

ML module #5

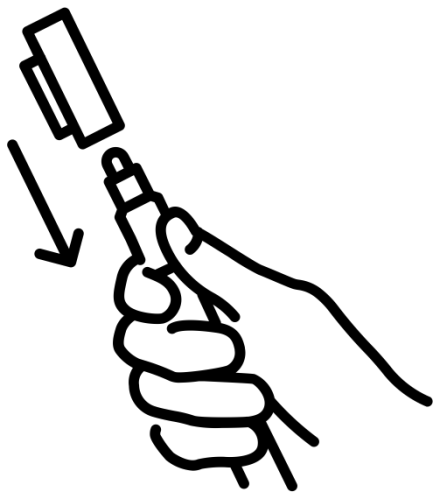
(Recap & Metrics)

Benjamin Sanchez-Lengeling

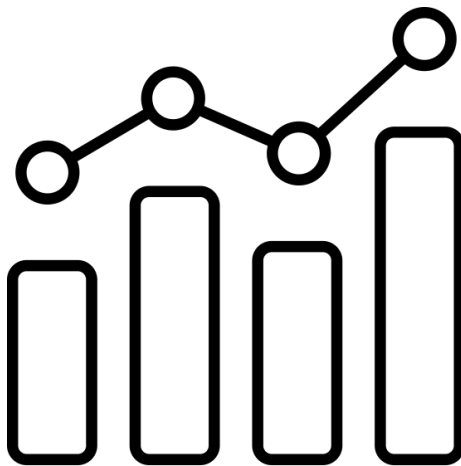
CrossTALK: Cross-Training in AI and Laboratory Knowledge for
Drug Discovery.



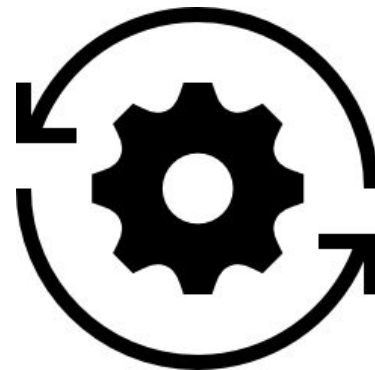
Panorama



Recap

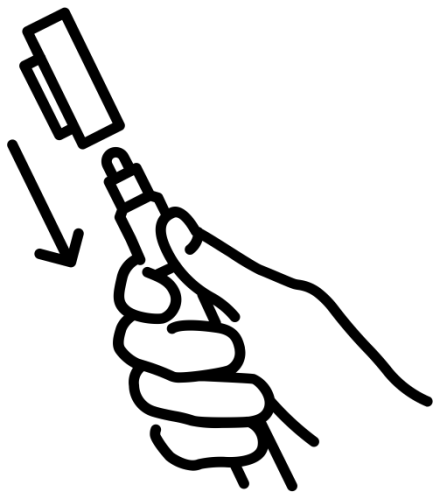


Metrics



Working sesh

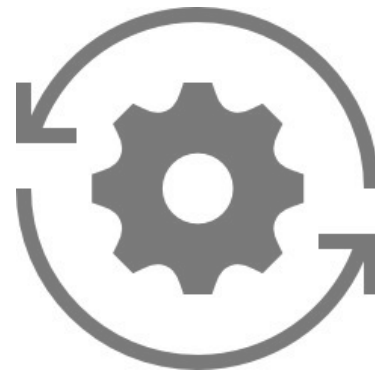
Panorama



Recap



Metrics

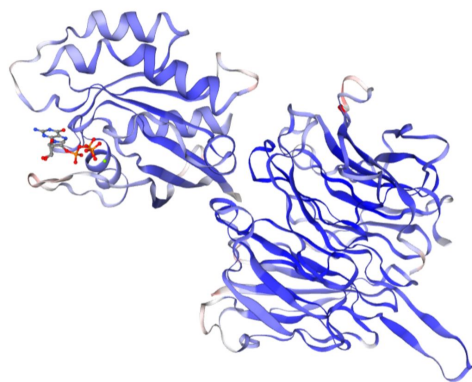


Working sesh

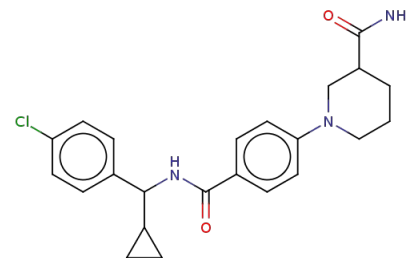
Recap: From Problem to Experiment

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)

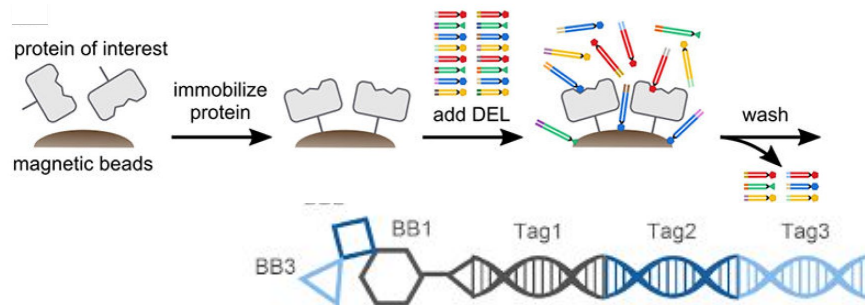


18 molecules with binding data

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

DEL Experiments

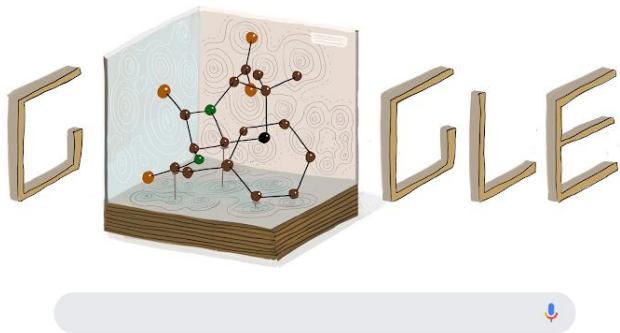
Allow us to test MASSIVE number of molecules



Recap: Computational solutions



Desired solution:
“Molecular search engine”



Filters:
Lipinski's rule of 5

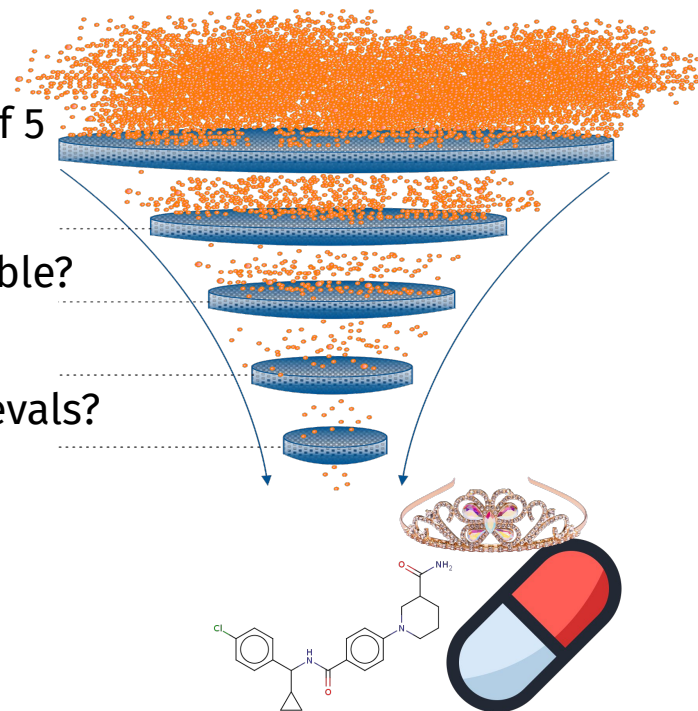
Does it bind?

Orally bioavailable?

Toxic?

Animal/Tissue evals?

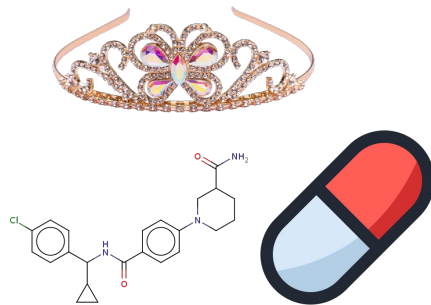
Human evals?



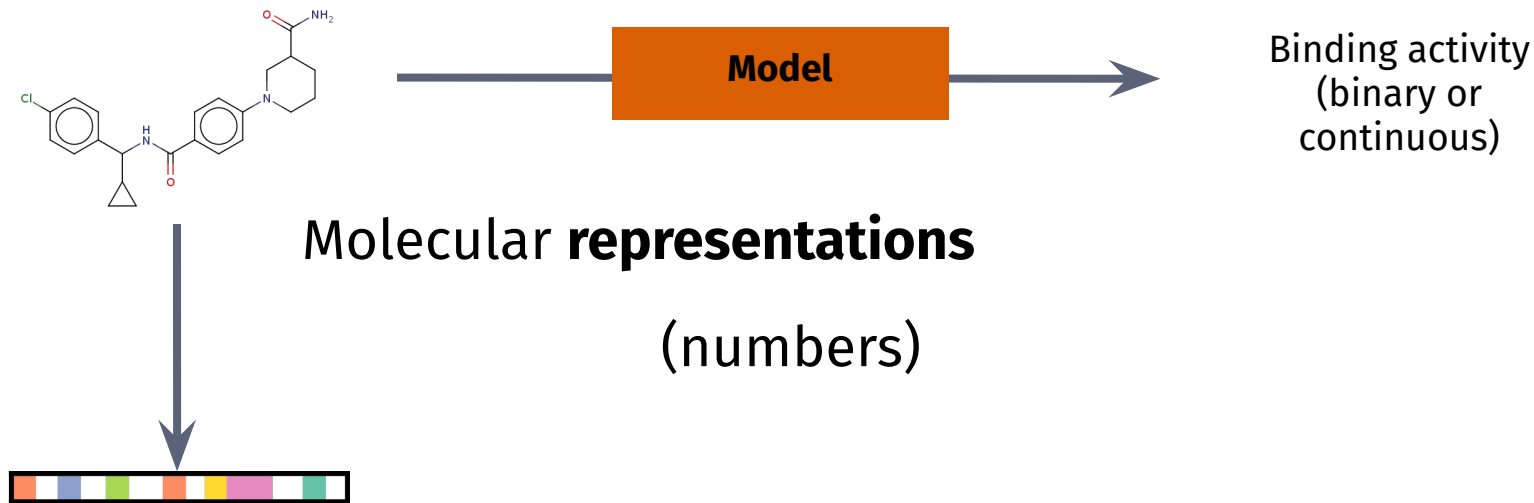
More hits, more shots on goal!

Recap: Caveats, Hits are not drugs

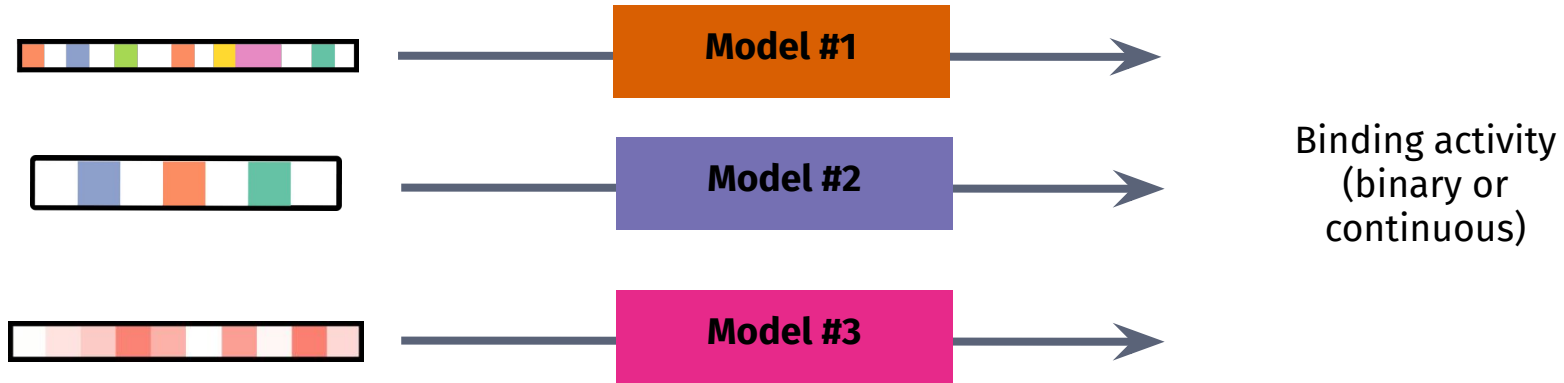
- Low potency
- Low specificity
- Insoluble in water
- Unstable
- Unable to get into cells
- False positives



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



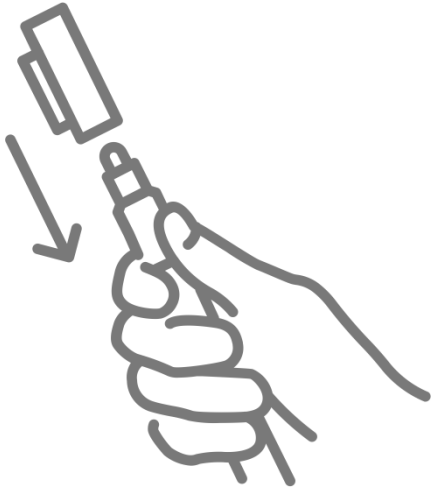
Recap: Which model is the best? (and why?)



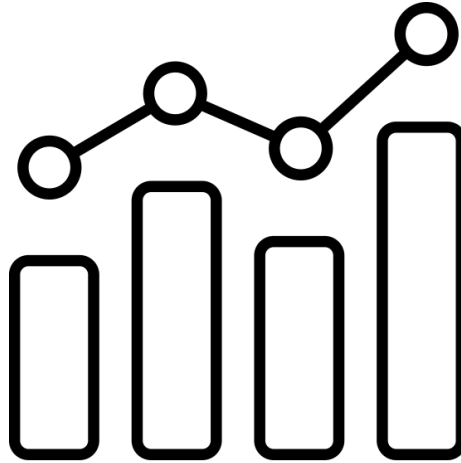
Metrics!

Panorama

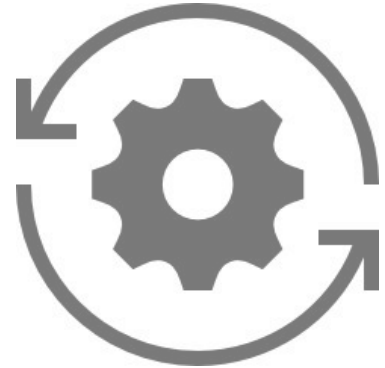
Section slides prepared by Cait Harrigan



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

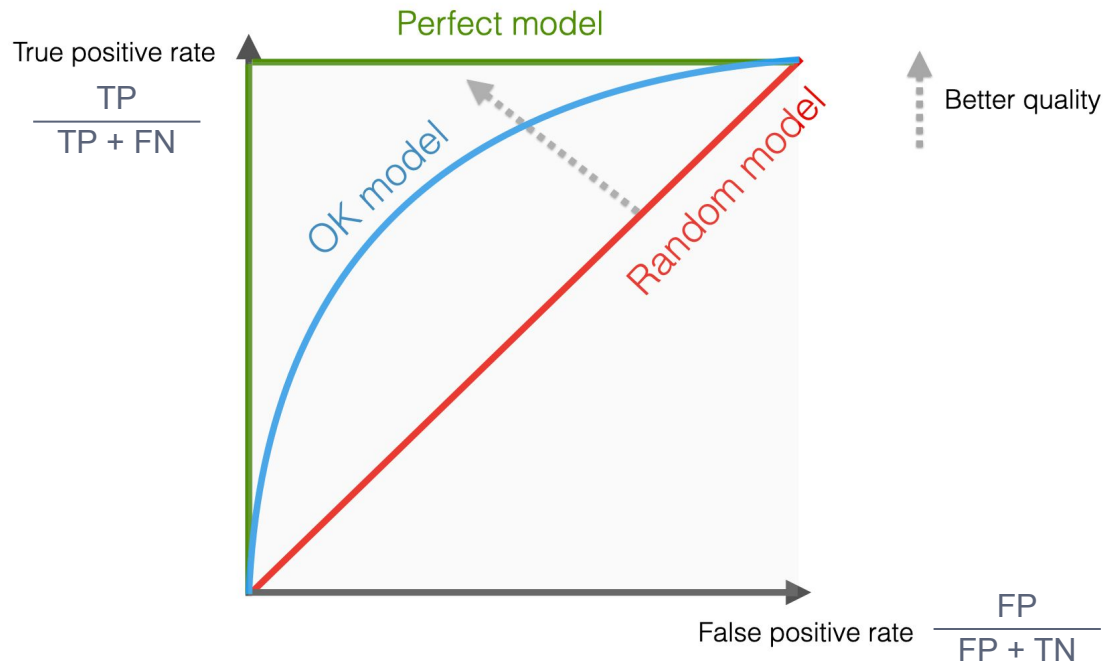
Perfect model	Predict hit	Predict no hit
ASMS hit	100%	0%
ASMS no hit	0%	100%

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



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

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

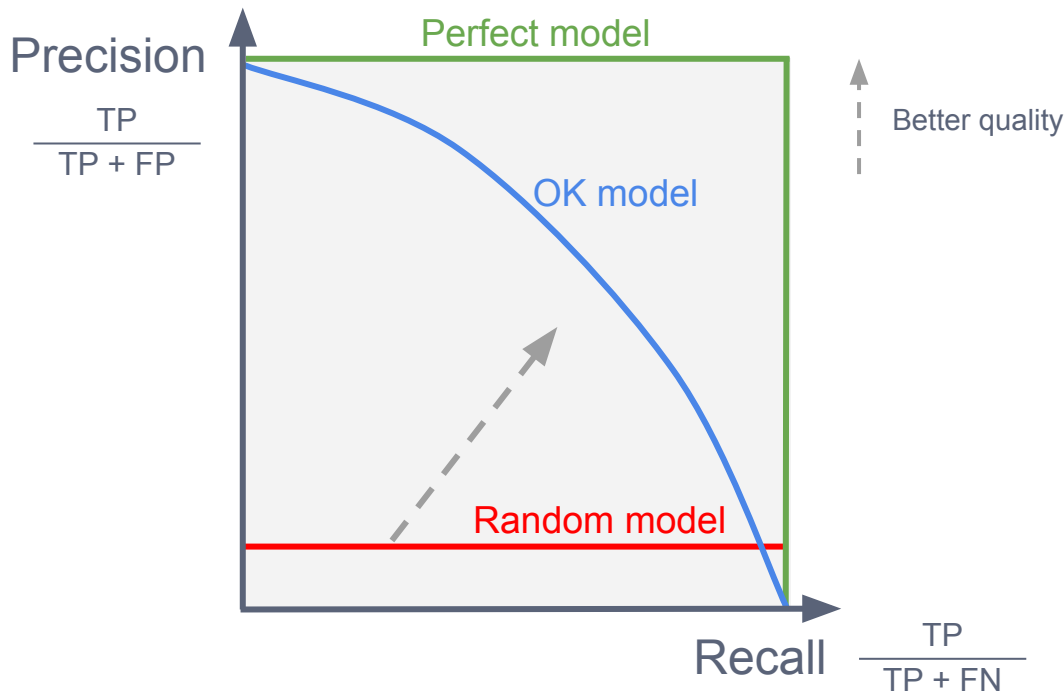
Precision

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 monday night)

Team	#	auroc	#	auprc	#	hits_at_200	#	precision_at_200	#	hits_at_2000	#	precision_at_2000
NOKEYUAN		0.894		0.003		0		0.00		7		0.00
SilvaLeung		0.889		0.004		2		0.01		18		0.01
data loaders		0.883		0.004		1		0.01		15		0.01
Nilooushka		0.863		0.004		3		0.02		15		0.01
Kevin Zhu		0.816		0.002		1		0.01		5		0.00
Lee MICKEY		0.791		0.002		3		0.02		9		0.00
Tech SunMead		0.776		0.001		0		0.00		3		0.00
dhawanikita		0.653		0.001		0		0.00		3		0.00
Sangwook Kim		0.108		0.000		0		0.00		0		0.00


Kaggle: Final Eval Metric has landed

 CROSSTALK - COMMUNITY PREDICTION COMPETITION - PRIVATE - 4 DAYS TO GO

[Submit Prediction](#) ...

CrossTalk_Round2

CrossTalk workshop @ UoT




[Overview](#) [Data](#) [Code](#) [Models](#) [Discussion](#) **[Leaderboard](#)** [Rules](#) [Team](#) [Submissions](#) [Settings](#)


Leaderboard

[Raw Data](#) [Refresh](#)




YOUR RECENT SUBMISSION

 **demo(1).csv** Score: 0.00000
Submitted by CrossTalk · Submitted 3 days ago

[Jump to your leaderboard position](#)

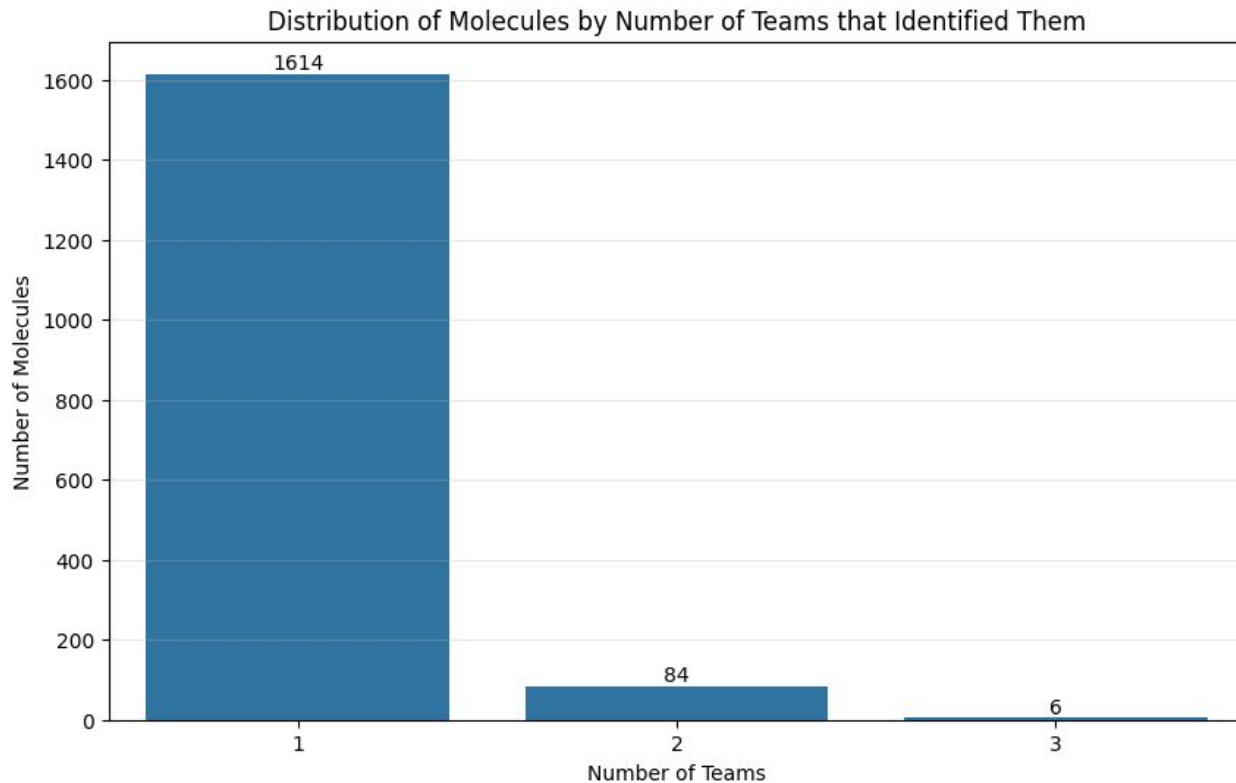
 Search leaderboard

This leaderboard is calculated with all of the test data.

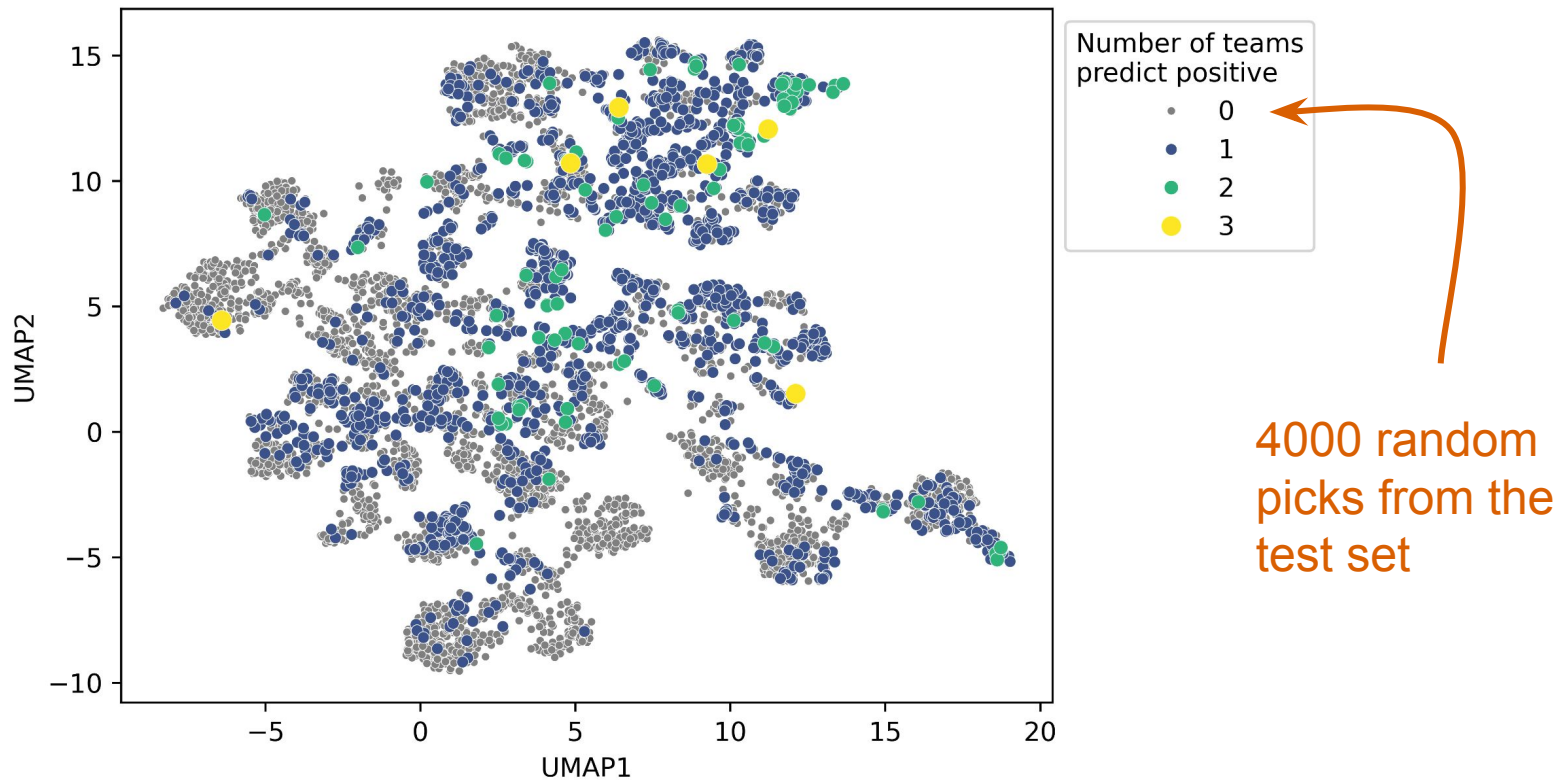
#	Team	Members	Score	Entries	Last	Join
1	dataloaders 		0.95399	3	21h	
2	SilvaLeung		0.92246	2	2d	

Hits @ 200

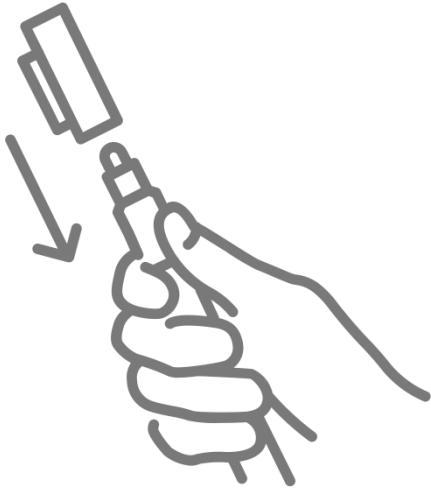
Teams agree on 6 molecules



UMAP on AVALON fingerprints



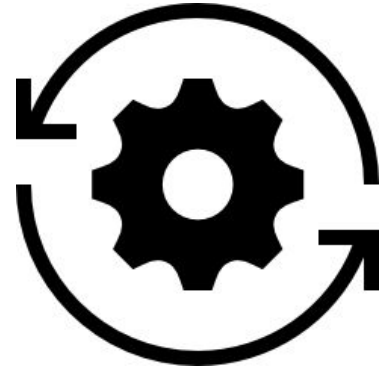
Panorama



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](#))

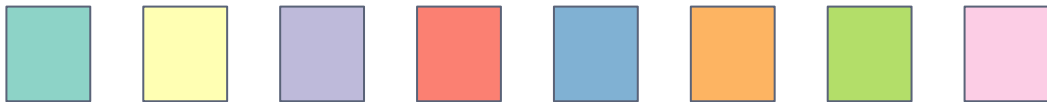
For color palette

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

Dark2



Set3



Cividis (continuous)



PiYG (divergent)

