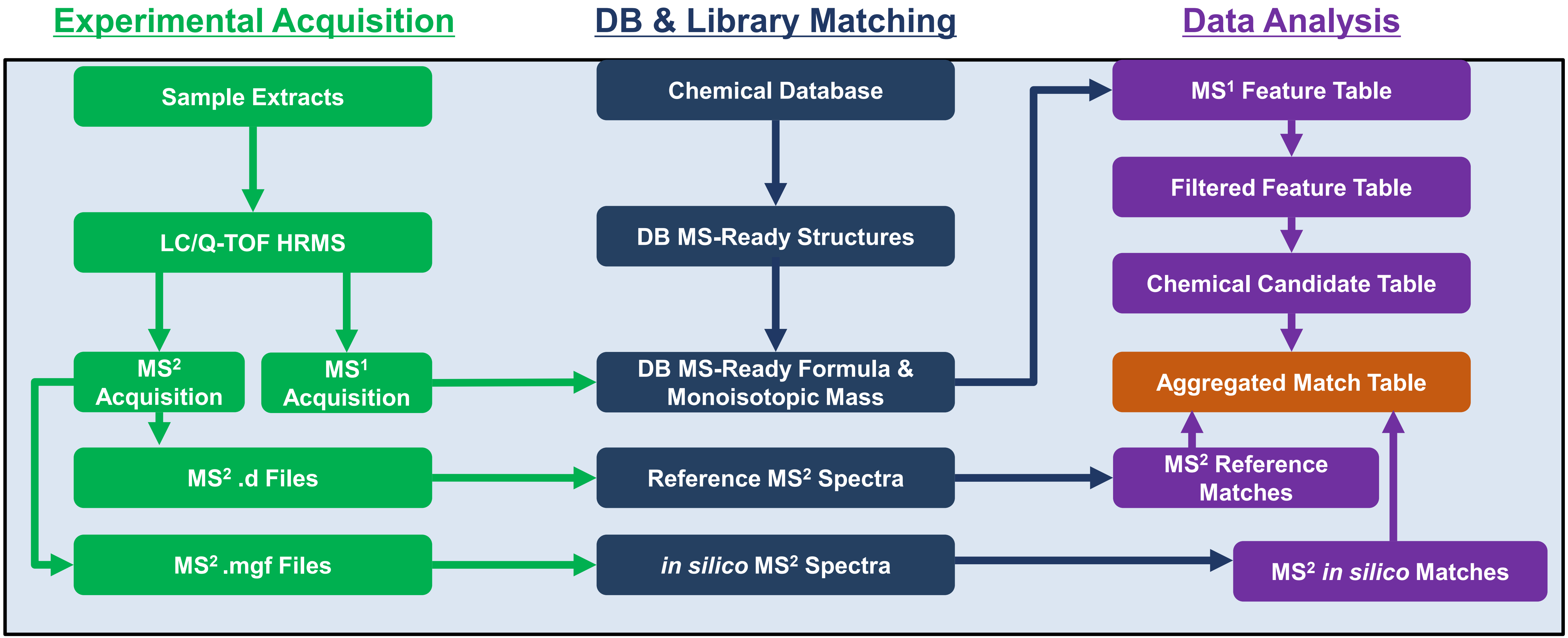
United States
Environmental Protection
Agency

Building a Non-Targeted Analysis Research Program at the U.S. EPA

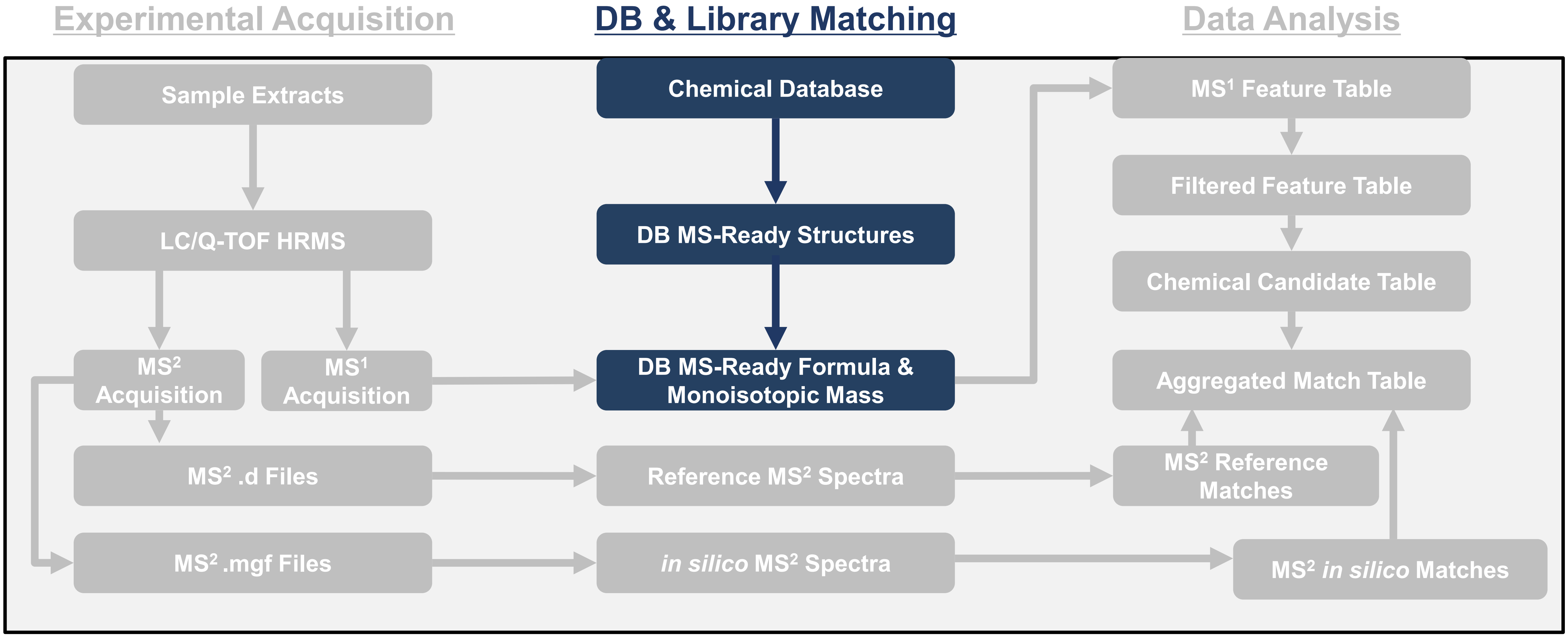
Jon R. Sobus, Ph.D. & the EPA/ORD NTA Team

Center for Computational Toxicology and Exposure
Research Triangle Park, NC

Agilent LC/Q-TOF Simplified Workflow

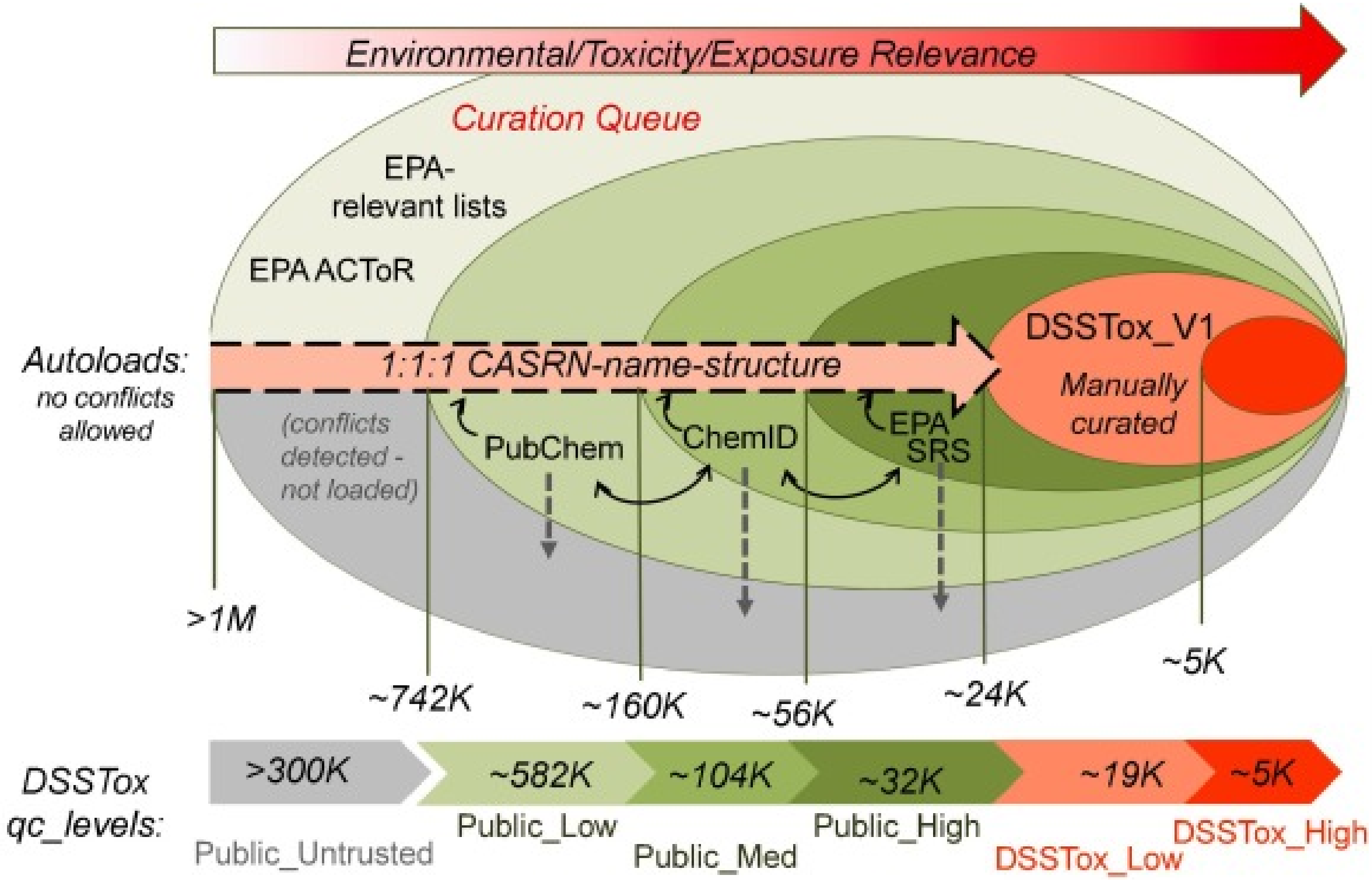


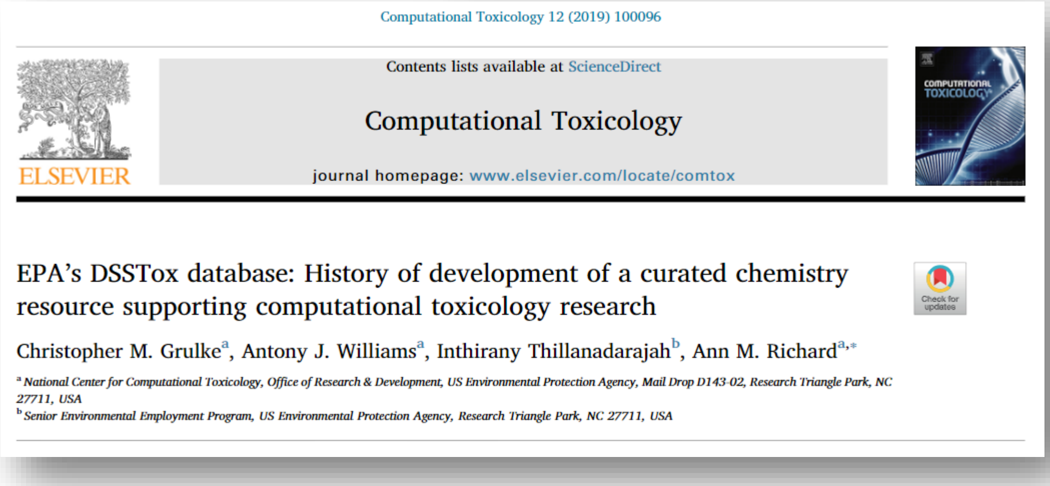
Agilent LC/Q-TOF Simplified Workflow



EPA
United States
Environmental Protection
Agency

Chemical Database = DSSTox





Computational Toxicology 12 (2019) 100096

Contents lists available at ScienceDirect

ELSEVIER

Computational Toxicology

journal homepage: www.elsevier.com/locate/comtox

EPA's DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research

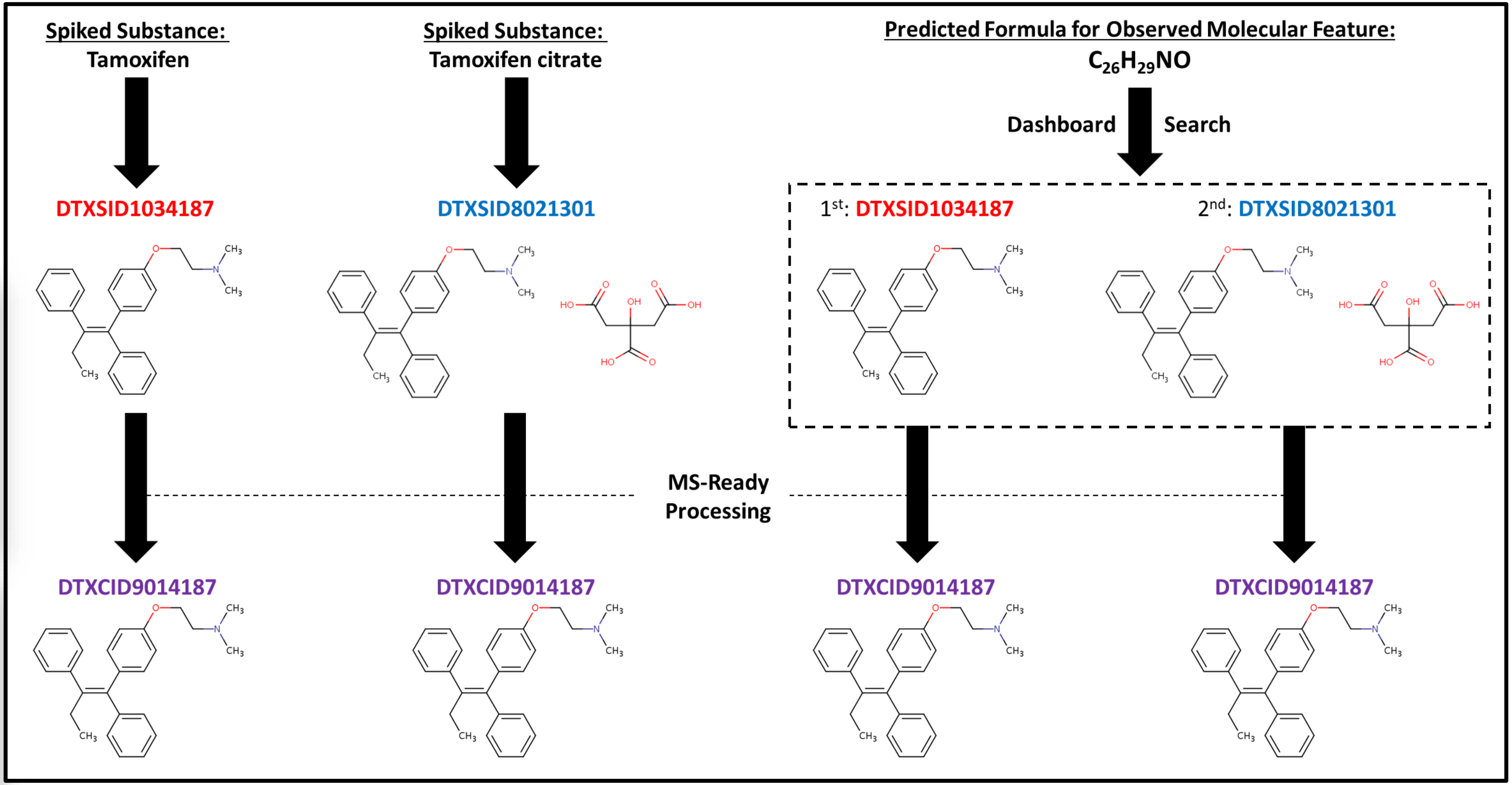
Christopher M. Grulke[a], Antony J. Williams[a], Inthirany Thillanadarajah[b], Ann M. Richard[a,*]

[a] National Center for Computational Toxicology, Office of Research & Development, US Environmental Protection Agency, Mail Drop D143-02, Research Triangle Park, NC 27711, USA

[b] Senior Environmental Employment Program, US Environmental Protection Agency, Research Triangle Park, NC 27711, USA

United States
Environmental Protection
Agency

MS-Ready Structures



Spiked Substance: Tamoxifen → DTXSID1034187

Spiked Substance: Tamoxifen citrate → DTXSID8021301

Predicted Formula for Observed Molecular Feature: $C_{26}H_{29}NO$

Dashboard Search

1st: DTXSID1034187

2nd: DTXSID8021301

MS-Ready Processing

DTXCID9014187

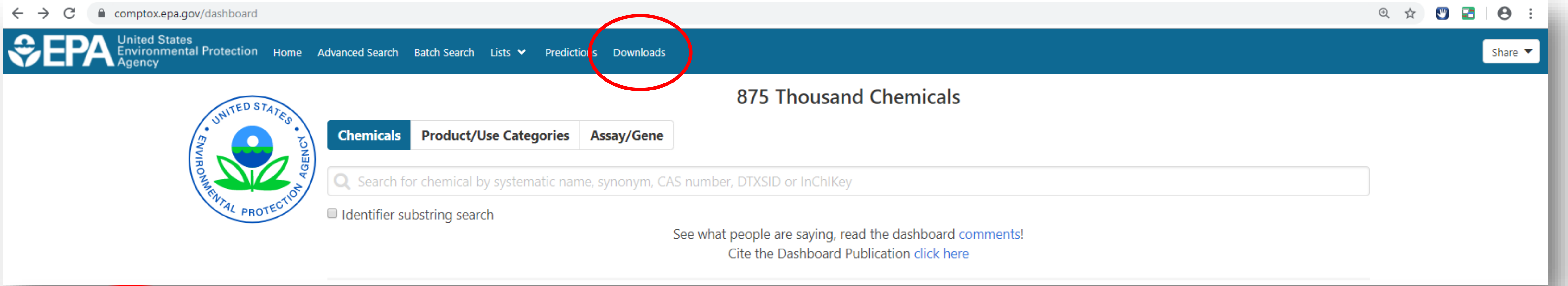
DTXCID9014187

DTXCID9014187

DTXCID9014187

United States Environmental Protection Agency

Dashboard Access



comptox.epa.gov/dashboard

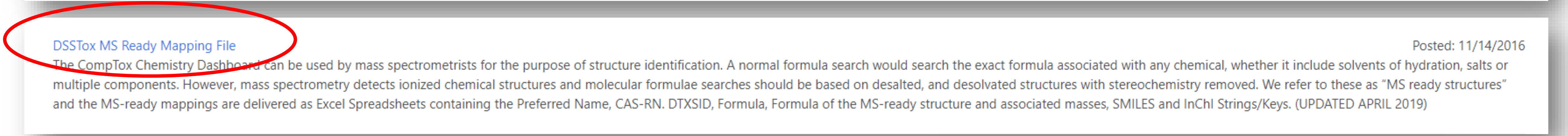
United States Environmental Protection Agency
Home
Advanced Search
Batch Search
Lists
Predictions
Downloads
Share

875 Thousand Chemicals

Chemicals
Product/Use Categories
Assay/Gene
Search for chemical by systematic name, synonym, CAS number, DTXSID or InChIKey

□ Identifier substring search

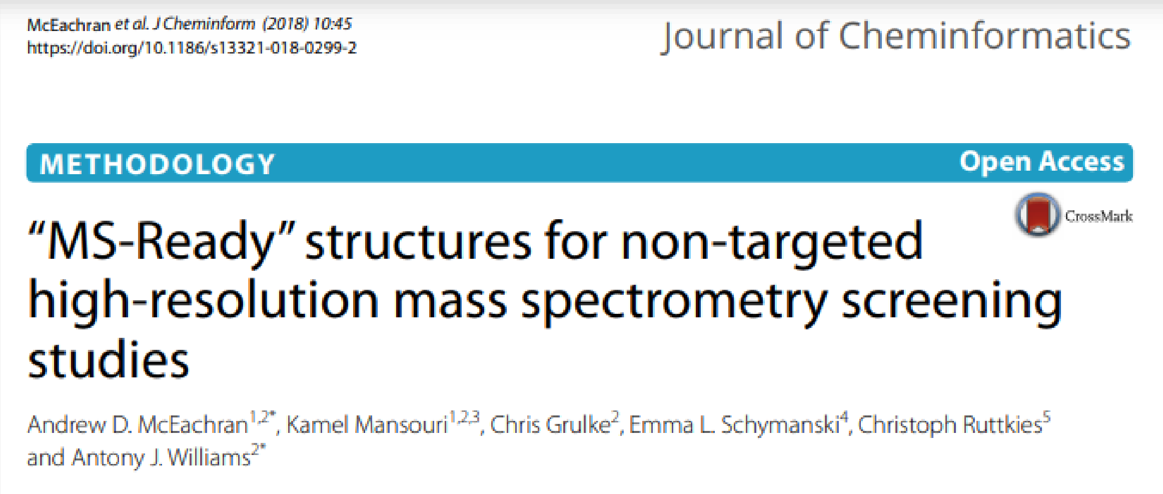
See what people are saying, read the dashboard comments!
Cite the Dashboard Publication click here



DSSTox MS Ready Mapping File

Posted: 11/14/2016

The CompTox Chemistry Dashboard can be used by mass spectrometrists for the purpose of structure identification. A normal formula search would search the exact formula associated with any chemical, whether it include solvents of hydration, salts or multiple components. However, mass spectrometry detects ionized chemical structures and molecular formulae searches should be based on desalted, and desolvated structures with stereochemistry removed. We refer to these as "MS ready structures" and the MS-ready mappings are delivered as Excel Spreadsheets containing the Preferred Name, CAS-RN. DTXSID, Formula, Formula of the MS-ready structure and associated masses, SMILES and InChI Strings/Keys. (UPDATED APRIL 2019)



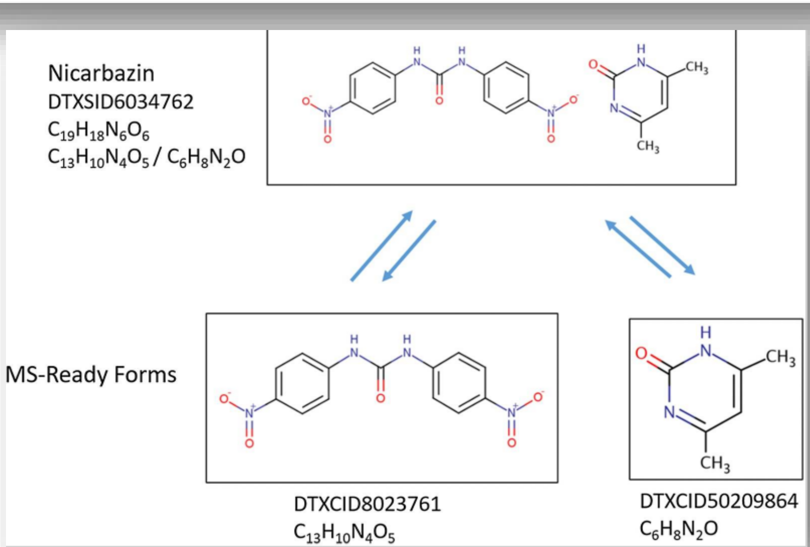
McEachran *et al. J Cheminform* (2018) 10:45
https://doi.org/10.1186/s13321-018-0299-2

Journal of Cheminformatics

METHODOLOGY Open Access

"MS-Ready" structures for non-targeted high-resolution mass spectrometry screening studies

Andrew D. McEachran[1,2*], Kamel Mansouri[1,2,3], Chris Grulke[2], Emma L. Schymanski[4], Christoph Ruttkies[5] and Antony J. Williams[2*]



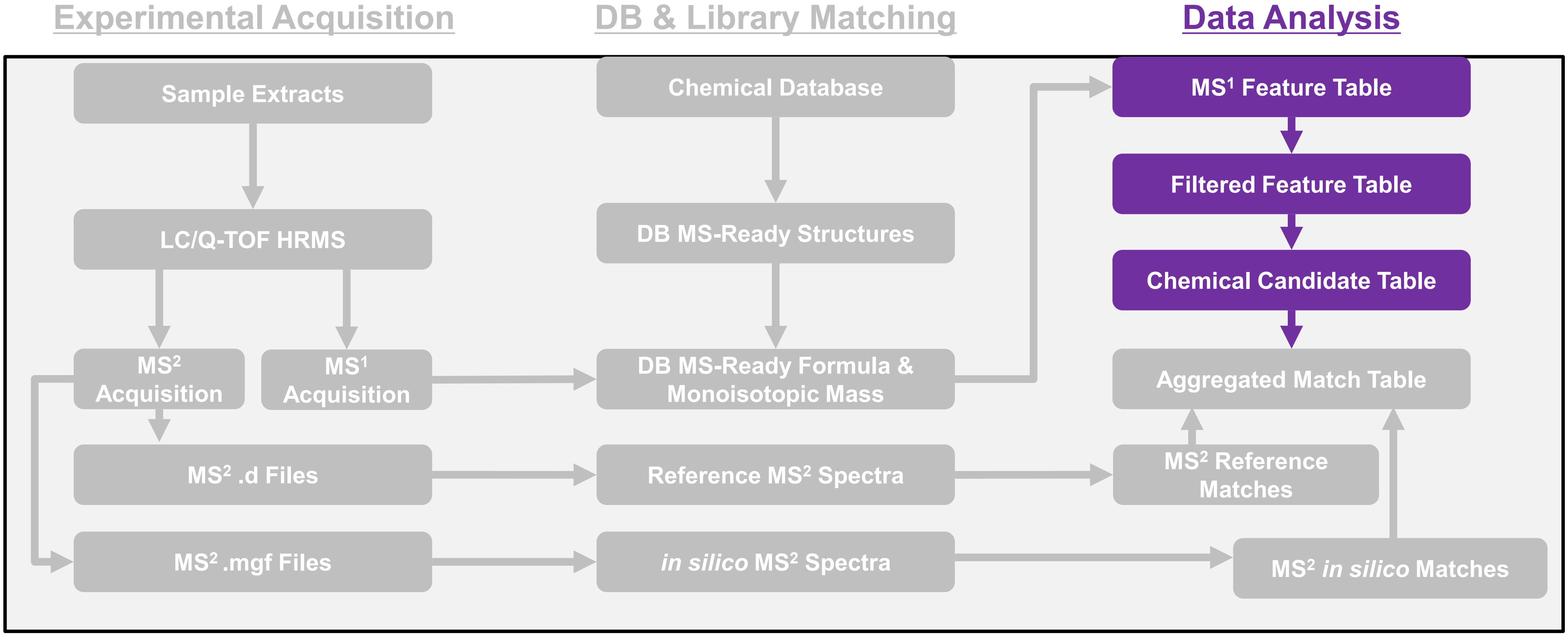
Nicarbazin
DTXSID6034762
$C_{19}H_{18}N_6O_6$
$C_{13}H_{10}N_4O_5$ / $C_6H_8N_2O$

MS-Ready Forms

DTXCID8023761
$C_{13}H_{10}N_4O_5$

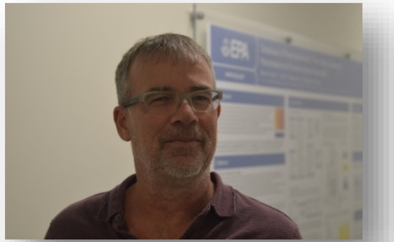
DTXCID50209864
$C_6H_8N_2O$

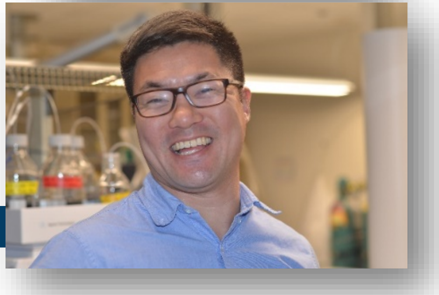
Agilent LC/Q-TOF Simplified Workflow



EPA
United States
Environmental Protection
Agency

EPA NTA WebApp





Feature Removal:

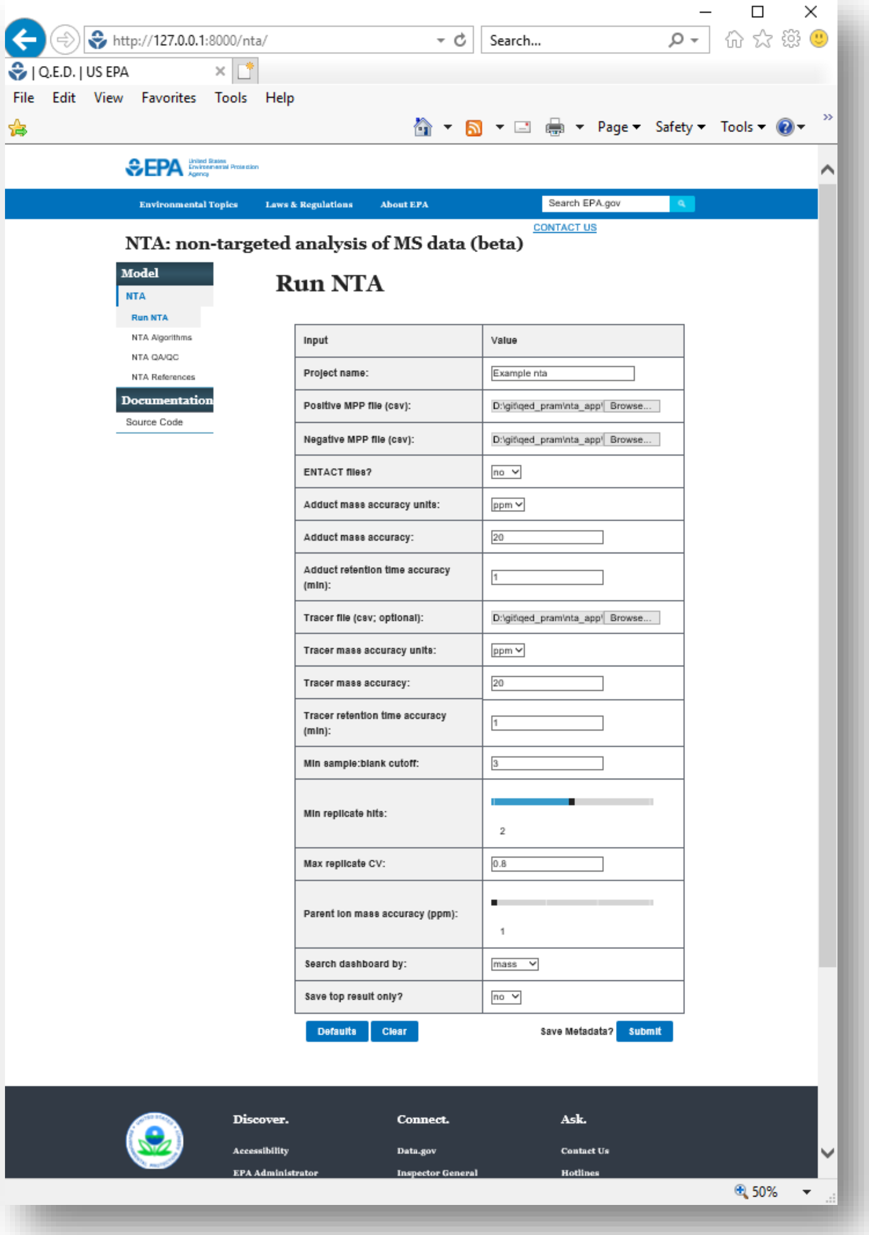
1) Duplicate features
2) Non-reproducible features
3) Blank features (sample:blank)
4) Non-responsive features (dilutions)

Feature Flagging:

1) Multi-mode hits (+ and -)
2) Meas. precision (CV threshold)
3) Formula match (score ≥ threshold)
4) Negative mass defect
5) Halogenation
6) Has/is adduct
7) Has/is neutral loss
8) Has/is multimer

Dashboard Integration:

1) Data source & pub counts
2) Bioactivity & exposure levels
3) Presence on lists
4) Product & use categories



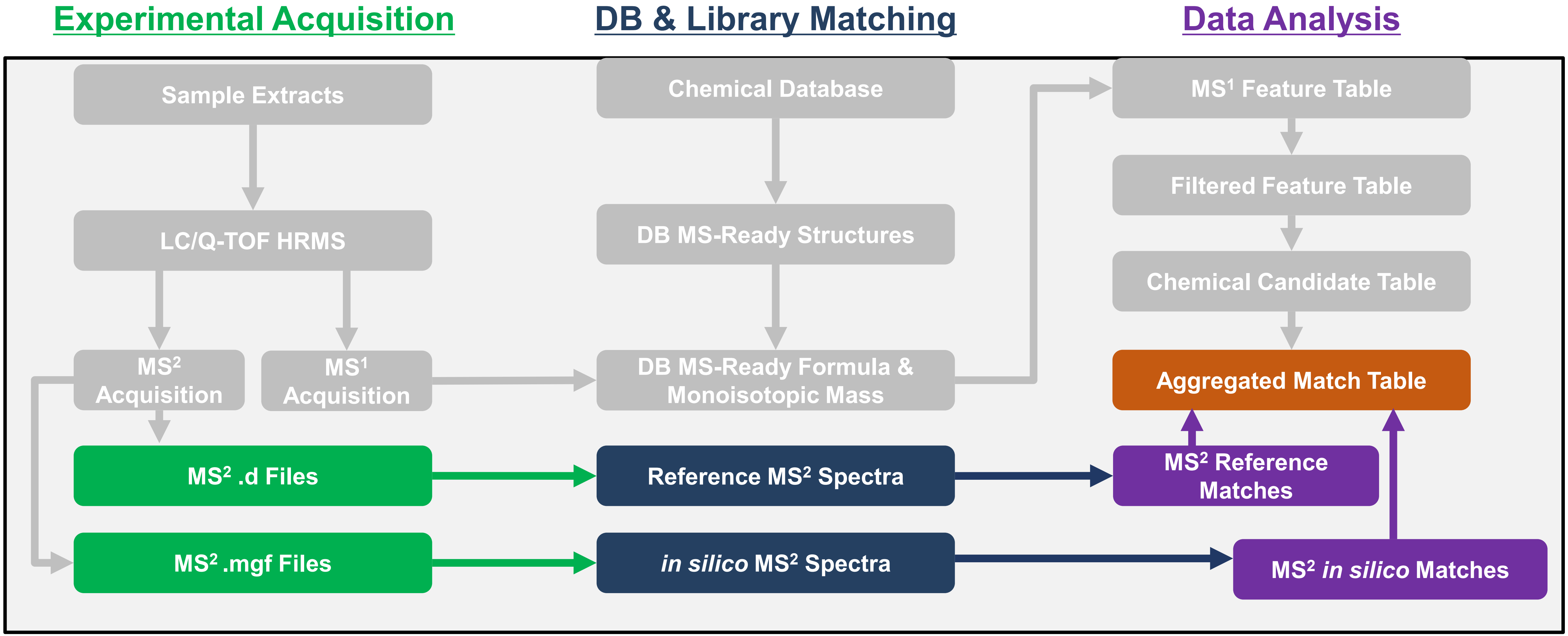
NTA: non-targeted analysis of MS data (beta)

Run NTA

Input	Value
Project name:	Example nta
Positive MPP file (csv):	D:\git\qed_pram\nta_app\ Browse...
Negative MPP file (csv):	D:\git\qed_pram\nta_app\ Browse...
ENTACT files?	no
Adduct mass accuracy units:	ppm
Adduct mass accuracy:	20
Adduct retention time accuracy (min):	1
Tracer file (csv; optional):	D:\git\qed_pram\nta_app\ Browse...
Tracer mass accuracy units:	ppm
Tracer mass accuracy:	20
Tracer retention time accuracy (min):	1
Min sample:blank cutoff:	3
Min replicate hits:	2
Max replicate CV:	0.8
Parent ion mass accuracy (ppm):	1
Search dashboard by:	mass
Save top result only?	no

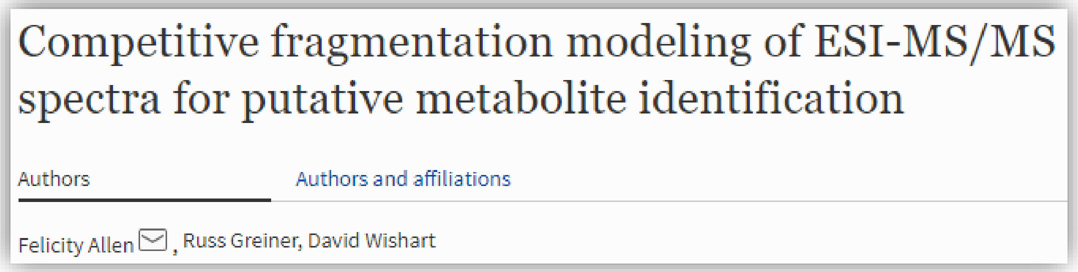
EPA United States Environmental Protection Agency

Agilent LC/Q-TOF Simplified Workflow



Generation of *in silico* Spectra

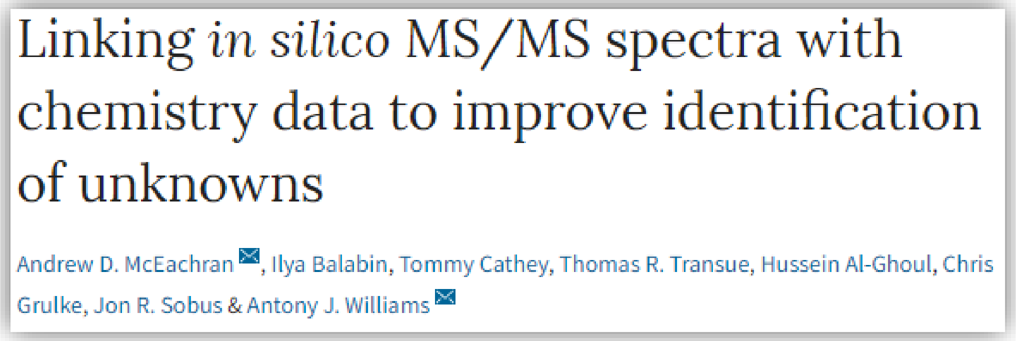
CFM-ID v2.0



Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification

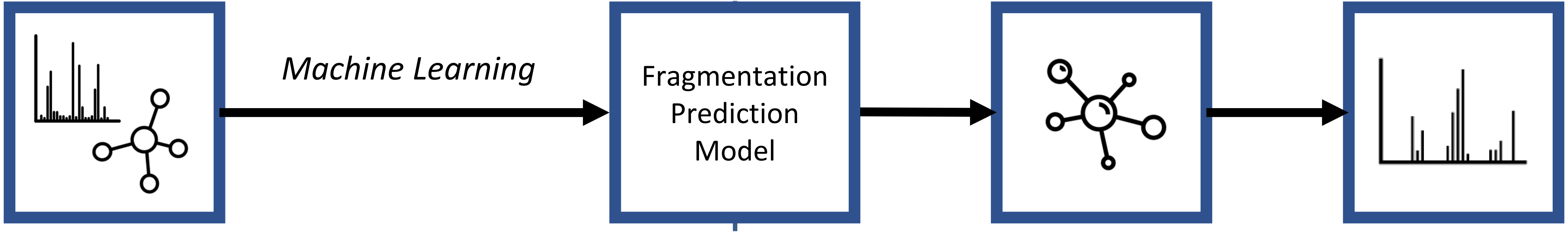
Authors | Authors and affiliations

Felicity Allen, Russ Greiner, David Wishart



Linking *in silico* MS/MS spectra with chemistry data to improve identification of unknowns

Andrew D. McEachran, Ilya Balabin, Tommy Cathey, Thomas R. Transue, Hussein Al-Ghoul, Chris Grulke, Jon R. Sobus & Antony J. Williams



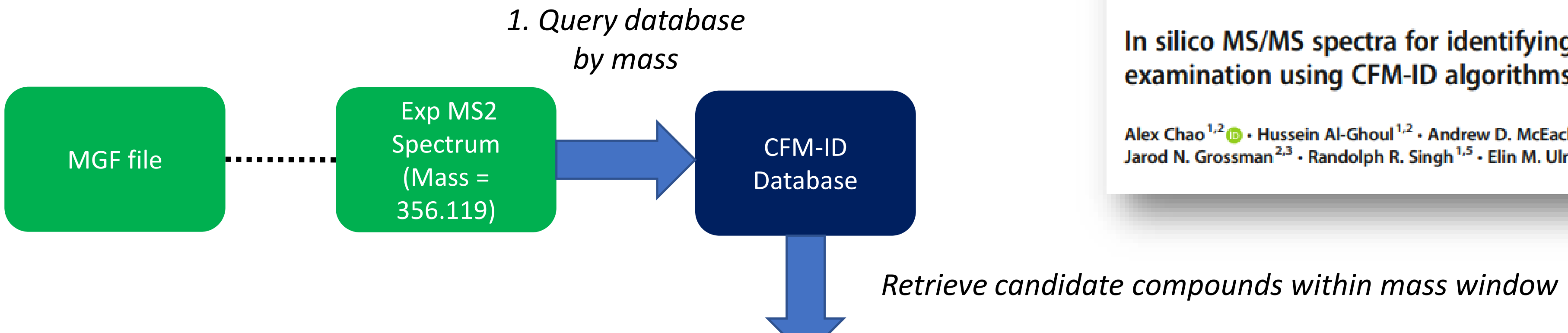
Training Set: Metlin MS2 spectra and structures

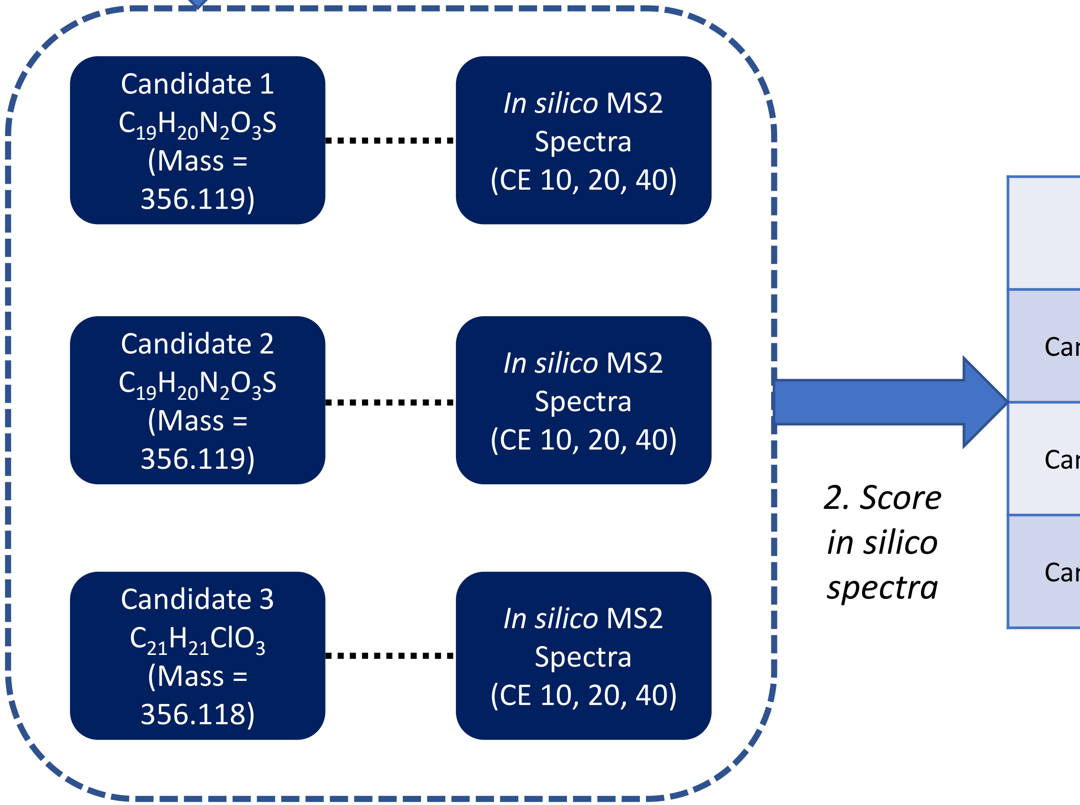
DSSTox MS-Ready Structures (~765,000)

DSSTox MS2 spectra (10, 20, 40v)

McEachran, Andrew D., et al. *Scientific data* 6.1 (2019): 1-9
Allen, Felicity, et al. *Metabolomics* 11.1 (2015): 98-110.

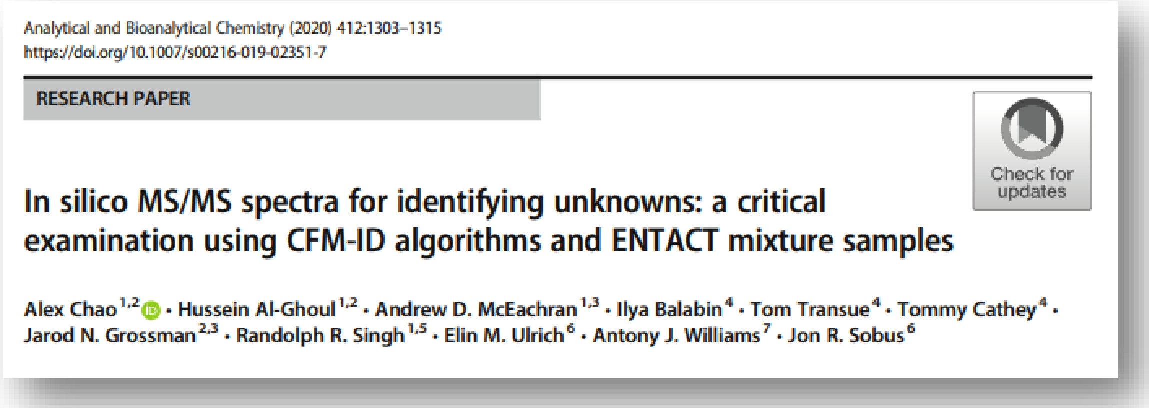
CFM-ID Database Matching


1. Query database by mass
MGF file
Exp MS2 Spectrum (Mass = 356.119)
CFM-ID Database
Retrieve candidate compounds within mass window


Candidate 1 $C_{19}H_{20}N_2O_3S$ (Mass = 356.119)
In silico MS2 Spectra (CE 10, 20, 40)
Candidate 2 $C_{19}H_{20}N_2O_3S$ (Mass = 356.119)
In silico MS2 Spectra (CE 10, 20, 40)
Candidate 3 $C_{21}H_{21}ClO_3$ (Mass = 356.118)
In silico MS2 Spectra (CE 10, 20, 40)
2. Score in silico spectra

CFM-ID Scores

	in silico CE 10	*in silico* CE 20	*in silico* CE 40
Candidate 1	0.5	0.3	0.1
Candidate 2	0.2	0.1	0.02
Candidate 3	0.1	0.05	0.01



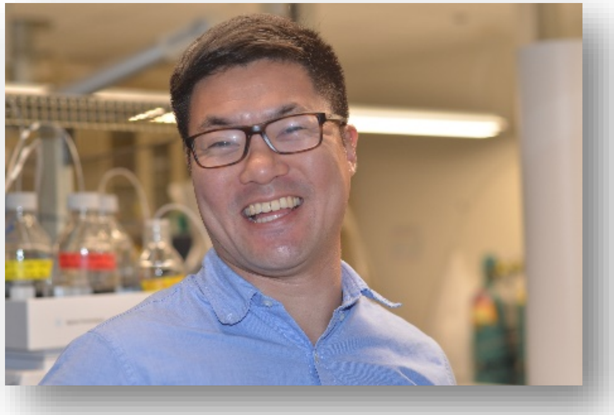
Analytical and Bioanalytical Chemistry (2020) 412:1303–1315
https://doi.org/10.1007/s00216-019-02351-7

RESEARCH PAPER

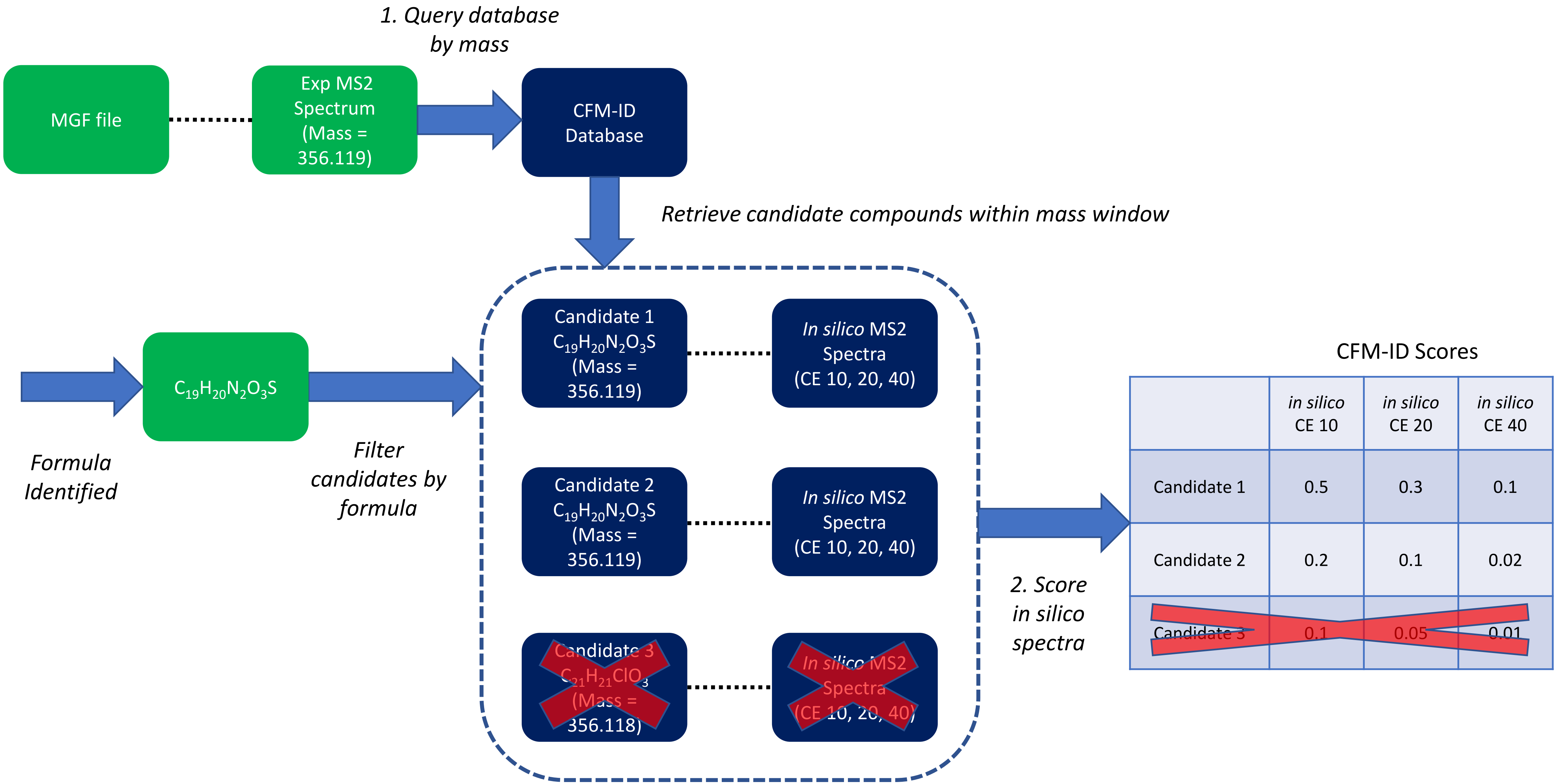
Check for updates

In silico MS/MS spectra for identifying unknowns: a critical examination using CFM-ID algorithms and ENTACT mixture samples

Alex Chao[1,2] · Hussein Al-Ghoul[1,2] · Andrew D. McEachran[1,3] · Ilya Balabin[4] · Tom Transue[4] · Tommy Cathey[4] · Jarod N. Grossman[2,3] · Randolph R. Singh[1,5] · Elin M. Ulrich[6] · Antony J. Williams[7] · Jon R. Sobus[6]



CFM-ID Database Matching (w/ Formula Information)



1. Query database by mass

MGF file

Exp MS2 Spectrum (Mass = 356.119)

CFM-ID Database

Retrieve candidate compounds within mass window

Formula Identified

$C_{19}H_{20}N_2O_3S$

Filter candidates by formula

Candidate 1 $C_{19}H_{20}N_2O_3S$ (Mass = 356.119)

In silico MS2 Spectra (CE 10, 20, 40)

Candidate 2 $C_{19}H_{20}N_2O_3S$ (Mass = 356.119)

In silico MS2 Spectra (CE 10, 20, 40)

~~Candidate 3 $C_{21}H_{21}ClO_3$ (Mass = 356.118)~~

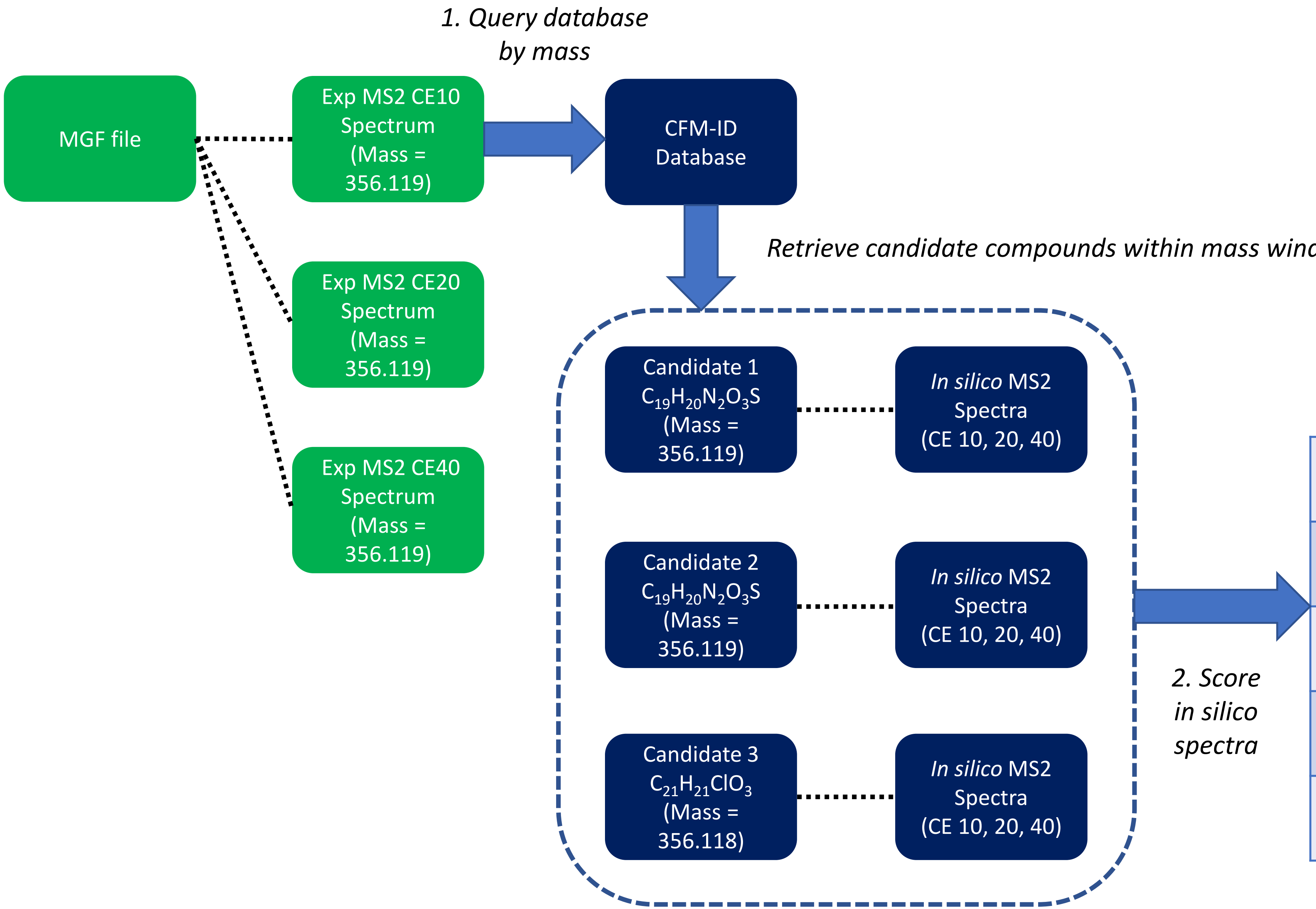
~~*In silico* MS2 Spectra (CE 10, 20, 40)~~

2. Score in silico spectra

CFM-ID Scores

	in silico CE 10	*in silico* CE 20	*in silico* CE 40
Candidate 1	0.5	0.3	0.1
Candidate 2	0.2	0.1	0.02
~~Candidate 3~~	~~0.1~~	~~0.05~~	~~0.01~~

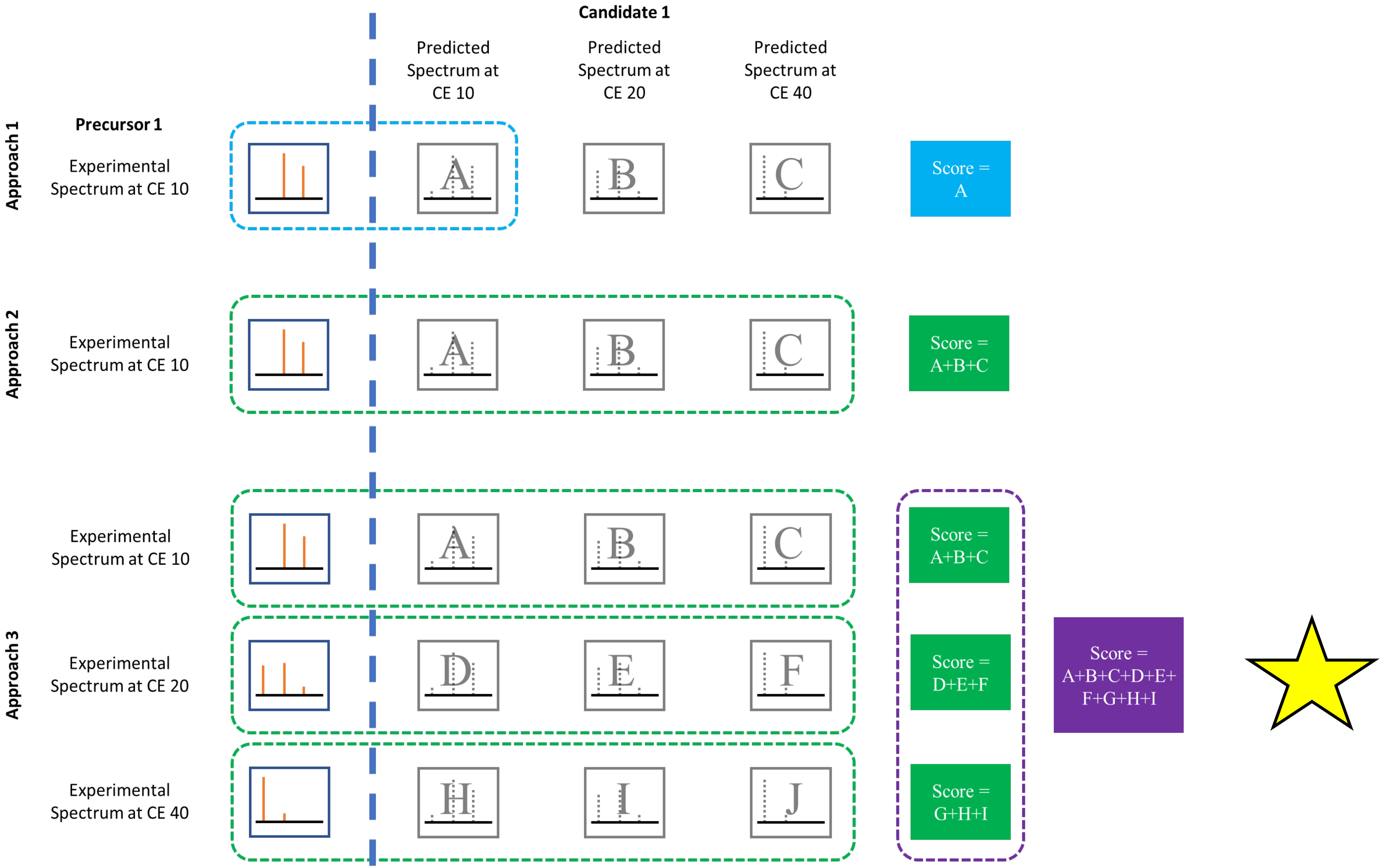
CFM-ID Database Matching (w/ Multiple CE$_{experimental}$)


1. Query database by mass
MGF file
Exp MS2 CE10 Spectrum (Mass = 356.119)
Exp MS2 CE20 Spectrum (Mass = 356.119)
Exp MS2 CE40 Spectrum (Mass = 356.119)
CFM-ID Database
Retrieve candidate compounds within mass window
Candidate 1 $C_{19}H_{20}N_2O_3S$ (Mass = 356.119)
Candidate 2 $C_{19}H_{20}N_2O_3S$ (Mass = 356.119)
Candidate 3 $C_{21}H_{21}ClO_3$ (Mass = 356.118)
In silico MS2 Spectra (CE 10, 20, 40)
In silico MS2 Spectra (CE 10, 20, 40)
In silico MS2 Spectra (CE 10, 20, 40)
2. Score in silico spectra

CFM-ID Scores

	in silico CE 10	*in silico* CE 20	*in silico* CE 40
Candidate 1 CE$_{exp}$ = 10	0.5	0.3	0.1
Candidate 1 CE$_{exp}$ = 20	0.4	0.5	0.12
Candidate 1 CE$_{exp}$ = 40	0.05	0.1	0.2
…	…	…	…

CFM-ID Scoring Approaches



Candidate 1

Predicted Spectrum at CE 10

Predicted Spectrum at CE 20

Predicted Spectrum at CE 40

Approach 1

Precursor 1

Experimental Spectrum at CE 10

Score = A

Approach 2

Experimental Spectrum at CE 10

Score = A+B+C

Approach 3

Experimental Spectrum at CE 10

Score = A+B+C

Experimental Spectrum at CE 20

Score = D+E+F

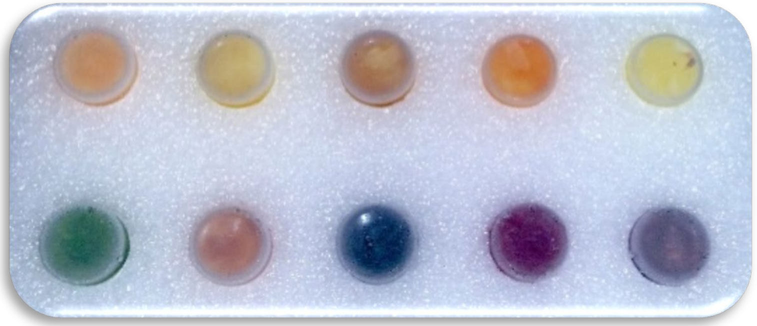
Experimental Spectrum at CE 40

Score = G+H+I

Score = A+B+C+D+E+F+G+H+I

EPA’S Non-Targeted Analysis Collaborative Trial

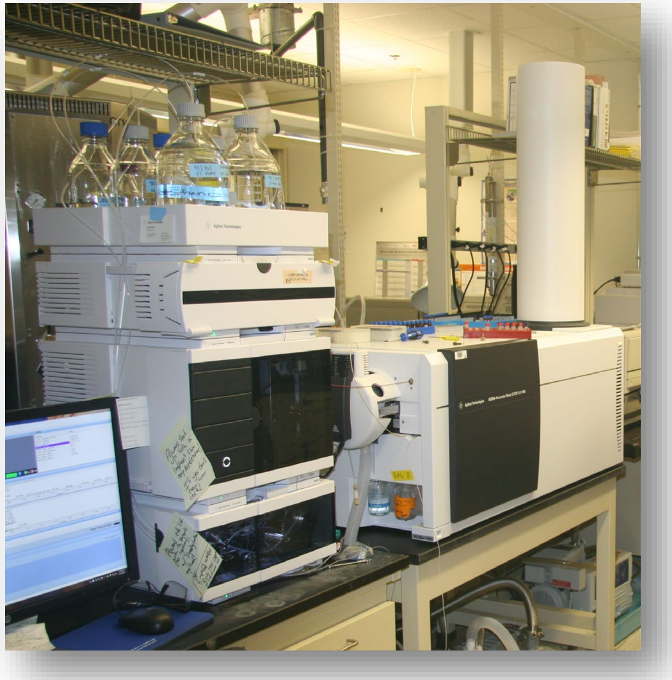
The Trial Mixtures:



10 Mixtures ranging from 95 to 365 compounds
(Total: 1,269 unique compounds)

“Pass” compounds = 377 with MS2 data

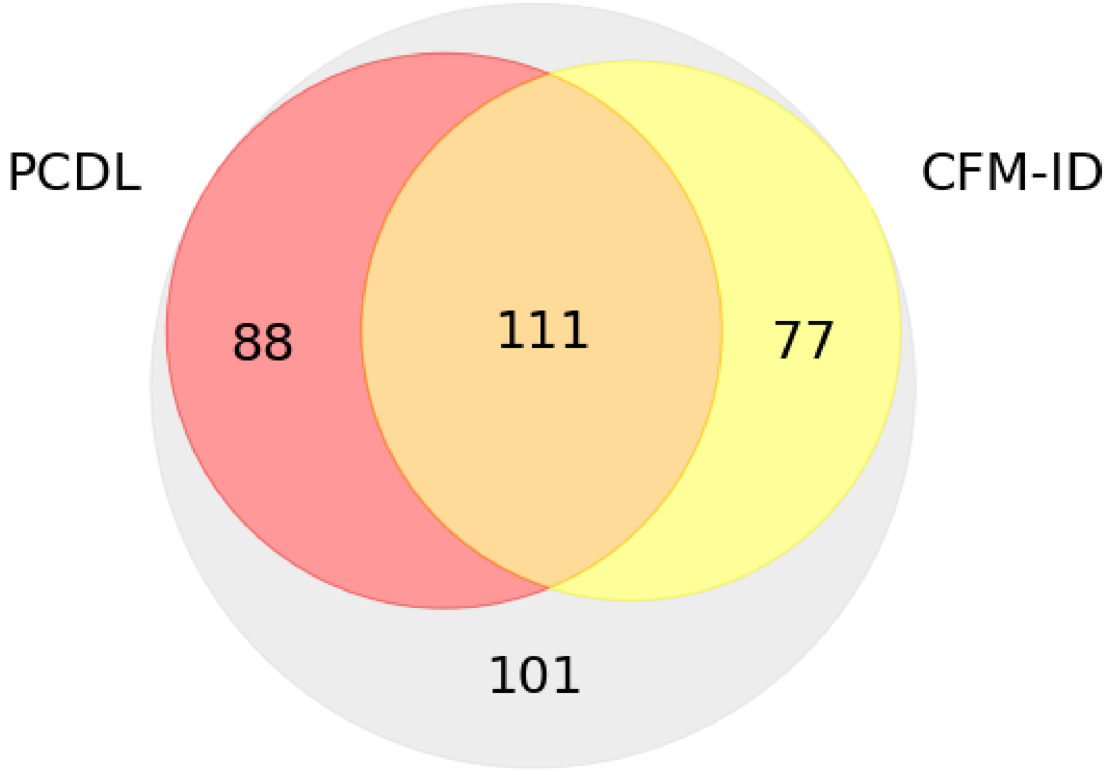
EPA Setup:



Agilent 1290 UPLC
Agilent 6530B Q-TOF with ESI source

Ulrich, Elin M., et al. *Analytical and bioanalytical chemistry* 411.4 (2019): 853-866.
Sobus, Jon R., et al. *Analytical and bioanalytical chemistry* 411.4 (2019): 835-851.

Reference vs. *in silico* Library Coverage



"Pass" Compounds

MS2 Library	% of "Pass" Compounds Identified
Agilent PCDL	53%
CFM-ID Top Hit	50%
PCDL and/or CFM-ID Top Hit	73%

PCDL → Agilent reference MS^2 library

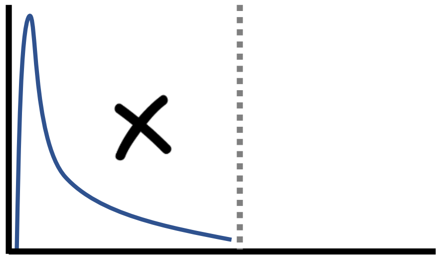
"Pass" compounds (n=377) → ENTACT chemicals observed with MS^2 data

NTA Workflows: Using CFM-ID Results as Filters

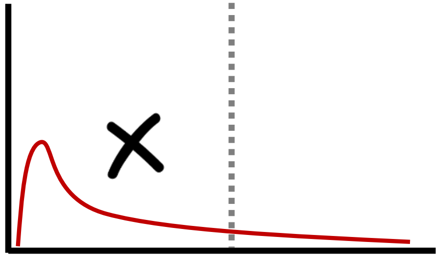
Score
Filter out candidates below score cutoff

Variability in score distribution

MS2 Spectrum 1



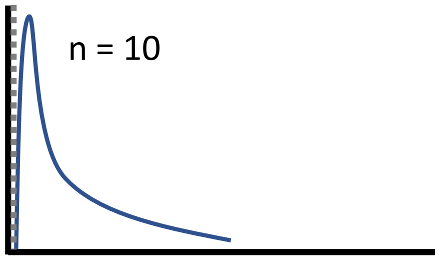
MS2 Spectrum 2



Rank
Filter out candidates above rank cutoff

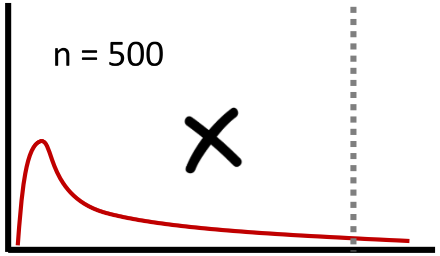
Variability in number of candidate compounds

MS2 Spectrum 1



Filter by Top 20

MS2 Spectrum 2



Normalizing CFM-ID Results Values

Score Quotient
Normalize score to the highest candidate compound score

Score Percentile
Normalize rank to the number of candidate compounds

	Rank	CFM-ID Score	Maximum Score	Score Quotient	Score Percentile
Candidate Compound 1	1	0.5	0.5	1	100
Candidate Compound 2	2	0.4	0.5	0.8	80
Candidate Compound 3	3	0.39	0.5	0.78	60
Candidate Compound 4	4	0.1	0.5	0.2	40
Candidate Compound 5	5	0.05	0.5	0.1	20

Score Quotient = Score / Maximum Score

NTA Workflows: Using CFM-ID Normalized Results as Filters

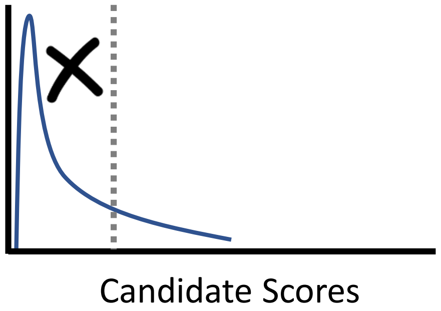
Score Quotient

Filter out candidates below score quotient cutoff

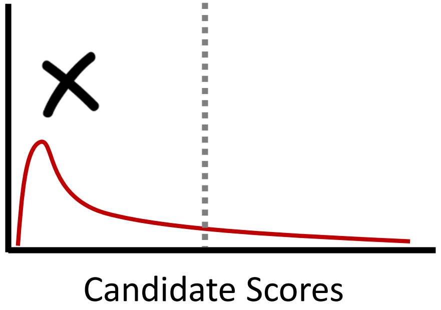
Score quotient cutoff = 0.5

Keep candidates scoring at least half of max score

MS2 Spectrum 1



MS2 Spectrum 2



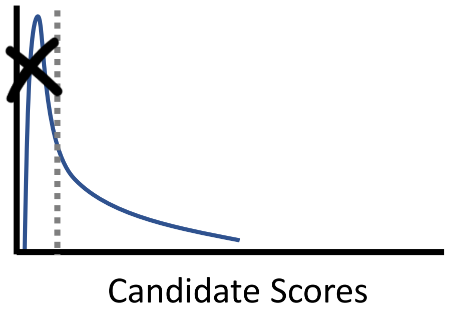
Score Percentile

Filter out candidates below percentile cutoff

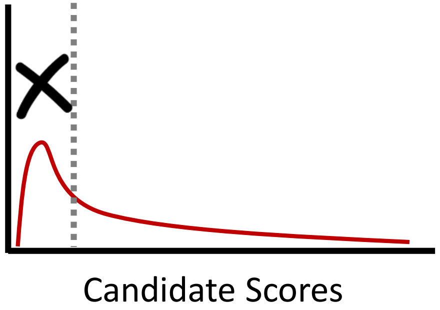
Score percentile cutoff = 0.5

Keep the top 50% of candidates

MS2 Spectrum 1

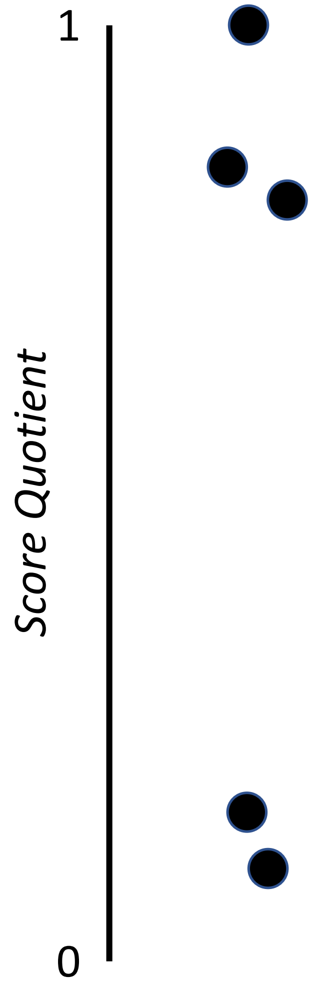


MS2 Spectrum 2



Applying Cut-off Filters to Data

	CFM-ID Score	Maximum Score	Score Quotient
Candidate Compound 1	0.5	0.5	1
Candidate Compound 2	0.4	0.5	0.8
Candidate Compound 3	0.39	0.5	0.78
Candidate Compound 4	0.1	0.5	0.2
Candidate Compound 5	0.05	0.5	0.1



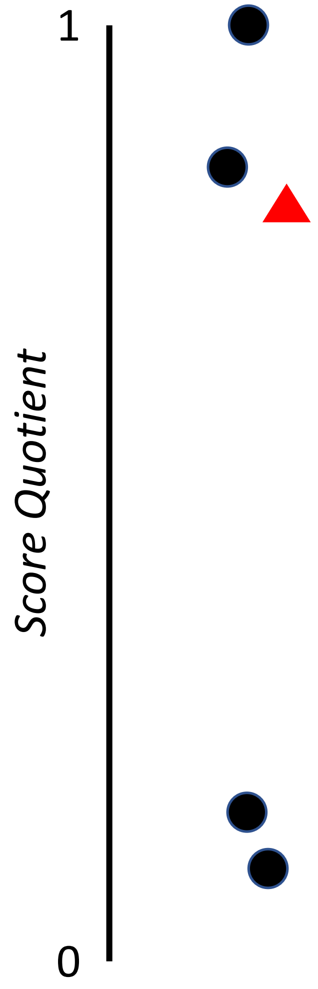
Applying Cut-off Filters to Data

		CFM-ID Score	Maximum Score	Score Quotient
●	Candidate Compound 1	0.5	0.5	1
●	Candidate Compound 2	0.4	0.5	0.8
▲	Candidate Compound 3	0.39	0.5	0.78
●	Candidate Compound 4	0.1	0.5	0.2
●	Candidate Compound 5	0.05	0.5	0.1

▲ True Compound

● Other Candidate Compounds

True Positives	
False Negatives	
True Negatives	
False Positives	



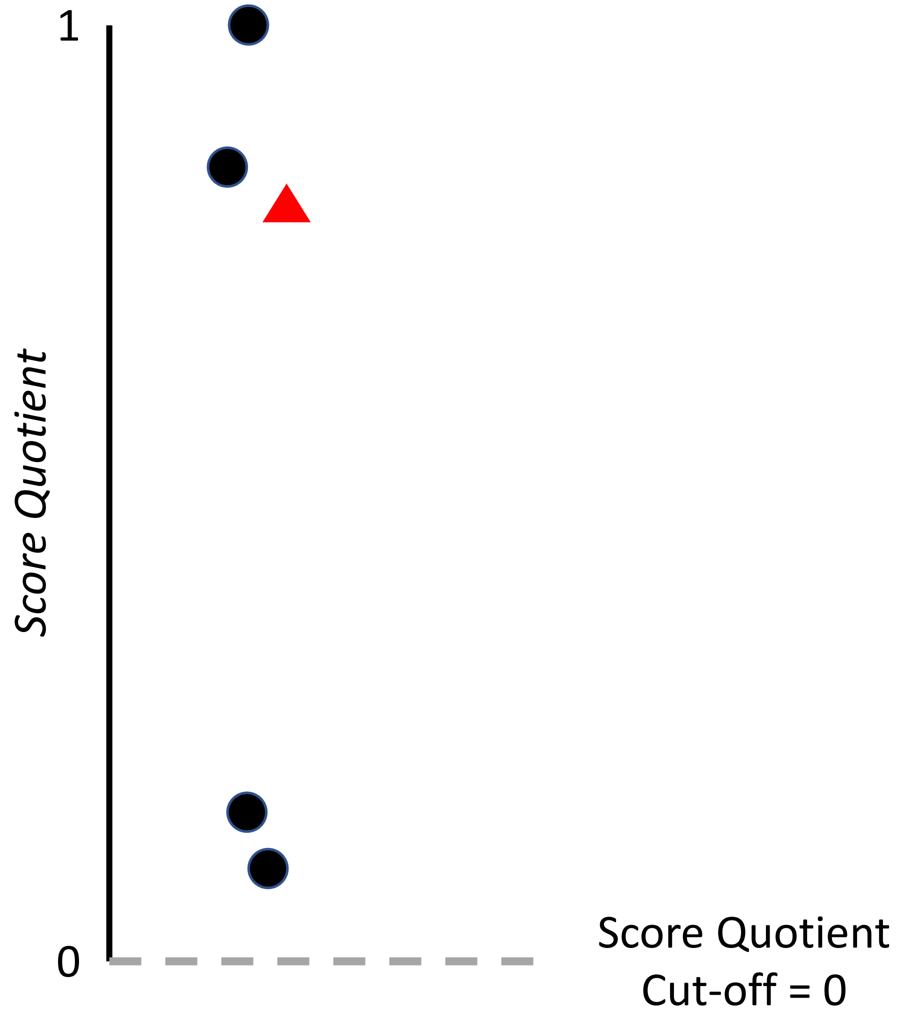
Applying Cut-off Filters to Data

		CFM-ID Score	Maximum Score	Score Quotient
●	Candidate Compound 1	0.5	0.5	1
●	Candidate Compound 2	0.4	0.5	0.8
▲	Candidate Compound 3	0.39	0.5	0.78
●	Candidate Compound 4	0.1	0.5	0.2
●	Candidate Compound 5	0.05	0.5	0.1

▲ True Compound

● Other Candidate Compounds

True Positives	1
False Negatives	0
True Negatives	0
False Positives	4



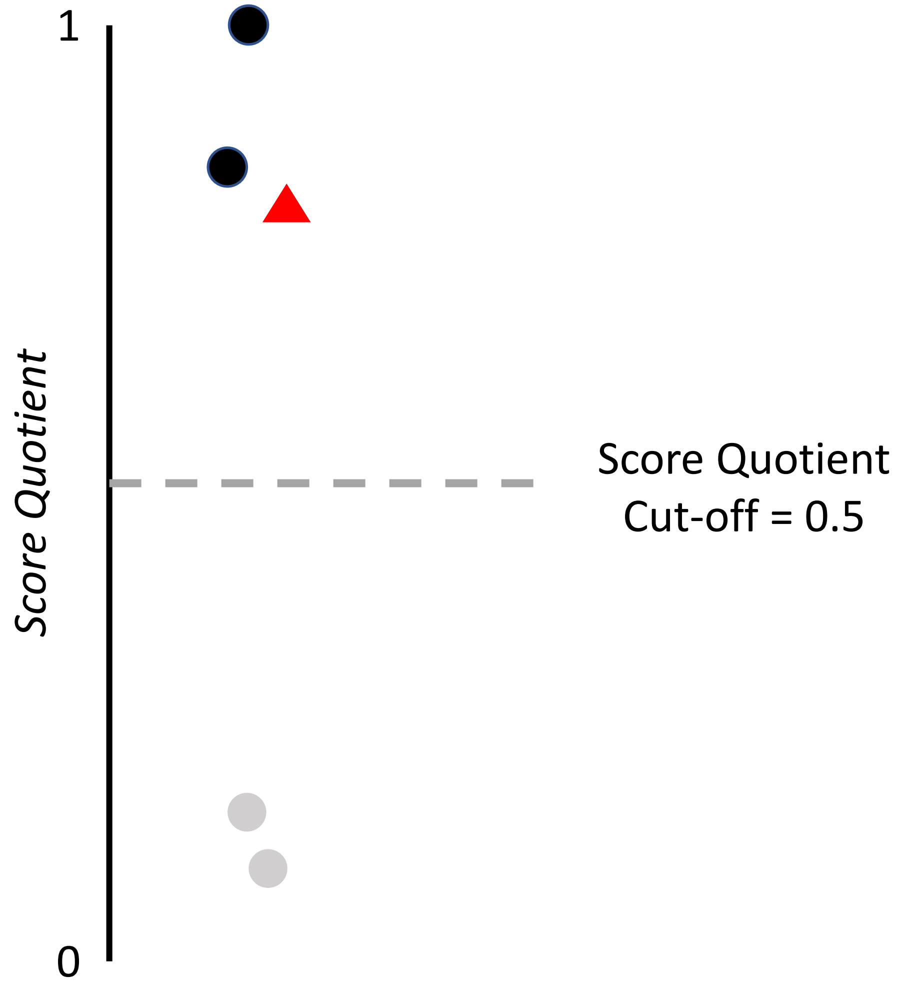
Applying Cut-off Filters to Data

		CFM-ID Score	Maximum Score	Score Quotient
●	Candidate Compound 1	0.5	0.5	1
●	Candidate Compound 2	0.4	0.5	0.8
▲	Candidate Compound 3	0.39	0.5	0.78
●	Candidate Compound 4	0.1	0.5	0.2
●	Candidate Compound 5	0.05	0.5	0.1

▲ True Compound

● Other Candidate Compounds

True Positives	1
False Negatives	0
True Negatives	2
False Positives	2



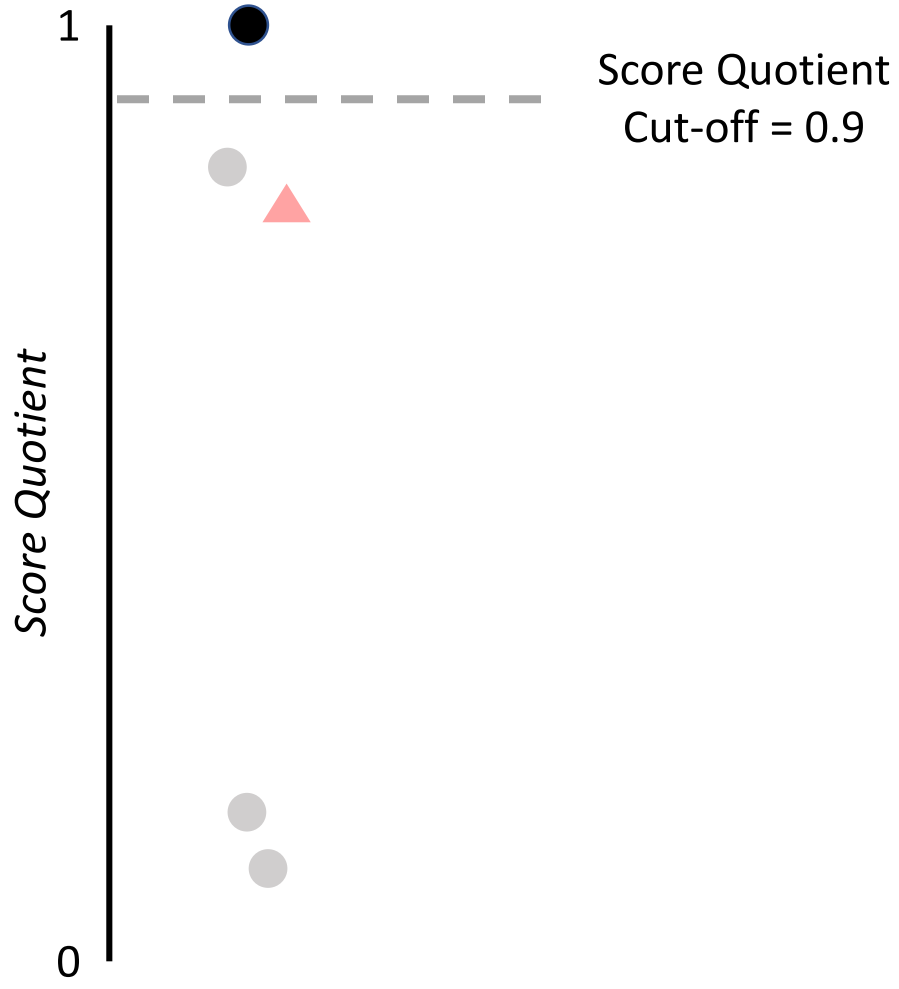
Applying Cut-off Filters to Data

		CFM-ID Score	Maximum Score	Score Quotient
●	Candidate Compound 1	0.5	0.5	1
●	Candidate Compound 2	0.4	0.5	0.8
▲	Candidate Compound 3	0.39	0.5	0.78
●	Candidate Compound 4	0.1	0.5	0.2
●	Candidate Compound 5	0.05	0.5	0.1

▲ True Compound

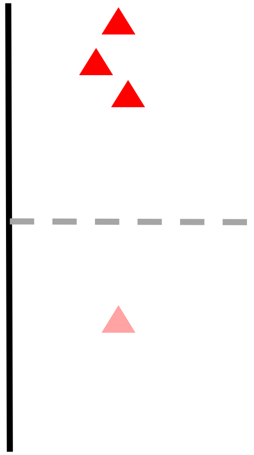
● Other Candidate Compounds

True Positives	0
False Negatives	1
True Negatives	3
False Positives	1



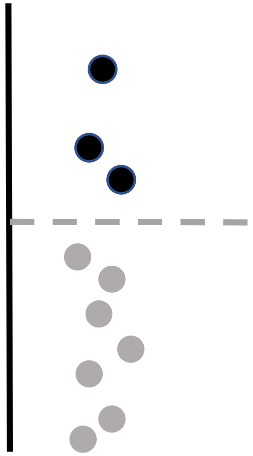
Balancing Cut-offs

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN}$$



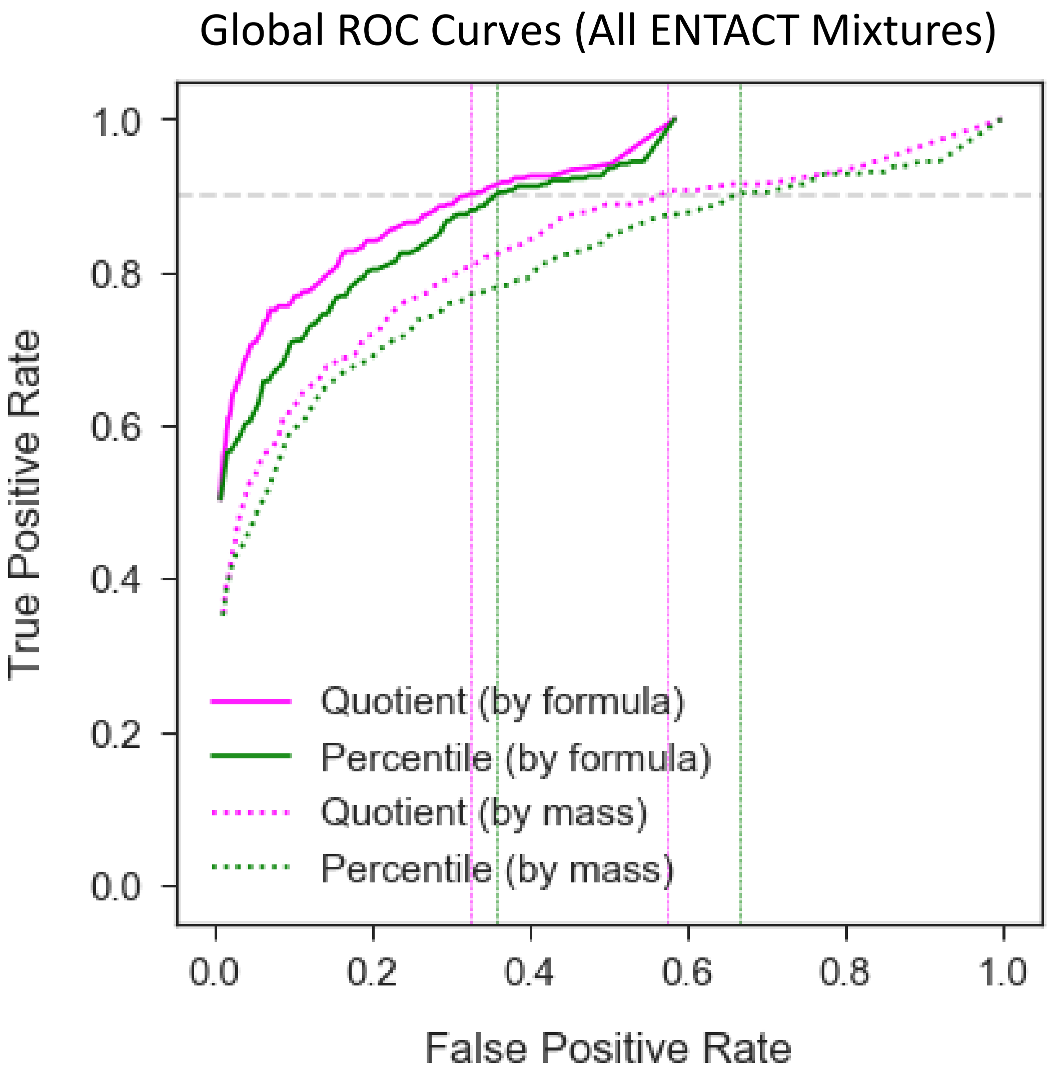
How many of the true compounds are we keeping?

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{FP + TN}$$



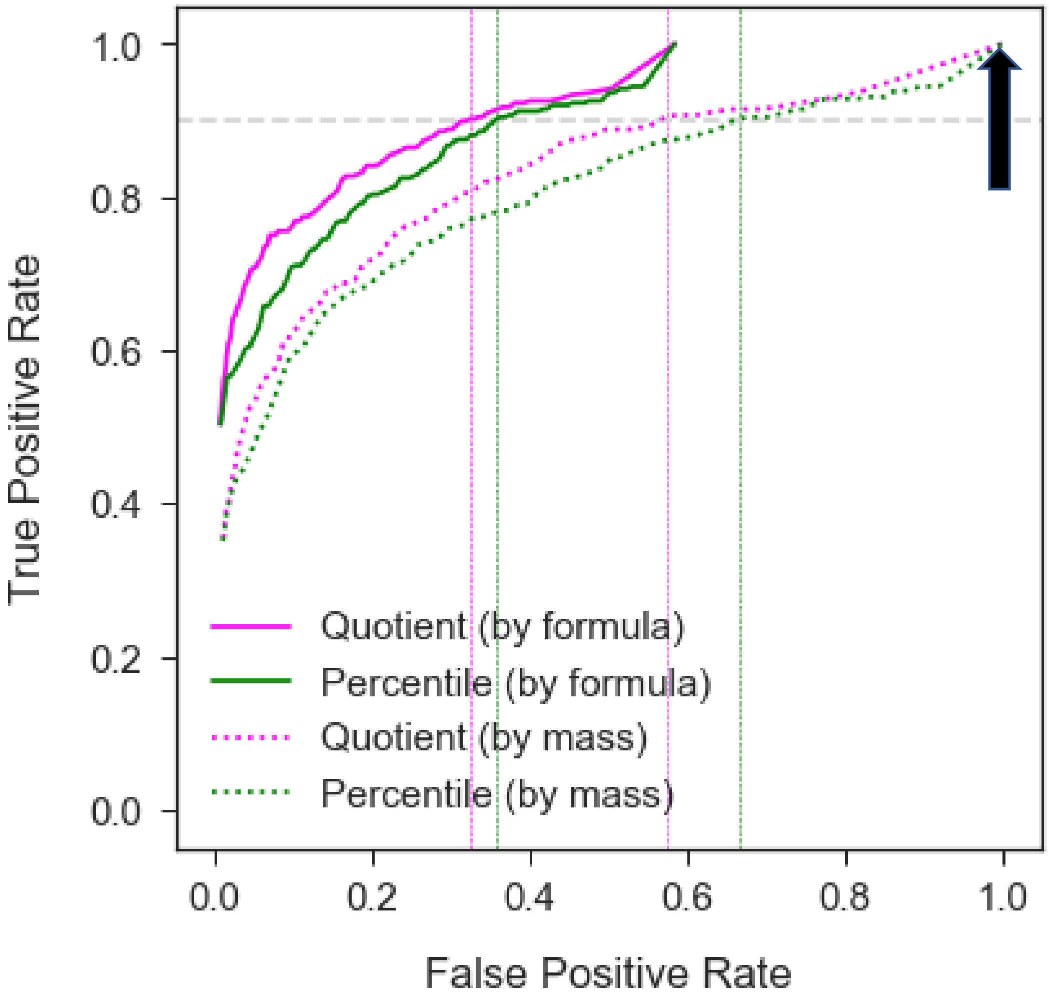
How much of the junk are we getting rid of?

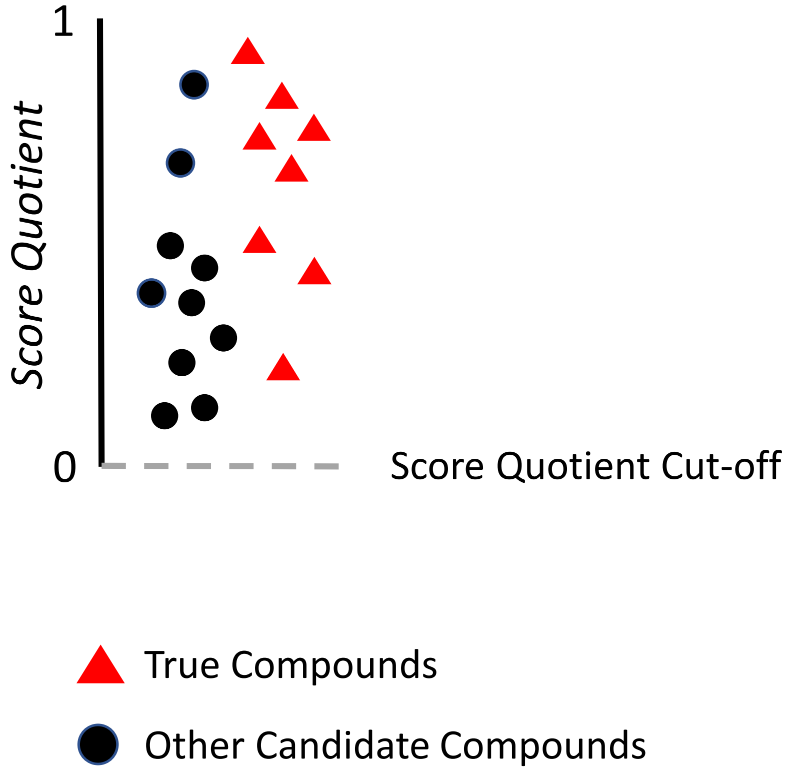
Quotient Vs. Percentile Cutoffs



Quotient Vs. Percentile Cutoffs

Global ROC Curves (All ENTACT Mixtures)



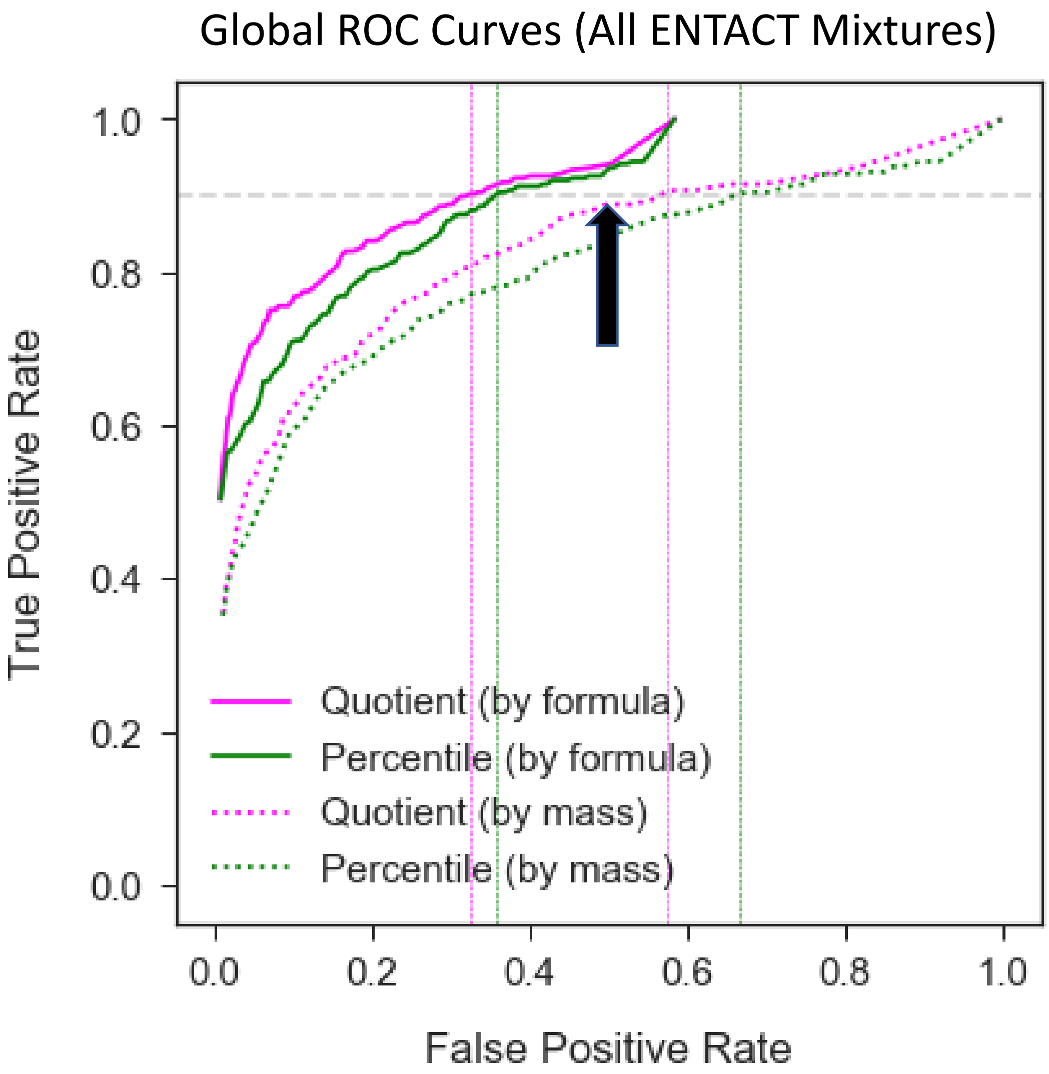


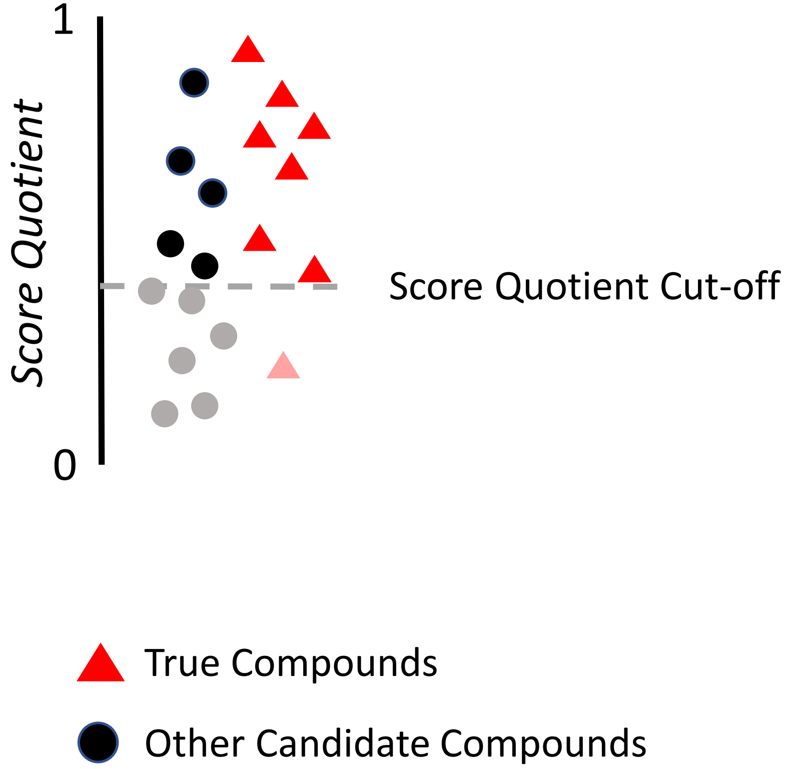
Score Quotient Cut-off

▲ True Compounds

● Other Candidate Compounds

Quotient Vs. Percentile Cutoffs



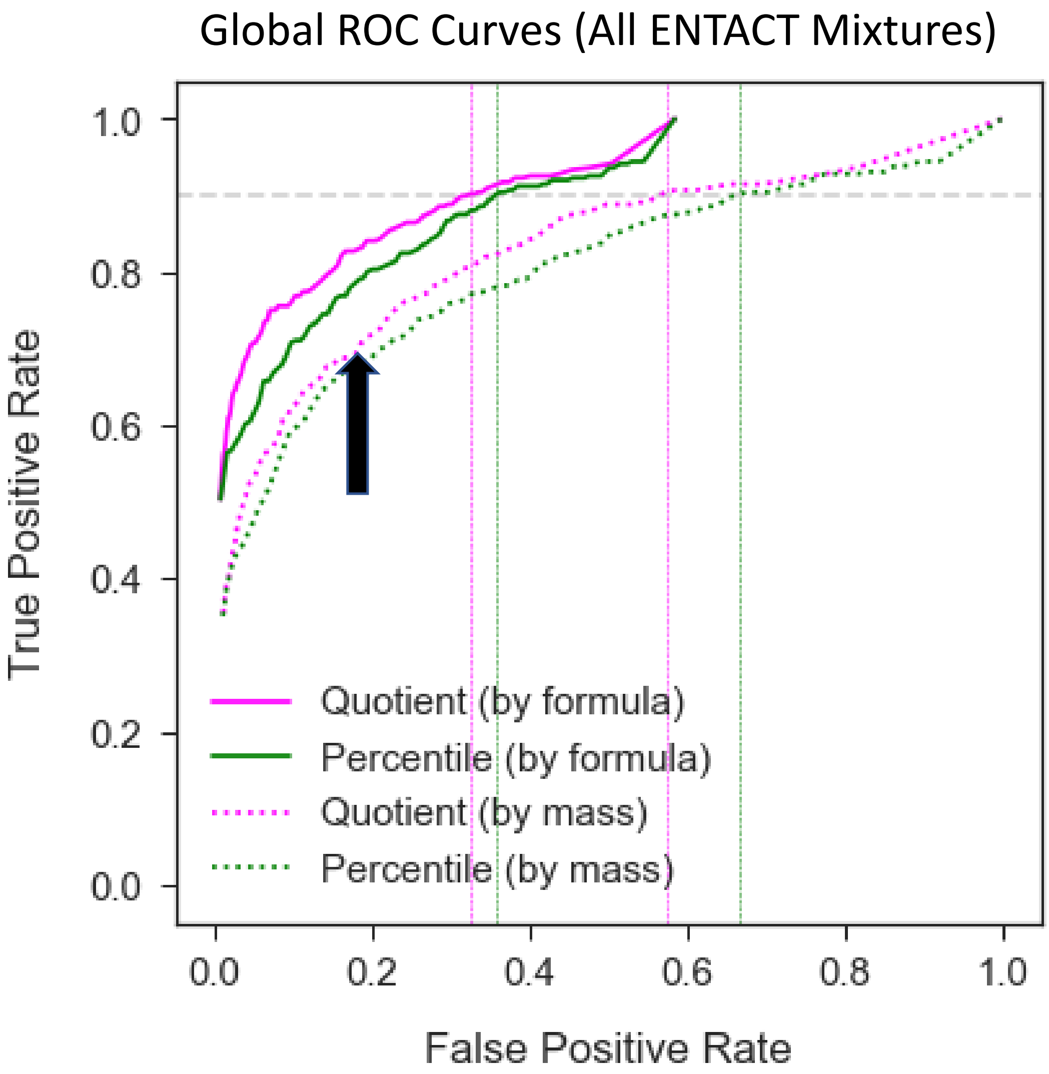


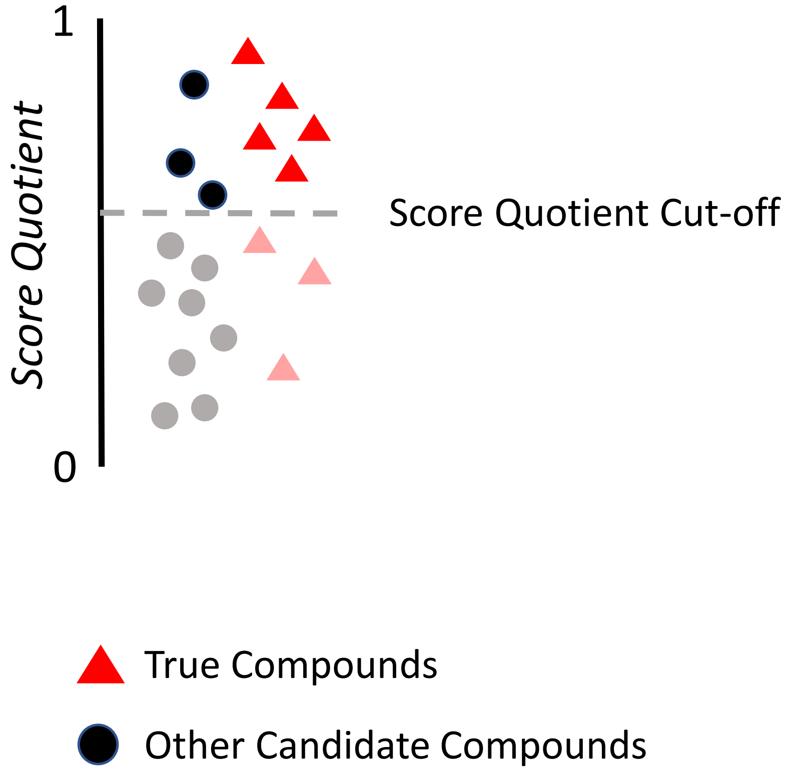
Score Quotient Cut-off

▲ True Compounds

● Other Candidate Compounds

Quotient Vs. Percentile Cutoffs



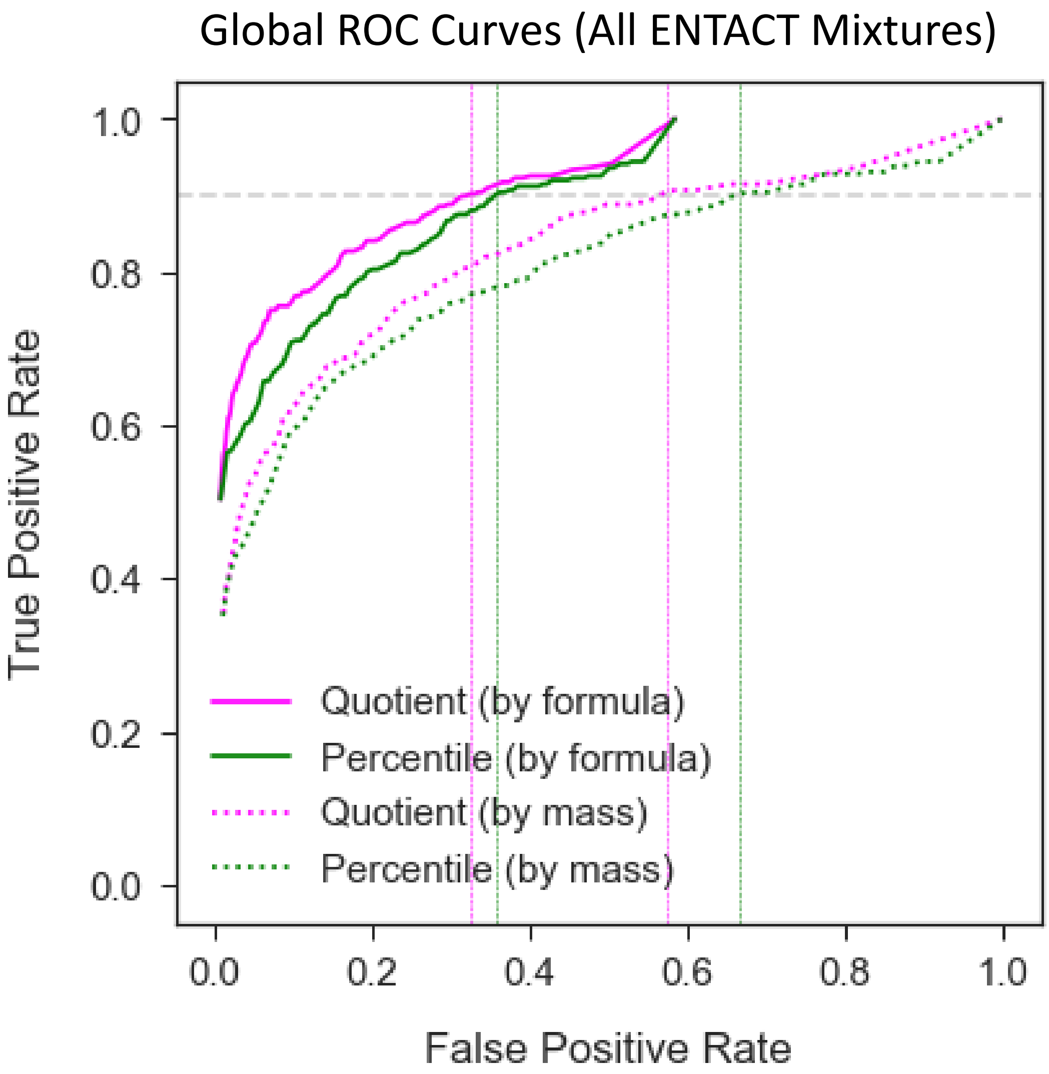


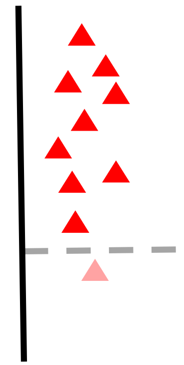
Score Quotient Cut-off

▲ True Compounds

● Other Candidate Compounds

Quotient Vs. Percentile Cutoffs



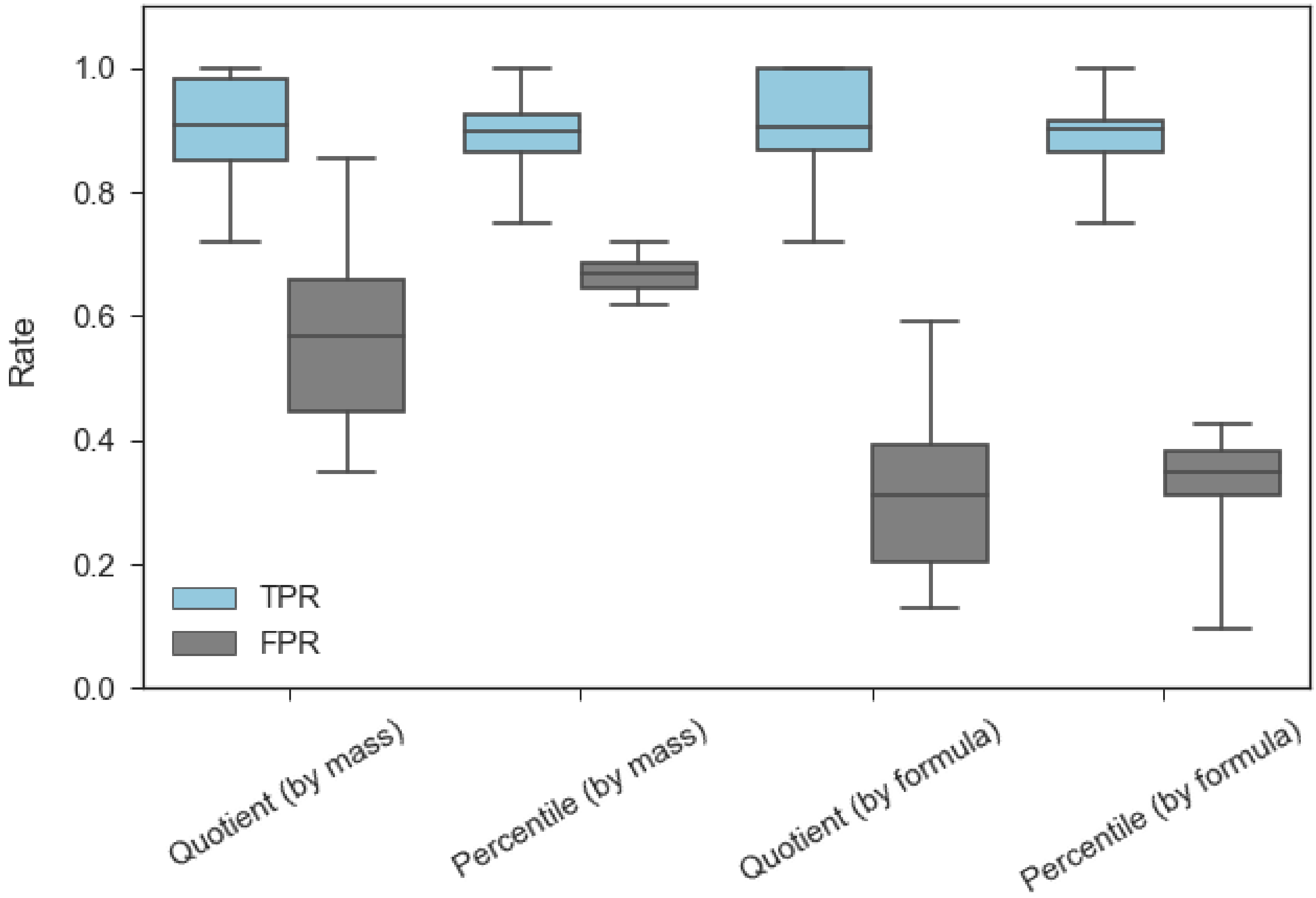


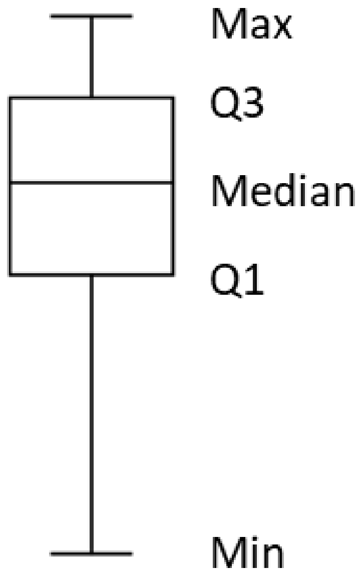
Cut-off Values for Global TPR = 0.9

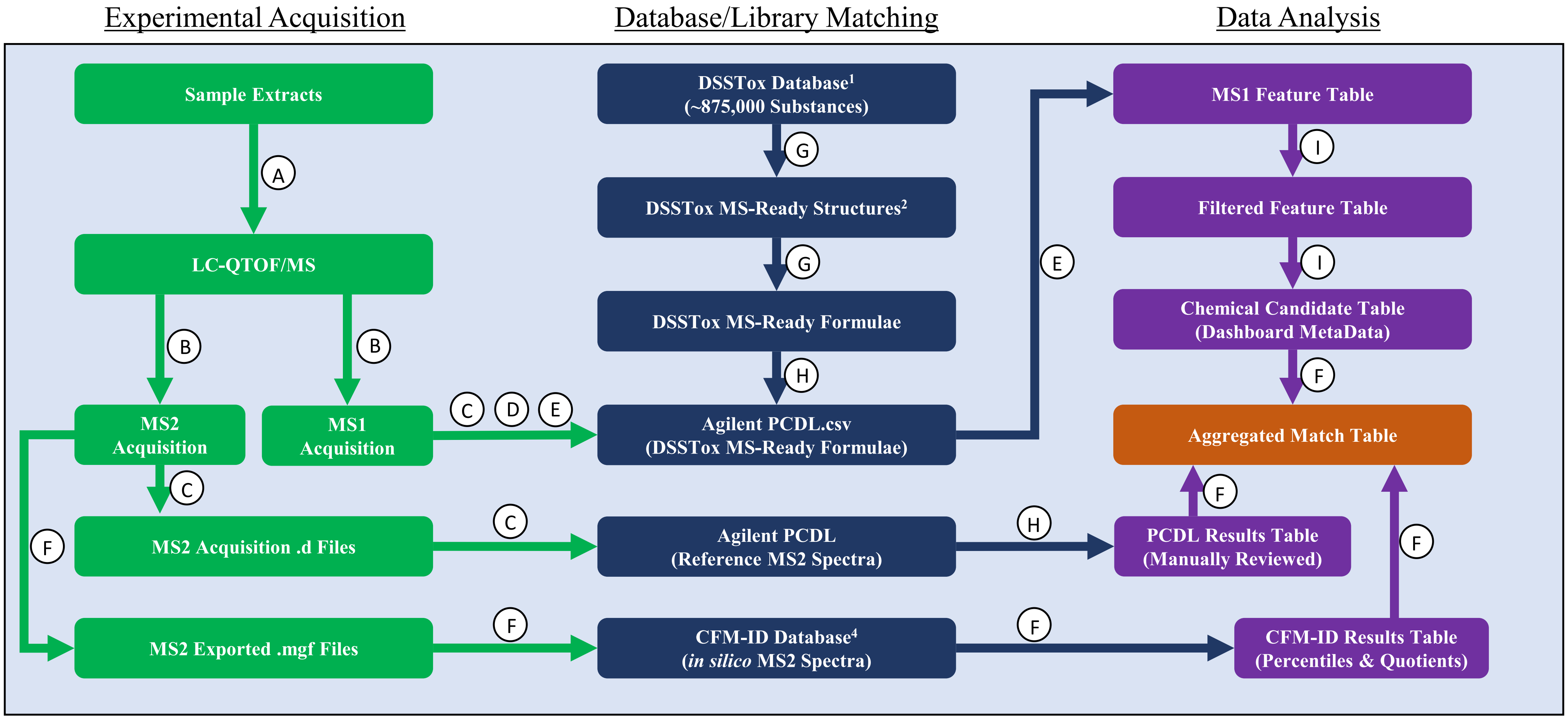
	Cut-off value
Quotient (by formula)	0.18
Percentile (by formula)	38
Quotient (by mass)	0.13
Percentile (by mass)	32

Apply to individual ENTACT mixtures

CFM-ID Cut-off Filtering: Individual ENTACT Mixtures







Experimental Acquisition

Sample Extracts

LC-QTOF/MS

MS2 Acquisition

MS1 Acquisition

MS2 Acquisition .d Files

MS2 Exported .mgf Files

Database/Library Matching

DSSTox Database[1] (~875,000 Substances)

DSSTox MS-Ready Structures[2]

DSSTox MS-Ready Formulae

Agilent PCDL.csv (DSSTox MS-Ready Formulae)

Agilent PCDL (Reference MS2 Spectra)

CFM-ID Database[4] (*in silico* MS2 Spectra)

Data Analysis

MS1 Feature Table

Filtered Feature Table

Chemical Candidate Table (Dashboard MetaData)

Aggregated Match Table

PCDL Results Table (Manually Reviewed)

CFM-ID Results Table (Percentiles & Quotients)

Software & Tools

A) Excel Macro (naming & randomization)
B) Agilent MassHunter Data Acquisition
C) Agilent MassHunter Qualitative Analysis
D) Agilent Profinder (peak picking & alignment)
E) Agilent Mass Profiler Professional (formula matching)
F) Python Script
G) CompTox Chemicals Dashboard[3]
H) Excel
I) EPA NTA WebApp

General Examples

Sobus et al. https://link.springer.com/article/10.1007%2Fs00216-018-1526-4

Newton et al. https://www.sciencedirect.com/science/article/pii/S026974911732691X?via%3Dihub

Hedgespeth et al. https://www.sciencedirect.com/science/article/pii/S004896971933298X?via%3Dihub

Specific References

[1]Grulke et al. https://www.sciencedirect.com/science/article/pii/S2468111319300234

[2]McEachran et al. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0299-2

[3]Williams et al. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-017-0247-6

[4]McEachran et al. https://www.nature.com/articles/s41597-019-0145-z

Contributing Researchers

This work was supported, in part, by ORD's Pathfinder Innovation Program (PIP) and an ORD EMVL award

Credit: the Research Triangle Foundation

EPA ORD

Hussein Al-Ghoul*
Alex Chao*
Jarod Grossman*
Kristin Isaacs
Sarah Laughlin*
Charles Lowe
James McCord
Jeff Minucci
Seth Newton
Katherine Phillips
Tom Purucker
Randolph Singh*
Mark Strynar
Elin Ulrich

* = ORISE/ORAU

EPA ORD (cont.)

Chris Grulke
Kamel Mansouri*
Andrew McEachran*
Ann Richard
John Wambaugh
Antony Williams

Agilent

Jarod Grossman
Andrew McEachran

GDIT

Ilya Balabin
Tom Transue
Tommy Cathey

Questions?

sobus.jon@epa.gov

The views expressed in this presentation are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

Credit: the Research Triangle Foundation