HANDLING NUISANCE COMPOUNDS

●●● AT CZ-OPENSCREEN

Adam Hanzlík

ENBIK 2025

UNIVERSITY OF
CHEMISTRY AND TECHNOLOGY
PRAGUE

CZ OPENSCREEN AT IMG

HTS - IDEAL PIPELINE

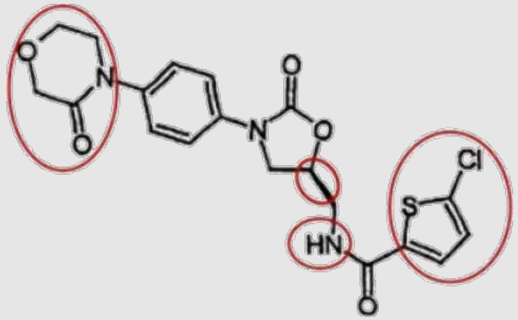
DIVERSE LIBRARY

EXPERIMENT DESIGN

AUTOMATION

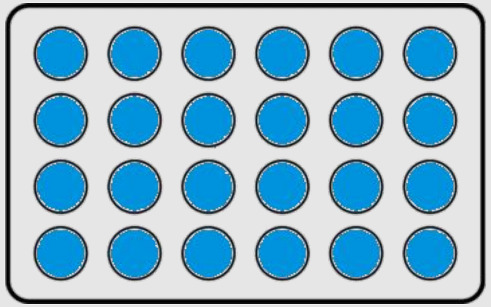
SCREENER

CHEMINFORMATICS



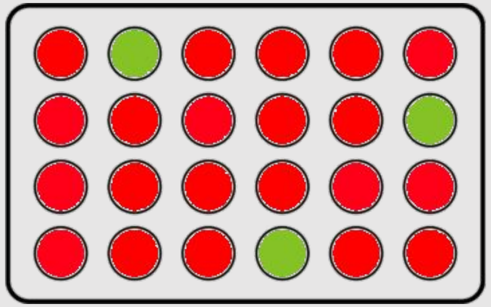
STRUCTURE-ACTIVITY RELATIONSHIP (SAR)

ANALYSIS



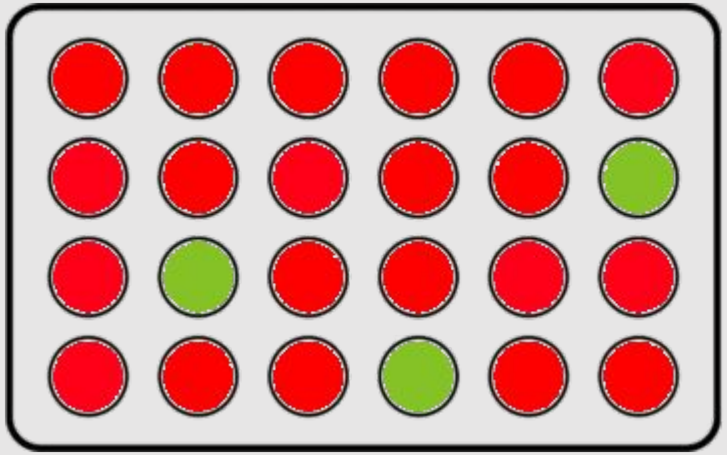
SCREENING SET

ASSAY



HITS

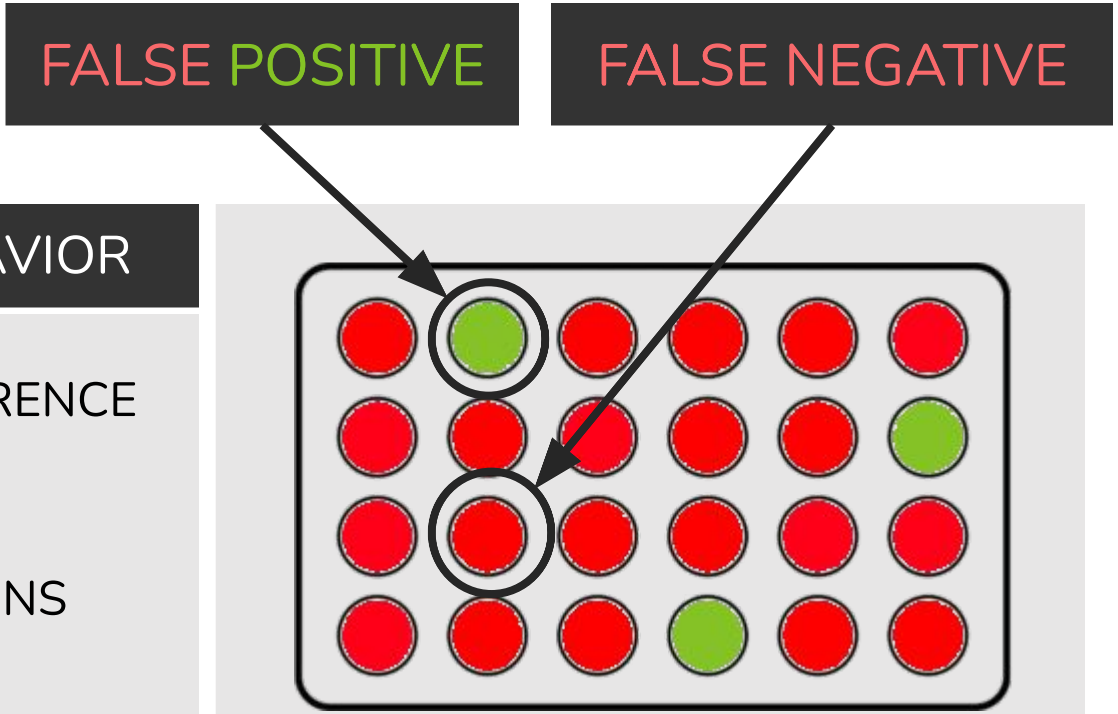
NUISANCE BEHAVIOR



TRUE ACTIVITY

NUISANCE BEHAVIOR

- ASSAY INTERFERENCE
- CYTOTOXICITY
- AGGREGATION
- REDOX REACTIONS
- NON-SPECIFIC INTERACTIONS



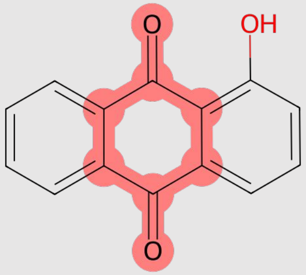
HITS

ANALYZING HITS RIDDLED WITH FALSE POSITIVES MAY LEAD TO MISLEADING SAR

TOOLS TO HANDLE NUISANCE

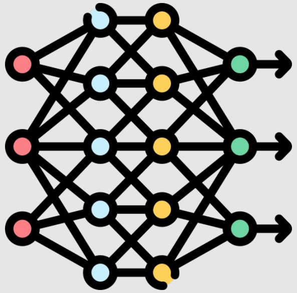
SUBSTRUCTURE FILTERS

- PAINS (Baell)
- GSK
- BMS
- LINT (Pfizer)
- and more (~2500)



MACHINE LEARNING MODELS

- HitDexter 3
- Luciferase Advisor
- BadApple
- SCAM Detective

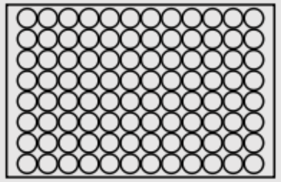


BAD COMPOUND LISTS

- Aggregator Advisor
- Nuisance Compounds in Cellular Assays
- CONS (Baell)
- Obsolete Compounds

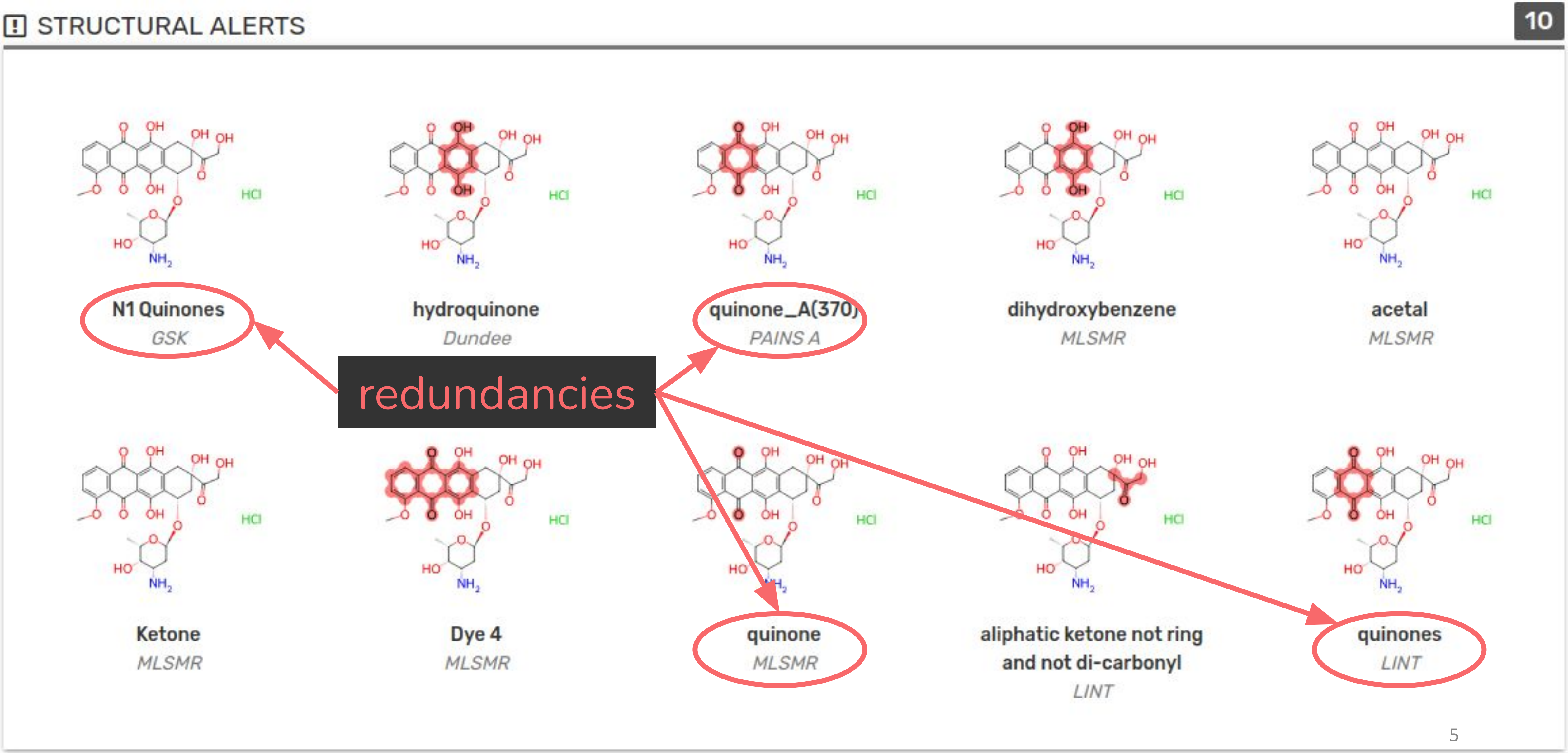
SECONDARY SCREENS

- ALARM NMR
- Orthogonal target
- Redox assays
- Technology counter assays



A compound that behaved badly before is likely to do so again

CURRENT CZ-OPENSCREEN IMPLEMENTATION



STRUCTURAL ALERTS

10

N1 Quinones *GSK*	**hydroquinone** *Dundee*	**quinone_A(370)** *PAINS A*	**dihydroxybenzene** *MLSMR*	**acetal** *MLSMR*
Ketone *MLSMR*	**Dye 4** *MLSMR*	**quinone** *MLSMR*	**aliphatic ketone not ring and not di-carbonyl** *LINT*	**quinones** *LINT*

redundancies

COMMON PITFALLS OF CURRENT TOOLS

- Bias for compounds set used to derive the method
- Lack of significance metric
- Redundancy when combining sources
- Lack of experimental context
- Hard to interpret
- SMARTS design sometimes not as intended

These pitfalls were carefully considered during the development of new tools to address nuisance behaviour at CZ-OPENSCREEN

TOOL 1 - UNIFIED SUBSTRUCTURE FILTER SET

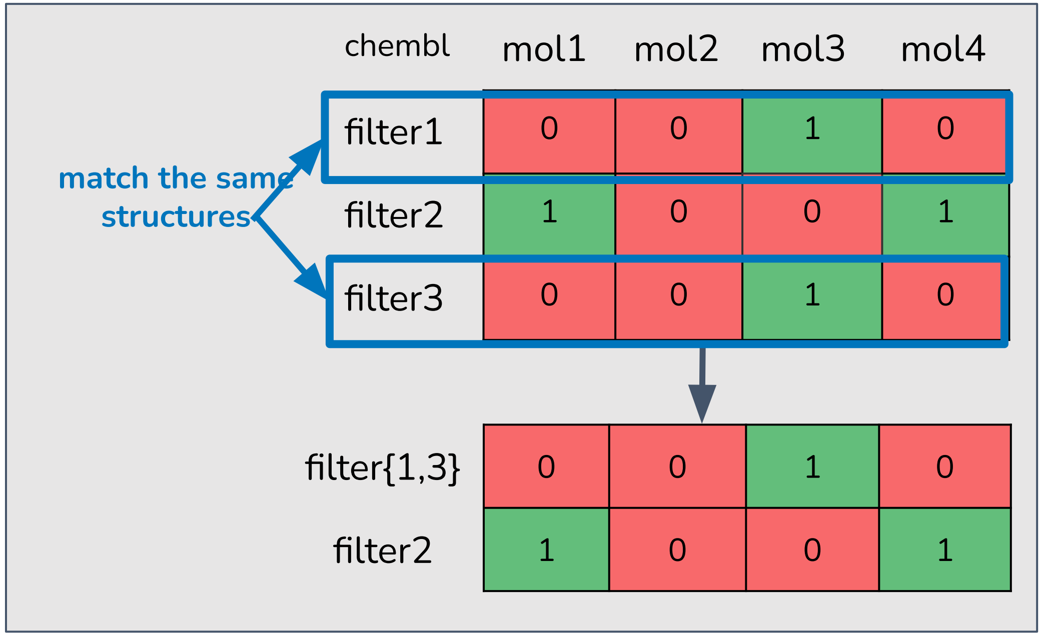
- Based on a collection of known filters provided by datamol-io/medchem
- Filters merged based on matching profile against Chembl

✓ 2500 filters reduced to 1500

✓ faster

✓ number of filters matched becomes more meaningful

✓ substructure filter fingerprint as input for ML



match the same structures

chembl	mol1	mol2	mol3	mol4
filter1	0	0	1	0
filter2	1	0	0	1
filter3	0	0	1	0

filter{1,3}	0	0	1	0
filter2	1	0	0	1

chembl	mol1	mol2	mol3	mol4
filter1	1	0	0	1
filter2	1	0	0	0

if filter1 then filter2

CZ-OPENSCREEN DATA AS SOURCE OF NEW KNOWLEDGE

- Around 4.5 million (sample X experiment X activity) measurements
- ~100K unique structures used in at least 15 experiments
- Comprehensive experiment metadata on every experiment:
  - target
  - method (Luminescence, Fluorescence intensity…)
  - assay format (cell based, biochemical)
  - reporter
- Quality Control performed on samples

THE ASSUMPTION

NUISANCE COMPOUND

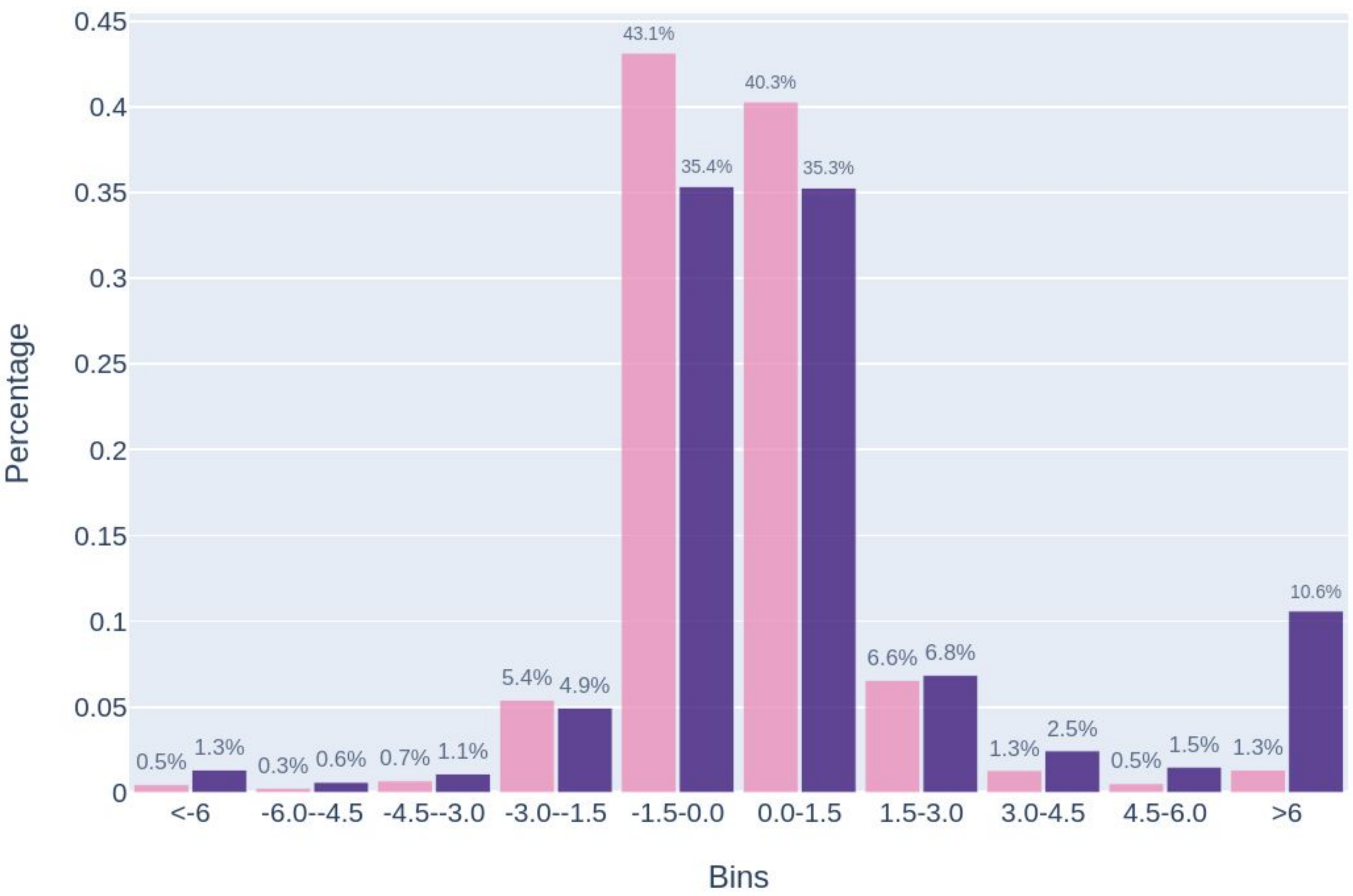
=

COMPOUND THAT APPEARS ACTIVE MORE OFTEN THAN EXPECTED IN DIFFERENT EXPERIMENTS

TO INVESTIGATE A COMPOUND, COMPARE THE DISTRIBUTION OF ITS READINGS TO THE GLOBAL DISTRIBUTION OF READINGS

TOOL 2 - HISTOGRAM + P-VALUE

Comparison of all autofluorescent-tagged against all in all primary assays. (p = 0.00e+00)



202956 structure_ids in full subset, (2882094 data points)

711 structure_ids in full subset, (46804 data points)

VISUAL

TWO SIDE-BY-SIDE HISTOGRAMS OF DISTRIBUTIONS

STATISTICAL

P-VALUE OF NONPARAMETRIC TEST (H0: SAME DISTRIBUTION)

●●● TOOL 2 - HISTOGRAM + P-VALUE

It does not matter wherever we look at structures, samples, libraries, sources, method subsets or all primary assay data.

We are always comparing two distributions of activity scores.

All readings are normalized using a modified Z-score algorithm (b-score, median instead of mean).

Row-wise, column-wise and plate-wise median polishing

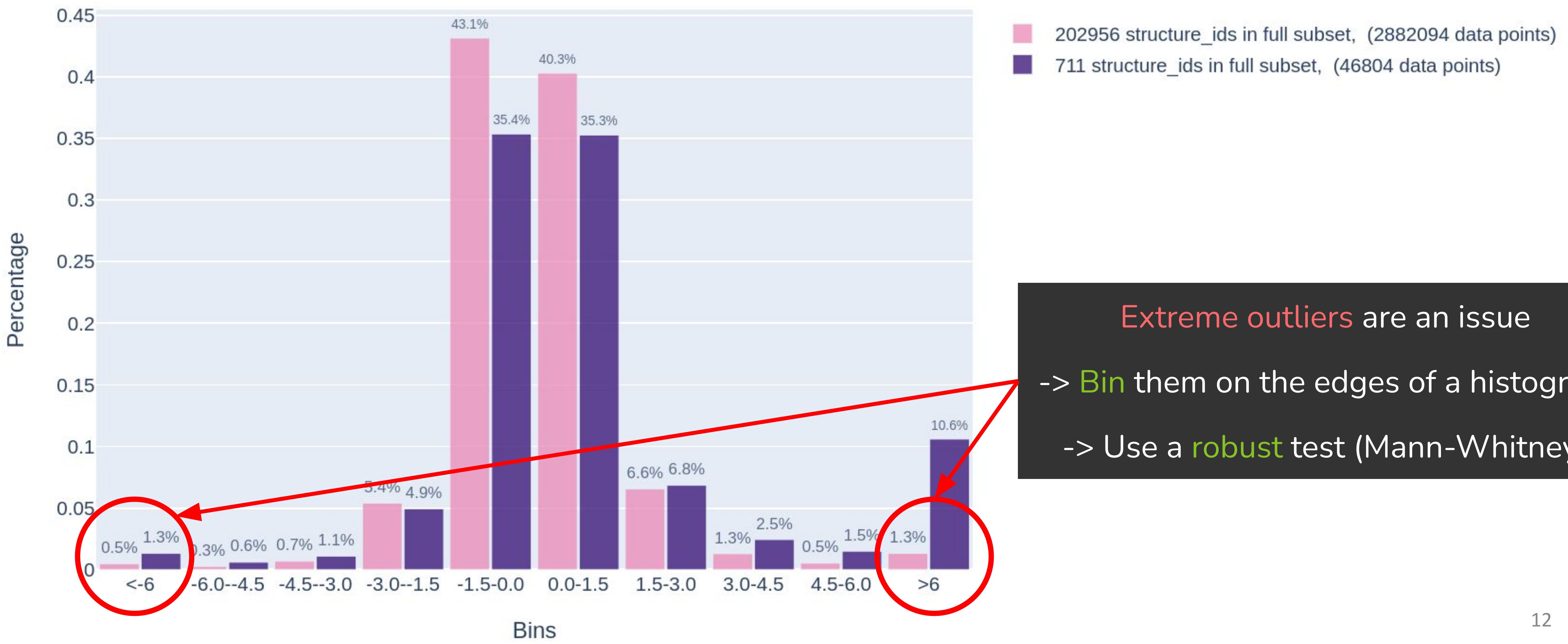
Extreme outliers are an issue

-> Bin them on the edges of a histogram

-> Use a robust test to compare the distributions

TOOL 2 - HISTOGRAM + P-VALUE

Comparison of all autofluorescent-tagged against all in all primary assays. (p = 0.00e+00)

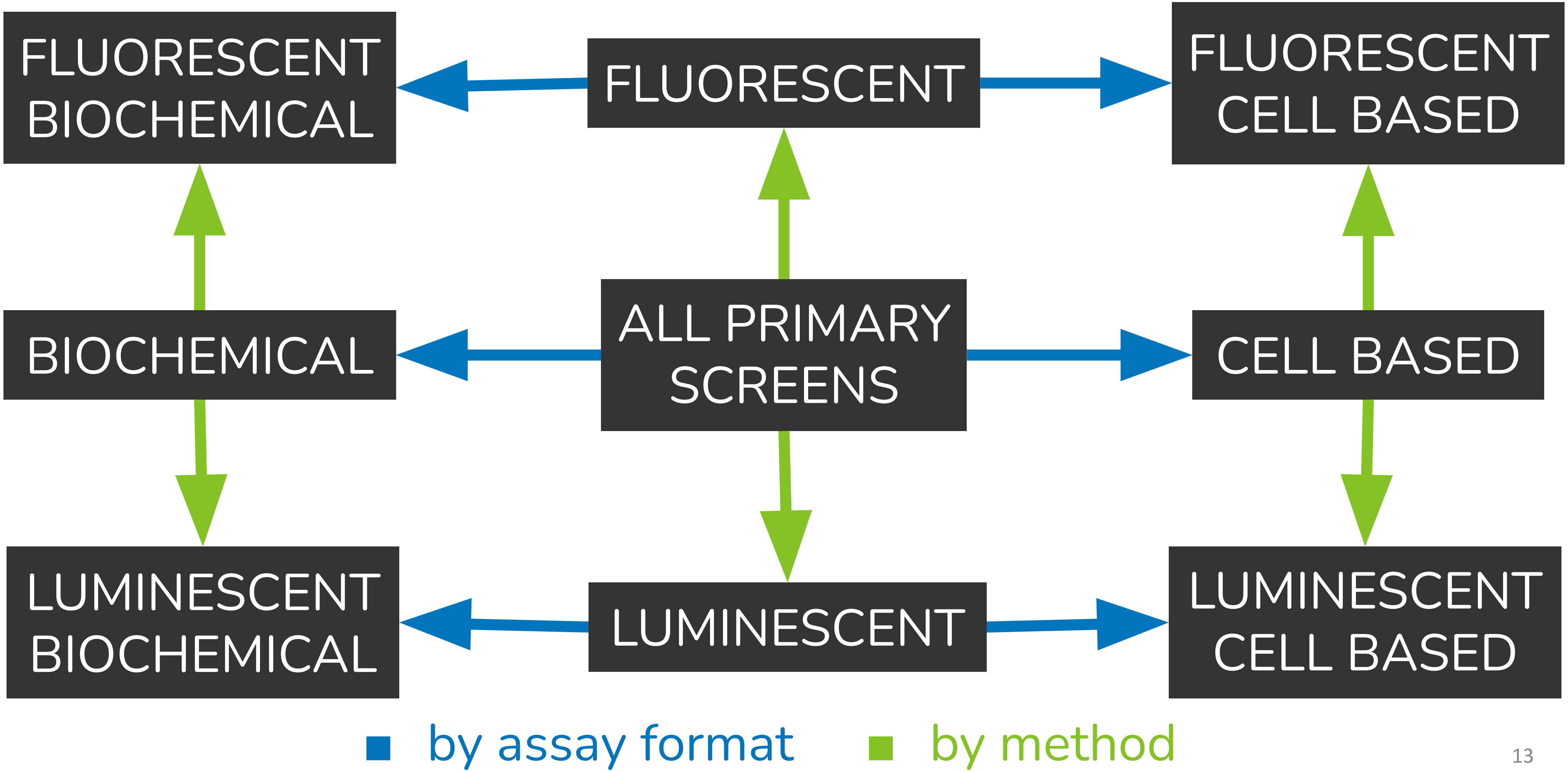


Extreme outliers are an issue

-> Bin them on the edges of a histogram

-> Use a robust test (Mann-Whitney)

STRATIFICATION BASED ON METADATA TO REDUCE NOISE



FLUORESCENT BIOCHEMICAL

FLUORESCENT

FLUORESCENT CELL BASED

BIOCHEMICAL

ALL PRIMARY SCREENS

CELL BASED

LUMINESCENT BIOCHEMICAL

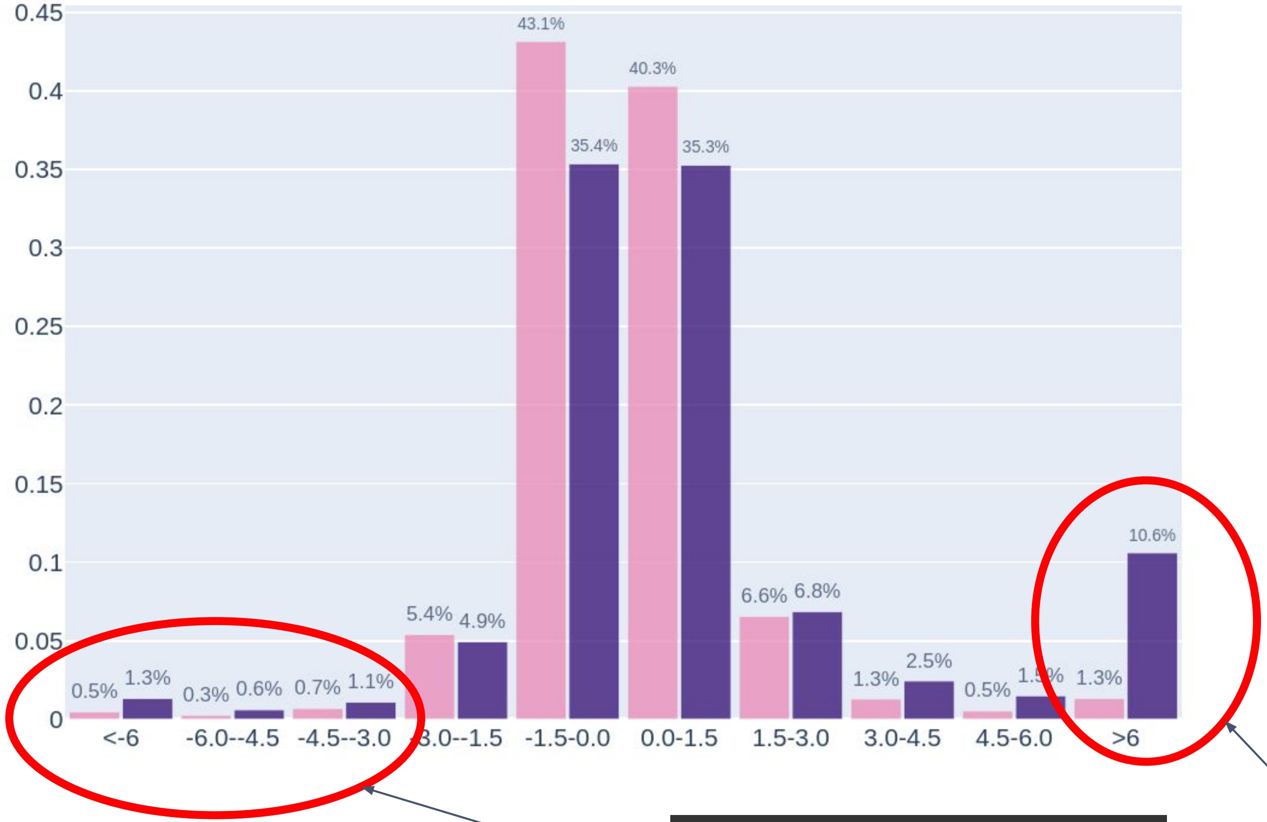
LUMINESCENT

LUMINESCENT CELL BASED

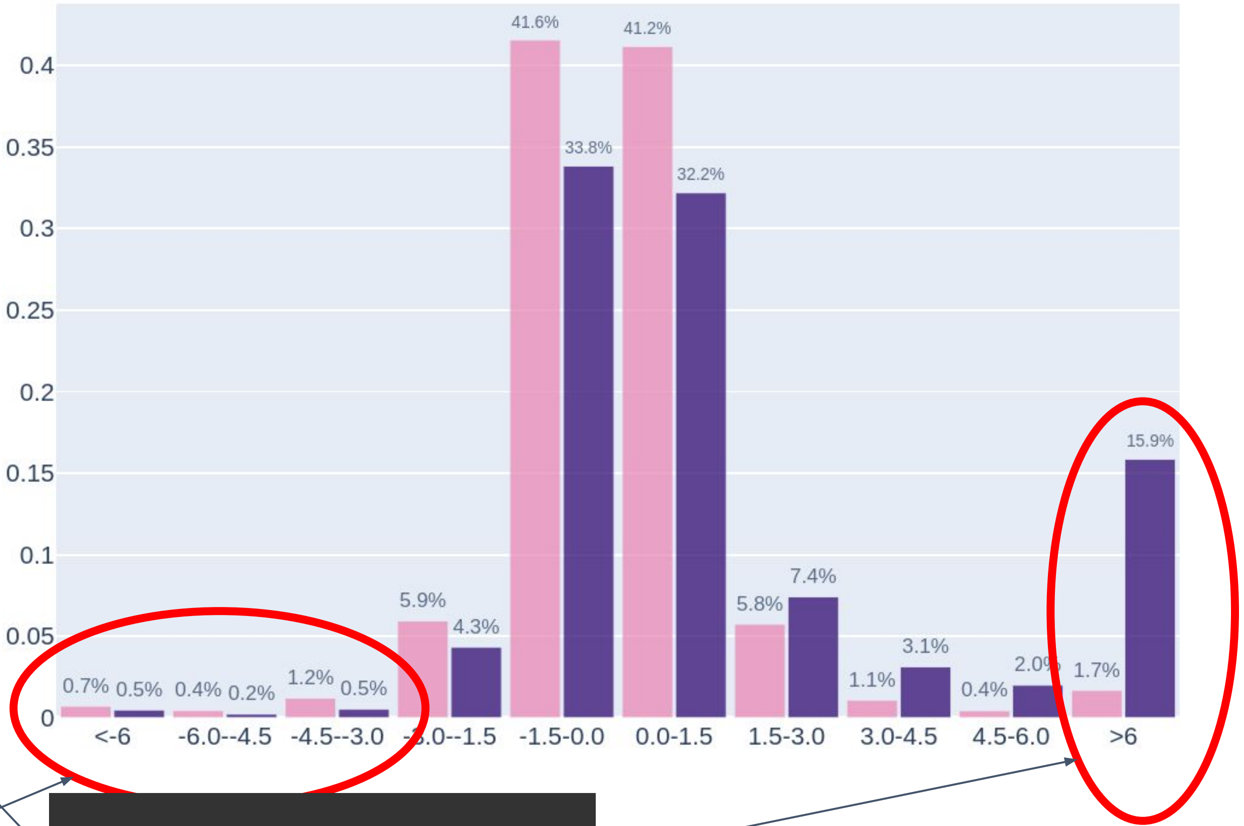
🟦 by assay format 🟩 by method

STRATIFICATION REDUCES NOISE

ALL METHODS AND FORMATS



FLUORESCENCE AND BIOCHEMICAL



LESS NOISE

ENRICHMENT

●●● TOOL 2 - HISTOGRAM + P-VALUE SUMMARY

- ✓ One method can assess ANY set of samples or structures globally or focused on a particular method and/or assay format subset
- ✓ Flexible and fast comparisons (slices can be preset and precalculated
- ✓ Visual + quantitative -> interpretation + significance
- ✓ Interactive - bin ranges, outlier thresholds…

TOOL 3 - NUISANCE FLAGS

SAMPLE OR STRUCTURE

FOCUSED ON

METHOD

FLUORESCENCE

OR

LUMINESCENCE

FORMAT

CELL-BASED

OR

BIOCHEMICAL

TARGET

SAME TARGET EXPERIMENTS POOLED

HITS

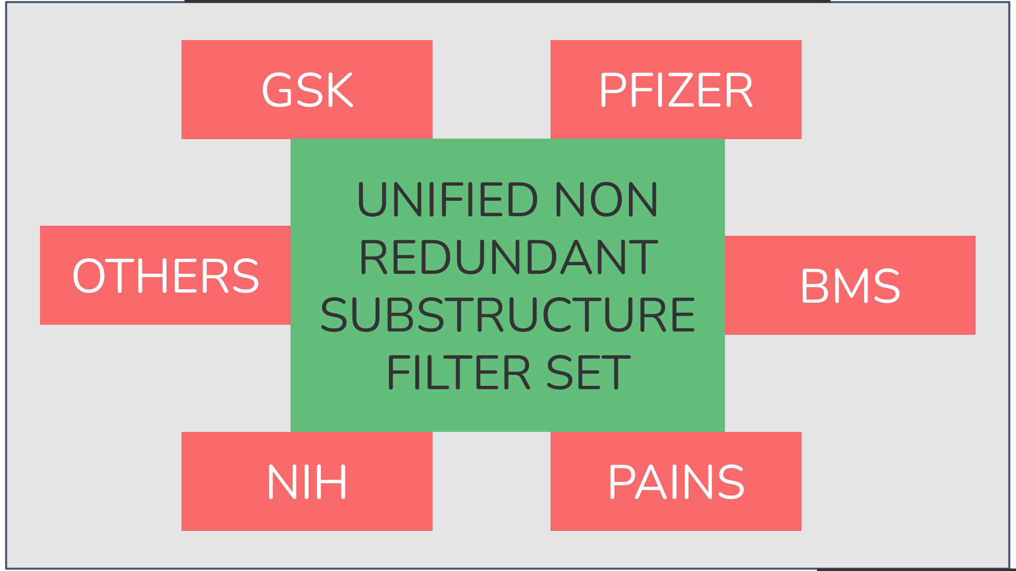
ACTIVITY IN IMPLIED EXPERIMENT DIRECTION

IF P VALUE OF STATISTICAL TEST LOWER THAN THRESHOLD

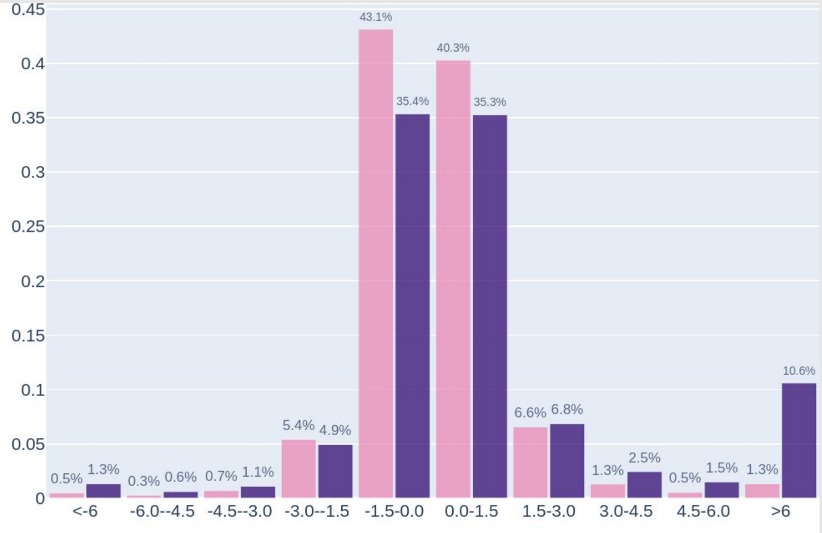
THEN FLAGGED AS

AUTOFLUORESCENT, AUTOLUMINESCENT, QUENCHER, CYTOTOXIC, PROMISCOUS (ACTIVE AGAINST MANY TARGETS), FREQUENT HITTER

UNIFIED FILTER SET



NUISANCE HISTOGRAM



NUISANCE FLAGS

ID 1 🚩 AUTOFLUORESCENCE

ID 6 🚩 CELL TOXICITY

ID 14 🚩 FREQUENT HITTER

⋮ ⋮

●●● SUMMARY

ACKNOWLEDGEMENTS

Milan Voršilák

supervisor

Ctibor Škuta

consultation

Petr Bartůněk

consultation

Tomáš Muller

consultation

Funding:
Supported by **National Infrastructure for Chemical Biology** (LM2023052), funded by the **Ministry of Education, Youth and Sports** under the **Large RDI Infrastructures Programme**, 2023–2026.
CZ-OPENSCREEN is the Czech node of **EU-OPENSCREEN ERIC**, providing open-access research in chemical biology and genetics.