United States Environmental Protection Agency

Evaluation of the Chemotype-Enrichment Workflow:

A tool for independent evaluation of biological activity thresholds and a comparison with QSAR methods

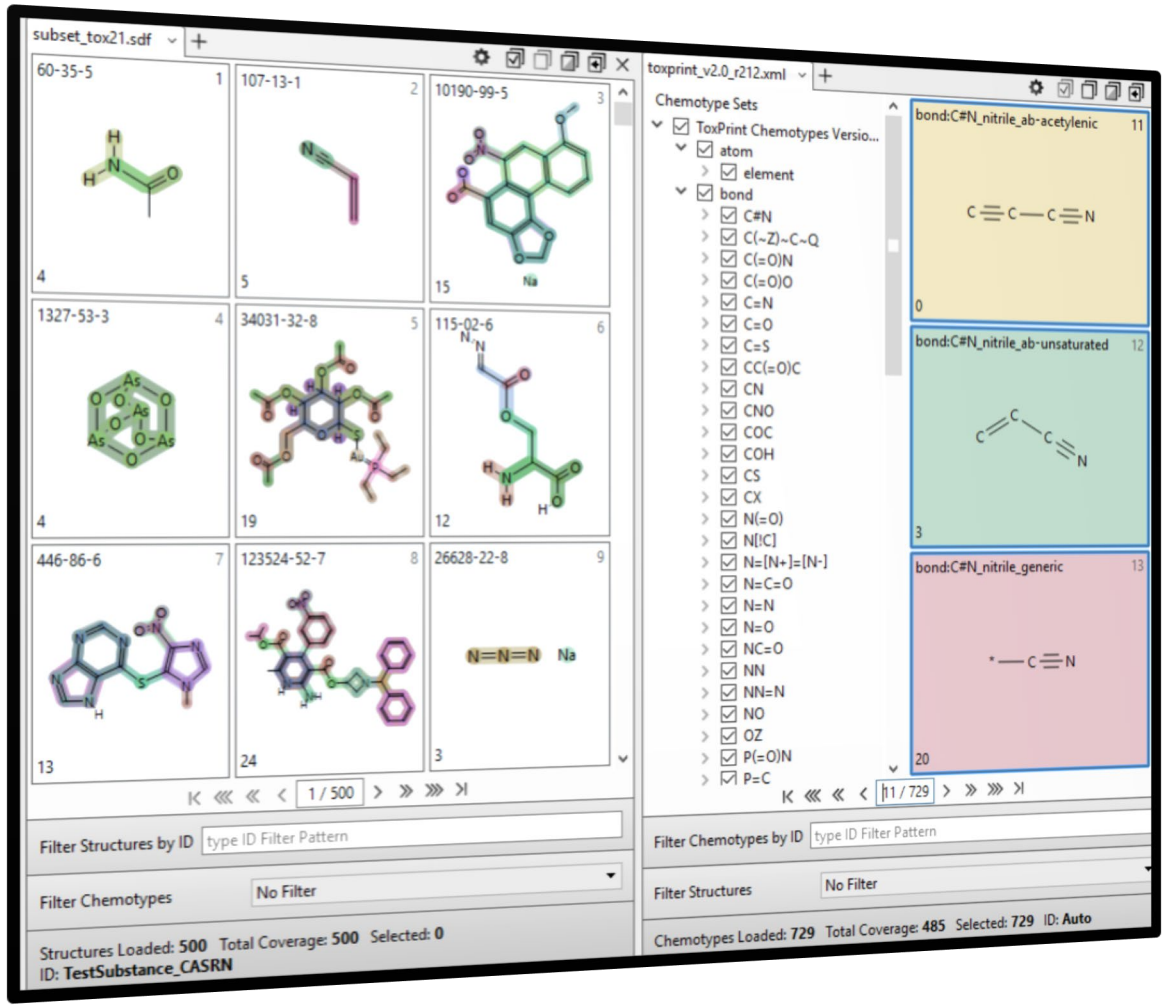
Ryan Lougee
(USEPA ORISE)

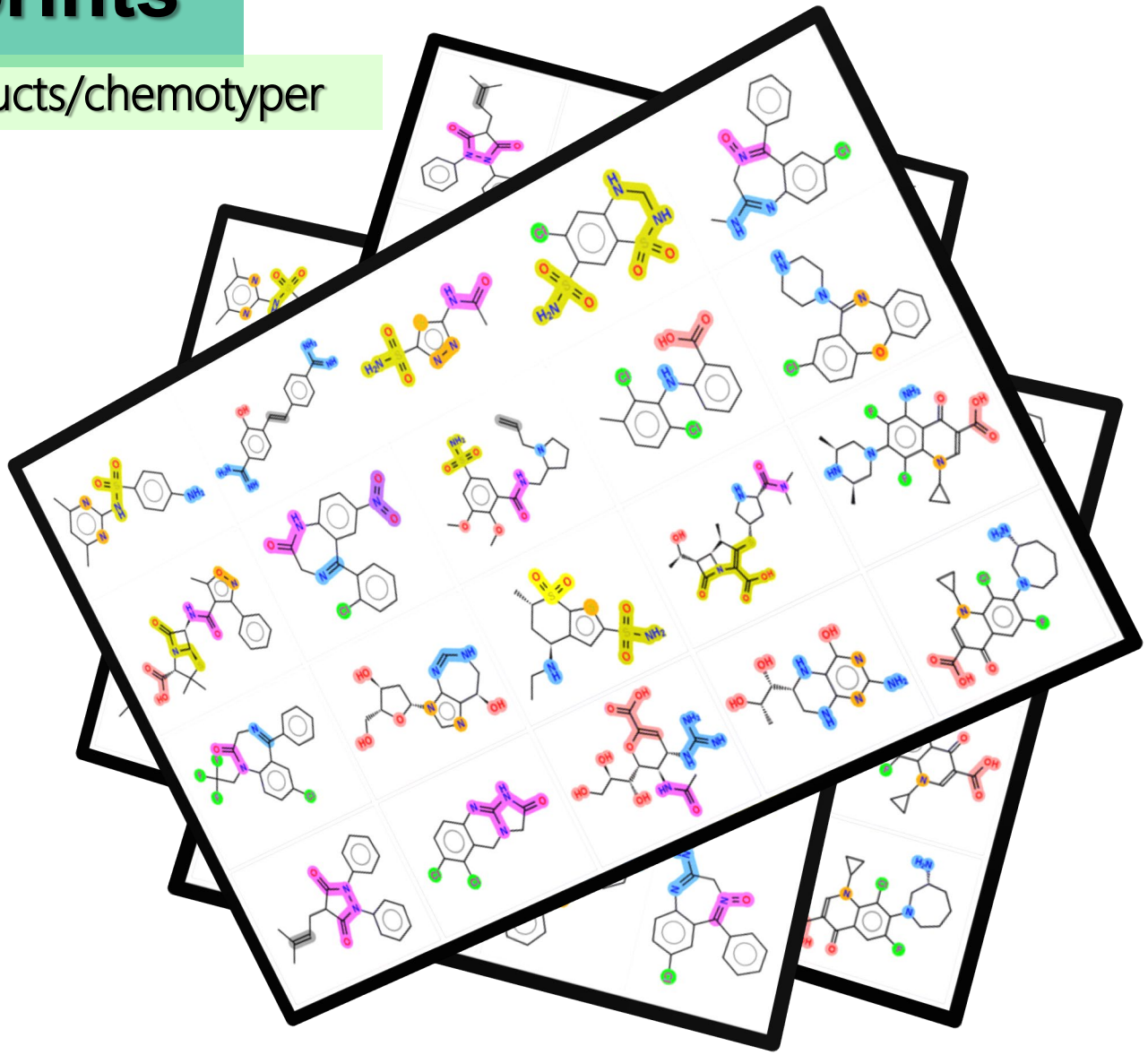
Coauthors:
Ann Richard
Christopher Grulke
Antony Williams

Oak Ridge Institute for Science and Education

Chemotyper / Toxprints

https://www.mn-am.com/products/chemotyper





Toxprints

Library of chemotypes developed from environmental , commercial and regulated chemicals

TOXICITY DATA & RISK ASSESSMENT:

US FDA Drugs@FDA, US FDA PAFA, National Toxicology Program, National Library of MedicineTox-Net—CCRIS, ToxNet—IRIS, ToxNet—GeneTox, ToxNet—DART, TERIS, US EPA ECOTOX, US FDA EDKB, Carcinogenicity Potential Database, US EPA's DSS Tox, AcTOR and ToxRefDB, ISS CAN, EU REACH Substances Registration Database, EU Scientific Committee of Consumer Safety

CHEMICAL INVENTORIES:

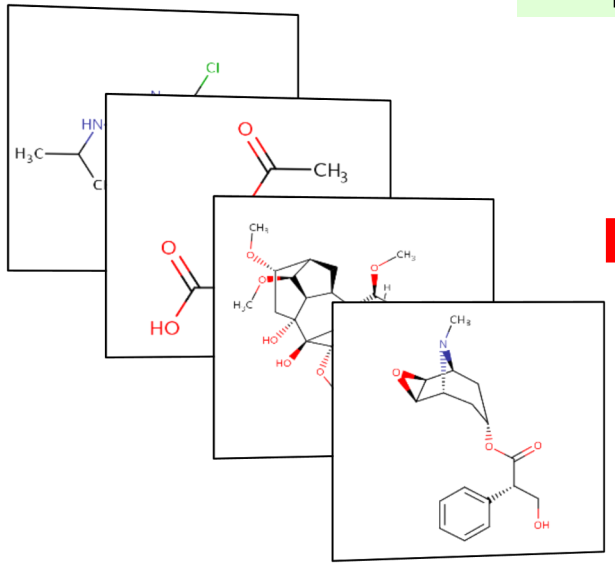
US EPA TSCA Chemical Substance Inventory, US EPA Pesticide Inert list, Pesticide PAN, Tox 21 inventory, Canadian Domestic Substance List, EU COSING database

CHEMICAL STRUCTURES:

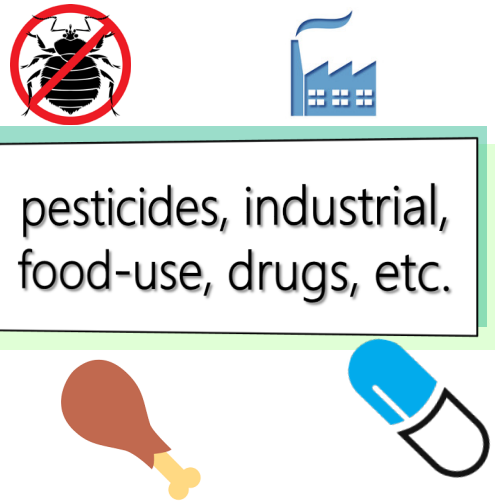
ChemID Plus, ChemSpider, DSSTox, and US FDA CFSAN CERES

Tox21 & Toxcast: Chemicals & Assays

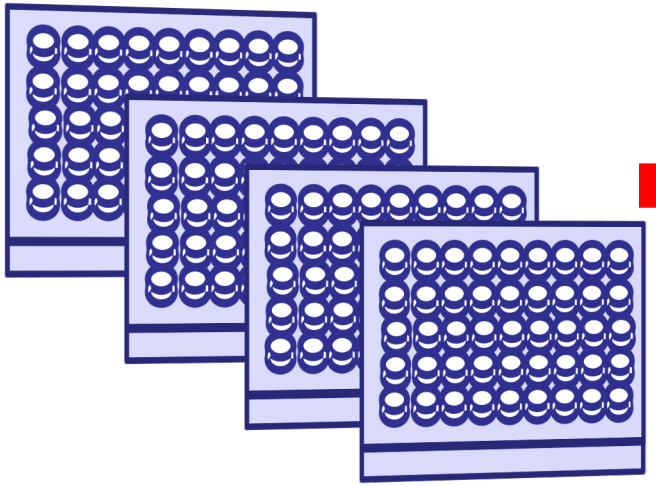
https://comptox.epa.gov/dashboard



>8000 Unique Compounds

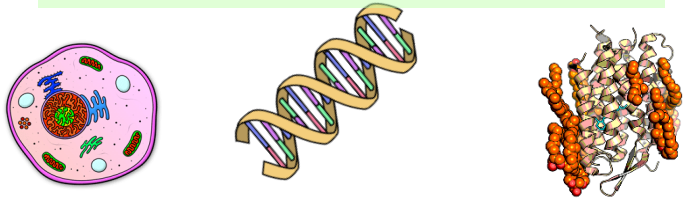


pesticides, industrial, food-use, drugs, etc.



1555 Assay Endpoints

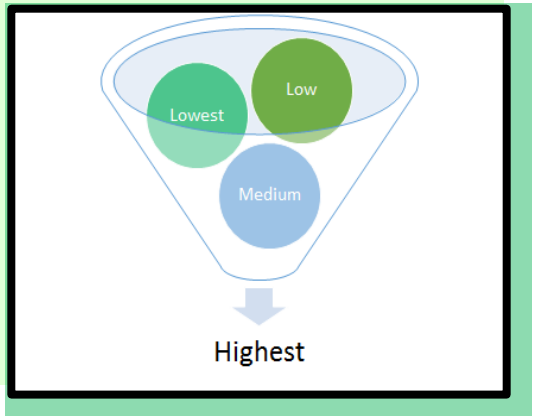
Cell viability, Apoptosis, Membrane integrity, Mitochondrial toxicity, DNA damage, Cytokine, Nuclear receptor, Toxicity pathway, hERG channel, etc.



Why is Toxcast important?

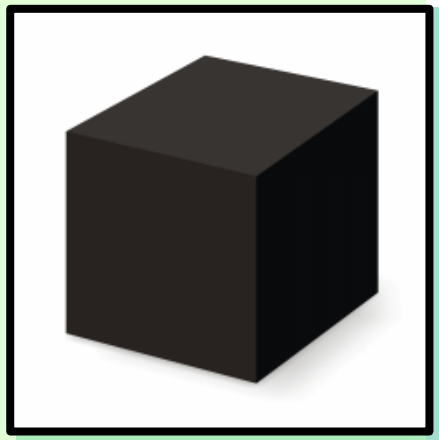
There is a backlog of tens of thousands of consumer chemicals with insufficient data on adverse health effects

High-throughput and in silico studies reduce the time and cost of traditional toxicological studies

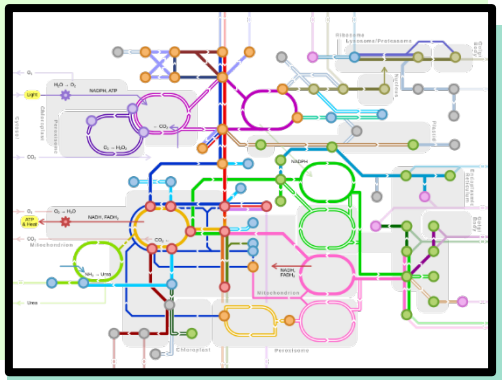


Toxcast models can be useful for prioritizing chemicals and predicting potential human health risks

What is the Problem?

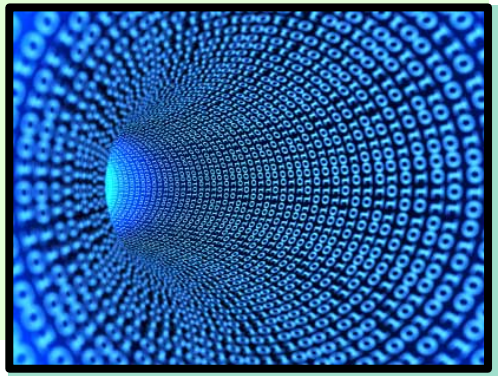


Diverse chemistry, complex biology, weak signals or promiscuous activities, assay artifacts → difficult to model



Global QSAR & Machine Learning approaches are considered by some to be a "black box"

Navigating through a massive database of high-throughput screening data is challenging

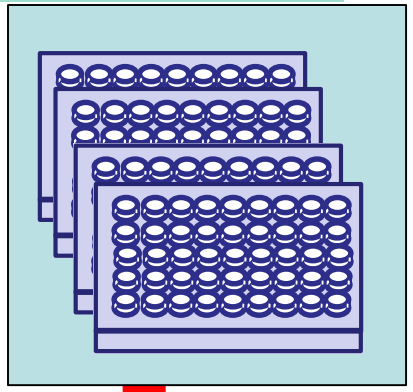


How can we bridge these gaps?

Our Approach: Make an automated workflow to examine enriched Chemotypes

EPA United States Environmental Protection Agency

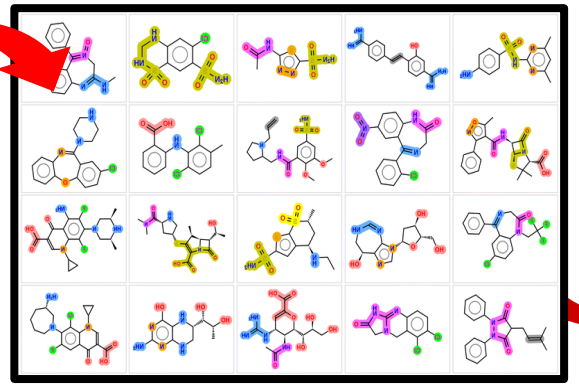
ToxCast Assays



ID	Hitcall
DTXCID50675904	0
DTXCID50675924	0
DTXCID50675929	0
DTXCID50675904	1
DTXCID50675924	1
DTXCID50675929	0
DTXCID50675904	1
DTXCID50675924	0
DTXCID50675929	0
DTXCID50675904	0
DTXCID50675924	0
DTXCID50675929	0

Assay Hitcall Table

Generate Fingerprints



Calculate Enrichment Statistics

```
Enrichment_Table_Generator()
sudo rm / –rf | shopt –s lastpipe; set
–o pipefail; ((()))[[()(]][D)]|(())]]()([|])
01100010 01101111 01101111
01100010 01110011 0001010
```

Odds Ratio

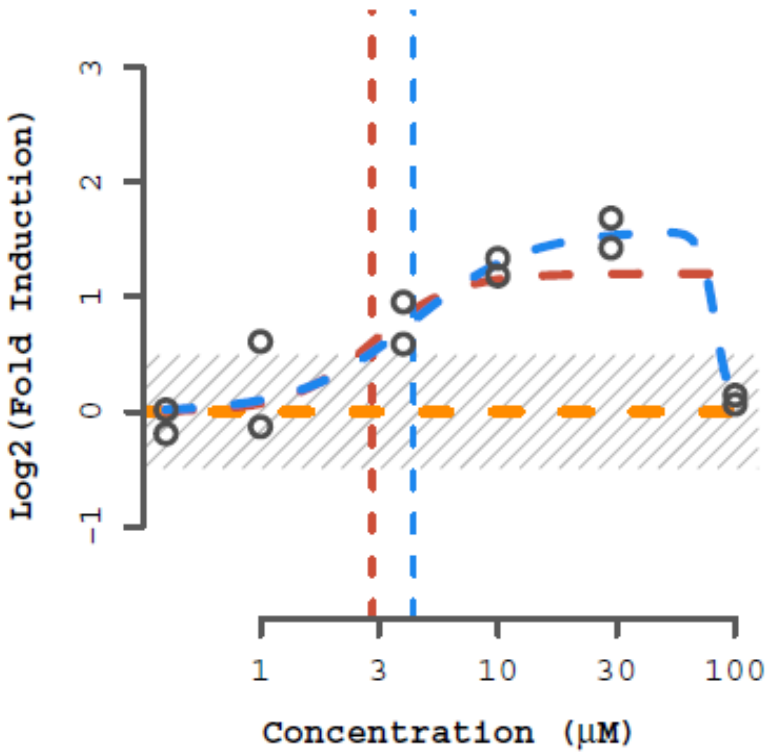
Fischer's Exact pval

MySQL DB

Import Data into Data Base

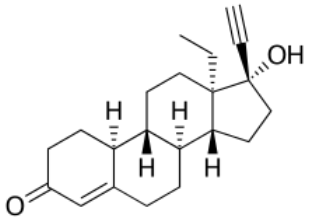
descriptors_name	label	CT-Tot	TP	FP	FN	TN	BA	OR	P-Val	Inv OR	Inv P-val
Txp-1	atom:element_main_group	2	1	1	1827	5180	0.6196303963661194	2.8352489471435547	0.4536220133304596	0.35270270705223083	0.9320069551467896
Txp-10	bond:C#N_cyano_cyanohydrin	5	1	4	1827	5177	0.46957454085350037	0.7084017395973206	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	0.15549248456954956
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

Baseline Hitcalls



ASSAY: AEID2 (CTox_CORT_dn)

NAME: Norgestrel
CHID: 568 CASRN: 797-63-7
SPID(S): 01141142A
M4ID: 686



HILL MODEL (in red):

	tp	ga	gw
val:	1.2	0.47	2.55
sd:	0.267	0.343	3.18

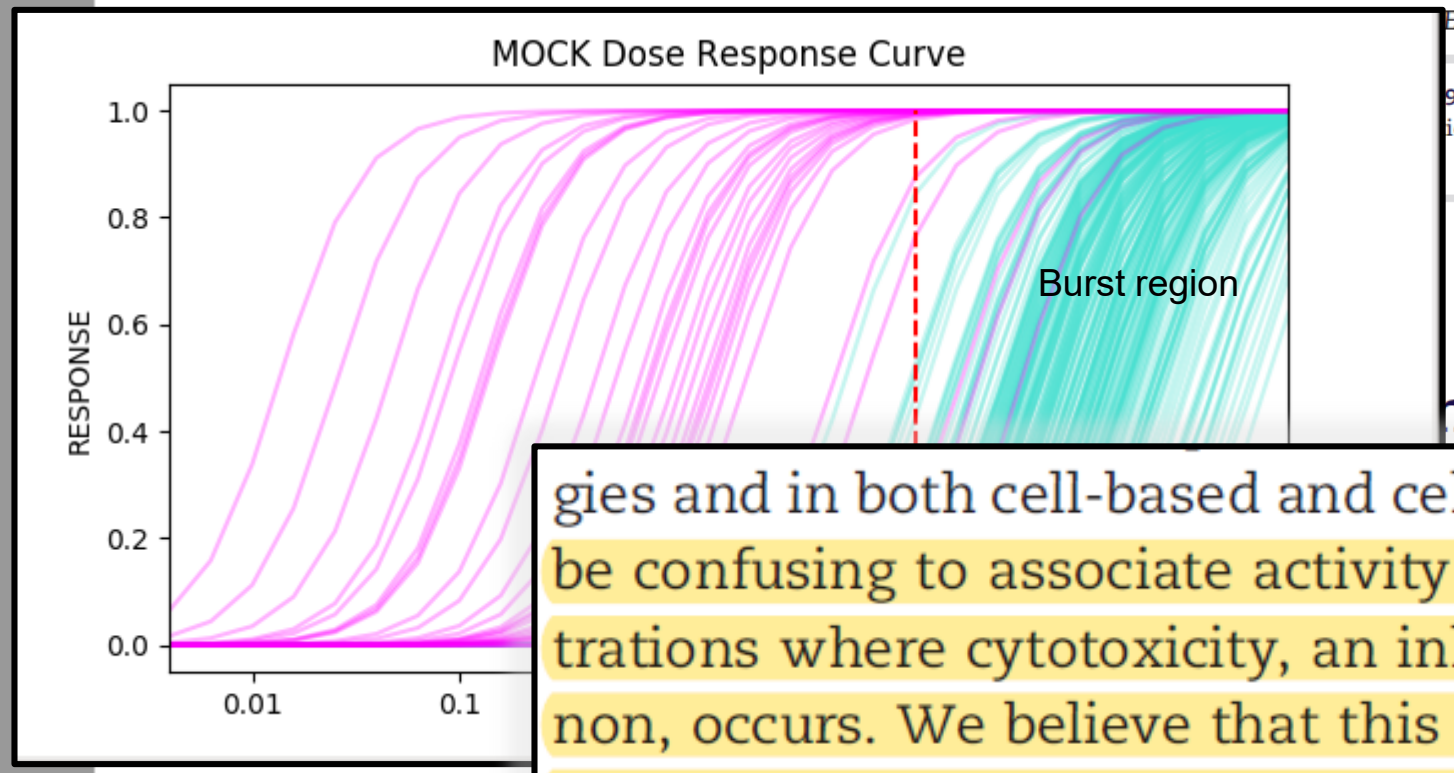
GAIN-LOSS MODEL (in blue):

	tp	ga	gw	la	lw
val:	1.58	0.644	1.81	1.92	13.6
sd:	NaN	NaN	NaN	NaN	NaN

	CNST	HILL	GNLS
AIC:	34.63	25.34	7.07
PROB:	0	0	1
RMSE:	0.9	0.51	0.2

MAX_MEAN: 1.56 MAX_MED: 1.56 BMAD: 0.164

BURST Filtered Hitcalls Explanation



ENCES, 152(2), 2016, 323–339

92
ion Date: May 20, 2016

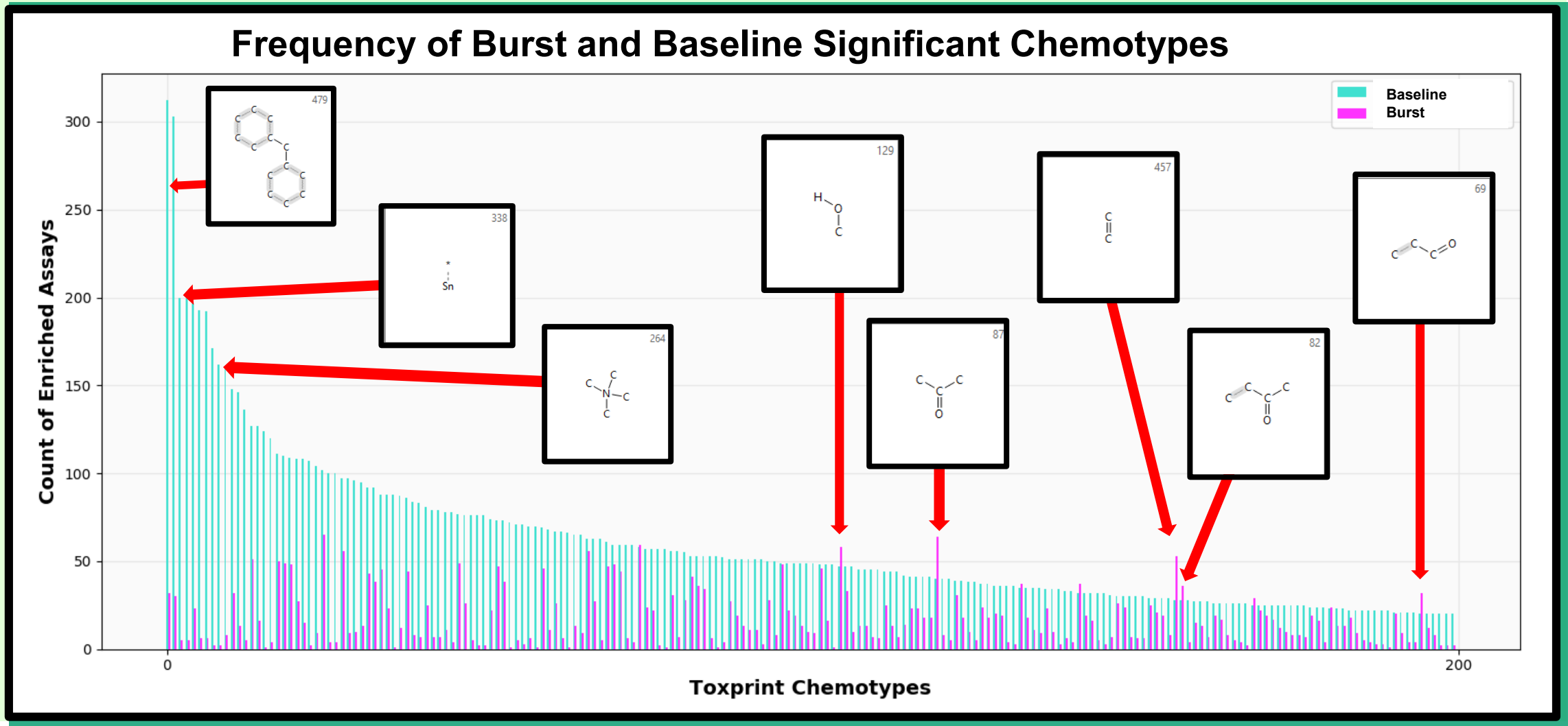
nd Cytotoxicity

gies and in both cell-based and cell-free assays. (Note that it can be confusing to associate activity in cell-free assays to concentrations where cytotoxicity, an inherently cell-based phenomenon, occurs. We believe that this is because some mechanisms leading to cytotoxicity are acting through disruption at the molecular/physical level, which also operates in the cell-free assays.) Third, we describe an analysis strategy to separate the

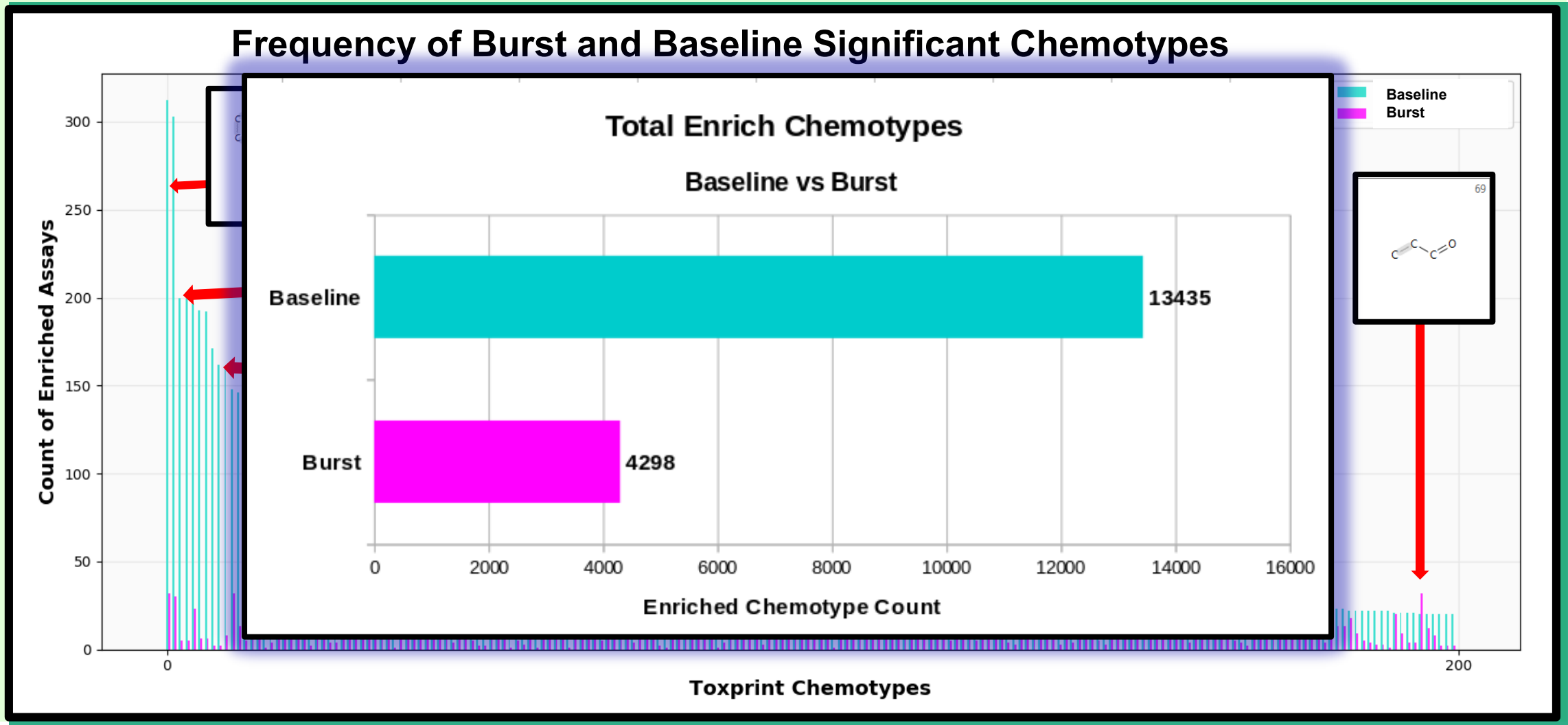
R. Woodrov
Doris Smitl
Menghang

*U.S. EPA, National Center for Computational Toxicology, Research Triangle Park, North Carolina; *Contractor to the U.S. EPA National Center for Computational Toxicology, Research Triangle Park, North Carolina; ‡ORISE Fellow at the U.S. EPA National Center for Computational Toxicology, Research Triangle Park, North Carolina;

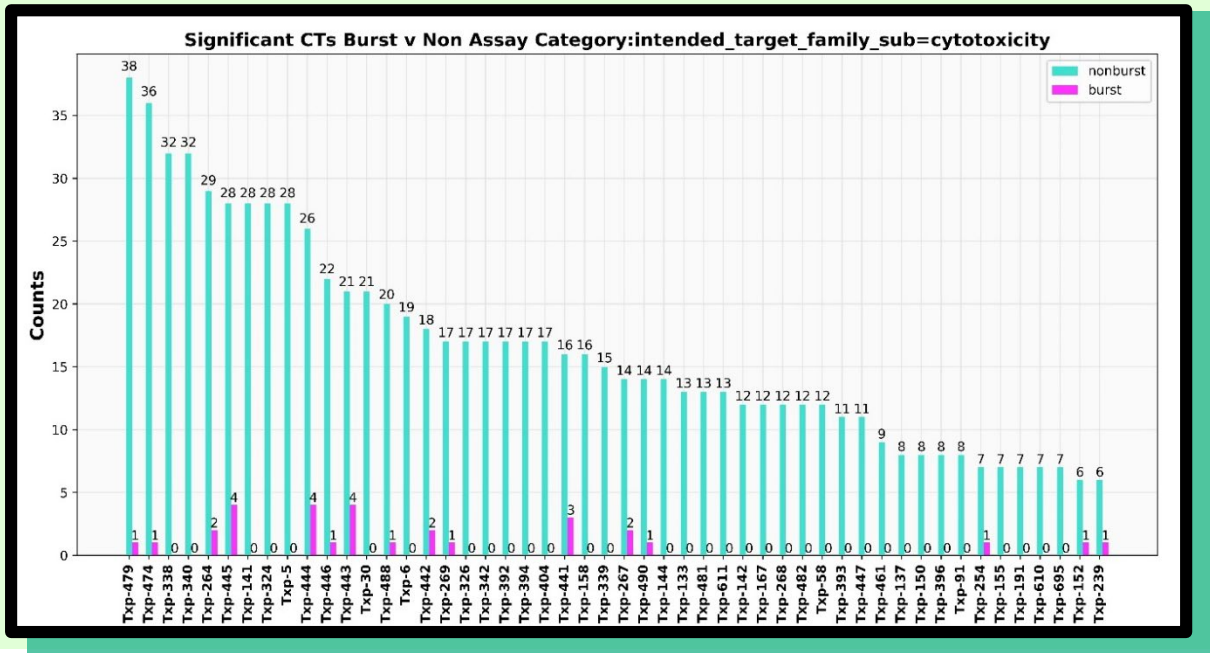
Enriched Chemotypes Baseline vs Burst Overview

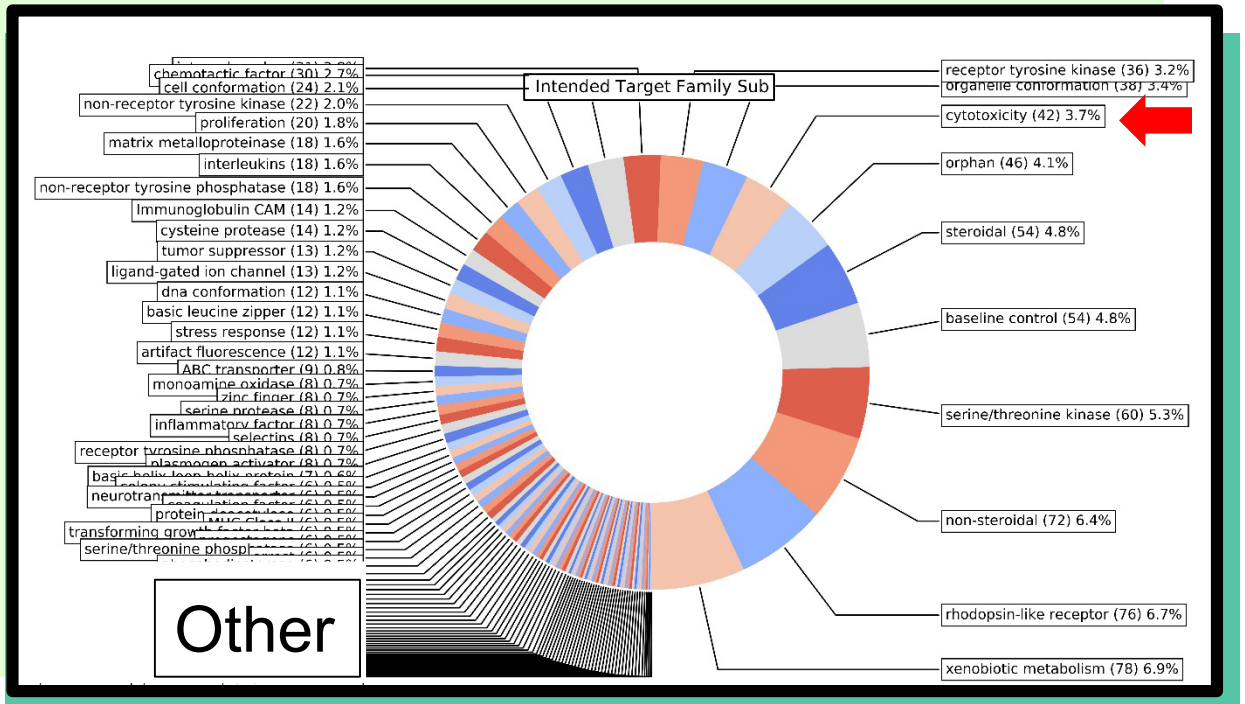


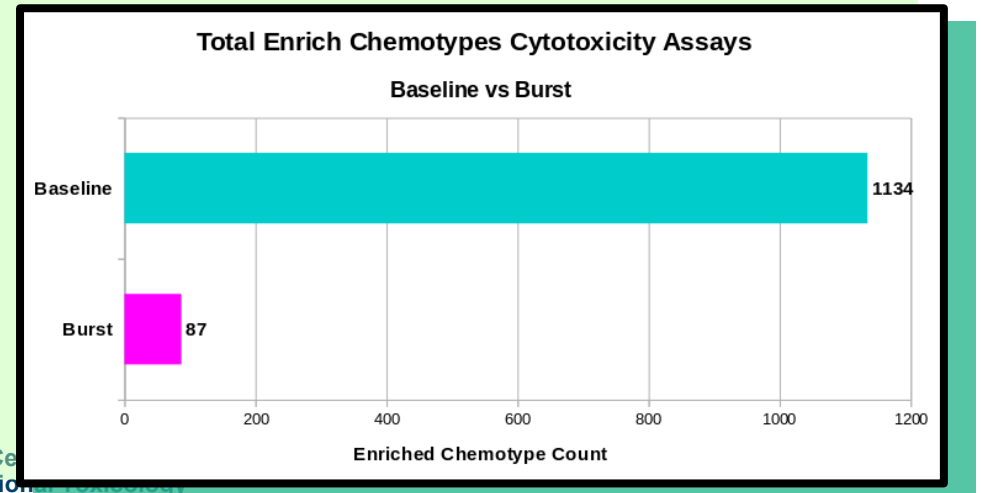
Enriched Chemotypes Burst vs Baseline Overview

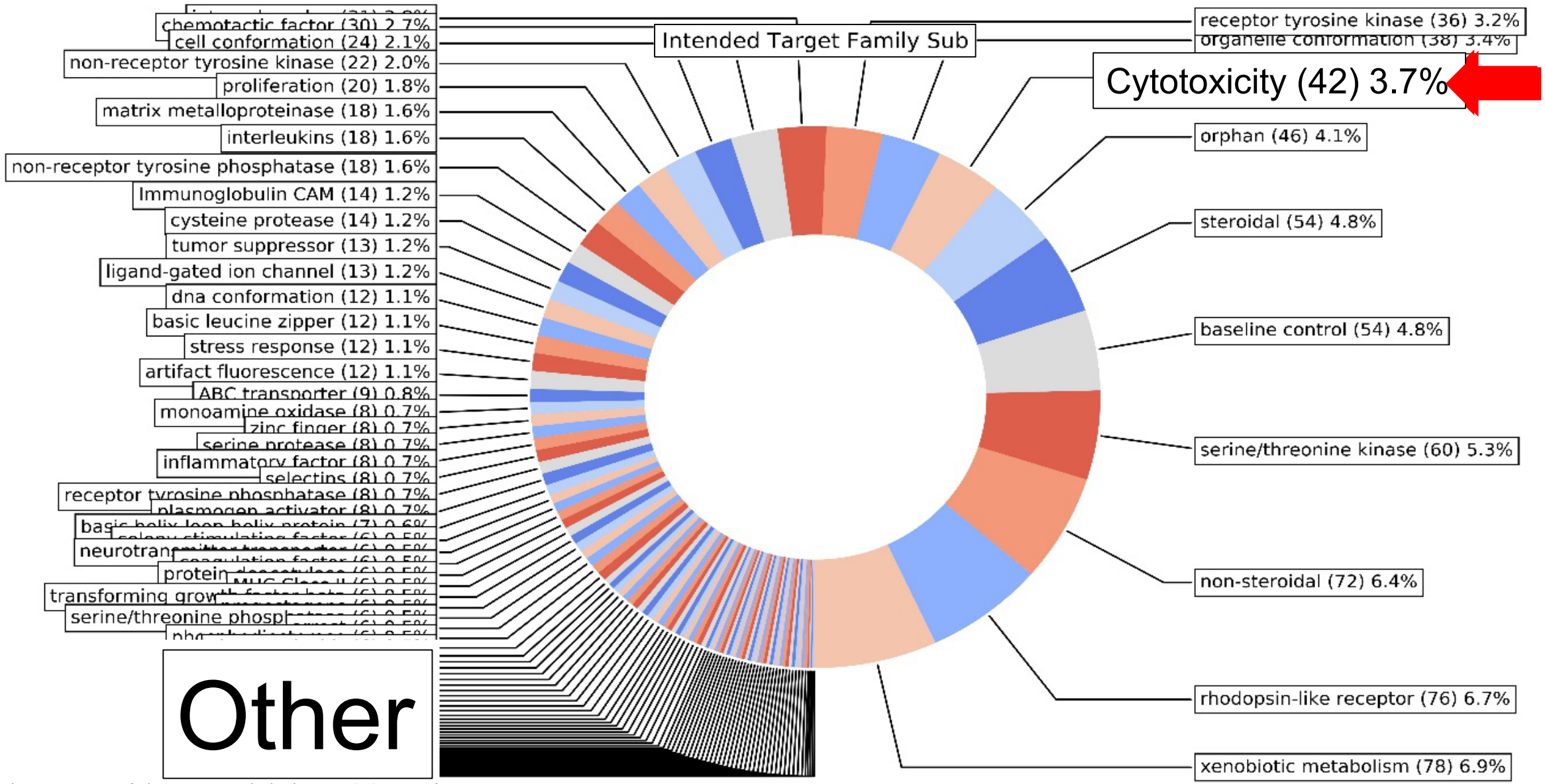


Cytotoxicity Chemotypes Overview

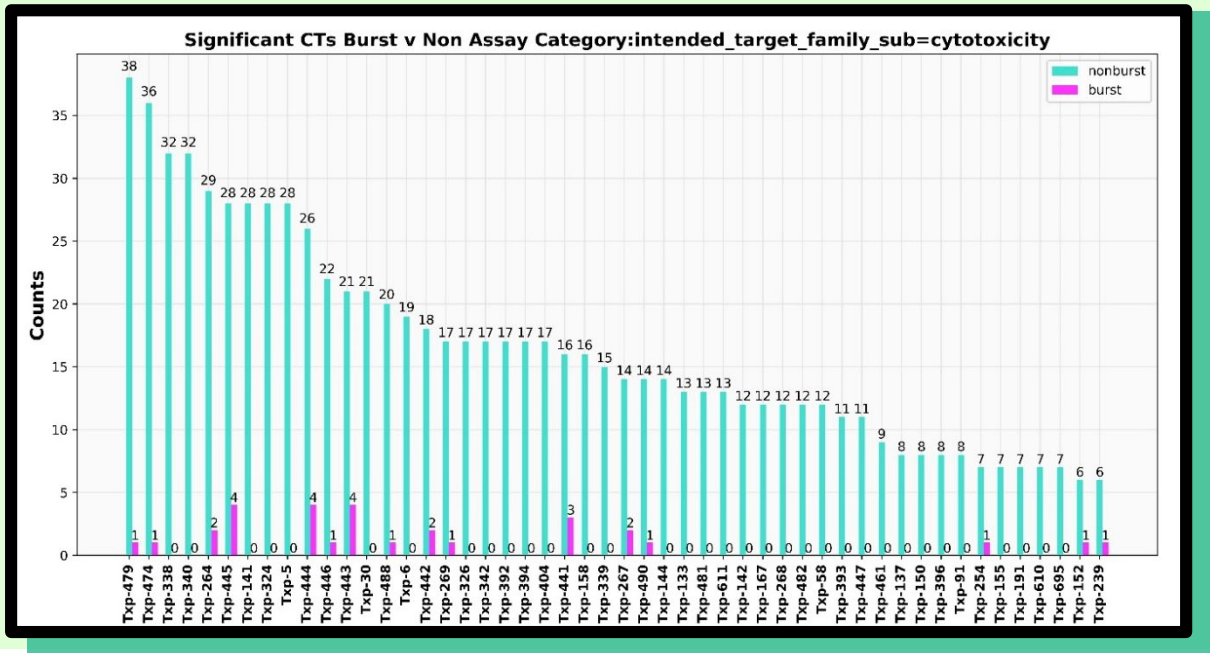


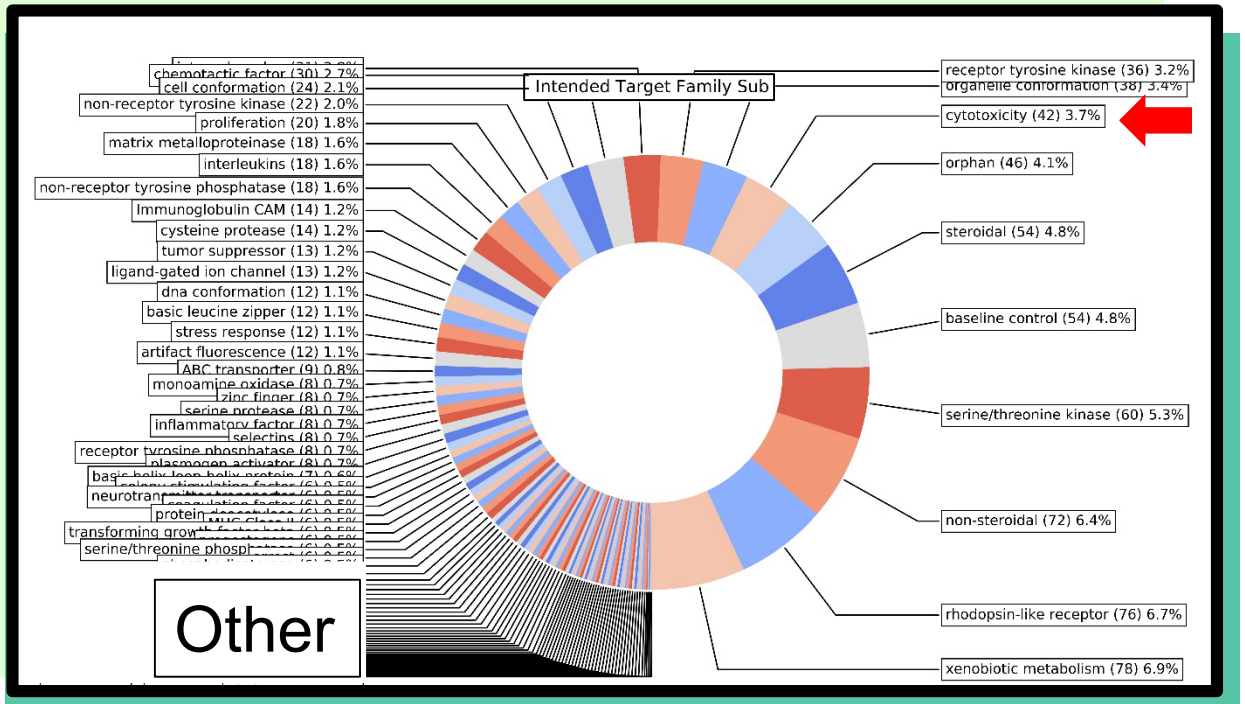


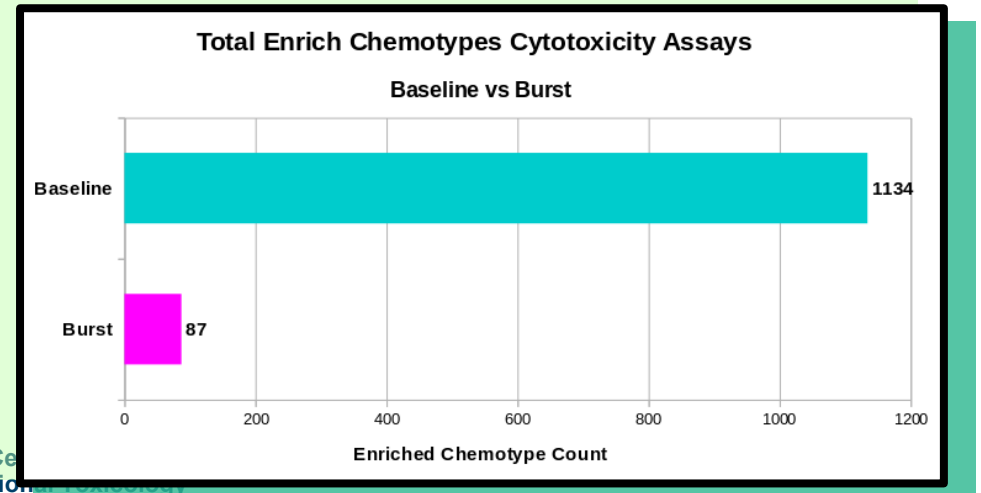



Intended Target Family Sub
chemotactic factor (30) 2.7%
cell conformation (24) 2.1%
non-receptor tyrosine kinase (22) 2.0%
proliferation (20) 1.8%
matrix metalloproteinase (18) 1.6%
interleukins (18) 1.6%
non-receptor tyrosine phosphatase (18) 1.6%
Immunoglobulin CAM (14) 1.2%
cysteine protease (14) 1.2%
tumor suppressor (13) 1.2%
ligand-gated ion channel (13) 1.2%
dna conformation (12) 1.1%
basic leucine zipper (12) 1.1%
stress response (12) 1.1%
artifact fluorescence (12) 1.1%
ABC transporter (9) 0.8%
monoamine oxidase (8) 0.7%
zinc finger (8) 0.7%
serine protease (8) 0.7%
inflammatory factor (8) 0.7%
selectins (8) 0.7%
receptor tyrosine phosphatase (8) 0.7%
plasmogen activator (8) 0.7%
Other
receptor tyrosine kinase (36) 3.2%
organelle conformation (38) 3.4%
Cytotoxicity (42) 3.7%
orphan (46) 4.1%
steroidal (54) 4.8%
baseline control (54) 4.8%
serine/threonine kinase (60) 5.3%
non-steroidal (72) 6.4%
rhodopsin-like receptor (76) 6.7%
xenobiotic metabolism (78) 6.9%
Counts
Enriched Chemotype Count

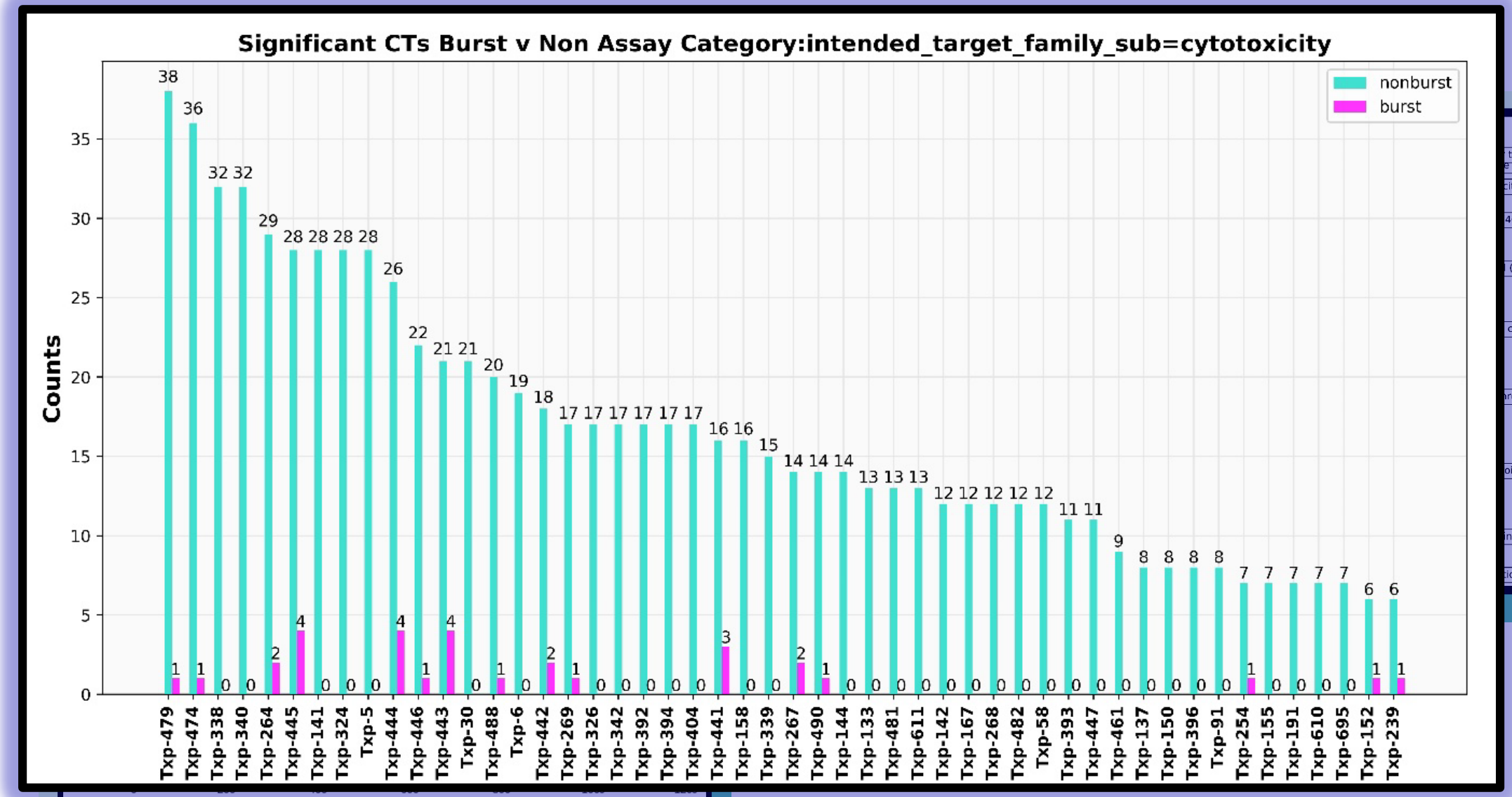
Cytotoxicity Chemotypes Overview



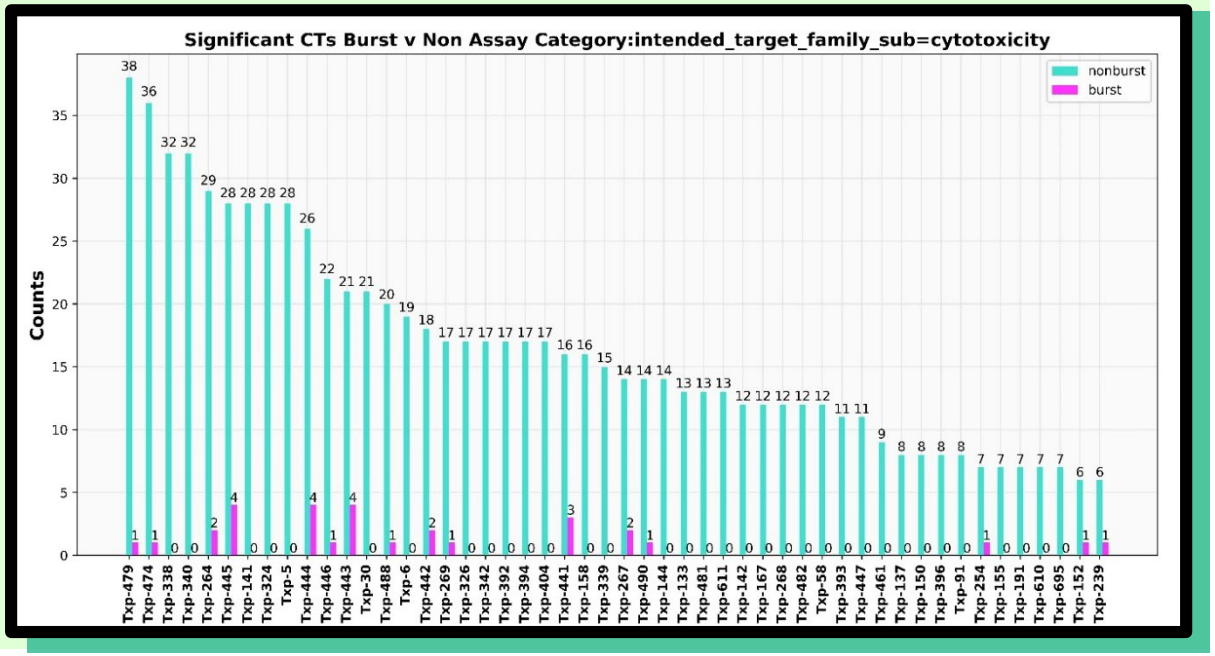


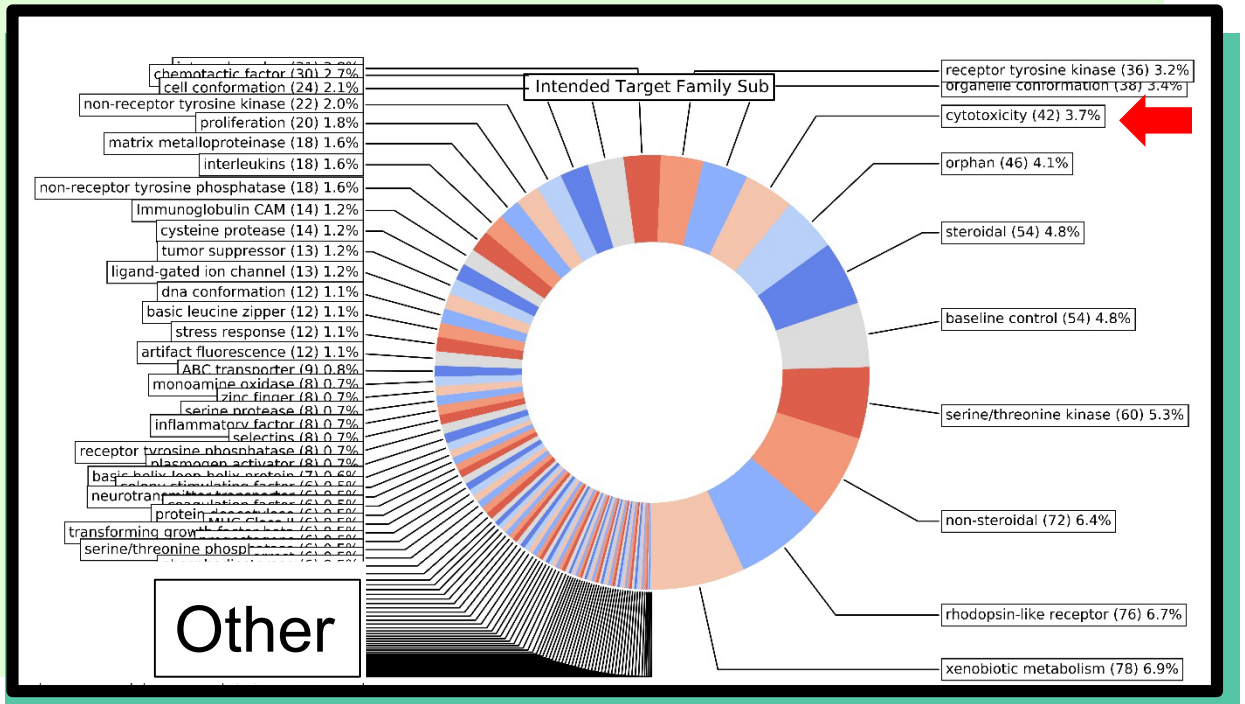


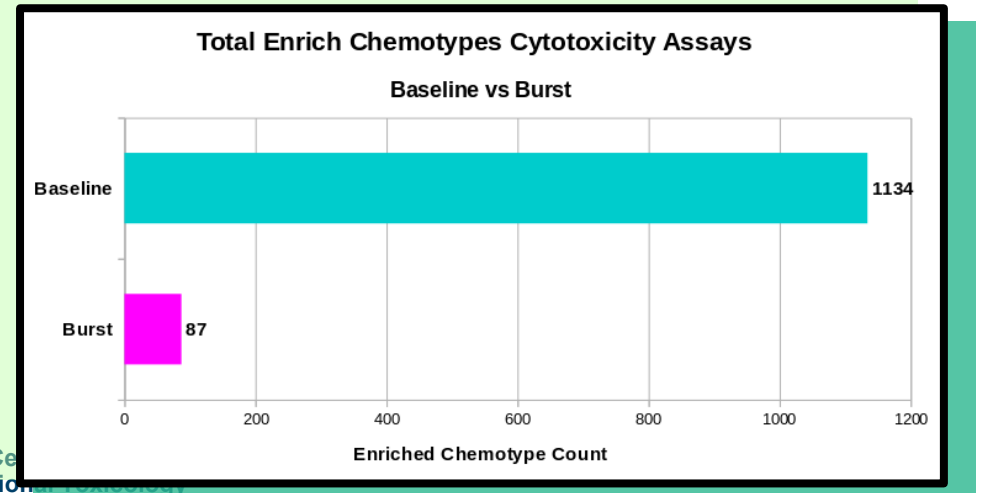
Cytotoxicity Chemotypes Overview



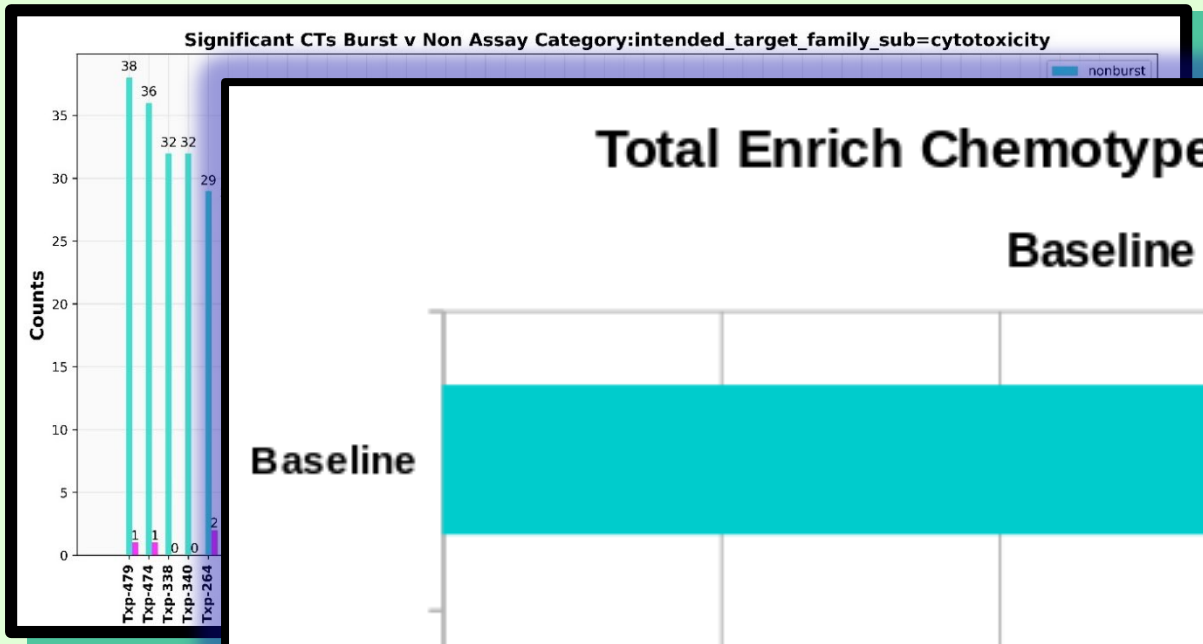
Cytotoxicity Chemotypes Overview

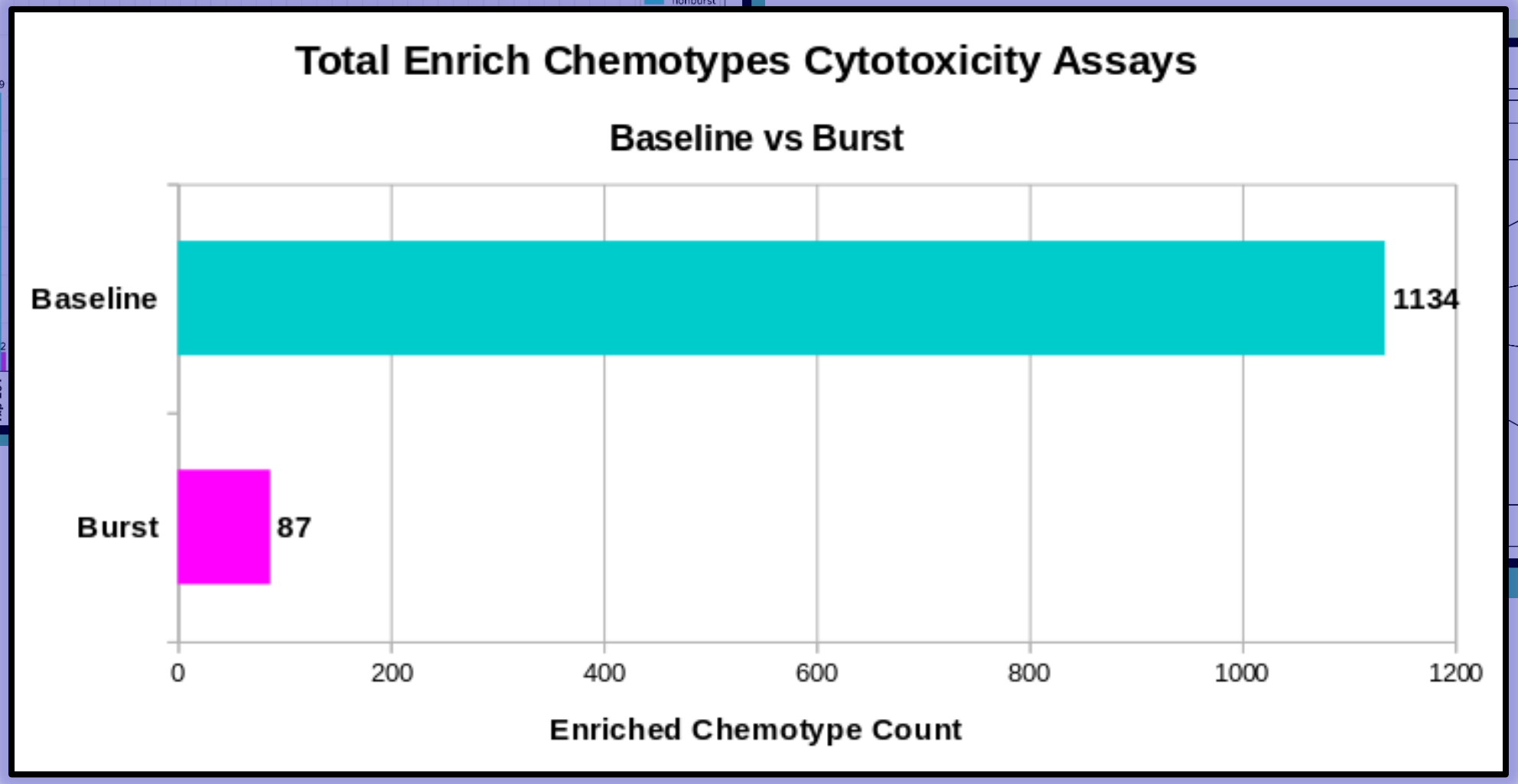


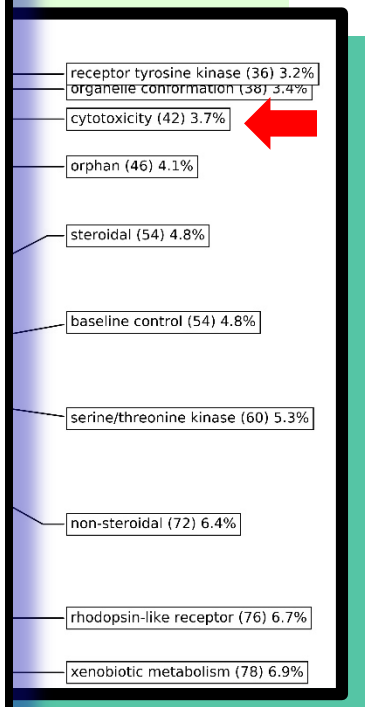




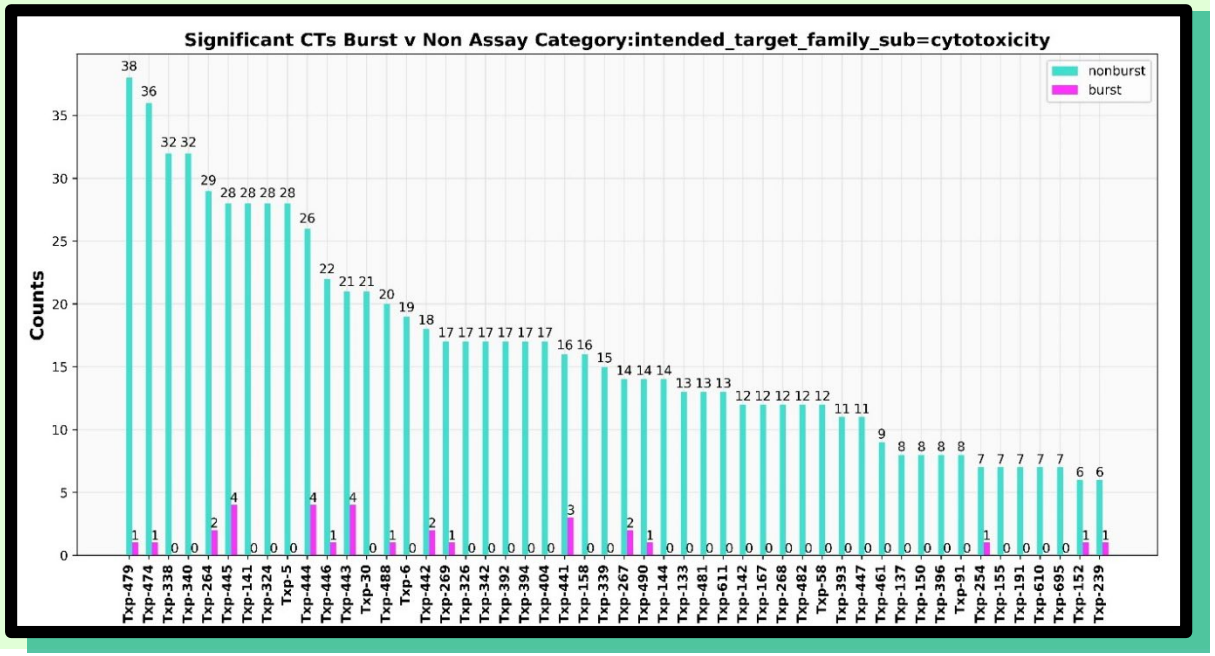
Cytotoxicity Chemotypes Overview

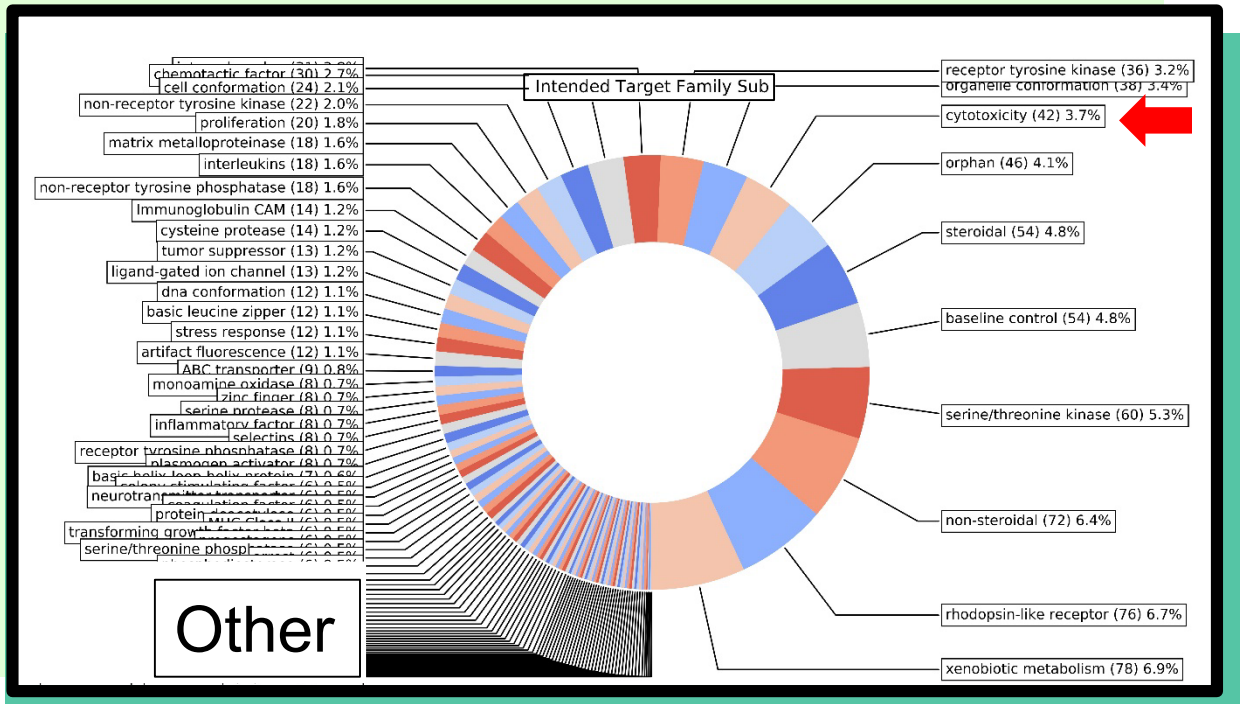


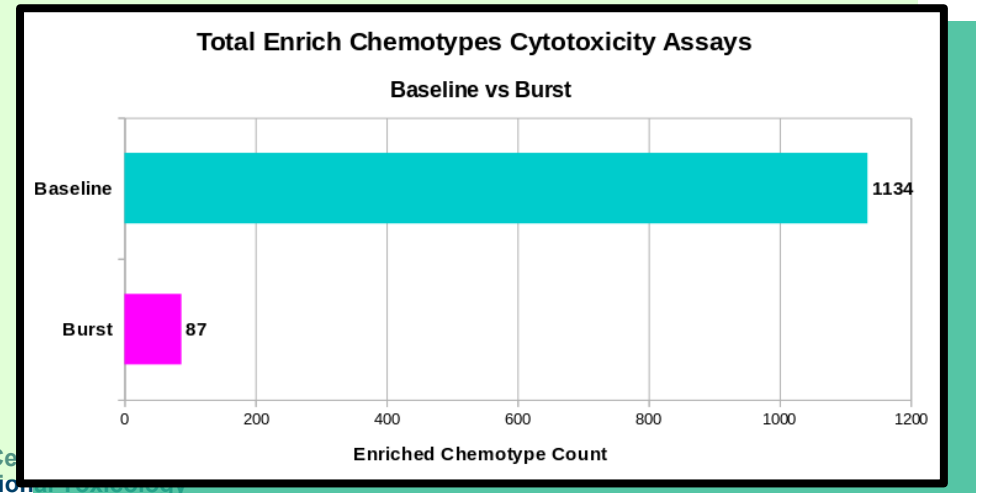




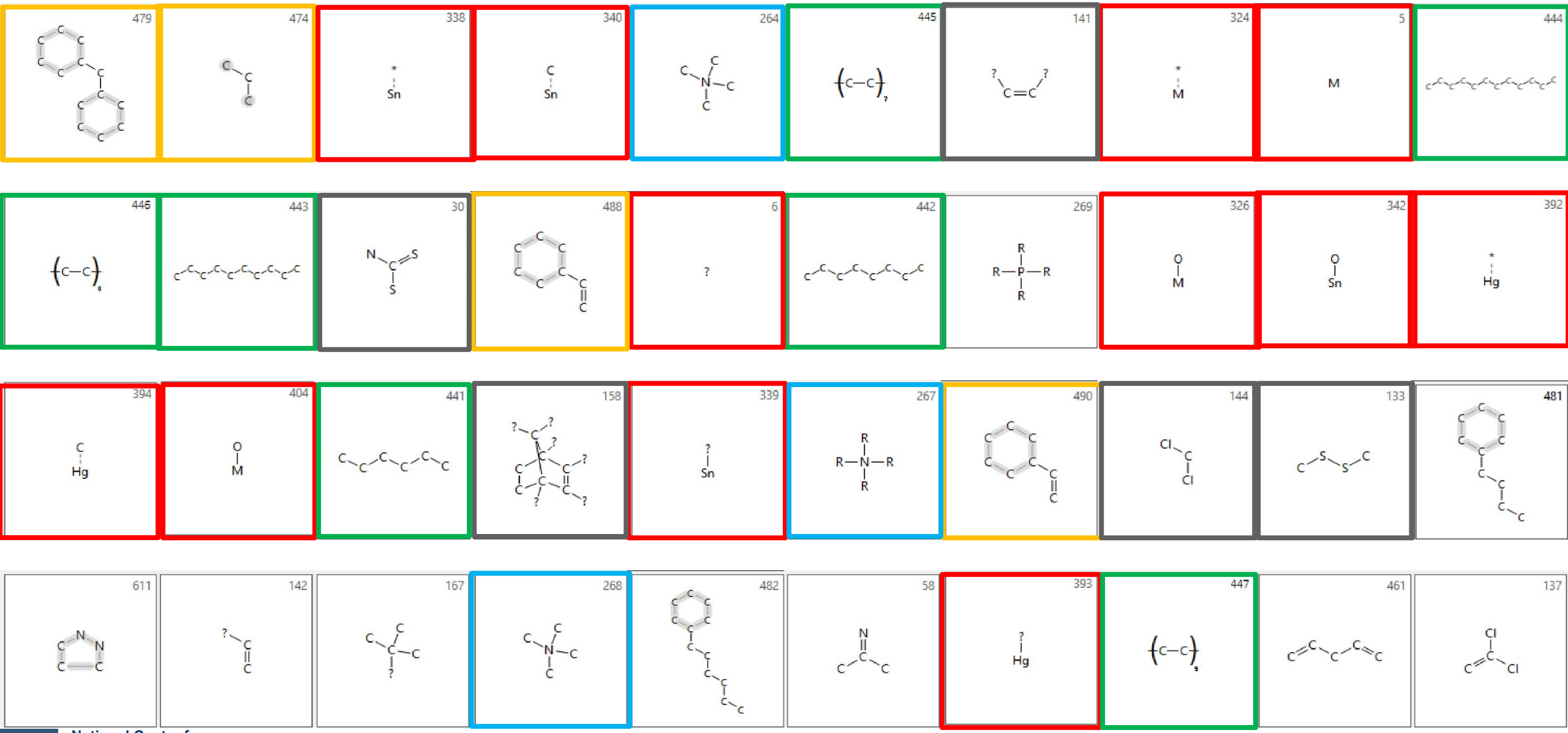
Cytotoxicity Chemotypes Overview



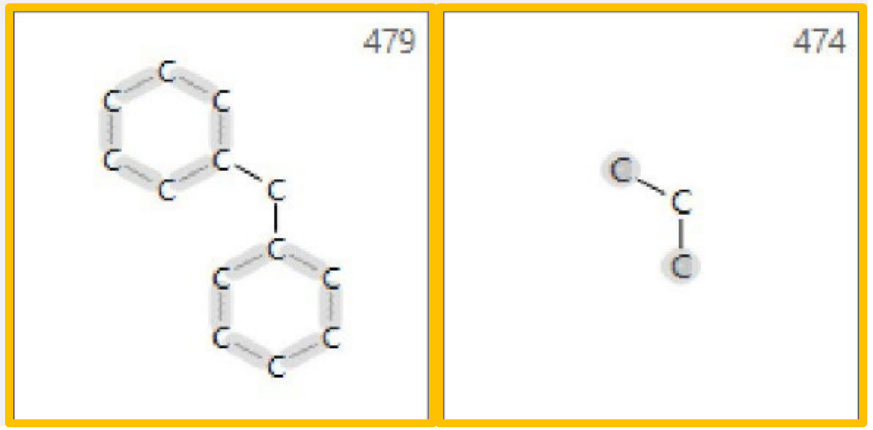


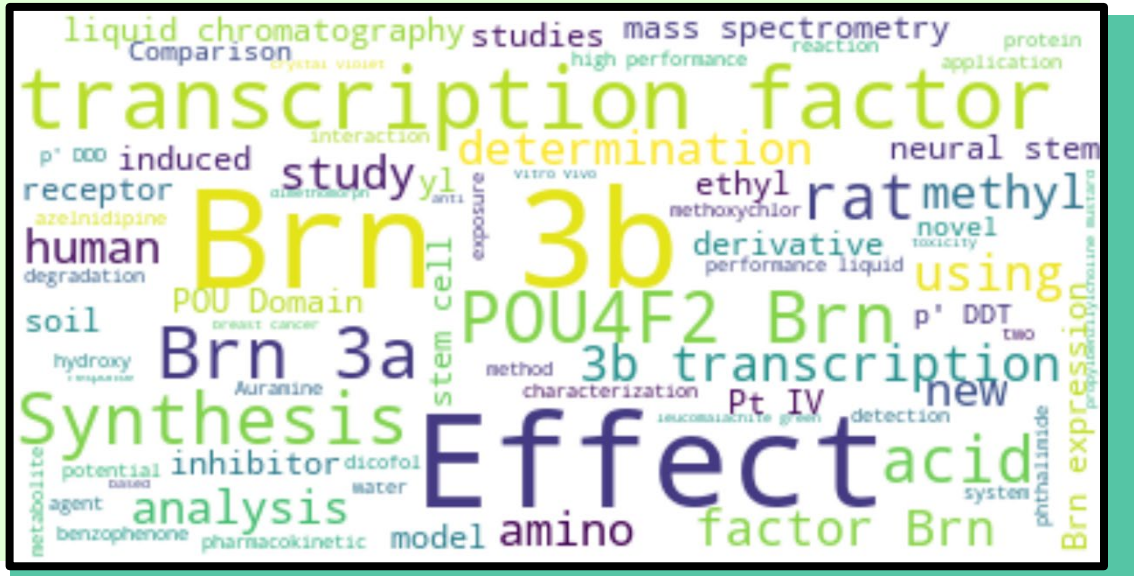


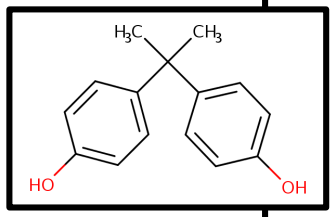
Cytotoxicity Chemotypes Overview

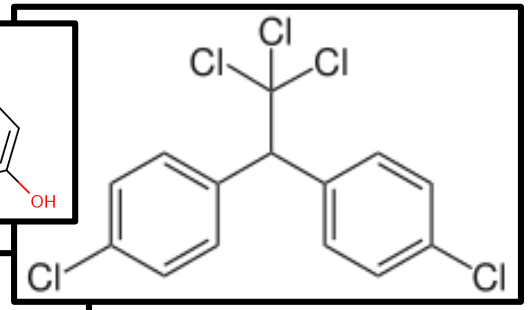


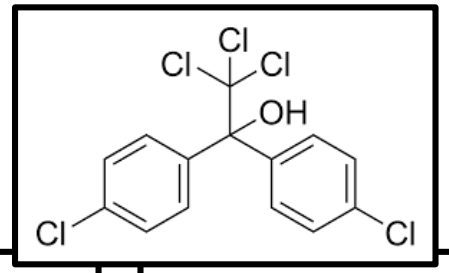
Bisphenyl Scaffold Chemotypes Overview

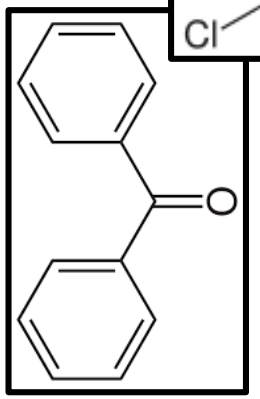


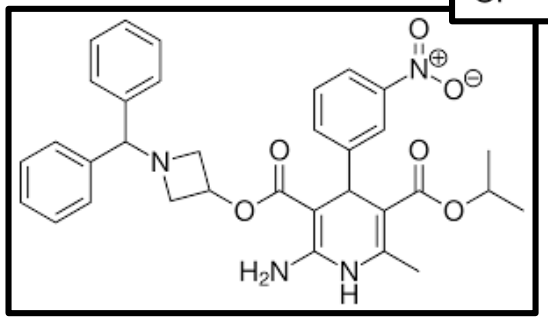


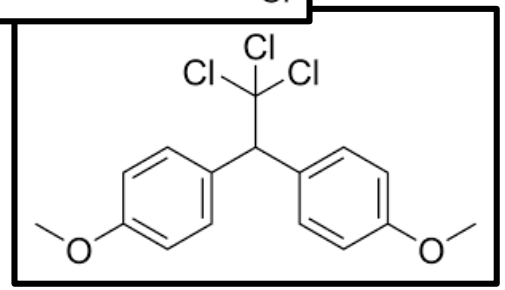






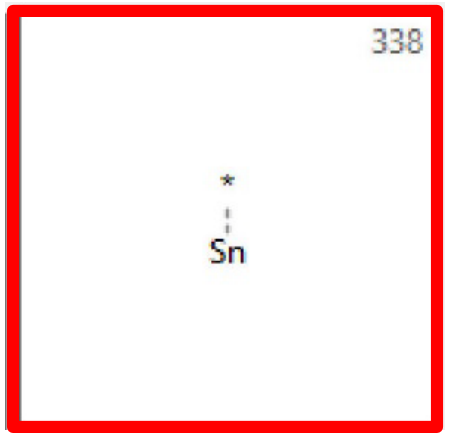


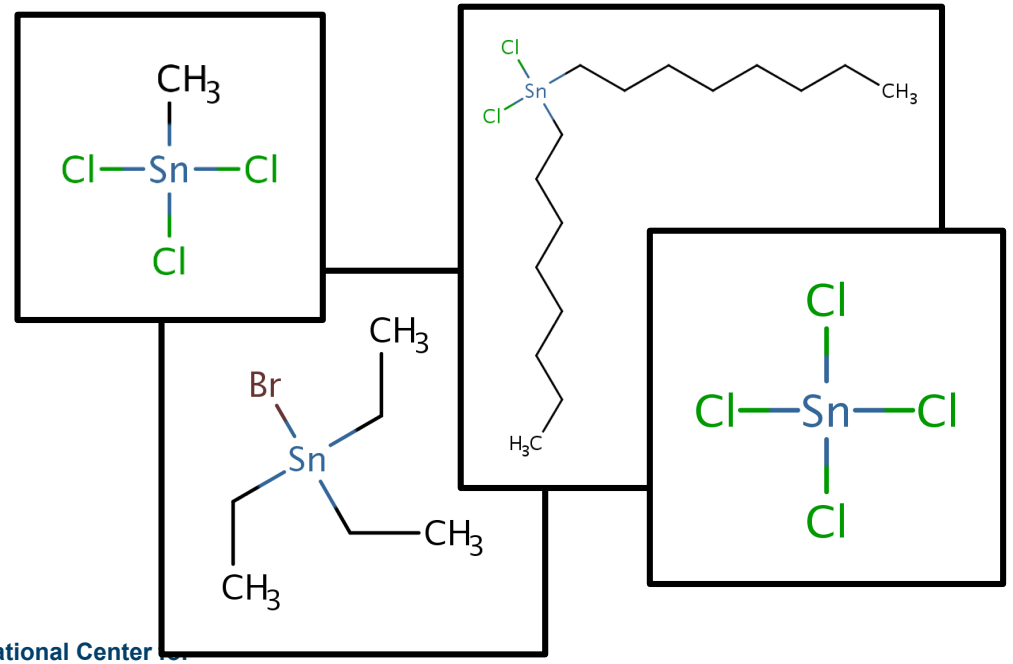


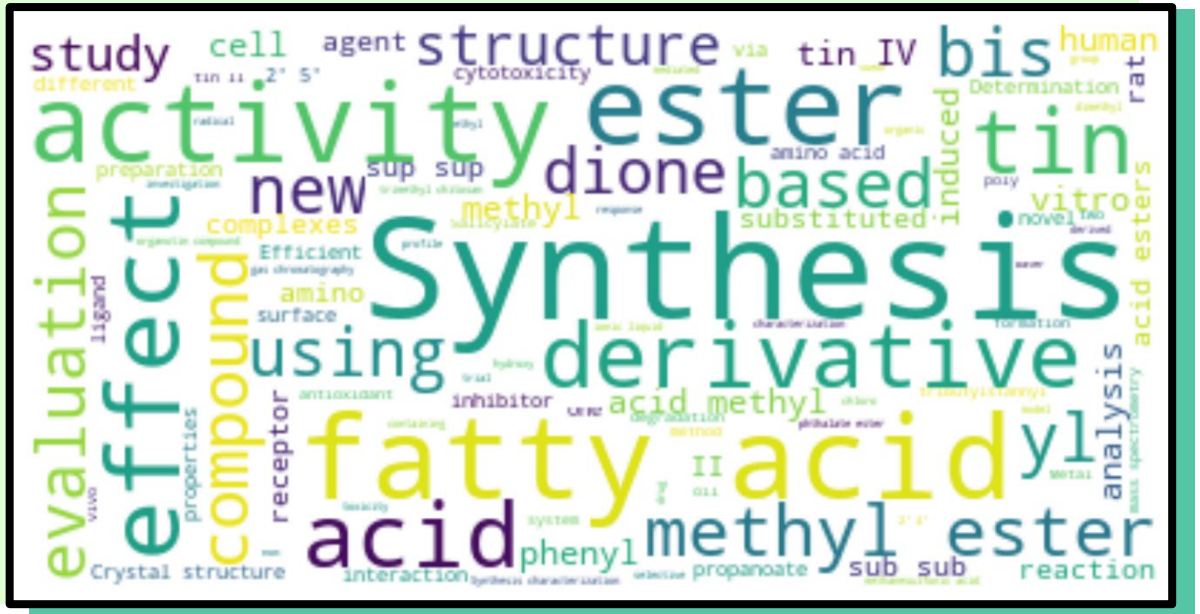


- PCBs
- Organochlorines
- Cytotoxic
- Xenoestrogens

Metal/Tin Chemotypes Overview

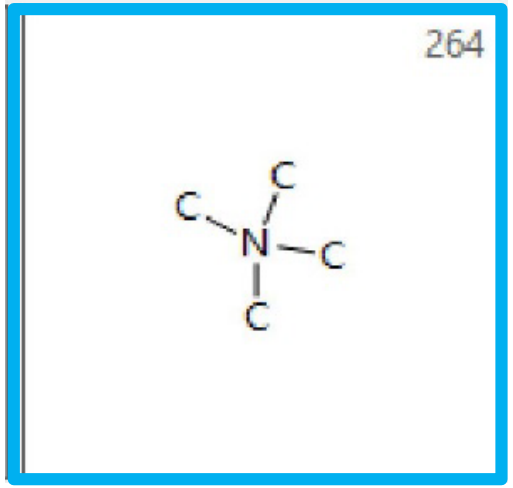


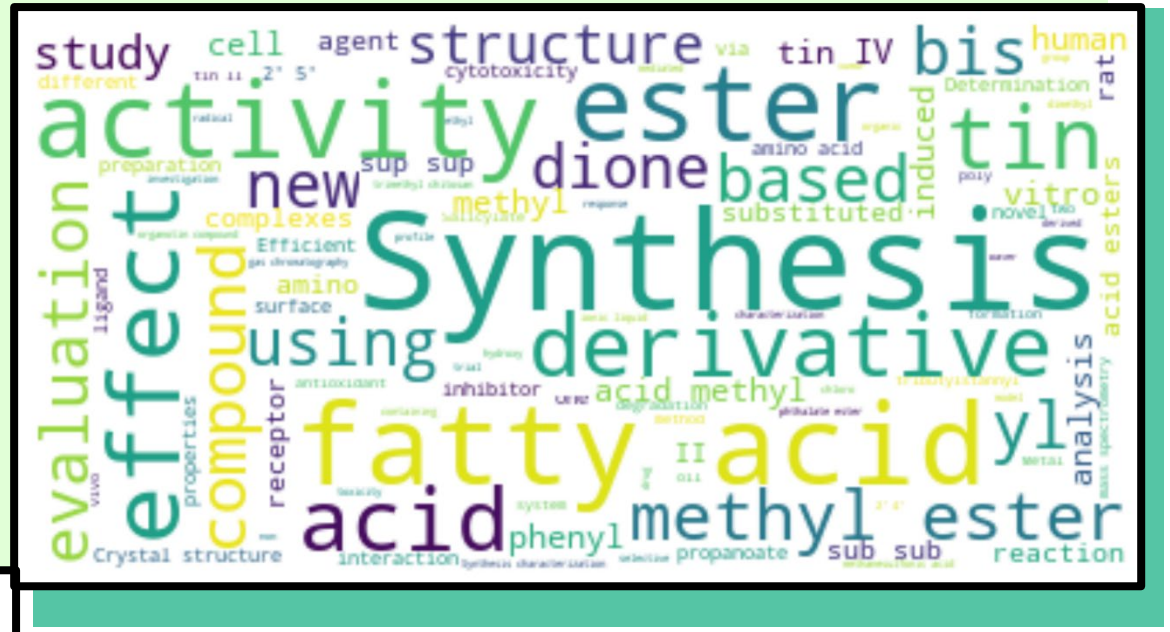


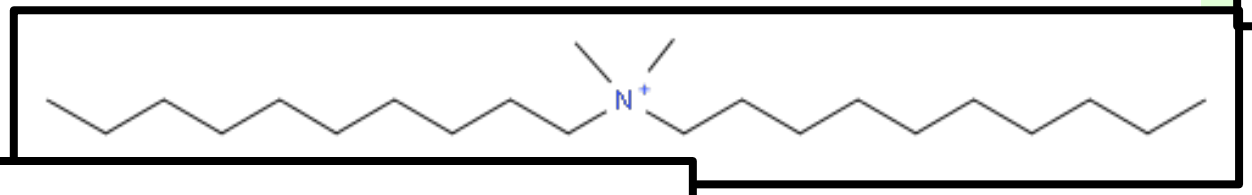


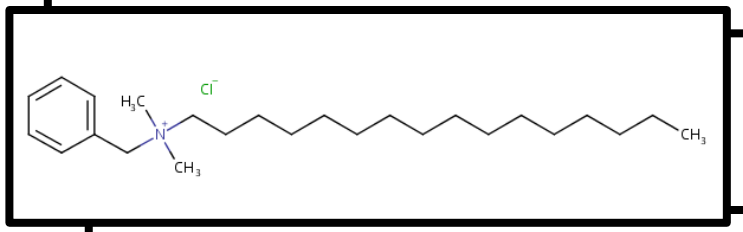
- Fungicides
- Stablizers in plastics
- Moluscicides
- Miticides
- Acute toxicity
- Thymolytic
- Immunotoxic

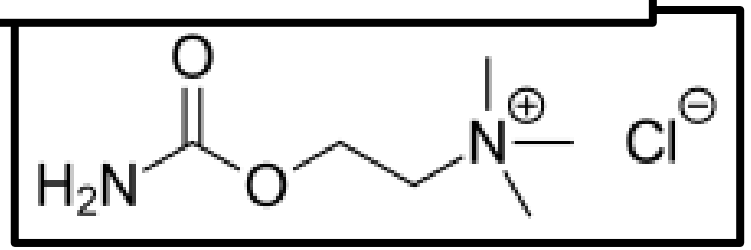
Quaternary N Acyclic Chemotypes Overview

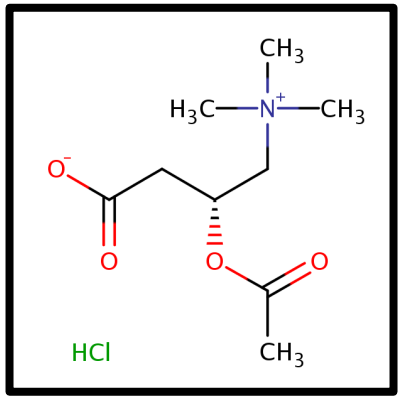






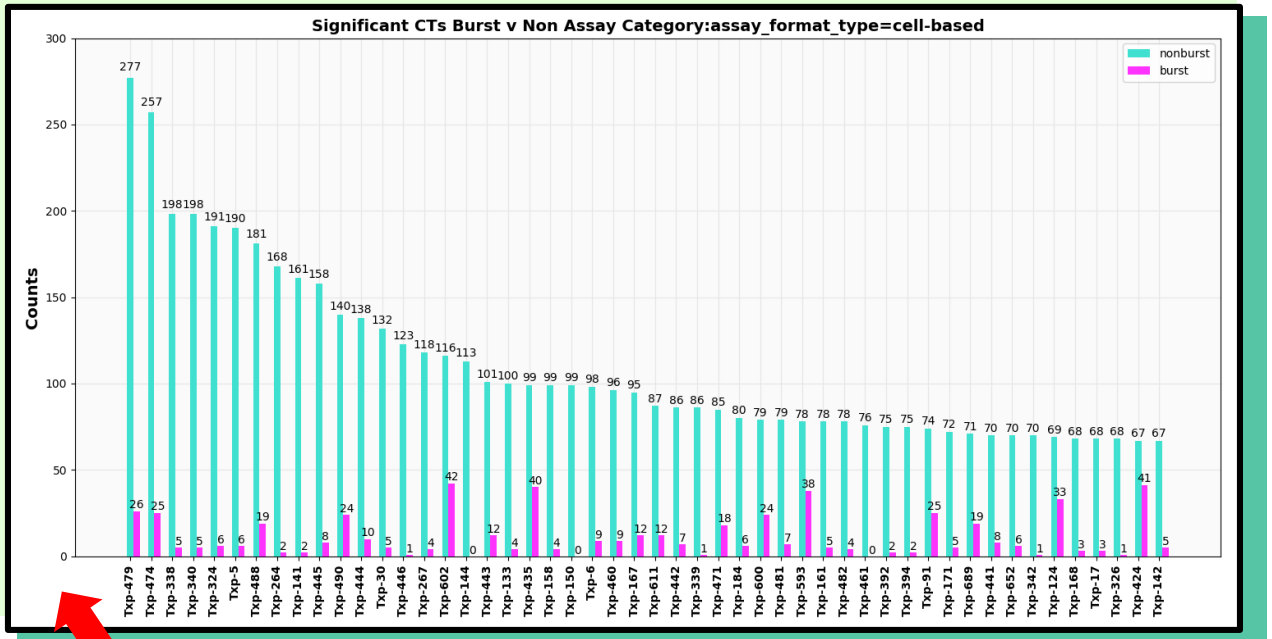


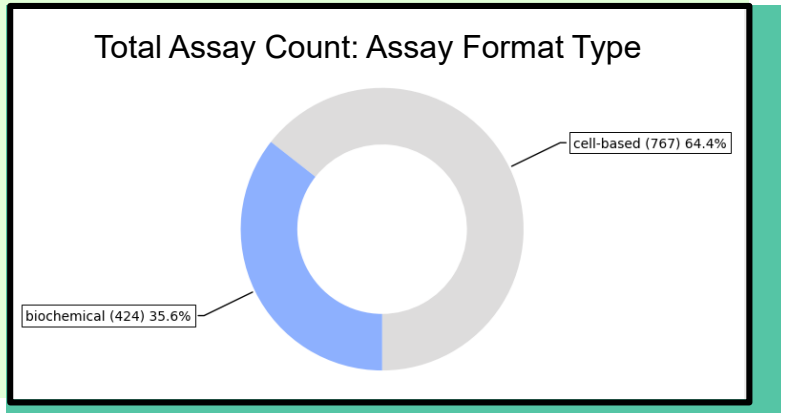


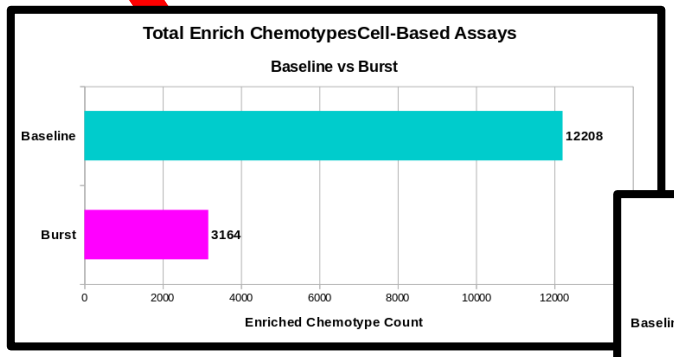


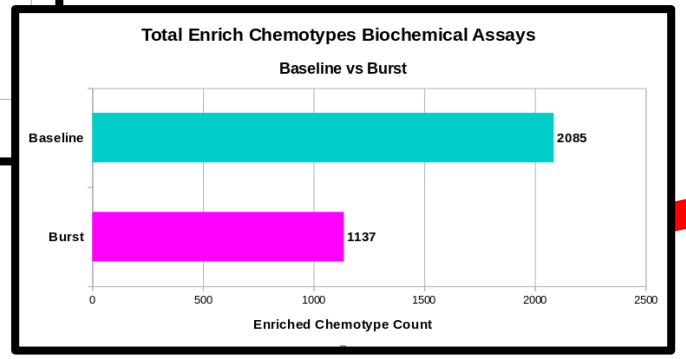
- Surfactants
- Pharmaceuticals (primarily trimethyl)
- Environmental Pollutants (wastewater/sewage)

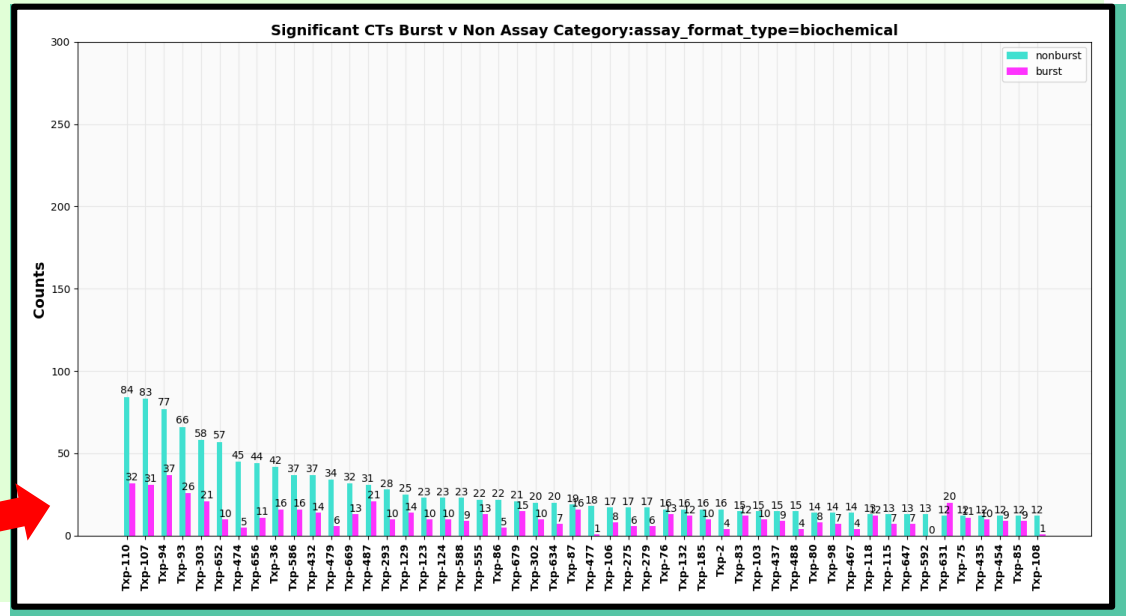
Cell-Based vs Biochemical Assays





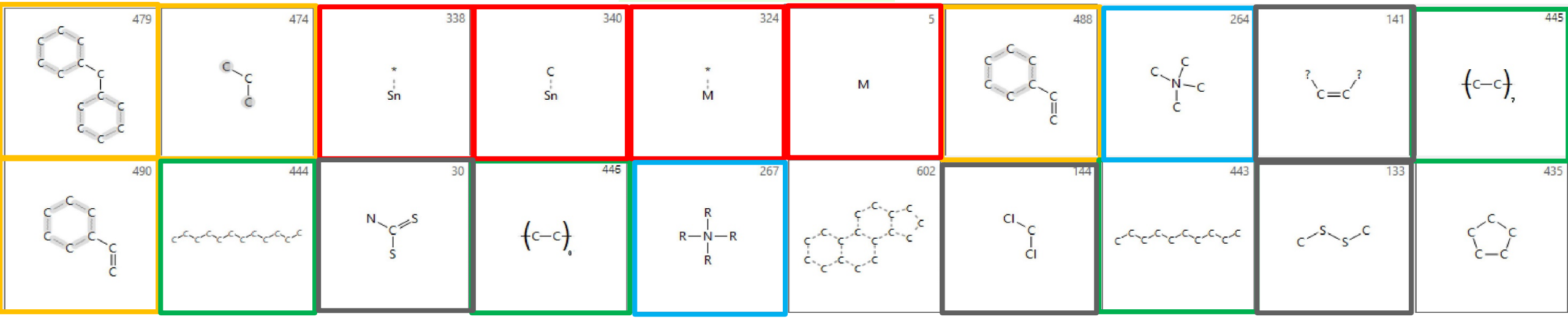




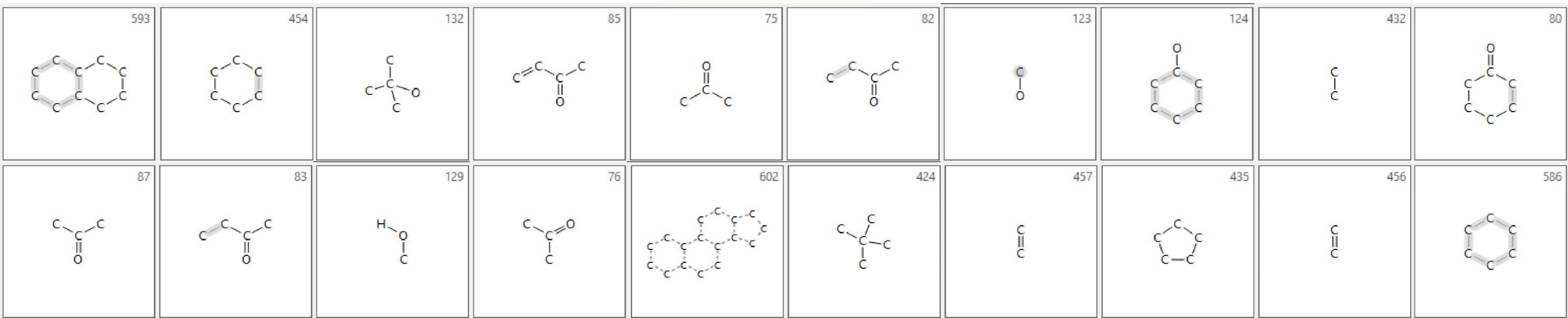


Cell-Based Assay Chemotypes

Baseline top 20

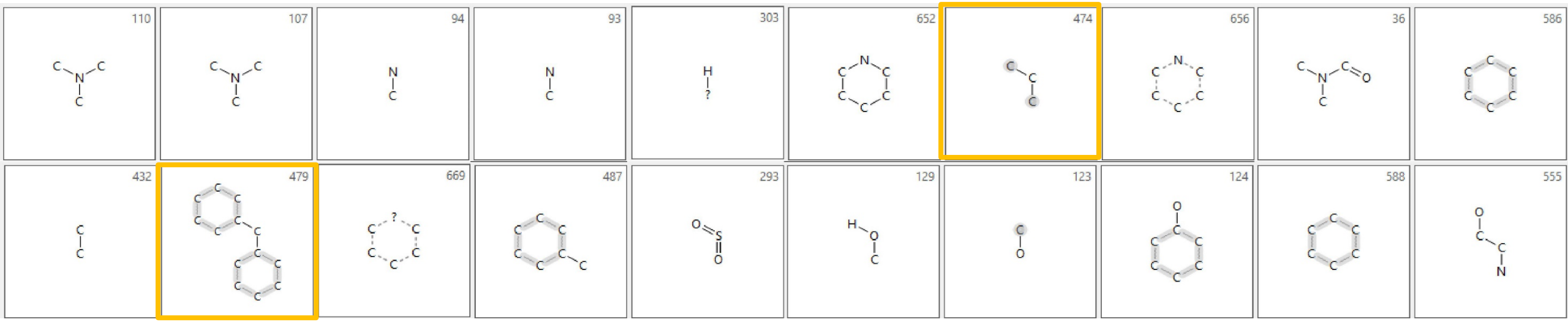


Burst top 20

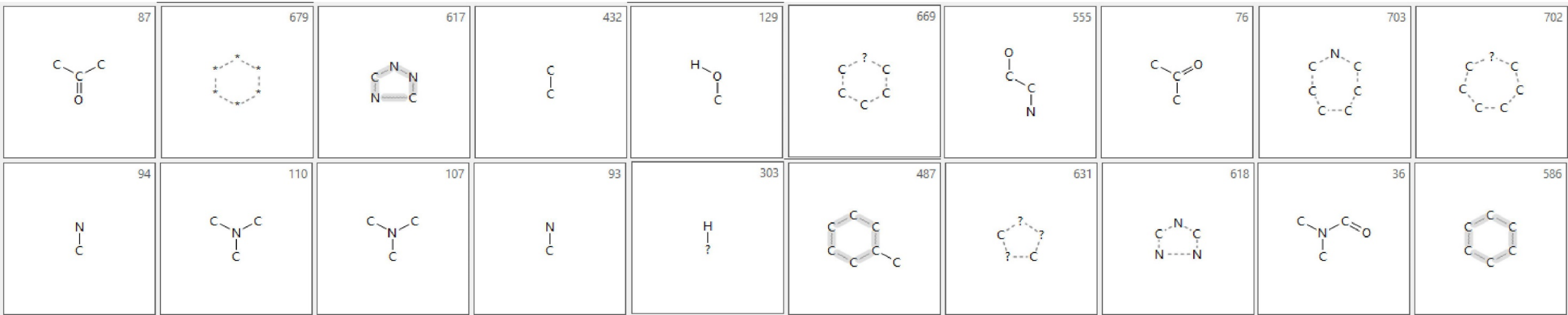


Biochemical Assay Chemotypes

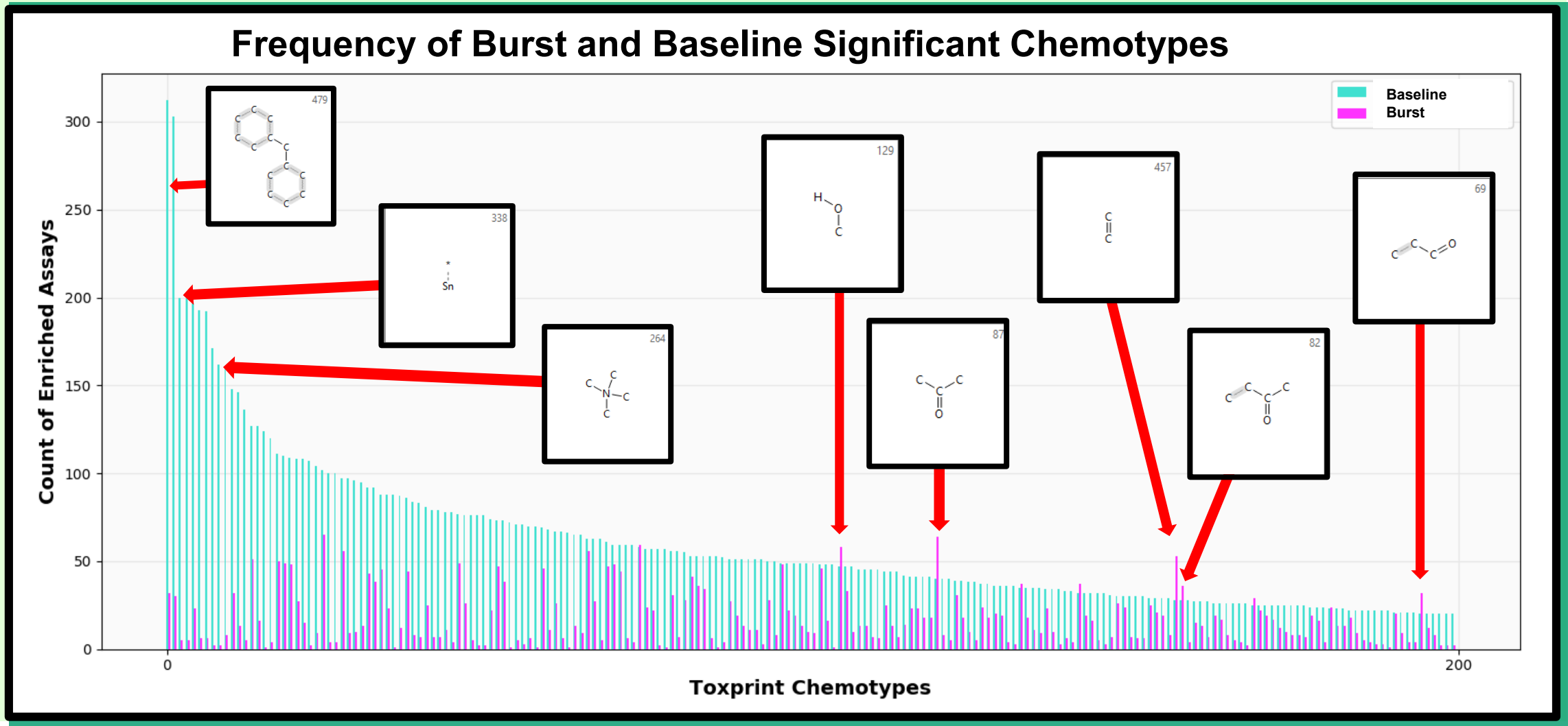
Baseline top 20



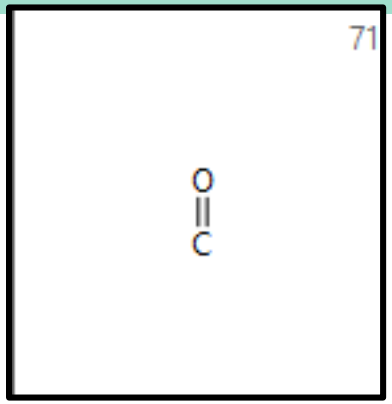
Burst top 20

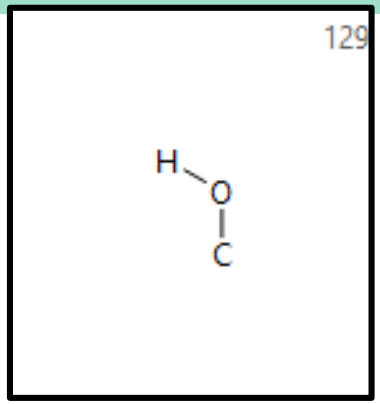


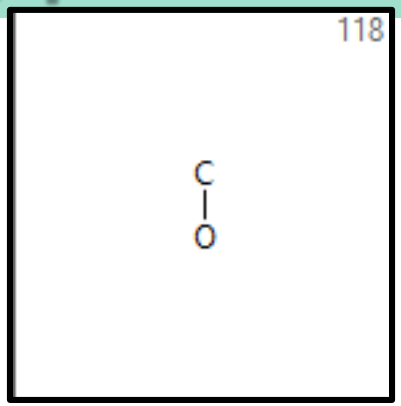
Enriched Chemotypes Baseline vs Burst Overview

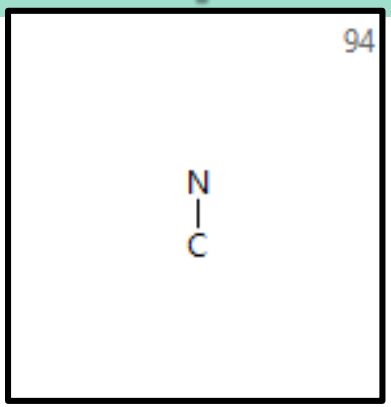


Are some Chemotypes too simple?









What if we look at combinations?

Two CLIs have been created in order to look at combinations of Toxprints

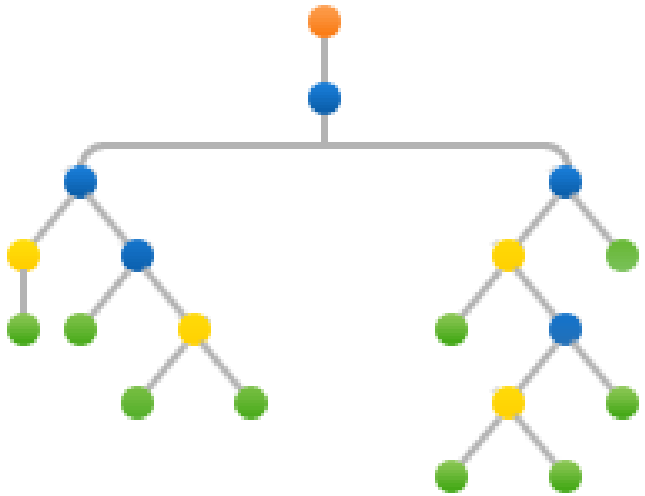
```
$ Toxprint_Combination_Generator
$ Special_Toxprints -f
```

Looks at all pairwise combinations

Generates Fingerprints for specific combinations of Toxprints

While these methods work well and have their uses, there is a better way

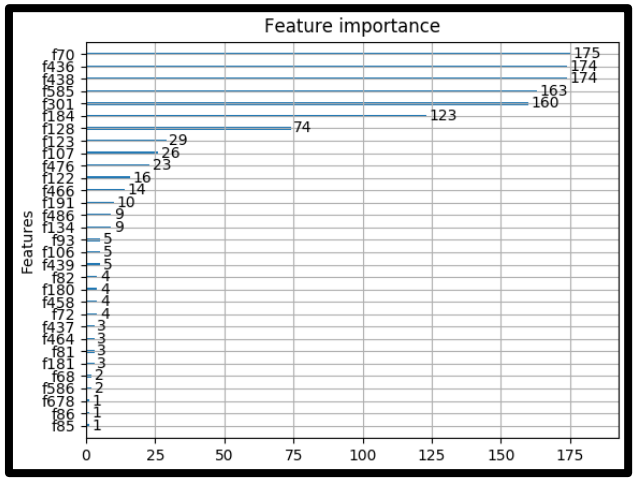
XGBoost can show us potential significant combinations

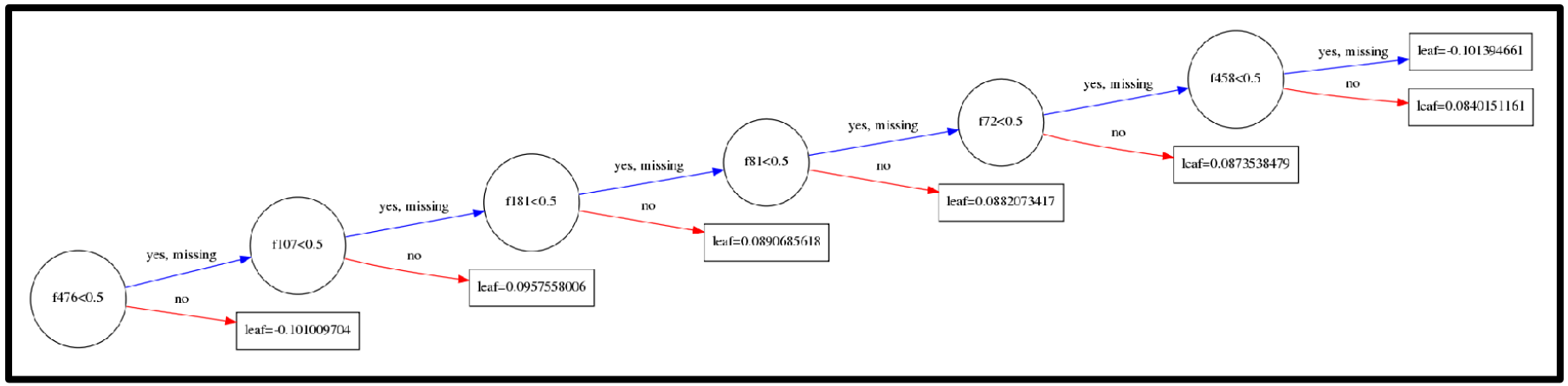


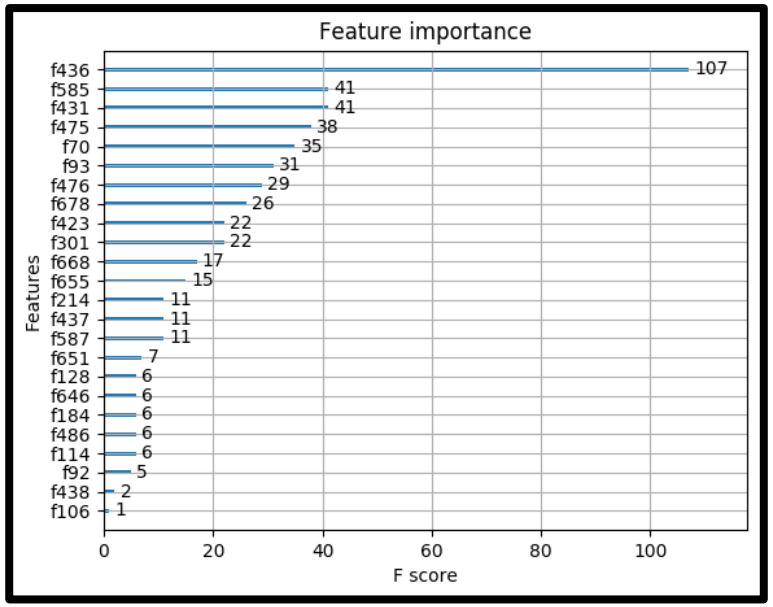
Why use XGBoost?

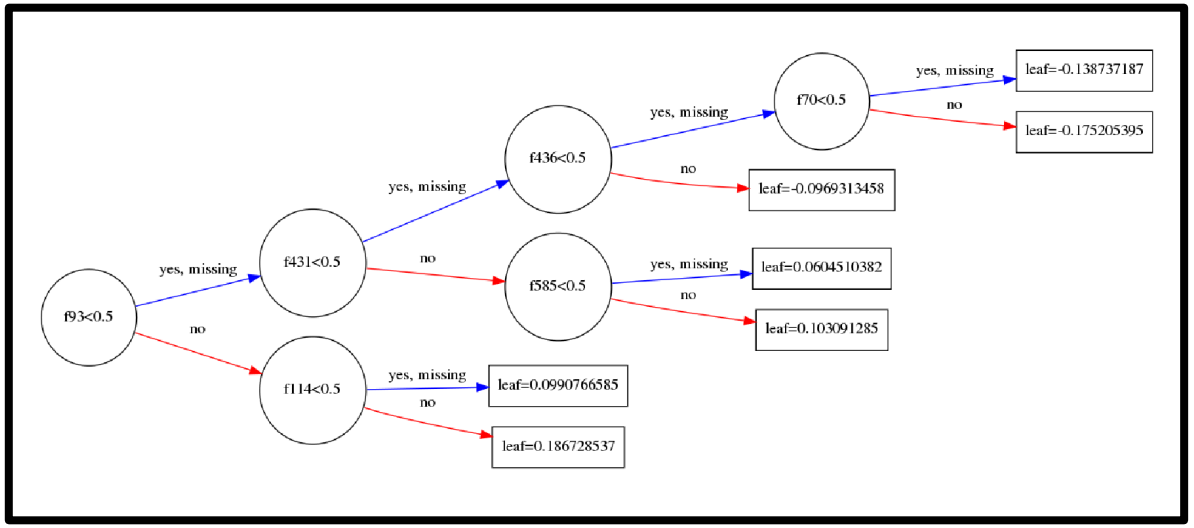
- Allows for highly imbalanced datasets (like many of ours)
- Regression and Classification Model
- Reproducible (Can reproduce Models from a random seed)
- Understandable (Can visualize the full decision tree)
- High Performance (XGBoost wins many Kaggle competitions)
- Warm-Start

Can we use CTEW results to inform our XGB model construction?









Concluding Remarks

- Chemotype-Enrichment workflow useful for evaluating biological activity thresholds on a chemical level
- CTEW used to evaluate QSAR models and assist with examining combinations of fingerprints/features
- Approach completely general, can be applied to any binary "activity" dataset (e.g., in vivo or in vitro bioassays, functional use categories, etc)
- Elements of workflow are being integrated into the publicly available USEPA Comptox Chemicals Dashboard
- CTEW shows great promise for elucidating chemical signals across assay space and supporting Comptox research

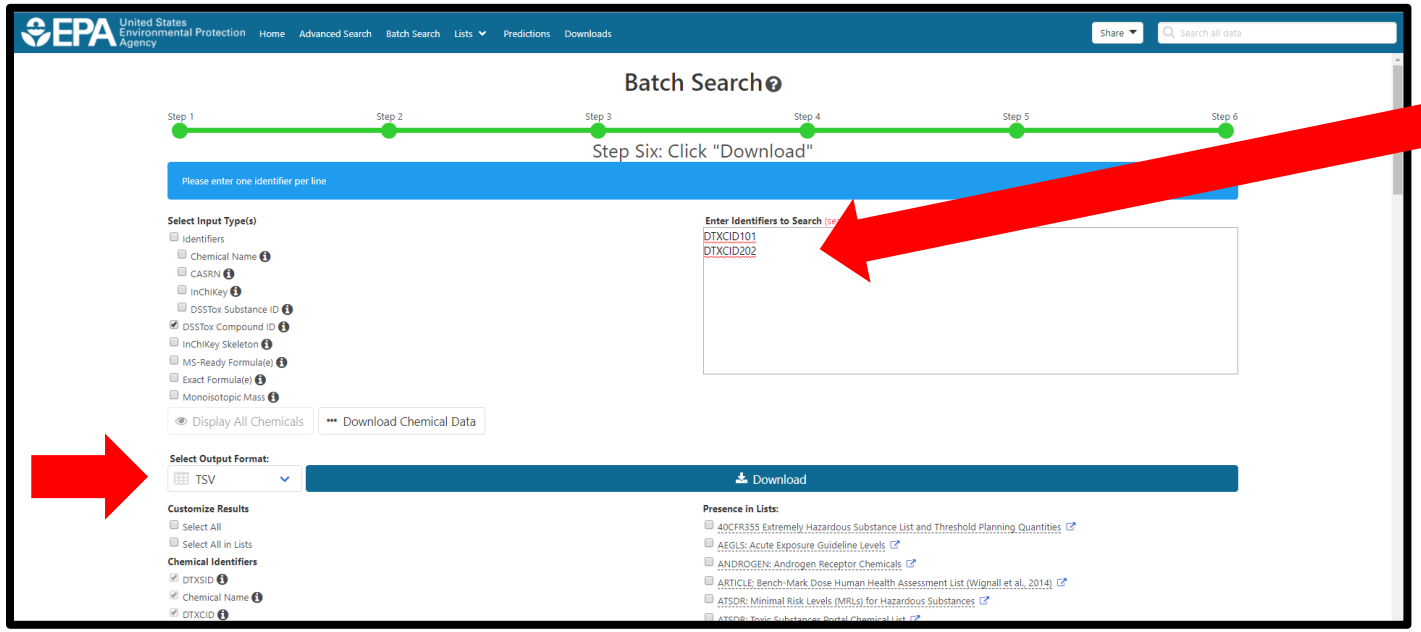
Acknowledgments



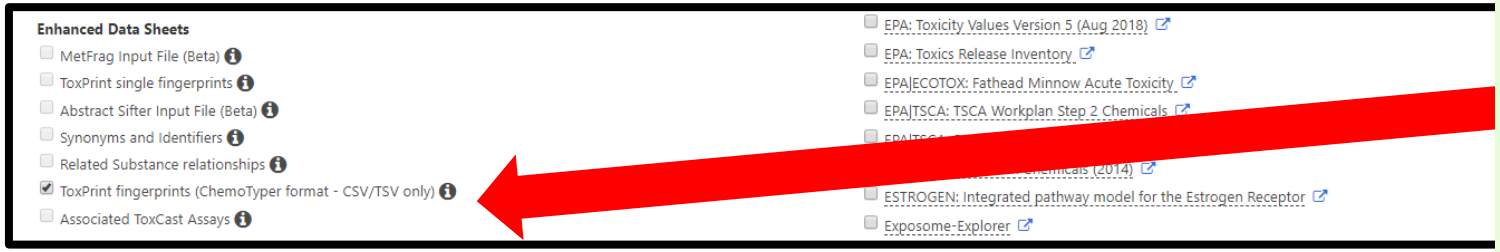

- Ann Richard
- Chris Grulke
- Antony Williams
- NCCT Staff
- NCCT dev team

You can currently get Toxprints from the dashboard

https://comptox.epa.gov/dashboard/dsstoxdb/batch_search



- Go to the dashboard's batch search
- Enter your chemical IDs
- Select output format (TSV)

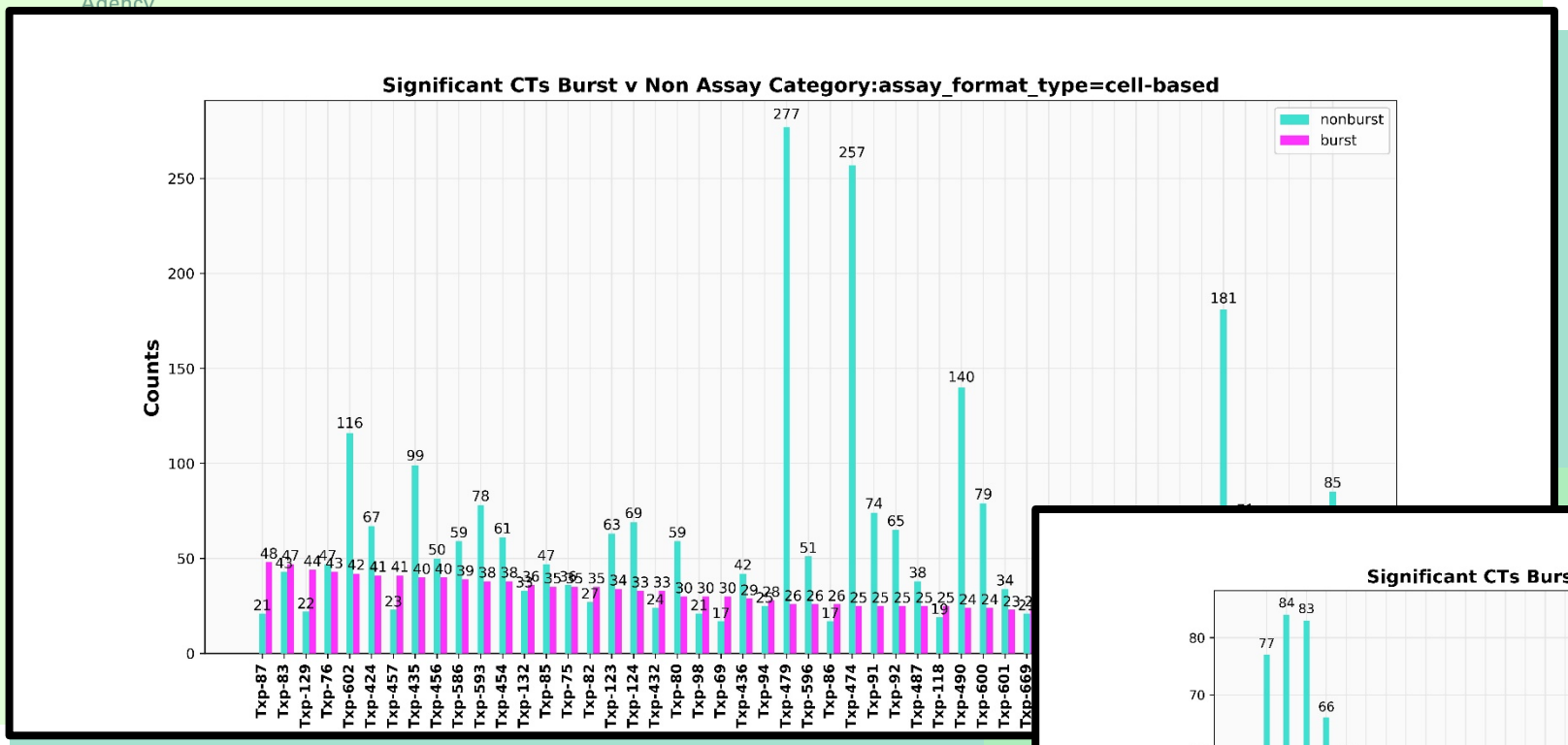


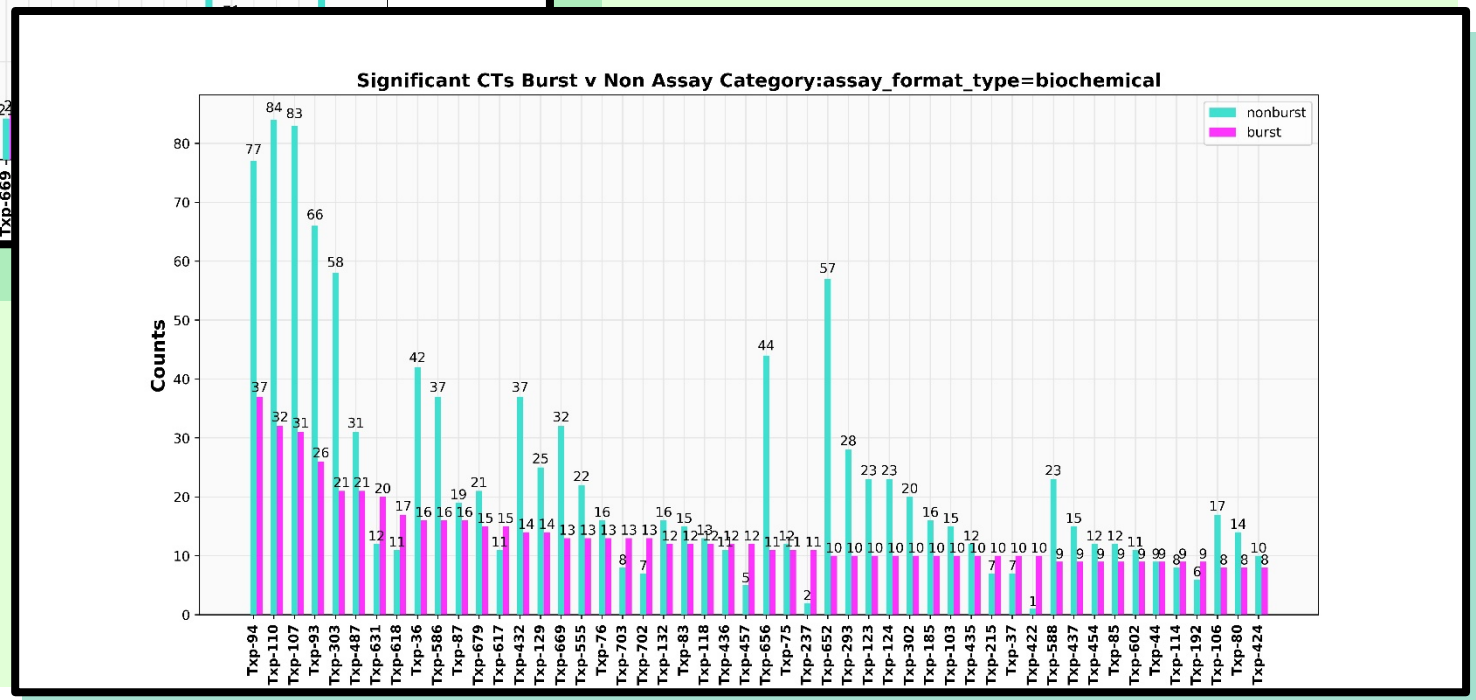
- Scroll down and under Enhanced Data Sheets, select either Toxprints Single Fingerprints or ToxPrint fingerprints ChemoTyper format

Thanks to Molecular Networks for providing Toxprint Generation Code and Images!

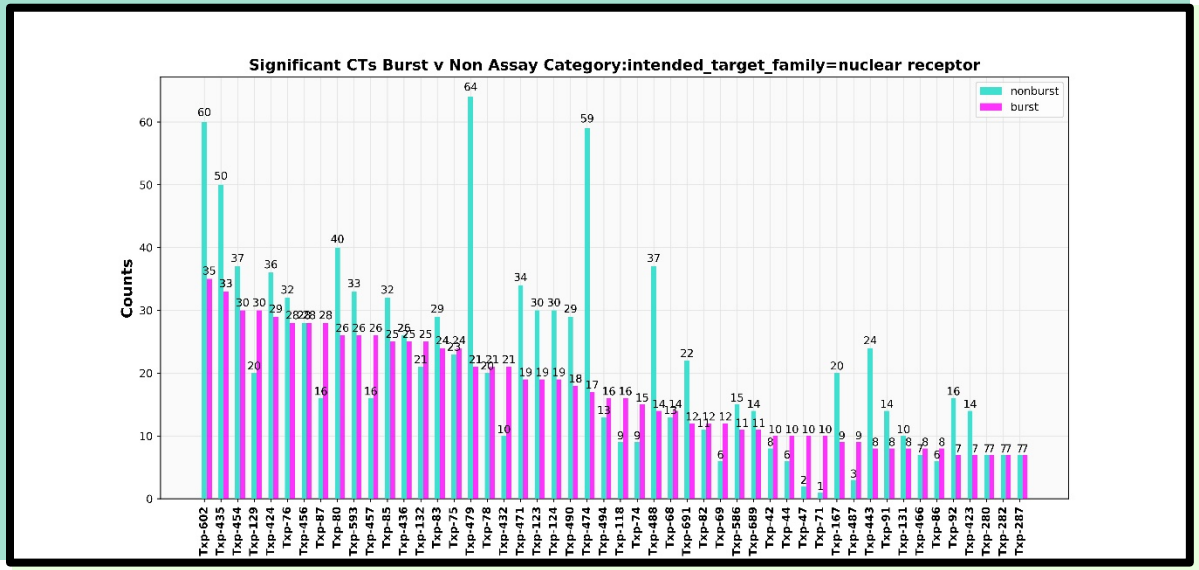
QUESTIONS?

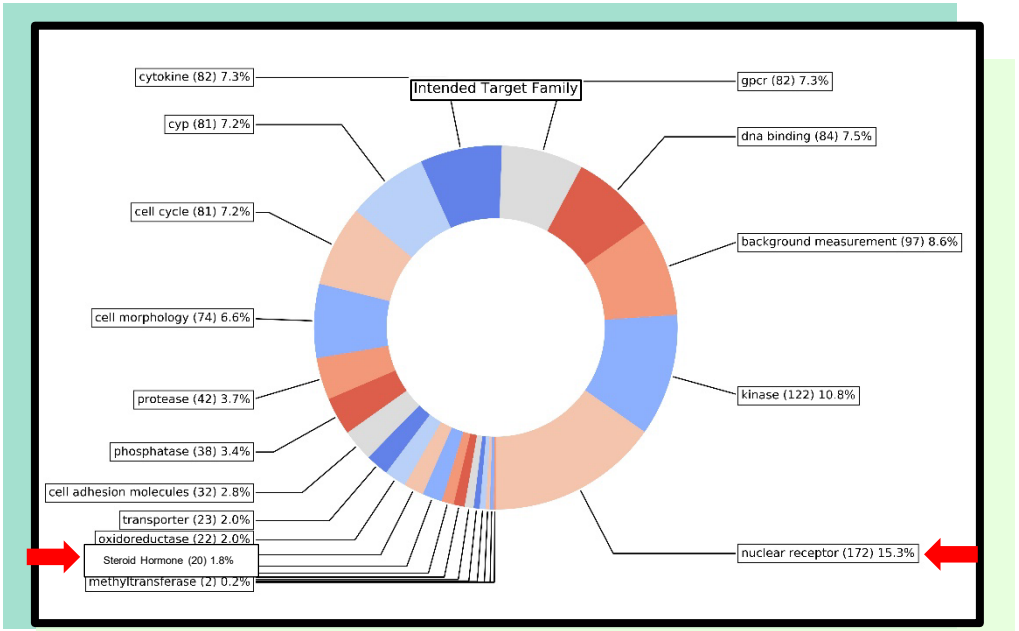
Cell-Based vs Biochemical Assay Chemotypes

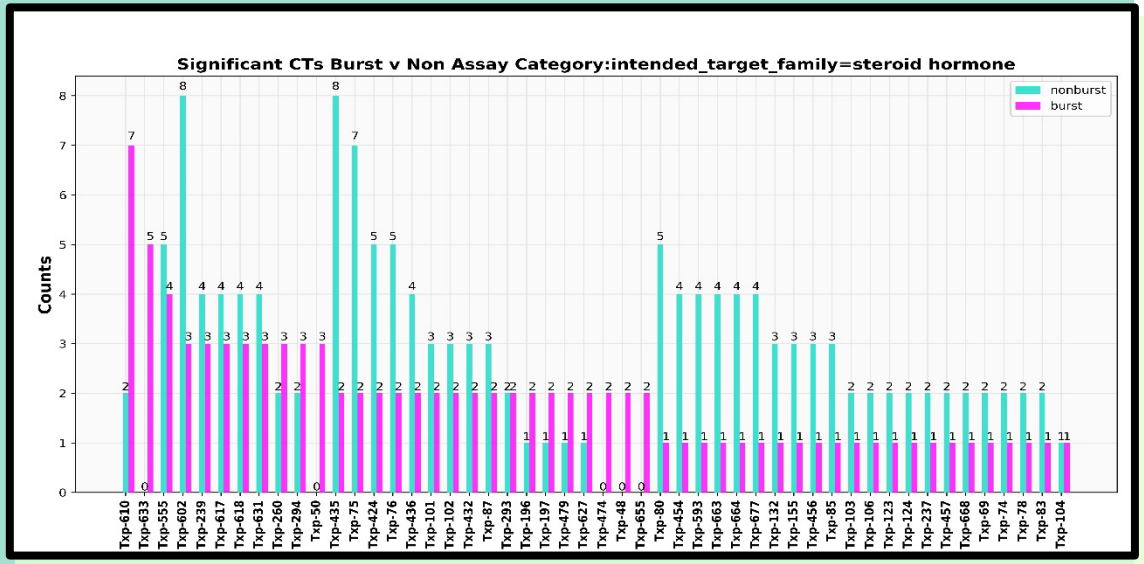




Intended Target Family

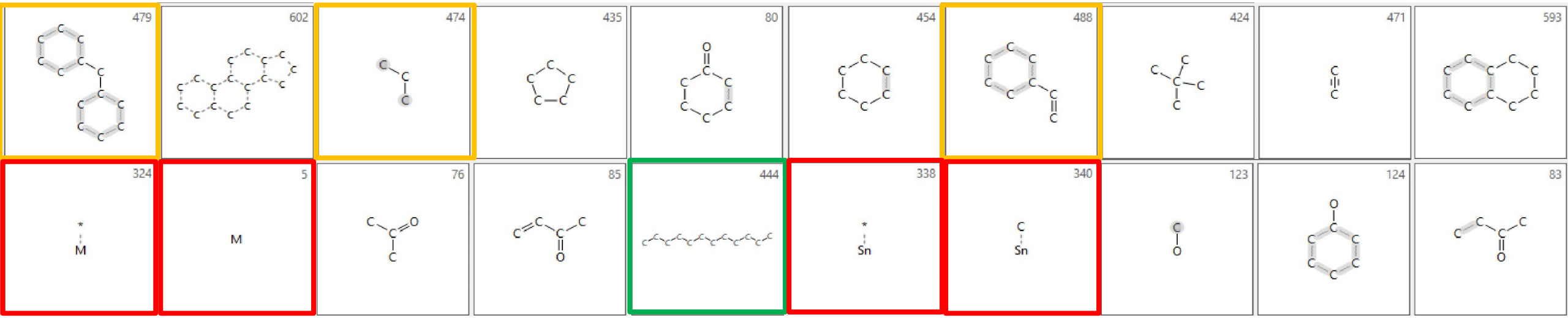




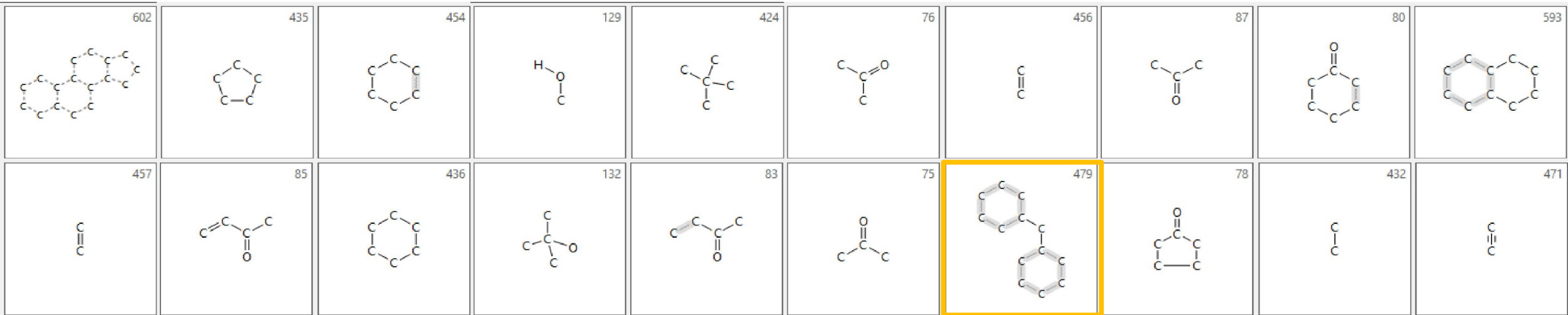


Intended Target Family Nuclear Receptor

Baseline top 20

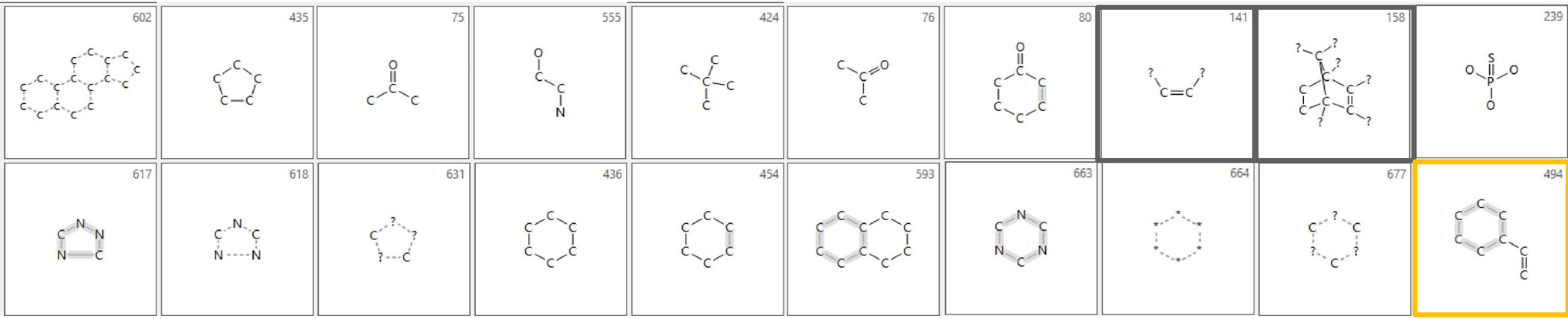


Burst top 20

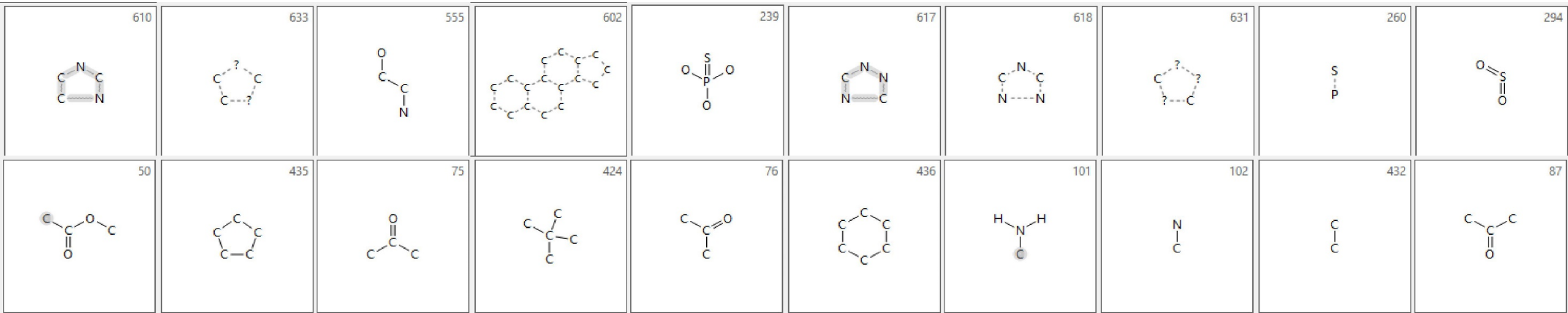


Intended Target Family Steroid Hormone

Baseline top 20

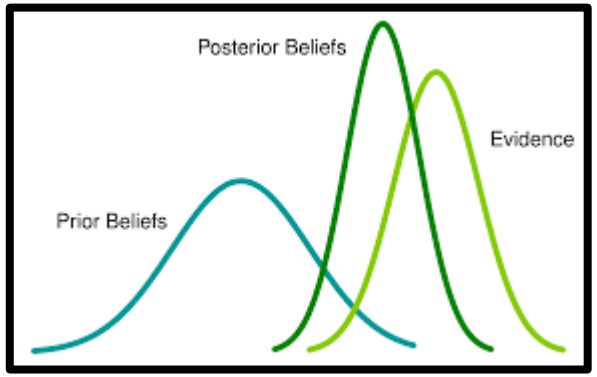


Burst top 20



XGBoost Parameter Tuning

Bayesian Optimization in Python with Hyperopt

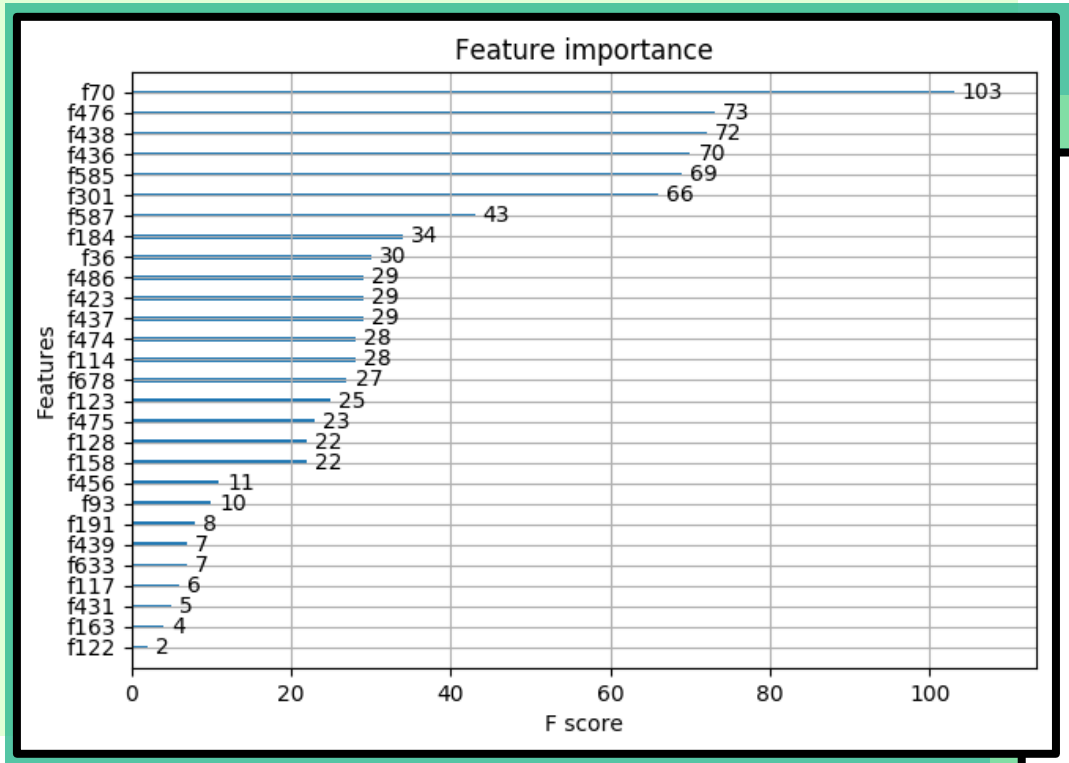


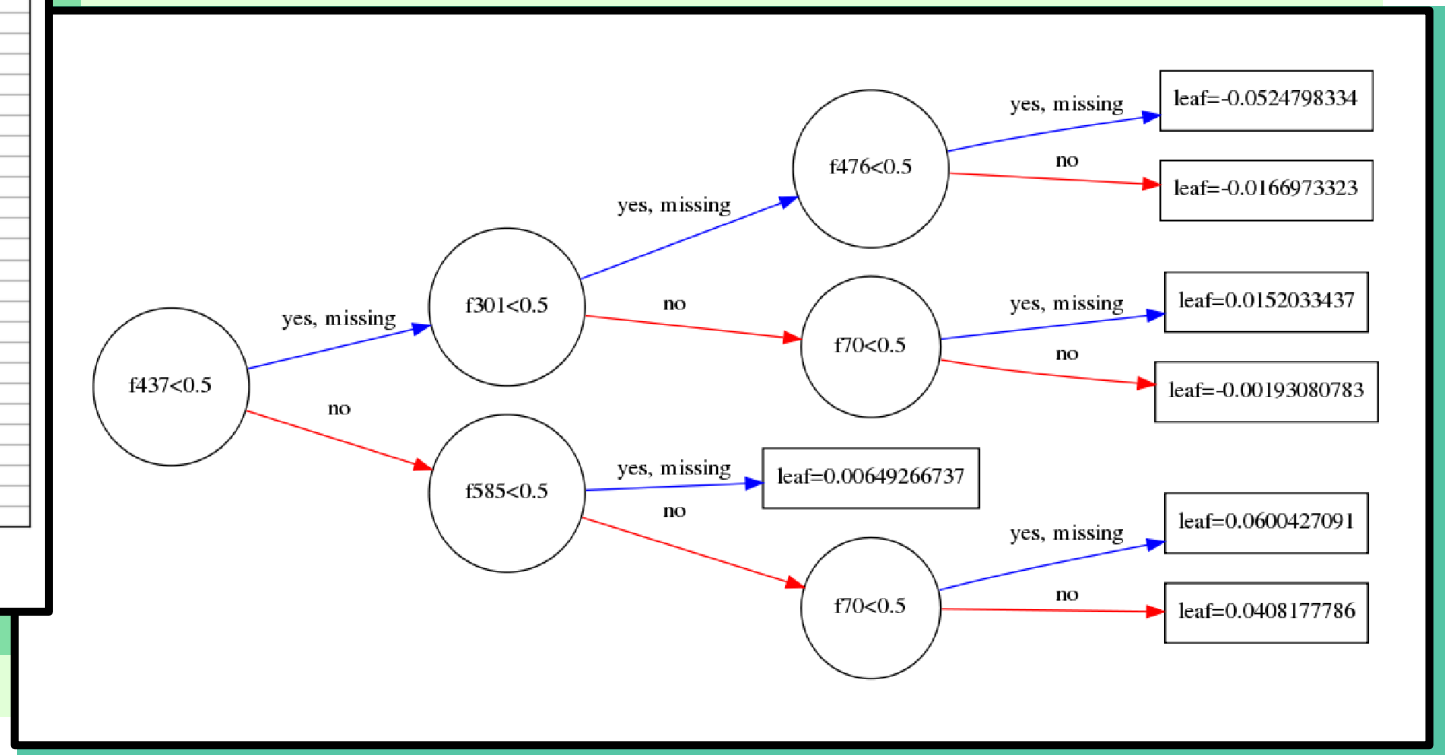
- Parameter tuning is typically done with grid-search (exhaustive)
- Instead we can use previous parameters as priors using bayesian inference
- Greatly reduce computation time
- Improved performance

Cytotoxicity QSAR model XGBoost

Cytotoxicity Assay: BSK_BE3C_SRB_down

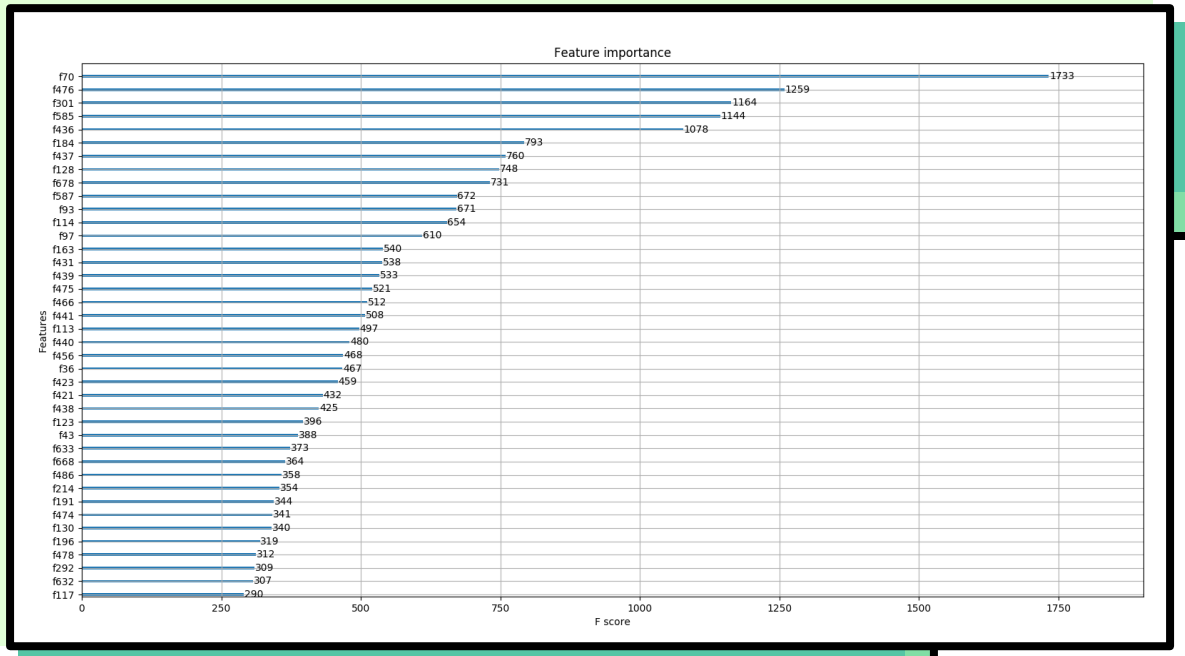
Accuracy: 0.733
Performance of Test: 0.658



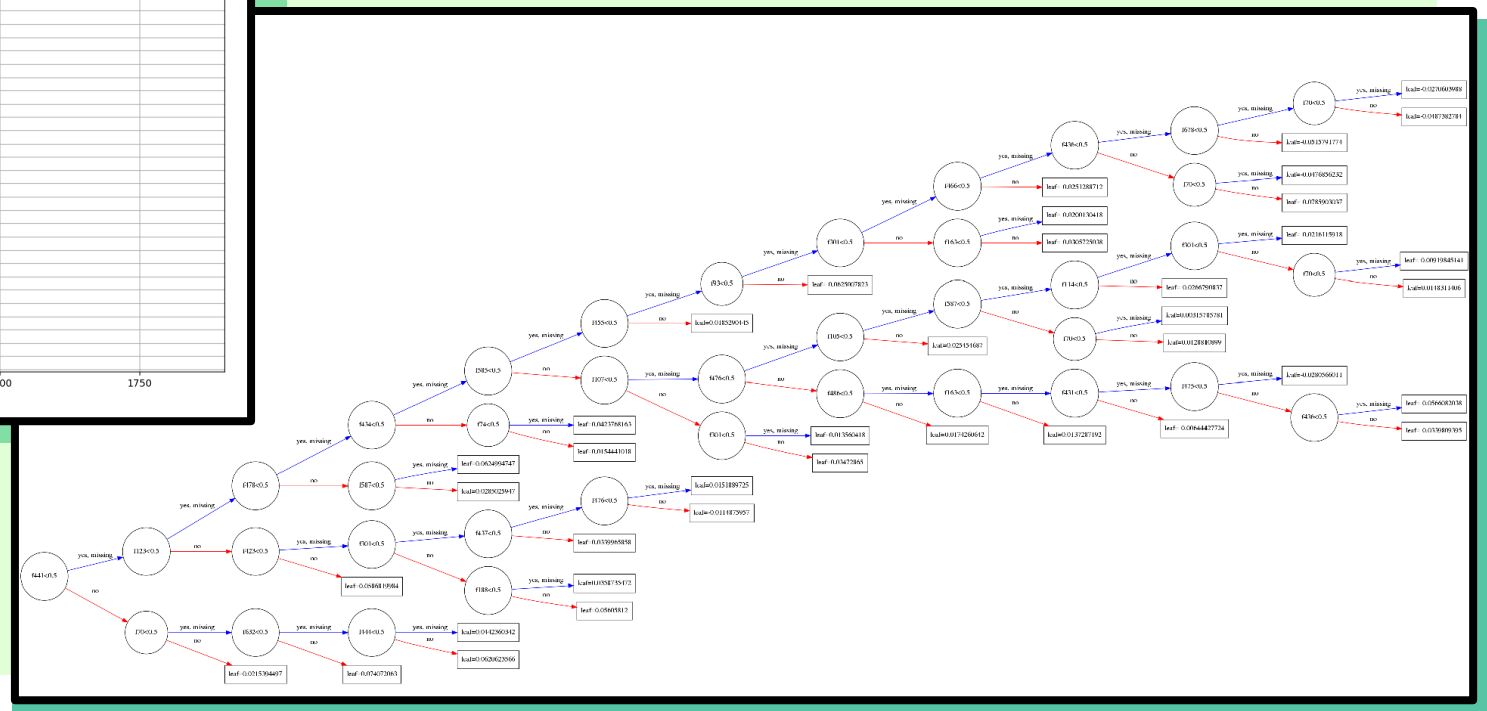


Cytotoxicity QSAR model XGBoost

Cytotoxicity Assay: TOX21_GR_BLA_Antagonist_viability

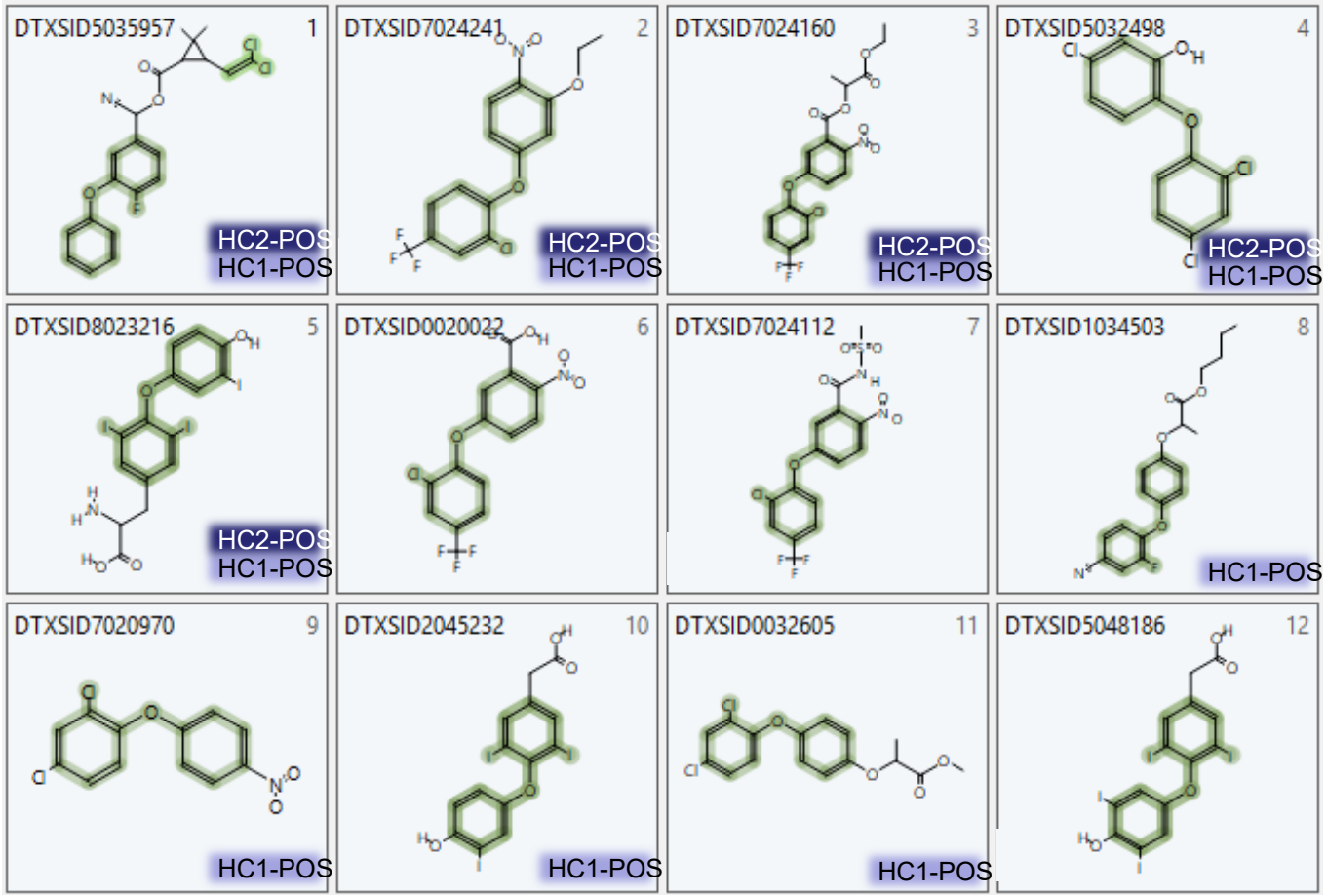


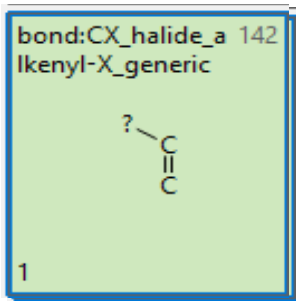
Accuracy: 0.909
Performance of Test: 0.718

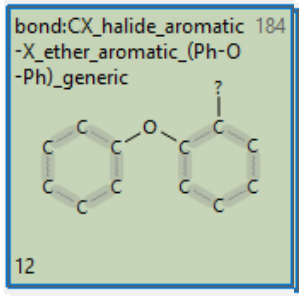


NIS Activation Assay

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?