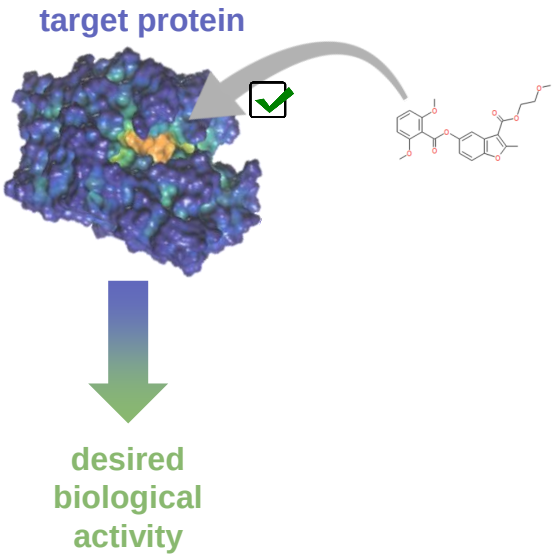
Building Chemogenomics Models From a Large-Scale Public Dataset and Applying them to Industrial Datasets

AstraZeneca

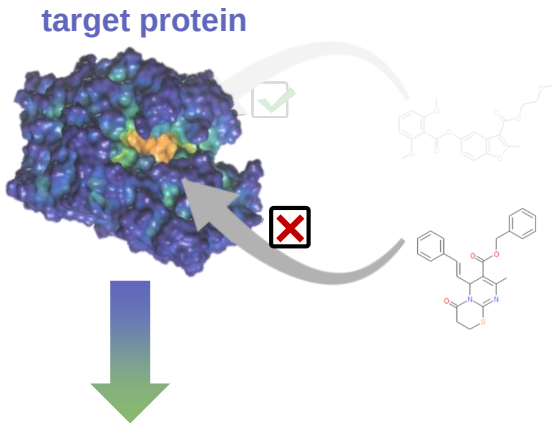
Funded by the European Union's Horizon 2020 Research and Innovation program.
Grant Agreement no. 654168

Noé Sturm – *Bled, Slovenia* – May 2019

Chemogenomics modelling

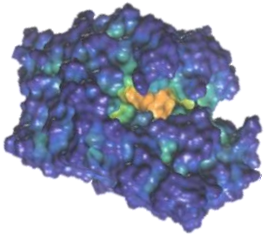


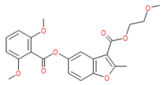
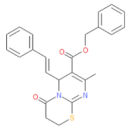
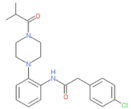
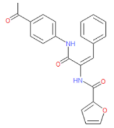
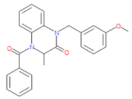
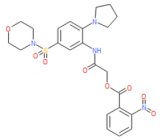
Chemogenomics modelling



Chemogenomics modelling

target protein



X	y_true	*value*
	✓	
	✗	
	✗	
	✓	
	✗	
	✗	

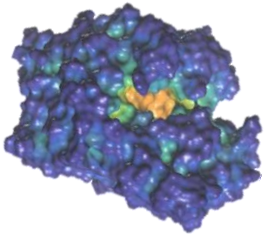
{ Molecule ; Activity }

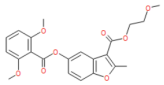
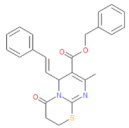
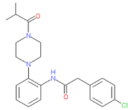
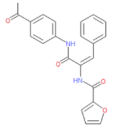
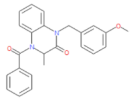
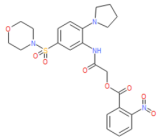
X ⇨ ***y_true***

Classification: *y_true* is binary
Regression: *y_true* is continuous

Chemogenomics modelling

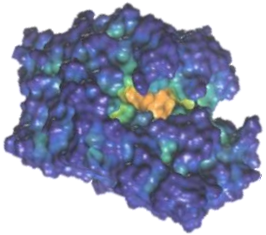
target protein



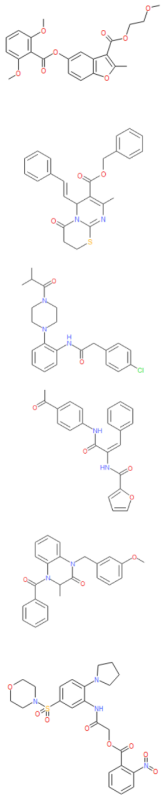
	target 1	target 2	target 3	target 4	.....	target N
	✓	✗	✗	✗		✓
	✗	✓	✗	✓		✗
	✗	✗	✗	✓		✗
	✓	✗	✗	✗		✗
	✗	✗	✗	✗		✓
	✗	✗	✓	✗		✗

Chemogenomics modelling

target protein



target 1 target 2 target 3 target 4 ..... target N



{ Molecule ; Activity ; Target }

X ⇨ ***y_true***

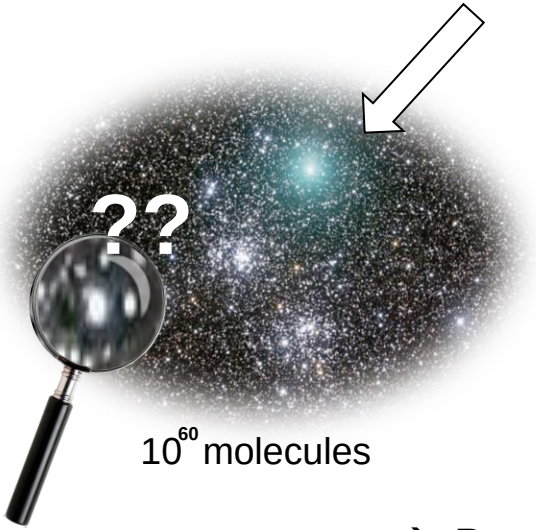
Multi-task: *y_true* is a vector

Aim of chemogenomics: Accelerate Drug Discovery

Drug Discovery Pipeline

Aim of chemogenomics: Accelerate Drug Discovery

Drug Discovery Pipeline

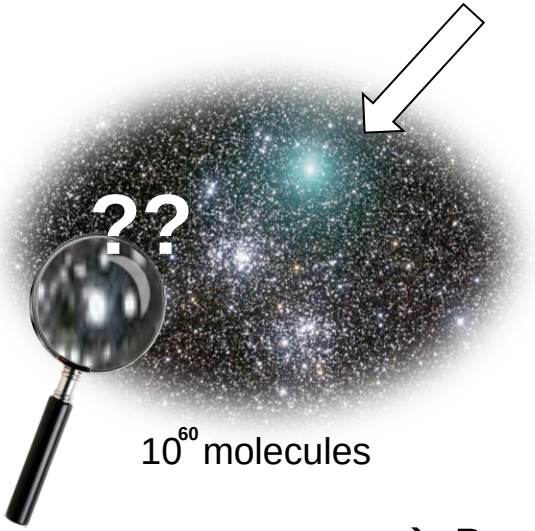


10^{60} molecules

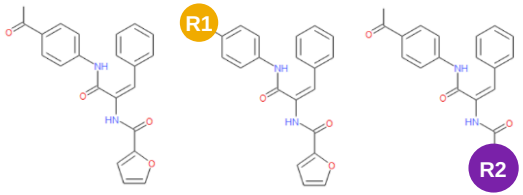
→ Predict activity of compounds from uncharted chemical space

Aim of chemogenomics: Accelerate Drug Discovery

Drug Discovery Pipeline



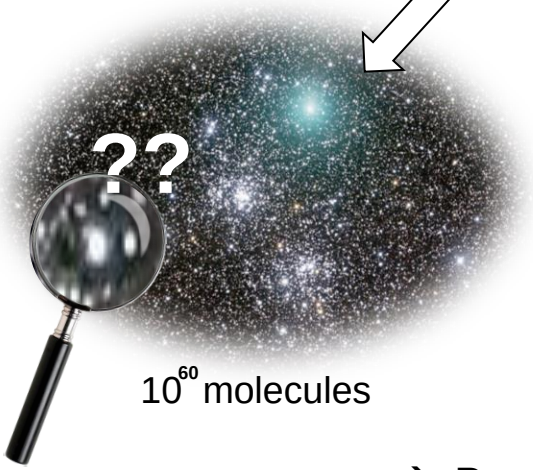
10^{60} molecules



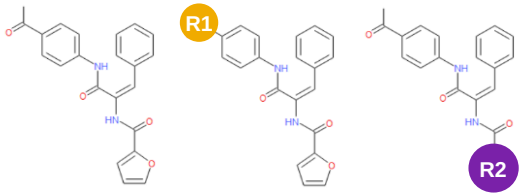
→ Predict activity of compounds from uncharted chemical space
→ Be sensitive enough to handle subtle chemical changes

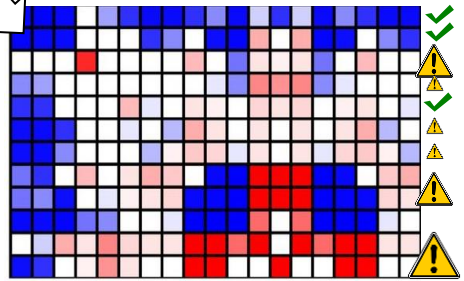
Aim of chemogenomics: Accelerate Drug Discovery

Drug Discovery Pipeline



10^{60} molecules

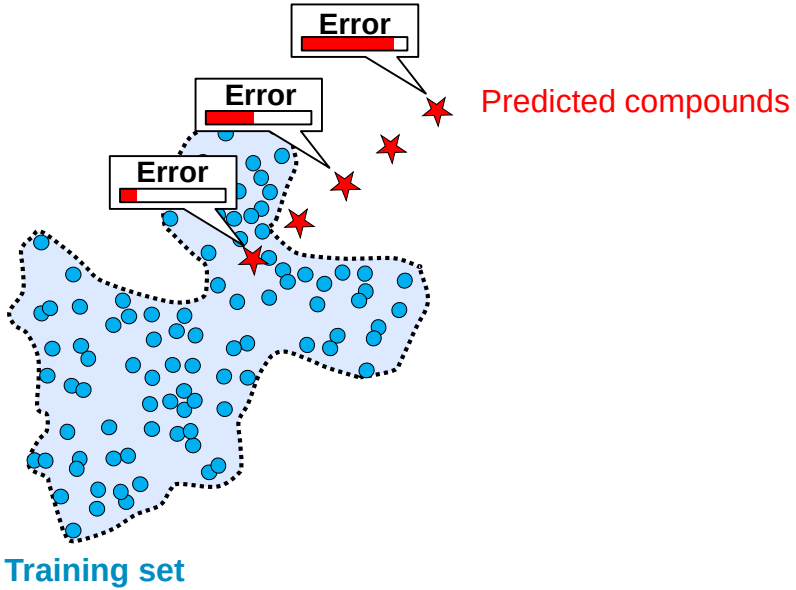




Safety profiling panels

→ Predict activity of compounds from uncharted chemical space
→ Be sensitive enough to handle subtle chemical changes
→ Be accurate across targets (MOA elucidation, safety)

Challenge*:* Application Domain

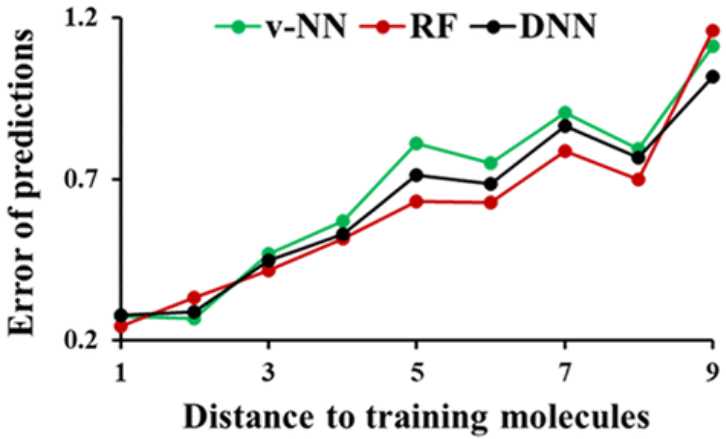


Challenge: Application Domain

Dissecting Machine-Learning Prediction of Molecular Activity: Is an Applicability Domain Needed for Quantitative Structure–Activity Relationship Models Based on Deep Neural Networks?

Ruifeng Liu[*,†], Hao Wang[†], Kyle P. Glover[§], Michael G. Feasel[§], and Anders Wallqvist[*,†]
[†] Department of Defense, Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, Fort Detrick, Maryland 21702, United States
[§] U.S. Army−Edgewood Chemical Biological Center, Aberdeen Proving Ground, Maryland 21010, United States

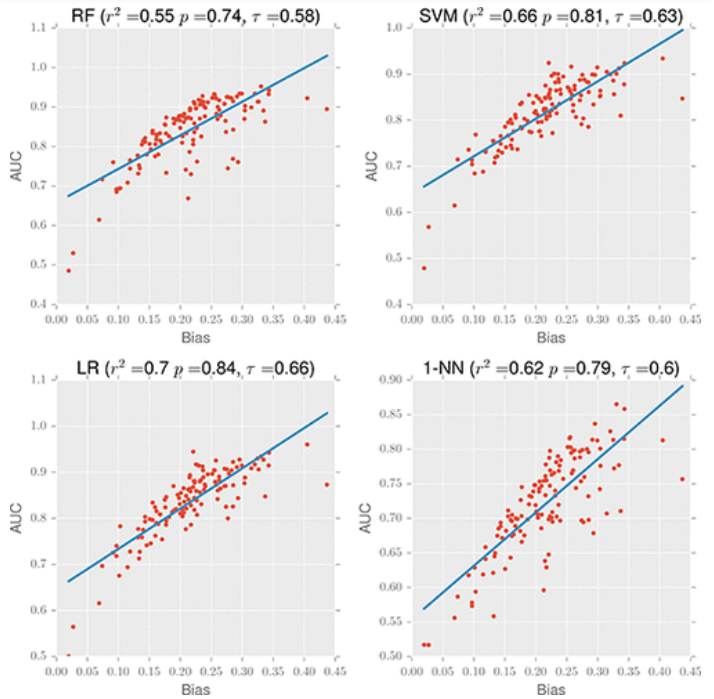
Distance to training molecules is the most important determinant of prediction accuracy



Liu et al, J. Chem. Inf. Model., **2019**, *59* (1), pp 117–126

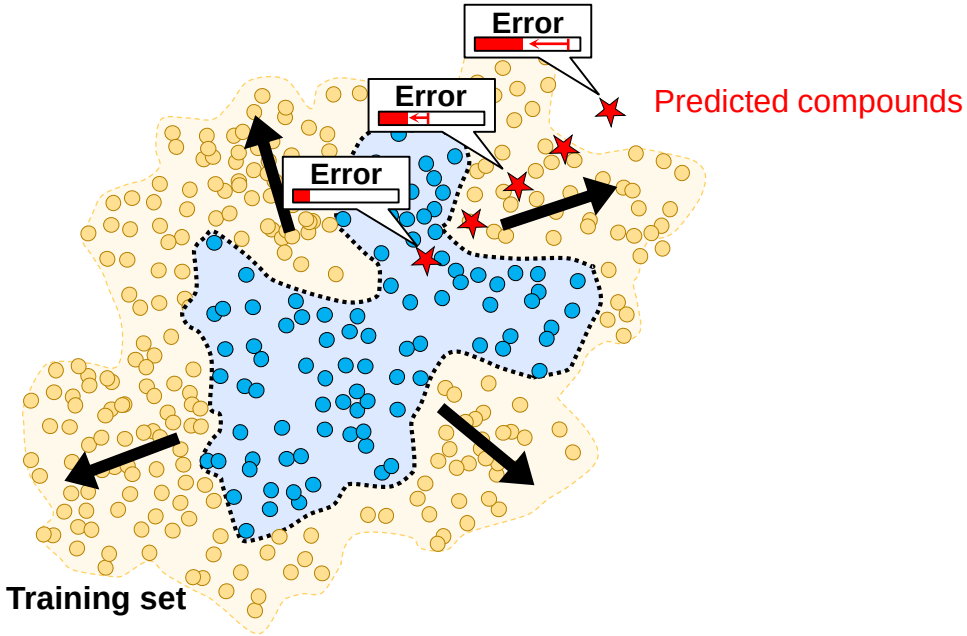
Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization

Izhar Wallach[*] and Abraham Heifets[*]
Atomwise Inc., 221 Main Street, Suite 1350, San Francisco, California 94105, United States



Wallach I., Heifets A., J. Chem. Inf. Model., **2018**, 58 (5), pp 916–932

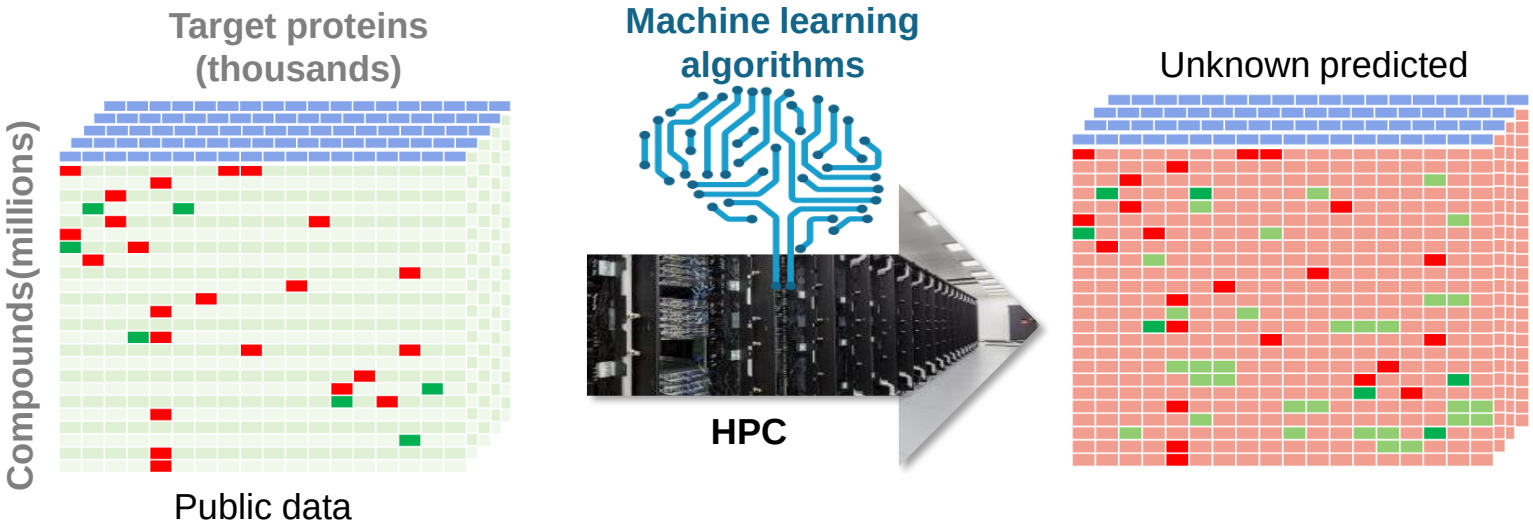
Challenge*:* Application Domain



One Possible solution:

→ Expand training set
→ Integrate many examples of (active, inactive)
→ public ML benchmark dataset

Exa-Scale Compound Activity Prediction Engine



But also,
predict new compounds!

Funded by the European Union's Horizon 2020 Research and Innovation program.
Grant Agreement no. 671555

Team

9 partners / 8 countries

Chemogenomics Data

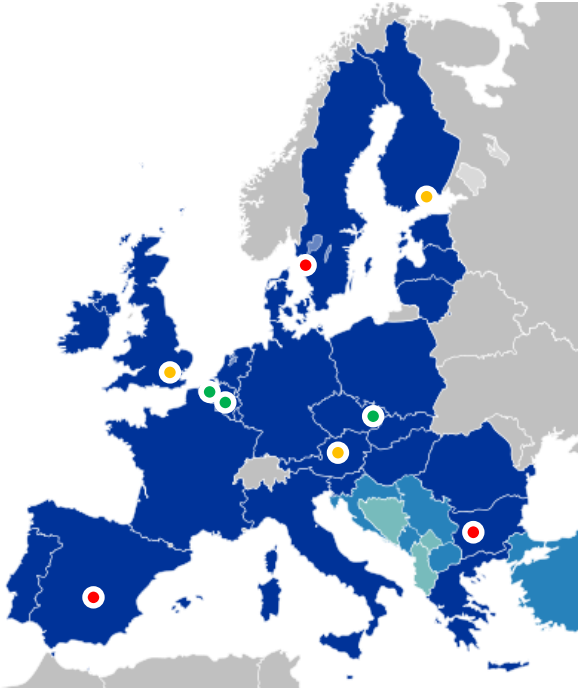
- **Compound activity**
- **Data preparation**
- **Benchmarking**

High Performance Computing

- **Cluster deployment**
- **Programming model**

Machine Learning Algorithms

- **Deep learning**
- **Matrix factorisation**
- **Conformal prediction**



www.excape-h2020.eu

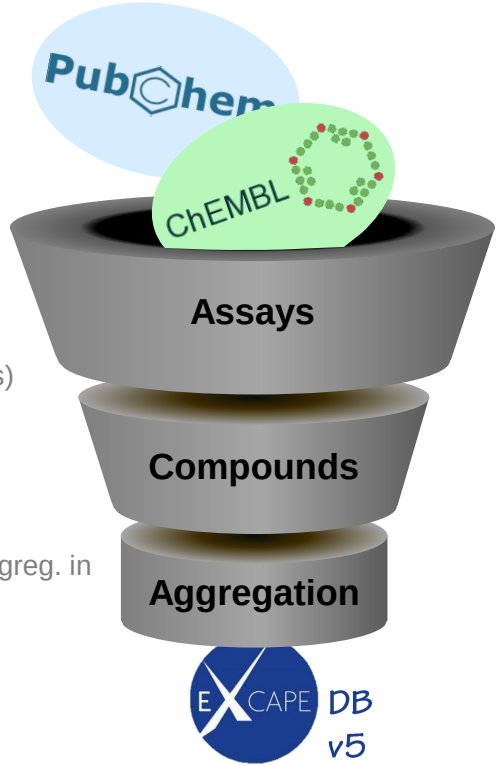
Funded by the European Union's Horizon 2020 Research and Innovation program.
Grant Agreement no. 671555

Overall Workflow

1. Dataset collection
2. Dataset split
3. Descriptors
4. Performance metrics
5. Algorithms
6. Hyperparameter selection
7. Retrospective model evaluation
8. Prospective model evaluation

Dataset choice: ChEMBL + PubChem = *ExCAPE-DB*

Jiangming Sun
Hongming Chen
Felip Golib
Vladimir Chupakhin
Nina Jeliazkova



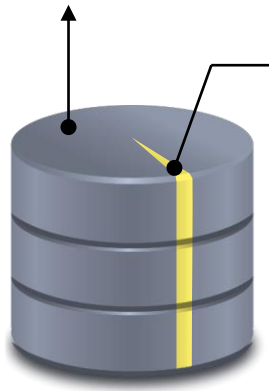
- ✓ Single protein assays
- ✓ Human, mouse, rat
- ✓ Activity ≤ 10 μM
- ✓ Inactive (flags from hts)

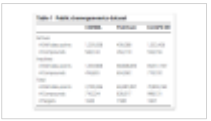
- ✓ Structure standization
- ✓ MW < 1000
- ✓ Heavy atoms > 12

- ✓ Compound activity aggreg. in pXC50

INACTIVES
Data pts 69,517,737
(719,192 cmpds)

ACTIVES
Data pts 1,332,426
(593,156 cmpds)





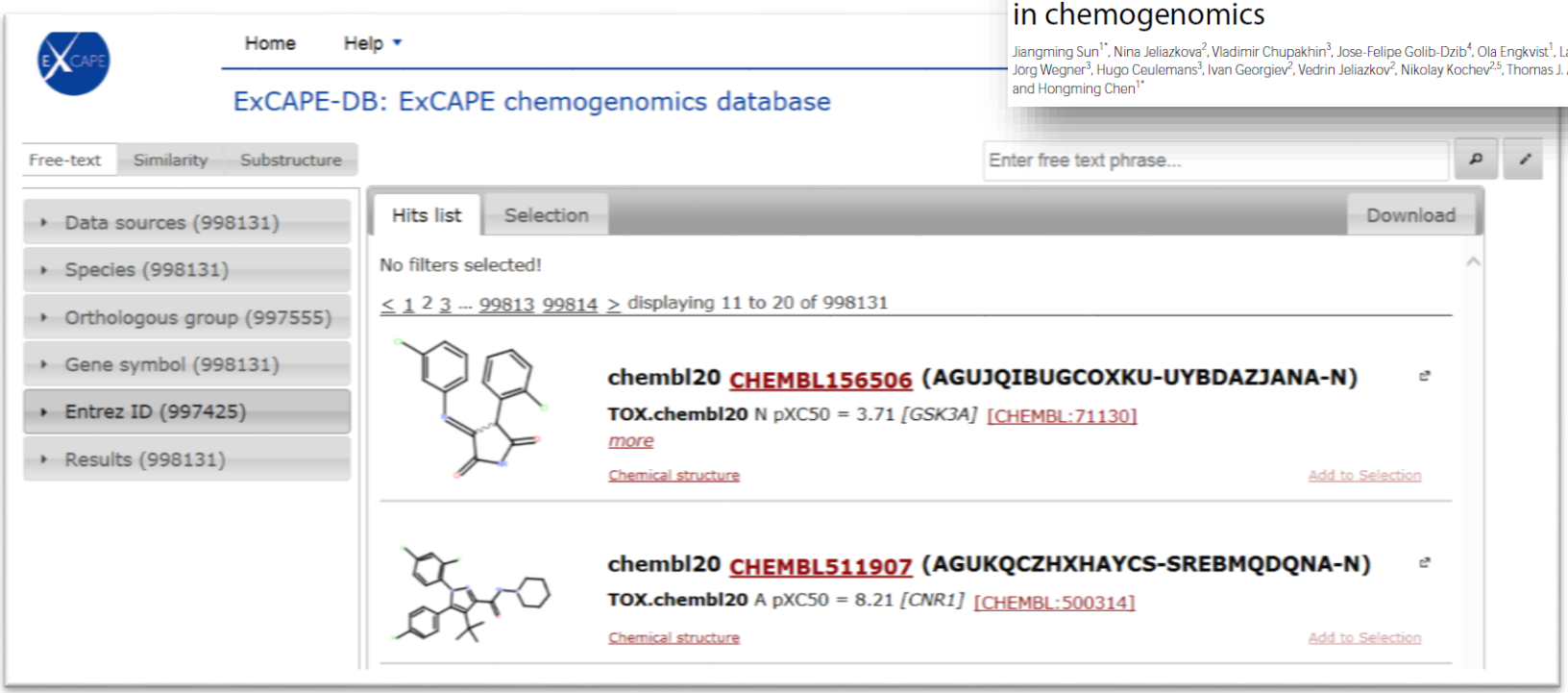
TOTAL

70,850,153	data points
998,131	unique compounds
1667	target proteins

Sun et al. J Cheminform, 2017, **9**:17

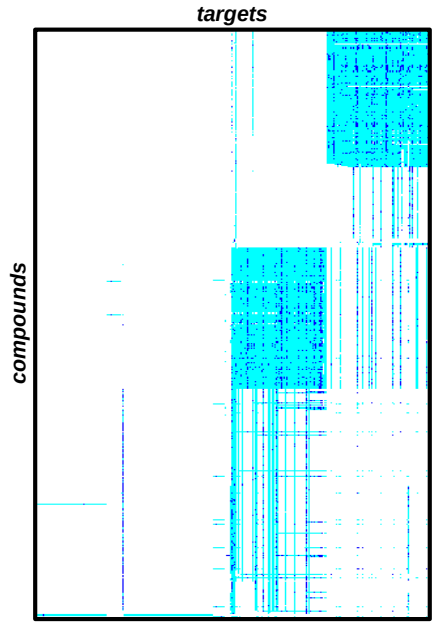
Chemogenomics data: *ExCAPE-DB*

https://solr.ideaconsult.net/search/excape/



Machine learning dataset: *ExCAPE-ML*

ExCAPE-DB v5



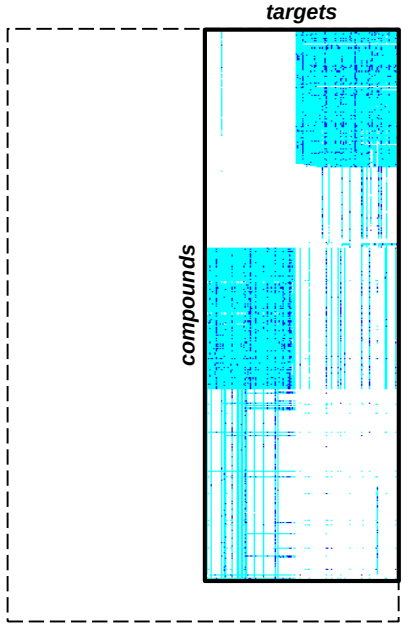
Sun et al. J Cheminform, 2017, **9**:17

pAct < 6 (inactive)
pAct ≥ 6 (active)
missing

FILTER

N ≥ 300
N_active ≥ 75
N_inactive ≥ 75

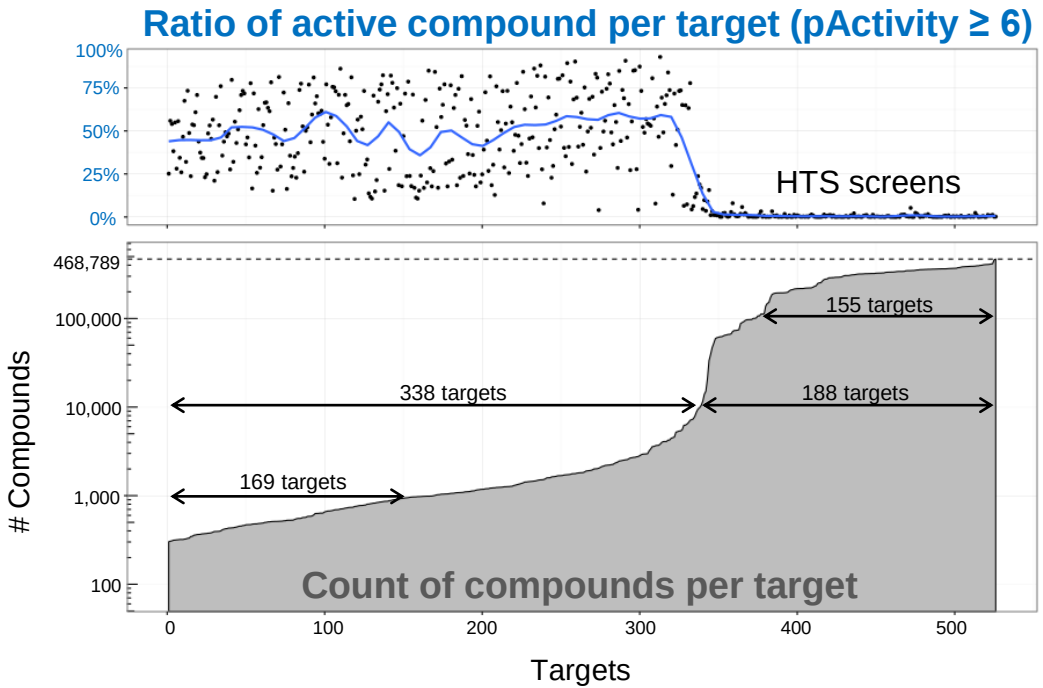
ExCAPE-ML v5



70,850,153 - data points (A:I = 69.5M: 1.3M)
998,131 - unique compounds
1667 - target proteins
4.3% - density

49,516,318 - data points (A:I = 48.8M: 0.5M)
955,386 - unique compounds
526 - target proteins
9.8% - density

Target dataset sizes and active ratio



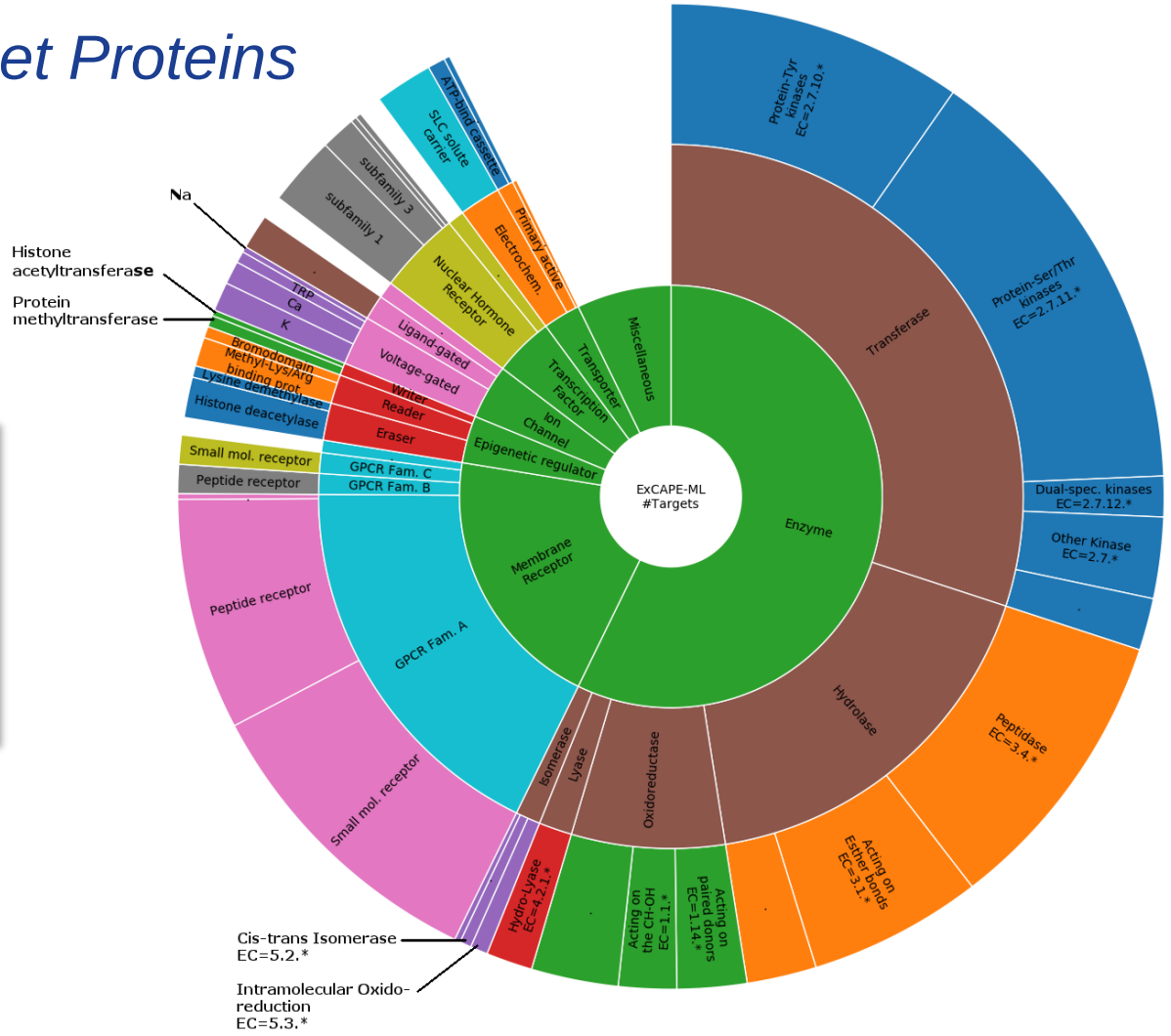
- 188 targets tested ≥10k
- 155 targets tested ≥100k

- 321 targets with ≥15% actives have ~1500 data points on average.

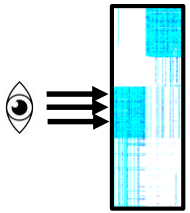
- 205 targets with <15% actives have ~238,000 data points on average.

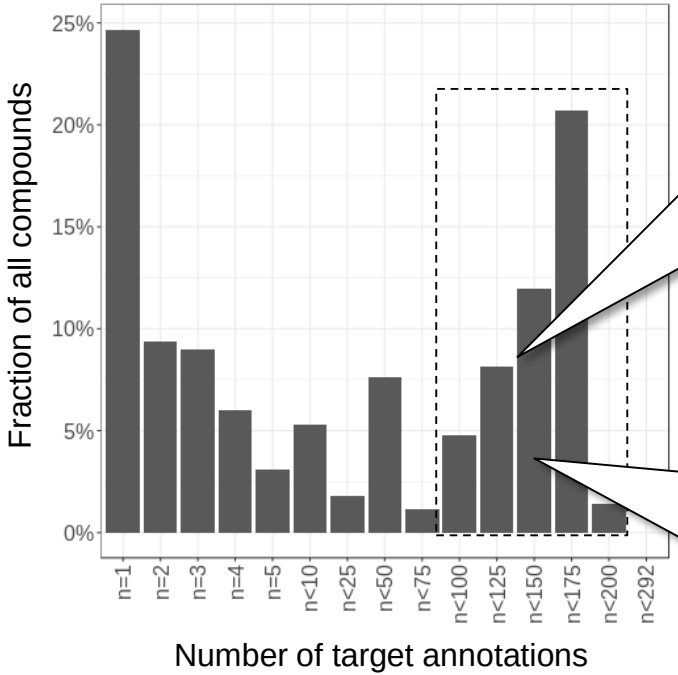
ExCAPE-ML: *Target Proteins*

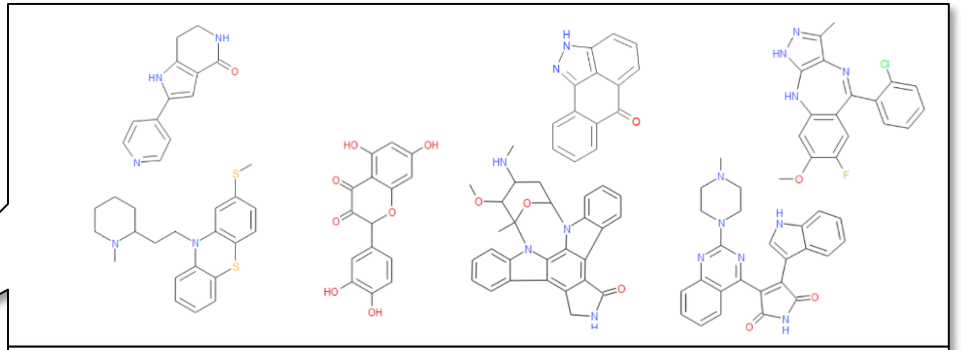
301 Enzymes
107 Membrane receptors
24 Transcription factors
22 Ion channels
19 Epigenetic regulators
15 Transporters



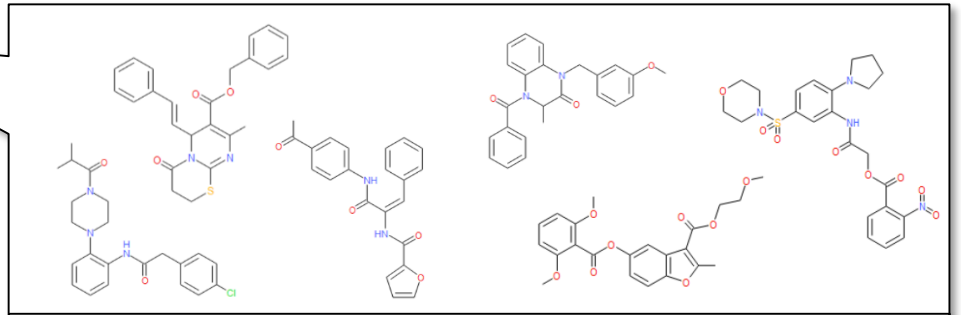
Compounds annotations in ExCAPE-ML







100 annotations , > 25 protein hits

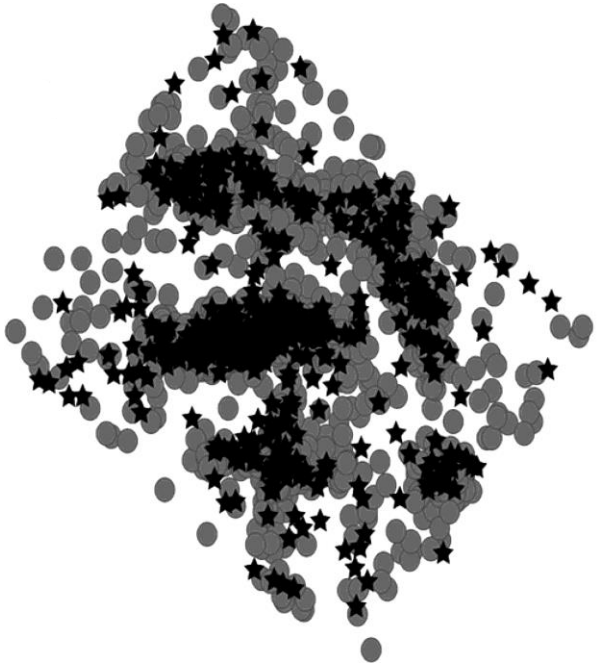


100 annotations , 1 protein hit

Overall Workflow

1. Dataset collection
2. Dataset split
3. Descriptors
4. Performance metrics
5. Algorithms
6. Hyperparameter selection
7. Retrospective model evaluation
8. Prospective model evaluation

Chemical Series Bias



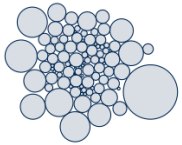
Random split

(Over optimistic performance)

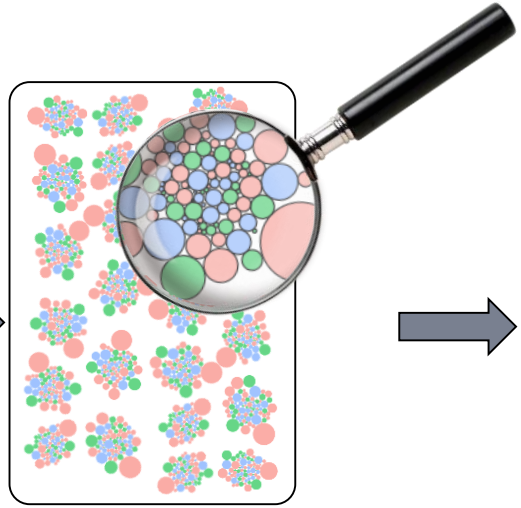
Martin et al., J. Chem. Inf. Model., **2017**, *57* (8), pp 2077–2088

Chemical Series Bias: *a Realistic Dataset Split*

955K Compounds
526 Targets



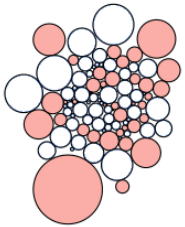
Cluster molecules with ECFP6



Search 3 folds with:

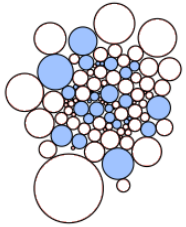
- 526 targets in each fold
- ~ even distribution of compounds

Fold-1



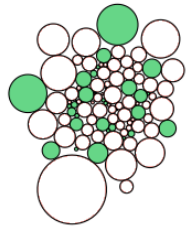
#molecules: 332,657

Fold-2



#molecules: 308,169

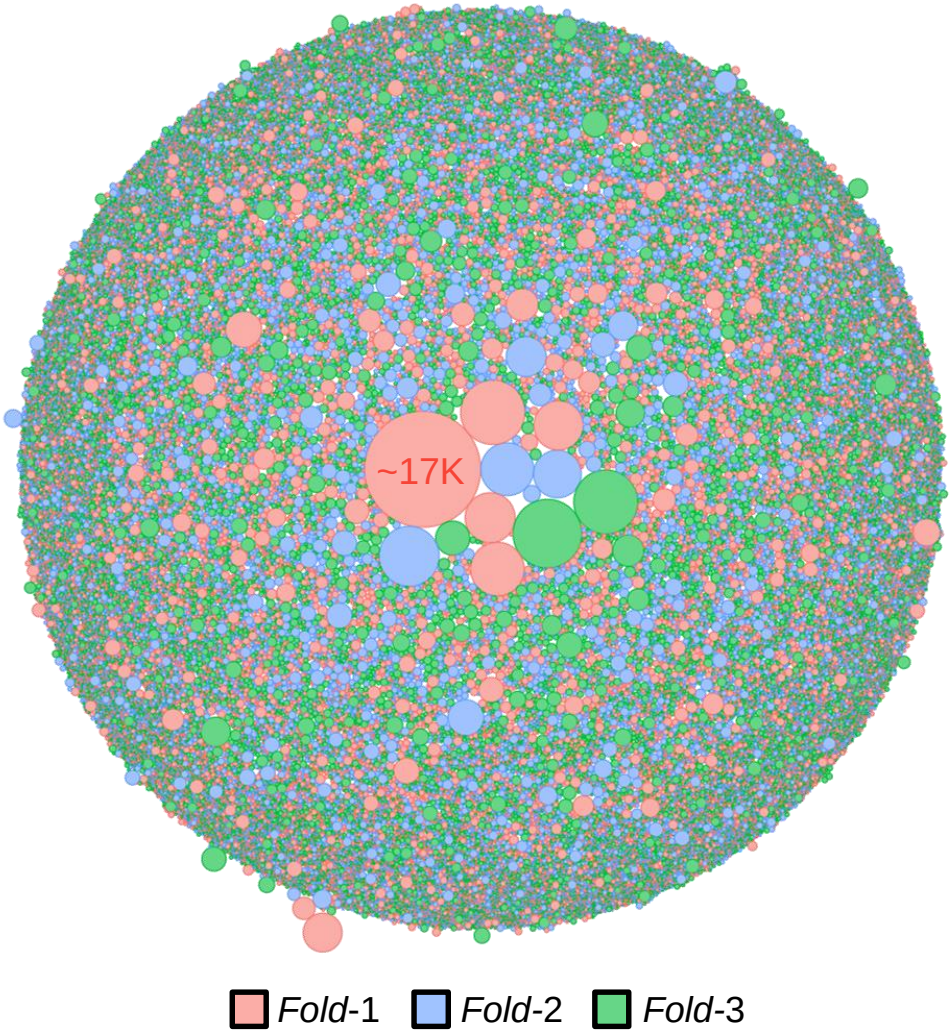
Fold-3



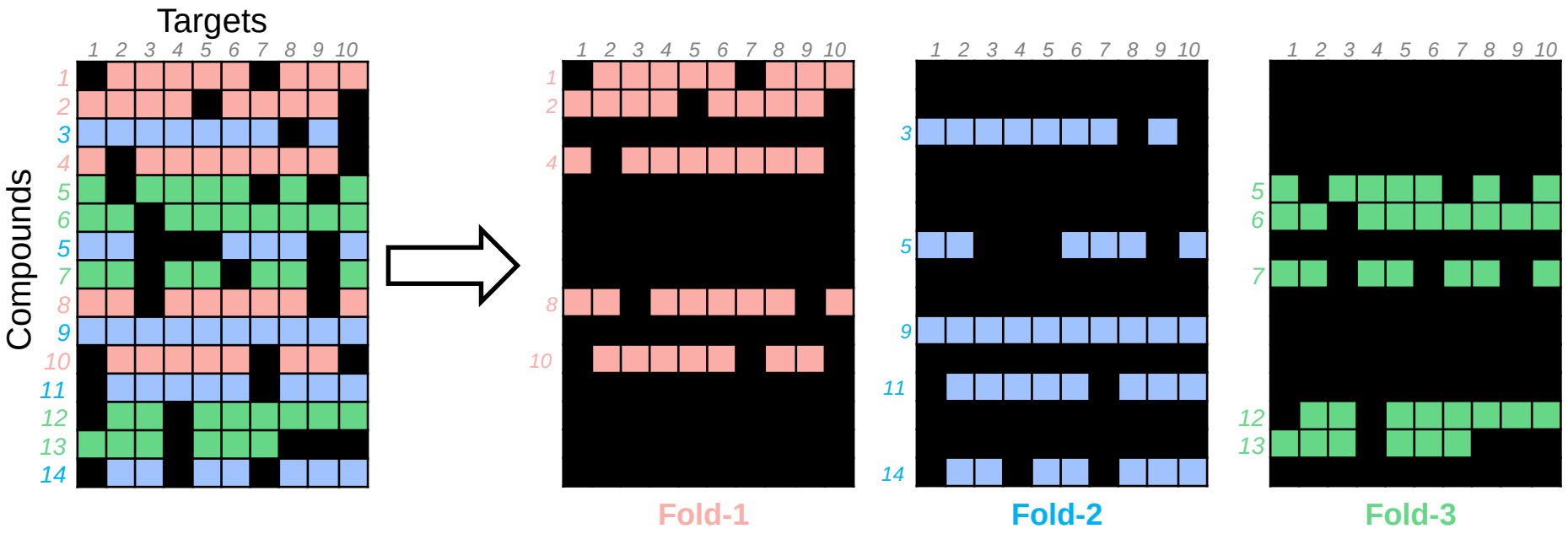
#molecules: 314,560

Cluster sizes

Fold	Clusters		Singletons	Total
	#cluster	#molec	#molec	#molec
1	20,614	320,451	12,206	332,657
2	21,204	295,557	12,612	308,169
3	20,592	302,330	12,230	314,560



Chemogenomics Matrix [C x P]

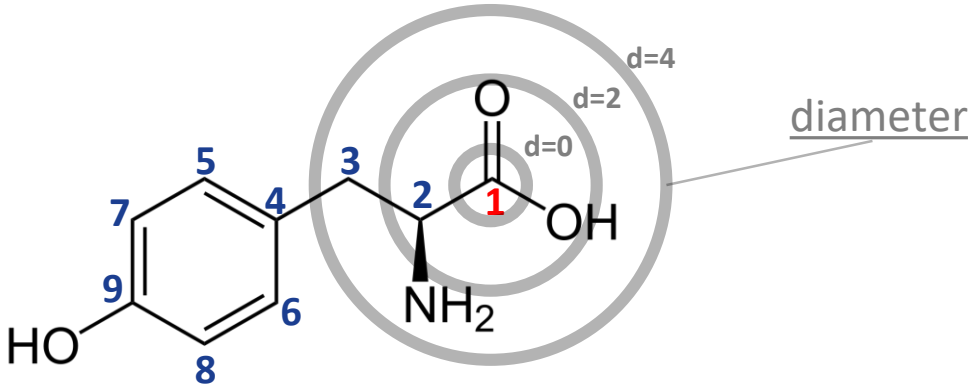


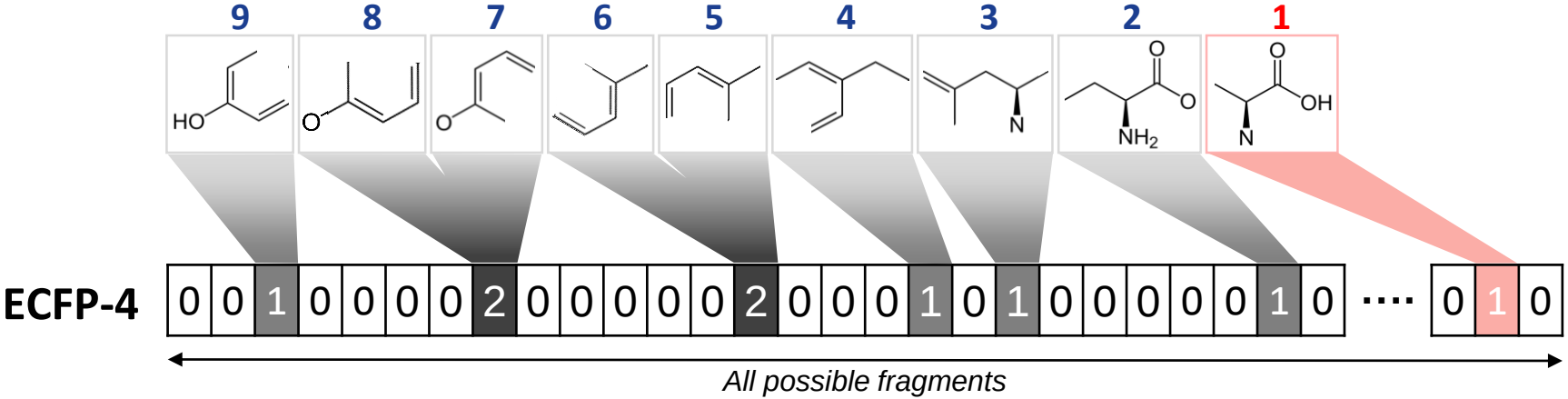
SciPy sparse matrices

Overall Workflow

1. Dataset collection
2. Dataset split
3. Descriptors
4. Performance metrics
5. Algorithms
6. Hyperparameter selection
7. Retrospective model evaluation
8. Prospective model evaluation

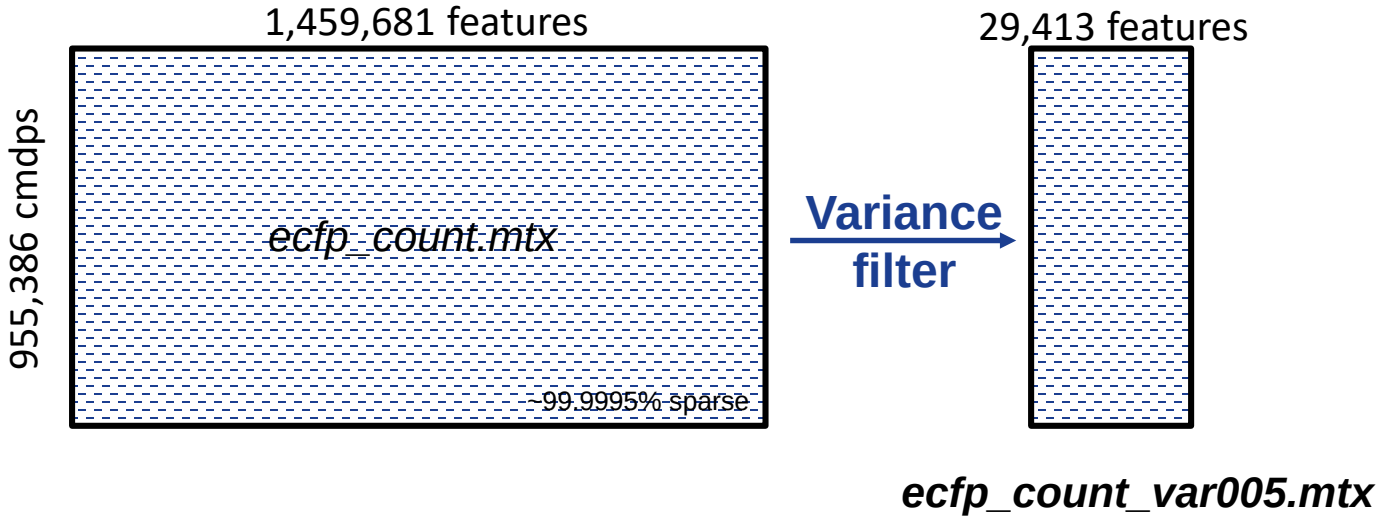
Molecular Features: *Extended Connectivity Fingerprint*





David Rogers and Mathew Hahn, J. Chem. Inf. Model., **2010**, *50* (5), pp 742–754

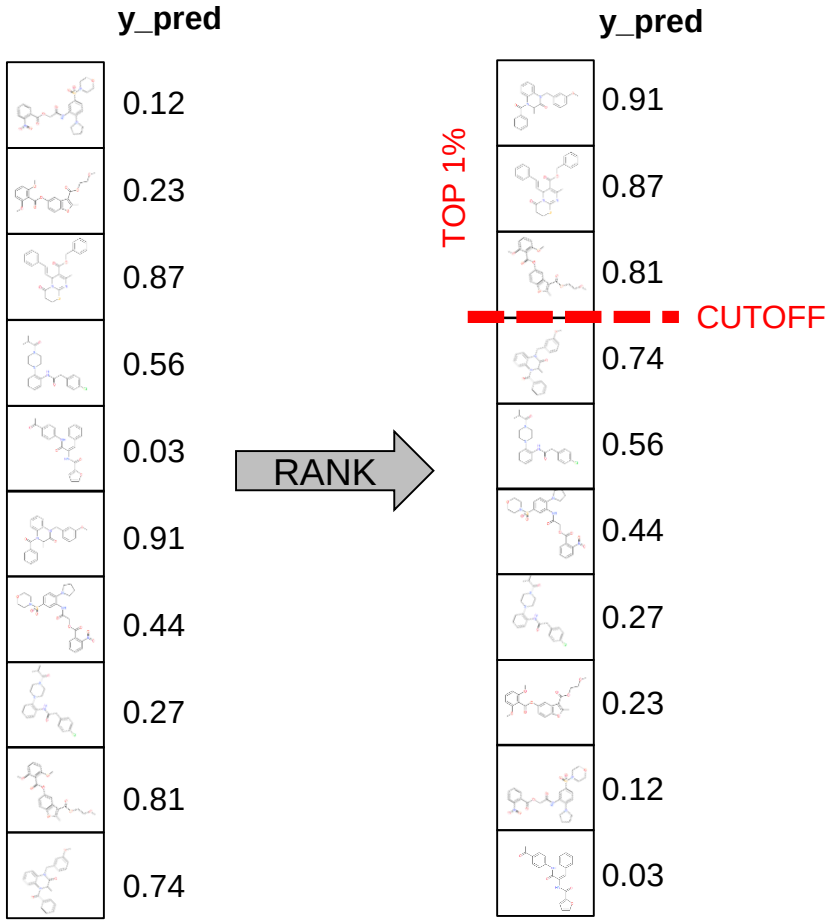
Molecular Features: *ECFP-6*



Overall Workflow

1. Dataset collection
2. Dataset split
3. Descriptors
4. Performance metrics
5. Algorithms
6. Hyperparameter selection
7. Retrospective model evaluation
8. Prospective model evaluation

Virtual Screening: *prioritize active compounds*

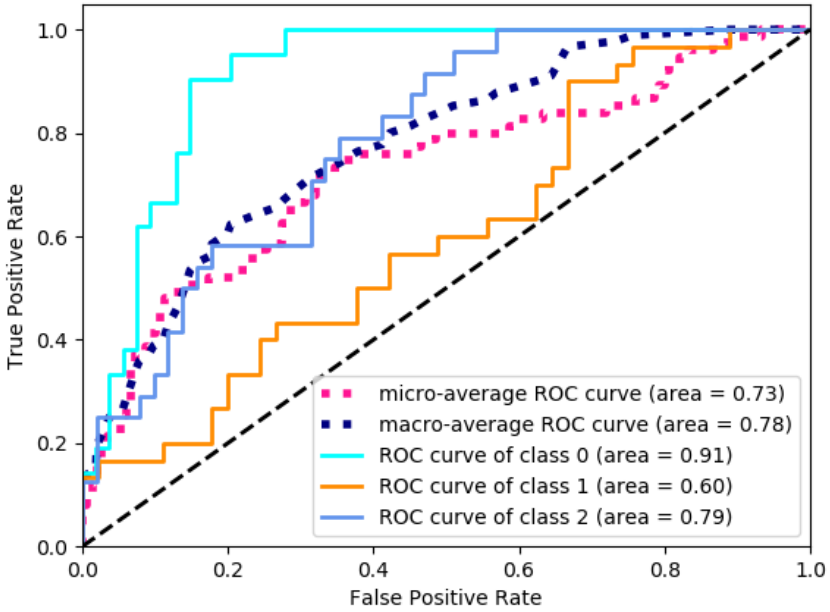


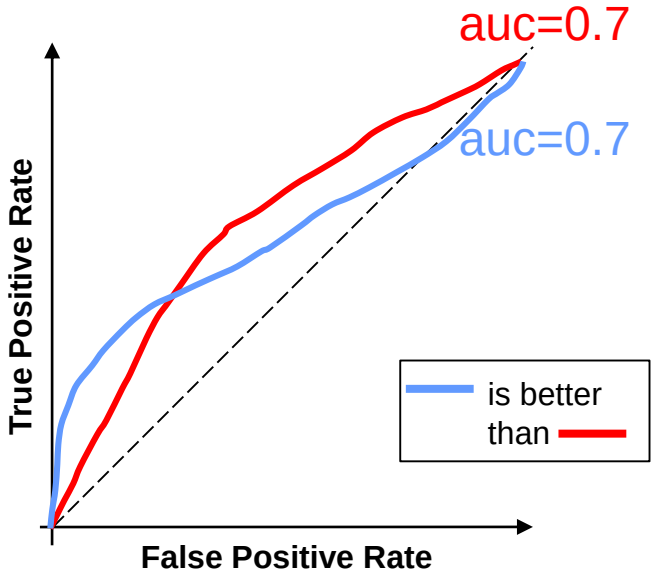
Active compound enrichment ***vs*** HTS overall hit rate

Performance metrics : *ROC-AUC*

$$\boldsymbol{ROCAUC} = \int_{-\infty}^{+\infty} TPR(T)FPR'(T)dT$$

TPR *: True Positive Rate* | ***FPR*** *: False Positive Rate*





https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

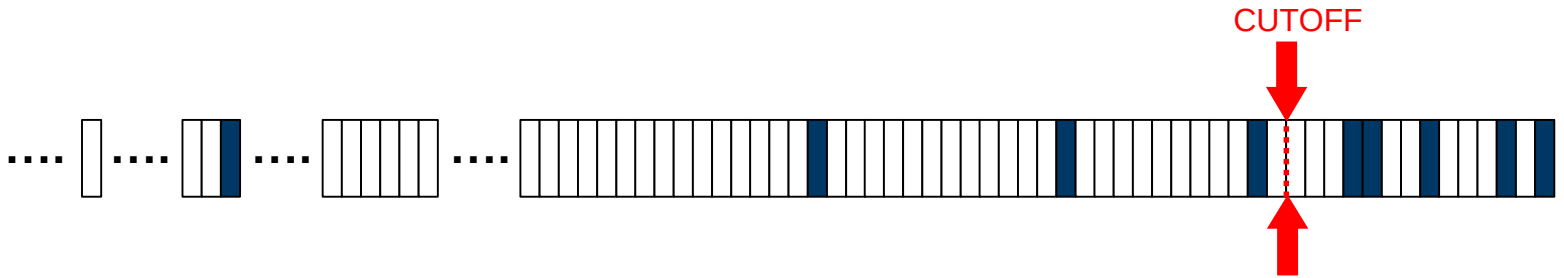
Performance metrics: *Confusion Matrix*

$$\textbf{Precision} = \frac{TP}{TP+FP}$$

$$\textbf{Recall} = \frac{TP}{TP+FN}$$

$$\textbf{F1-score} = 2 \times \frac{precision\ .\ recall}{precision + recall}$$

TP *: True Positive* | ***FP*** *: False Positive* | ***FN*** *: False Negative* | ***TN*** *: True Negative*

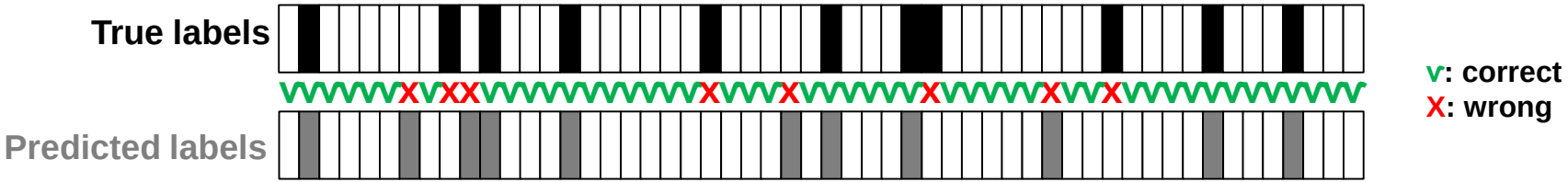


Performance metrics: *Cohen-Kappa*

$$\kappa = \frac{p_0 - p_e}{1 - p_e}$$

Inter-classification agreement

P_0 : *Agreement between two classifiers (accuracy)*
P_e : *Expected agreement by chance of two classifiers*



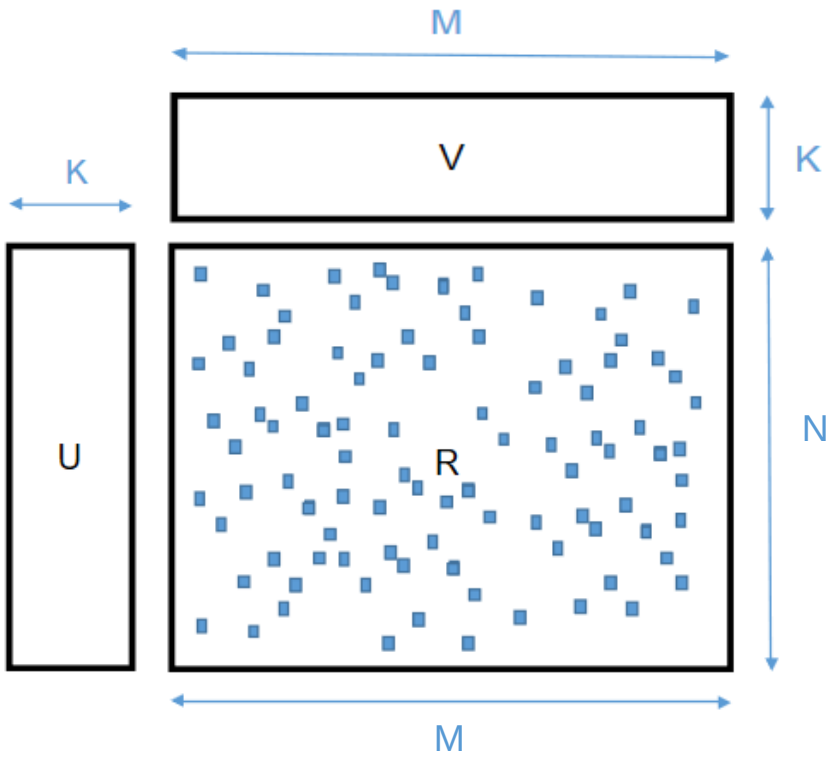
Overall Workflow

1. Dataset collection
2. Dataset split
3. Descriptors
4. Performance metrics
5. Algorithms
6. Hyperparameter selection
7. Retrospective model evaluation
8. Prospective model evaluation

Matrix Factorization : *SMURFF / Macau*

imec *Tom Vander*
Tom Ashby

Molecular descriptors (ECFP-6)

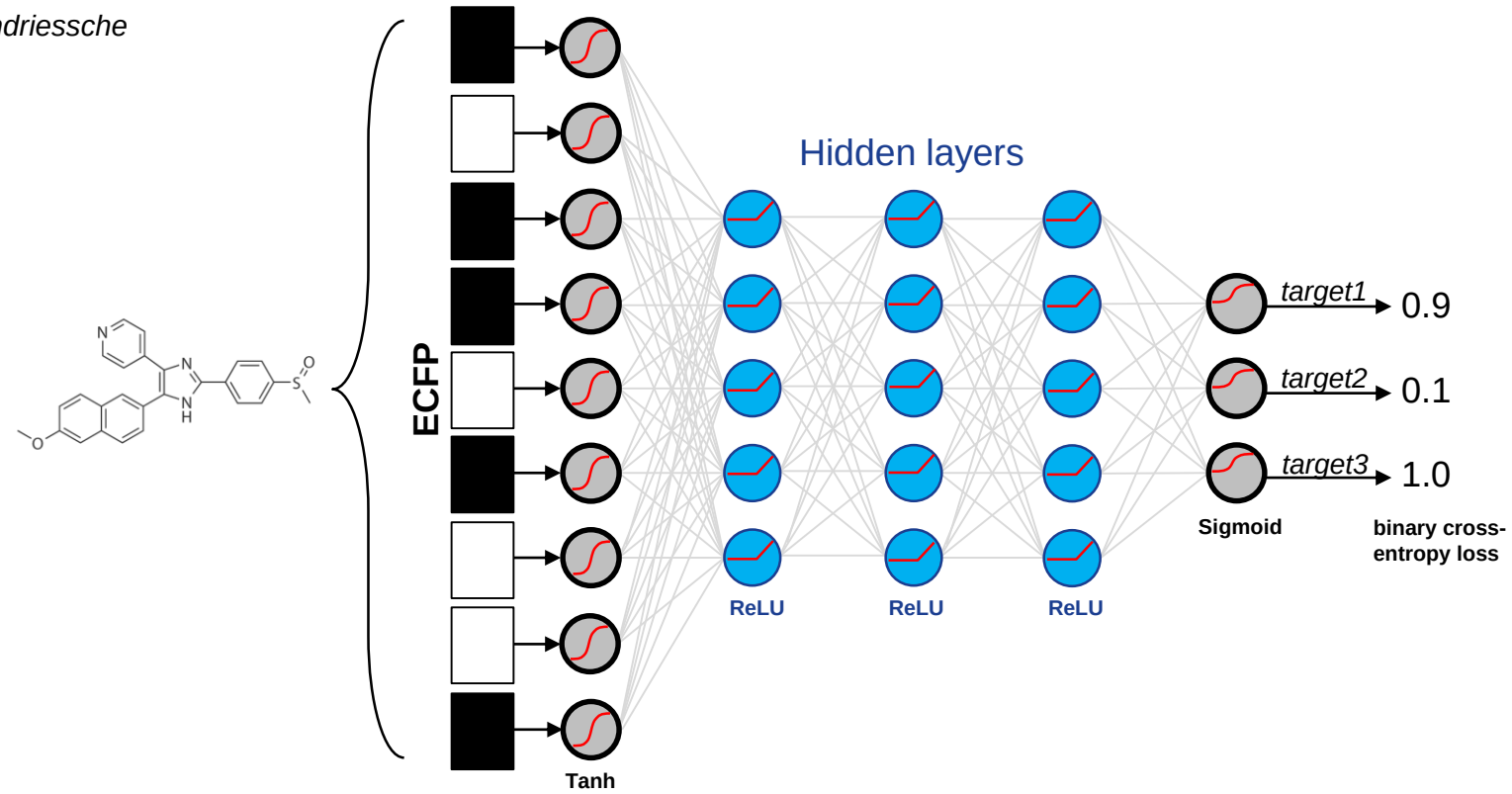


https://github.com/ExaScience/smurff

Deep Learning algorithm: *ExNET*

Andreas Mayr

Yves Vandriessche



Overall Workflow

1. Dataset collection
2. Dataset split
3. Descriptors
4. Performance metrics
5. Algorithms
6. Hyperparameter selection
7. Retrospective model evaluation
8. Prospective model evaluation

Hyperparameters & algorithms

SMURFF: MACAU	
Parameters	Considered values
Dim latent space D	{8, 16, 32, 64}
Precision obs. TP	{1.0, 5.0, 10.0}
Precision feat. BP	{1.0, 5.0, 10.0}
# Samples	{100, 200, …, 1700}

XGboost	
Parameters	**Considered values**
Objective	binary:logistic
Booster	gbtree
Learning rate	0.05
Scale positive weight	{1, 5, 10}
Nbr estimators	{50, 100, 200}
Maximum depth	{5, 10}

Random Forest	
Parameters	**Considered values**
Max features	auto
Class weight	balanced
Nbr estimators	{125, 256, 512}

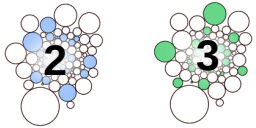
SVM: LIBLINEAR	
Parameters	**Considered values**
kernel	linear
C	{ 300, 100, 30,10, 1, 0.1, 0.05, 0.01, 0.001 }
penalty_loss_dual	(“l2”, ”squared_hinge”, *True*) (“l2”, ”squared_hinge”, *False*) (“l2”, ”hinge”, *True*) (“l1”, ”squared_hinge”, *True*)
Sampling	*none*
Feature scaling	MaxAbsScaler

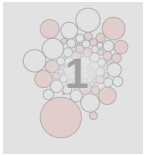
ExNET	
Parameters	**Considered values**
Architecture	[1024x1024x1024], [2048x2048], [2048x2048x2048], [4096x4096x4096]
Learning Rate	{0.1, 0.01}
Input dropout	{0.0, 0.2}
Dropout	0.5
Momentum	{0.0, 0.4}
Activation Fnct	ReLU
Minibatch size	64

Model evaluation workflow: *Nested Cross-Validation*

Model evaluation workflow: *Nested Cross-Validation*

INNER FOLDS
Hyperparameter search

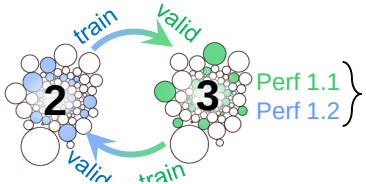




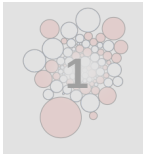
1. Leave one out

Model evaluation workflow: *Nested Cross-Validation*

INNER FOLDS
Hyperparameter search

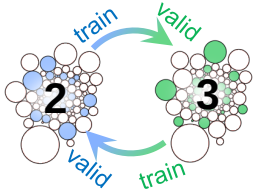


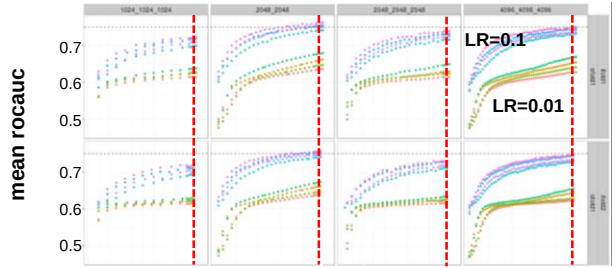
HP selection
best mean



1. Leave one out
2. Select hyperparam.

Hyperparameter Selection: *ExNET*

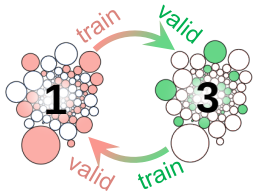


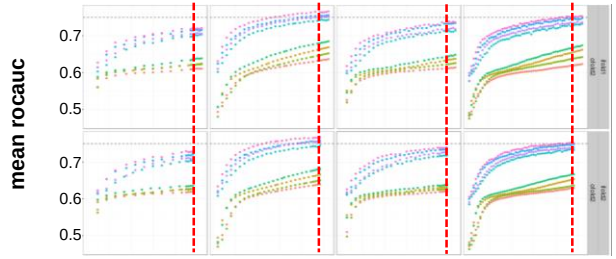


ofold	model	AUC
1	2048_2048-0_1-0_5-0_8-0_4	0.744942
1	4096_4096_4096-0_1-0_5-0_8-0_4	0.733502
1	2048_2048-0_1-0_5-0_8-0_0	0.731866

HP selection **best mean**

Fold-1

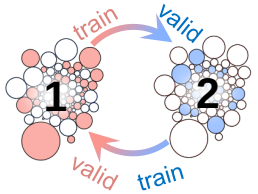


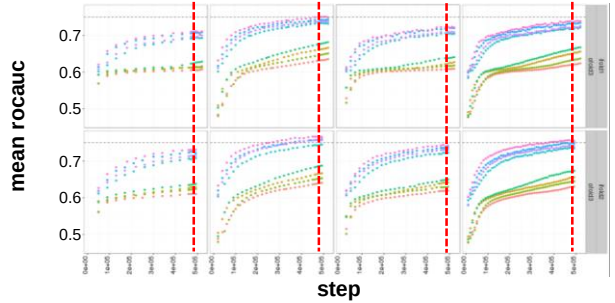


ofold	model	AUC
2	2048_2048-0_1-0_5-0_8-0_4	0.747402
2	4096_4096_4096-0_1-0_5-0_8-0_4	0.736152
2	2048_2048-0_1-0_5-0_8-0_0	0.735000

HP selection **best mean**

Fold-2





ofold	model	AUC
3	2048_2048-0_1-0_5-0_8-0_4	0.742804
3	4096_4096_4096-0_1-0_5-0_8-0_4	0.731517
3	2048_2048-0_1-0_5-0_8-0_0	0.730214

HP selection **best mean**

Fold-3

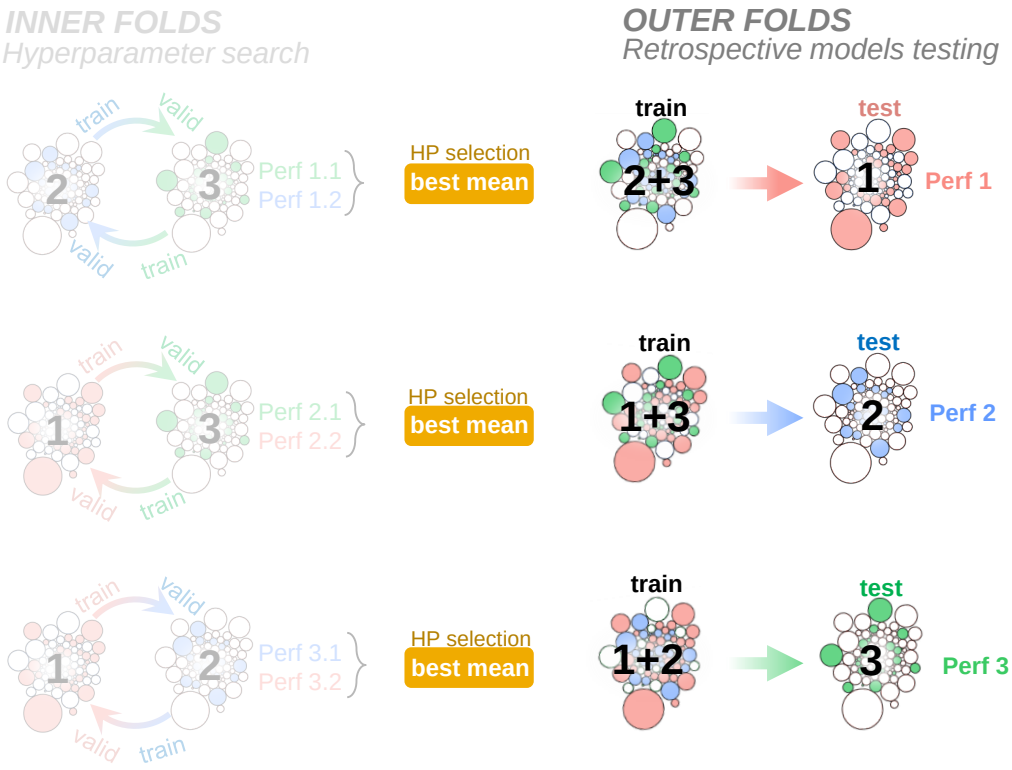
Last checkpoint in common (480k steps)

hyperSet
0_01-0_5-0_8-0_0 0_01-0_5-1_0-0_0 0_1-0_5-0_8-0_0 0_1-0_5-1_0-0_0
0_01-0_5-0_8-0_4 0_01-0_5-1_0-0_4 0_1-0_5-0_8-0_4 0_1-0_5-1_0-0_4

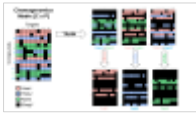
Overall Workflow

1. Dataset collection
2. Dataset split
3. Descriptors
4. Performance metrics
5. Algorithms
6. Hyperparameter selection
7. Retrospective model evaluation
8. Prospective model evaluation

Model evaluation workflow: *Nested Cross-Validation*

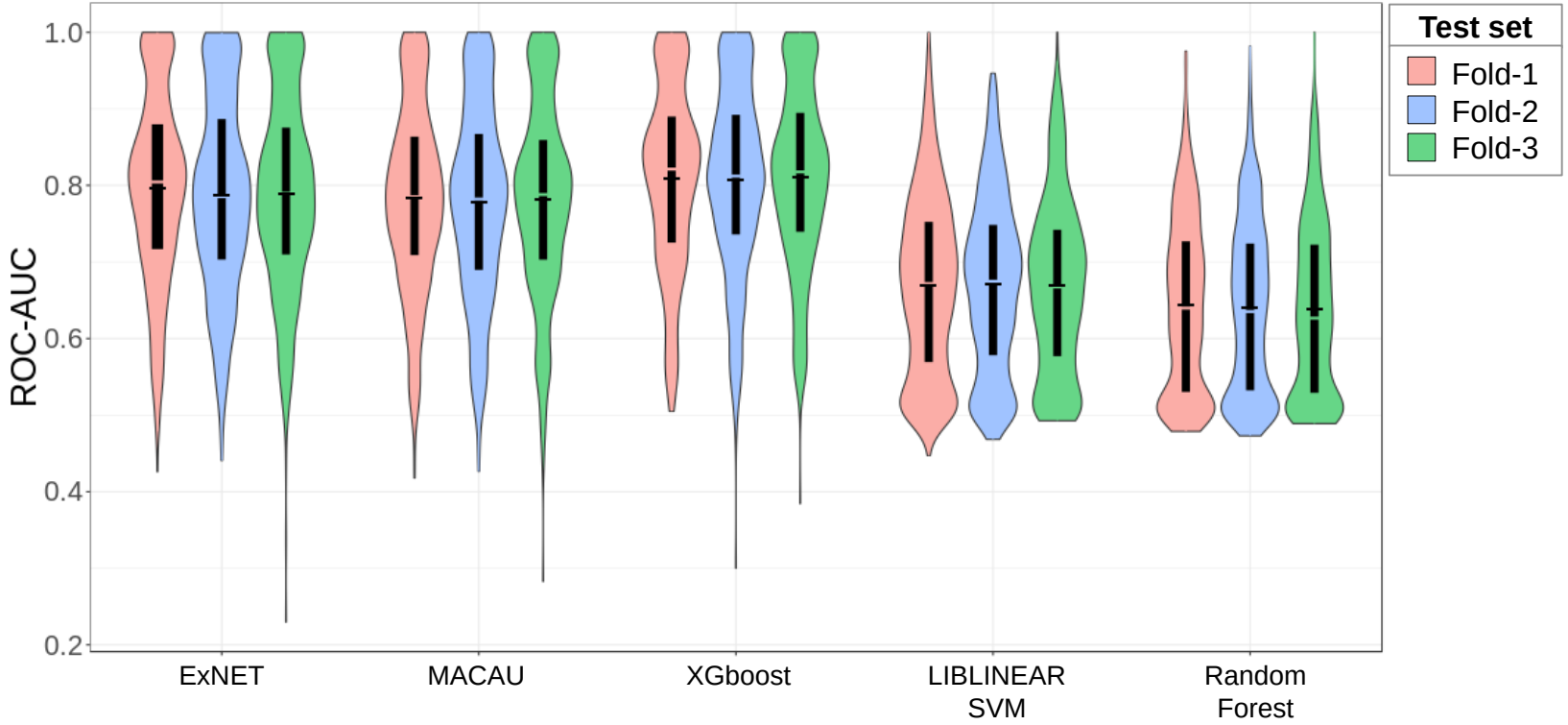


1. Leave one out
2. Select hyperparam.
3. Test models

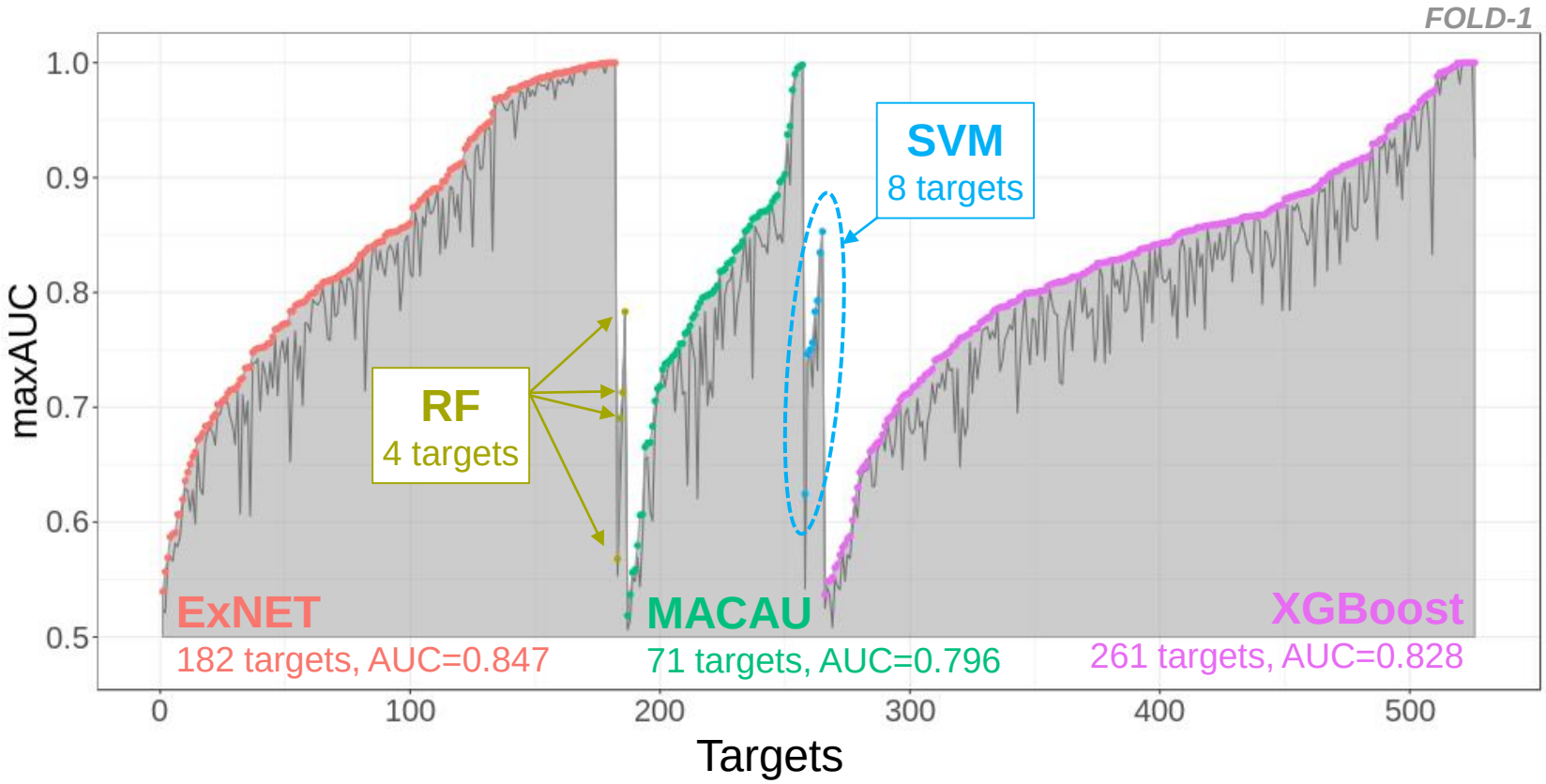


Test performances

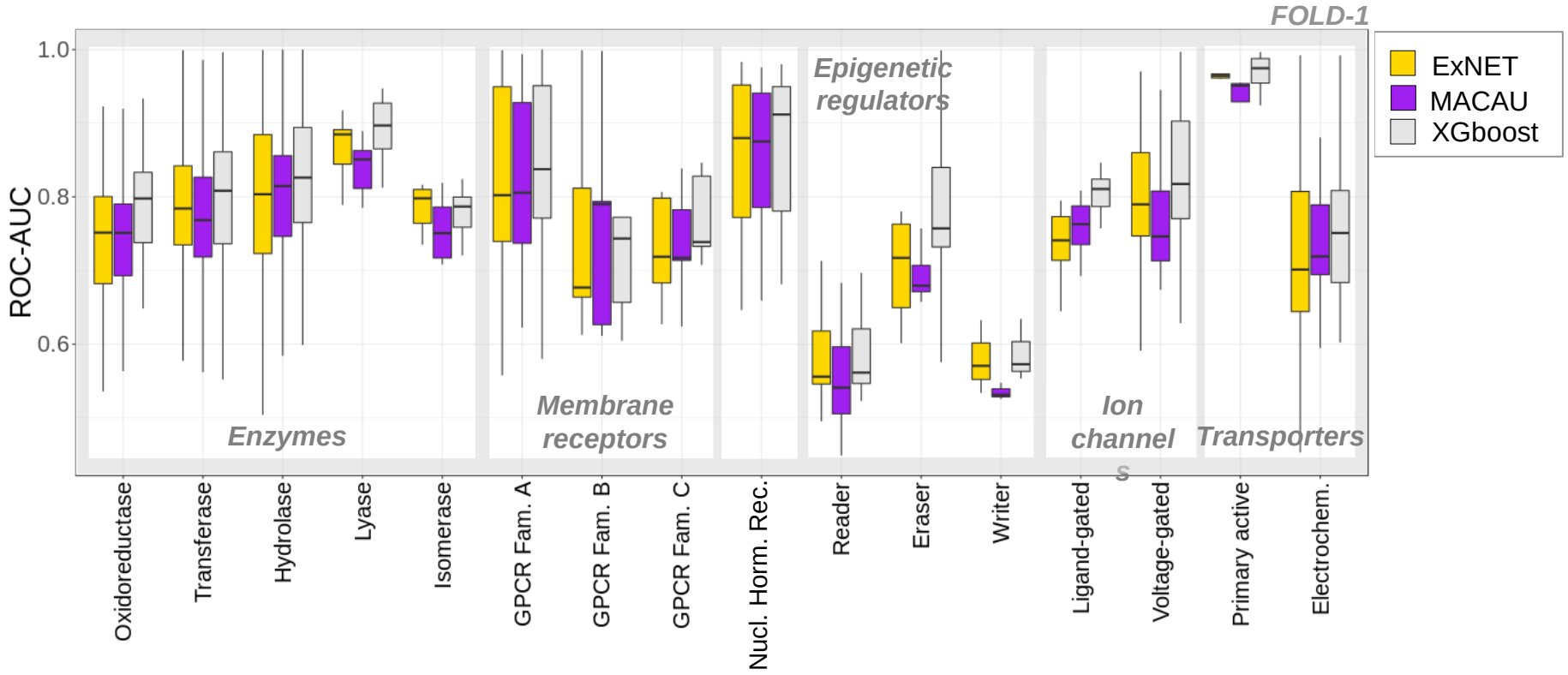
Mean ROC-AUCs: 0.7962 ± 0.004 | 0.7828 ± 0.002 | 0.8086 ± 0.002 | 0.6700 ± 0.001 | 0.6411 ± 0.003



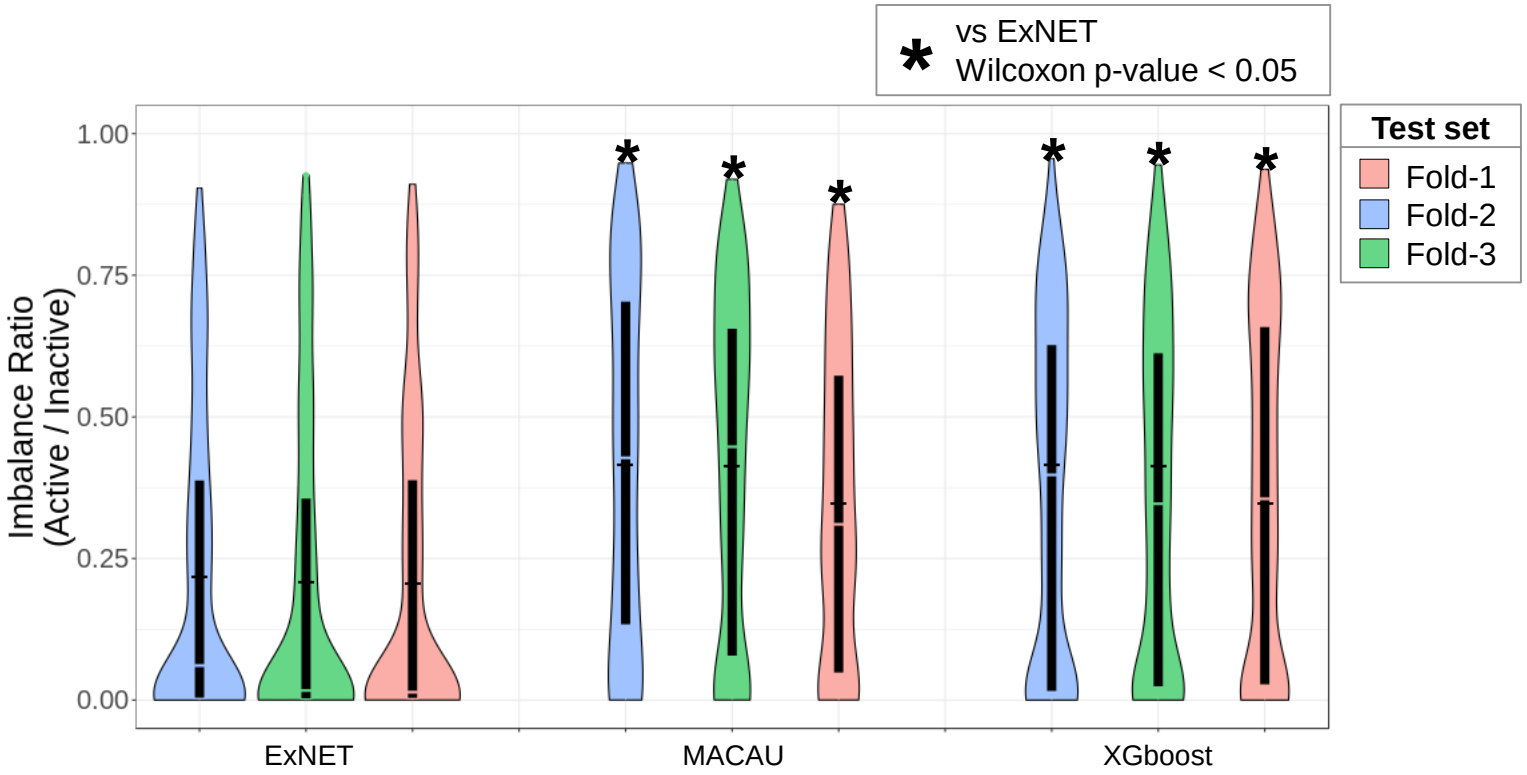
Winning algorithms



Performance per protein family



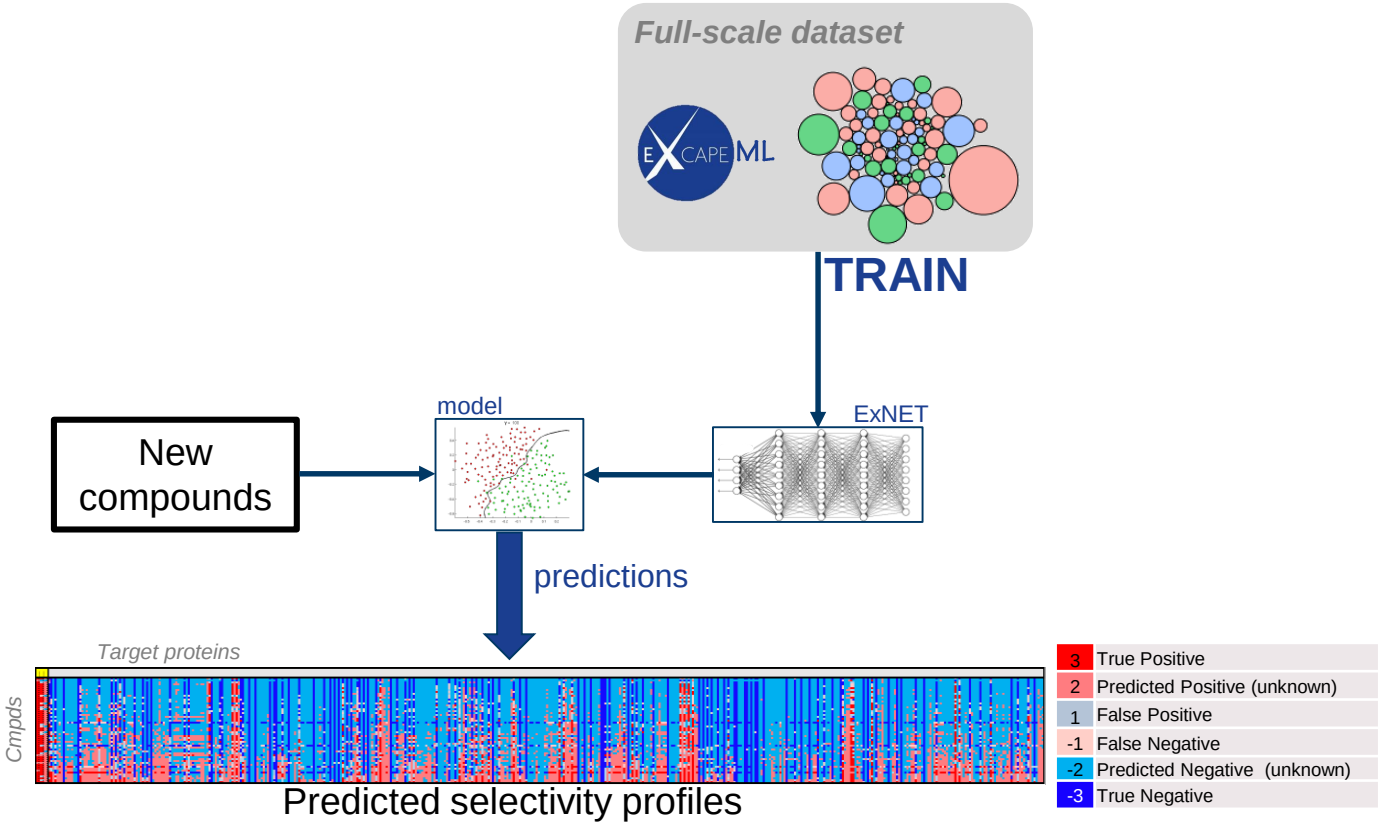
Imbalance Ratio of Target Datasets by Winning Algo.



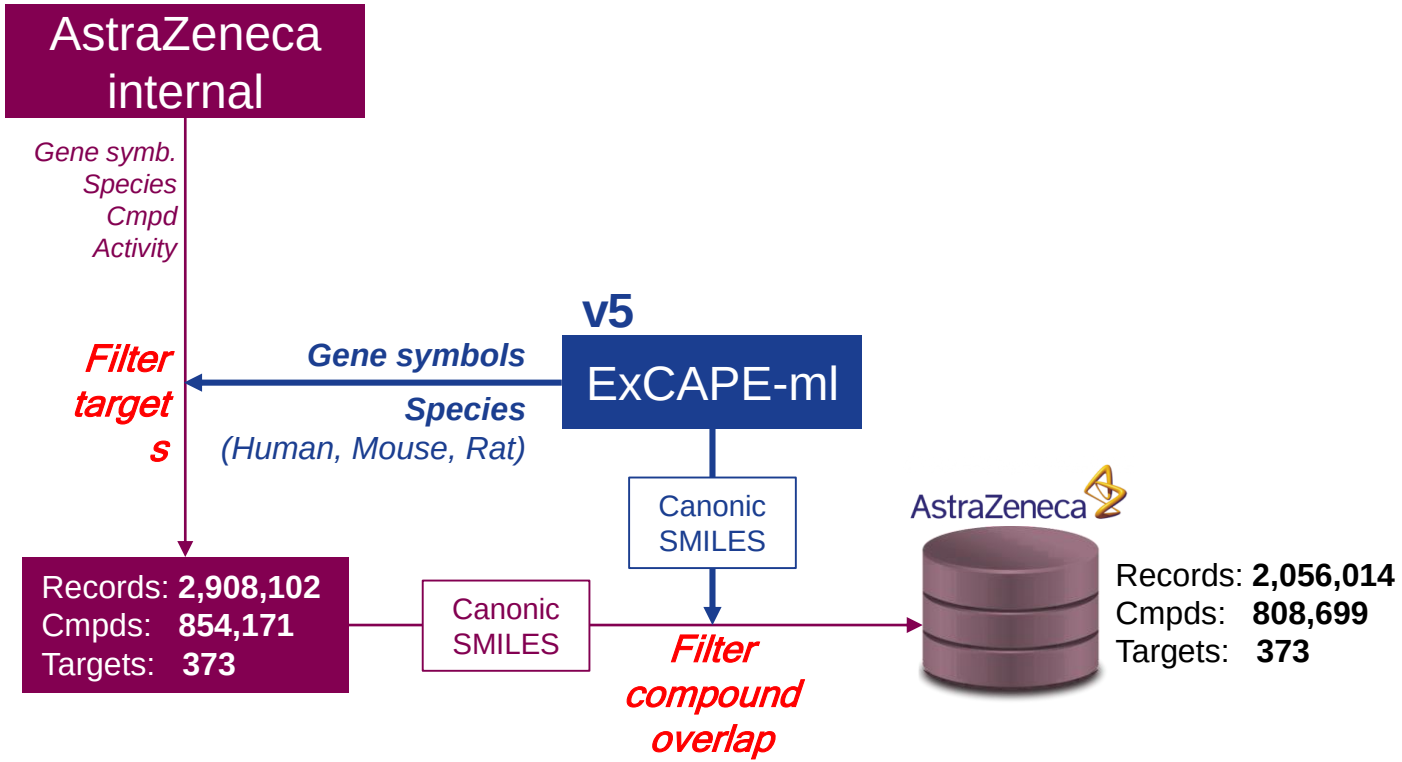
Overall Workflow

1. Dataset collection
2. Dataset split
3. Descriptors
4. Performance metrics
5. Algorithms
6. Hyperparameter selection
7. Retrospective model evaluation
8. Prospective model evaluation

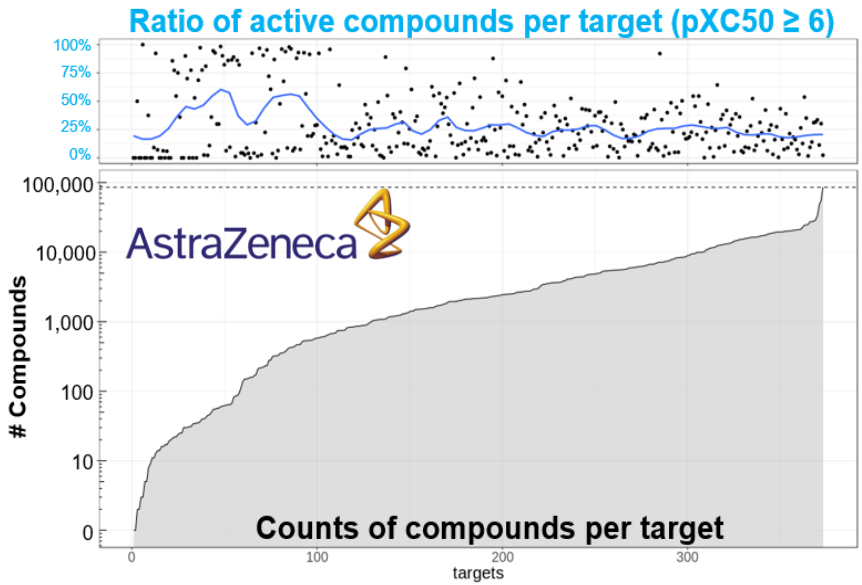
Prospective predictions



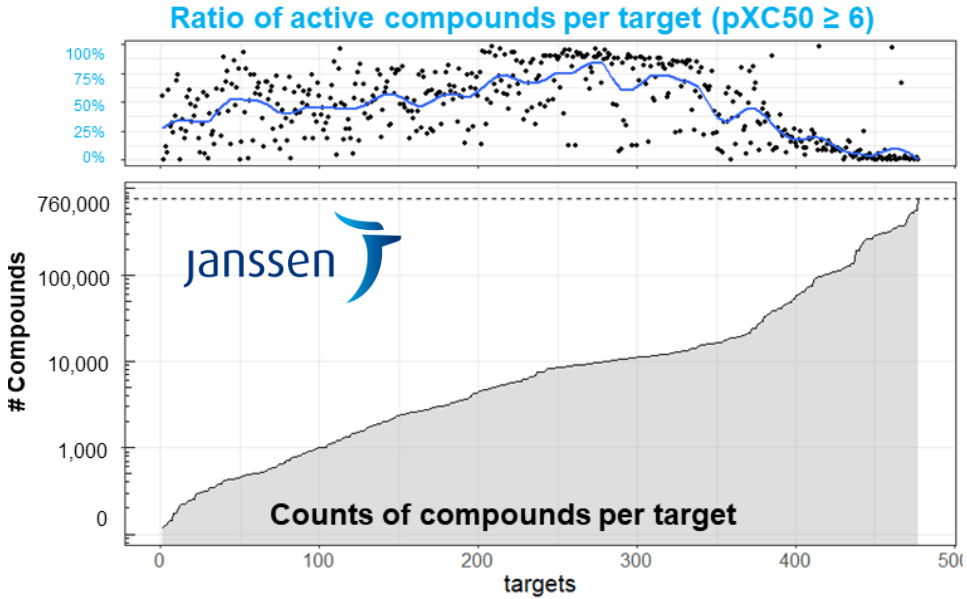
Dataset preparation: *AZ*



Industrial dataset profiles

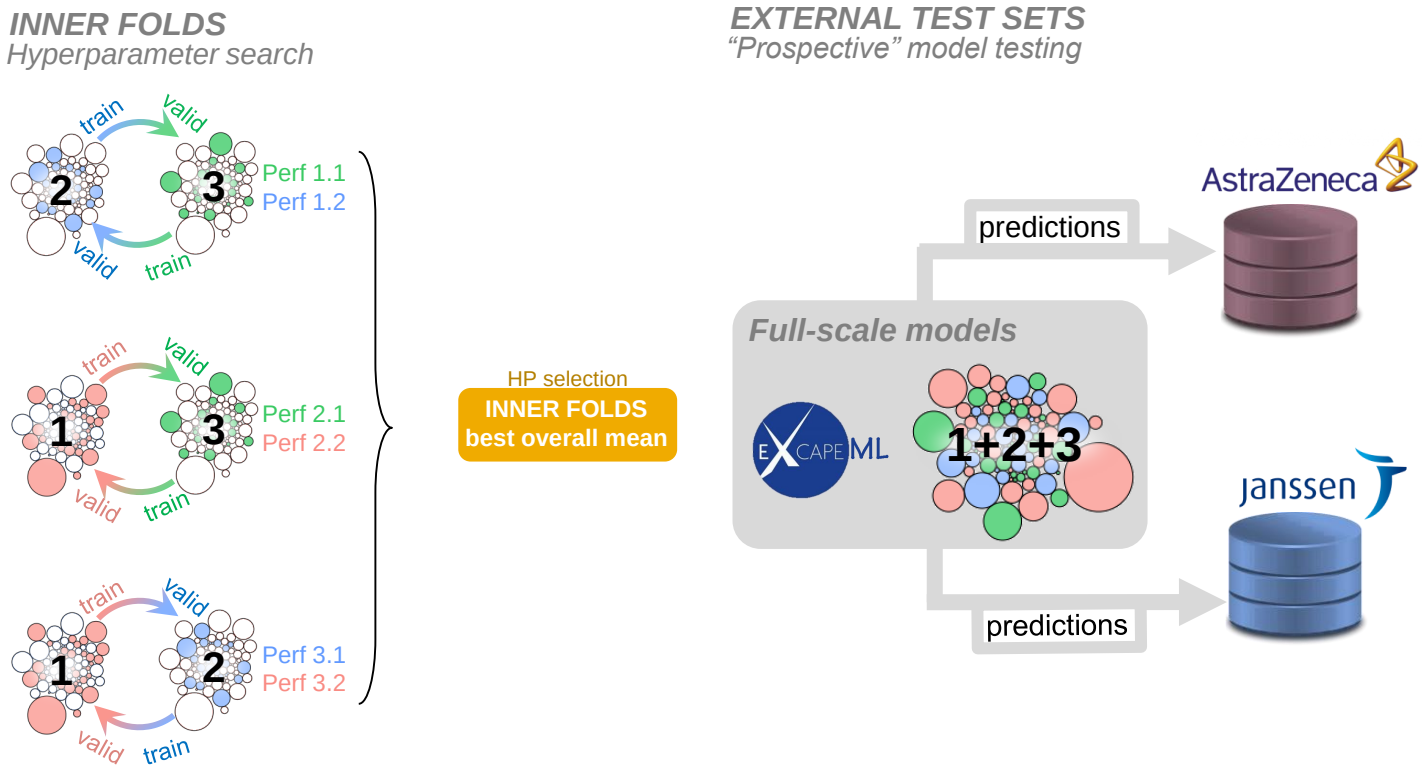


~800K compounds
373 targets

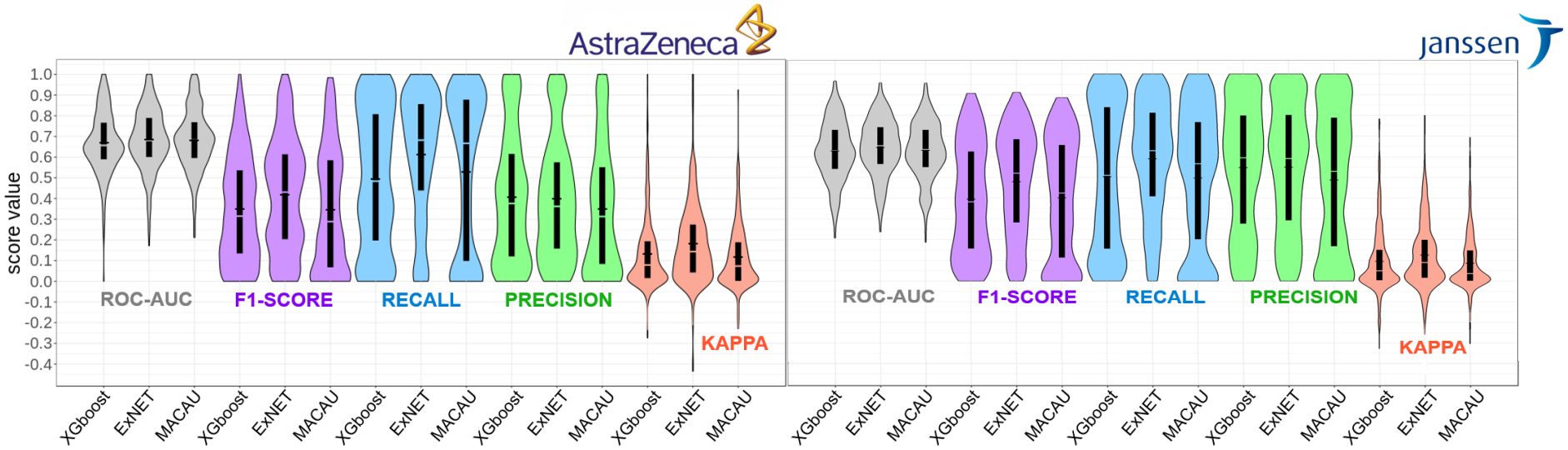


~1.7M compounds
467 targets

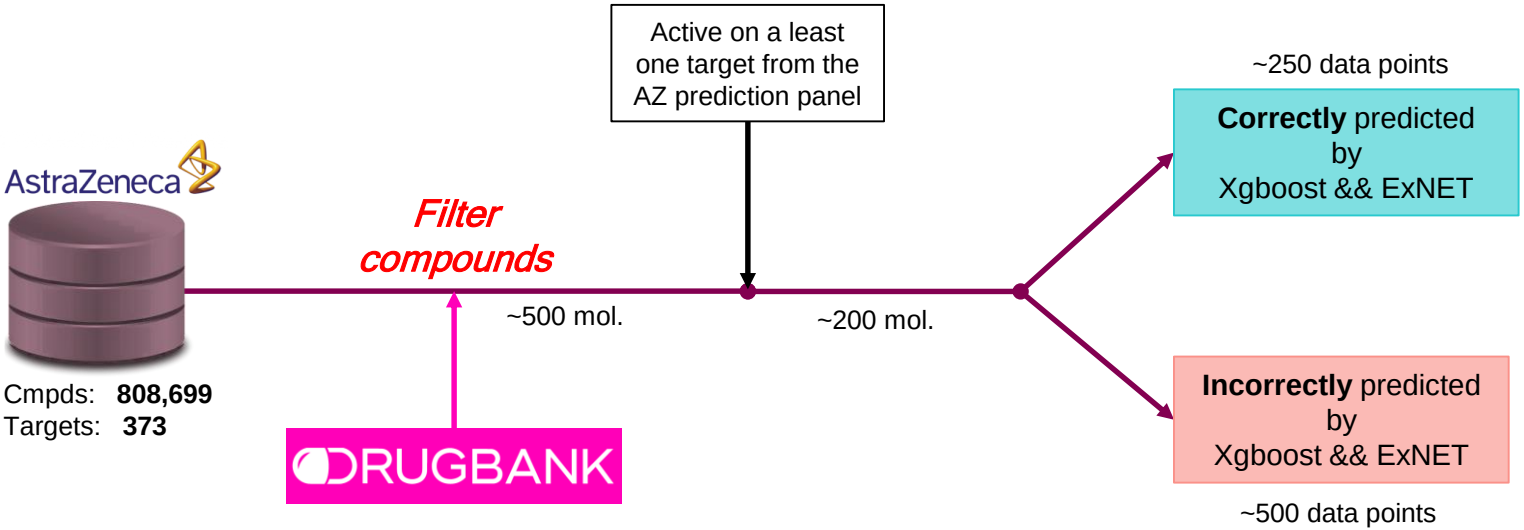
Application of Models to Industrial Datasets



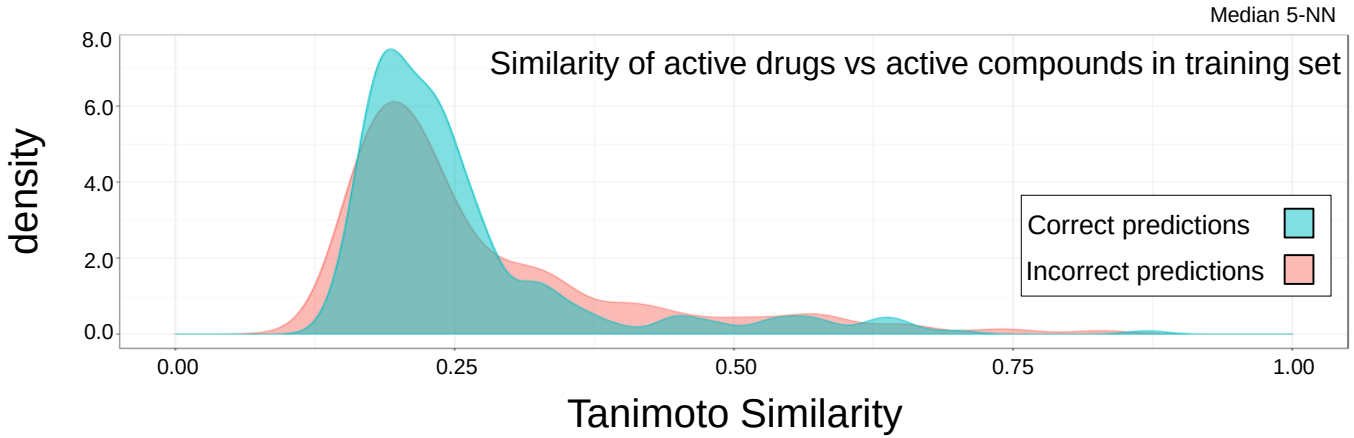
ExCAPE-ML full-scale model performance



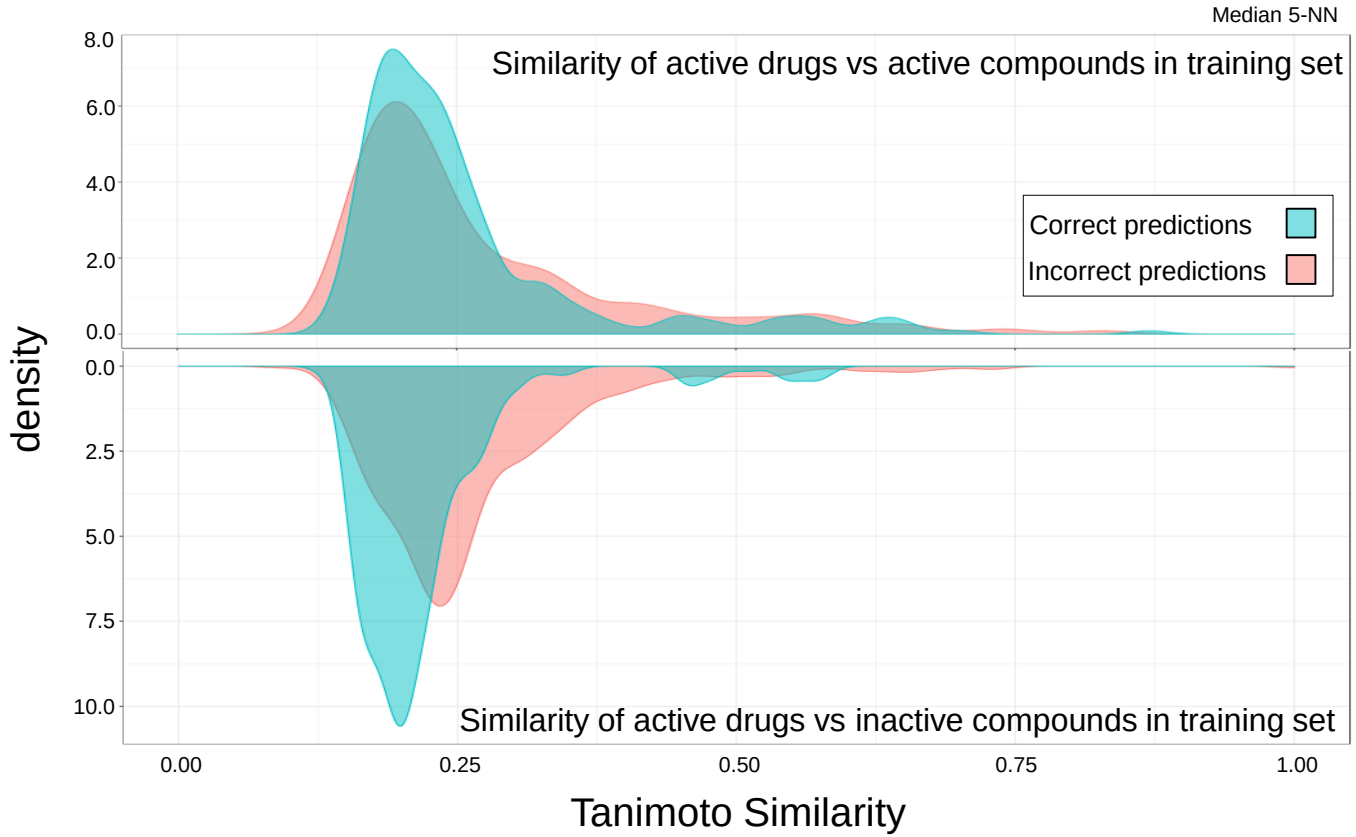
Investigating missed predictions…



Investigating missed predictions…



Investigating missed predictions…



Acknowlegments

ExCAPE consortium

AstraZeneca
Jiangming Sun
Hongming Chen
Lars A Carlsson
Ernst Ahlberg Helgee
Ola Engkvist

janssen PHARMACEUTICAL COMPANIES OF Johnson&Johnson
Thanh Le Van
Felipe Golib
Vladimir Chupakhin

IDEA consult
Nina Jeliazkova

intel
Yves Vandriessche

IT4I
Vojtech Cima

imec
Tom Vander
Tom Ashby

JKU JOHANNES KEPLER UNIVERSITÄT LINZ
Andreas Mayr
Günter Klambauer

ROYAL HOLLOWAY UNIVERSITY OF LONDON
Paolo Toccaceli

Aalto-yliopisto
Xiangju Qin
Samuel Kaski