ACS Spring Meeting, Orlando, FL Mar 31- April 4, 2019
CINF: Applications of Cheminformatics to Environmental Science

United States Environmental Protection Agency

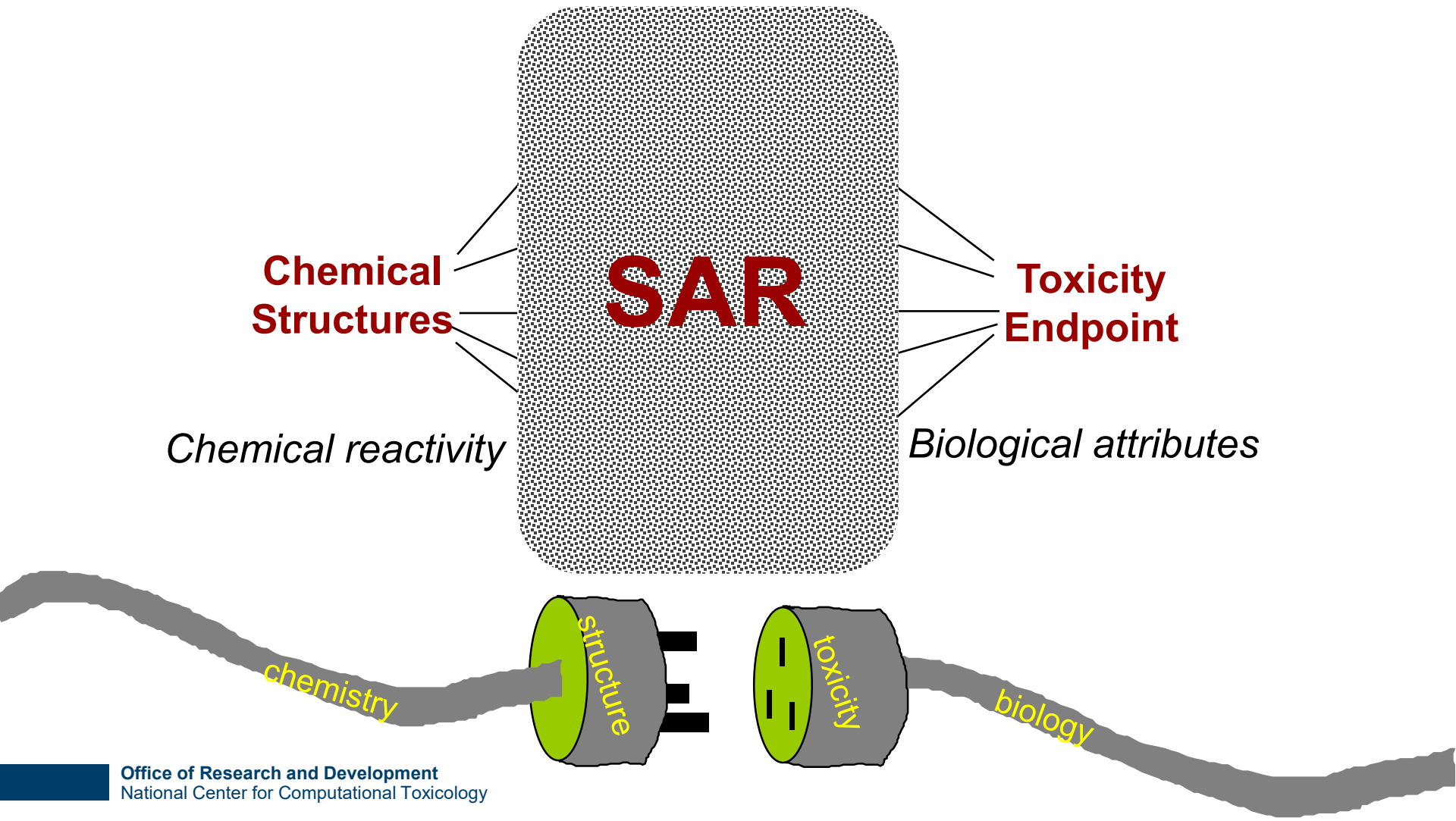
Applications of a Chemotype-Enrichment Approach to the ToxCast data landscape and beyond: *Inverting the SAR paradigm*

Coauthors:
Ryan Lougee
Christopher Grulke
Nancy Baker
Jun Wang
Antony Williams

Presenter: Ann Richard
National Center for Computational Toxicology
Office of Research & Development
richard.ann@epa.gov

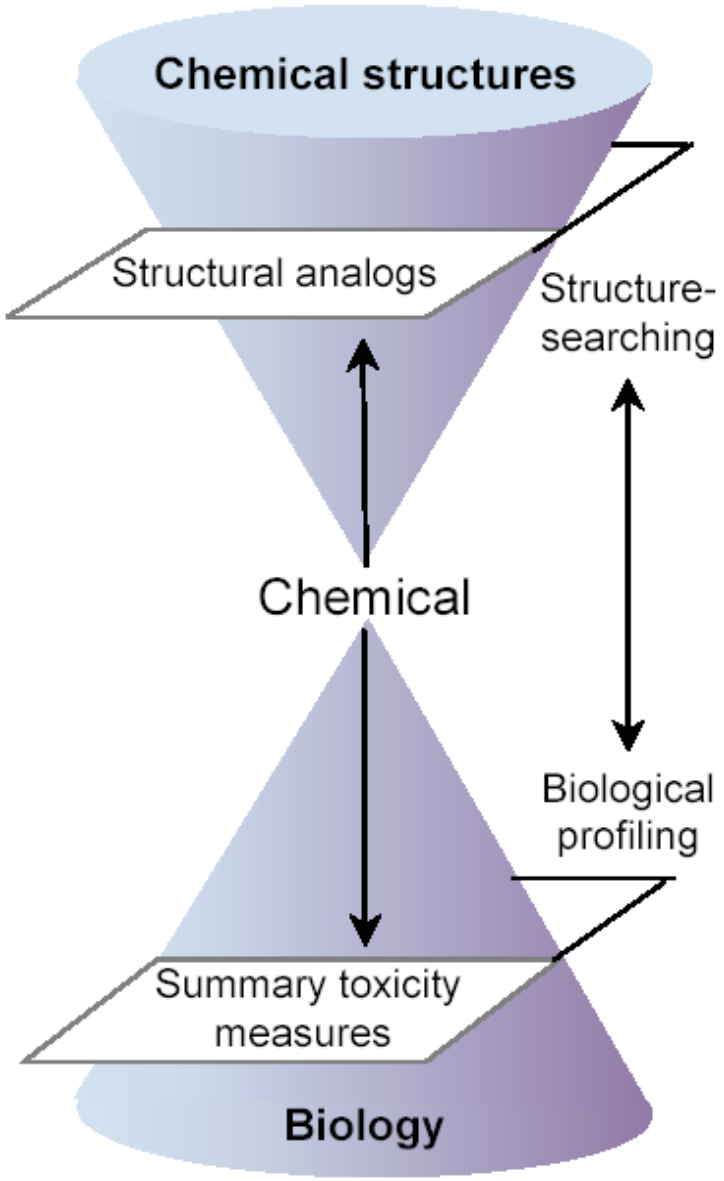
Office of Research and Development
National Center for Computational Toxicology

Global vs. Local SAR models

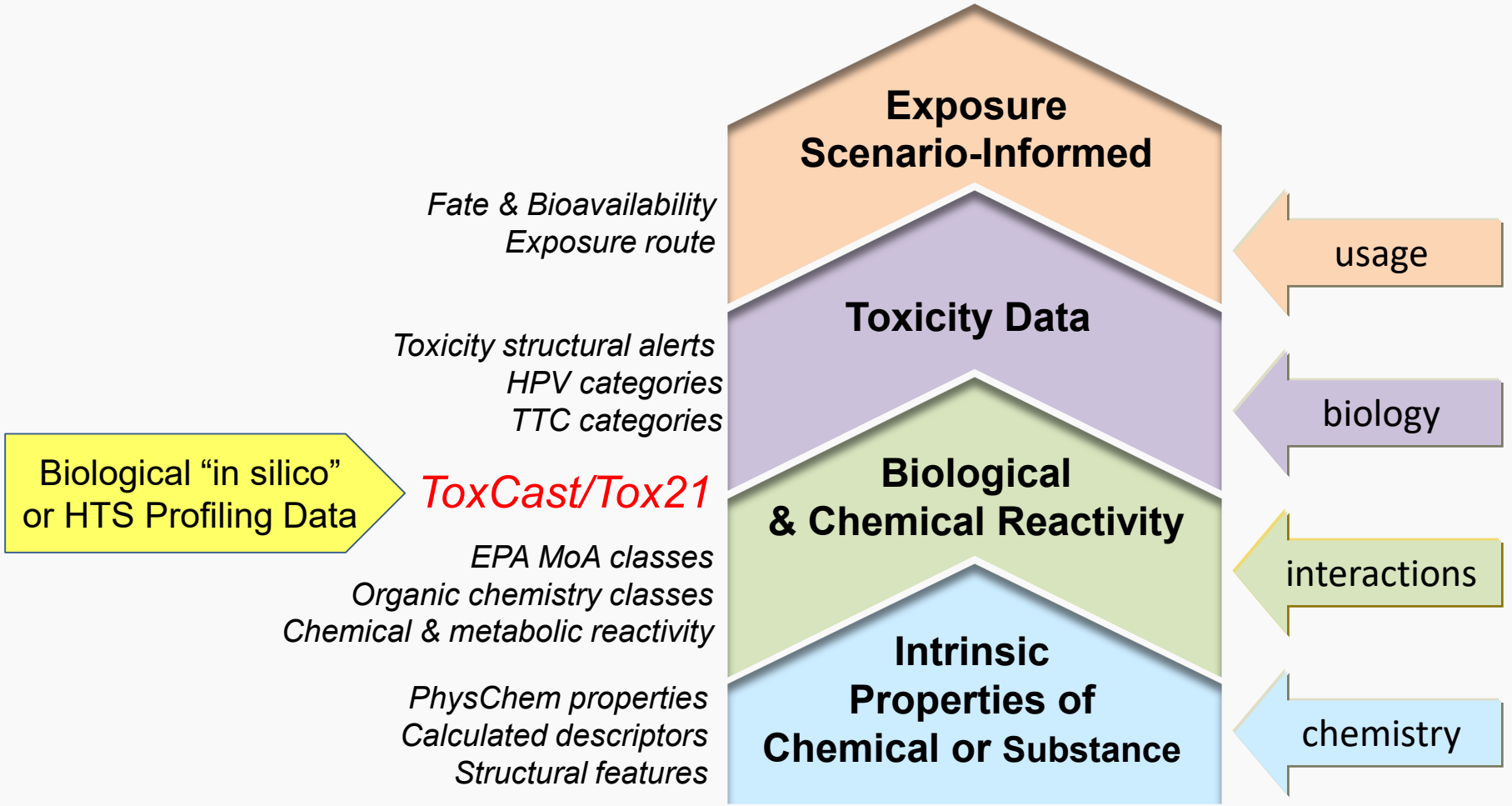


“Lensing”

For any given chemical-assay dataset, what is the optimal level of chemical and biological specificity to discover useful SAR associations?

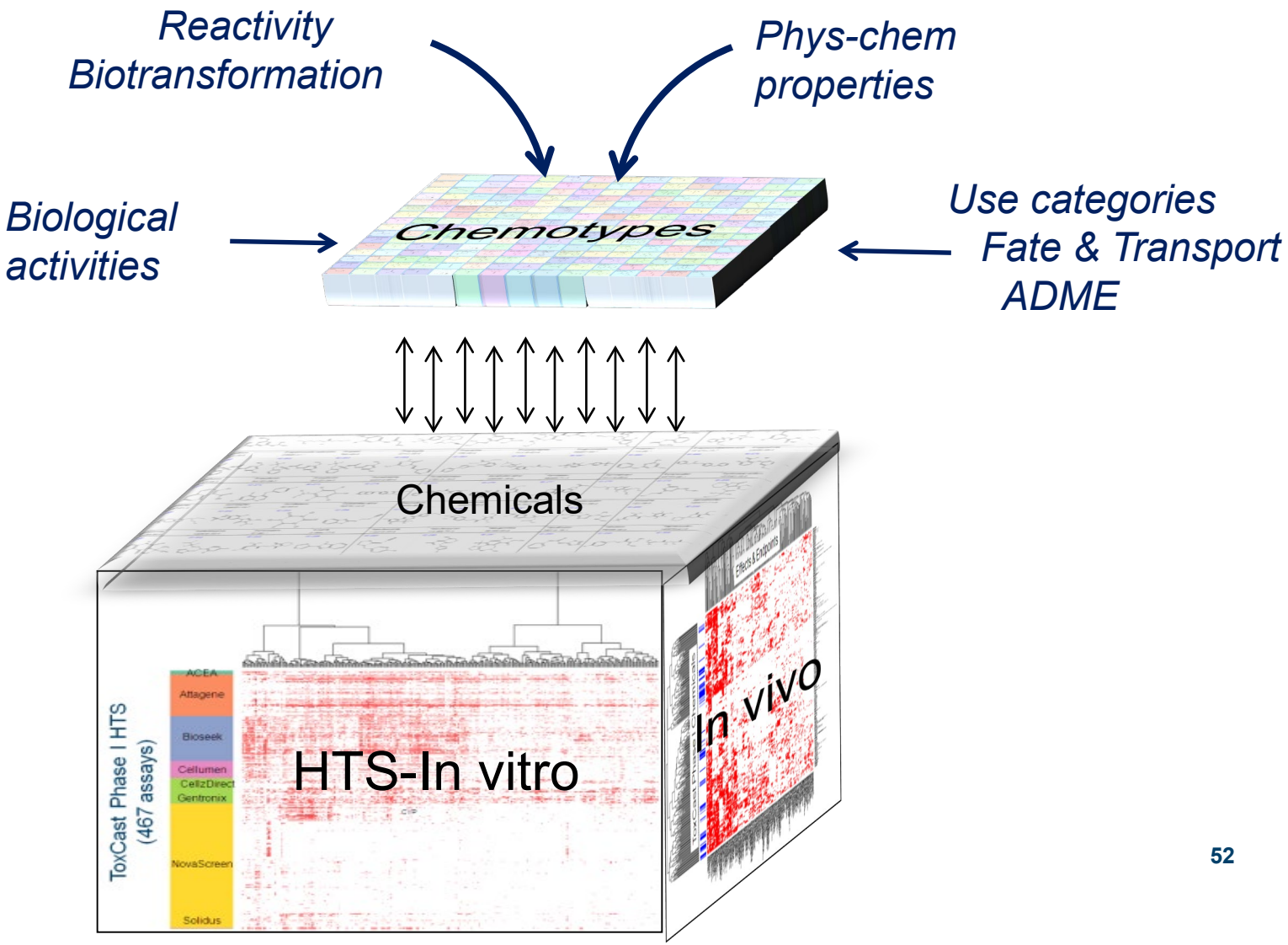


Chemical/Substance representations for hazard & risk evaluation



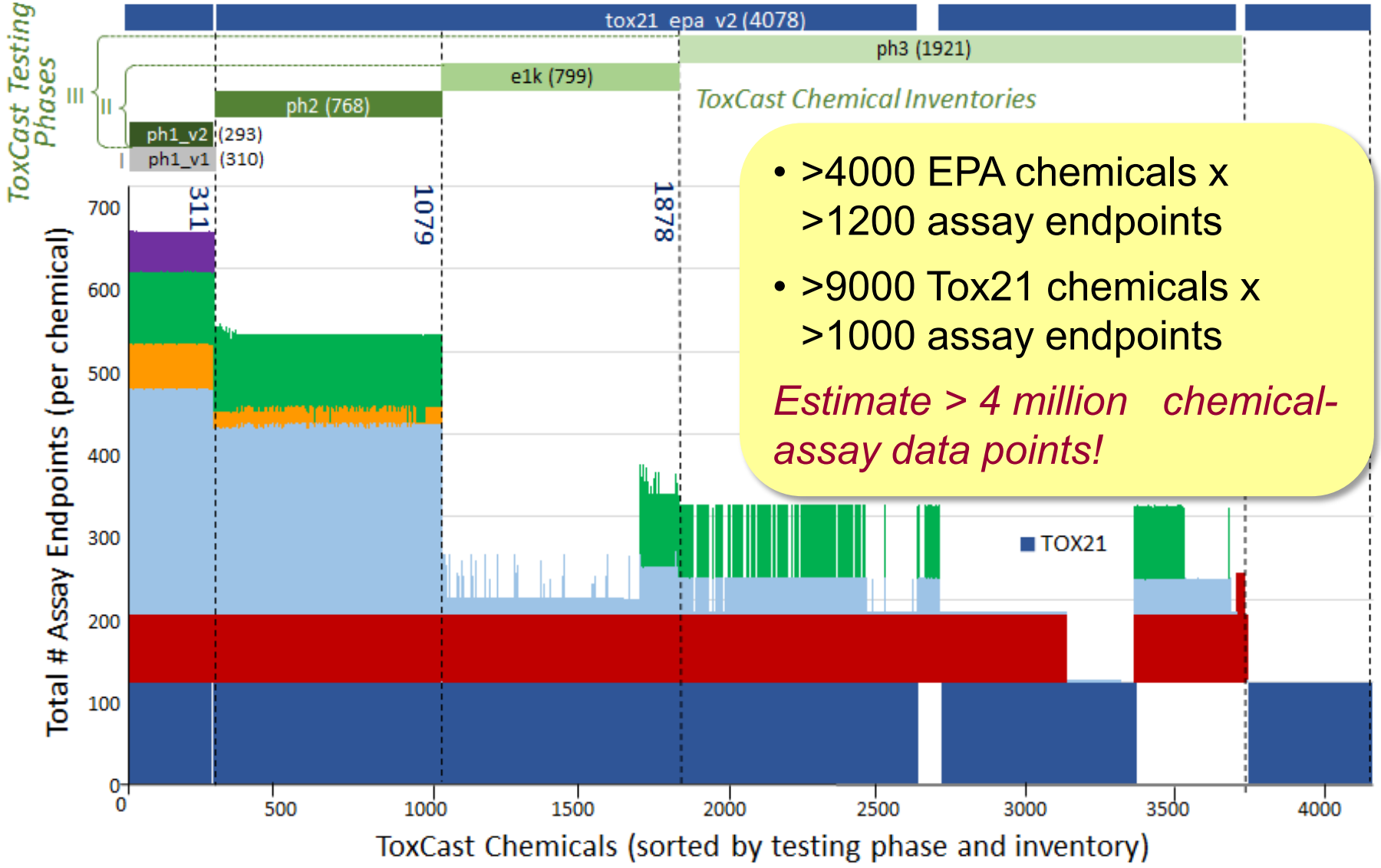
EPA workshop, April 2011

Building a public chemotype "knowledge- base"



United States Environmental Protection Agency

ToxCast HTS data

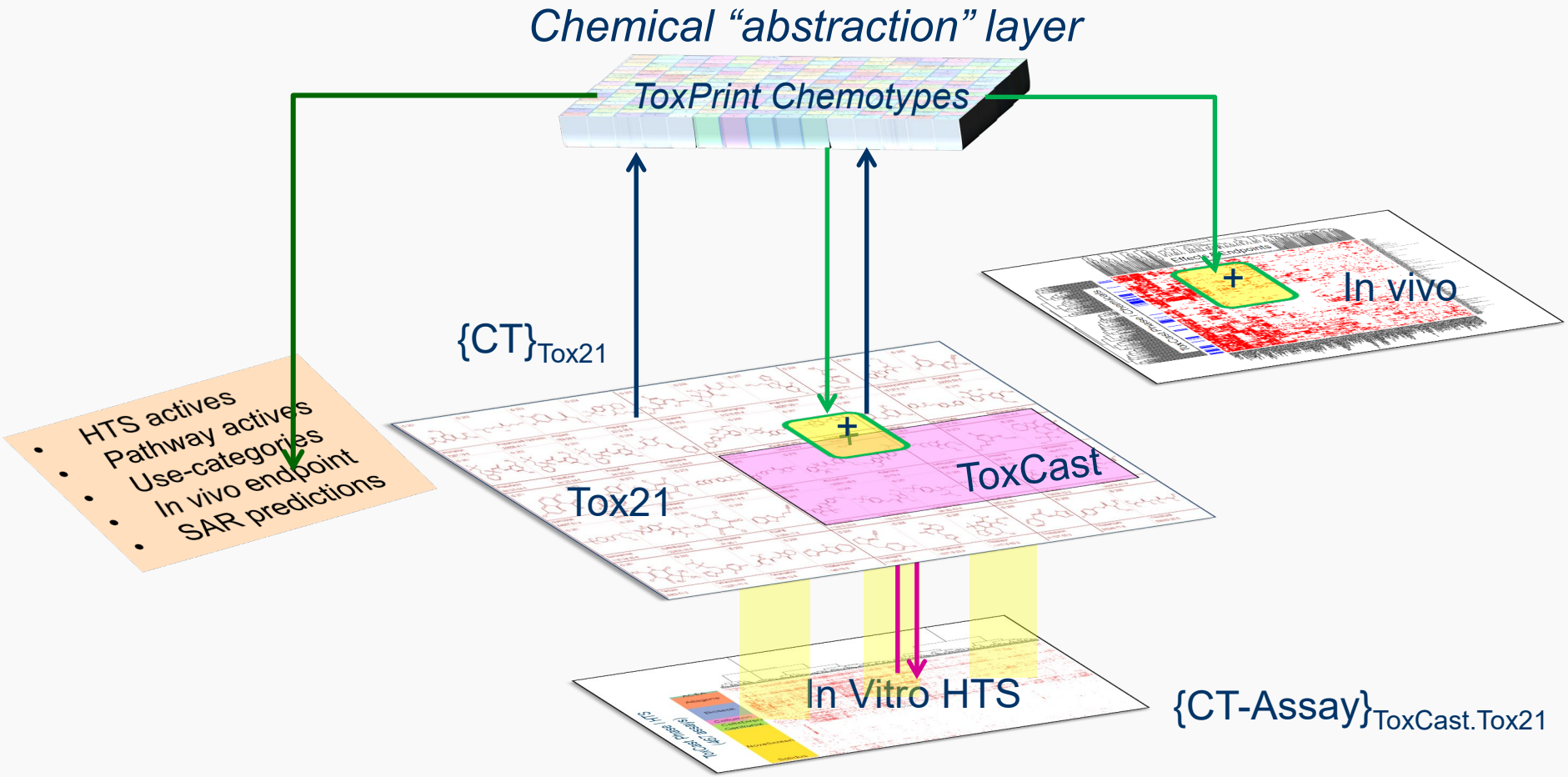


- >4000 EPA chemicals x >1200 assay endpoints
- >9000 Tox21 chemicals x >1000 assay endpoints

Estimate > 4 million chemical-assay data points!

Chemotype-Activity Enrichments

- ✓ Create {CT-Activity} enrichment profiles for any “active” subspace of a test set
- ✓ Focus studies in local CT domains & compare enrichments across data domains



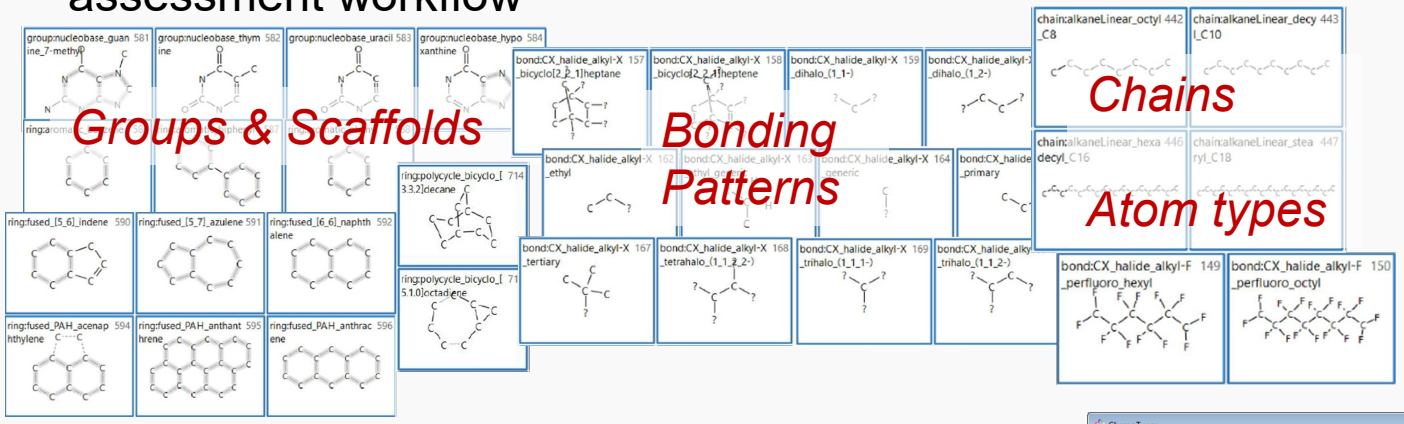
ToxPrints: A Public Set of Chemotypes

EPA United States Environmental Protection Agency

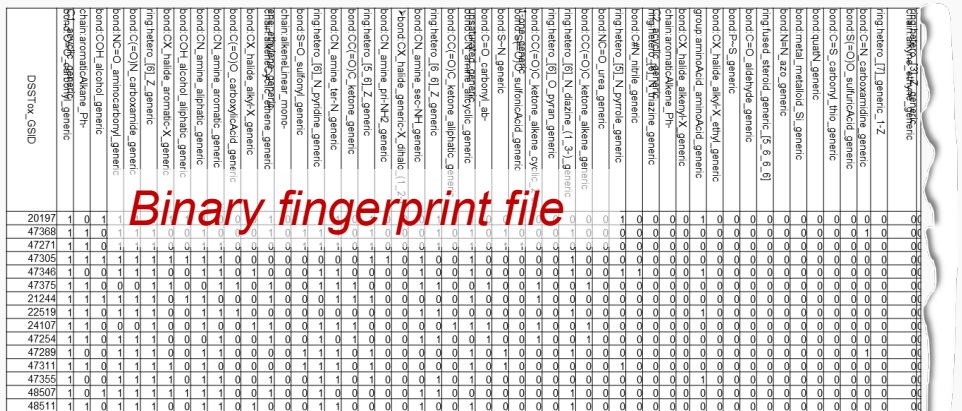
→ *Clear, reproducible means for defining & visualizing local chemistry*

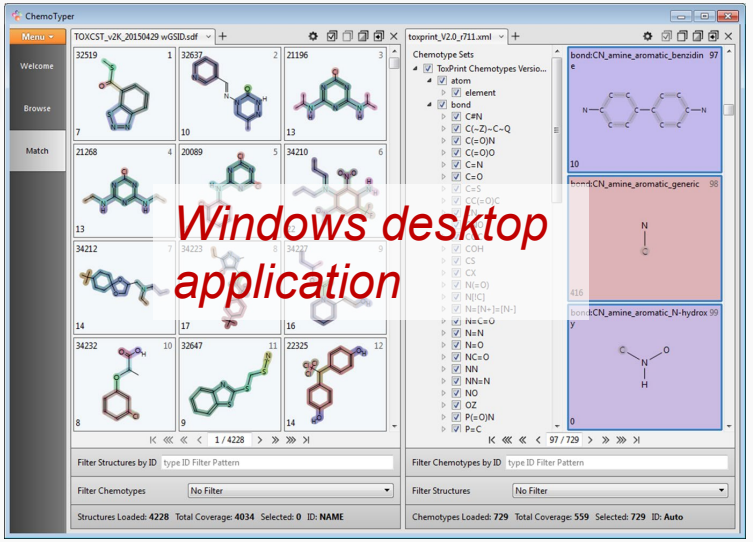
ToxPrints: http://www.toxprint.org

729 features important to EPA & FDA's "chemical exposure" landscape and safety assessment workflow



Chemotyper: http://www.chemotyper.org

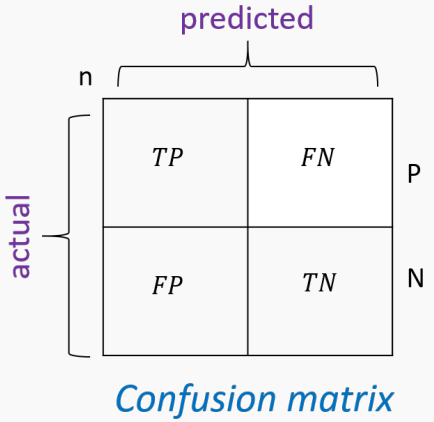




Computing CT-Assay "Enrichments"

Set statistical thresholds & filters for significance to support data-mining objectives:

TP_ID	ToxPrint_CT_name[2]	CT_{Tot}	T_{pos}	F_{pos}	F_{neg}	T_{neg}	Odd's Ratio	Fischer's pval
423	chain:alkaneBranch_t-butyl_C4	41	24	17	294	693	**3.3**	**2.0E-04**
479	chain:aromaticAlkane_Ph-C1-Ph	39	27	12	291	698	**5.4**	**6.5E-07**
303	bond:X[any_!C]_halide_inorganic	28	17	11	301	699	**3.6**	**9.0E-04**



Confusion matrix

TestSet = # Pos + # Neg = # chems tested

CT_{Tot} = total # chems in TestSet w/ CT (Pos or Neg)
TP (T_{pos}) = # Pos in TestSet w/ CT
FP (F_{pos}) = # Neg in TestSet w/ CT
FN (F_{neg}) = # Pos in TestSet w/o CT
TN (T_{pos}) = # Neg in TestSet w/o CT

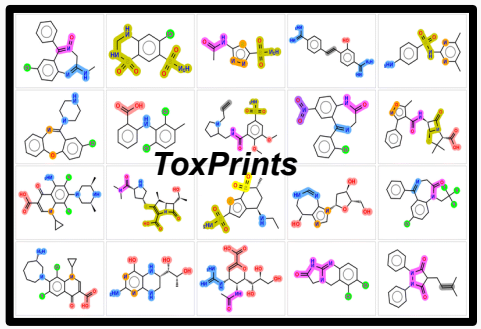
- Odds Ratio ≥ 3, *conveys simple fractional enrichment*
- Fischer's exact p value ≤ 0.05, *takes into account size of dataset*
- T_{pos} (TP) ≥ 3, *require at least 3 chemicals with CT in Positives*

Automated Chemotype-Enrichment Workflow (CTEW)

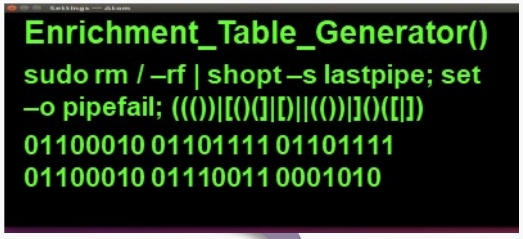
ID	Endpoint
DTXCID50675904	0
DTXCID50675924	0
DTXCID50675929	0
DTXCID50675904	1
DTXCID50675924	1
DTXCID50675929	0
DTXCID50675904	1
DTXCID50675924	0
DTXCID50675929	0
DTXCID50675904	0
DTXCID50675924	0
DTXCID50675929	0

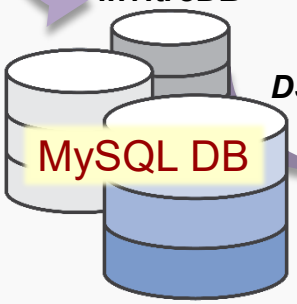
Generate binary "Hitcall Matrix"

Generate fingerprints



Calculate enrichment statistics





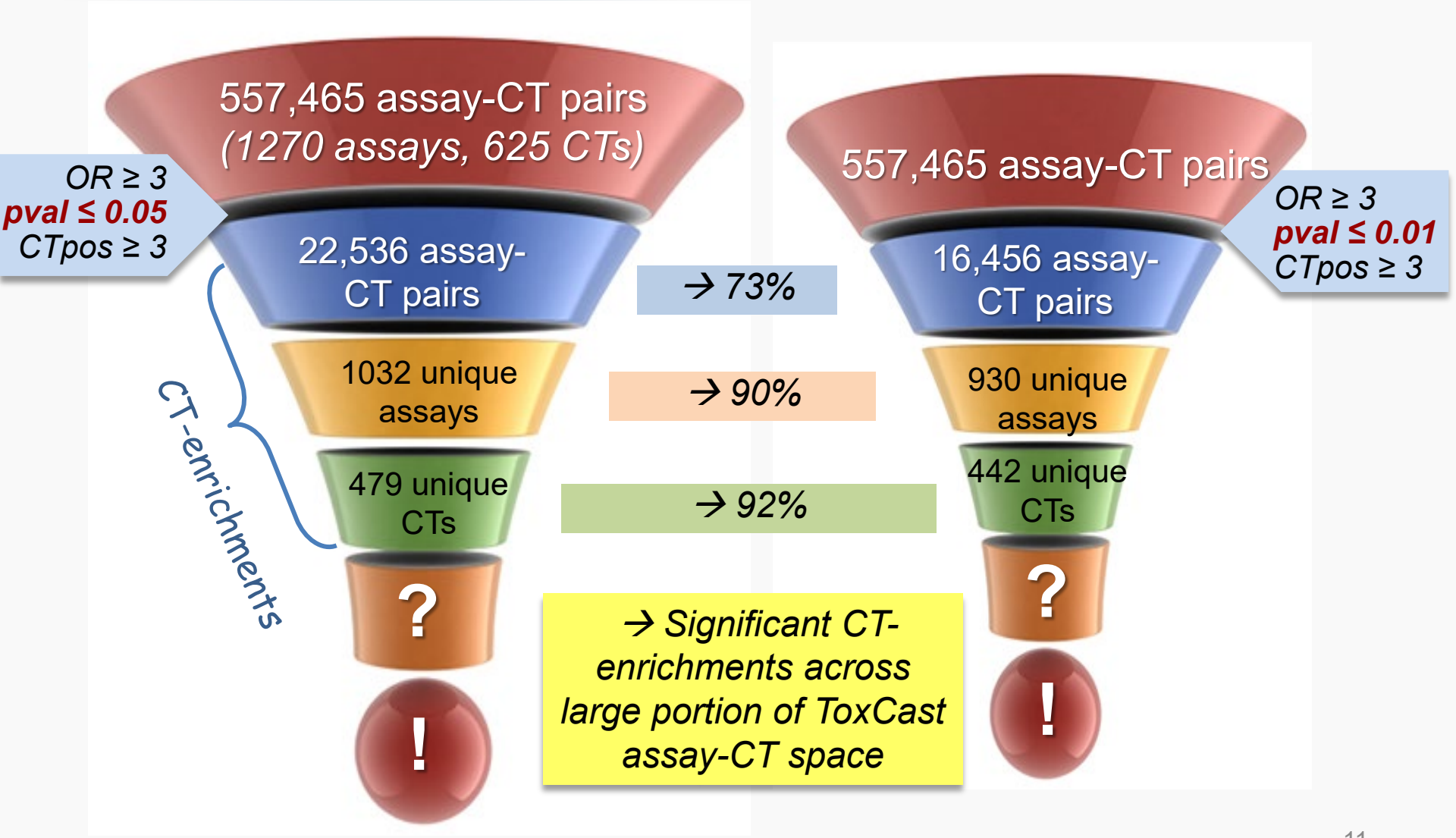
Generate CT-Enrichment Table Results

descriptors_name	label	CT-Tot	TP	FP	FN	TN	BA	OR	P-Val	Inv OR	Inv P-val
Txp-1	atom:element_main_group	2	1	1	1827	5180	0.6196303963661194	2.8352489471435547	0.4536220133304596	0.35270270705223083	0.9320069551467896
Txp-10	bond:C#N_cyano_cyanohydrin	5	1	4	1827	5177	0.46957454085350037	0.7084017395973206	0.7794185876846313	1.4116283655166626	0.6100181937217712
Txp-100	bond:CN_amine_pri-NH2_alkyl	112	25	87	1803	5094	0.4808981418609619	0.8118652701377869	0.8468714356422424	1.2317314147949219	0.21217265725135803
Txp-101	bond:CN_amine_pri-NH2_aromatic	341	84	257	1744	4924	0.4923933148384094	0.9228215217590332	0.7525069713592529	1.0836331844329834	0.28962212800979614
Txp-102	bond:CN_amine_pri-NH2_generic	459	110	349	1718	4832	0.4886806607246399	0.8864842653274536	0.8697083592414856	1.1280516386032104	0.1554924845695495
Txp-103	bond:CN_amine_sec-NH_alkyl	199	44	155	1784	5026	0.4795689284801483	0.7997395992279053	0.9174827933311462	1.2504069805145264	0.11147873103618622
Txp-104	bond:CN_amine_sec-NH_aromatic	151	36	115	1792	5066	0.4885549545288086	0.8849766850471497	0.7643690705299377	1.129973292350769	0.298209011554718
Txp-105	bond:CN_amine_sec-NH_aromatic_aliphatic	97	22	75	1806	5106	0.4827597141265869	0.82932448387146	0.8107256293296814	1.2058006525039673	0.2608259320259094
Txp-106	bond:CN_amine_sec-NH_generic	253	58	195	1770	4986	0.48362982273101807	0.8378617763519287	0.8932034373283386	1.1935142278671265	0.13699662685394287
Txp-107	bond:CN_amine_ter-N_aliphatic	449	161	288	1667	4893	0.5522294044494629	1.6408655643463135	0.00000143775969263515 44	0.6094344258308411	0.9999991655349731
Txp-108	bond:CN_amine_ter-N_aromatic	167	66	101	1762	5080	0.568841278553009	1.8839976787567139	0.00008360712672583759	0.5307862162590027	0.9999573826789856

Store CTEW results in DB

Slide courtesy of Ryan Lougee

CTEW applied to all ToxCast assays in *invitroDB (Dec-2017)*



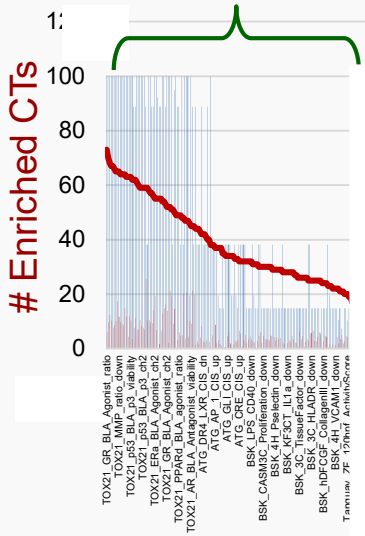
Assays x Enriched CT count

Enriched CT-counts across 1032 assays

*Random Forest models based on ToxPrint CT descriptors, validated using independent Test Set & Y-randomization, with Training (100A,100I) & Test (25A,25I) Set minimums (J. Fitzpatrick)

"Global" QSAR models* for 167 ToxCast Assays

Avg % Median BA = 0.67



Tested Chems (scaled)

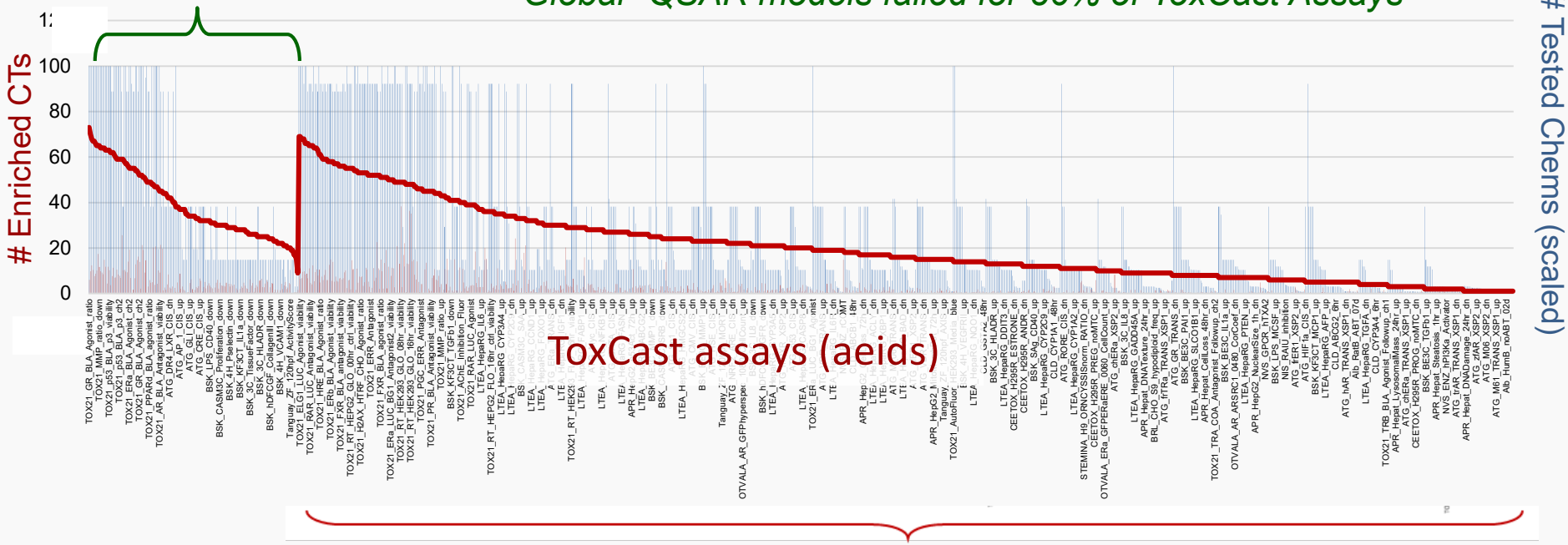
Assays x Enriched CT count

Enriched CT-counts across 1032 assays

*Random Forest models based on ToxPrint CT descriptors, validated using independent Test Set & Y-randomization, with Training (100A,100I) & Test (25A,25I) Set minimums (J. Fitzpatrick)

"Global" QSAR models* for 167 ToxCast Assays
Avg % Median BA = 0.67

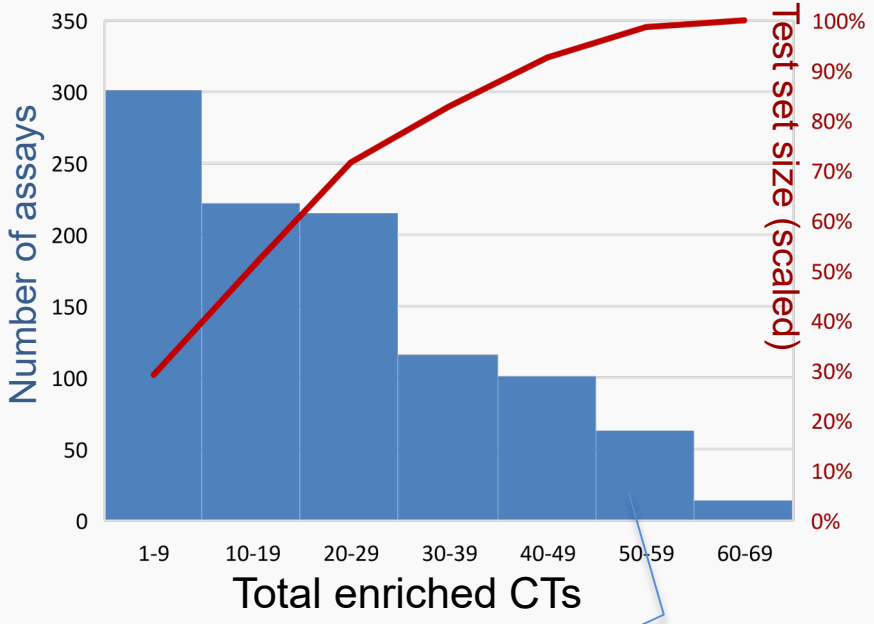
"Global" QSAR models failed for 86% of ToxCast Assays



Significant CT-enrichments across ToxCast assay space considered "unmodelable" by traditional "global" QSAR methods

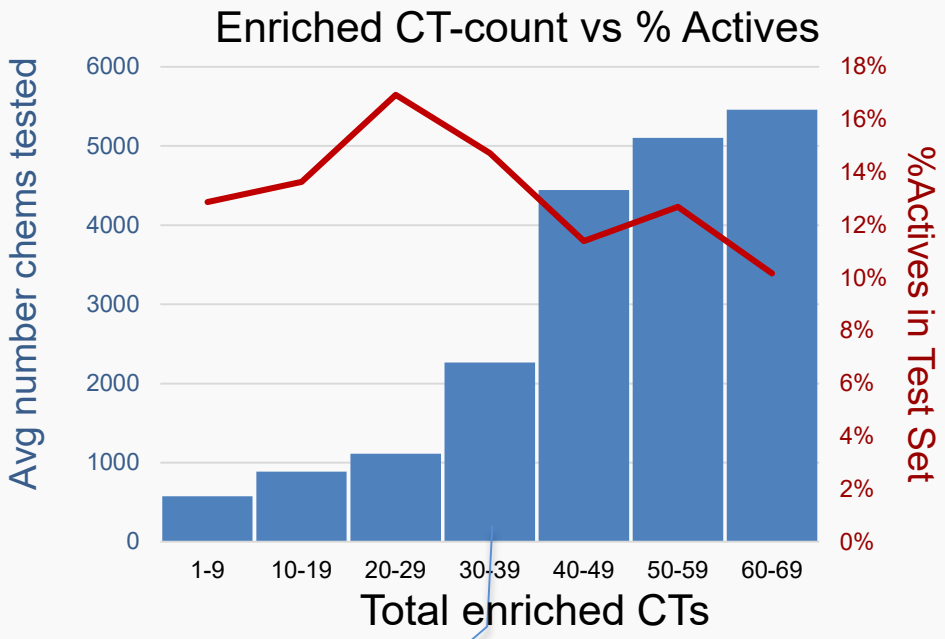
Enriched CT-count per assay trends (1032 assays)

Does # enriched CTs depend on test set size? YES!



63 assays enriched with 50-59 CTs

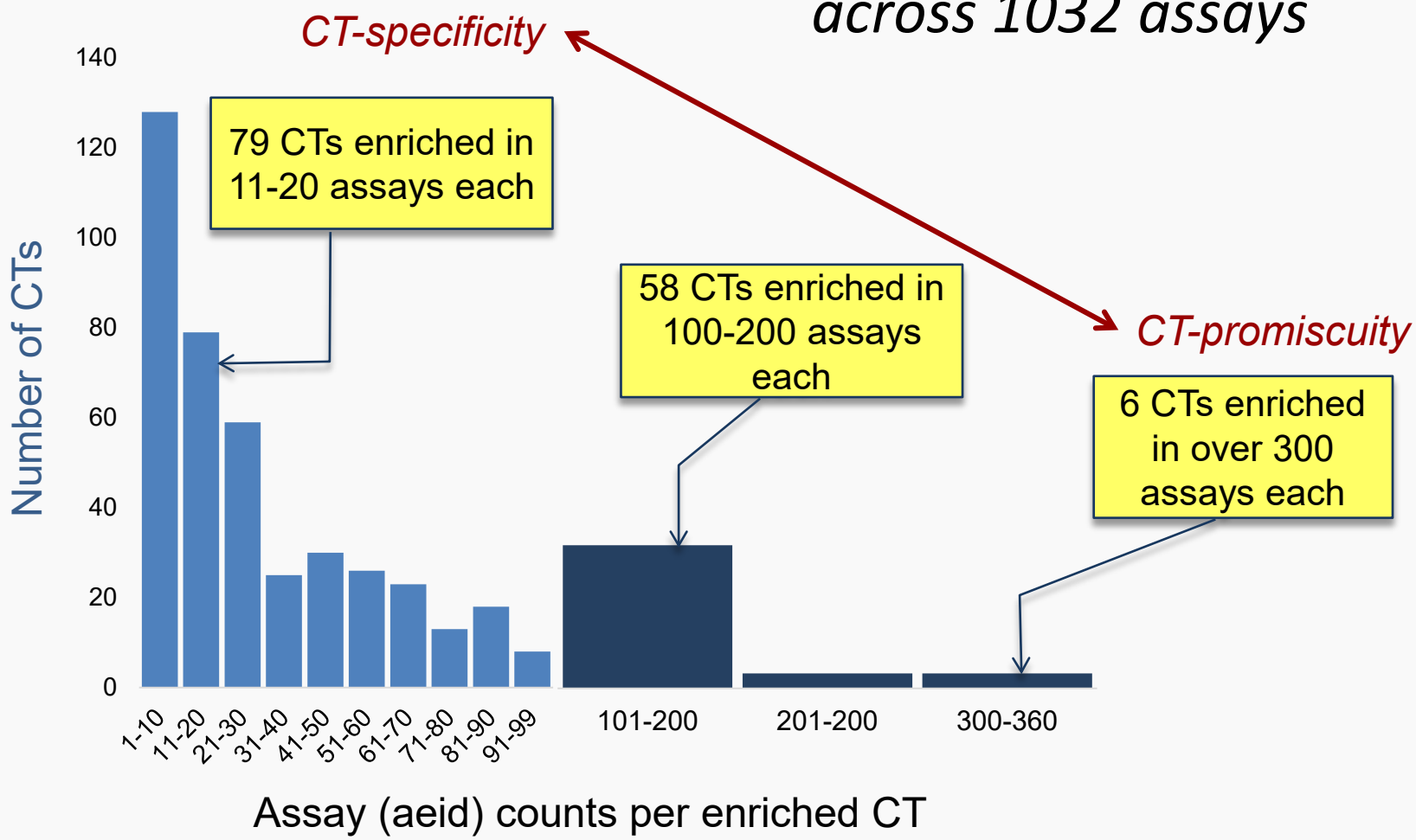
Does # enriched CTs depend on %Actives? Not so much.



30-39 CTs enriched in assays with avg Test Set size 2264 chems

Assay-counts per enriched CT

Enriched CT-counts across 1032 assays



Tuning biology to pick up chemical "signal"

CT Enrichments of Time-series* "Assays"

aenm_870unique	TP_Count	%actives	tested.chnm
APR_Hepat_Apoptosis_1hr_up	1	3.2	310
APR_Hepat_Apoptosis_24hr_up	15	14.8	310
APR_Hepat_Apoptosis_48hr_up	6	16.8	310
APR_Hepat_DNADamage_1hr_up	8	3.9	310
APR_Hepat_DNADamage_24hr_up	7	18.4	310
APR_Hepat_DNADamage_48hr_up	14	20.0	310

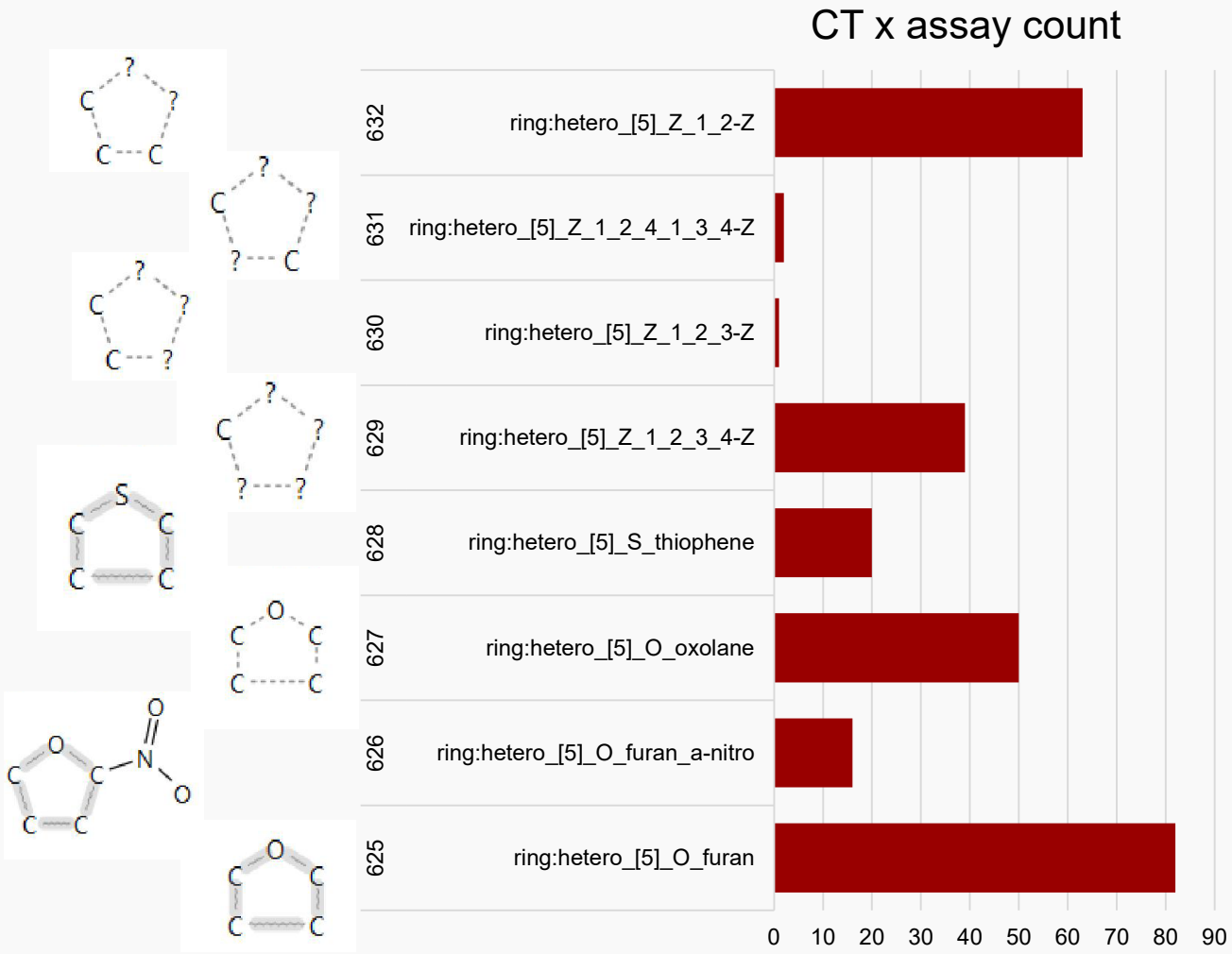
** 95 Assays out of 1032 total are within time-series groups*

"Tuning" biology to increase CT enrichments → increases biological-chemical "signal"

- How are {CT-assay} enrichments affected by activity threshold assumptions?
- How are {CT-assay} enrichments affected by cell toxicity ("burst") filters? (R. Lougee)

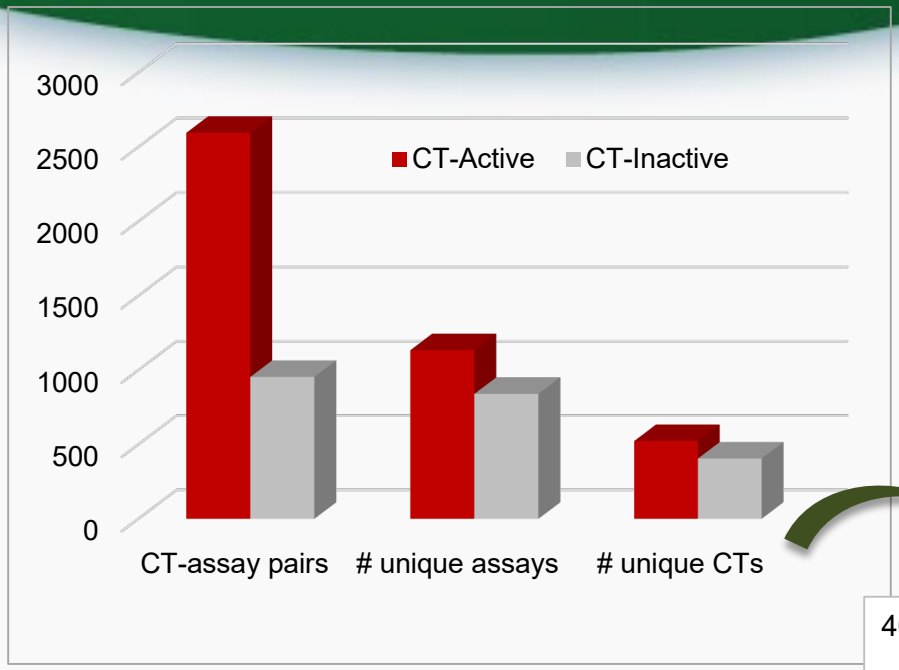
Tuning chemistry to pick up biological "signal"

Assay Enrichments of chemistry-series groups



"Tuning" chemistry specificity to increase chemical-assay "signal"

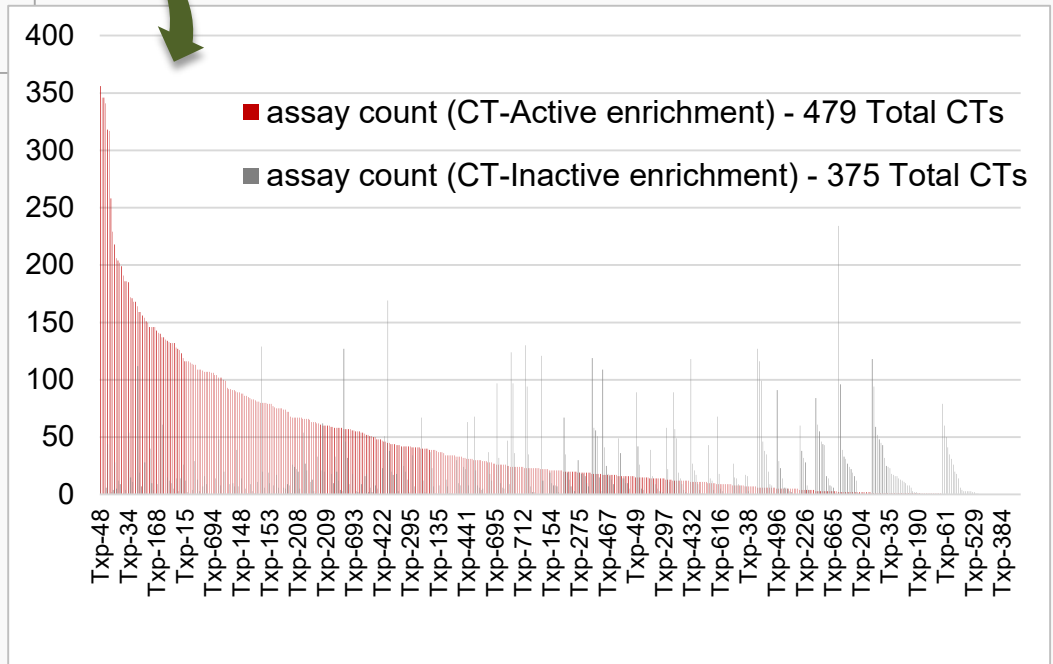
What about the "Inactives"?



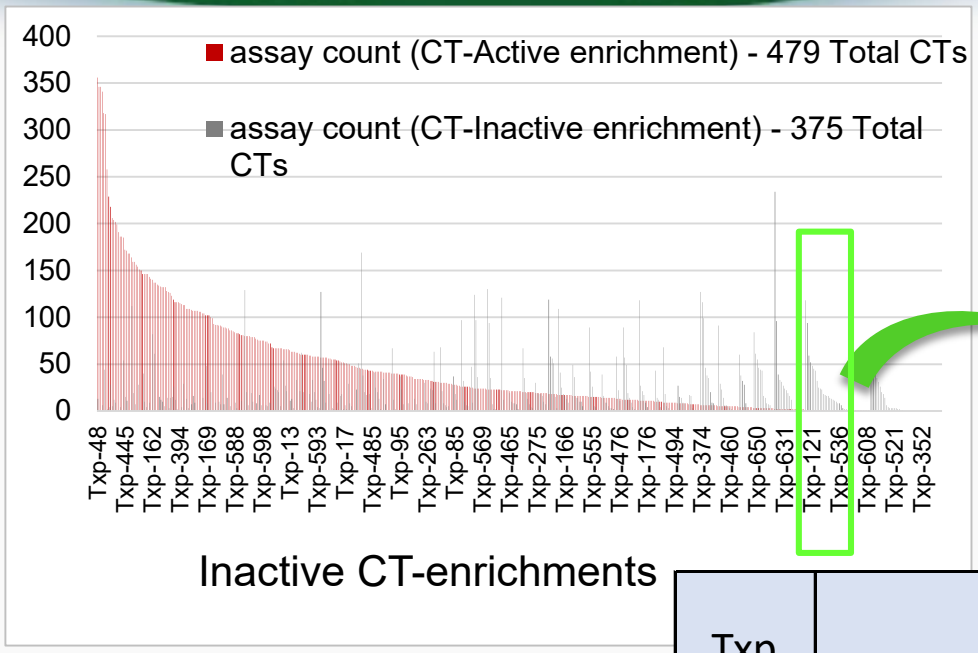
More "CT-signal" in Actives, but significant amount of signal in Inactives

Inactive CT-enrichments

- Span ToxCast assay space
- More likely to occur in assays with fewer Active CT-enrichments
- May be due to several factors:
  - *True inactivity*
  - *Assay artifacts*
  - *QC failure*



Top 10 enriched CT-Inactives (skewed from actives)

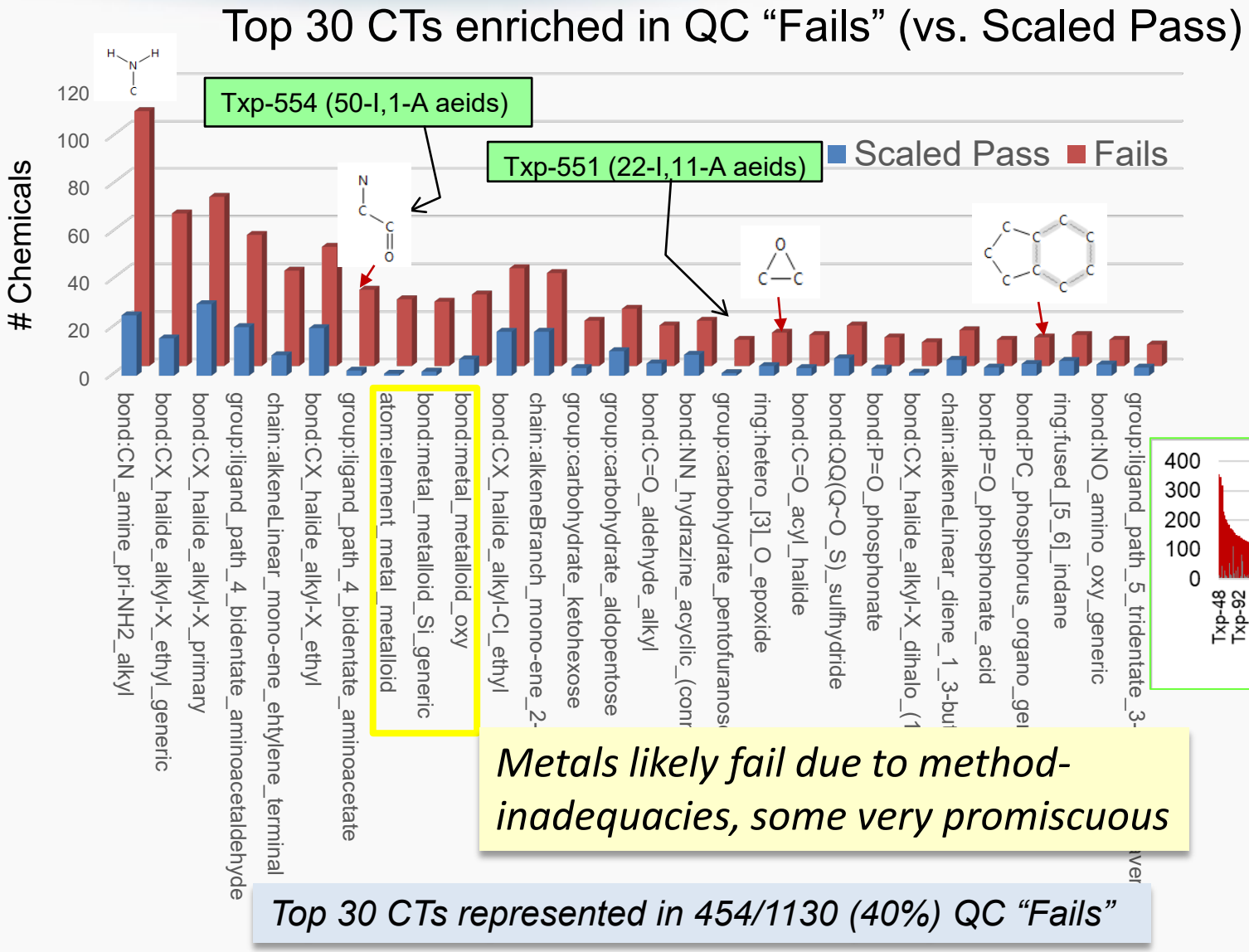


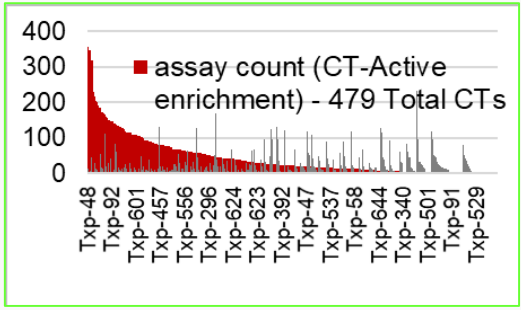
CT-Inactive >75 assays, CT-active <10 assays

- *True inactivity?*
- *Assay artifacts?*
- ***QC failure?***

Txp	CT Name	assay count (CT-Active enrichment)	assay count (CT-Inactive enrichment)
Txp-101	bond:CN_amine_pri-NH2_aromatic	3	84
Txp-145	bond:CX_halide_alkyl-Cl_ethyl	0	79
Txp-260	bond:P~S_generic	1	94
Txp-362	bond:metal_metalloid_oxy	2	96
Txp-372	bond:metal_metalloid_Si_organo	6	99
Txp-374	bond:metal_metalloid_Si _oxy	6	116
Txp-496	chain:oxy-alkaneLinear_ethyleneOxide_EO1(O)	5	91
Txp-497	chain:oxy-alkaneLinear_ethyleneOxide_EO2	6	127
Txp-607	ring:hetero_[4]_N_beta_lactam	1	118
Txp-663	ring:hetero_[6]_N_triazine_(1_3_5-)	2	234

ToxPrints enriched in Fails

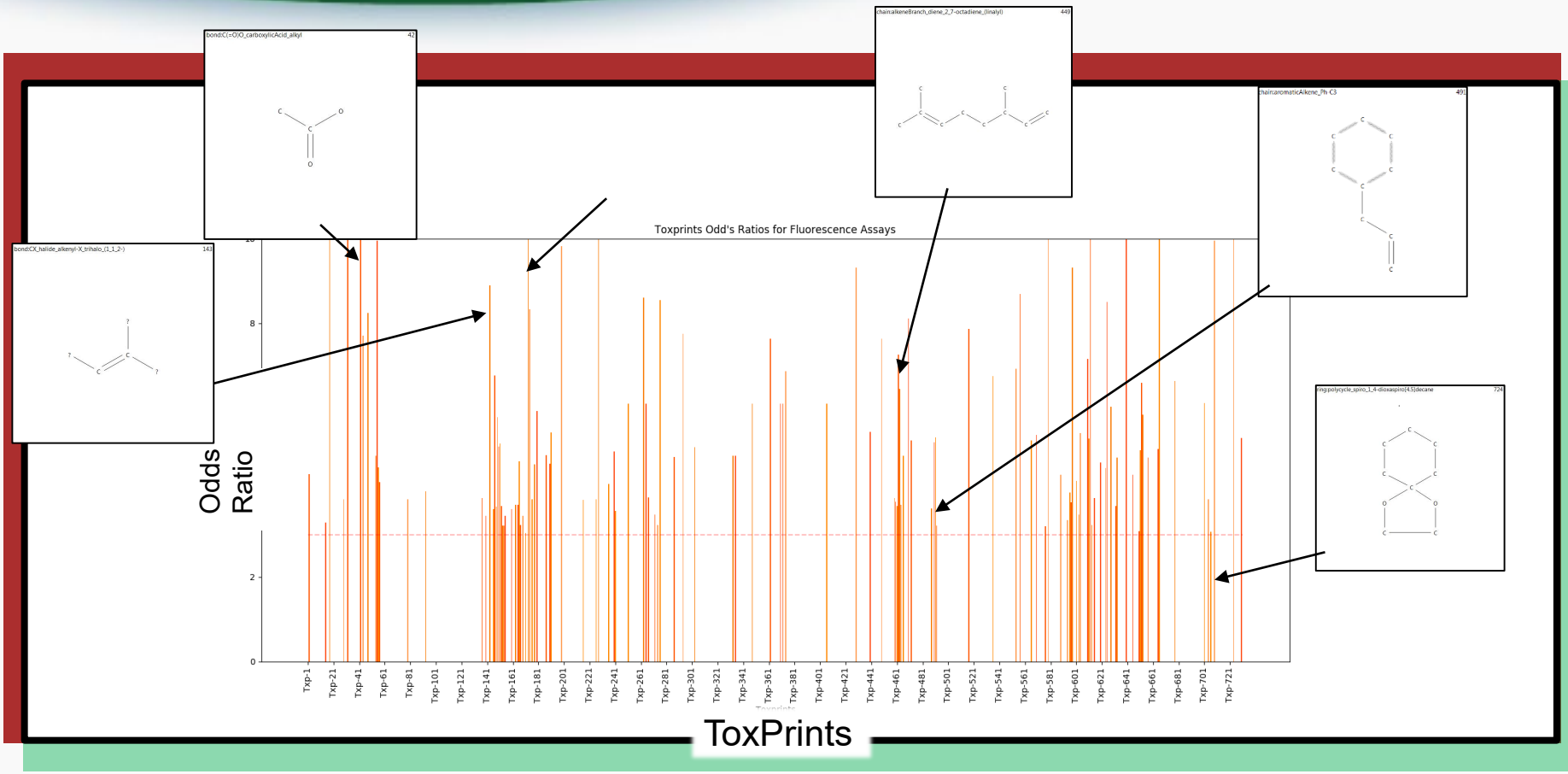




Metals likely fail due to method-inadequacies, some very promiscuous

Top 30 CTs represented in 454/1130 (40%) QC "Fails"

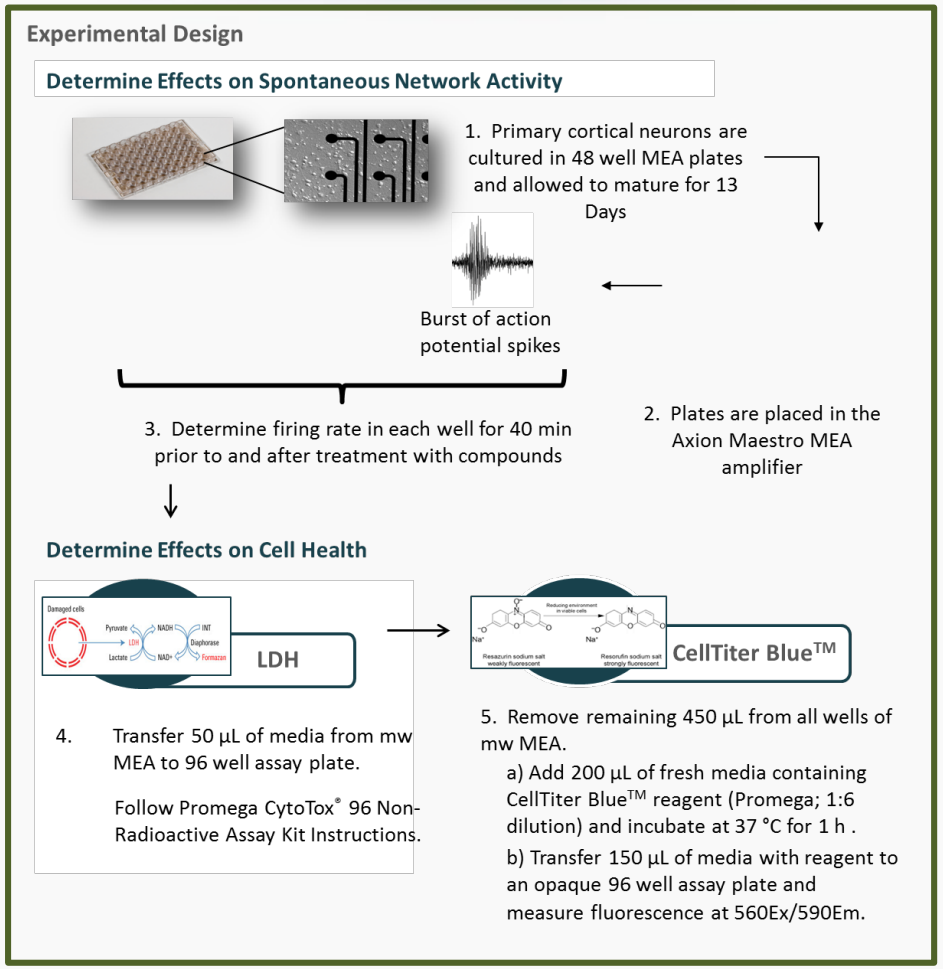
CT Enrichments in Fluorescence Assays



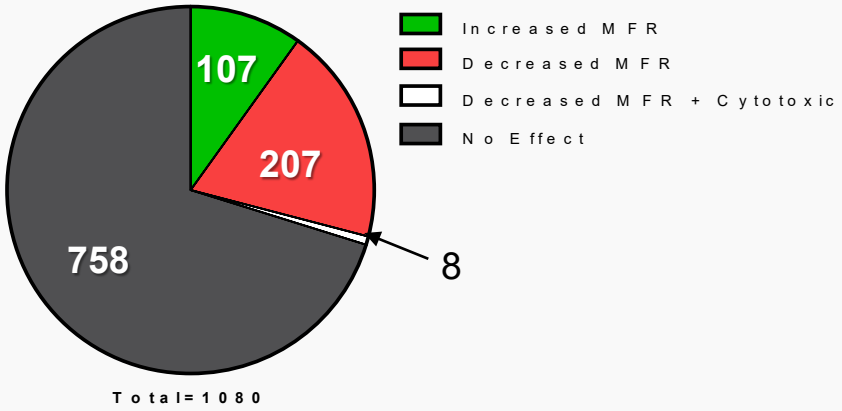
CT's enriched (much more likely to be active) in "Fluorescence Assays" than in remaining assays

Slide courtesy of Ryan Lougee

Multi-electrode array (MEA) neurotoxicity assay



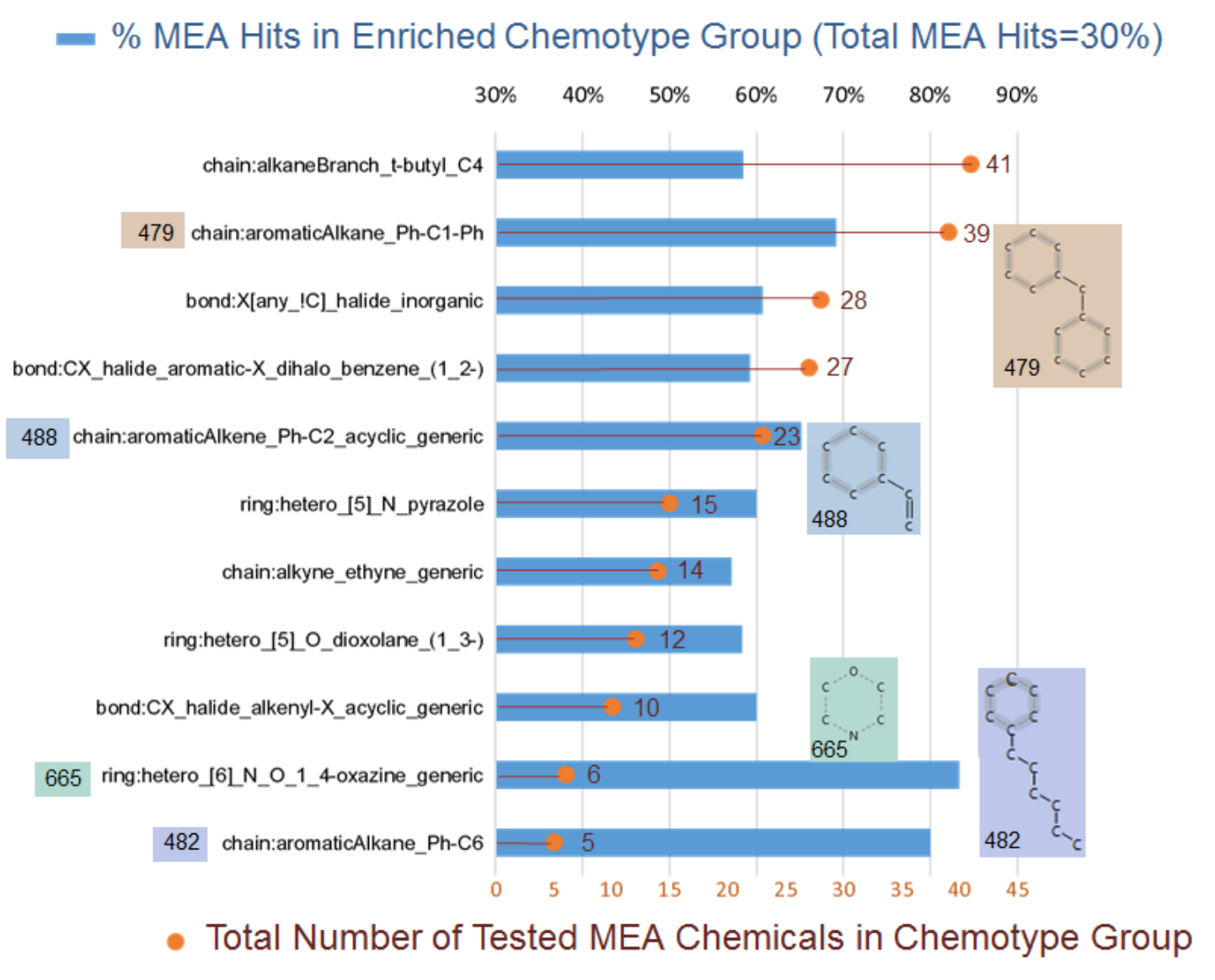
- Screened 1080 ToxCast Phase II chemicals
- Activity measured as Mean Firing Rate (MFR) above or below threshold



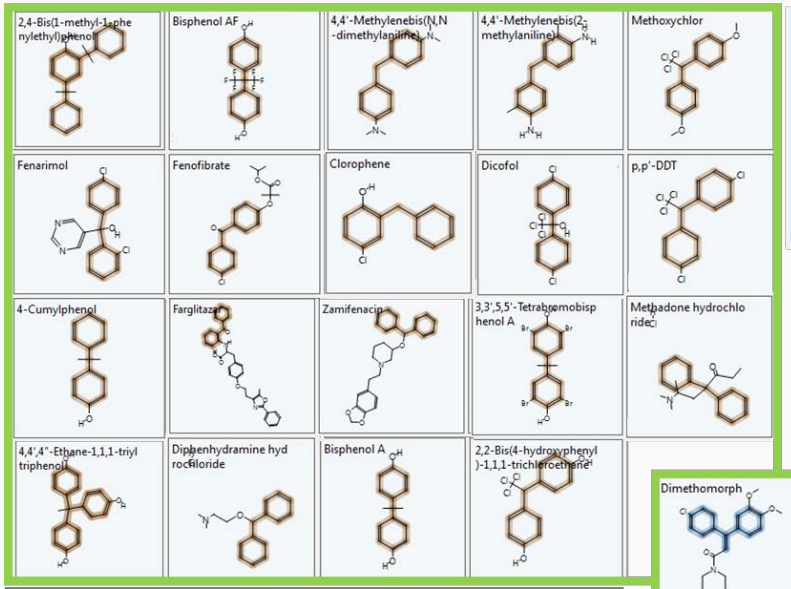
- 314 total “Actives” or Hits
- 758 “Inactives” or No-hits (8 cytotoxic)
- Total Tested = 1080 (30% hit rate)

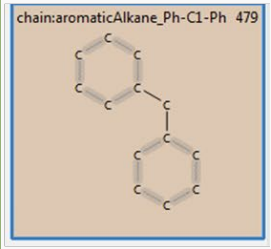
Strickland et al, Arch. Toxicol., 92:487-500, 2018

MEA CT-Enrichment Analysis:
Overall coverage



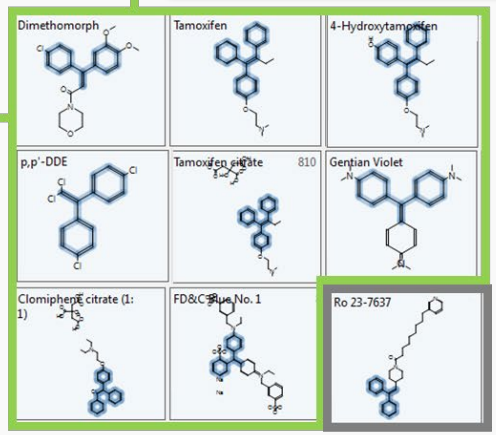
MEA CT-Enrichment Analysis:
Explore SAR within local CT domains

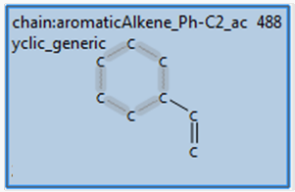


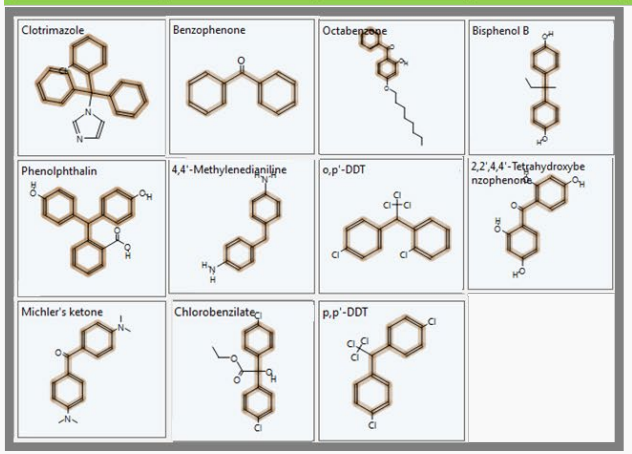


OR=5.4 (27/39 Hits)

OR=18.3 (8/9 hits)







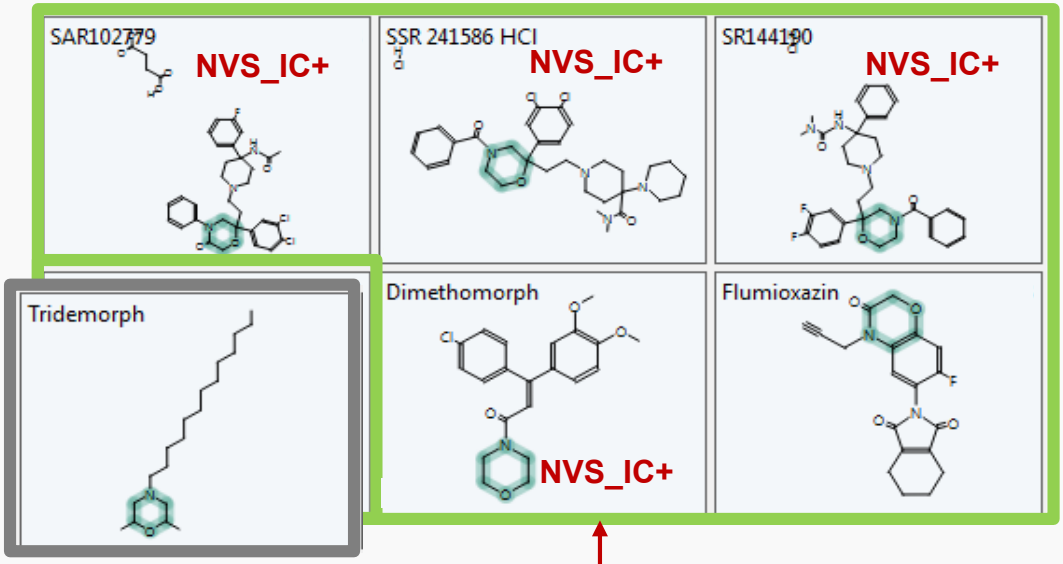
MEA Hits

MEA Non-Hits

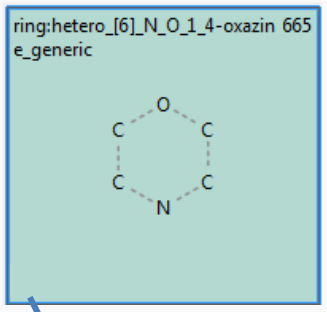
MEA CT-Enrichment Analysis:
Use CTs to build biological linkages

MEA Hits

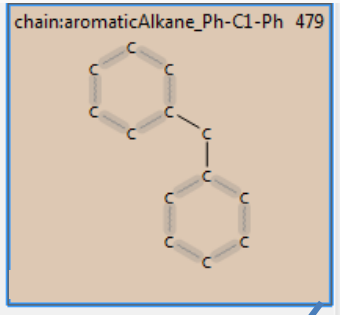
MEA Non-Hits



OR=11.3 (5/6 Hits)



OR=5.4 (27/39 Hits)



ToxCast assays enriched with hits in CT665 (OR > 4)

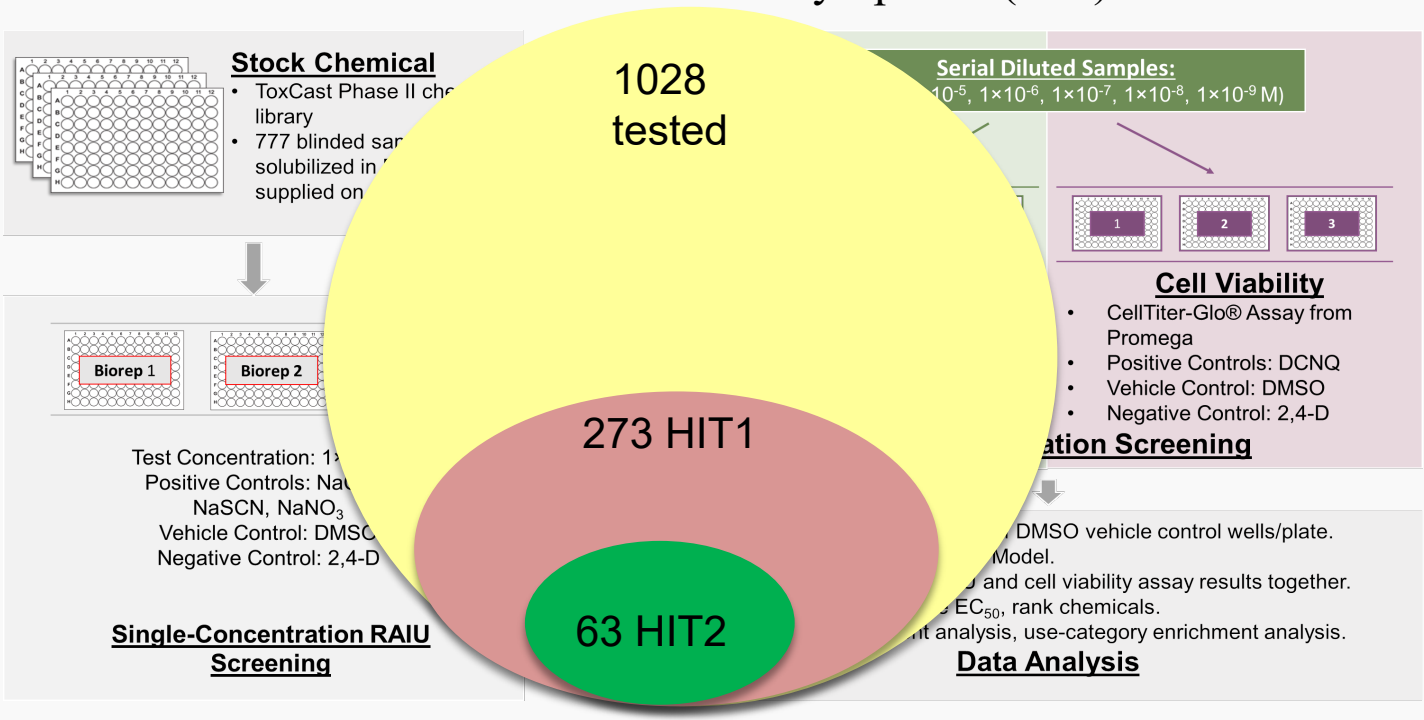
- NVS_GPCR_hOpiate_mu
- NVS_GPCR_rOpiate_NonSelective
- NVS_GPCR_rOpiate_NonSelectiveNa
- NVS_IC_rCaBTZCHL
- NVS_IC_rCaDHPRCh_L
- NVS_IC_rNaCh_site2

4/5 MEA hits with CT665 also MEA hits in Ion Channel (IC) assays

16/27 MEA hits in CT479 also hits in NVS_IC assays

NIS Inhibition Assay
(with Jun Wang, Susan Laws et al.)

High-Throughput Screening and Chemotype-Enrichment Analysis of ToxCast Phase II Chemicals Evaluated for Human Sodium-Iodide Symporter (NIS) Inhibition



1028 ToxCast Phase II chemicals screened

→ 27% positive at 20% threshold in single screen …**Hit1**

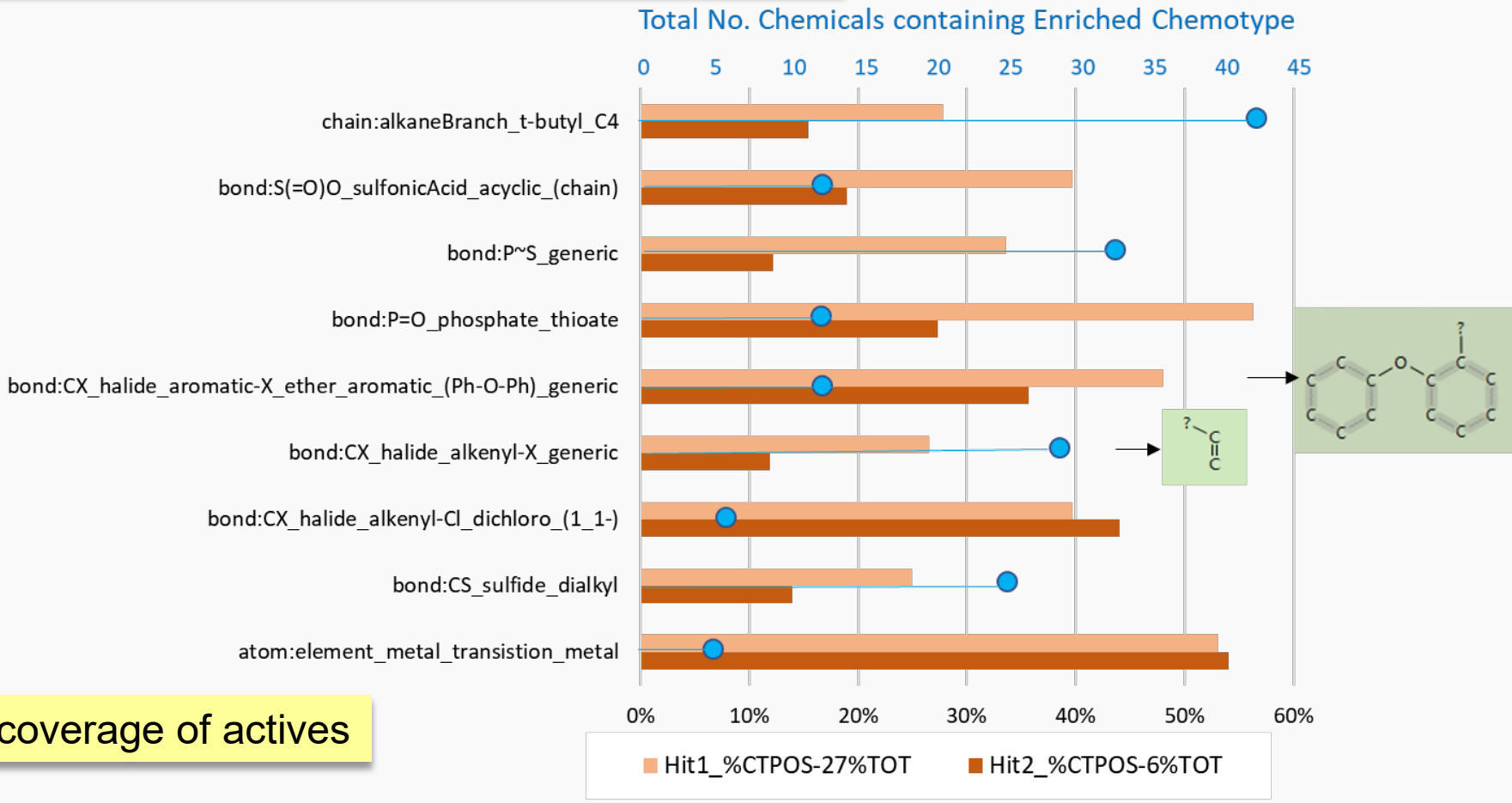
→ 6% positive in multiscreen, cell viability filtered results …**Hit2**

Wang et al., Environ. Int, 2019, 126:377-386

NIS Activation Assay

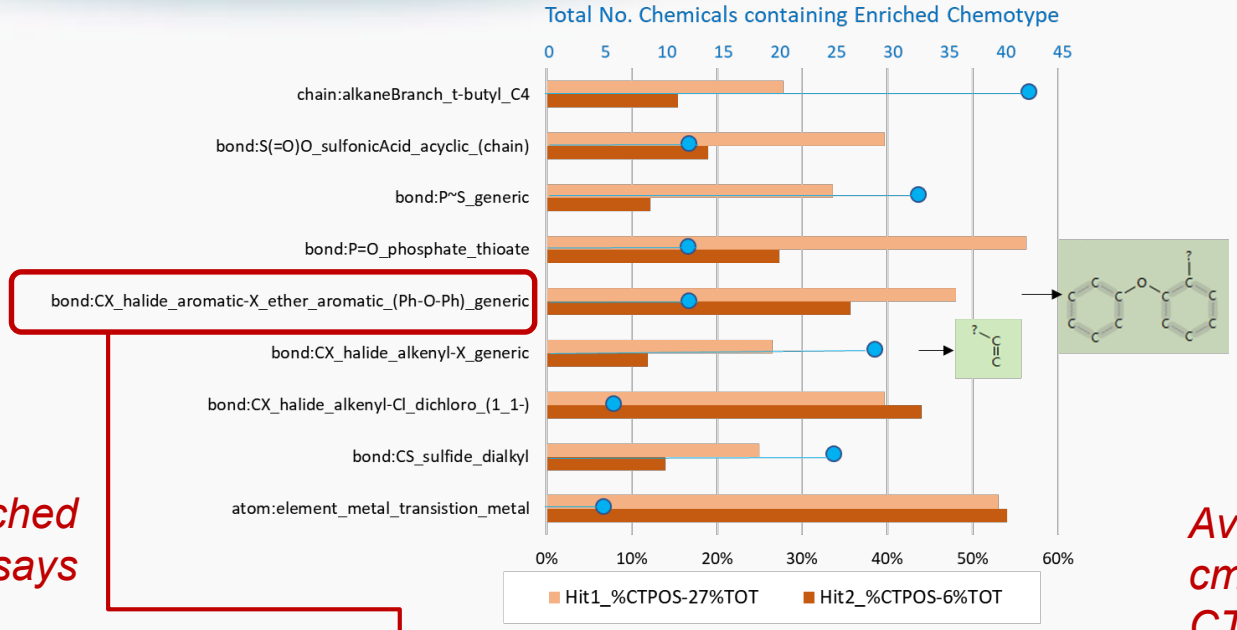
Overall coverage

Top 9 CT's enriched in both Hit1 and Hit2 datasets



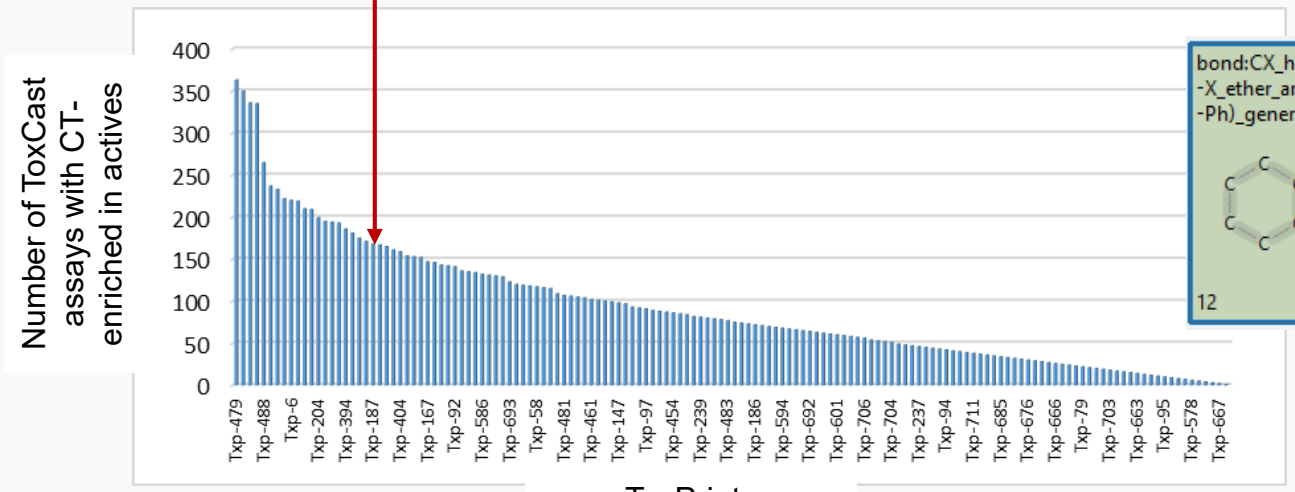
60% coverage of actives

Linking global inferences to local CT enrichment results



CT Enriched in 167 assays

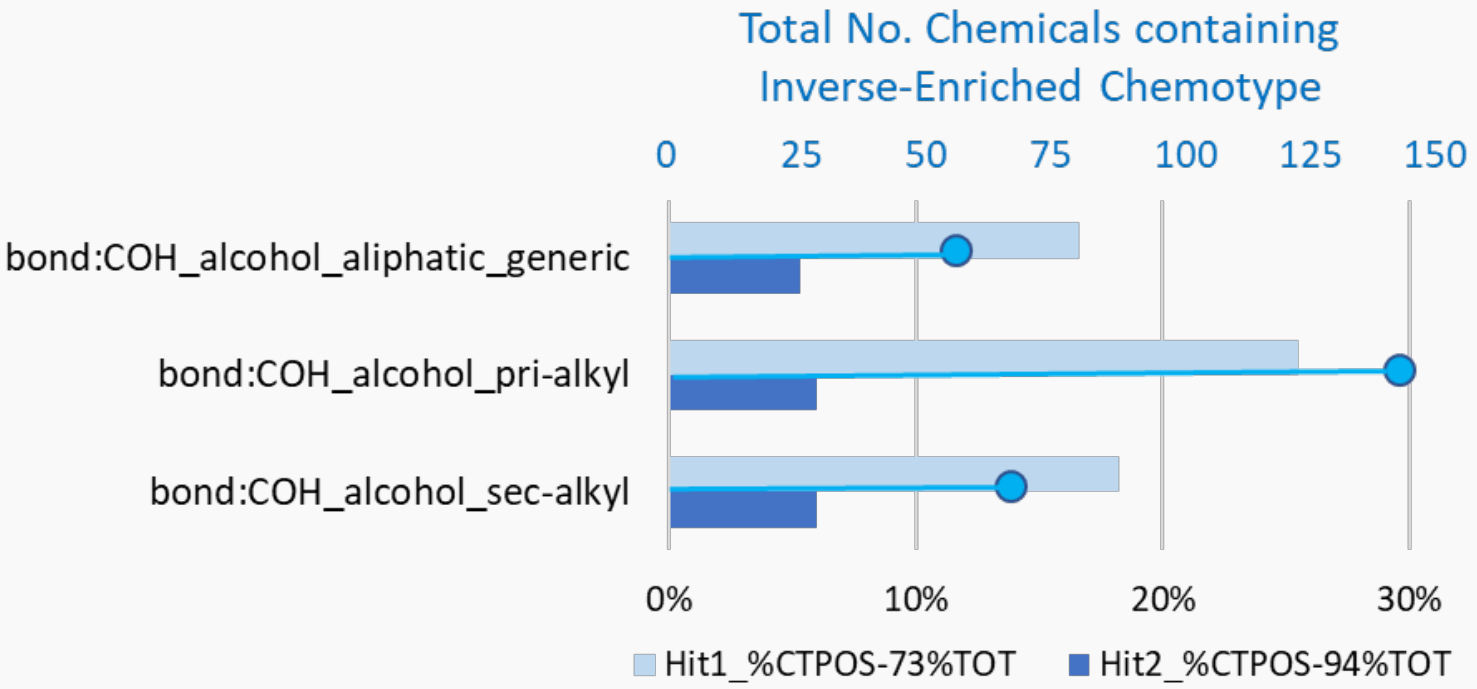
Avg Hit rate for 12 cmpds containing CT = 12%



NIS Activation Assay

Inverse activity enrichments

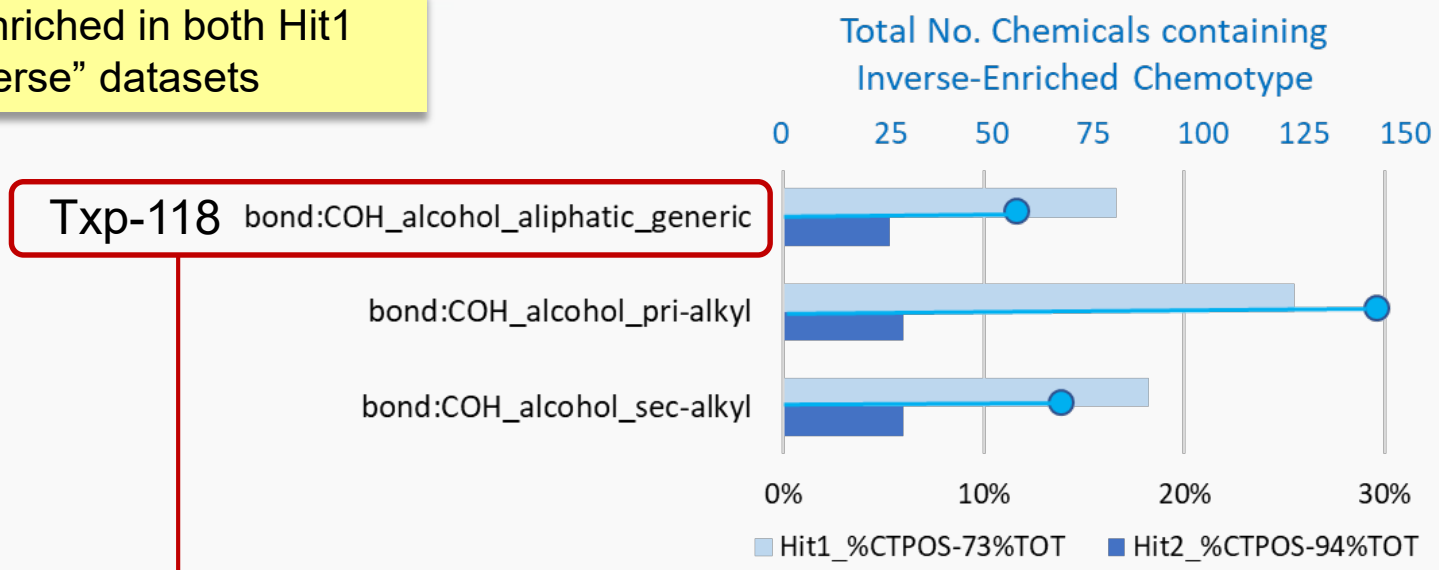
Top 3 CT's enriched in both Hit1 and Hit2 "Inverse" datasets

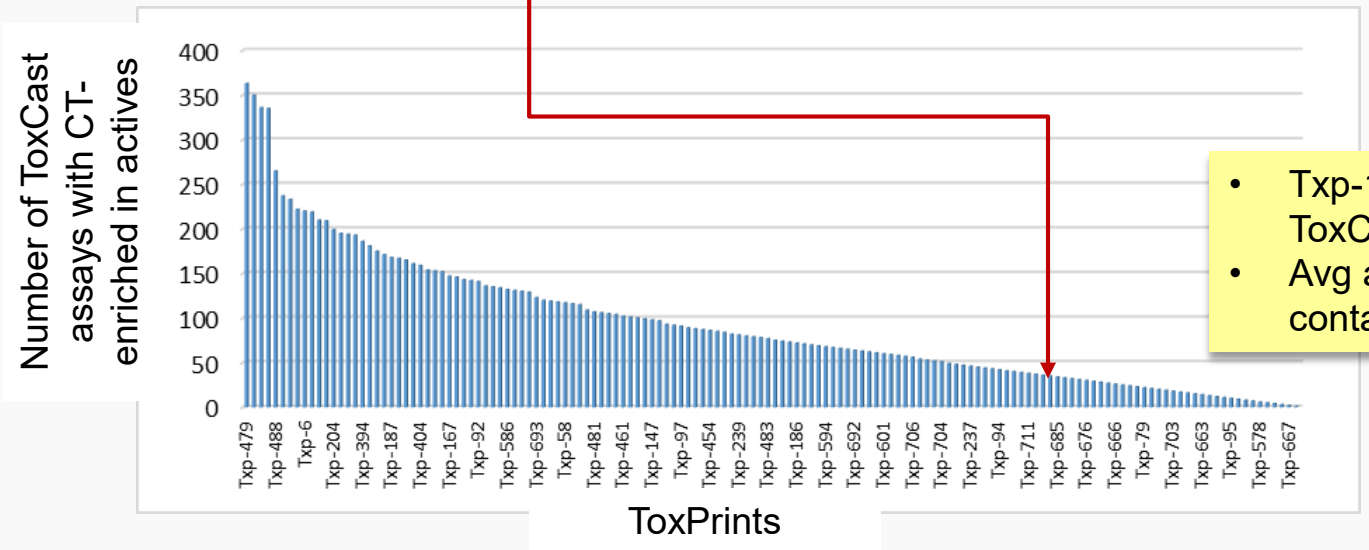


NIS Activation Assay

Inverse activity enrichments

Top 3 CT's enriched in both Hit1 and Hit2 "Inverse" datasets

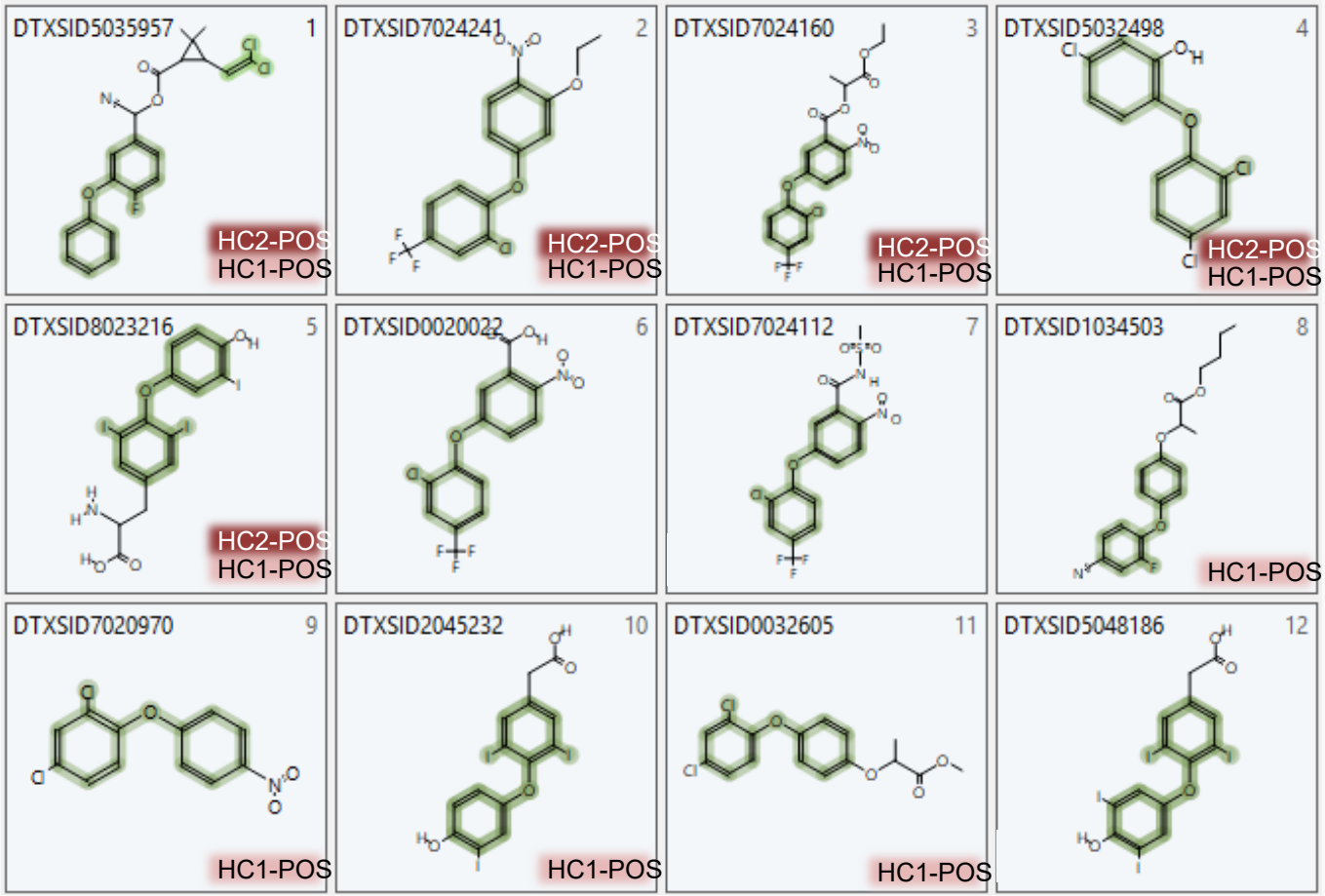


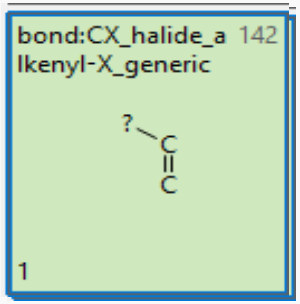


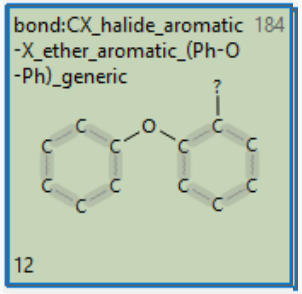
- Txp-118 enriched in 42 ToxCast assays
- Avg assay hits per chemical containing Txp-118 is 8%

NIS Activation Assay

Exploring activity within CT domain







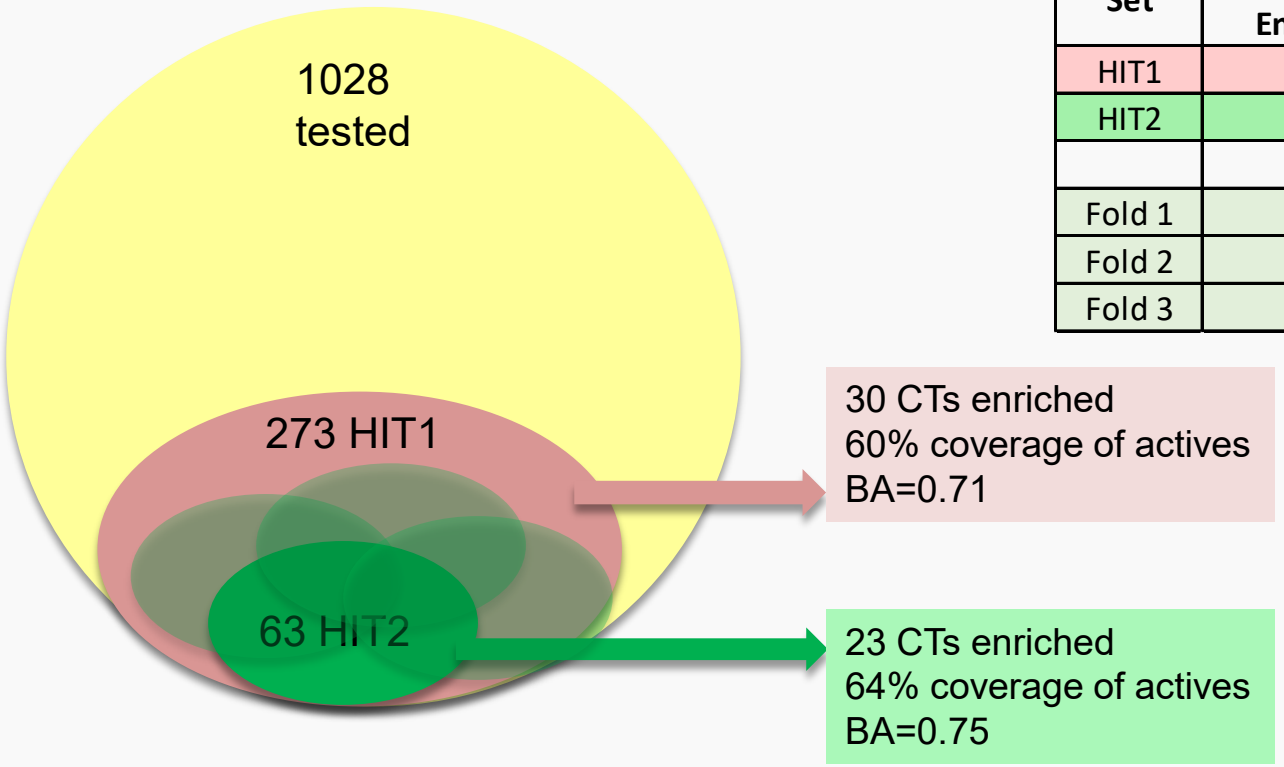
? (halogen) = F, Cl, I

- What distinguishes inactives in CT-subspace?
- What distinguishes the multiscreen Hit2 actives (HC2) from the single screen Hit1 (HC1) actives?

Random folds to select Hit2 set

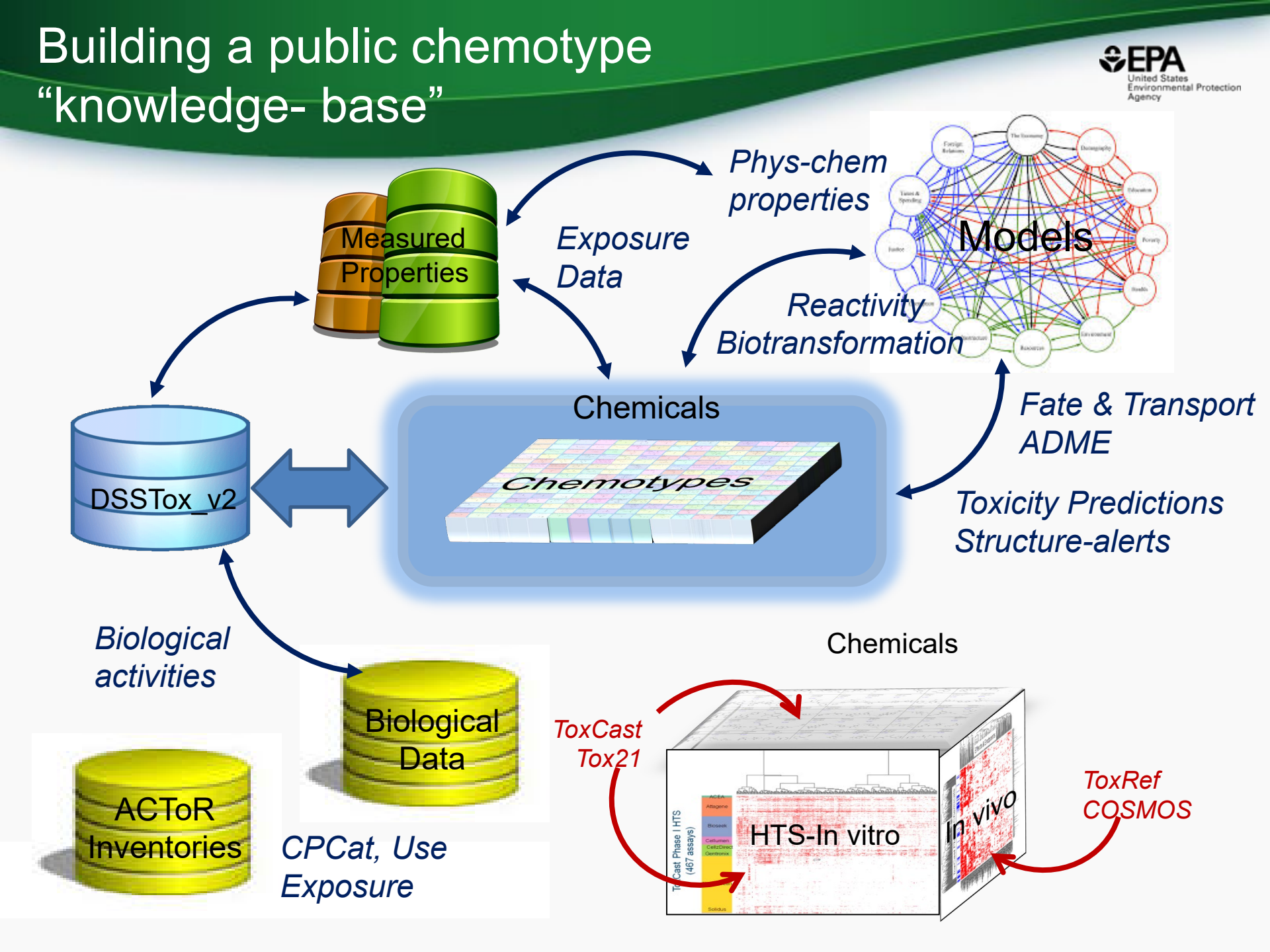
→ 3 Random folds to select 63 HIT2 candidates from HIT2 set

Set	Total CT-Enrich	BA (full model)
HIT1	30	0.71
HIT2	23	0.75
Fold 1	15	0.71
Fold 2	19	0.71
Fold 3	16	0.71



HIT1: Single Screen
HIT2: Multiscreen + cell viability filters

Building a public chemotype "knowledge- base"



Acknowledgements:

- EPA NCCT Chemistry Team
  - Ryan Lougee (CTEW workflow coding)
  - Chris Grulke, Antony Williams (DSSTox, Chemotypes, Dashboard)
  - Jeremy Fitzpatrick (Global QSARs)
- ToxCast case studies
  - Tim Shafer, Jenna Strickland, Susan Laws, Jun Wang, Tammy Stoker
- ToxPrint Chemotypes
  - Chihae Yang, Jim Rathman & colleagues, Molecular Networks & Altamira
- NCCT & ToxCast Team

EPA
United States
Environmental Protection
Agency

Thank you for your attention