ML module #4
(Recap & Metrics)

Benjamin Sanchez-Lengeling

CrossTALK: Cross-Training in AI and Laboratory Knowledge for Drug Discovery.

VECTOR INSTITUTE

UNIVERSITY OF TORONTO

Panorama

Recap

Metrics

Working sesh

Panorama

Recap

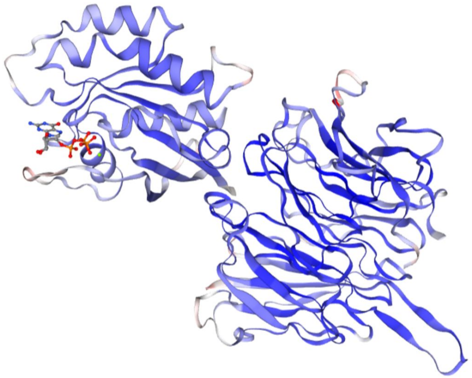
Metrics

Working sesh

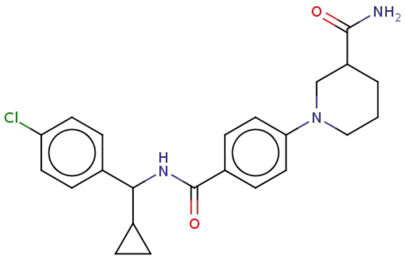
Recap: From Problem to Experiment

Simplified:

Some diseases are caused by malfunctioning Proteins, To “fix them” we need to physically interact with it using a “drug molecule”.



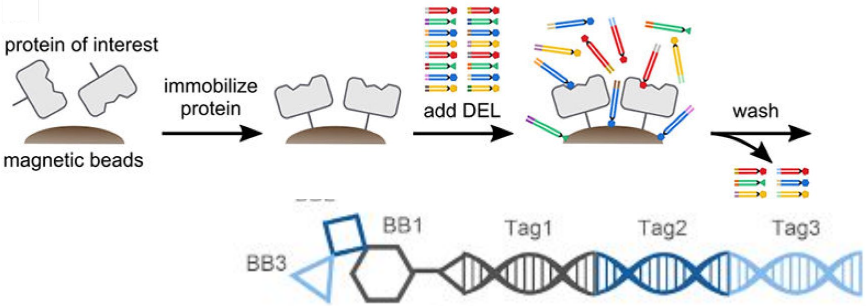
A4D1P6 (WDR91_HUMAN)



18 molecules with binding data

(https://www.bindingdb.org/uniprot/A4D1P6)

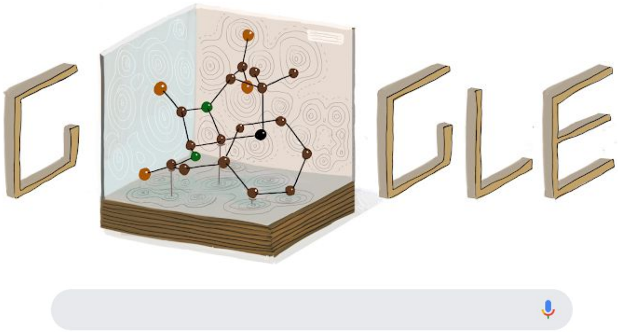
DEL Experiments
Allow us to test MASSIVE number of molecules

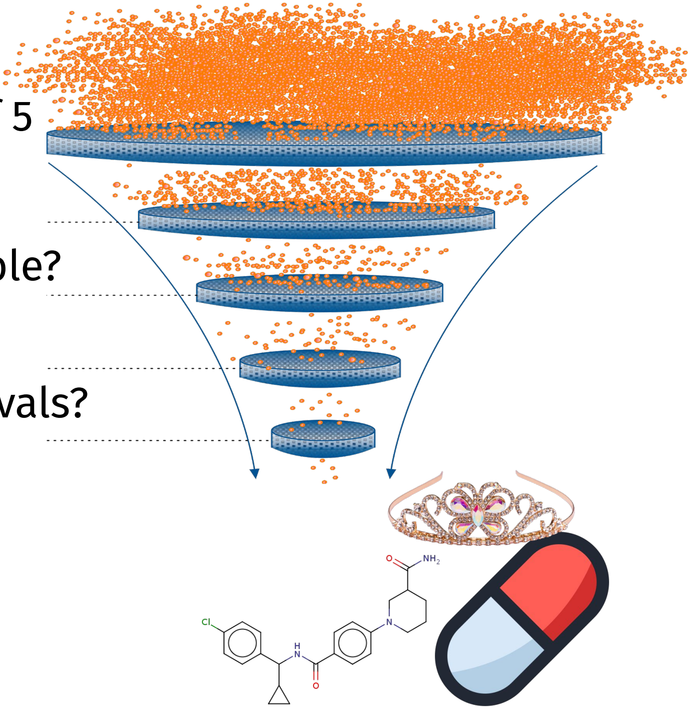


Recap: Computational solutions

Enamine

Desired solution:
“Molecular search engine”





Filters:
Lipinski’s rule of 5

Does it bind?

Orally bioavailable?

Toxic?

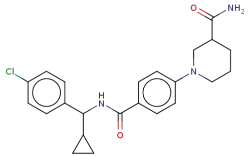
Animal/Tissue evals?

Human evals?

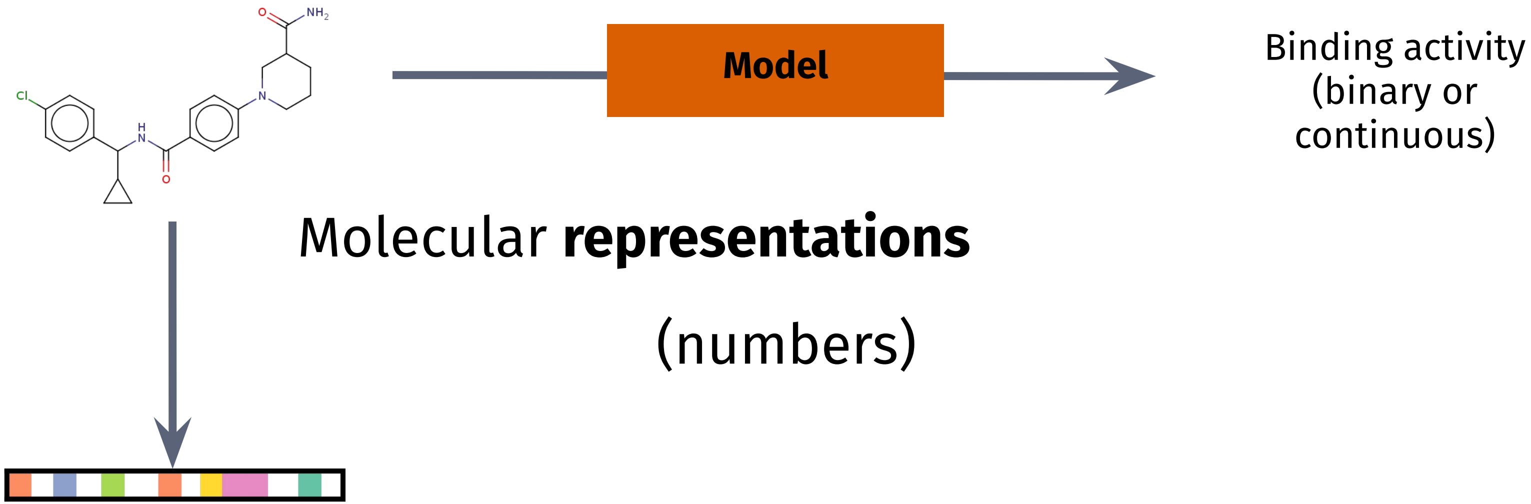
More hits, more shots on goal!

Recap: Caveats, Hits are not drugs

- Low potency
- Low specificity
- Insoluble in water
- Unstable
- Unable to get into cells
- False positives



Recap: Starting from the “simplest” problem (but still hard!)



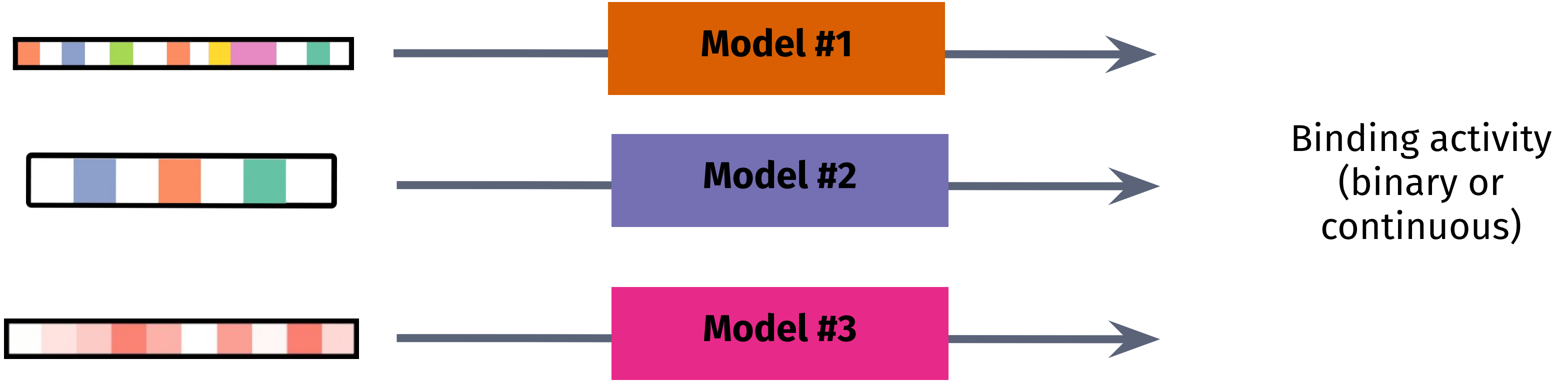
Model

Binding activity (binary or continuous)

Molecular **representations**

(numbers)

Recap: Which model is the best? (and why?)



Metrics!

Panorama

Section slides prepared by Cait Harrigan

Recap

Metrics

Working sesh

AUROC - area under receiver operating characteristic

Asks: what is the probability that a random true positive will be ranked higher than a random true negative? *Measures ranking at all thresholds*

	Predict hit	Predict no hit
ASMS hit	TP	FN
ASMS no hit	FP	TN

$$\frac{TP}{TP + FN}$$ True positive rate aka recall

$$\frac{FP}{FP + TN}$$ False positive rate

AUROC - area under receiver operating characteristic

Asks: what is the probability that a random true positive will be ranked higher than a random true negative? *Measures ranking at all thresholds*

	Predict hit	Predict no hit
ASMS hit	TP	FN
ASMS no hit	FP	TN

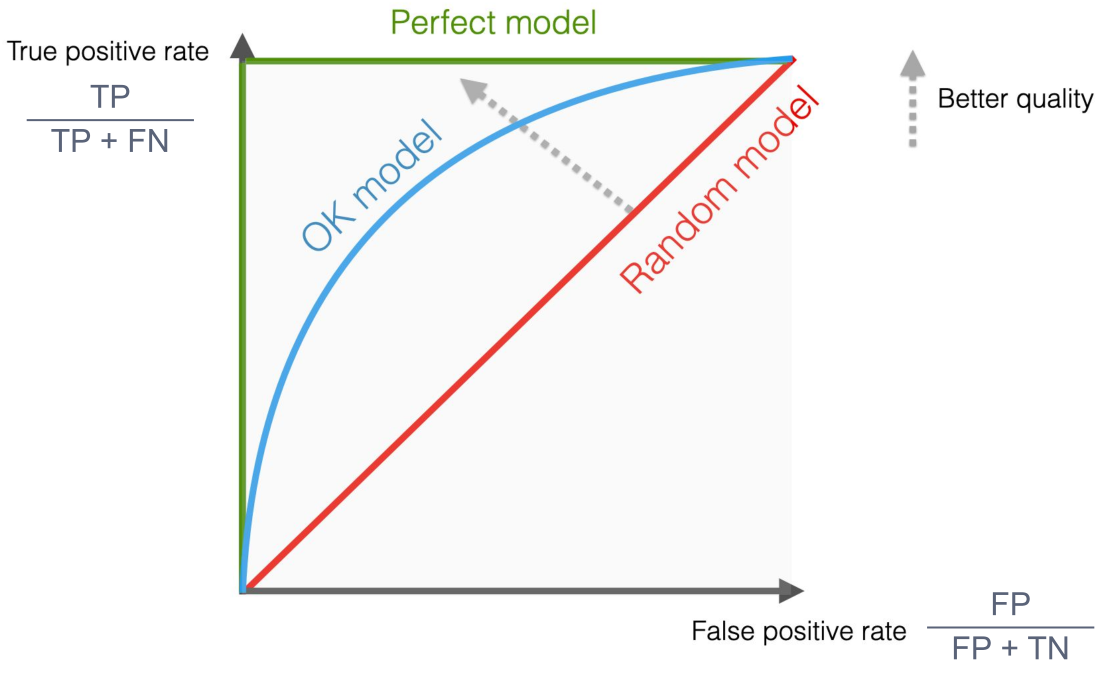
Perfect model	Predict hit	Predict no hit
ASMS hit	**100%**	**0%**
ASMS no hit	**0%**	**100%**

Random model	Predict hit	Predict no hit
ASMS hit	**50%**	**50%**
ASMS no hit	**50%**	**50%**

Row percentages

AUROC - area under receiver operating characteristic

	Predict hit	Predict no hit
ASMS hit	TP	FN
ASMS no hit	FP	TN



https://www.evidentlyai.com/classification-metrics/explain-roc-curve

AUPRC - area under precision recall curve

Asks: how hit-rich are my top ranked predictions? *Measures expected precision at all thresholds*

	Predict hit	Predict no hit
ASMS hit	TP	FN
ASMS no hit	FP	TN

$$\frac{TP}{TP + FN}$$

True positive rate aka recall

$$\frac{TP}{TP + FP}$$

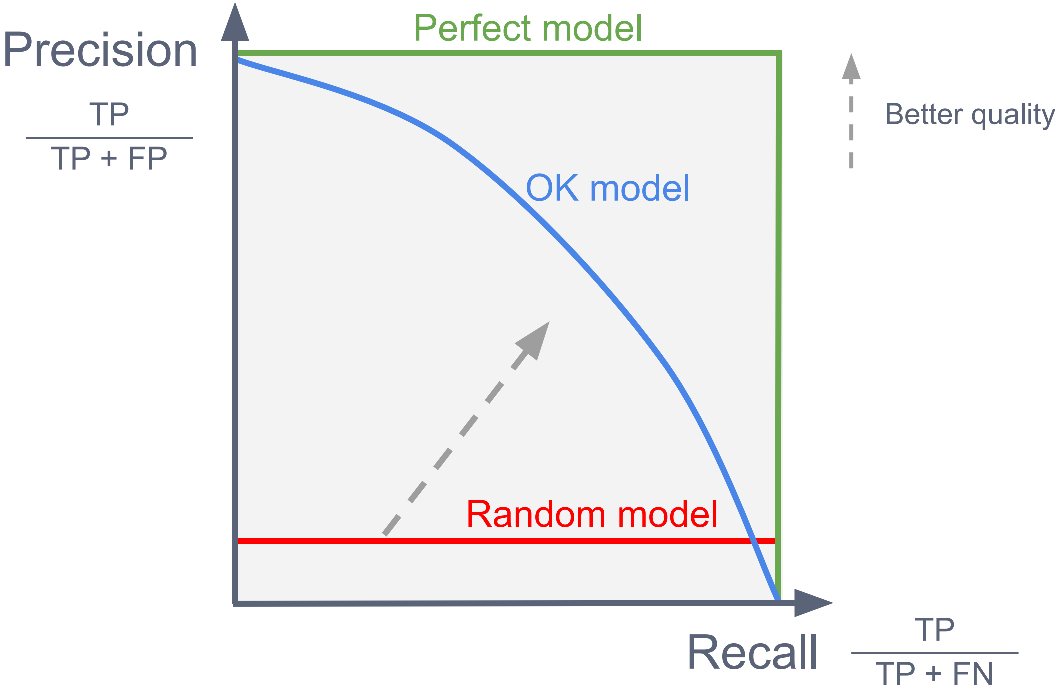
Precision

Interested in a row % and a column %

AUPRC - area under precision recall curve

Asks: how hit-rich are my top ranked predictions? *Measures expected precision at all thresholds*

	Predict hit	Predict no hit
ASMS hit	TP	FN
ASMS no hit	FP	TN



We care most about the top ranked molecules *not necessarily performance at all thresholds*

Predictions from model

Molecule	Predicted probability
E	0.65
B	0.40
F	0.20
A	0.12
C	0.03
D	0.01

Test labels from ASMS

Molecule	ASMS Hit (ground truth)
A	1
B	1
C	0
D	1
E	1
F	0

Hits at 3 *How many TP are in top 3?*

Molecule	Predicted probability
E	0.65
B	0.40
F	0.20
A	0.12
C	0.03
D	0.01

Molecule	ASMS Hit (ground truth)
A	1
B	1
C	0
D	1
E	1
F	0

Hits at 3 = 2

Molecule	Predicted probability	
E	0.65	✅
B	0.40	✅
F	0.20	
A	0.12	
C	0.03	
D	0.01	

Molecule	ASMS Hit (ground truth)
A	1
B	1
C	0
D	1
E	1
F	0

Precision at 3 *what % of top 3 are TP?*

Molecule	Predicted probability
E	0.65
B	0.40
F	0.20
A	0.12
C	0.03
D	0.01

Molecule	ASMS Hit (ground truth)
A	1
B	1
C	0
D	1
E	1
F	0

Precision at 3 = 2/3 = 0.66

Molecule	Predicted probability	
E	0.65	✅
B	0.40	✅
F	0.20	❌
A	0.12	
C	0.03	
D	0.01	

Molecule	ASMS Hit (ground truth)
A	1
B	1
C	0
D	1
E	1
F	0

Recall at 3 *what % of TP are in the top 3?*

Molecule	Predicted probability
E	0.65
B	0.40
F	0.20
A	0.12
C	0.03
D	0.01

Molecule	ASMS Hit (ground truth)
A	1
B	1
C	0
D	1
E	1
F	0

Recall at 3 = 2/4 = 0.5

Molecule	Predicted probability	
E	0.65	✅
B	0.40	✅
F	0.20	
A	0.12	❌
C	0.03	
D	0.01	❌

Molecule	ASMS Hit (ground truth)
A	1
B	1
C	0
D	1
E	1
F	0

Why not threshold? It′s too stringent

Want to get credit for ranking B highly!

Molecule	Predicted probability	Pred proba >0.5	
E	0.65	1	✅
B	0.40	0	❌
F	0.20	0	
A	0.12	0	❌
C	0.03	0	
D	0.01	0	❌

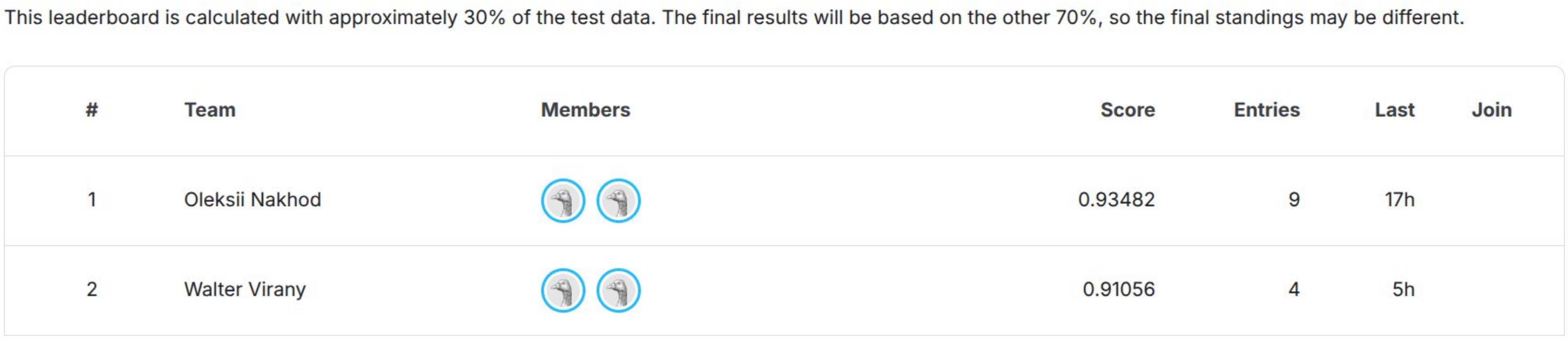
Molecule	ASMS Hit (ground truth)
A	1
B	1
C	0
D	1
E	1
F	0

Summary

- **AUROC** measures ranking ability at all thresholds
- **AUPRC** measures expected precision at all thresholds
- **Hits @ K** measures number of True Positives in top K
- **Precision @ K** measures percentage of top K which are True Positives
- **Recall @ K** measures percentage of True Positives which are in the top K

Drumrolls

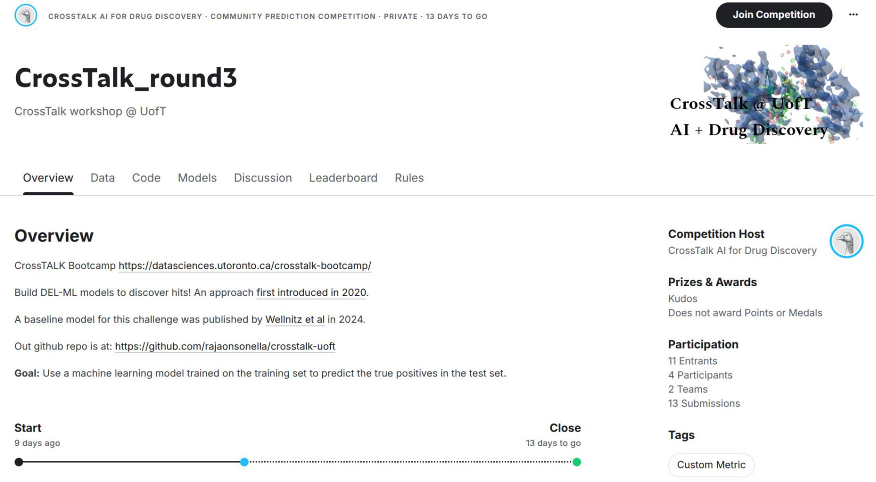
Leaderboard! (as of thursday night)



This leaderboard is calculated with approximately 30% of the test data. The final results will be based on the other 70%, so the final standings may be different.

#	Team	Members	Score	Entries	Last	Join
1	Oleksii Nakhod		0.93482	9	17h	
2	Walter Virany		0.91056	4	5h	

Kaggle: Final Eval Metric has landed



CROSSTALK AI FOR DRUG DISCOVERY · COMMUNITY PREDICTION COMPETITION · PRIVATE · 13 DAYS TO GO

CrossTalk_round3

CrossTalk workshop @ UofT

Overview

CrossTALK Bootcamp https://datasciences.utoronto.ca/crosstalk-bootcamp/

Build DEL-ML models to discover hits! An approach first introduced in 2020.

A baseline model for this challenge was published by Wellnitz et al in 2024.

Out github repo is at: https://github.com/rajaonsonella/crosstalk-uoft

Goal: Use a machine learning model trained on the training set to predict the true positives in the test set.

Start
9 days ago

Close
13 days to go

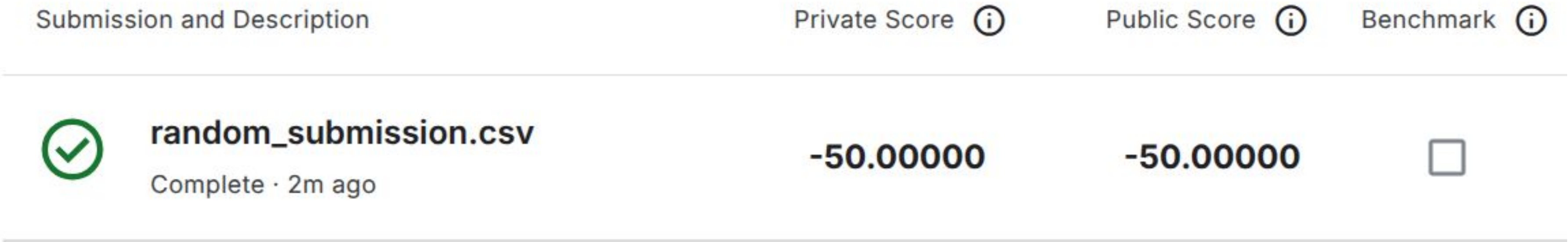
Competition Host
CrossTalk AI for Drug Discovery

Prizes & Awards
Kudos
Does not award Points or Medals

Participation
11 Entrants
4 Participants
2 Teams
13 Submissions

Tags
Custom Metric

“Penalized”
Hits @ 200



Submission and Description	Private Score	Public Score	Benchmark
✓ random_submission.csv Complete · 2m ago	-50.00000	-50.00000	□

Panorama

Recap

Metrics

Working sesh

Some “ML tricks”

- XGboost Tricks / Feature engineering
- Ensembles are always better, many models and average predictions
- Uncertainties help to re-rank stuff
- Hyperparameter tuning
- “Balanced” /”Adversarial” splits
- Compute, get a 24GB RAM computer
- Internet / Reddit
  - r/MachineLearning
  - Kaggle forums (example)

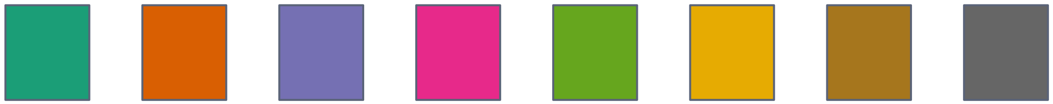
Next sessions: TabPFN + Ranking

- TabPFN is a tabular foundation model
- Ranking is a “ordered” classification

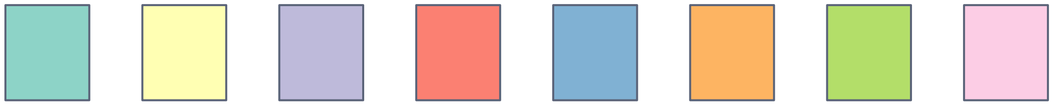
For color palette

https://colorbrewer2.org/#type=sequential&scheme=BuGn&n=3

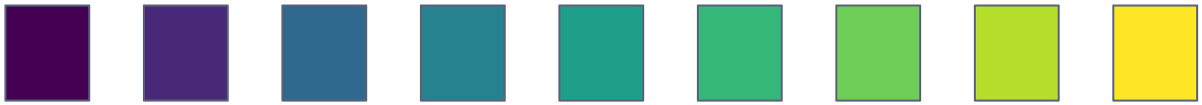
Dark2



Set3



Cividis (continuous)



PiYG (divergent)