What if We've Been Looking at the Wrong Data? Reimagining Clinical Trial Success Prediction using AI

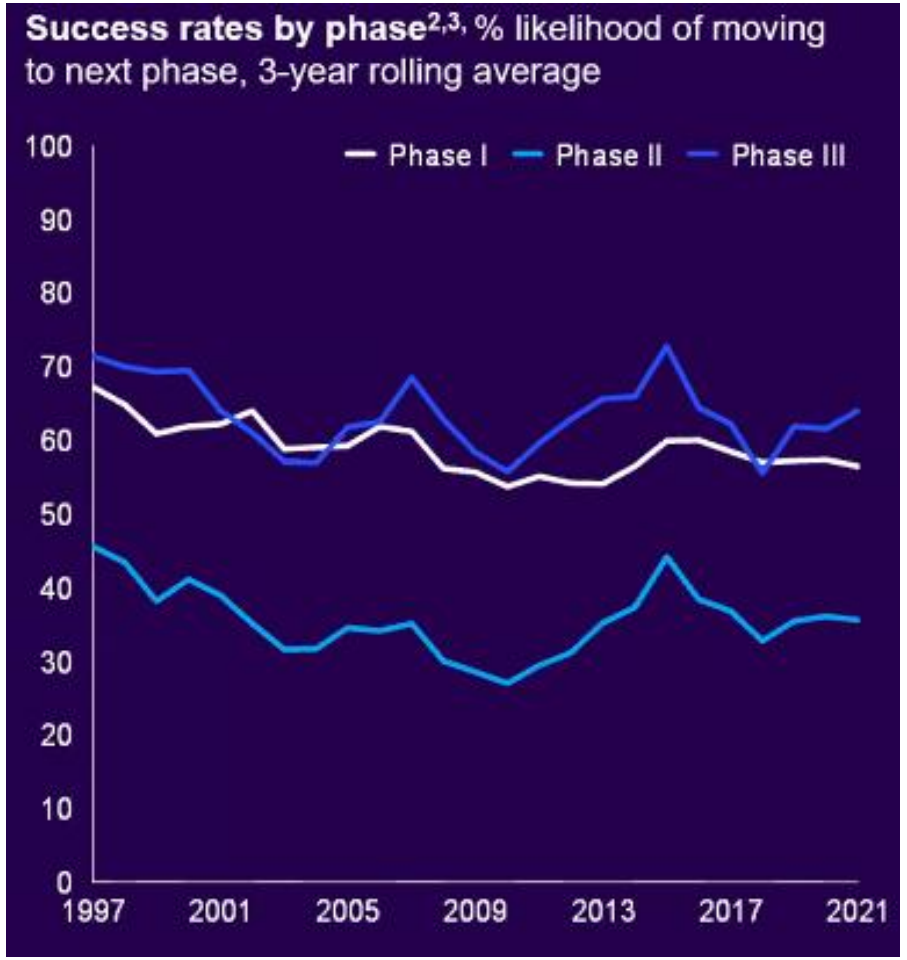
Léo Fournier, Juan Martinez, Tarun Naithani, Krishna Bellamkonda, Wenting Wang, Nils Ternès & Christelle Reynès

June 16th, 2026
PSI

01 Brief introduction

Clinical trials are likely to fail!

- Overall success probability of clinical programs is relatively low (~ 10%)
  - Unnecessary cost and development time
  - Inadequate patient exposure
- Predicting probability of clinical trial success (PoS) could support decision making
  - Go/No Go decision, indication prioritization, compound positioning
- Phase II studies around 35% success
- Increasing research in this area following advancement in AI capabilities

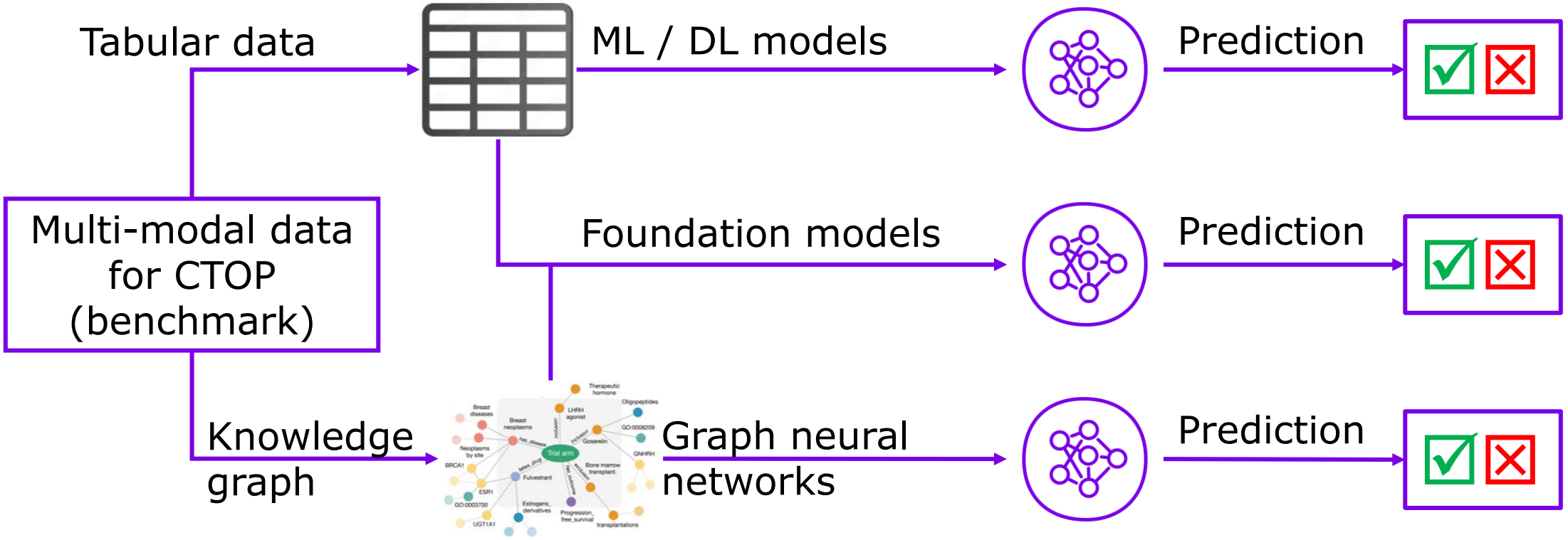


PharmaProjects, July 2022

02 Existing methods and benchmark data for clinical trial outcome prediction (CTOP)

Methods | AI-based approaches for Clinical Trial Outcome Prediction (CTOP)

Growing efforts and researches, especially with the emergence of machine and deep learning (ML/DL) and the large amount of data available



Data | Existing multi-modal benchmarks and some models for CTOP

DATA

TOP (2022)
- ✓ Manual curation: ~12,000 trials

CTO (2025)
- ✓ Automatic curation: ~100,000 trials

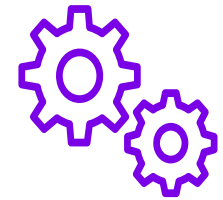
TrialPanorama (2025)
- ✓ Automatic curation: ~53,000 trials

MODELS

HINT (2022)
- ✓ Graph Convolutional Network

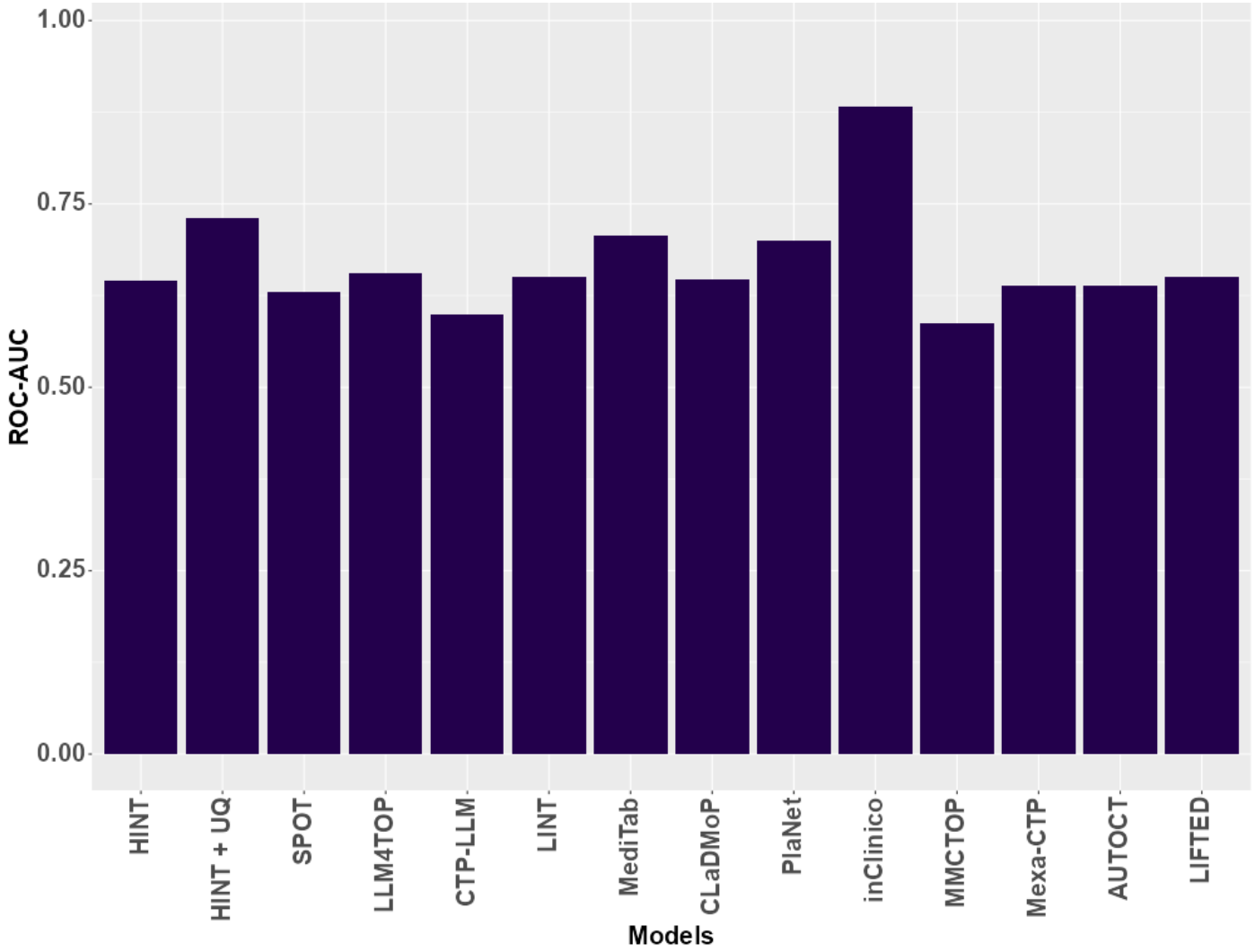
SPOT (2023)
- ✓ Recurrent Neural Network with k-means

MediTab (2025)
- ✓ Foundation model



Models | Augmenting complexity, performance plateau

ROC-AUC of different models for CTOP ordered by increasing complexity



Our current hypothesis:

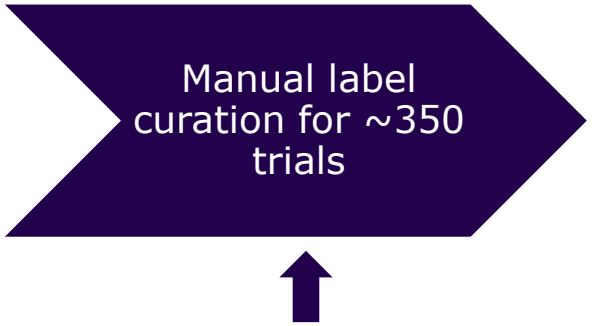
- Issue with the benchmarks?
- Lack of relevant information for appropriate prediction?
- Sub-optimal model architecture?
- Anyway, the prediction task is too complex?

03 Current effort and next steps

Hypothesis 1:
Problem with outcome labels?

Data | Our automation process for outcome labeling

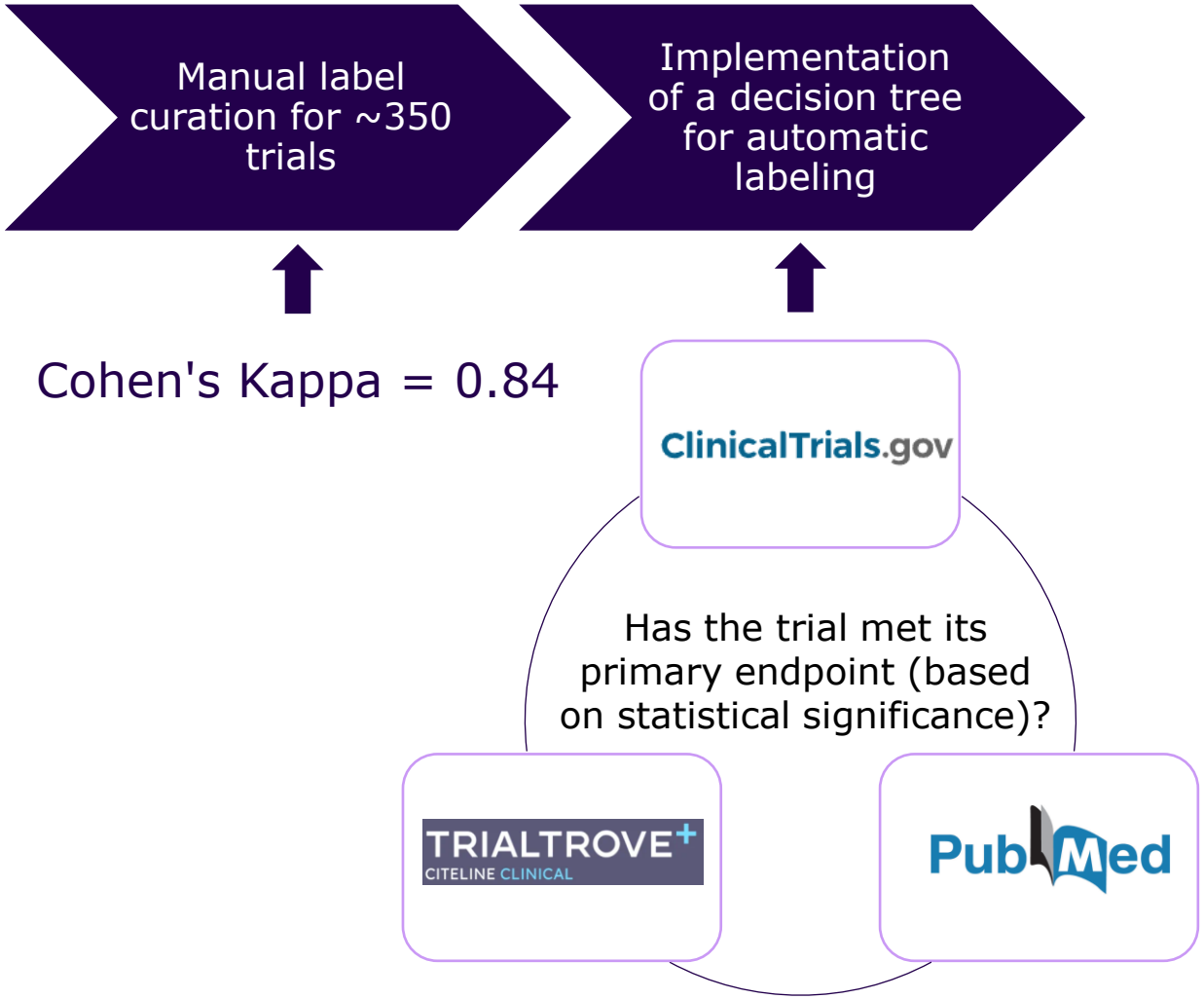
Our overall process (focused mostly on phase II trials):



Cohen's Kappa = 0.84

Data | Our automation process for outcome labeling

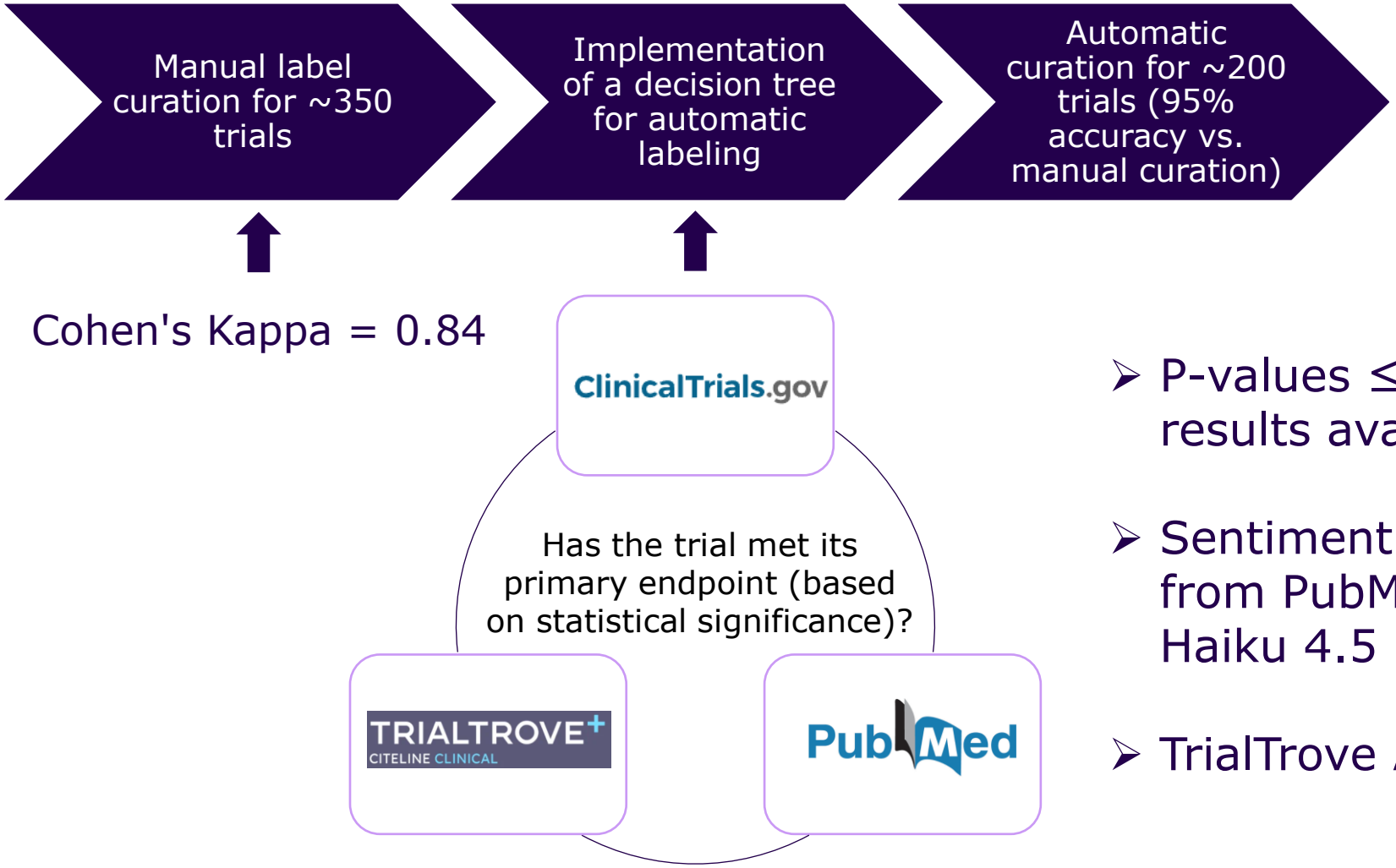
Our overall process (focused mostly on phase II trials):



- P-values ≤ 0.05 for CT.gov when results available
- Sentiment analysis on abstracts from PubMed using Anthropic Haiku 4.5 LLM
- TrialTrove API

Data | Our automation process for outcome labeling

Our overall process (focused mostly on phase II trials):

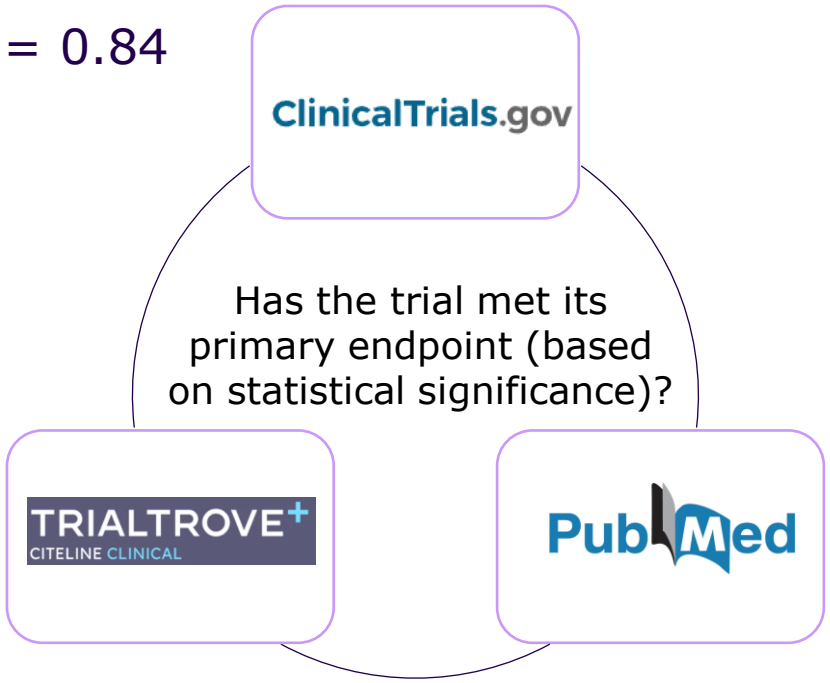


- P-values ≤ 0.05 for CT.gov when results available
- Sentiment analysis on abstracts from PubMed using Anthropic Haiku 4.5 LLM
- TrialTrove API

Data | Our automation process for outcome labeling

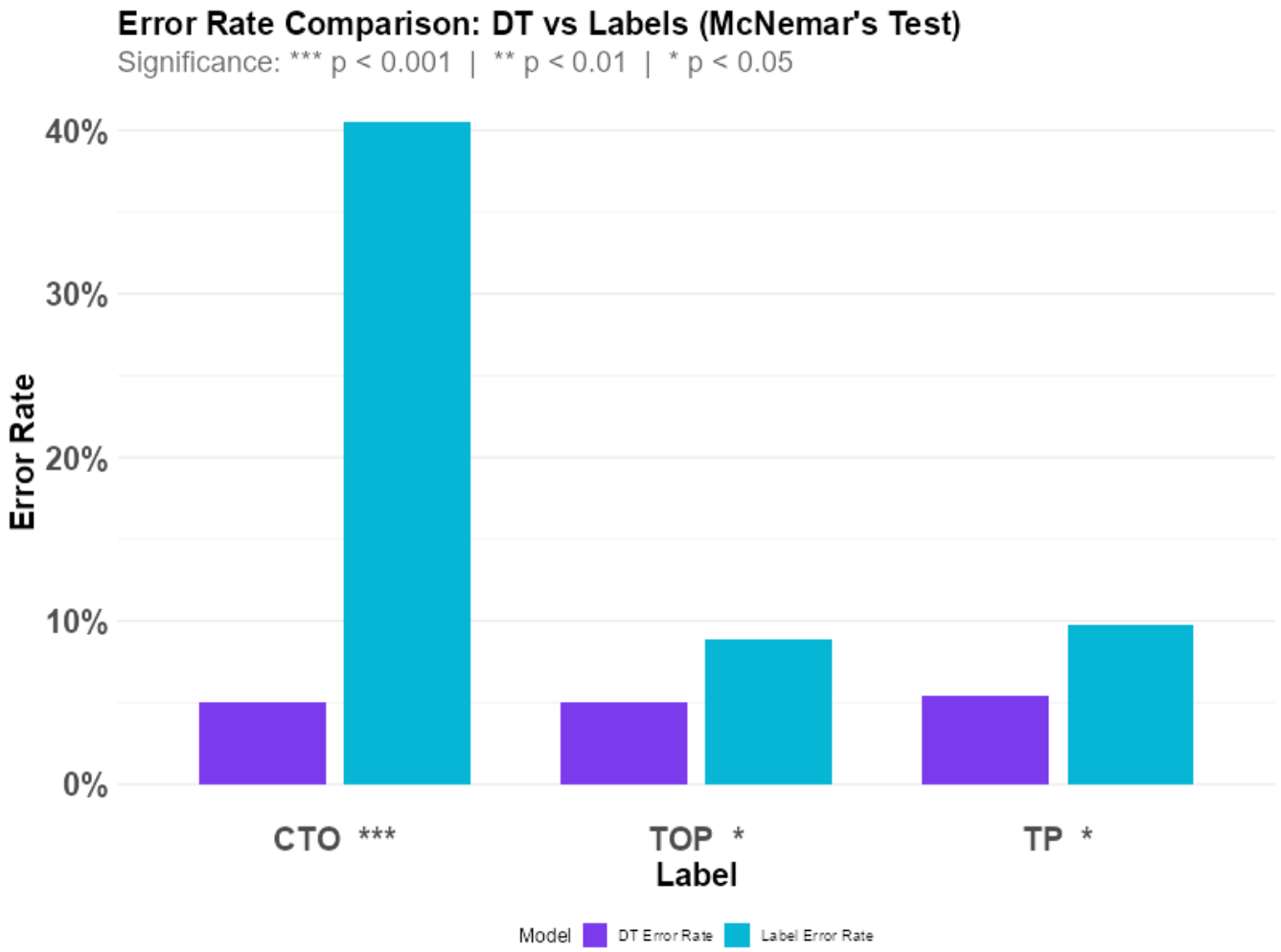
Our overall process (focused mostly on phase II trials):

Cohen's Kappa = 0.84



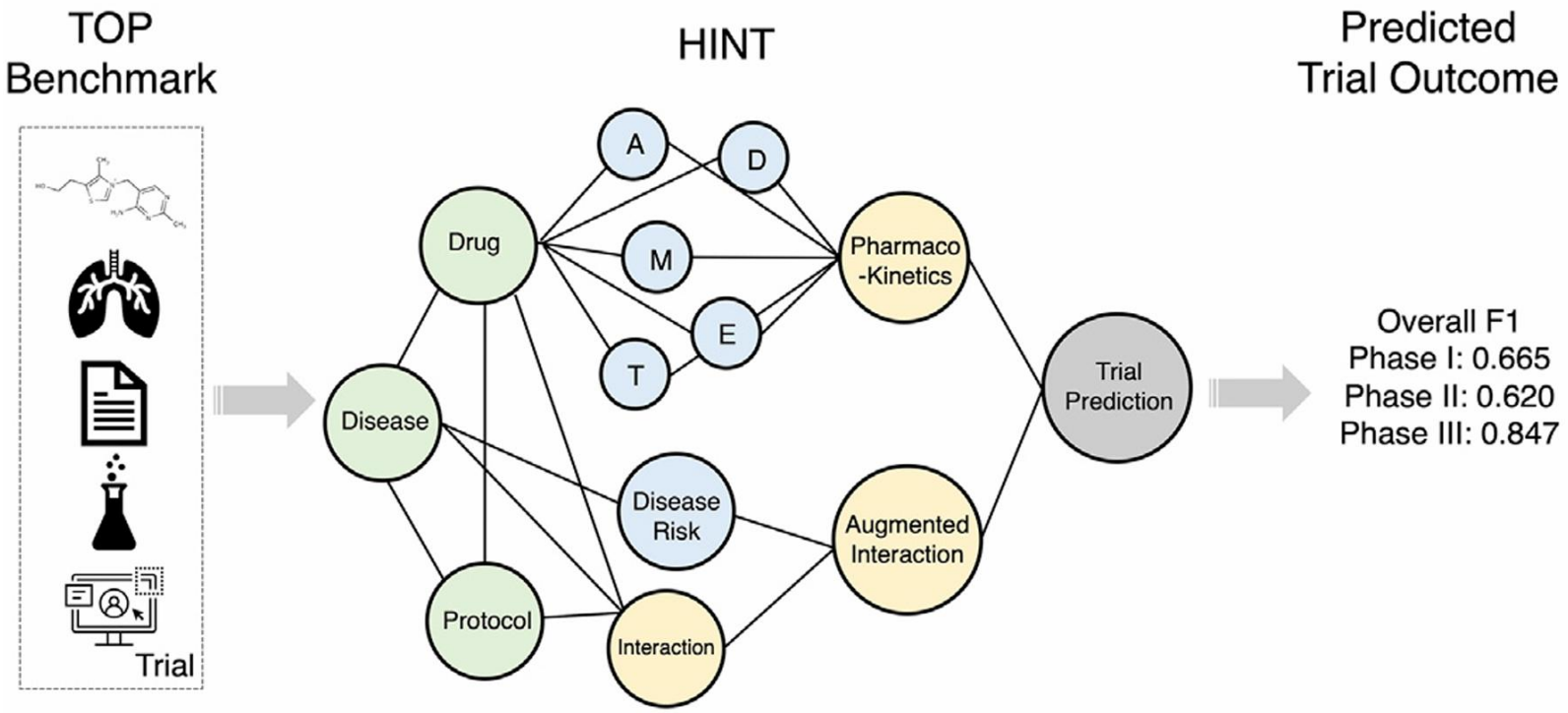
- P-values ≤ 0.05 for CT.gov when results available
- Sentiment analysis on abstracts from PubMed using Anthropic Haiku 4.5 LLM
- TrialTrove API

Data | Performance of our decision tree vs other benchmarks



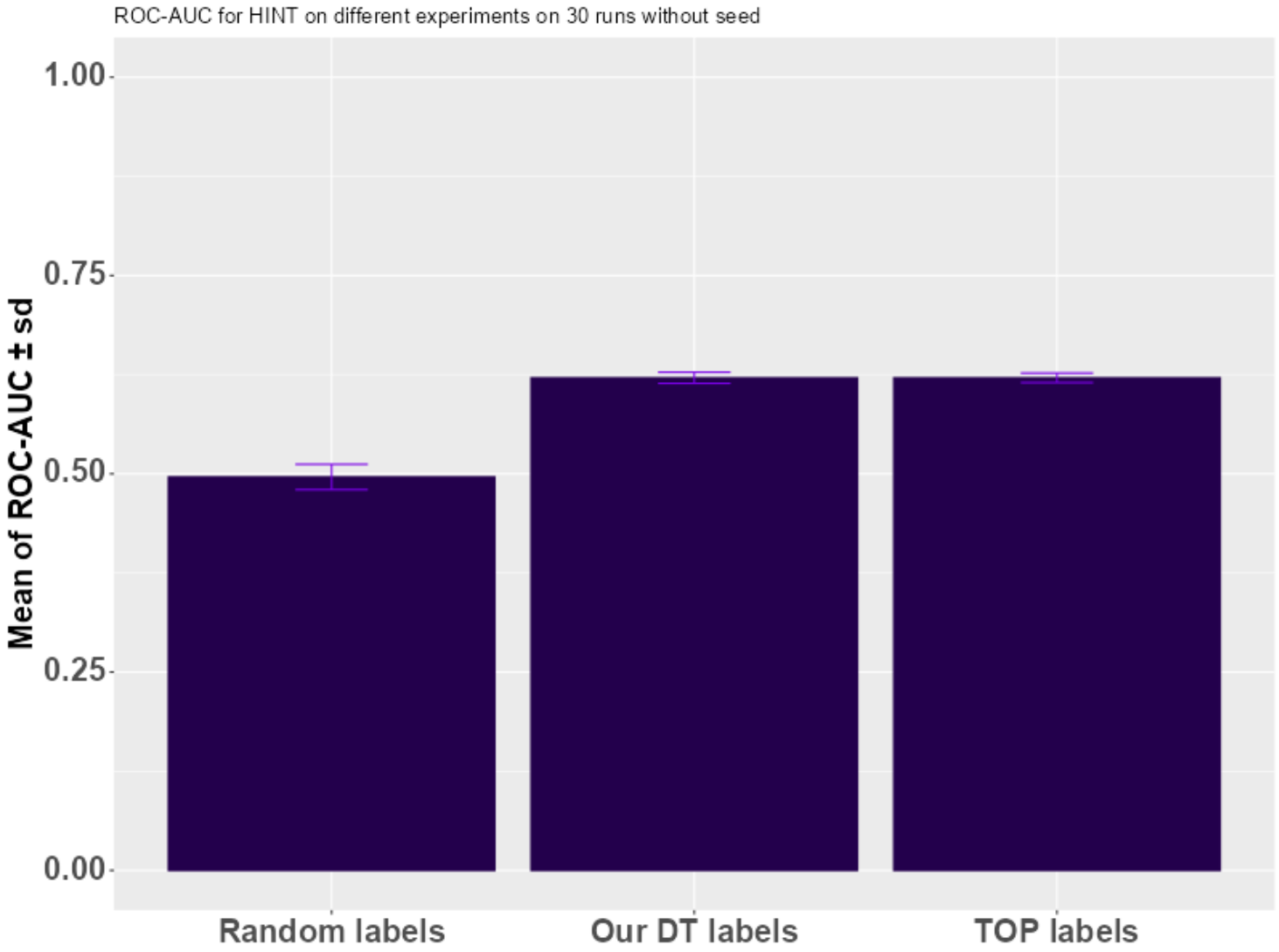
- Our decision tree (DT) reaches significantly lower classification error rate vs other benchmarks
- Better performance, massive time savings
- Applied to TOP dataset (~6000 studies) : 7 w/o predicted labels, 305 different labels

Models | Deep dive into HINT



- Graph Convolutional Network
- Inputs: SMILES, Disease name, I/E criteria, ADMET SMILES-based, historical risk rate
- Output: PoS

Models| Relaunching HINT



- Our labels reach same performances and is better than hazard model
- Labels automatically given instead of TOP labels manually curated

Hypotheses 2 & 3:
Lack of relevant information?
Sub-optimal model architecture?

Data | Quality assessment of existing benchmarks

Quality of the information used for predicting clinical trial outcome is not enough, e.g.:

- ❑ SMILES & PK data not accurate and/or relevant

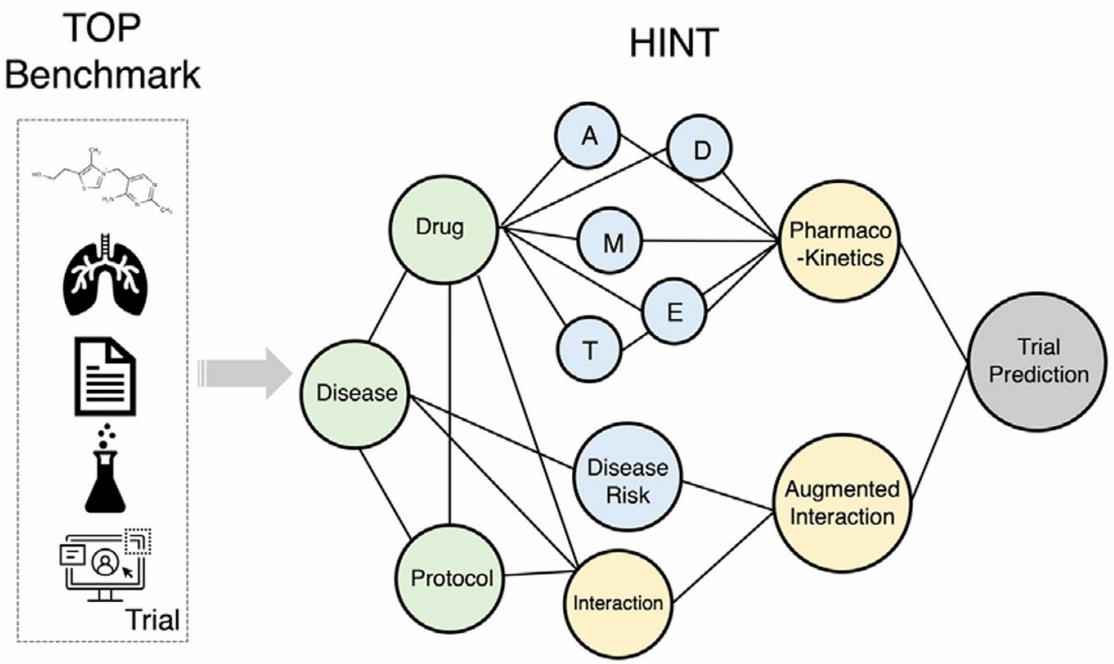
nctid str	drugs str	smiless str
NCT00405938	['bevacizumab', '…	['[H][N]1([H])[C@@H]2CCCC[C@H]2[N]([H])([H])[Pt]11OC(=O)C(=O)O1',…
NCT00583622	['bevacizumab', '…	['[H][N]1([H])[C@@H]2CCCC[C@H]2[N]([H])([H])[Pt]11OC(=O)C(=O)O1',…
NCT00661778	['bevacizumab', '…	['[H][N]1([H])[C@@H]2CCCC[C@H]2[N]([H])([H])[Pt]11OC(=O)C(=O)O1',…

For Bevacizumab : SMILES corresponds to Oxaliplatin

```
CN(C)C(=O)C=1/N=C(/CNC(=O)CN)N(N=1)C2=C/C=C(/[C1])C=C2C(=O)C3=CC=C/C=C3/[C1]    1
NC(=O)CN2CC1=CC=CC=C1OC2=O  1
CCC(=O)[C@@]2(CCN(CCCC(=O)C1=C/C=C(/[F])C=C1)CC2)N3CCCCC3   1
OCC[C@@]4([H])CCN(CC/C=C2/C1=C/C=C(/[F])C=C1[S]C3=CC=C(C=C23)C([F])([F])[F])CC4 1
```

Example of distribution dataset for ADMET model:
0/1 label for ability of passing through BBB: cannot explain the entire process of distribution.

Models | Deep dive into HINT



<u>Multiple areas of improvement in existing model architecture, e.g.,:</u>

- Graph Convolutional Network with dropout rate of 0.6, adjacency matrix with huge diagonal values, only 5 epochs with minimal learning rate (3e-4), weight decay implemented but not used.
- Bootstrap (20 resamplings) only on test to add variability to performance metrics on test set.
- All types of trials treated the same (oncology, respiratory, …)

Models | Hypothesis and next steps

	SOLUTION
Lack of relevant information	Create dataset of BM, PK, PC, etc.
Sub-optimal model architecture	Hyperparameter optimization
Lack of Explainability	Change the model

Next steps: Create a new model based on explainable ML algorithms (random forests...) combined with feature selection algorithms

Discussion

- CTOP: hot topic ⇨ potential to support strategic clinical development

  Fit-for-purpose data is the foundation of any AI-related work

- Despite increasing research, performances plateau :
  - Labels hard to get and often biased
  - Lack of relevant information
  - Sub-optimal architecture

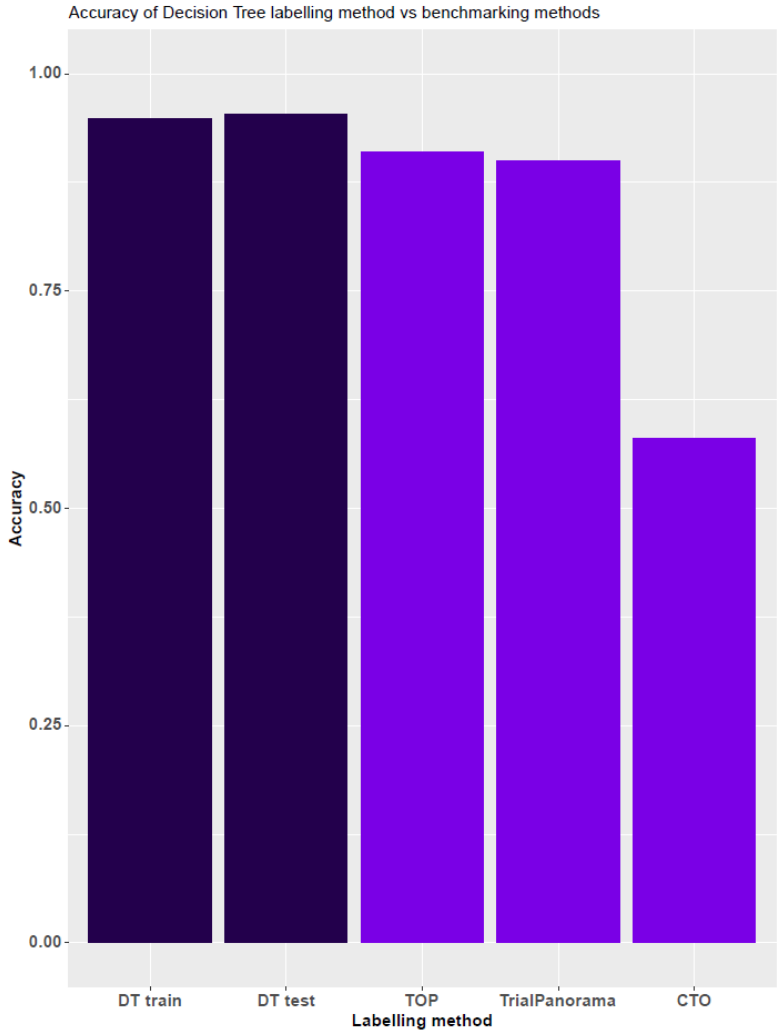
- We are building our fit-for-purpose foundations:
  - Outcome labels
  - Data for predicting outcome success
  - Trials for model building
  - Model architecture

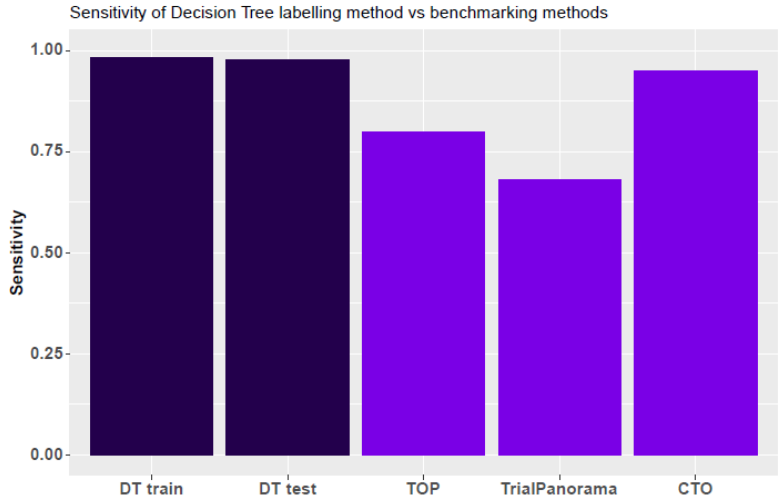
- Once the foundation are set, we will develop a fit-for-purpose methodology, considering explainability as a central piece

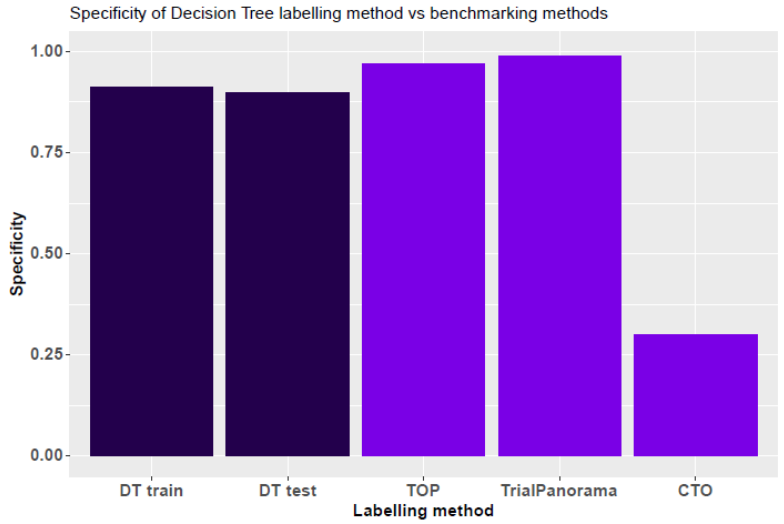
Thank you !

04 Back Up

Data | Performance of decision tree vs other benchmarks

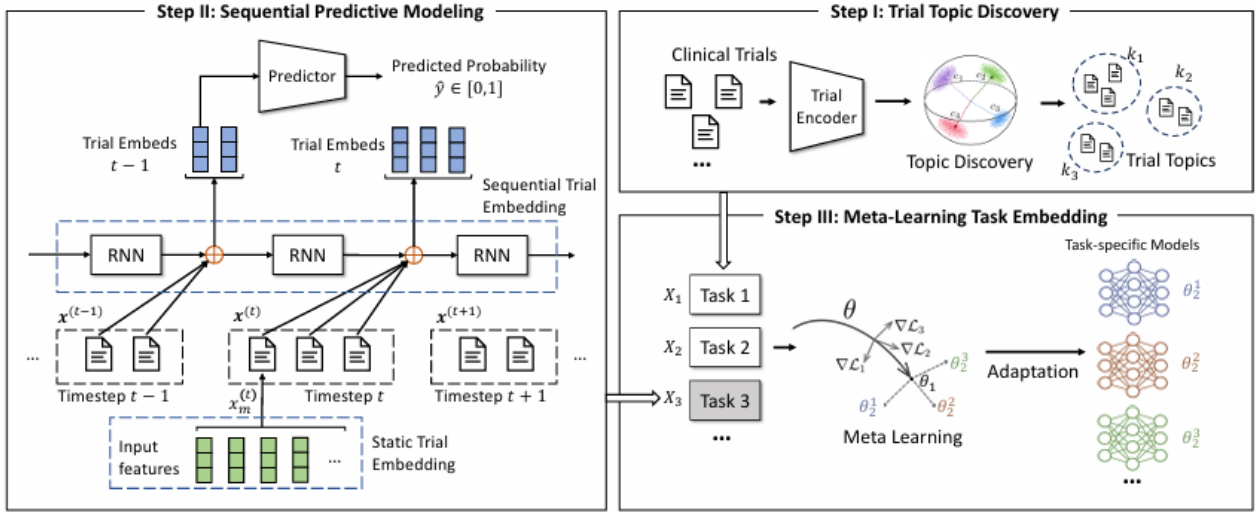






- Decision tree reaches better performance in terms of Accuracy and Sensitivity
- Bit less on Specificity
- CTO is floundering !

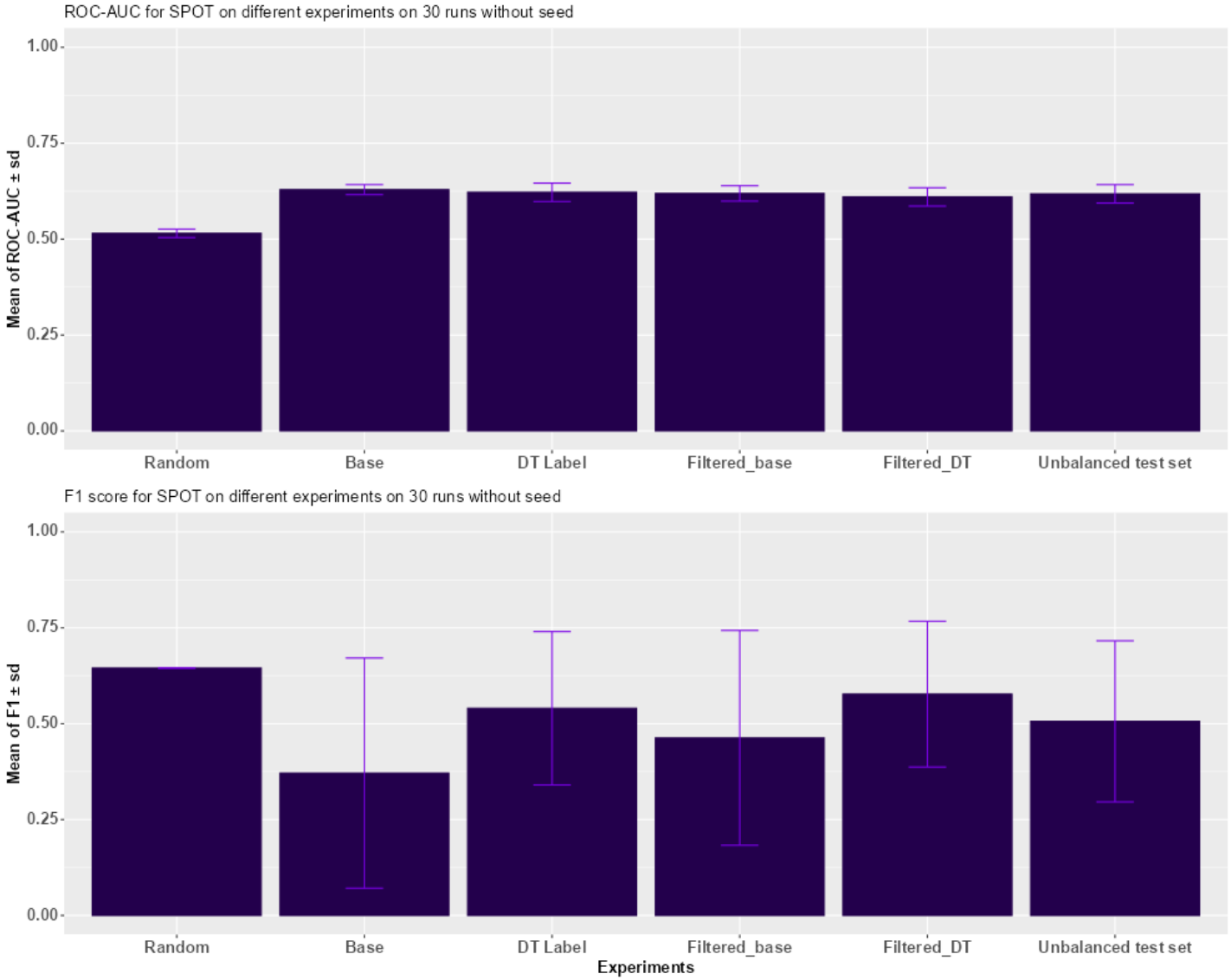
Models | Deep dive into SPOT



Some odd features in SPOT

- Same as HINT for data
- Kmeans algorithm used for clustering of trials but no way to optimize k (number of clusters) -> k=50
- 50 sub models with 10 epochs each.

Models| Relaunching SPOT

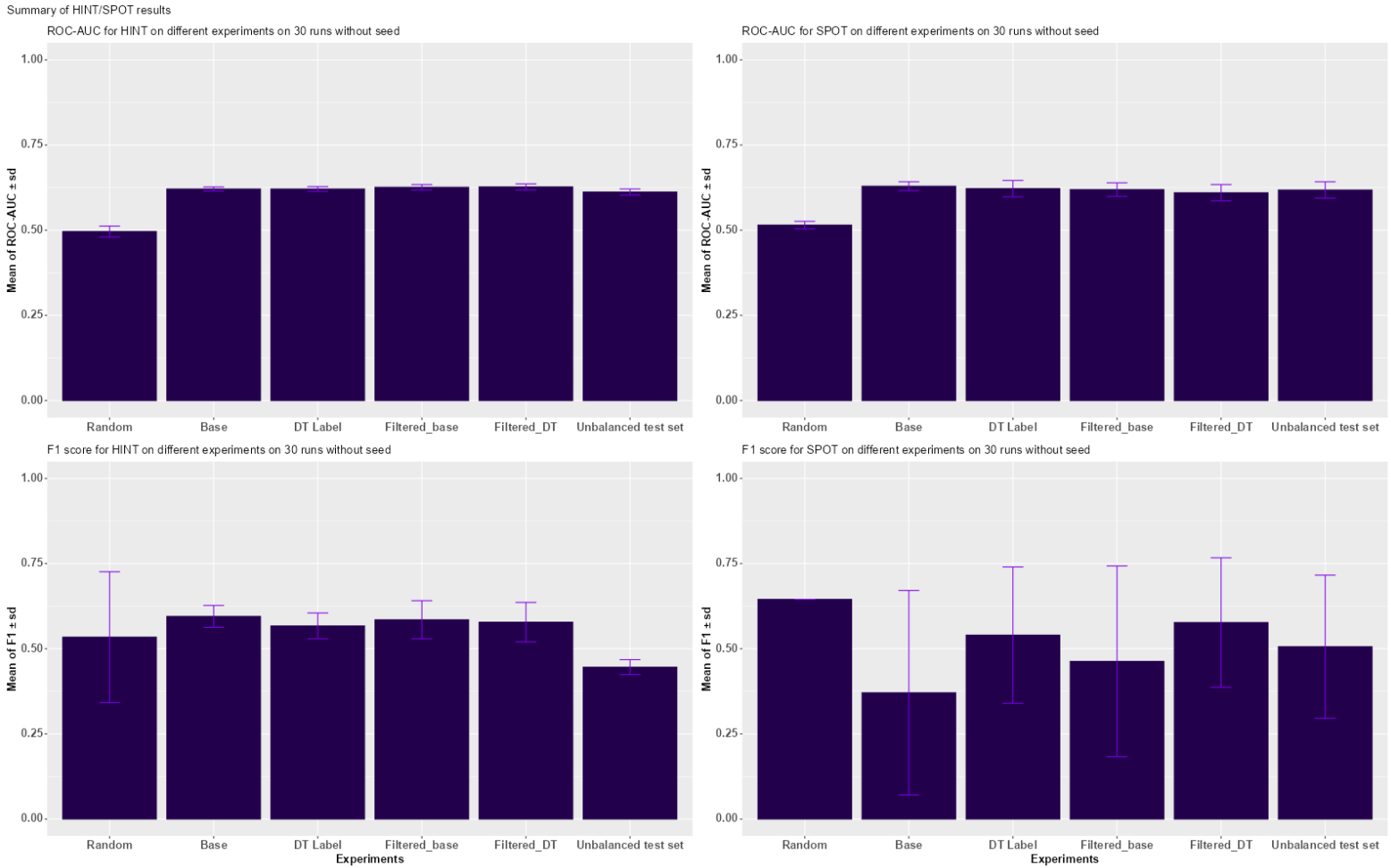


- Our labels reach same or better performances
- Filtering trials doesn't seem to add value
- F1 with very high variability

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

F1 doesn’t consider the true negatives

Models| Summary of HINT / SPOT



- Our labels reach same or better performances
- Labels are automatically given instead of TOP labels that were manually curated
- Filtering trials doesn't seem to add value
- Having a test set with 70% Failure / 30% Success, as the reality, worsen the predictions