Statistical approach enabling technology-specific assay interference prediction from large screening data sets

30 April 2024

Vincenzo Palmacci

Problem statement

-Thanks to their sensitivity and efficiency, **fluorescence-based assays are the most widely employed technology** for the high-throughput-screening (HTS) of compounds [1, 2].

-Despite the technical advantages brought to the field, fluorescence-based assays result in a significant number of **false positive readouts caused by assay interference** [3].

- If false readouts remain undetected, they may **trigger costly follow-up studies** that may eventually turn out as futile.

1. Macarron R. et al. *Impact of high-throughput screening in biomedical research*. Nat. Rev. Drug Discovery
2. https://atrandi.com/blog/fluorescence-assays-dominate-life-science-research
3. Sink R. et al. *False positives in the early stages of drug discovery.* Curr. Med Chem

The dominant readout: Fluorescence

“Fluorescence-based detection is the most used detection method in HTS: it is highly sensitive and has good signal-to-noise ratios.
Fluorescence assays don't require special setups and can be miniaturized, which makes them ideal tools for screening applications.”[3]

3. https://atrandi.com/blog/fluorescence-assays-dominate-life-science-research

Fields of application

-HITS TRIAGING:

The practice of selecting a compound series with a promising efficacy profile that meets basic safety requirements and to justify investment in its optimization [4].

-NEGATIVE DESIGN:

Battery of methods that are usually employed to eliminate molecules with undesired properties [5].

4. Vincent F. et al. *Hit Triage and Validation in Phenotypic Screening: Considerations and Strategies.* Cell Chem Bio
5. Yang Z. et al. *Application of Negative Design To Design a More Desirable Virtual Screening Library.* J. Med Chem

Agenda

PART 1. (~5 mins): **Introduction**

Introduction to miniaturized fluorescent assays in high-throughput-screening

Overview of the main mechanisms of assay interference

PART 2. (~5 mins): **State-of-art**

Addressing assay interference: experimental and in-silico methodologies

Addressing assay interference: pitfalls of existing approaches

PART 3. (~20 mins): **A new pipeline to predict assay interference**

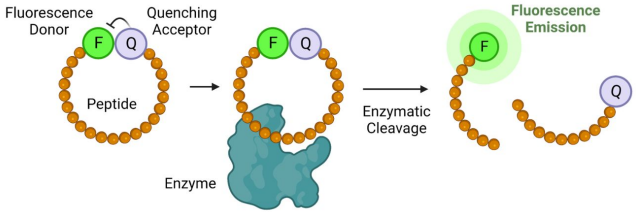
Building a comprehensive dataset from primary HTS screenings

Identifying interfering compounds through analysis of compound activity rates

Prediction of compounds likely to interfere with the assay technology

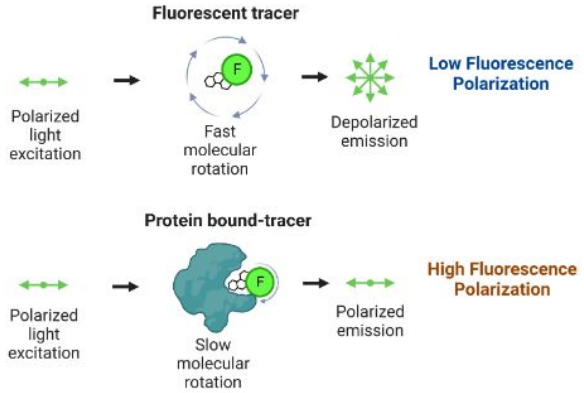
Fluorescence Intensity Assays (FLINT)

Fluorogenic assays



- Convenient for screening enzymatic inhibitors
- Fluorescent emission upon enzymatic cleavage

Fluorescence polarization (FP)



- Detect dynamic interaction between the biological target and the ligand
- Fluorescent emission upon interaction

Credits: https://bpsbioscience.com/

Other popular fluorescence-based assay formats

- Fluorescence Resonance Energy Transfer (FRET)

  Measures the energy transfer between a donor-acceptor pair. For the energy transfer to work donor and acceptor must be in close proximity.

- Time-Resolved FRET (TR-FRET)

  Measures the time a fluorophore spends in the excited state before it reverts to its ground state by emitting a photon (FLT).

Pro of fluorescence-based assays

High specificity.

High sensitivity with low background noise.

Simple operation.

Credits: https://bpsbioscience.com/

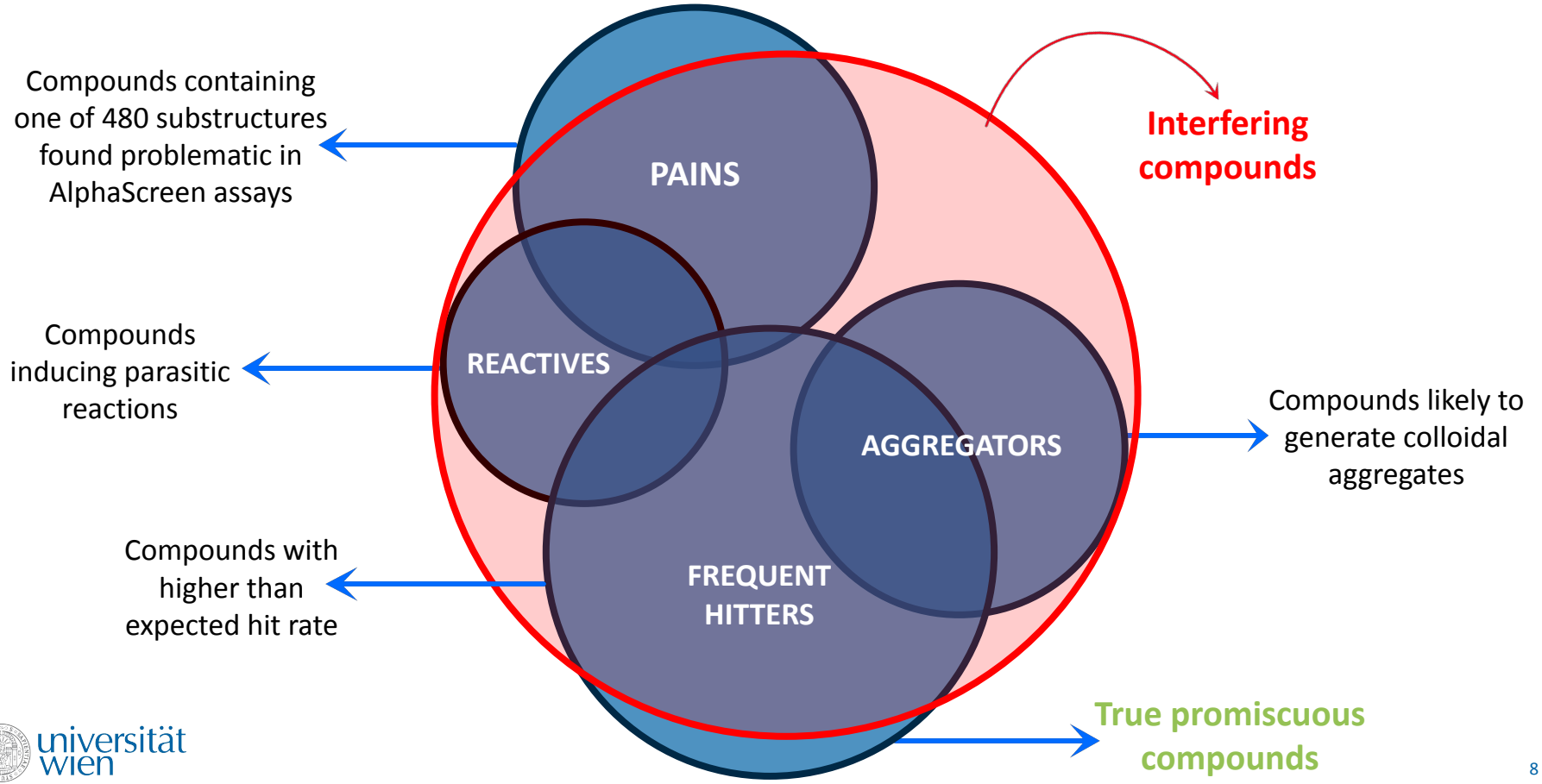
No one’s safe: false positive readouts

“Many hits are artefacts - their activity does not depend on a specific, drug-like interaction between molecule and protein. Artefacts have subversive reactivity that masquerades as drug-like binding and yields false signals across a variety of assays.”[6]

universität wien

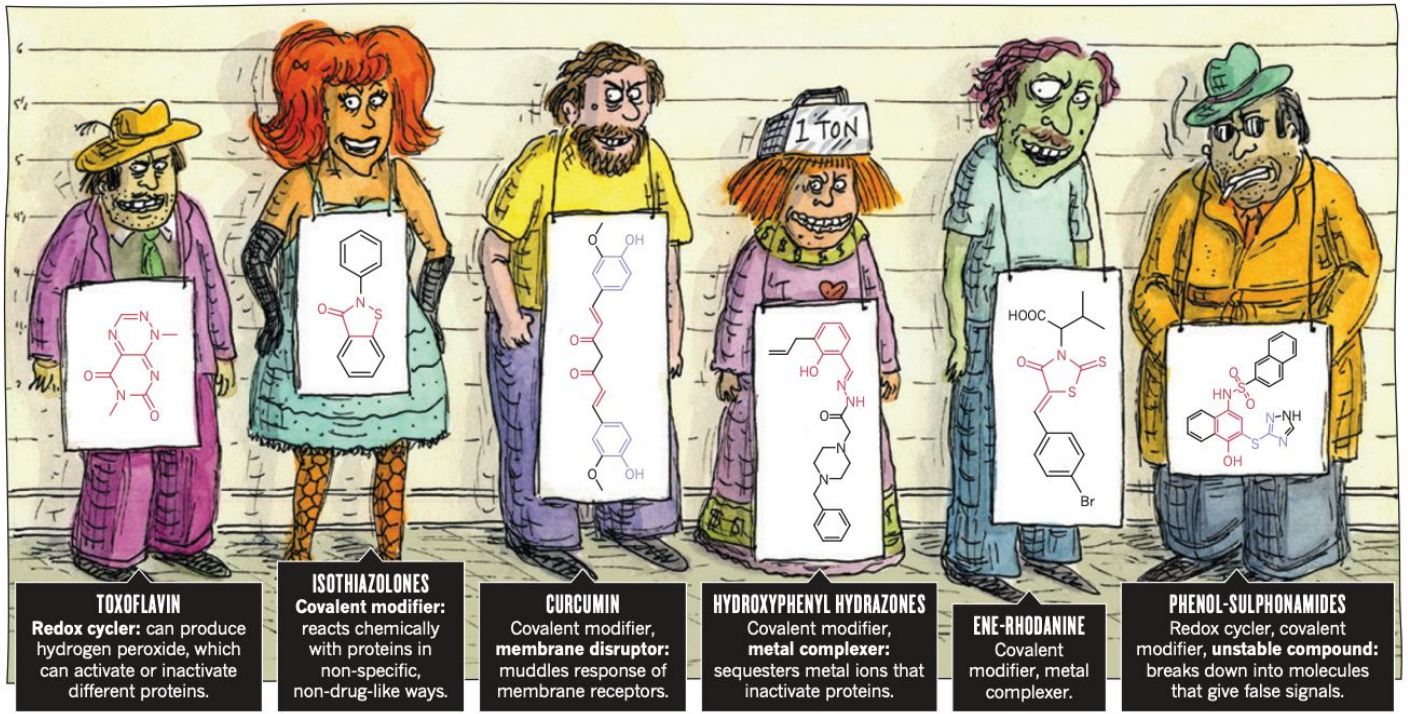
6. Baell J. et al. *Chemical con artists foil drug* discovery. Nature

Interfering compounds and interference mechanisms



universität wien

Interfering compounds and interference mechanisms



6. Baell J. et al. *Chemical con artists foil drug* discovery. Nature

Dealing with assay interference: prevention measures and hit-triaging

Experimental countermeasures

- Screening with non-ionic detergents to prevent compounds aggregation
- Use of novel fluorophores emitting in a different region of the spectrum
- Use orthogonal assays to confirm the primary hits
- Implementation of counter-screen assays to identify interfering compounds

In-silico methodologies

Global methods:

HitDexter3, Pan-Assay interference compounds (PAINS)*

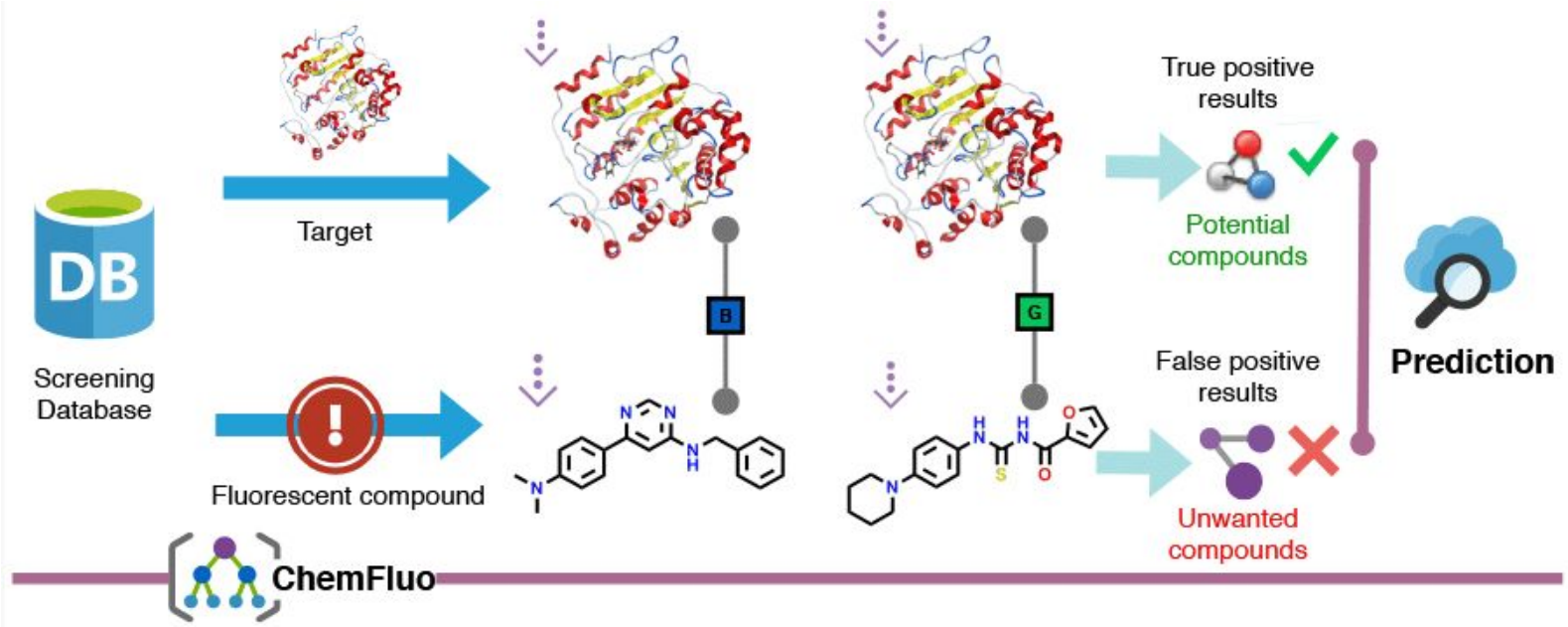
Specialized methods:

InterPred, ChemFluo, AZ (TR-)FRET interference classifiers

universität wien

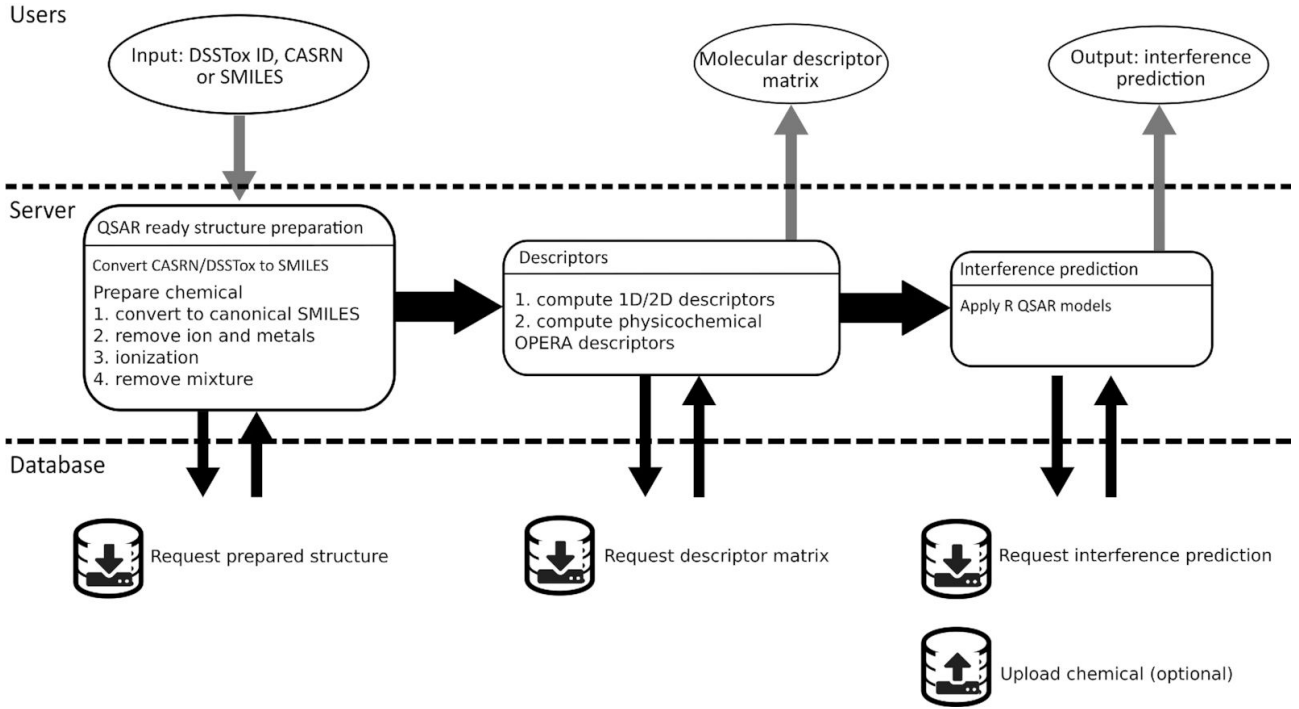
Why we do *in-silico* modelling? To avoid running expensive experiments!

Specialized methods: ChemFluo



7. Yang Z. et al. *ChemFLuo: a web-server for structure analysis and identification of fluorescent compounds.* Briefings in Bioinformatics

Specialized methods: InterPred



8. Borrel A. et al. *InterPred: a webtool to predict chemical autofluorescence and luminescence interference.* Nucleic Acids Research.

Overview of existing methods

ChemFluo

- Two Extreme Gradient Boosting (XGBoost) ensembles
- Prediction of autofluorescence in **blue and green channels**
- Trained and tested on **counter-screen data**
- Underwhelming performances on external test set (**MCC=0.34**) using the **same experimental evidence used in training**

AZ (TR-)FRET interference classifiers

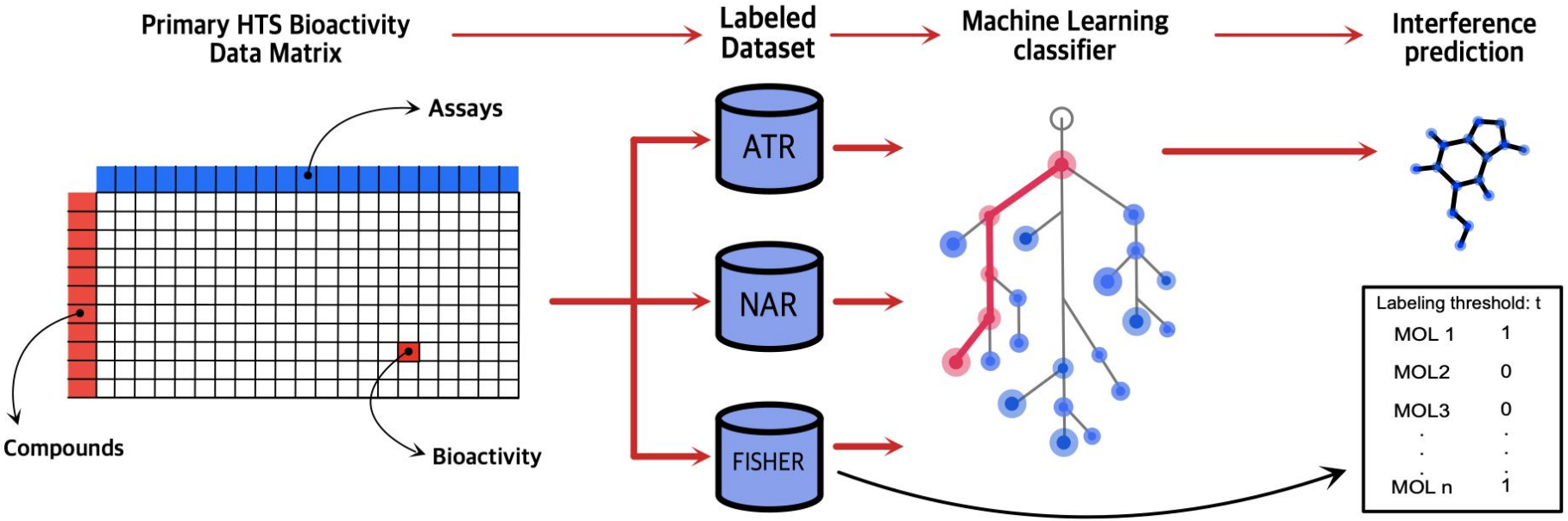
- Random Forest Classifier (RFC)
- Prediction of i**nterference in AlphaScreen, FRET and TR-FRET** assays
- Trained and tested on **counter-screen data** from AstraZeneca
- Underwhelming performances on public test dataset (**MCC=0.20**)

InterPred

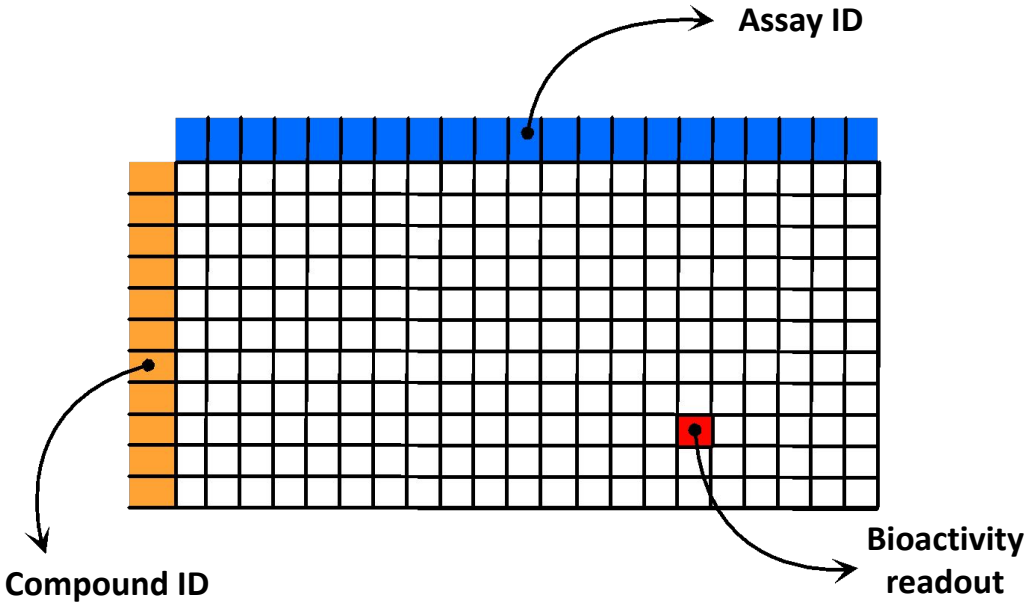
- 13 Random Forest Classifiers (RFCs)
- Prediction of autofluorescence in the **blue, green and red channels**
- Trained validated and tested on random splits of the same dataset
- **Uses counter-screen data produced ad-hoc**
- Each is model specific for one assay
- **Not possible to assess their performances** due to testing strategy used

A new methodology to identify compounds interfering with fluorescent assays

Overview


Primary HTS Bioactivity Data Matrix
Labeled Dataset
Machine Learning classifier
Interference prediction
Assays
Compounds
Bioactivity
ATR
NAR
FISHER
Labeling threshold: t
MOL 1 1
MOL2 0
MOL3 0
MOL n 1

Data collection: Bayer AG HTS historical data



Dataset composition

- **Compounds:**

  More than 5 millions compounds represented by canonical SMILES strings.

- **Assays:**

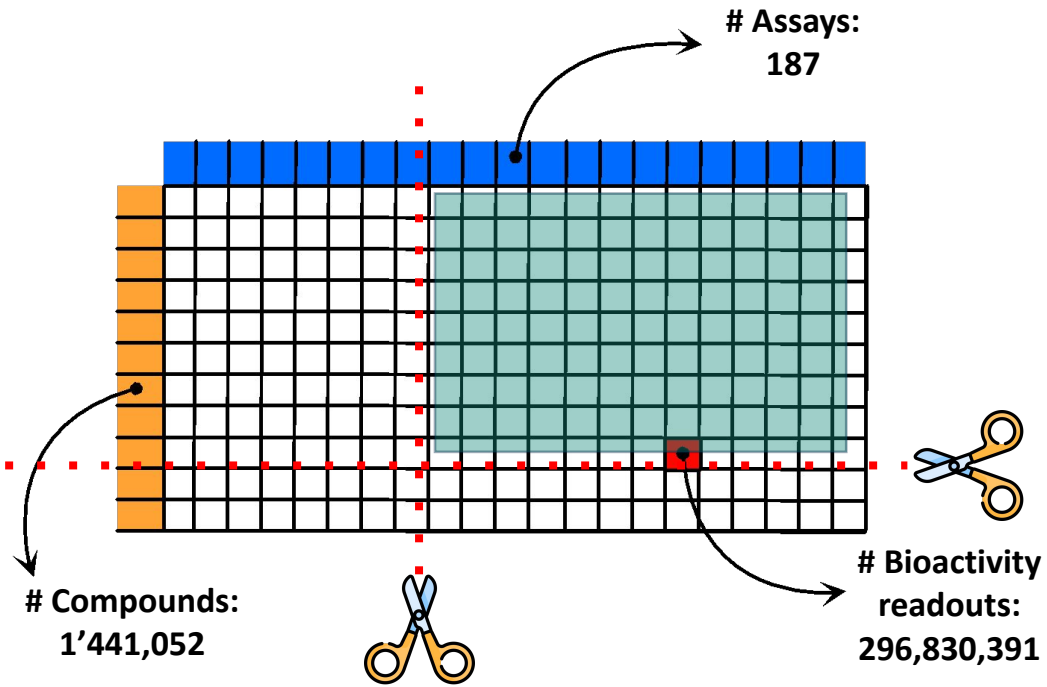
  500 different assays (different technology, biological target).

- **Readouts:**

  Preprocessed bioactivity readouts received from experimentalists as Z-scores

To my knowledge, this represent the largest HTS primary screening dataset available both in public and private domain.

Data preprocessing: Bayer AG HTS historical data



Preprocessing pipeline

1. Assays must have bioactivity recorded for at least 80% of the compounds

2. Compounds must have bioactivity recorded for at least 80% of the assays

Dataset composition after step 1 and 2:

205 assays, 1’488’407 compounds

3. Assays must be annotated

4. Compounds must be unique (SMILES strings matching)

5. Binarize Z-scores following experimentalist indications

Note: Readouts were available for both main and background signal

Data preprocessing: Bayer AG HTS historical data

Dataset split prior to modelling

1. Compute Murcko scaffolds

2. Group molecules sharing the same scaffold

3. Random assignment of grouped molecules to training (80%), validation (10%) and test (10%) set.

Splits sizes:

Train: 1’130’711

Validation: 135’ 830

Test: 174’493

	# of compounds	# of Murcko scaffolds
Training set	1,130,711	197,095
Validation set	135,830	48,559
Test set	174,493	58,932

Dataset characterization: Bayer AG HTS historical data

Assays space:

FLINT: 56 Blue, 23 Green, 8 Red

FRET: 16

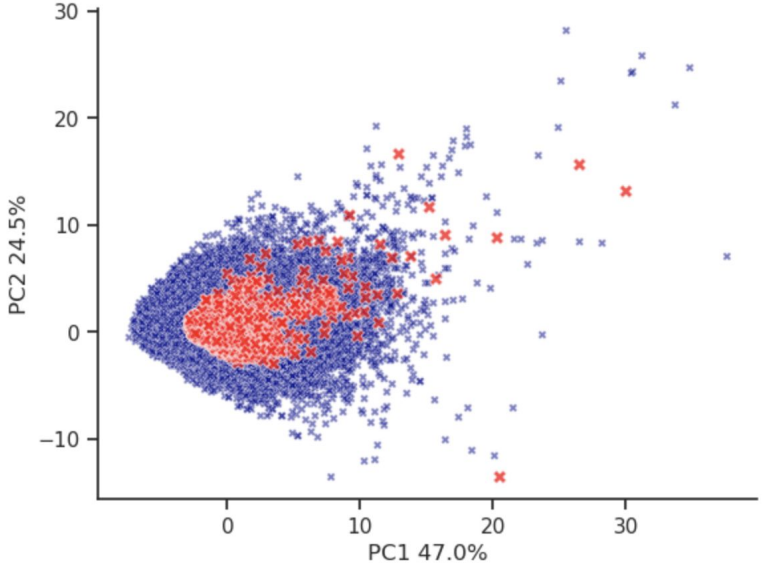
TR-FRET: 10

Bioluminescence: 76

Cell-based: 88

Biochemical: 99

Chemical space



PCA comparing the training set chemical space (BLUE) and the DrugBank approved drugs space (RED)

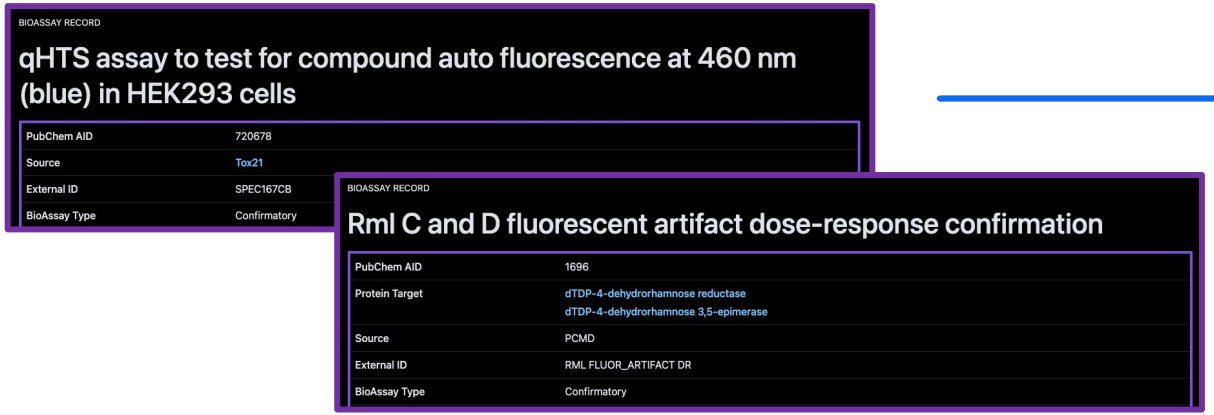
universität wien

Data collection and preprocessing: PubChem test set

Table S1. The information of the external validation sets.	
Set	**AID**
Blue fluorescent external validation set	AID 1696, AID 720678, AID 720678 and AID 720678

ChemFluo blue fluorescence external validation set.

Data were collected from PubChem Bioassay database as described by the authors*



# of Compounds:

10’691

# of Compounds after preprocessing:

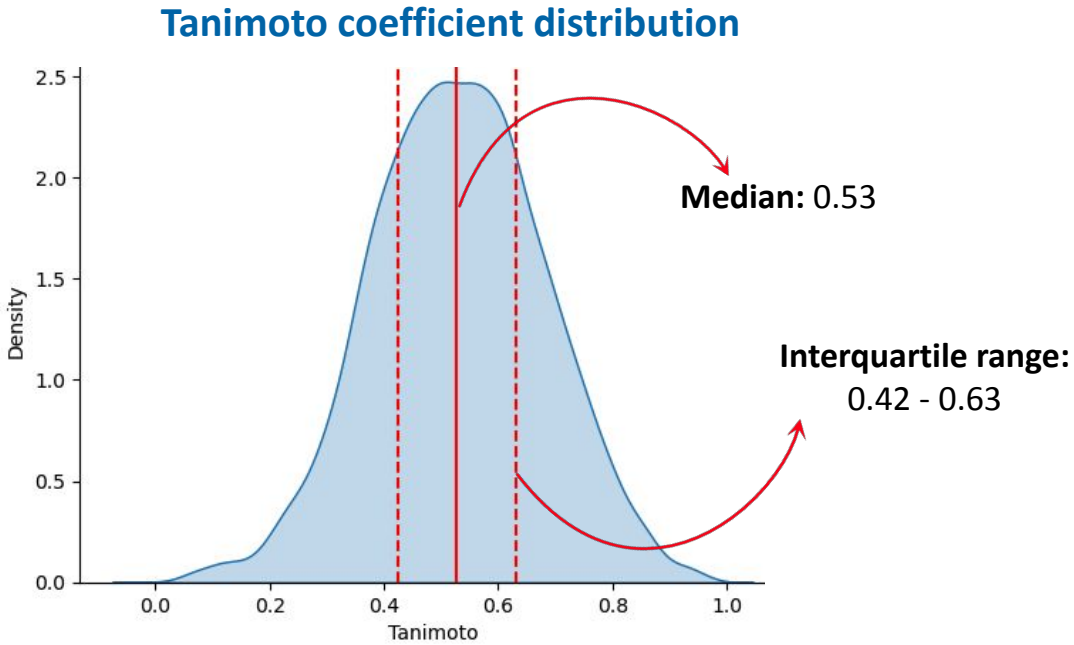
10’031

Preprocessing:

1. Removal of salt moieties
2. Removal of duplicated structures (0)
3. Removal of structures matching with the training set (660)

*As you may notice someone did a mess and wrote 3 times the same assay ID. This show that Supplementary materials should never be overlooked.

Dataset characterization: PubChem test set



Distribution of compounds' maximum Tanimoto coefficient computed between the training and the PubChem test set.

universität wien

Core work: labelling compounds likely to interfere with fluorescence-based assays

Labelling compounds likely to interfere with the assay technology: compute interference metrics

Activity-to-tested ratio (ATR) reloaded

Activity-to-tested ratio is computed as:

$$ATR^i = \frac{\#\ active\ readouts}{\#\ times\ tested}$$

Where i is the i-th compound in the dataset

1. Compute compounds ATR:

- ATR in fluorescence assays
- ATR in other assays

2. Apply threshold to obtain binary interference labels

Noise-to-active ratio (NAR)

Noise-to-active ratio is computed as:

$$NAR^i = \frac{\#\ active\ back \cap \#\ active\ main}{\#\ times\ tested}$$

Where i is the i-th compound in the dataset

1. Compute compounds NAR considering only fluorescent assays

2. Apply threshold to obtain binary interference labels

Fisher exact test

For the compound contingency table X:

	0	**1**
Fluorescent assays	a	b
Other technologies	c	d

1. Compute compounds p-values applying Fisher –exact test

2. Apply threshold to obtain binary interference labels

universität wien

Labelling compounds likely to interfere with the assay technology: compute binary labels

Thresholds applied to compute binary interference labels

	Percentage of likely interference compounds (thresh)			
	2%	**5%**	**10%**	**20%**
ATR	5.00	3.00	1.00	0.90
NAR	0.10	0.07	0.04	0.03
p-value from Fisher's exact test	0.01	0.07	0.17	0.35

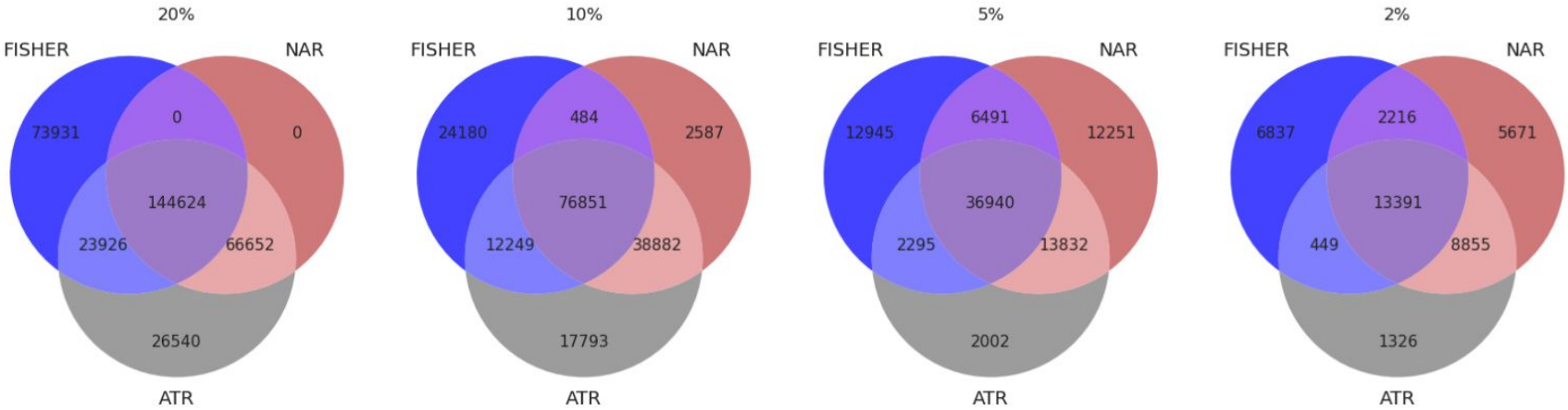
Rationale applied to interference metrics to compute binary labels

$$ATR^i_{Fluo} \geq mean(ATR_{Other}) + thresh * std(ATR_{Other})$$

$$NAR^i \geq thresh$$

$$p^i \leq thresh$$

Labelling compounds likely to interfere with the assay technology: comparison of labelling strategies



20% threshold:

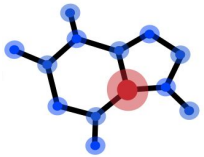
- ATR is a superset of NAR
- Fisher recognize a different set of compounds as interfering

2% threshold:

- Higher relative size of overlap among all methods
- NAR is not a subset of ATR anymore

Development of machine learning classifiers for assay interference prediction

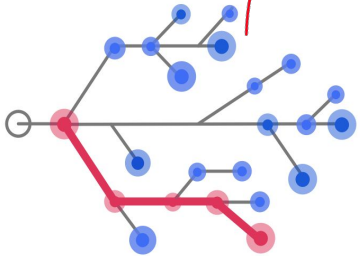
mol



GetMorganFingerprintAsBitVect

Xi
Yi
Pi
Ki
Hi

Morgan2(nbits=2048)

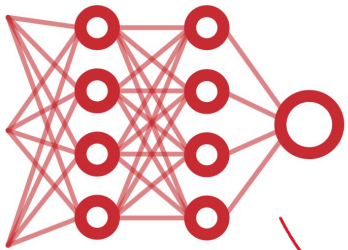


RFC

BalancedRandomForest (imbalanced-learn)
BayesOpt 50 iterations

Hyperparameters optimized:

- n_estimators
- max_depth
- bootstrap



MLP

MLP (PyTorchLightning)

ELU activation function
BinaryCrossEntropyLoss
WeightedRandomSampler

Optuna 50 iterations

Hyperparameters optimized:

- n_layers
- n_units
- dropout
- learning_rate

universität wien

Optimized hyperparameters: RFC

Have fun reading the tables

ATR

	2%	5%	10%	20%
n_estimators	1928	228	1633	3900
max_depth	36	40	48	48
bootstrap	False	True	False	True

NAR

	2%	5%	10%	20%
n_estimators	4490	566	3085	4557
max_depth	48	41	46	50
bootstrap	True	False	True	True

FISHER

	2%	5%	10%	20%
n_estimators	471	337	2270	420
max_depth	39	40	62	52
bootstrap	True	False	False	True

Optimized hyperparameters: MLP

Have fun reading the tables

ATR

	2%	5%	10%	20%
n_layers	5	2	5	4
n_units	2034	1025	1894	1385
dropout	0.8	0.5	0.8	0.7
learning_rate	3.4e-4	1.0e-4	5.0e-4	1.4e-4

NAR

	2%	5%	10%	20%
n_layers	5	3	5	5
n_units	1125	1154	1420	1270
dropout	0.6	0.7	0.8	0.3
learning_rate	2.5e-4	1.9e-4	1.4e-4	1.3e-4

FISHER

	2%	5%	10%	20%
n_layers	2	5	3	5
n_units	1249	1801	1881	1835
dropout	0.7	0.7	0.7	0.8
learning_rate	1.9e-4	1.3e-3	1.3e-4	1.0e-4

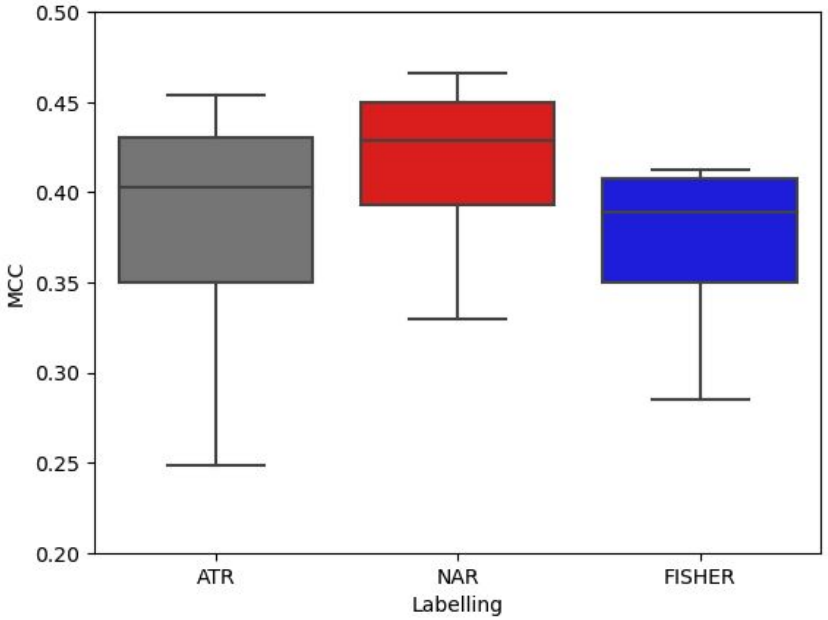
Model performances on the Bayer AG test set

	Labelling method	Threshold	Bayer AG test set			
			MCC	AUC	Precision	Recall
RFC	ATR	20%	0.42	0.87	0.34	0.81
		10%	0.45	0.87	0.40	0.75
		5%	0.38	0.91	0.23	0.81
		2%	0.25	0.94	0.09	0.86
	NAR	20%	0.47	0.85	0.47	0.72
		10%	0.44	0.88	0.36	0.77
		5%	0.41	0.91	0.27	0.80
		2%	0.33	0.94	0.15	0.86
	Fisher	20%	0.41	0.80	0.49	0.61
		10%	0.41	0.86	0.32	0.73
		5%	0.37	0.89	0.23	0.76
		2%	0.29	0.92	0.12	0.80

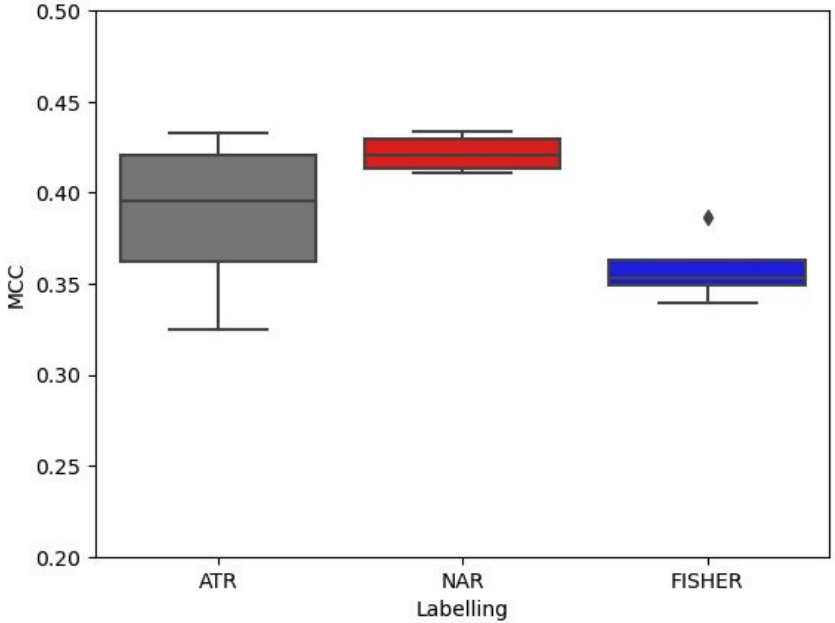
	Labelling method	Threshold	Bayer AG test set			
			MCC	AUC	Precision	Recall
MLP	ATR	20%	0.38	0.84	0.30	0.80
		10%	0.42	0.84	0.37	0.71
		5%	0.43	0.89	0.42	0.50
		2%	0.32	0.91	0.17	0.70
	NAR	20%	0.43	0.82	0.47	0.65
		10%	0.41	0.85	0.34	0.73
		5%	0.43	0.88	0.34	0.67
		2%	0.41	0.90	0.39	0.47
	Fisher	20%	0.34	0.76	0.40	0.65
		10%	0.39	0.82	0.36	0.58
		5%	0.36	0.86	0.23	0.71
		2%	0.35	0.88	0.31	0.44

Model performances on the Bayer AG test set

RFC MCC for different labelling methods



MLP MCC for different labelling methods



universität wien

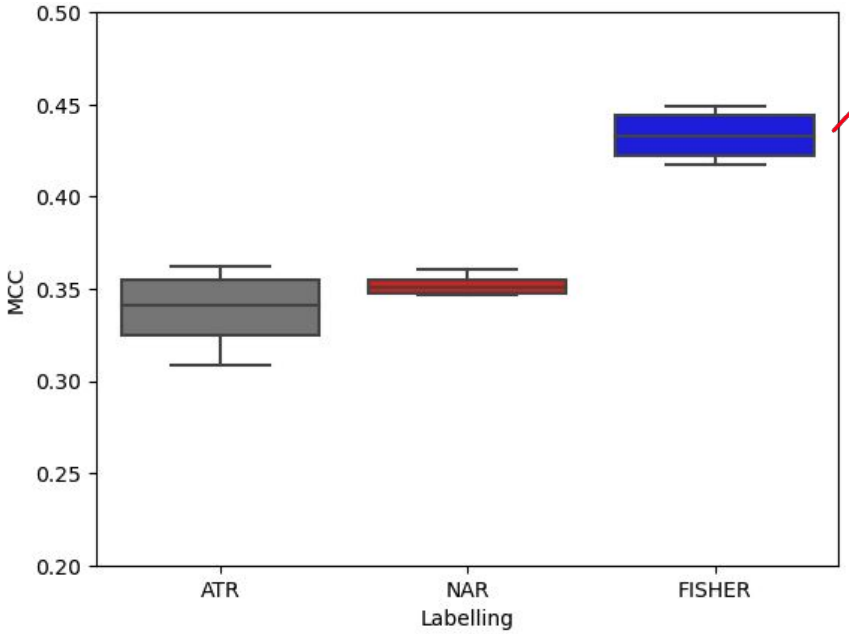
Model performances on the PubChem derived test set

	Labelling method	Threshold	PubChem test set			
			MCC	AUC	Precision	Recall
RFC	ATR	20%	0.35	0.91	0.20	0.75
		10%	0.36	0.91	0.21	0.75
		5%	0.33	0.91	0.18	0.73
		2%	0.31	0.90	0.17	0.73
	NAR	20%	0.35	0.91	0.19	0.79
		10%	0.35	0.91	0.19	0.81
		5%	0.36	0.91	0.21	0.75
		2%	0.35	0.91	0.19	0.78
	Fisher	20%	0.41	0.93	0.31	0.63
		10%	0.42	0.93	0.29	0.71
		5%	0.45	0.94	0.34	0.66
		2%	0.44	0.94	0.32	0.69

	Labelling method	Threshold	PubChem test set			
			MCC	AUC	Precision	Recall
MLP	ATR	20%	0.21	0.75	0.15	0.53
		10%	0.25	0.78	0.19	0.50
		5%	0.22	0.79	0.32	0.20
		2%	0.26	0.72	0.32	0.29
	NAR	20%	0.25	0.78	0.19	0.52
		10%	0.25	0.78	0.12	0.55
		5%	0.23	0.77	0.28	0.26
		2%	0.21	0.78	0.37	0.15
	Fisher	20%	0.17	0.73	0.12	0.58
		10%	0.24	0.77	0.19	0.45
		5%	0.25	0.72	0.25	0.34
		2%	0.21	0.75	0.39	0.14

Model performances on the PubChem derived test set

RFC MCC for different labelling methods

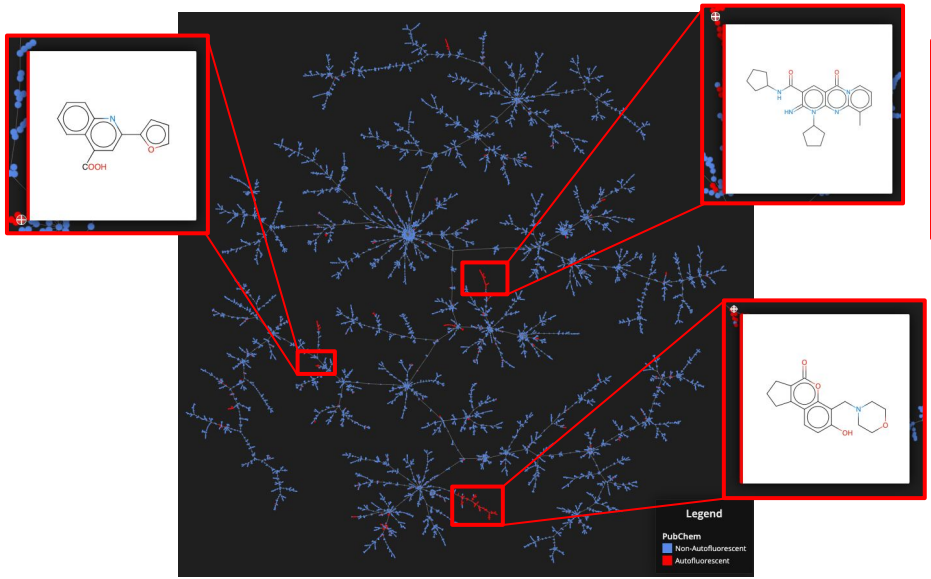


Benchmark against available methods

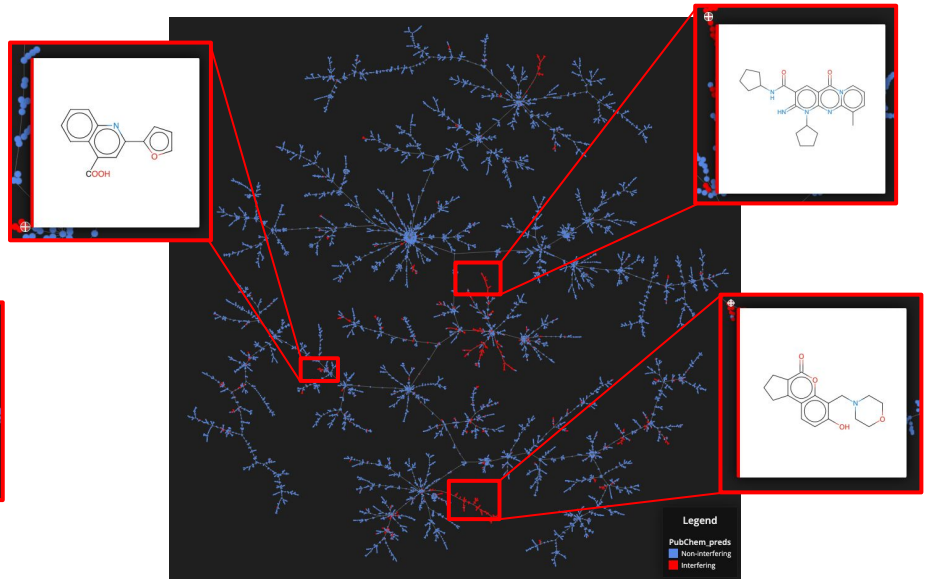
MODEL	MCC	ROC-AUC	Precision	Recall
HitDexter3.0	0.25	0.84	0.18	0.84
ChemFluo	0.34	0.82	0.23	0.75
FI-RF	0.45	0.94	0.34	0.66

universität wien

Analysis of performances on the PubChem derived test set



TMAP of PubChem test set compounds colored by PubChem activity label (BLUE non-autofluorescent, RED autofluorescent)



TMAP of PubChem test set compounds colored by predicted interference label (BLUE non-interfering, RED interfering)

universität wien

Conclusions

- **Single-dose HTS** data can be used with very **little preprocessing** to address assay interference
- We show that statistically derived labels can be used to train ML models for prediction of assay interference (best model reaching **MCC=0.47** on the internal test set)
- The interference labels obtained using ATR, NAR, and Fisher exact test can **approximate experimental evidence**
- Our best model **outperforms existing methods** for the prediction of autofluorescent compounds (**MCC=0.45** on the external test set)

Additional experiments

- Explore if the models can predict other type of interference (e.g. aggregation)
- Find additional public available datasets to further assess the models applicability
- Extend the approach to other assay technologies