. Both use identical encoder structures that process clinical trial protocol, drug, and disease data, with one ADMET encoder pretrained from HINT. (b) Our "classroom" trains student and teacher in parallel using separate losses. (c) The student uses binary cross-entropy loss backpropagated through both model and encoders. (d) The teacher's loss combines BCE and MSE between the teacher and the student's negative gradient. This ULE approach avoids overpowering background correlations in the loss landscape. We innovate with dULE by dynamically adjusting lambda as student and teacher diverge. (e) Using the negative student gradient penalizes the teacher from pursuing the same correlations. (f) We apply dULE to our graph models, computing gradients against the inputs (our encoding outputs X_s and X_t).

LLM Labeling



We developed a system using GPT-3.5 to classify clinical trial outcomes as successful (1) or failed (0), addressing the manual annotation bottleneck. Our hierarchical classification framework incorporated domain-specific knowledge about endpoint achievement, statistical significance, and termination reasons. Through iterative refinement, we optimized prompts to maximize accuracy, achieving 85% on validation and 71% on a manually-labeled test set. This approach enabled us to process 7,897 trials that would have been impractical to annotate manually.
While manual annotation offers slightly higher accuracy, our LLM method gave us a significantly larger, more diverse dataset in a fraction of the time, a valuable tradeoff for model training.

Results

Future Work

Improving Clinical Trial Outcome Prediction: Causal Machine Unlearning Approach and LLM Augmented Dataset Enhancement

Teddy Ganea[1,2], Simon Pritchard[3,4], Jacob Rubenstein[1,5]

(tganea@stanford.edu, skpritch@stanford.edu, jacobr1@stanford.edu)

1. Department of Mathematics, 2. Department of Classics, 3. Department of Biology, 4. Department of Statistics, 5. Department of Policy; Stanford University

Abstract

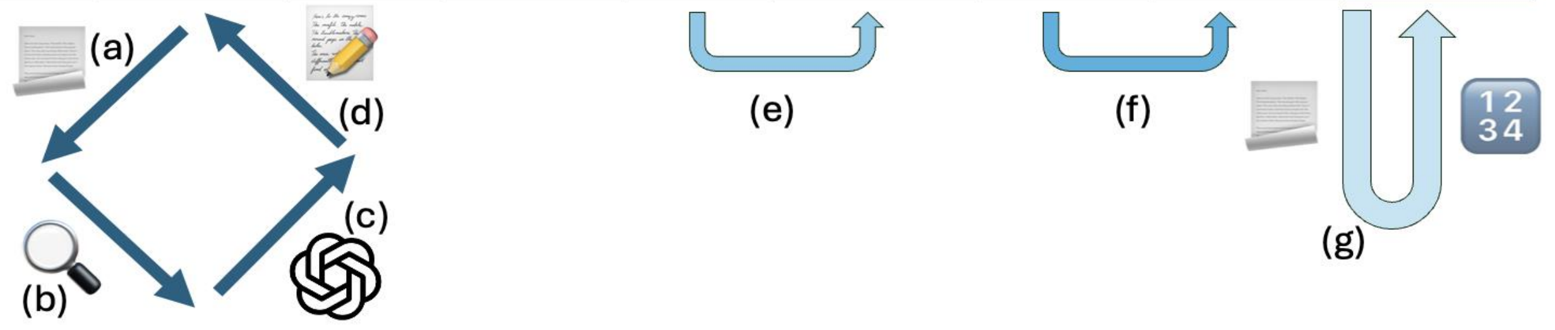
Clinical trials are essential to pharmaceutical development, but **only 13.8% of all drug candidates eventually lead to approval, with an average of $2.6 billion dollars spent on each drug eventually brought to market.** Labeling clinical trial outcomes requires specialized knowledge and is often ambiguous, making it difficult to create large high-quality datasets. **To address this, we expanded the HINT dataset by using large language models to label additional trials, more than doubling the available training data while updating coverage to recent trials.** Additionally, a significant challenge in clinical trial outcome prediction involves addressing spurious correlations that may lead models to focus on confounding factors rather than those that are relevant. **In our project, we extended the Hierarchical Interaction Network (HINT) architecture via dynamic UnLearning from Experience (dULE), a novel student-teacher causal unlearning approach designed to identify and mitigate spurious correlations.**

Dataset + Augmentation

Original HINT Dataset – 17,538 Trials								
NCT ID	Label	Status	Phase	Disease	ICD-10	Drugs	SMILES	Criteria
Unique Trial Identifier	Binary Success or Failure	Completed or Terminated	Trial Phase Classification	Medical Conditions Involved	Standardized Medical Codes	Medications Involved	Chemical Structure	✓ / X Inclusion and Exclusion Requirements

Additional Hand Labeled Trials – 7,638 Trials								
NCT ID	Label	Status	Phase	Disease	ICD-10	Drugs	SMILES	Criteria

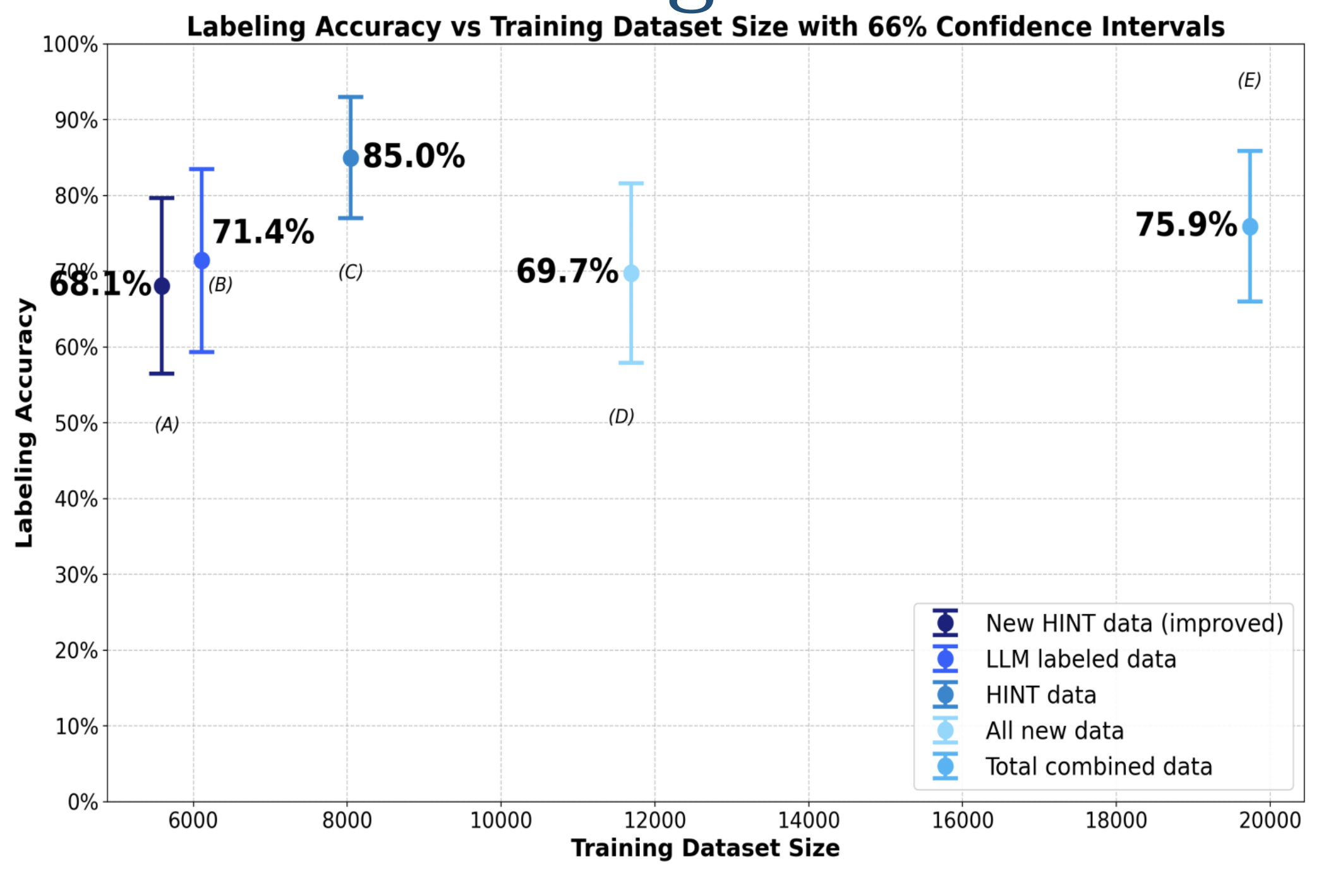
Pulled From Clinicaltrials.gov – 7,897 Trials								
NCT ID		Status	Phase	Disease		Drugs		Criteria



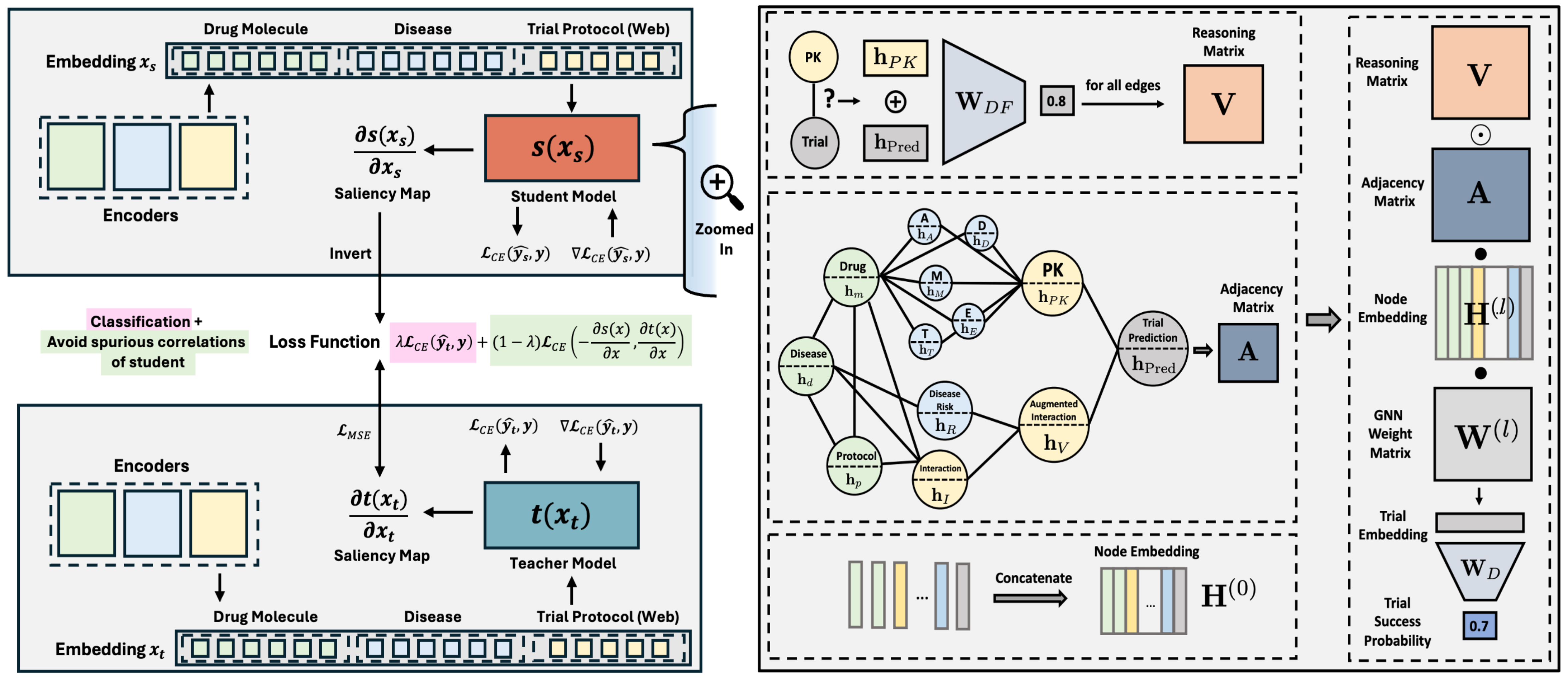
(a) Created trial-specific instructions for GPT-3.5; **(b)** Identified key data signals and refined prompts; **(c)** Applied LLM to trials with 71% accuracy; **(d)** Incorporated binary success/failures into data; **(e)** Created standardized mapping from diseases to ICD-10 codes; **(f)** Pulled from drugbank API to convert drugs to their chemical structures; **(g)** Text to vectors, reduce dimensions, maintain consistent features.

LLM Labeling

- GPT-3.5 classified trial outcomes as success/failure based on endpoints, significance, and termination reasons.
- Framework achieved 85% validation accuracy and 71% on manual test sets.
- Enabled efficient processing of 7,897 trials, trading slight accuracy reduction for larger, more diverse dataset.



Causal Unlearning Methodology



DULE-HINT Model Architecture: Building on HINT architecture, we implement dynamic UnLearning from Experience with a custom loss function in a "classroom" student-teacher environment. **(a)** Teacher and student have identical architecture: embeddings X_s and X_t are processed through a one-layer GNN with self-attention and a causal adjacency matrix to predict trial success or failure. Both use identical encoder structures that process clinical trial protocol, drug, and disease data, with one ADMET encoder pretrained from HINT. **(b)** Our "classroom" trains student and teacher in parallel using separate losses. **(c)** The student uses binary cross-entropy loss backpropagated through both model and encoders. **(d)** The teacher's loss combines BCE and MSE between the teacher and the student's negative gradient. This ULE approach avoids overpowering background correlations in the loss landscape. We innovate with dULE by dynamically adjusting lambda as student and teacher diverge. **(e)** Using the negative student gradient penalizes the teacher from pursuing the same correlations. **(f)** We apply dULE to our graph models, computing gradients against the inputs (our encoding outputs X_s and X_t).

Results #1

Results #2

Discussion

References

Tianfan Fu, Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Hint: Hierarchical interaction net-work for clinical-trial-outcome predictions. Patterns, 3(4):100445, Feb 2022.

Jeff Mitchell, Jesus Martınez del Rincon, and Niall McLaughlin. Unlearning from experience to avoid spurious correlations, 2024.

Original HINT Dataset – 17,538 Trials

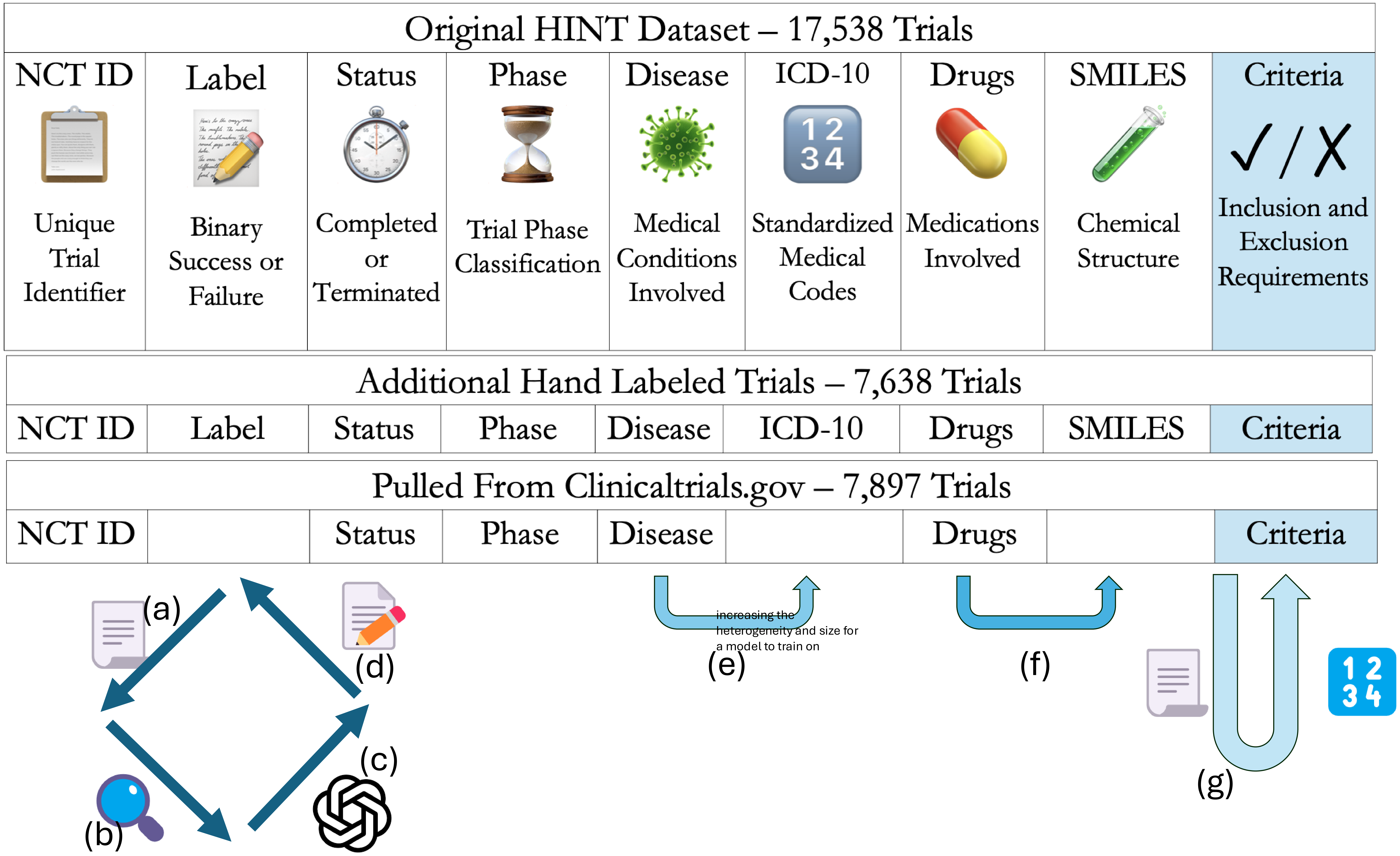
NCT ID	Label	Status	Phase	Disease	ICD-10	Drugs	SMILES	Criteria
Unique Trial Identifier	Binary Success or Failure	Completed or Terminated	Trial Phase Classification	Medical Conditions Involved	Standardized Medical Codes	Medications Involved	Chemical Structure	✓ / ✗ Inclusion and Exclusion Requirements

Additional Hand Labeled Trials – 7,638 Trials

ICT ID	Label	Status	Phase	Disease	ICD-10	Drugs	SMILES	Criteria

Pulled From Clinicaltrials.gov – 7,897 Trials

ICT ID		Status	Phase	Disease		Drugs		Criteria



Original HINT Dataset – 17,538 Trials								
NCT ID	Label	Status	Phase	Disease	ICD-10	Drugs	SMILES	Criteria
Unique Trial Identifier	Binary Success or Failure	Completed or Terminated	Trial Phase Classification	Medical Conditions Involved	Standardized Medical Codes	Medications Involved	Chemical Structure	✓ / X Inclusion and Exclusion Requirements

Additional Hand Labeled Trials – 7,638 Trials								
NCT ID	Label	Status	Phase	Disease	ICD-10	Drugs	SMILES	Criteria

Pulled From Clinicaltrials.gov – 7,897 Trials								
NCT ID		Status	Phase	Disease		Drugs		Criteria

(a)
(d)
(b)
(c)
increasing the heterogeneity and size for a model to train on
(e)
(f)
1 2 3 4
(g)

Improving Clinical Trial Outcome Prediction: Causal Machine Unlearning Approach and LLM Augmented Dataset Enhancement

Teddy Ganea[1,2], Simon Pritchard[3,4], Jacob Rubenstein[1,5]

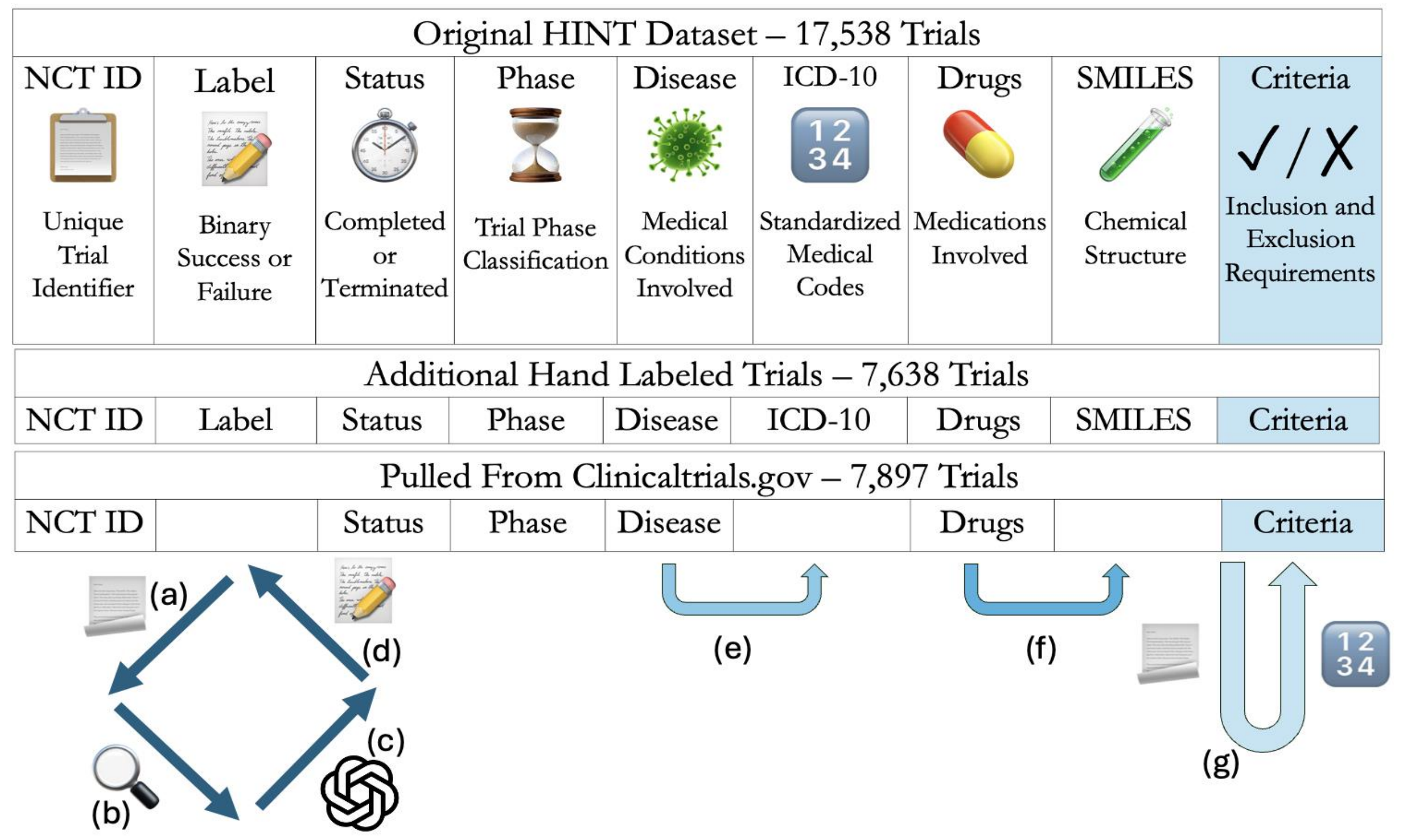
(tganea@stanford.edu, skpritch@stanford.edu, jacobr1@stanford.edu)

1. Department of Mathematics, 2. Department of Classics, 3. Department of Biology, 4. Department of Statistics, 5. Department of Public Policy; Stanford University

ABSTRACT

Clinical trials are essential to pharmaceutical development, but **only 13.8% of all drug candidates eventually lead to approval, with an average of $2.6 billion dollars spent on each drug eventually brought to market.** Labeling clinical trial outcomes requires specialized knowledge and is often ambiguous, making it difficult to create large high-quality datasets. **To address this, we expanded the HINT dataset by using LLMs to label additional trials, more than doubling the available training data while updating coverage to recent trials.** Additionally, a significant challenge in clinical trial outcome prediction involves addressing spurious correlations that may lead models to focus on confounding factors rather than those that are relevant. **In our project, we extended the Hierarchical Interaction Network (HINT) architecture via dynamic UnLearning from Experience (dULE), a novel student-teacher causal unlearning approach designed to identify and mitigate spurious correlations.**

HINT DATASET AND AUGMENTATION



Original HINT Dataset – 17,538 Trials								
NCT ID	Label	Status	Phase	Disease	ICD-10	Drugs	SMILES	Criteria
Unique Trial Identifier	Binary Success or Failure	Completed or Terminated	Trial Phase Classification	Medical Conditions Involved	Standardized Medical Codes	Medications Involved	Chemical Structure	Inclusion and Exclusion Requirements

Additional Hand Labeled Trials – 7,638 Trials								
NCT ID	Label	Status	Phase	Disease	ICD-10	Drugs	SMILES	Criteria

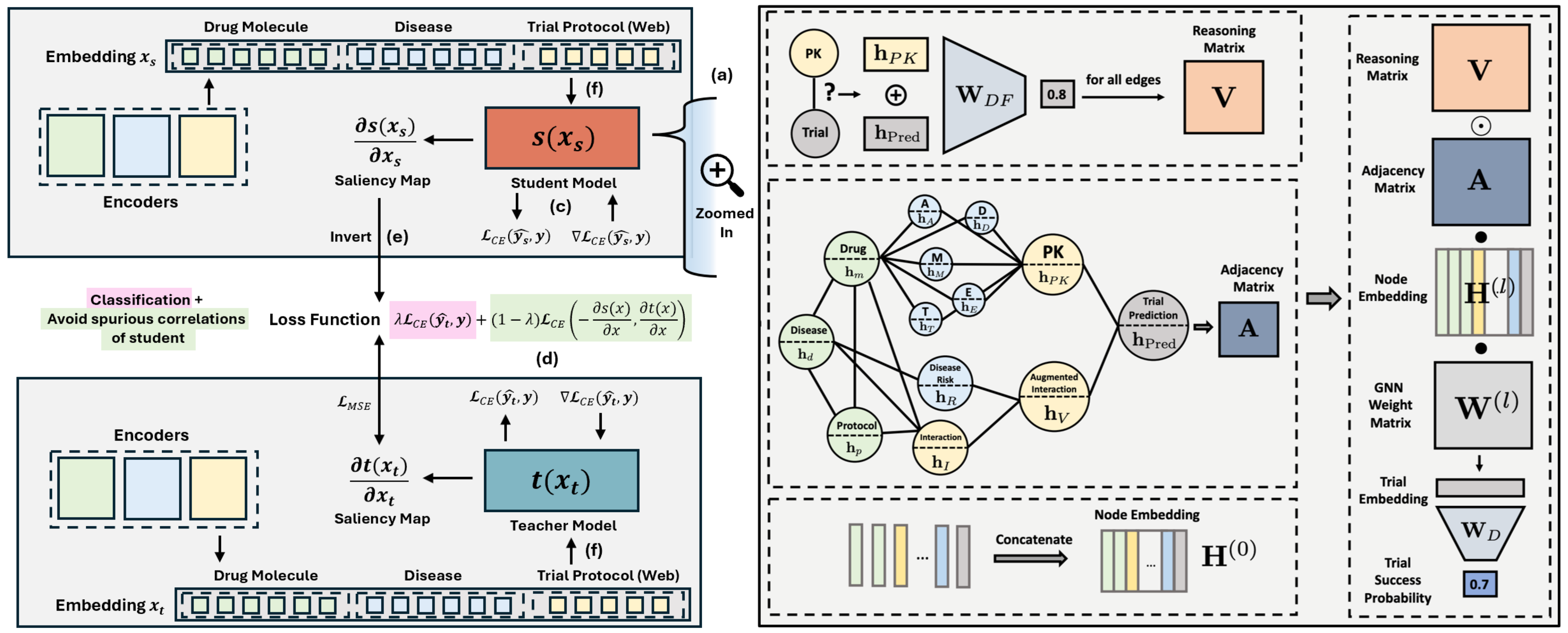
Pulled From Clinicaltrials.gov – 7,897 Trials								
NCT ID		Status	Phase	Disease		Drugs		Criteria

Figure 1: **(a)** Created instructions for GPT-3.5; **(b)** Engineered relevant prompts; **(c)** Applied LLM to trials with 71% accuracy; **(d)** Incorporated binary success/failures into data; **(e)** Mapped diseases to ICD10 codes; **(f)** Pulled from drugbank API to convert drugs to their chemical structures; **(g)** Text to vectors, reduce dimensions

LLM DATA LABELING

Figure 2: Our self-supervised LLM labeling achieves marginally lower accuracy while increasing dataset size; 66% confidence intervals.

dULE-HINT MODEL ARCHITECTURE



Building on HINT architecture, we implement **d**ynamic **U**n**L**earning from **E**xperience with a custom loss function in a "classroom" student-teacher environment. **(a)** Teacher and student have identical architecture: embeddings X_s and X_t are processed through a one-layer GNN with self-attention and a causal adjacency matrix to predict trial success or failure. Both X_s and X_t use identical encoder structures, with one ADMET encoder pretrained from HINT. **(b)** Our "classroom" trains student and teacher in parallel. **(c)** Student uses cross-entropy loss backpropagated through both GNN and encoders. **(d)** Teacher loss combines CE and MSE between the teacher and the student's negative gradient. We innovate with dULE by dynamically adjusting lambda over epochs. **(e)** Using the negative student gradient penalizes the teacher from pursuing the same (potentially spurious) correlations. **(f)** We apply dULE to our graph models, computing gradients against the inputs (our encoding outputs X_s and X_t).

dULE-TEACHER TENDS TO OUTPERFORM STUDENT

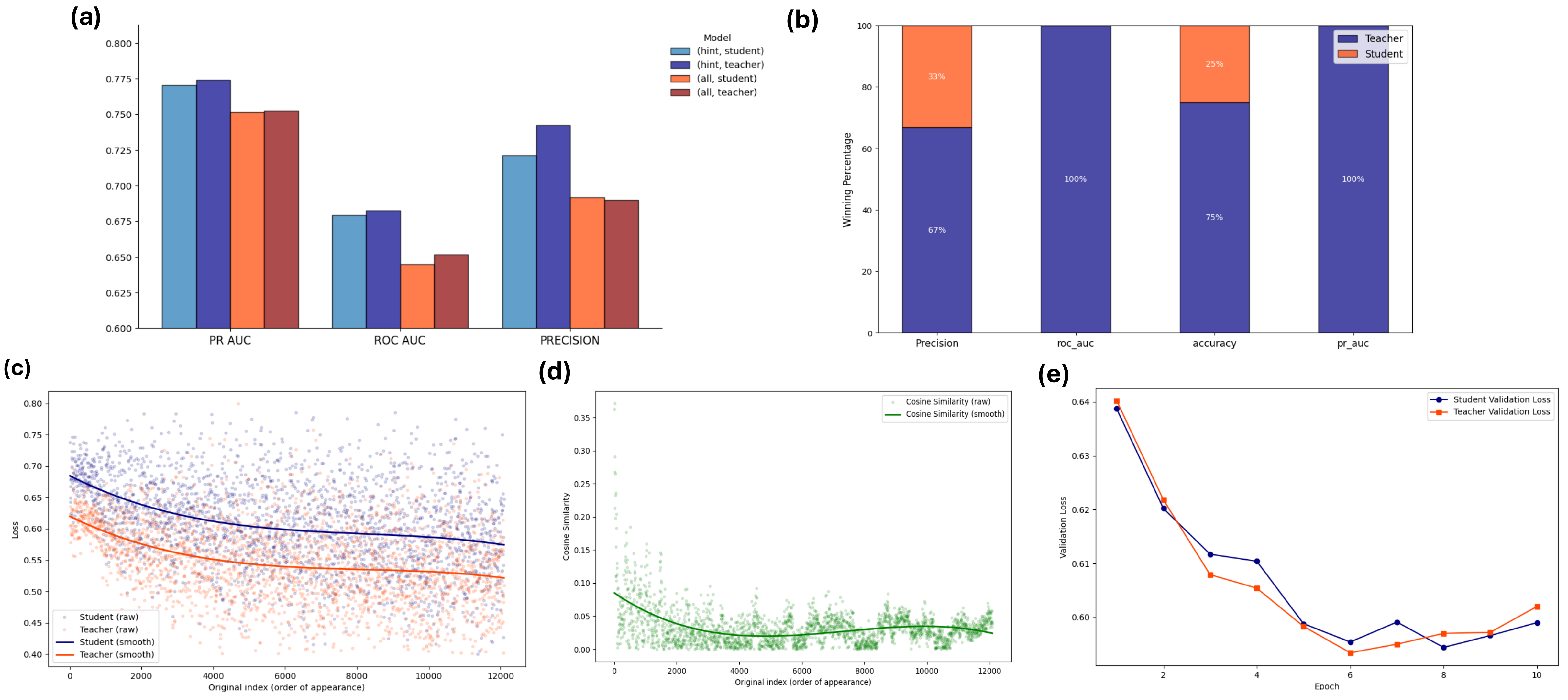


Figure 4: dULE-HINT model results and training behavior. **(a)** Respective performances of best model (cos_hint_0.9_0.0) on PR_AUC, ROC_AUC, and Precision. Performance tends to be similar across models. **(b)** Testing student/teacher pairs for superior performance within the top 50 models across various statistics. **(c)** Teacher, student training loss across epochs. **(d)** Teacher/student cosine similarities across epochs. **(e)** Teacher, Student validation loss across epochs.

Robustness of Hyperparameter Tuning

DISCUSSION

- In using out method of self-supervised labeling of clinical trials, we increased the dataset heterogeneity and size—despite roughly doubling the inaccuracy. Greater prompt engineering or use of more advanced LLMs could potentially avoid this cost.
- While the teacher model outperformed the baseline, the modest improvement suggests that this is too complex of a task for a one-layer graph neural network. However, it remains a promising finding that we achieved better performance statistics while training nearly orthogonal to the student's gradient.

- Modify the loss function by taking a product of MSE and cosine-similarity, combining both a directionality and magnitude penalty to help our teacher model better explore the loss landscape.
- Compare with classical ML models such as Naïve Bayes, Random Forest, or XGBoost to evaluate performance in the field.

References

1. Tianfan Fu, Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Hint: Hierarchical interaction net-work for clinical-trial-outcome predictions. Patterns, 3(4):100445, Feb 2022.
2. Jeff Mitchell, Jesus Martınez del Rincon, and Niall McLaughlin. Unlearning from experience to avoid spurious correlations, 2024.