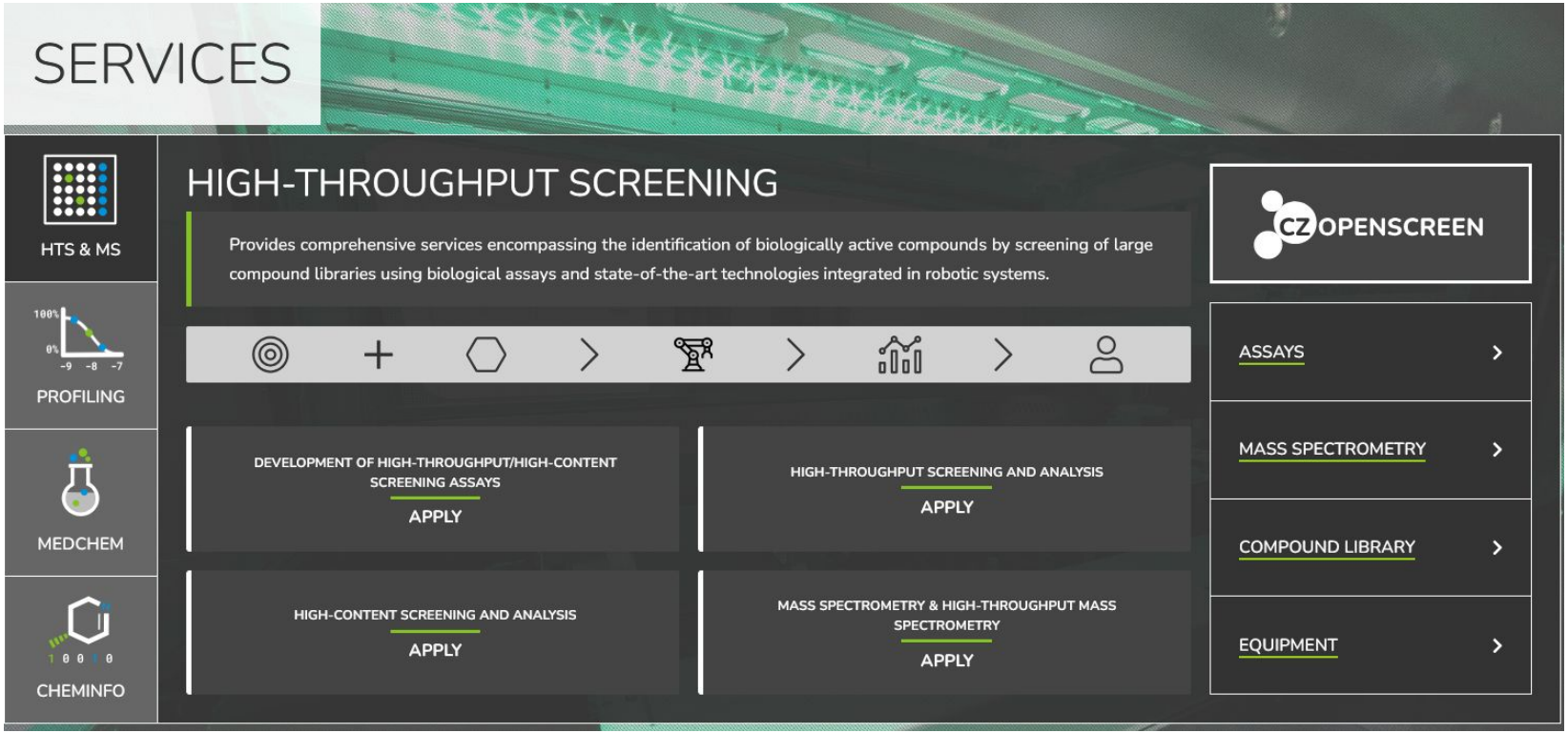
Handling Nuisance Compounds at CZ-OPENSCREEN

Adam Hanzlík

CZ-OPENSCREEN, Institute of Molecular Genetics CAS · UCT Prague · 2026

CZ-OPENSCREEN



HTS Pipeline

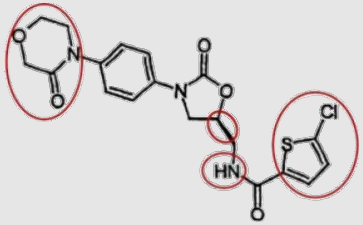
DIVERSE LIBRARY

EXPERIMENT DESIGN

AUTOMATION

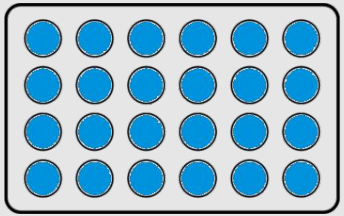
SCREENER

CHEMINFORMATICS



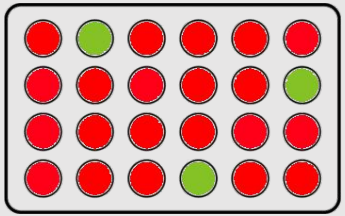
STRUCTURE-ACTIVITY RELATIONSHIP (SAR)

ANALYSIS



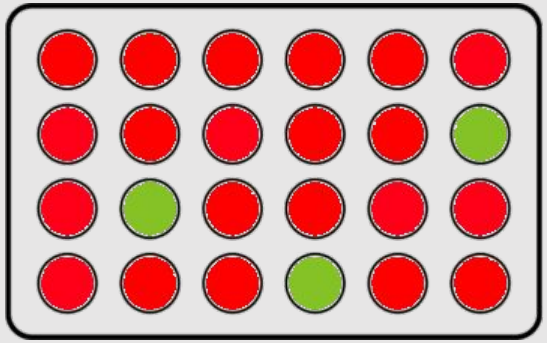
SCREENING SET

ASSAY



HITS

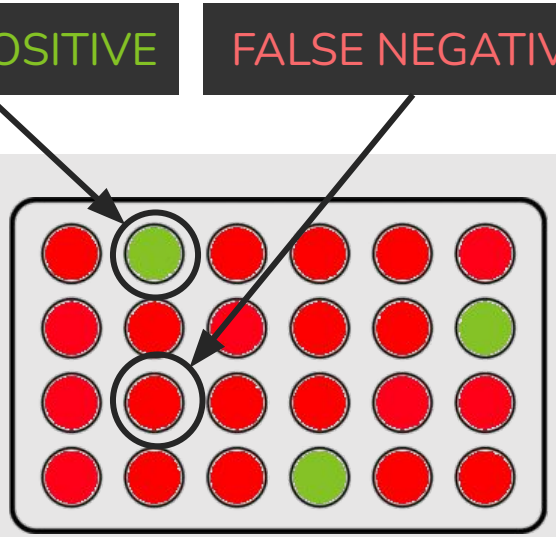
Nuisance Behaviour



TRUE ACTIVITY

NUISANCE BEHAVIOR

- ~95% of hits not genuine
- PAINS, frequent hitters
- aggregation, autofluorescence, cytotoxicity
- format-dependent



HITS

ANALYZING HITS RIDDLED WITH FALSE POSITIVES
LEADS TO SIGNIFICANT RESOURCE LOSS

Existing Tools

BAD COMPOUND LISTS

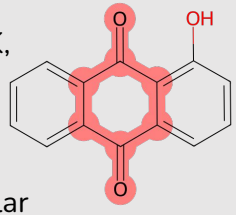
- Nuisance Compounds in Cellular Assays
- most reliable

SIMILARITY SEARCH

- against curated lists
- can't advise on novel chemistry
- AggregatorAdvisor

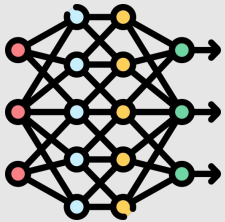
SUBSTRUCTURE FILTERS

- PAINS, Eli Lilly, GSK, BMS
- usually SMARTS
- easy to share, popular



MACHINE LEARNING MODELS

- e.g. HitDexter
- can capture more complex relationships
- hard to share



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

Weaknesses of SMARTS structure filters

No experimental context

In what context should the filter be considered
Fluorescence or luminescence?
Cells or biochemical?

No quantification

Filter = binary classification
Some filters are much more serious than others
Warning vs. Alert

Imperfect redundancy

Different facilities run into the same issues
but describe them slightly differently
-> Equivalent filters that do not cluster trivially

SMARTS quirks

creating good SMARTS filters is not simple
some non-obvious rdkit shenanigans

→ **At CZ-OPENSCREEN experiments record mandatory metadata about FORMAT and METHOD used**

Our approach

I Empirical pipeline

Nuisance flags from our own screening history

II Unified filter library

Non-redundant, labeled SMARTS filter set

annotates

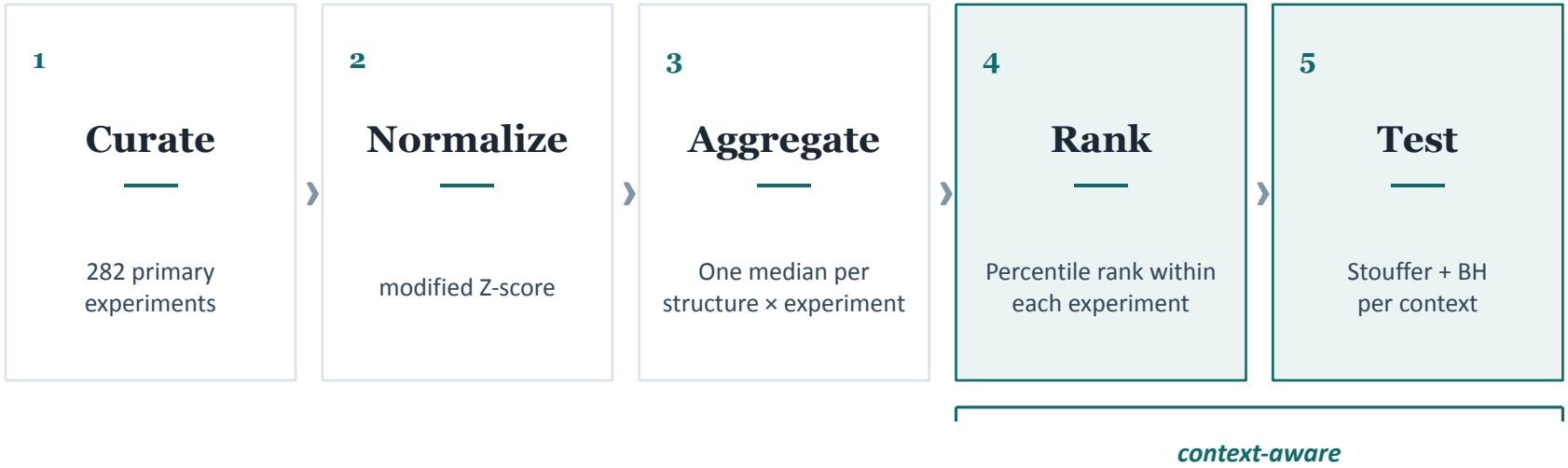
provides scaffolds

Context-aware substructure-filter library

each warning carries its empirical nuisance evidence per assay context

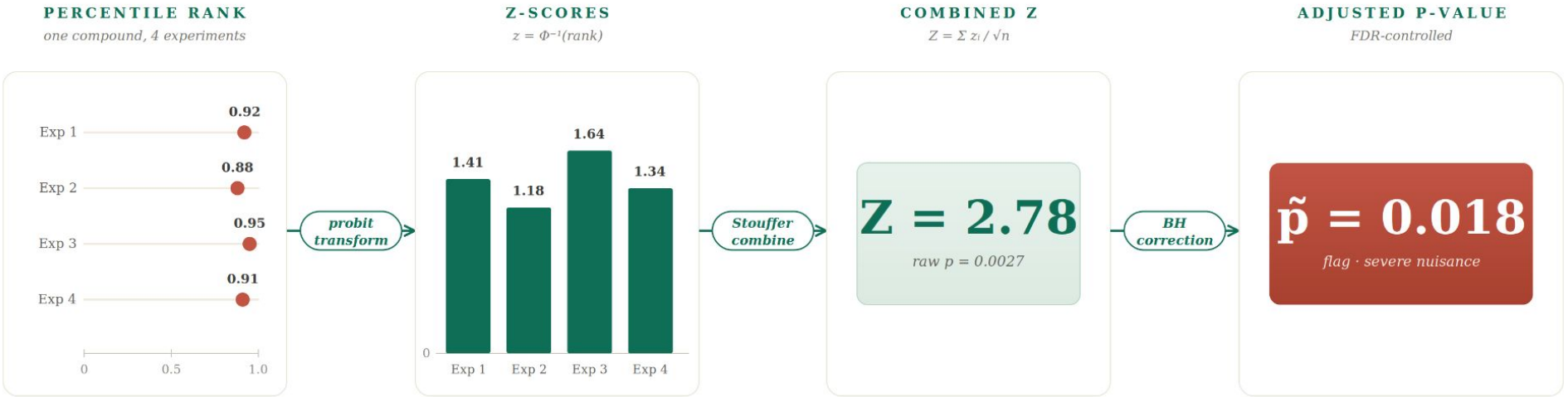
Context-aware annotation pipeline

Context = detection method + assay format + direction



Scope: 282 primary experiments ~7.9 M datapoints · 202 K structures · 10 assay contexts

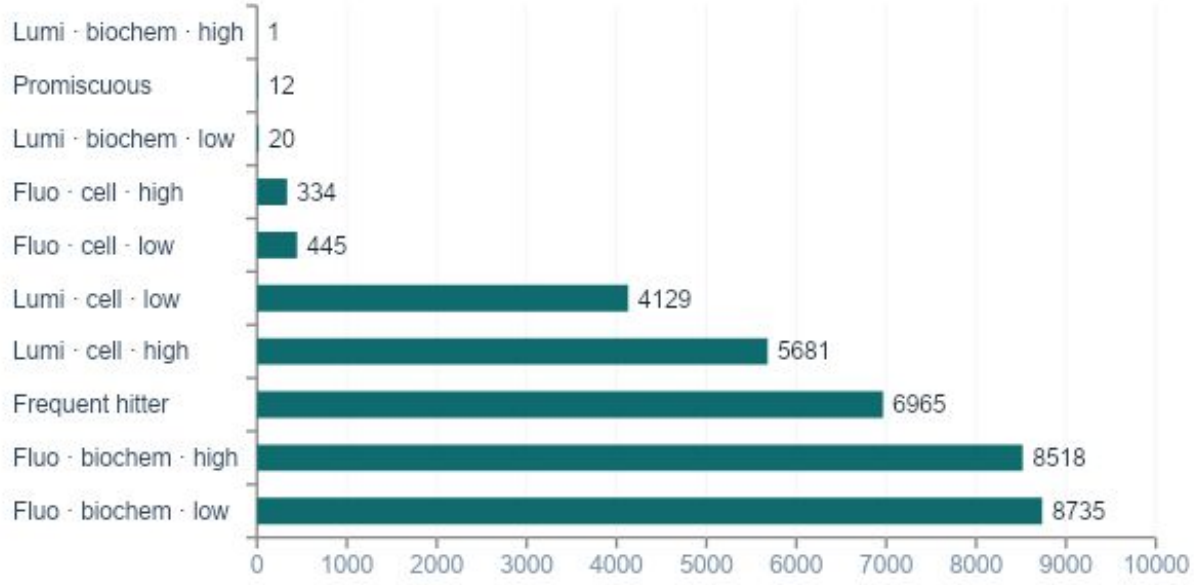
Identifying anomalous compounds



Per-experiment ranks → z-scores → combined Z → adjusted p-value

Severe flags across assay contexts

117 K samples annotated · 18 764 severe flags



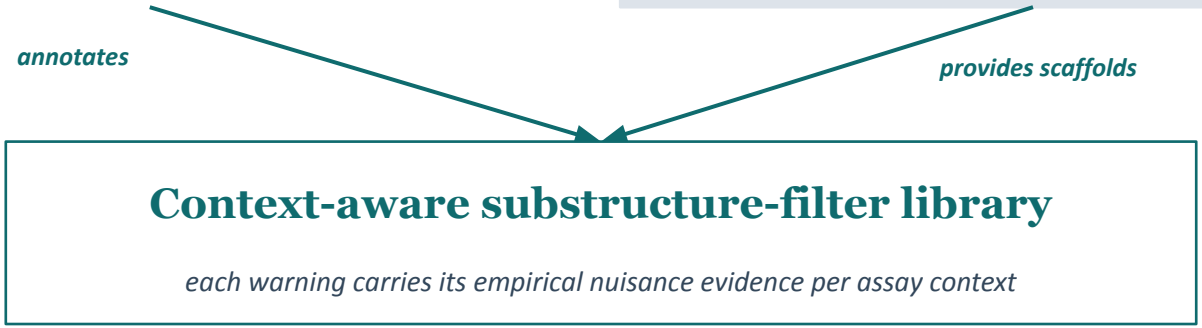
Our approach

I Empirical pipeline

Nuisance flags from our own screening history

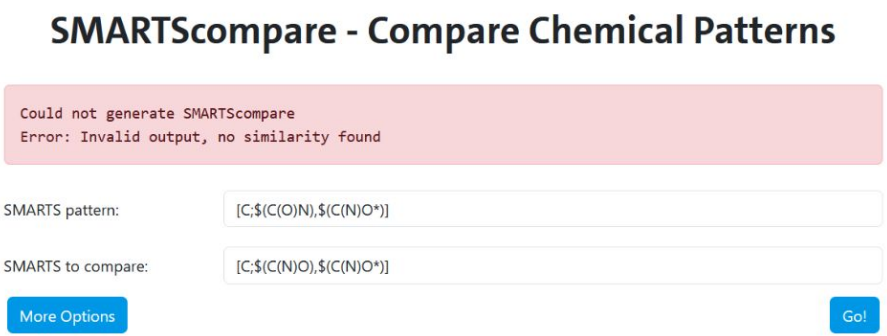
II Unified filter library

Non-redundant, labeled SMARTS filter set



Clustering by matching sets

Compared by what they match, not by how they are written.



Equivalent patterns, not recognised

THE APPROACH

Matching every filter against ChEMBL and PubChem

matching set = vector

Compare filters by their sets

Jaccard distance from matching set pairs

Hierarchical Clustering on the similarity

filters form equivalent groups

2 902 raw filters → **2 051** non-redundant clusters

SMARTS matching is not trivial

Naive use of canonical PAINS filters: 355 of 480 match nothing in ChEMBL.

1

Hydrogen suppression

require MergeQueryHs=True
(not the default)

2

qmol typecasting

SMARTS-only constructs need it;
SMILES-like patterns break with it

3

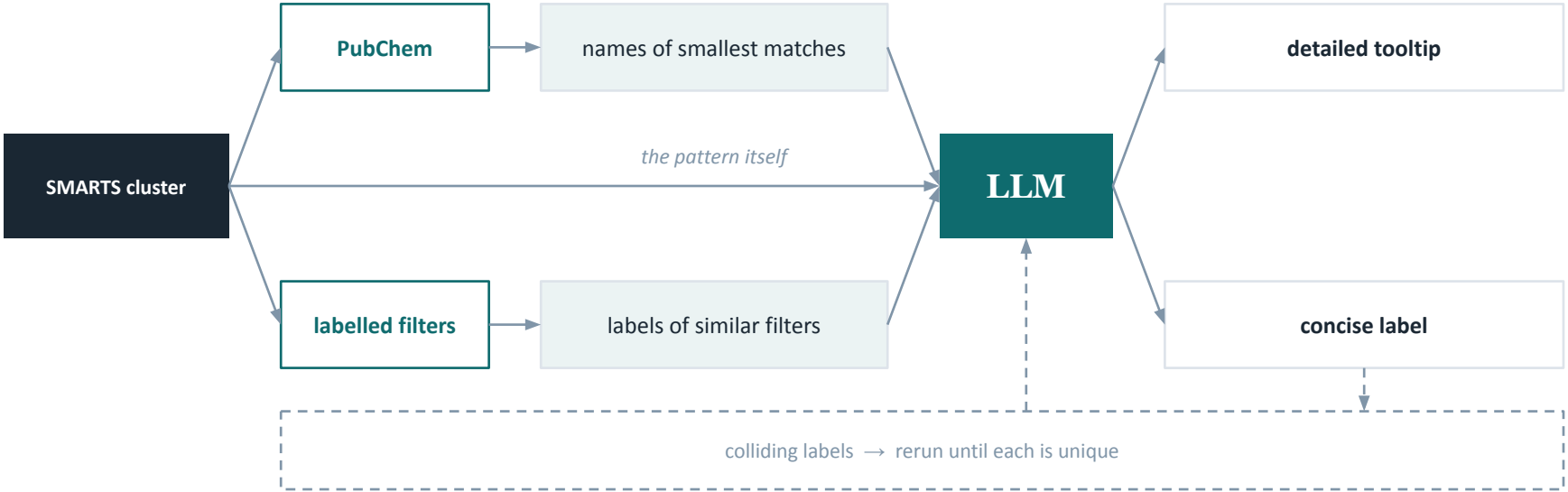
Hydrogens in OR statements

rdkit currently can’t suppress them

→ ***recommended reading: John Mayfield - PAINS in the butt presentation***

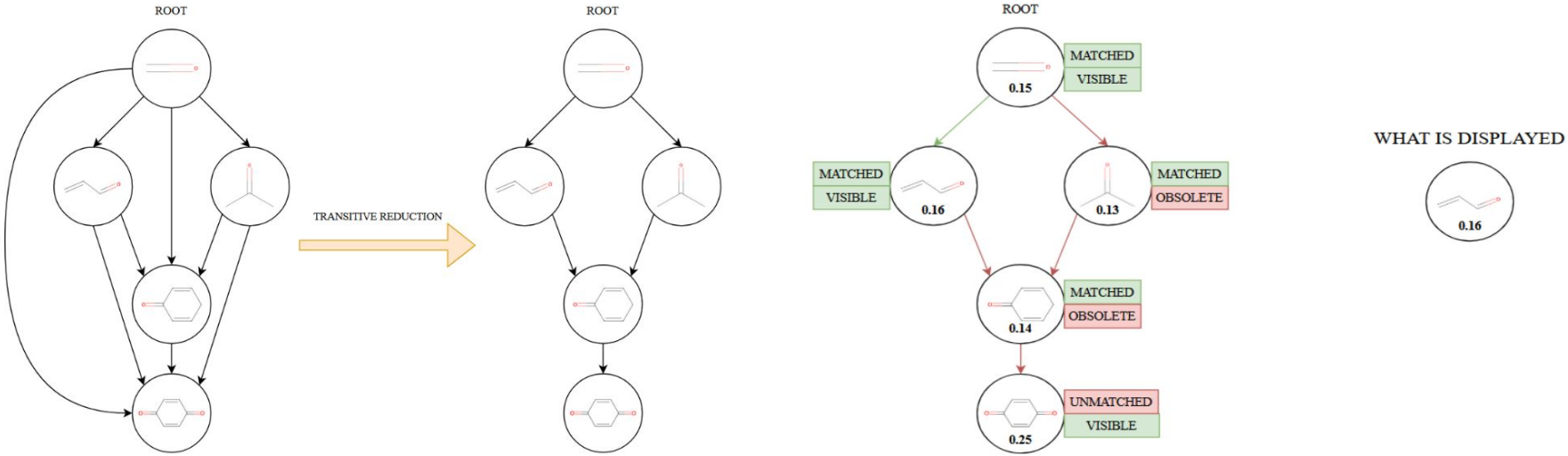
SMARTS LLM Annotator

The model reads the SMARTS; the prompt gives it evidence and examples to lean on.



→ **ongoing benchmarking and implementation · github.com/hanzlika/smarts_llm_annotator**

Hierarchical organization of filters



Only the most specific matched, non-obsolete node is shown

2 051 clusters → **375** statistically validated (adjusted pval < 0.05 AND enrichment > 2x in ≥ 1 context)

Our approach

I Empirical pipeline

Nuisance flags from our own screening history

II Unified filter library

Non-redundant, labeled SMARTS filter set

annotates

provides scaffolds

Context-aware substructure-filter library

each warning carries its empirical nuisance evidence per assay context

Integration

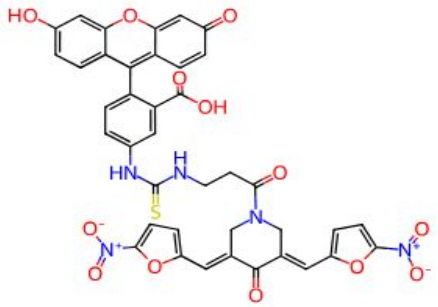
ANALYSIS RESULTS

Data Uploaded | Echo Log | Normalization B_score | Hits Selected 795 | Warnings 303

ID 36867 105563 B-score: **2.491**	ID 63493 Z805962858 lumi/cell B-score: **9.298**	ID 100357 9121396 B-score: **3.721**	ID 73736 Z1270867534 lumi/cell B-score: **27.729**	ID 90123 5721039 lumi/cell B-score: **8.221**
ID 65548 Z285682368 lumi/cell	ID 86029 Z51931897	ID 49166 Z300233624 lumi/cell	ID 65551 Z383577378	ID 36882 105307

ONE BADGE PER STRUCTURE

fluo/biochem

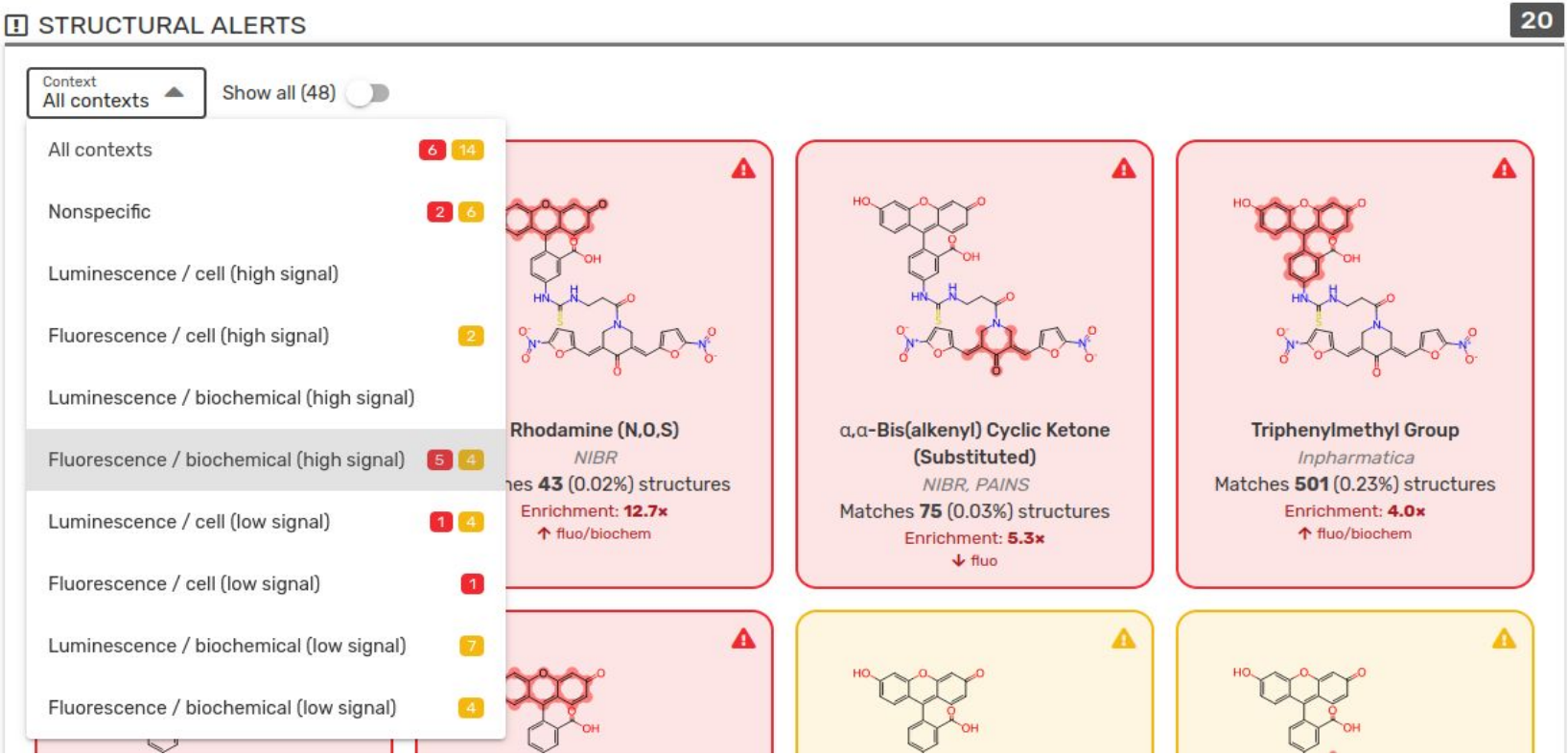


DETAILED TOOLTIPS

nonspecific

Frequent hitter: active in 53.4% of assays screened (71/133, p < 0.0001)
Frequent high signal in fluorescence / biochemical assays: 61.5% (n=65, p < 0.0001)
Poly-target: active against 37.5% of targets tested (9/24, p = 0.0601)
Substructure match: 17.6% of matched compounds are associated with frequent high signal in fluorescence / cell-based assays.
Substructure match: 16.8% of matched compounds are associated with frequent low signal in luminescence / biochemical assays.
Substructure match: 16.5% of matched compounds are associated with frequent high signal in fluorescence / biochemical assays.
Substructure match: 15.7% of matched compounds are associated with frequent signal in all assay types (nonspecific).
Substructure match: 15.3% of matched compounds are associated with frequent low signal in fluorescence / cell-based assays.
Substructure match: 14.6% of matched compounds are associated with frequent low signal in luminescence / cell-based assays.
Substructure match: 14.3% of matched compounds are associated with frequent low signal in fluorescence / biochemical assays.

Integration



What’s next?

Gap analysis

- Look at structures with confirmed nuisance that are not covered by good filters
- Design new filters
- Improve existing filters via editing to increase enrichment
- EU-OPENSCREEN/ECBD level validation and implementation
- validation against a real assay

Thank you — questions welcome.

Acknowledgement

Milan Voršilák

MSc. supervisor

Ctibor Škuta

PhD supervisor

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.