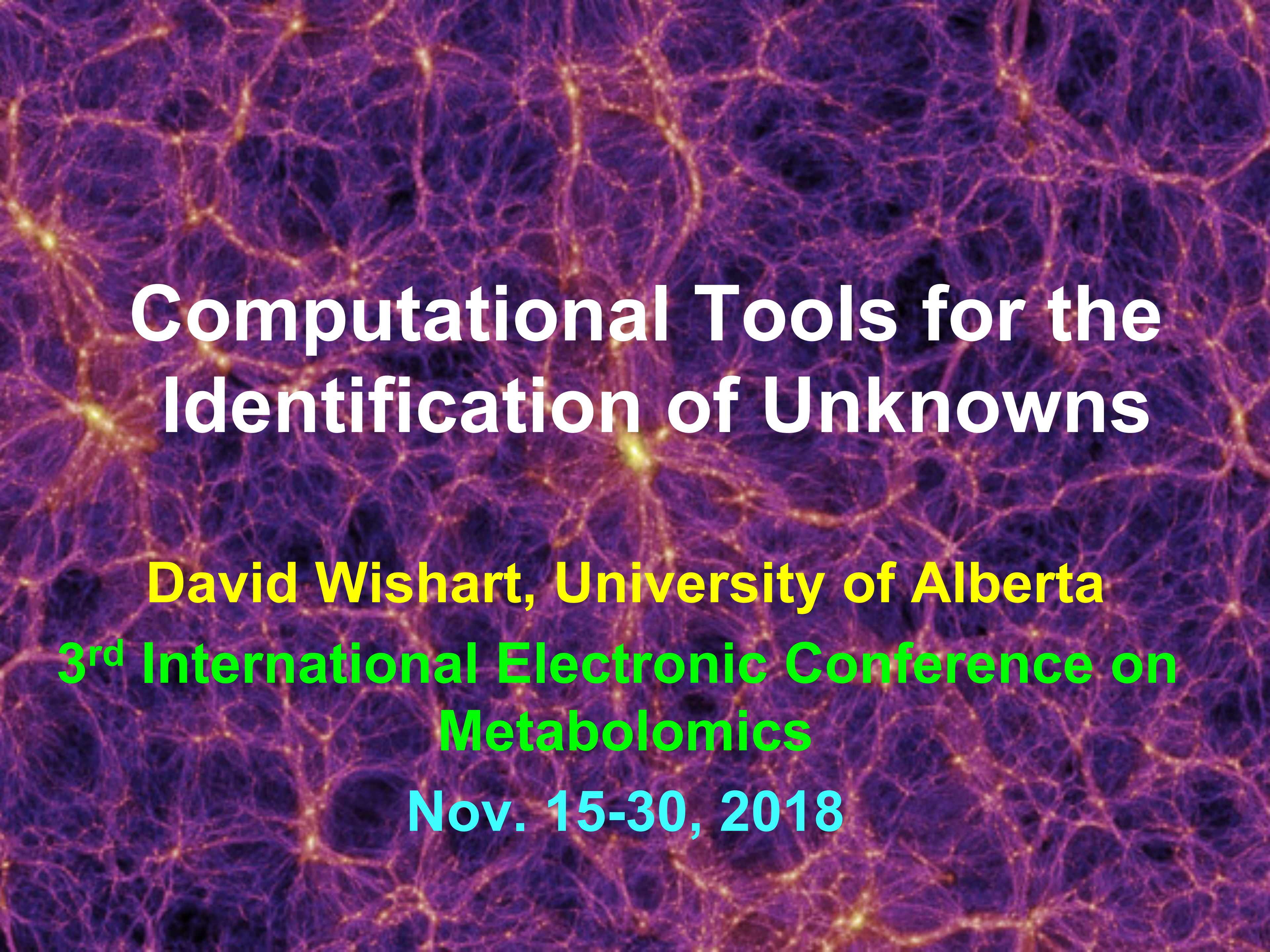


Computational Tools for the Identification of Unknowns

David Wishart, University of Alberta

3rd International Electronic Conference on Metabolomics

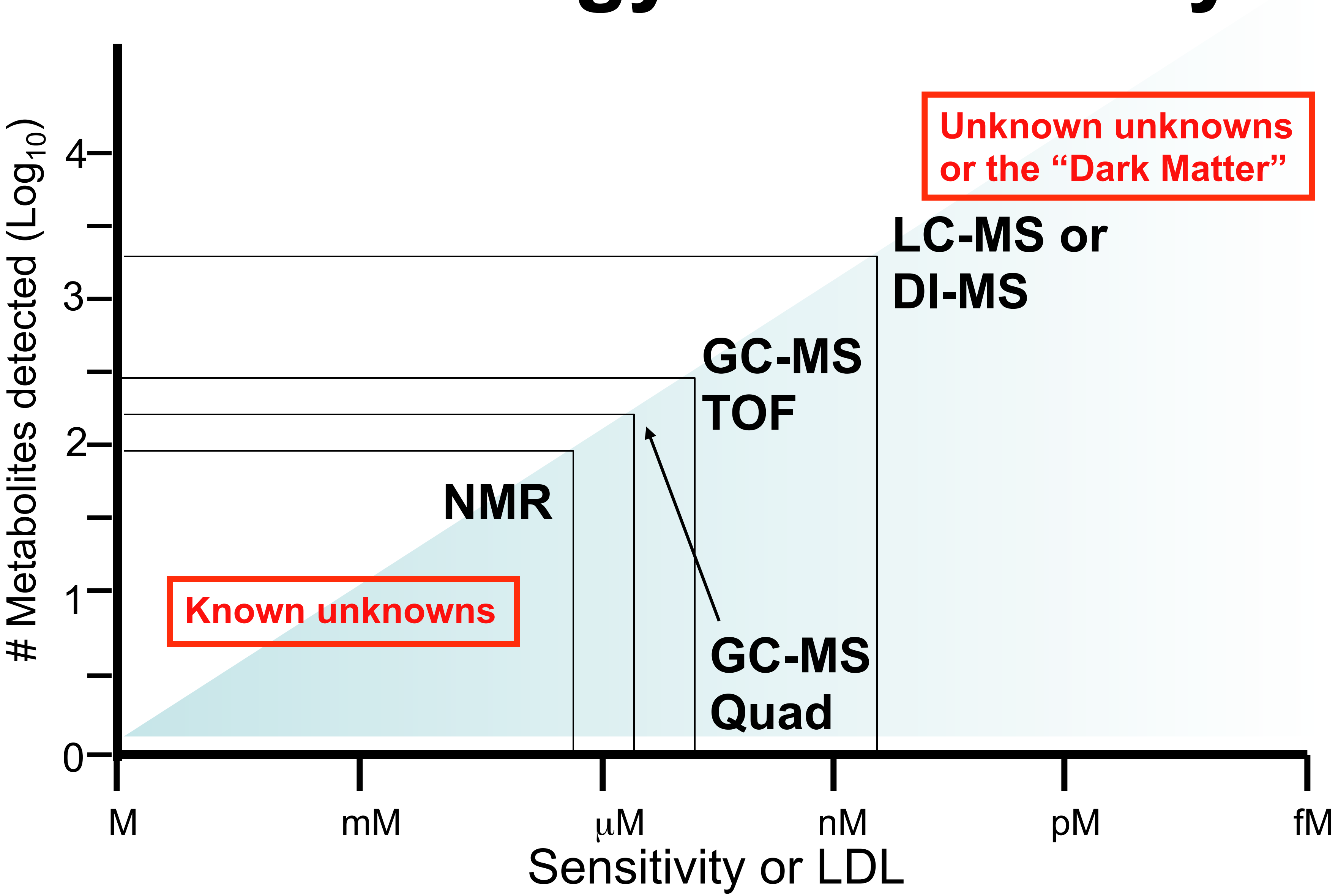
Nov. 15-30, 2018

What We Don’t Know

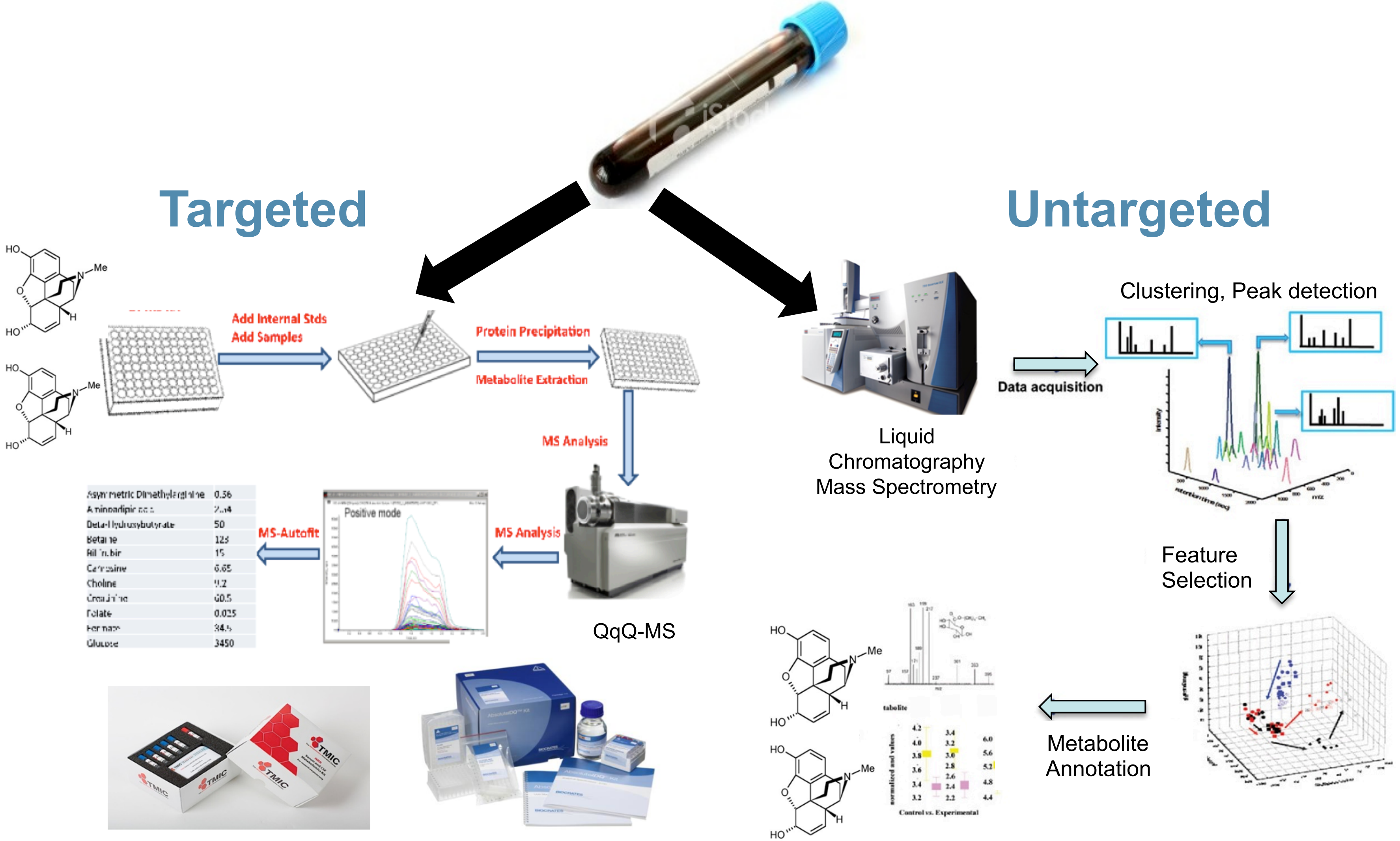
- “…*there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns – the ones we don't know we don't know.*”

Donald Rumsfeld, US Secretary of Defense - Nov. 2001

Technology & Sensitivity

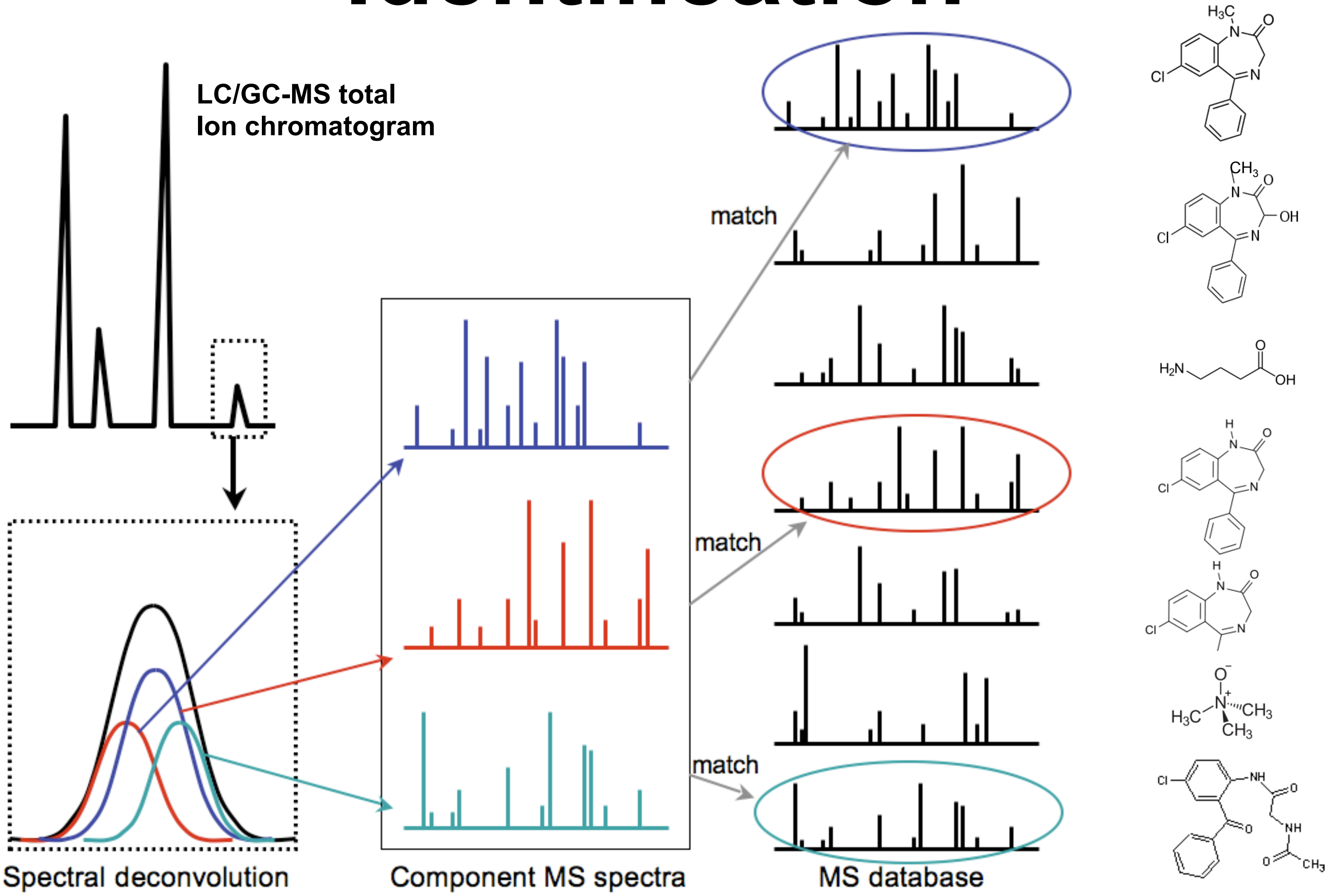


2 Routes to Metabolomics



Asymmetric Dimethylarginine	0.36
Aminoadipic acid	2.4
Beta-Hydroxybutyrate	50
Betaine	123
Bilirubin	15
Carnosine	6.65
Choline	9.2
Creatinine	60.5
Folate	0.025
Formate	34.5
Glucose	3450

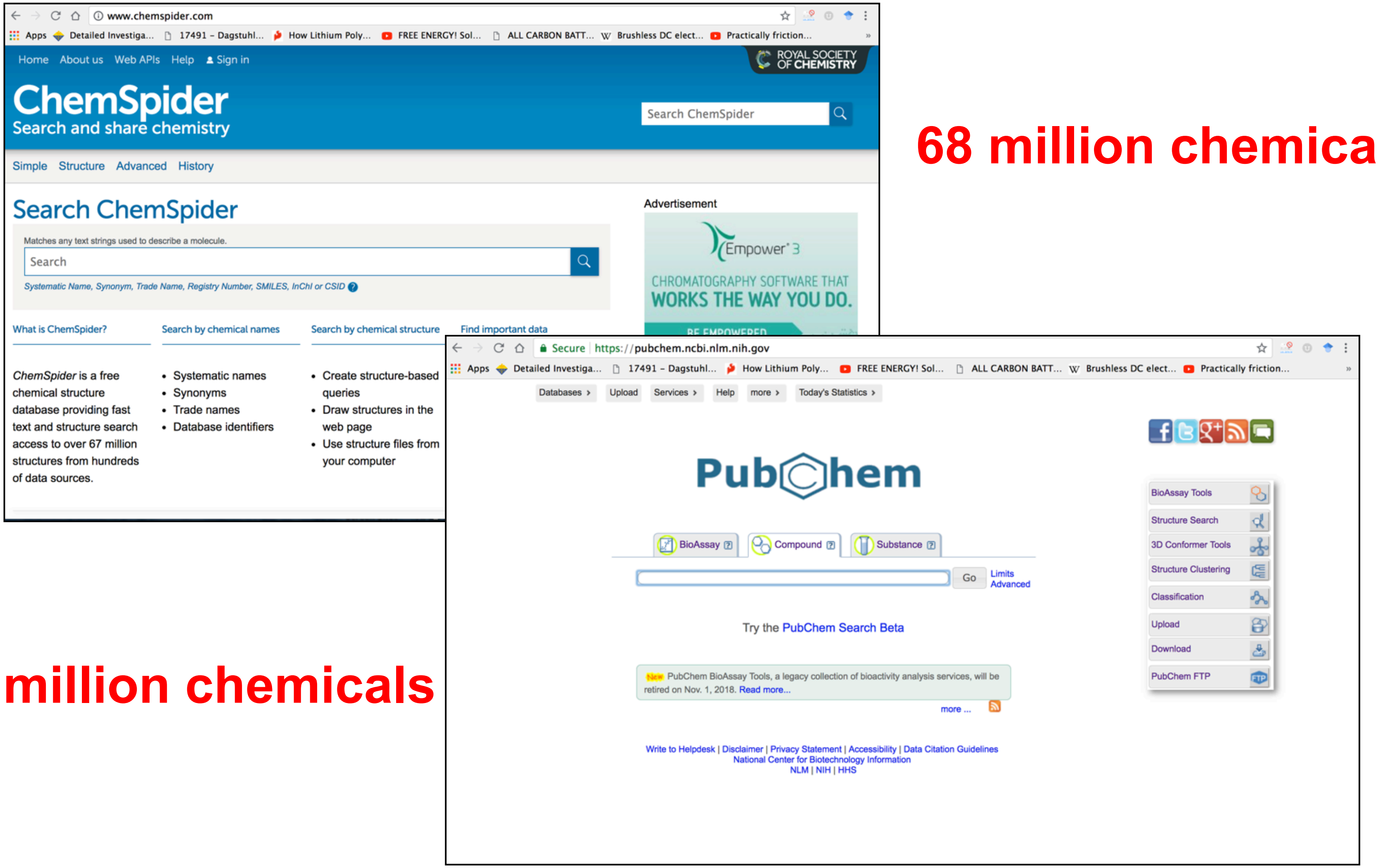
Untargeted MS Compound Identification



Levels of Metabolite ID for Untargeted Metabolomics

- **4 levels of metabolite identification**
- **Level 1 - Positively identified compounds**
  - Confirmed by MS/MS match and RT match to an actual/authentic standard
- **Level 2 - Putatively identified compounds**
  - Match to EI-MS + RT or MS/MS + RT from a reference database
- **Level 3 - Compounds putatively identified via molecular formula or m/z matching**
  - Match to high resolution m/z and nothing else
- **Level 4 - Unknown compounds**

Compound Identification (Formula Matching – Level 3 ID)



Search ChemSpider

Matches any text strings used to describe a molecule.

Search

Systematic Name, Synonym, Trade Name, Registry Number, SMILES, InChI or CSID

What is ChemSpider?

ChemSpider is a free chemical structure database providing fast text and structure search access to over 67 million structures from hundreds of data sources.

Search by chemical names

- Systematic names
- Synonyms
- Trade names
- Database identifiers

Search by chemical structure

- Create structure-based queries
- Draw structures in the web page
- Use structure files from your computer

Find important data

Try the PubChem Search Beta

New PubChem BioAssay Tools, a legacy collection of bioactivity analysis services, will be retired on Nov. 1, 2018. Read more...

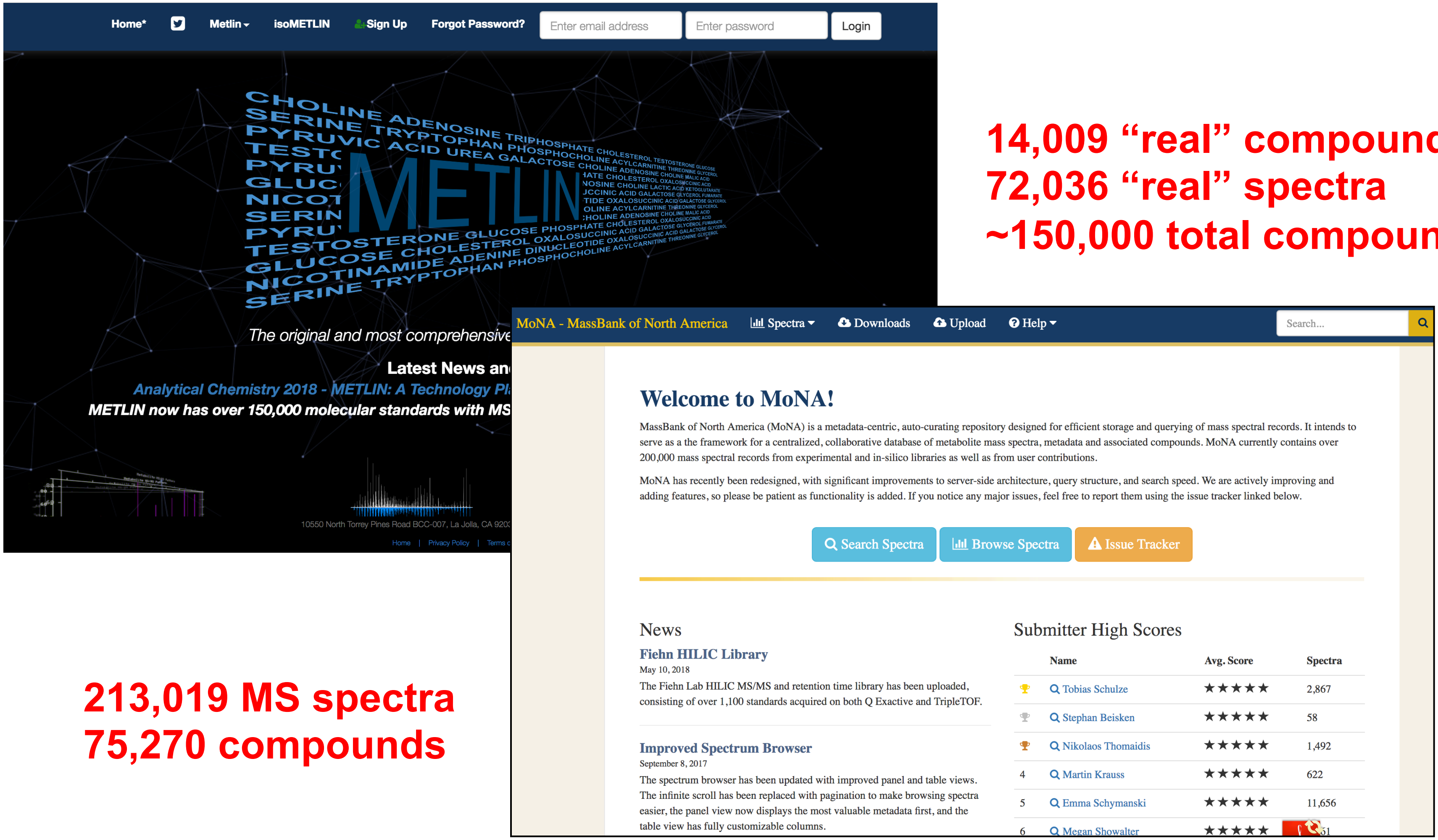
more ...

Write to Helpdesk | Disclaimer | Privacy Statement | Accessibility | Data Citation Guidelines
National Center for Biotechnology Information
NLM | NIH | HHS

68 million chemicals

96 million chemicals

Compound Identification (Spectral Matching – Level 2 ID)



The original and most comprehensive

Latest News an

Analytical Chemistry 2018 - METLIN: A Technology Pl

METLIN now has over 150,000 molecular standards with MS

10550 North Torrey Pines Road BCC-007, La Jolla, CA 9203

Home | Privacy Policy | Terms c

Welcome to MoNA!

MassBank of North America (MoNA) is a metadata-centric, auto-curating repository designed for efficient storage and querying of mass spectral records. It intends to serve as a the framework for a centralized, collaborative database of metabolite mass spectra, metadata and associated compounds. MoNA currently contains over 200,000 mass spectral records from experimental and in-silico libraries as well as from user contributions.

MoNA has recently been redesigned, with significant improvements to server-side architecture, query structure, and search speed. We are actively improving and adding features, so please be patient as functionality is added. If you notice any major issues, feel free to report them using the issue tracker linked below.

News

Fiehn HILIC Library

May 10, 2018

The Fiehn Lab HILIC MS/MS and retention time library has been uploaded, consisting of over 1,100 standards acquired on both Q Exactive and TripleTOF.

Improved Spectrum Browser

September 8, 2017

The spectrum browser has been updated with improved panel and table views. The infinite scroll has been replaced with pagination to make browsing spectra easier, the panel view now displays the most valuable metadata first, and the table view has fully customizable columns.

Submitter High Scores

	Name	Avg. Score	Spectra
	Tobias Schulze	★★★★★	2,867
	Stephan Beisken	★★★★★	58
	Nikolaos Thomaidis	★★★★★	1,492
4	Martin Krauss	★★★★★	622
5	Emma Schymanski	★★★★★	11,656
6	Megan Showalter	★★★★★	[illegible]

14,009 "real" compounds
72,036 "real" spectra
~150,000 total compounds

213,019 MS spectra
75,270 compounds

Other Resources for MS/MS Spectral Matching

- **HMDB** – 302,219 spectra, 114,100 cmpds
- **mzCloud** – 191,722 spectra, 8304 cmpds
- **NIST17 MS/MS** – 652,475 spectra, 14,351 cmpds
- **MassBank** – 28,185 spectra, 11,500 cmpds
- **Wiley LC-MS^n** – >10,000 spectra, 4500 poisons
- **ReSpect** – 9017 spectra, 3595 cmpds
- **GNPS** – 221,000 spectra, 18,163 cmpds

How Well Do We Do?

- **Untargeted LC-MS of human biofluids – 100-250 compounds positively ID'd (level 2) out of 10,000+ features (1%), 700-1000 tentatively ID'd (level 3) out of 10,000+ features (8%)**
- **Untargeted LC-MS of river water – 649 compounds identified (level 1 or 2) out of 8535 features (8%) *(Schymanski et al. Anal Bioanal Chem (2015) 407:6237–6255)***

Overall we are doing pretty badly

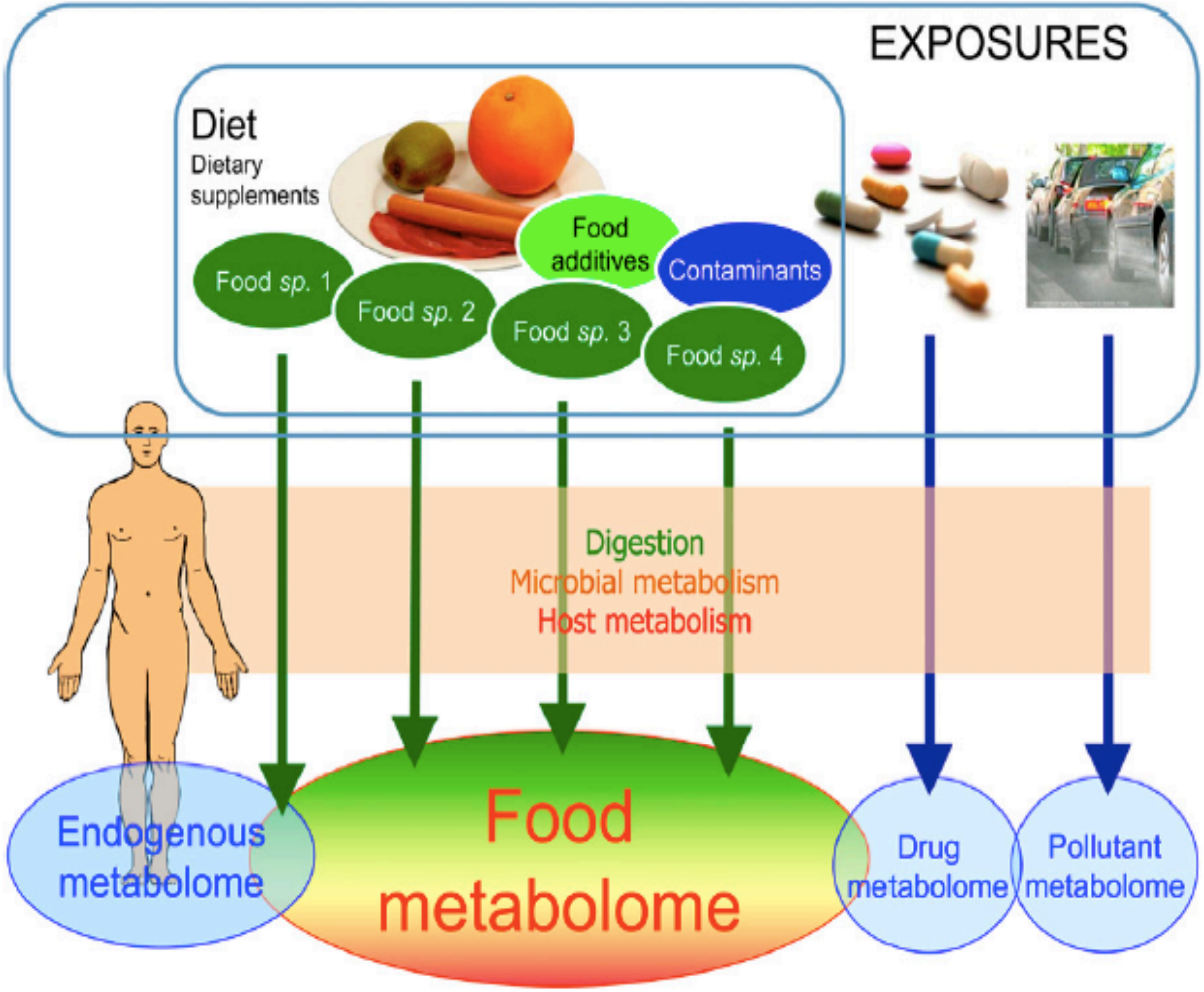
Why Are We Doing So Badly?

- 32,000,000,000 chemical formulae (<2000 Da)
- 2,400,000,000 chemically feasible formulae
- 96,500,000 chemicals in PubChem
- 1,500,000 LC/GC-MS spectra collected on ~15 different platforms
- 80,000 chemicals with EI-MS spectra
- ~20,000 chemicals with high resolution MS spectra
- ~1500 chemicals that are biologically relevant

Why Are We Doing So Badly?

- Using larger databases (PubChem, ChemSpider) and m/z matching is leading to many, many false positives
- <0.2% of compounds in PubChem or ChemSpider have ever left the laboratory or are likely to be found in humans or in the environment (Bigger isn’t better)
- Only 1500 “meaningful” chemicals are routinely available from vendors (Synthetic chemistry is hard)
- Enormous resources going to collect lots of MS spectra on a tiny number of chemicals (Measuring the same thing over doesn’t make the problem go away)
- Most of the unknowns are not in PubChem or Chemspider, or anywhere else (Where are they?)

What Are These Unknowns?



→ Mostly Known

→ Mostly Unknown

What To Do?

- Obtain or synthesize all commonly available xenobiotics (HPVs, drugs, pollutants, foods, etc.), prepare or synthesize their metabolites and collect their NMR, LC-MS and GC-MS spectra
  COST: 5 million cmpds X $1000/cmpd = $5 billion
- OR
- Do this entire exercise computationally

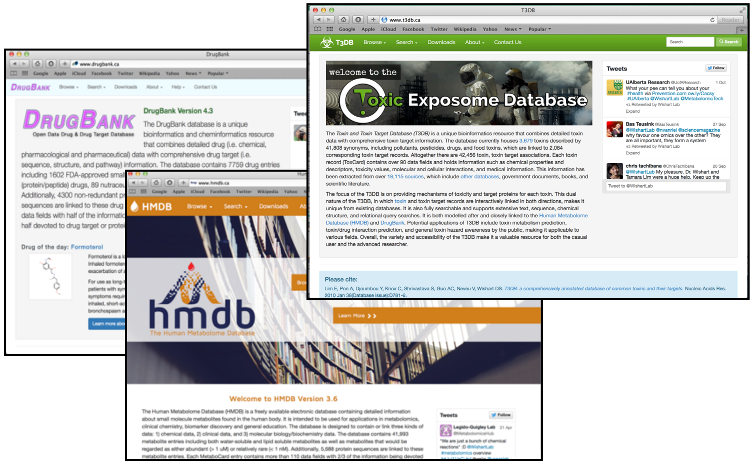
COST - 5,000,000 cmpds X $0.10/cmpd = $500,000

In Silico Metabolomics

- An emerging concept to facilitate metabolite ID of unknown unknowns
- Realization that all metabolites will never be synthesized or isolated and most will never have reference MS/MS or NMR spectra collected
- Based on *in silico* prediction of biologically feasible metabolites
- Based on *in silico* prediction of observables (RI, RT, NMR spectra, IR, CCS, MW, MS/MS spectra)

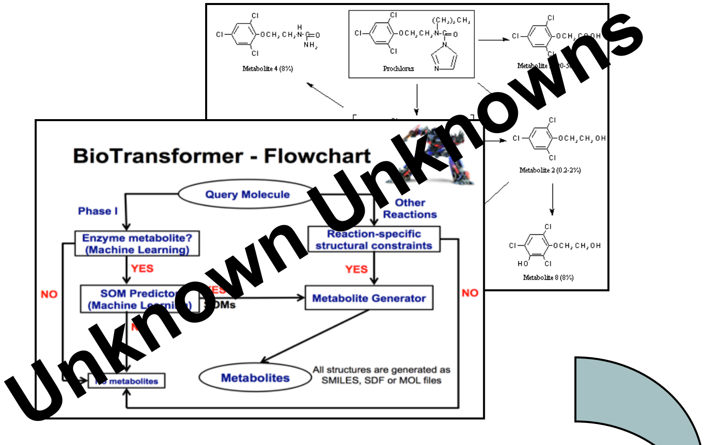
In Silico Metabolomics

HMDB/DrugBank/T3DB


DrugBank
DRUGBANK
Open Data Drug & Drug Target Database
DrugBank Version 4.3
Drug of the day: Formoterol
HMDB
hmdb
The Human Metabolome Database
Welcome to HMDB Version 3.6
T3DB
welcome to the
Toxic Exposome Database
Tweets
Please cite:

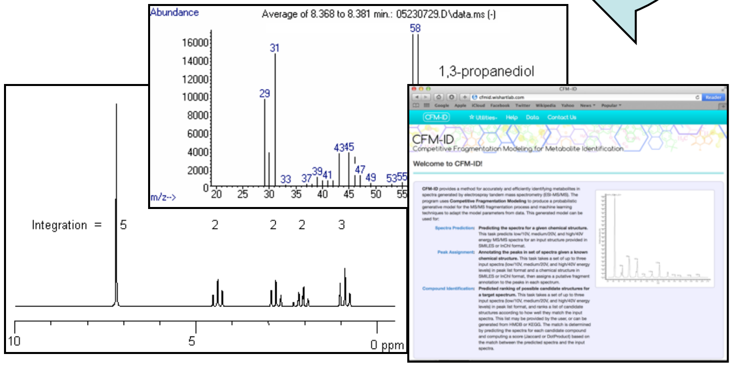
Known compounds (250,000)

BioTransformer

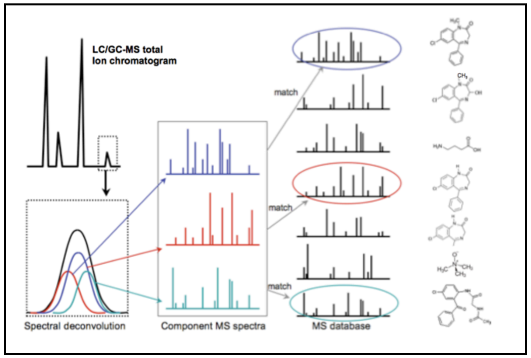

Metabolite 4 (8%)
Prochloraz
Metabolite 2 (0.2-2%)
Metabolite 8 (8%)
BioTransformer - Flowchart
Query Molecule
Phase I
Other Reactions
Enzyme metabolite? (Machine Learning)
Reaction-specific structural constraints
YES
YES
NO
NO
SOM Predictor (Machine Learning)
SOMs
Metabolite Generator
Metabolites
metabolites
All structures are generated as SMILES, SDF or MOL files
Unknown Unknowns

Predicted biotransformations
(250,000 --> 5,000,000)

CFM-ID/NMRPred



Predicted MS/MS, NMR, GC-MS
spectra of knowns + biotransformed



Match observed spectra
to predicted spectra to ID compounds

What Is Known?

UofA Metabolomics Databases

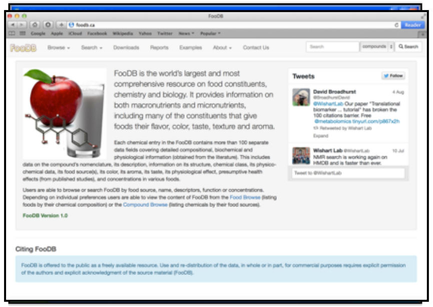
www.hmdb.ca

www.drugbank.ca

www.ymdb.ca

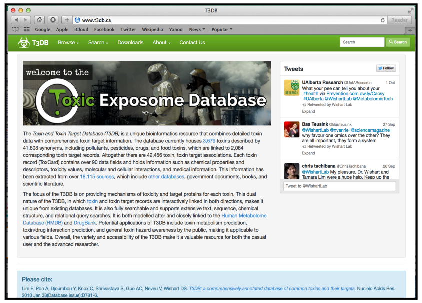
www.phenol-explorer.eu

www.ecmdb.ca



www.foodb.ca

www.cowmetdb.ca



www.t3db.ca

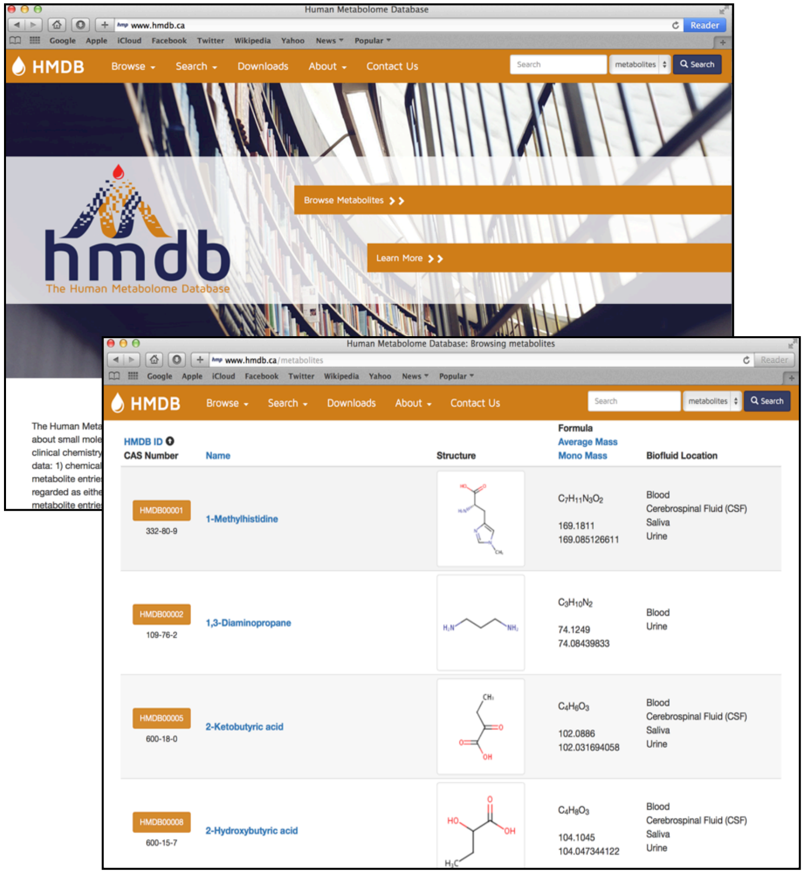
www.smpdb.ca

www.csfmetabolome.ca

www.serummetabolome.ca

www.urinemetabolome.ca

The New Human Metabolome Database (HMDB)

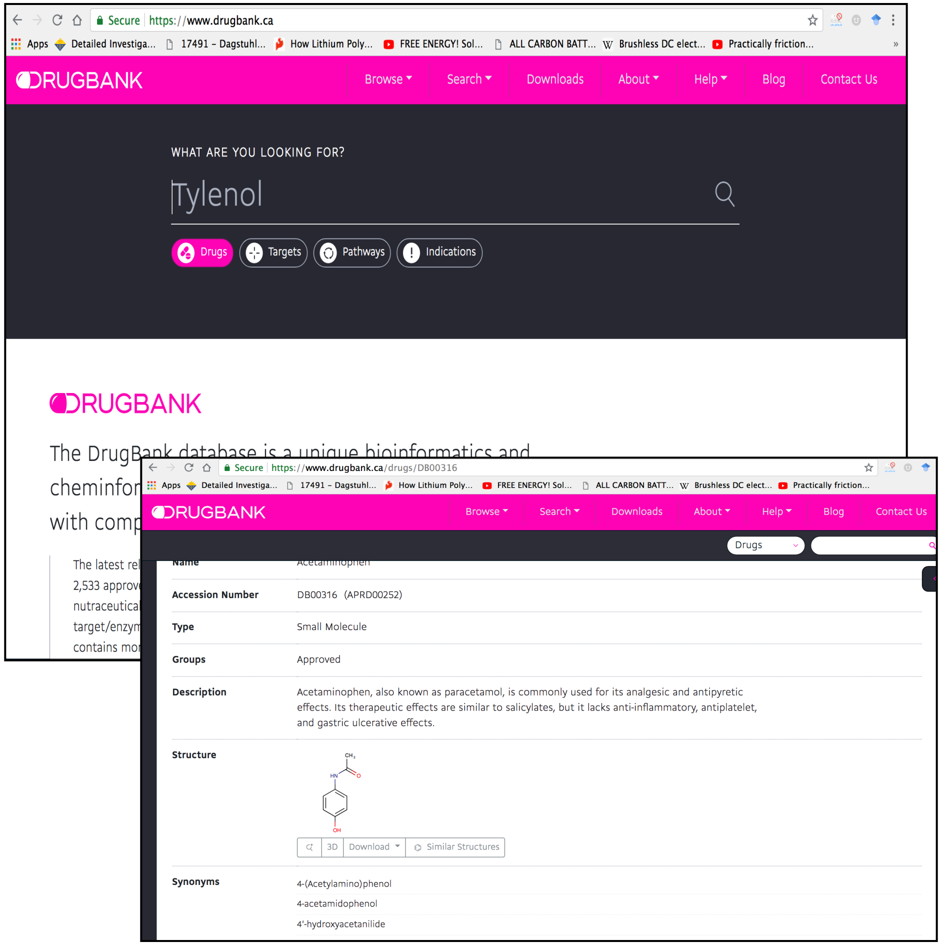

Human Metabolome Database
www.hmdb.ca
Reader
Google Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular
HMDB Browse Search Downloads About Contact Us
Search
metabolites
Search
hmdb
The Human Metabolome Database
Browse Metabolites >>
Learn More >>
The Human Meta
about small mole
clinical chemistry
data: 1) chemical
metabolite entrie
regarded as eithe
metabolite entrie
Human Metabolome Database: Browsing metabolites
www.hmdb.ca/metabolites
Reader
Google Apple iCloud Facebook Twitter Wikipedia Yahoo News Popular
HMDB Browse Search Downloads About Contact Us
Search
metabolites
Search

HMDB ID CAS Number	Name	Structure	Formula Average Mass Mono Mass	Biofluid Location
HMDB00001 332-80-9	1-Methylhistidine		$C_7H_{11}N_3O_2$ 169.1811 169.085126611	Blood Cerebrospinal Fluid (CSF) Saliva Urine
HMDB00002 109-76-2	1,3-Diaminopropane		$C_3H_{10}N_2$ 74.1249 74.08439833	Blood Urine
HMDB00005 600-18-0	2-Ketobutyric acid		$C_4H_6O_3$ 102.0886 102.031694058	Blood Cerebrospinal Fluid (CSF) Saliva Urine
HMDB00008 600-15-7	2-Hydroxybutyric acid		$C_4H_8O_3$ 104.1045 104.047344122	Blood Cerebrospinal Fluid (CSF) Saliva Urine

http://www.hmdb.ca

- HMDB 4.0 has 114,100 “quantified”, “detected”, “expected” and “predicted” metabolites (3X more than version 3.0)
- HMDB 3.0 had 442 biological pathways, HMDB 4.1 has 48,627 (100X more than version 3.0)
- New version has >500,000 MS/MS & GC-MS spectra, 3900 NMR spectra
- New version has 6800 metabolite-SNP associations, 2500 metabolite-drug associations and 2900 metabolite-age/gender associations
- 78,000 new lipids/peptides to be added in late 2018 – total = 192,000

The New Drug Database (DrugBank v. 5.0)



The DrugBank database is a unique bioinformatics and
cheminfor
with comp

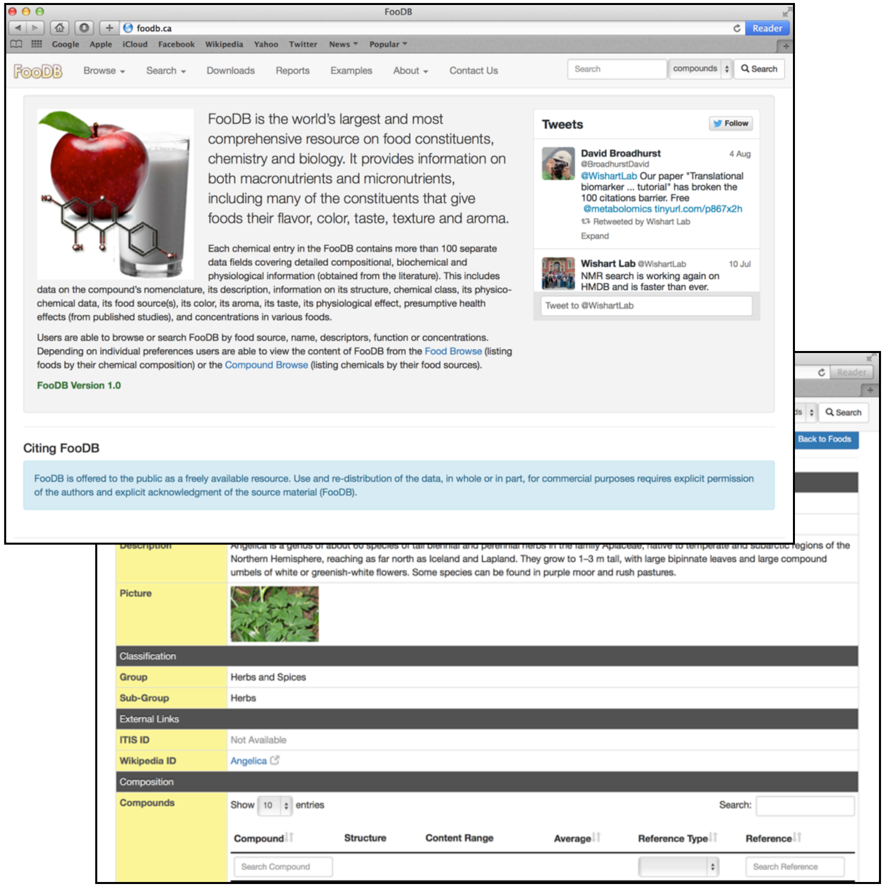
The latest rel
2,533 approv
nutraceutica
target/enzym
contains mor

Name	Acetaminophen
Accession Number	DB00316 (APRD00252)
Type	Small Molecule
Groups	Approved
Description	Acetaminophen, also known as paracetamol, is commonly used for its analgesic and antipyretic effects. Its therapeutic effects are similar to salicylates, but it lacks anti-inflammatory, antiplatelet, and gastric ulcerative effects.
Structure	
Synonyms	4-(Acetylamino)phenol 4-acetamidophenol 4'-hydroxyacetanilide

- A comprehensive database of drugs, drug actions and drug targets
- 2533 small molecule drugs
- >5700 experimental drugs
- Detailed ADMET, MOA and pharmacokinetic data
- >3850 drugs with metabolizing enzyme data
- >1360 drug metabolites
- >6000 MS+NMR spectra
- >5130 unique drug targets
- 215 data fields/drug
- *Published on Jan. 1, 2018*

http://www.drugbank.ca

The Food Database (FooDB)



FooDB is the world's largest and most comprehensive resource on food constituents, chemistry and biology. It provides information on both macronutrients and micronutrients, including many of the constituents that give foods their flavor, color, taste, texture and aroma.

Each chemical entry in the FooDB contains more than 100 separate data fields covering detailed compositional, biochemical and physiological information (obtained from the literature). This includes data on the compound's nomenclature, its description, information on its structure, chemical class, its physico-chemical data, its food source(s), its color, its aroma, its taste, its physiological effect, presumptive health effects (from published studies), and concentrations in various foods.

Users are able to browse or search FooDB by food source, name, descriptors, function or concentrations. Depending on individual preferences users are able to view the content of FooDB from the Food Browse (listing foods by their chemical composition) or the Compound Browse (listing chemicals by their food sources).

FooDB Version 1.0

David Broadhurst @BroadhurstDavid 4 Aug
@WishartLab Our paper "Translational biomarker ... tutorial" has broken the 100 citations barrier. Free @metabolomics tinyurl.com/p867x2h
Retweeted by Wishart Lab

Wishart Lab @WishartLab 10 Jul
NMR search is working again on HMDB and is faster than ever.

Citing FooDB

FooDB is offered to the public as a freely available resource. Use and re-distribution of the data, in whole or in part, for commercial purposes requires explicit permission of the authors and explicit acknowledgment of the source material (FooDB).

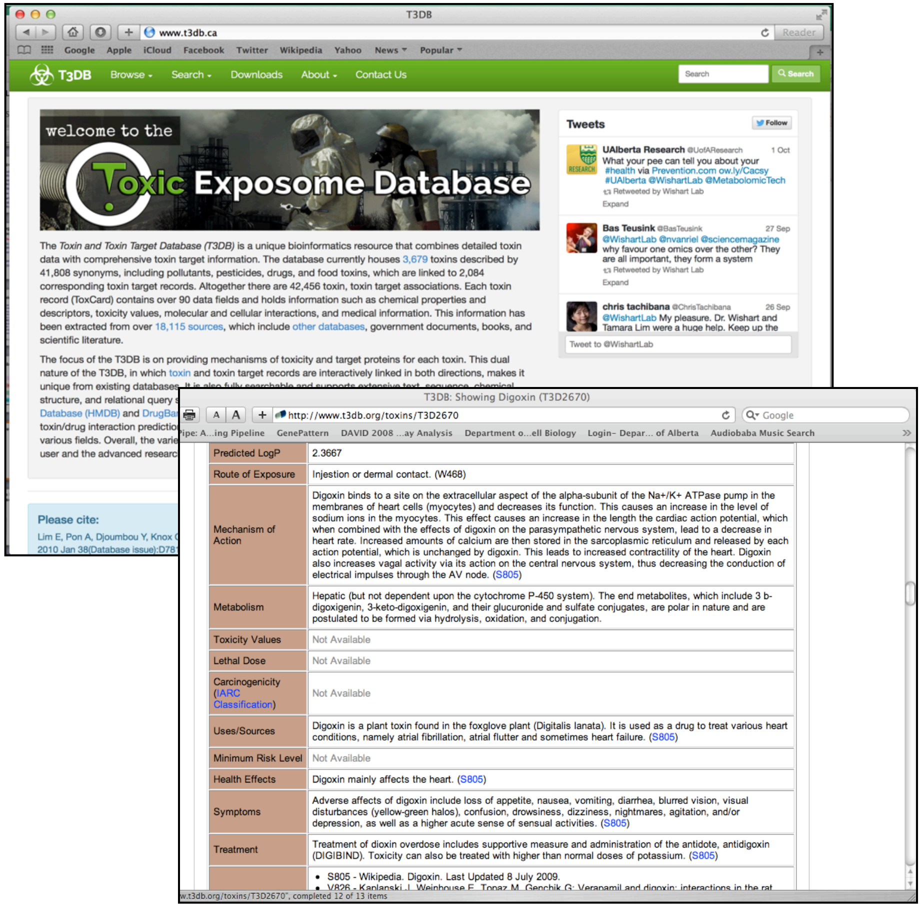
Description	Angelica is a genus of about 60 species of tall biennial and perennial herbs in the family Apiaceae, native to temperate and subarctic regions of the Northern Hemisphere, reaching as far north as Iceland and Lapland. They grow to 1–3 m tall, with large bipinnate leaves and large compound umbels of white or greenish-white flowers. Some species can be found in purple moor and rush pastures.
Picture	
Classification	
Group	Herbs and Spices
Sub-Group	Herbs
External Links	
ITIS ID	Not Available
Wikipedia ID	Angelica
Composition	
Compounds	Show 10 entries Search:

Compound	Structure	Content Range	Average	Reference Type	Reference
Search Compound					Search Reference

- A comprehensive food composition database (more than polyphenols)
- 28,771 compounds
- 718,405 concentration values for 722 raw/processed foods
- 31,791 references
- 1435 cmpds with health effects
- 2692 cmpds w flavour attributes
- 2000+ reference MS/NMR spectra
- Structure & text searches
- >100 data fields/compound
- *Publicly released on Jan. 1, 2018, manuscript being prepared*

www.foodb.ca

The Toxic Exposome Database (T3DB)



The *Toxin and Toxin Target Database (T3DB)* is a unique bioinformatics resource that combines detailed toxin data with comprehensive toxin target information. The database currently houses 3,679 toxins described by 41,808 synonyms, including pollutants, pesticides, drugs, and food toxins, which are linked to 2,084 corresponding toxin target records. Altogether there are 42,456 toxin, toxin target associations. Each toxin record (ToxCard) contains over 90 data fields and holds information such as chemical properties and descriptors, toxicity values, molecular and cellular interactions, and medical information. This information has been extracted from over 18,115 sources, which include other databases, government documents, books, and scientific literature.

The focus of the T3DB is on providing mechanisms of toxicity and target proteins for each toxin. This dual nature of the T3DB, in which toxin and toxin target records are interactively linked in both directions, makes it unique from existing databases. It is also fully searchable and supports extensive text, sequence, chemical structure, and relational query s
Database (HMDB) and DrugBa
toxin/drug interaction predictio
various fields. Overall, the varie
user and the advanced researc

Please cite:

Lim E, Pon A, Djoumbou Y, Knox C
2010 Jan 38(Database issue):D781

UAlberta Research @UofAResearch 1 Oct
What your pee can tell you about your #health via Prevention.com ow.ly/Cacsy #UAlberta @WishartLab @MetabolomicTech
Retweeted by Wishart Lab

Bas Teusink @BasTeusink 27 Sep
@WishartLab @nvanriel @sciencemagazine why favour one omics over the other? They are all important, they form a system
Retweeted by Wishart Lab

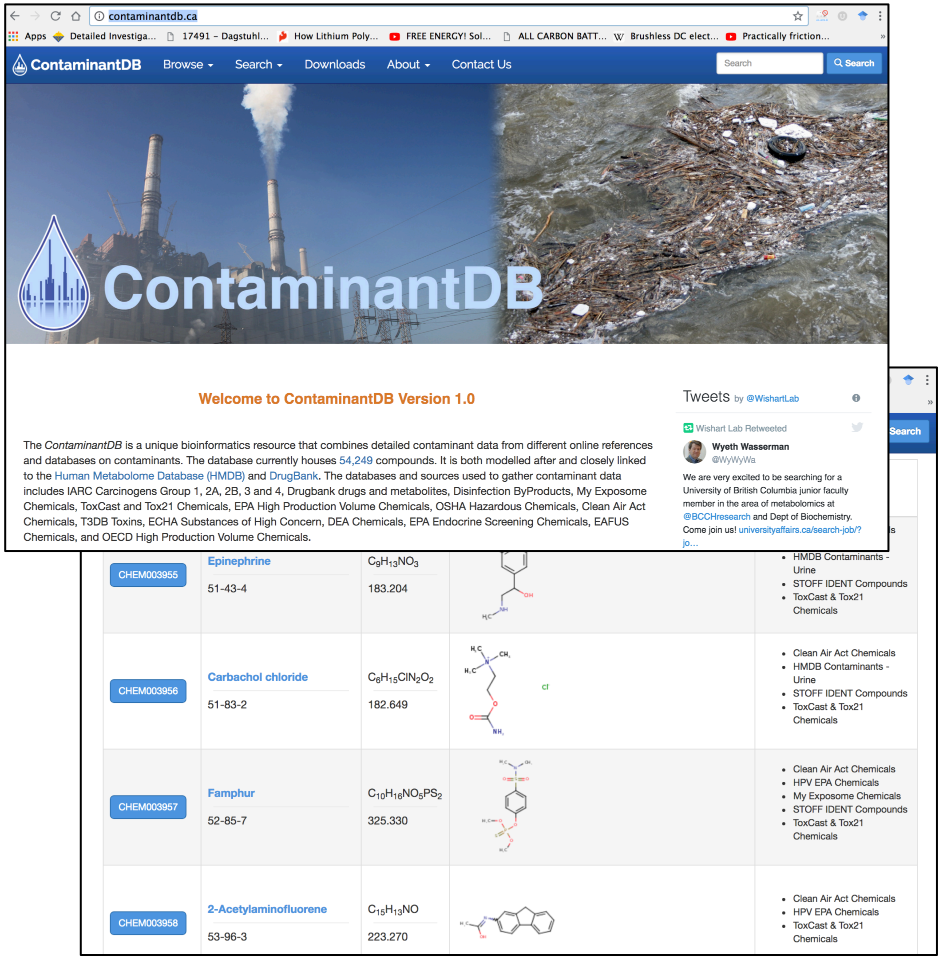
chris tachibana @ChrisTachibana 26 Sep
@WishartLab My pleasure. Dr. Wishart and Tamara Lim were a huge help. Keep up the

Predicted LogP	2.3667
Route of Exposure	Injestion or dermal contact. (W468)
Mechanism of Action	Digoxin binds to a site on the extracellular aspect of the alpha-subunit of the Na+/K+ ATPase pump in the membranes of heart cells (myocytes) and decreases its function. This causes an increase in the level of sodium ions in the myocytes. This effect causes an increase in the length the cardiac action potential, which when combined with the effects of digoxin on the parasympathetic nervous system, lead to a decrease in heart rate. Increased amounts of calcium are then stored in the sarcoplasmic reticulum and released by each action potential, which is unchanged by digoxin. This leads to increased contractility of the heart. Digoxin also increases vagal activity via its action on the central nervous system, thus decreasing the conduction of electrical impulses through the AV node. (S805)
Metabolism	Hepatic (but not dependent upon the cytochrome P-450 system). The end metabolites, which include 3 b-digoxigenin, 3-keto-digoxigenin, and their glucuronide and sulfate conjugates, are polar in nature and are postulated to be formed via hydrolysis, oxidation, and conjugation.
Toxicity Values	Not Available
Lethal Dose	Not Available
Carcinogenicity (IARC Classification)	Not Available
Uses/Sources	Digoxin is a plant toxin found in the foxglove plant (Digitalis lanata). It is used as a drug to treat various heart conditions, namely atrial fibrillation, atrial flutter and sometimes heart failure. (S805)
Minimum Risk Level	Not Available
Health Effects	Digoxin mainly affects the heart. (S805)
Symptoms	Adverse affects of digoxin include loss of appetite, nausea, vomiting, diarrhea, blurred vision, visual disturbances (yellow-green halos), confusion, drowsiness, dizziness, nightmares, agitation, and/or depression, as well as a higher acute sense of sensual activities. (S805)
Treatment	Treatment of dioxin overdose includes supportive measure and administration of the antidote, antidigoxin (DIGIBIND). Toxicity can also be treated with higher than normal doses of potassium. (S805)
	• S805 - Wikipedia. Digoxin. Last Updated 8 July 2009. • V826 - Kaplanski J, Weinhouse E, Topaz M, Genchik G: Verapamil and digoxin: interactions in the rat

http://www.t3db.ca

- Comprehensive data on toxic compounds (drugs, pesticides, herbicides, endocrine disruptors, drugs, solvents, carcinogens, etc.)
- Detailed mechanisms, binding constants, target info, lots of ToxCast data
- >3600 toxic compounds
- >1900 reference spectra
- ~2100 toxic targets
- Supports sequence, spectral, structure, text searches as well as compound browsing
- Full data downloads

ContaminantDB



Welcome to ContaminantDB Version 1.0

The *ContaminantDB* is a unique bioinformatics resource that combines detailed contaminant data from different online references and databases on contaminants. The database currently houses 54,249 compounds. It is both modelled after and closely linked to the Human Metabolome Database (HMDB) and DrugBank. The databases and sources used to gather contaminant data includes IARC Carcinogens Group 1, 2A, 2B, 3 and 4, Drugbank drugs and metabolites, Disinfection ByProducts, My Exposome Chemicals, ToxCast and Tox21 Chemicals, EPA High Production Volume Chemicals, OSHA Hazardous Chemicals, Clean Air Act Chemicals, T3DB Toxins, ECHA Substances of High Concern, DEA Chemicals, EPA Endocrine Screening Chemicals, EAFUS Chemicals, and OECD High Production Volume Chemicals.

Tweets by @WishartLab

Wishart Lab Retweeted

Wyeth Wasserman
@WyWyWa

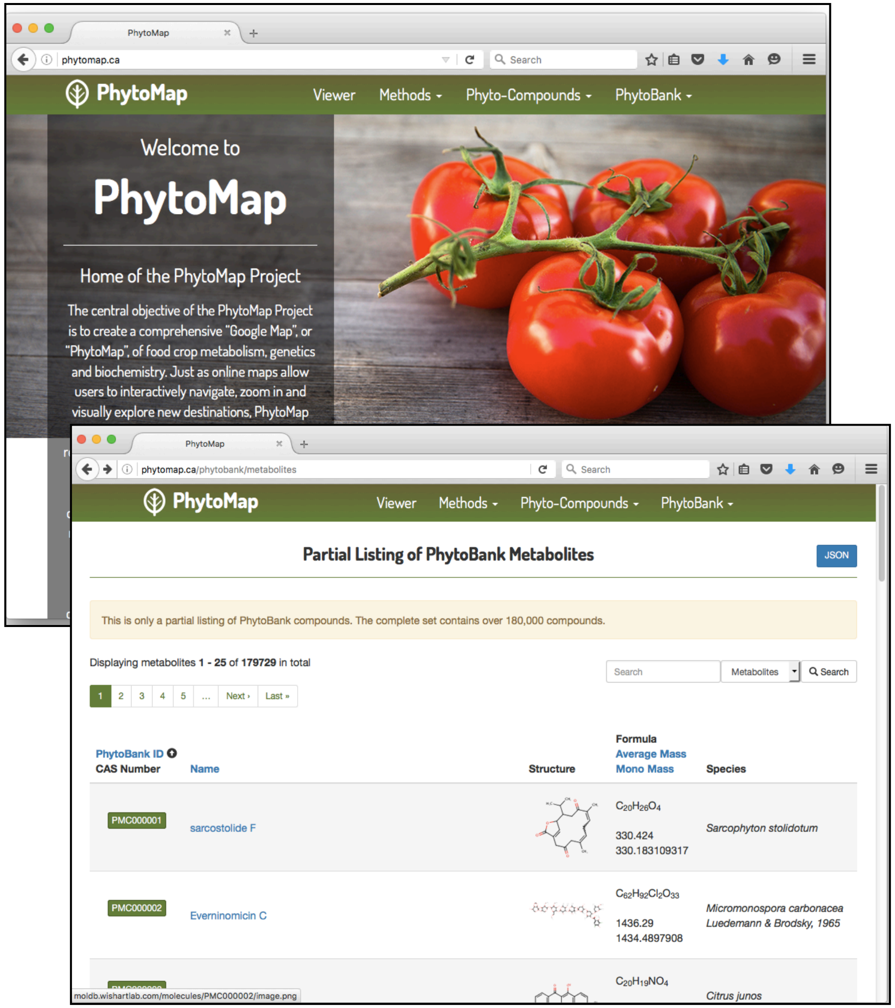
We are very excited to be searching for a University of British Columbia junior faculty member in the area of metabolomics at @BCCHresearch and Dept of Biochemistry. Come join us! universityaffairs.ca/search-job/?jo...

CHEM003955	Epinephrine 51-43-4	$C_9H_{13}NO_3$ 183.204		HMDB Contaminants - Urine STOFF IDENT Compounds ToxCast & Tox21 Chemicals
CHEM003956	Carbachol chloride 51-83-2	$C_6H_{15}ClN_2O_2$ 182.649		Clean Air Act Chemicals HMDB Contaminants - Urine STOFF IDENT Compounds ToxCast & Tox21 Chemicals
CHEM003957	Famphur 52-85-7	$C_{10}H_{16}NO_5PS_2$ 325.330		Clean Air Act Chemicals HPV EPA Chemicals My Exposome Chemicals STOFF IDENT Compounds ToxCast & Tox21 Chemicals
CHEM003958	2-Acetylaminofluorene 53-96-3	$C_{15}H_{13}NO$ 223.270		Clean Air Act Chemicals HPV EPA Chemicals ToxCast & Tox21 Chemicals

www.contaminantdb.ca

- Data on 54,249 probable or known chemical contaminants
- Expected to grow to 80,000+ by Sept. 2018
- Exp. MS data for 5000+ cmpds
- Pred. 54,000 EI-MS spectra, 150,000 ESI-MS/MS spectra
- Source or role information for most compounds
- >40% of the compounds in ContaminantDB are not found in PubChem or ChemSpider
- Supports spectral, structure and text searches as well as compound browsing

PhytoBank



Welcome to

PhytoMap

Home of the PhytoMap Project

The central objective of the PhytoMap Project is to create a comprehensive "Google Map", or "PhytoMap", of food crop metabolism, genetics and biochemistry. Just as online maps allow users to interactively navigate, zoom in and visually explore new destinations, PhytoMap

Partial Listing of PhytoBank Metabolites

This is only a partial listing of PhytoBank compounds. The complete set contains over 180,000 compounds.

Displaying metabolites **1 - 25** of **179729** in total

PhytoBank ID CAS Number	Name	Structure	Formula Average Mass Mono Mass	Species
PMC000001	sarcostolide F		$C_{20}H_{26}O_4$ 330.424 330.183109317	*Sarcophyton stolidotum*
PMC000002	Everninomicin C		$C_{62}H_{92}Cl_2O_{33}$ 1436.29 1434.4897908	*Micromonospora carbonacea Luedemann & Brodsky, 1965*
			$C_{20}H_{19}NO_4$	*Citrus junos*

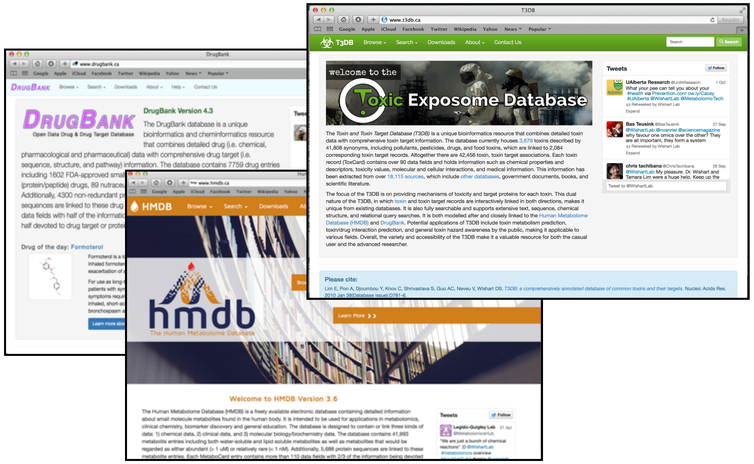
moldb.wishartlab.com/molecules/PMC000002/image.png

- **179,729 plant-derived compounds from more than 23,700 plant species including >8,318 food/ crop plants and >2,439 medicinal plants**
- **>33% of the compounds in PhytoBank are not found in PubChem or ChemSpider**
- **Will offer same resources as HMDB, DrugBank, etc.**

What Can We Predict?

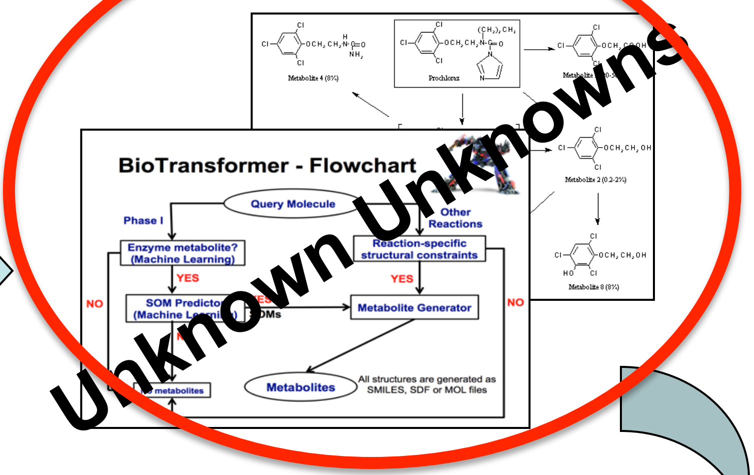
In Silico Metabolomics

HMDB/DrugBank/T3DB



Known compounds (250,000)

BioTransformer



BioTransformer - Flowchart

Query Molecule

Phase I

Other Reactions

Enzyme metabolite? (Machine Learning)

Reaction-specific structural constraints

YES

YES

NO

NO

SOM Predictor (Machine Learning)

Metabolite Generator

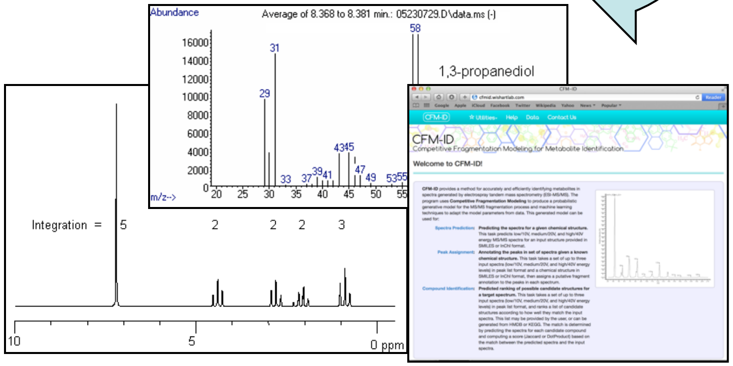
metabolites

Metabolites

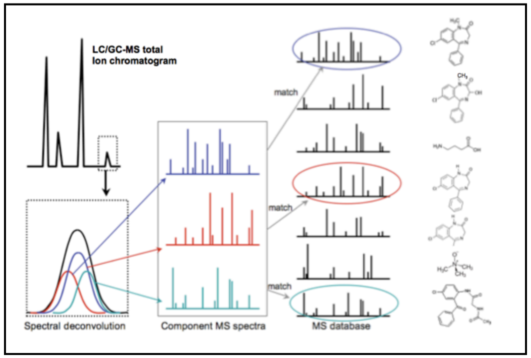
All structures are generated as SMILES, SDF or MOL files

Predicted biotransformations
(250,000 --> 5,000,000)

CFM-ID/NMRPred

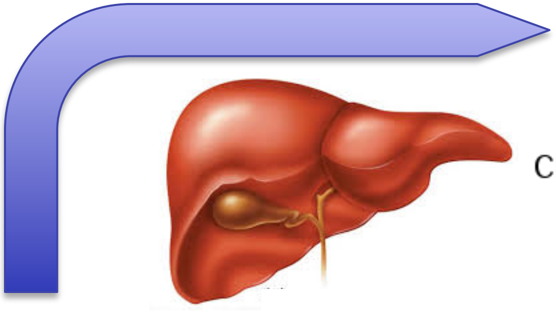


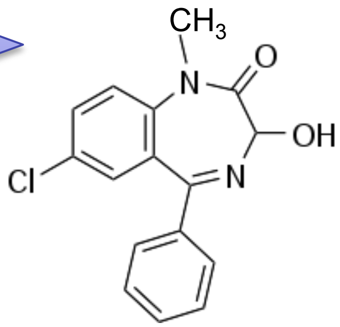
Predicted MS/MS, NMR, GC-MS
Spectra of knowns + biotransformed



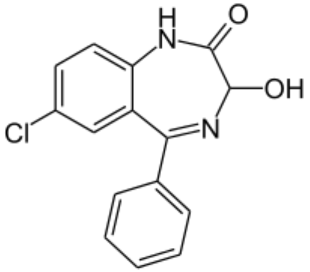
Match observed spectra
to predicted spectra to ID compounds

Examples of Biotransformation

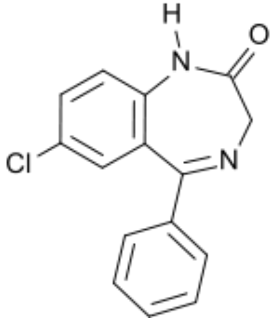




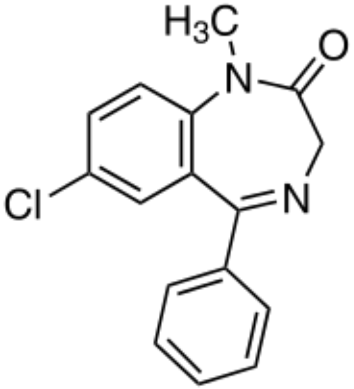
Tempazepam



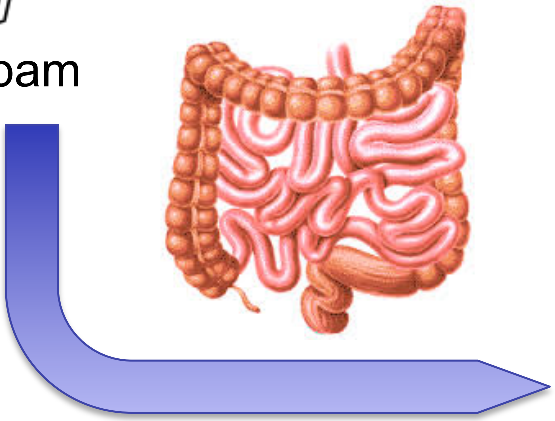
Oxazepam

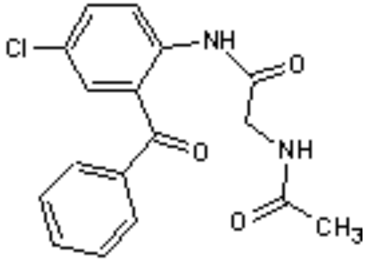


Nordazepam



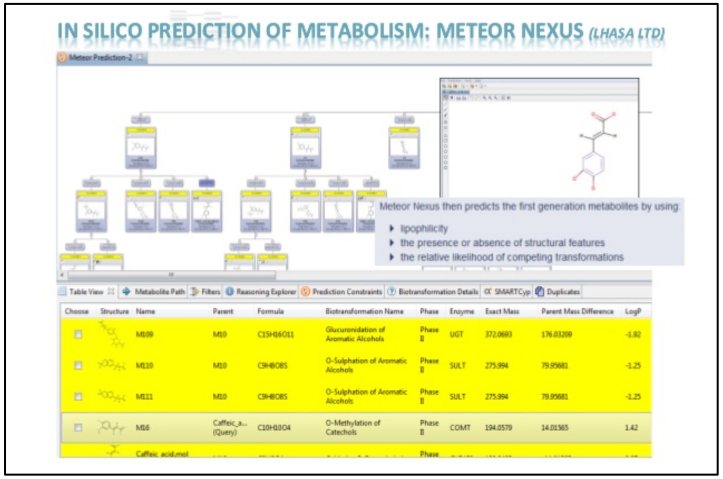
Diazepam





N-(2-Benzoyl-4-chlorophenyl)-2-acetamidoacetamide

Commercial Tools



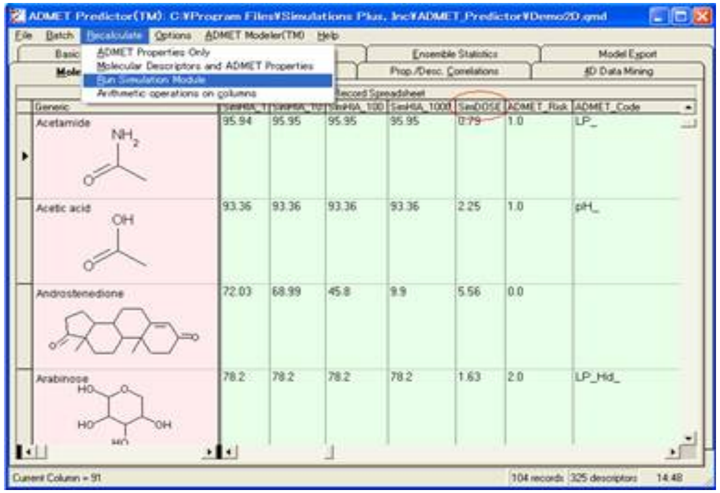
IN SILICO PREDICTION OF METABOLISM: METEOR NEXUS *(LHASA LTD)*

Meteor Nexus then predicts the first generation metabolites by using:

- lipophilicity
- the presence or absence of structural features
- the relative likelihood of competing transformations

Choose	Structure	Name	Parent	Formula	Biotransformation Name	Phase	Enzyme	Exact Mass	Parent Mass Difference	LogP
□		M109	M10	C15H16O11	Glucuronidation of Aromatic Alcohols	Phase II	UGT	372.0693	176.03209	-1.92
□		M110	M10	C9H8O8S	O-Sulphation of Aromatic Alcohols	Phase II	SULT	275.994	79.95681	-1.25
□		M111	M10	C9H8O8S	O-Sulphation of Aromatic Alcohols	Phase II	SULT	275.994	79.95681	-1.25
□		M16	Caffeic_a... (Query)	C10H10O4	O-Methylation of Catechols	Phase II	COMT	194.0579	14.01565	1.42
		Caffeic_acid.mol				Phase				

Meteor-Nexus

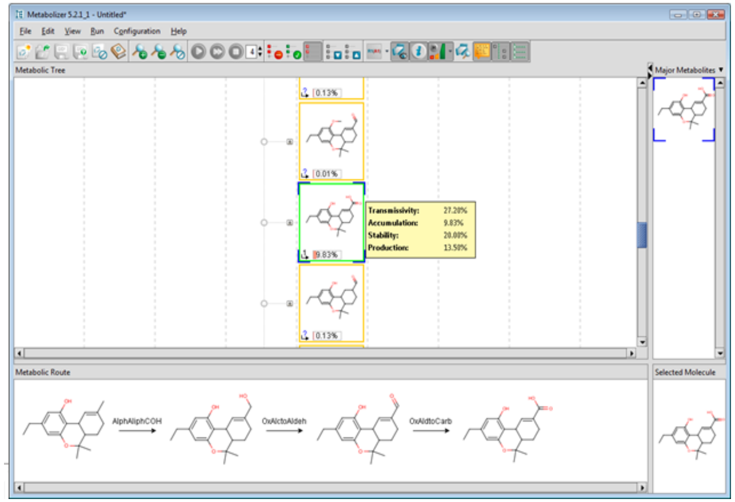


ADMET Predictor(TM): C:¥Program Files¥Simulations Plus, Inc¥ADMET_Predictor¥Demo2D.gmd

Generic	SimHIA_1	SimHIA_10	SimHIA_100	SimHIA_1000	SimDOSE	ADMET_Risk	ADMET_Code
Acetamide	95.94	95.95	95.95	95.95	0.79	1.0	LP_
Acetic acid	93.36	93.36	93.36	93.36	2.25	1.0	pH_
Androstenedione	72.03	68.99	45.8	9.9	5.56	0.0	
Arabinose	78.2	78.2	78.2	78.2	1.63	2.0	LP_Hid_

Current Column = 91 | 104 records | 325 descriptors | 14:48

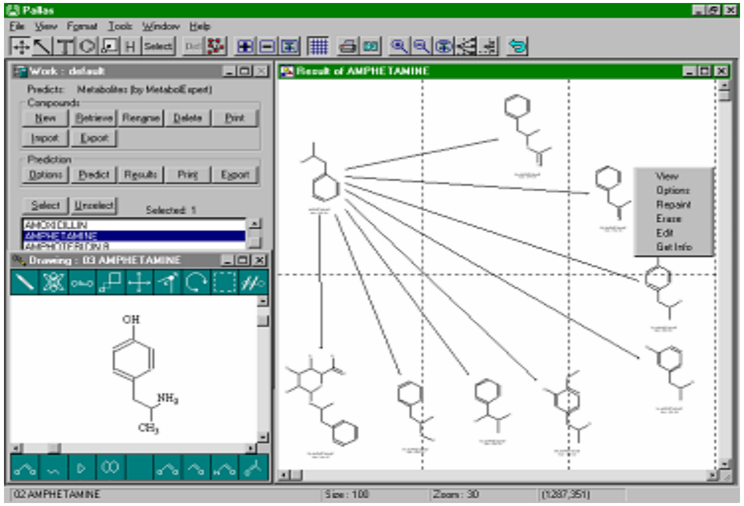
ADMET Predictor



Metabolizer 5.2.1_1 - Untitled*

Transmissivity: 27.20%
Accumulation: 9.83%
Stability: 20.00%
Production: 13.50%

Metabolizer



Predicts: Metabolites (by MetabolExpert)

Compounds

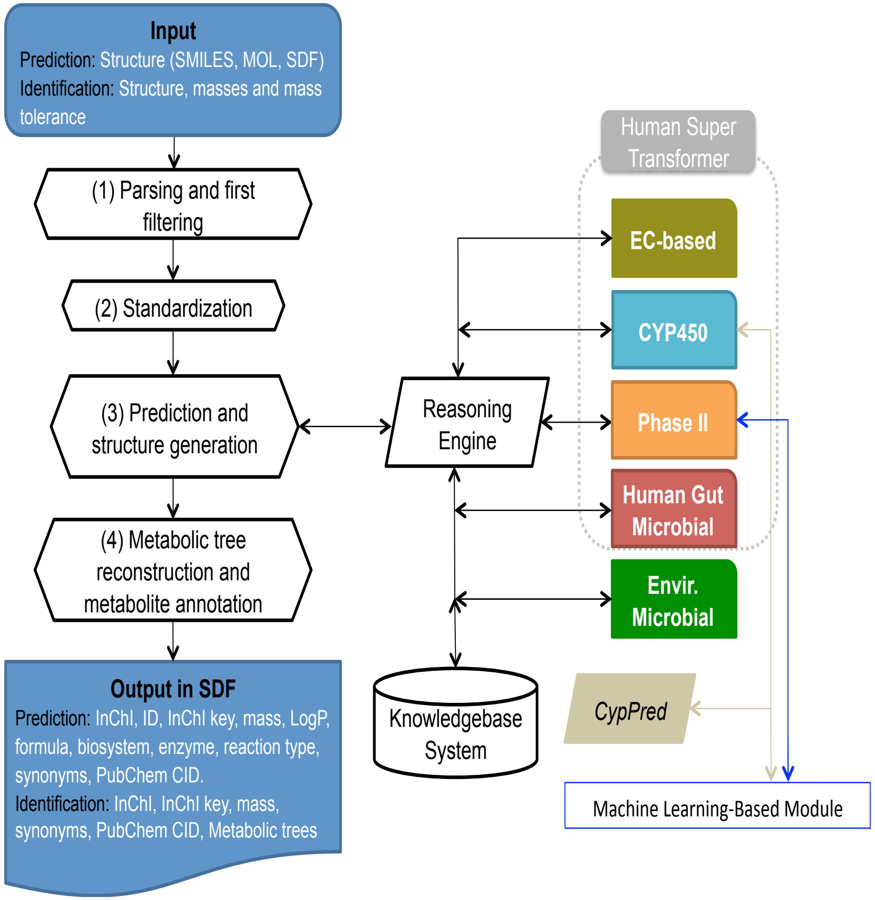
Selected: 1

AMOXICILLIN
AMPHETAMINE
AMPHOTERICIN B

02 AMPHETAMINE | Size : 100 | Zoom : 30 | (1287,351)

MetabolExpert

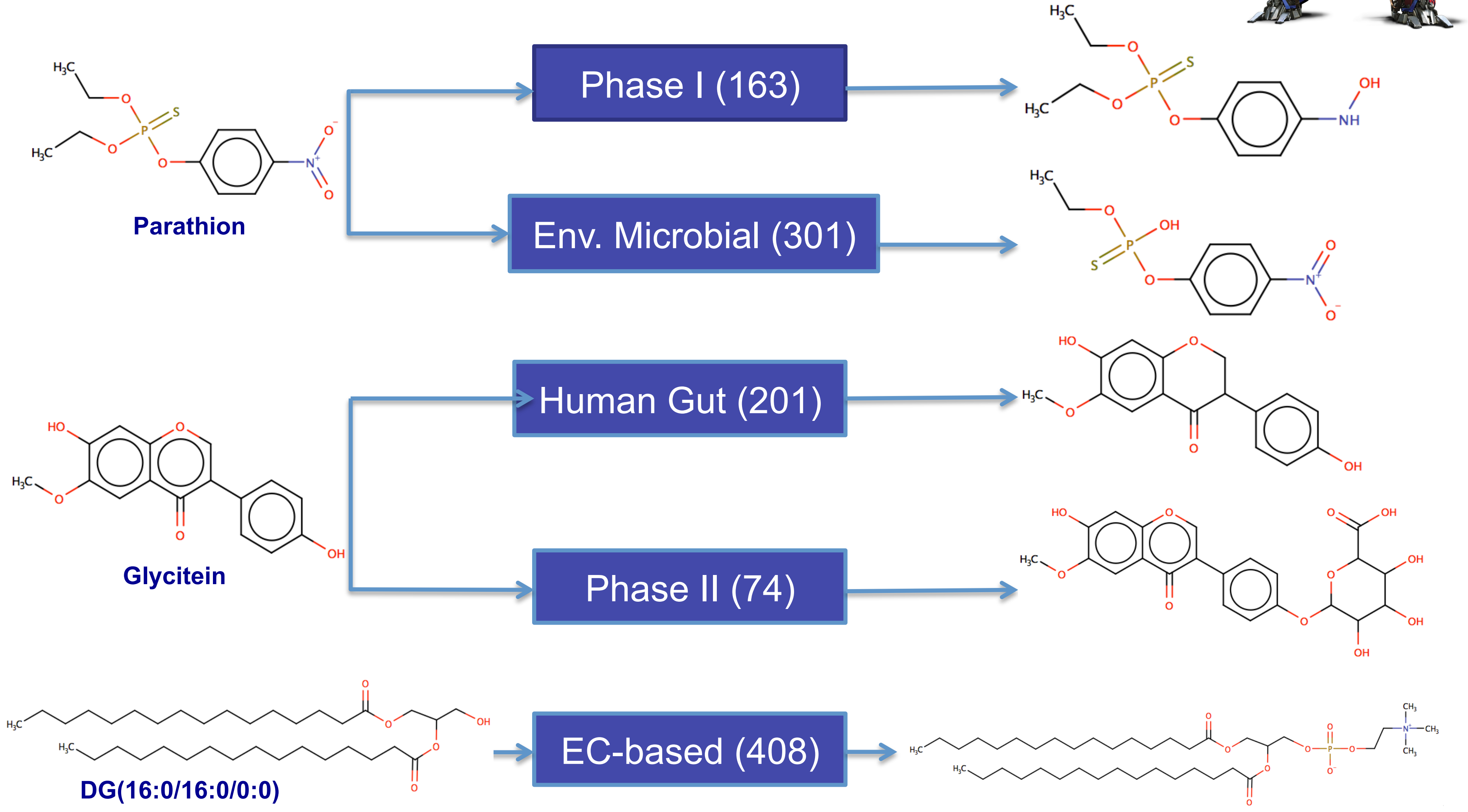
BioTransformer (Free)



Input
Prediction: Structure (SMILES, MOL, SDF)
Identification: Structure, masses and mass tolerance
(1) Parsing and first filtering
(2) Standardization
(3) Prediction and structure generation
(4) Metabolic tree reconstruction and metabolite annotation
Output in SDF
Prediction: InChI, ID, InChI key, mass, LogP, formula, biosystem, enzyme, reaction type, synonyms, PubChem CID.
Identification: InChI, InChI key, mass, synonyms, PubChem CID, Metabolic trees
Reasoning Engine
Knowledgebase System
Human Super Transformer
EC-based
CYP450
Phase II
Human Gut Microbial
Envir. Microbial
CypPred
Machine Learning-Based Module

- A comprehensive tool for predicting metabolite structures that have been biotransformed through phase I, phase II, microbial, promiscuous and environmental processes
- Uses a large knowledgebase and a large set of heuristic (and machine-learned) biotransformation rules
- Performs much better than well-known commercial tools
- *Publicly released, manuscript submitted*

BioTransformer



Parathion

Phase I (163)

Env. Microbial (301)

Glycitein

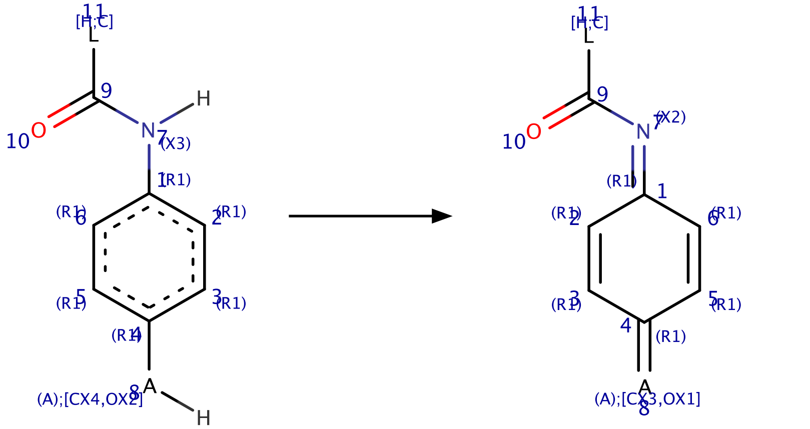
Human Gut (201)

Phase II (74)

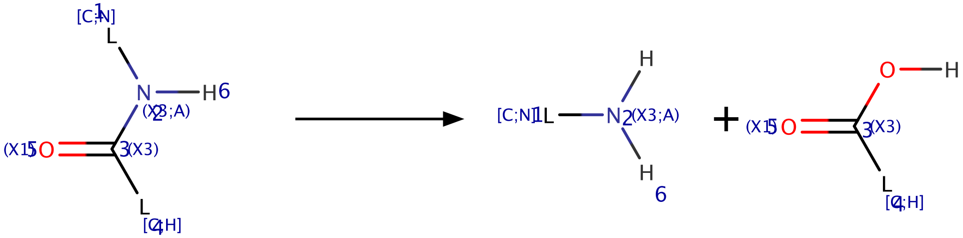
DG(16:0/16:0/0:0)

EC-based (408)

BioTransformer



Oxidation of p-substituted anilides (BTMR1018)
Human (Liver) Phase I Metabolism

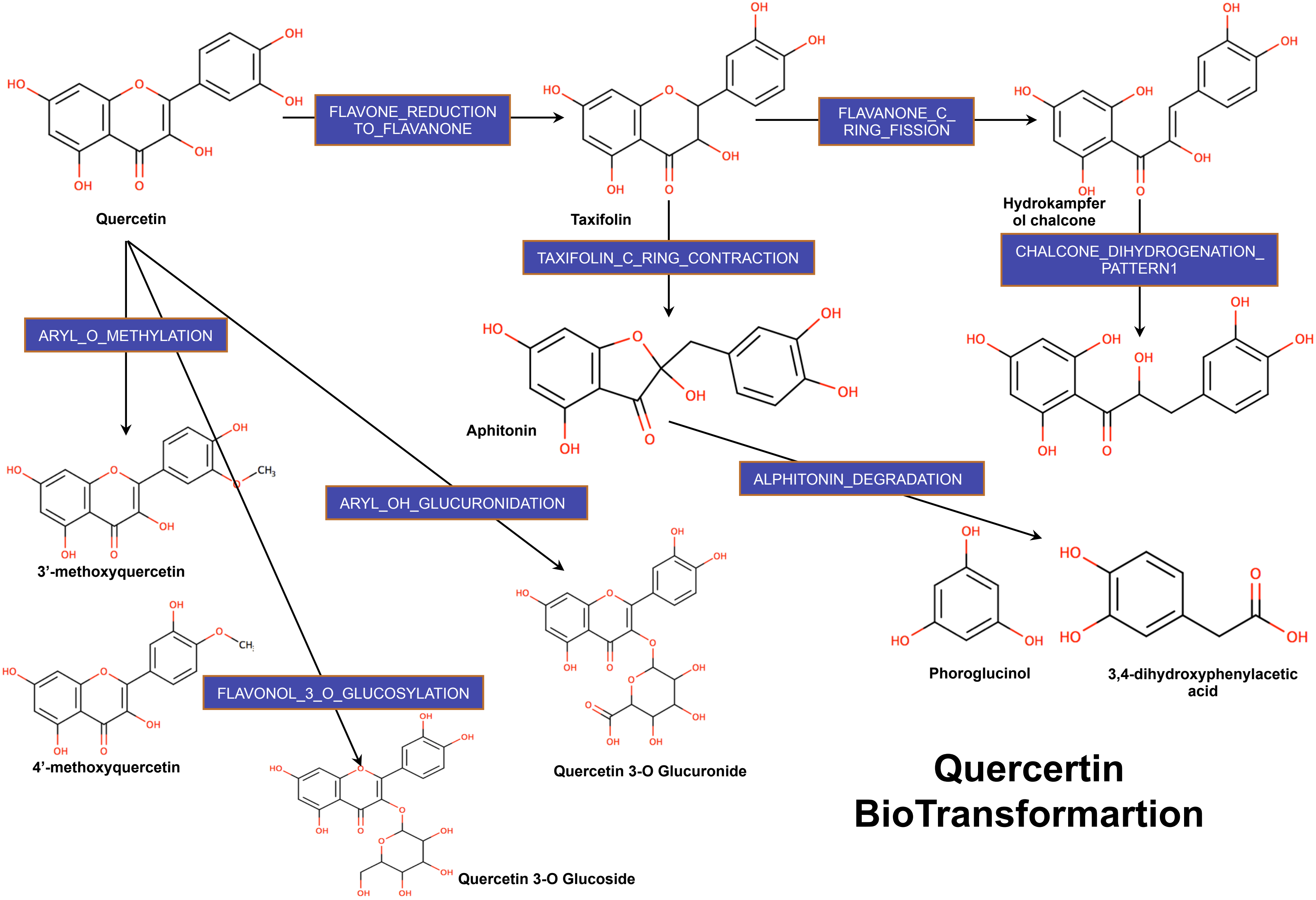


Hydrolysis of secondary amide (BTMR0704)
Environmental microbial (EAWAG-BBD)

	No. of enzymes	No. of biotransformation rules	No. of enzyme-rule associations	No. of covered biosystems
EC-based	285	408	459	2
CYP450	9	163	712	1
Human gut micro.	53	201	204	2
Phase II	9	74	81	2
Envir. microbial	1	301	301	1

Overall, 1,240 biotransformation rules were manually designed and tested. Overall, 2,150 reaction preference rules were implemented.

Quercertin BioTransformartion



Quercetin

FLAVONE_REDUCTION_TO_FLAVANONE

Taxifolin

FLAVANONE_C_RING_FISSION

Hydrokampferol chalcone

TAXIFOLIN_C_RING_CONTRACTION

CHALCONE_DIHYDROGENATION_PATTERN1

ARYL_O_METHYLATION

Aphitonin

ALPHITONIN_DEGRADATION

3’-methoxyquercetin

ARYL_OH_GLUCURONIDATION

Phoroglucinol

3,4-dihydroxyphenylacetic acid

4’-methoxyquercetin

FLAVONOL_3_O_GLUCOSYLATION

Quercetin 3-O Glucuronide

Quercetin 3-O Glucoside

BioTransformer Evaluation

	BioTransformer	Meteor
True Positives	188	153
False Positives	198	281
False Negatives	35	70
Total No. of Predictions	386	431
Precision	0.49	0.35
Recall	0.84	0.69
# of reported metabolites	223	

1. Test set: 40 compounds (incl. drugs and pesticides)
2. Metabolism data was retrieved from >60 references
3. BioTransformer (v. 1.0.4) and Meteor Nexus (v. 3.0.1) were used for 1- step prediction of **human metabolism**

BioTransformer Evaluation

1.Test set: 60 compounds (incl. drugs, pesticides, food compounds, steroids)
2.Metabolism data was retrieved from 60+ references
3.BioTransformer (v. 1.0.4) and ADMET Predictor (v. 8.5.11) were used for 1- step prediction of **human CYP450 metabolism**

	BioTransformer	ADMET Predictor
True Positives	162	110
False Positives	188	122
False Negatives	18	70
Total No. Predictions	350	232
Precision	0.46	0.47
Recall	0.9	0.61
No. of Reported Metabolites	180	

1.Test set: 20 compounds (incl. endogenous metabolites, phytochemicals, and other xenobiotics)
2.Metabolism data was retrieved from >50 references
3.BioTransformer (v. 1.0.4) was used for 1-step prediction of **human and gut microbial metabolism**

BioTransformer	
True Positives	111
False Positives	49
False Negatives	17
Total No. Predictions	160
Precision	0.69
Recall	0.87
No. of Reported Metabolites	128

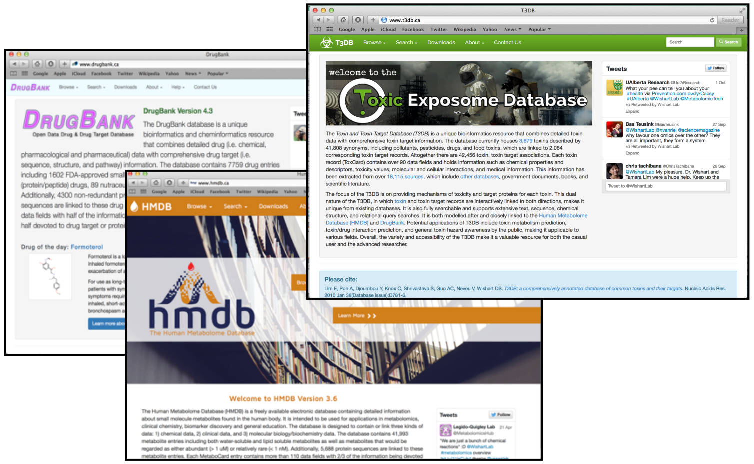
BioTransformer Updates

- **Added/updated:**
  - Available as a web server and a program, added metabolite ID option via m/z or molecular formula
- **Number of compounds expected**
  - 1-step reactions generate 5X as many compounds while 2-step reactions generate 20X as many compounds
  - Expect HMDB+BioTransformer will generate 2.2 million new compounds, all DBs+BioTransformer = 5 million cpds
- **Benchmark for computing time**
  - 1,000 FooDB compounds generate 5,071 human and gut microbial metabolites in 1 step (all enzymes)
  - 1h 29 mins (~5.35 s/compound per processor)

If you want compounds processed now, send your queries to
Yannick Djoumbou Feunang --- djoumbou@ualberta.ca

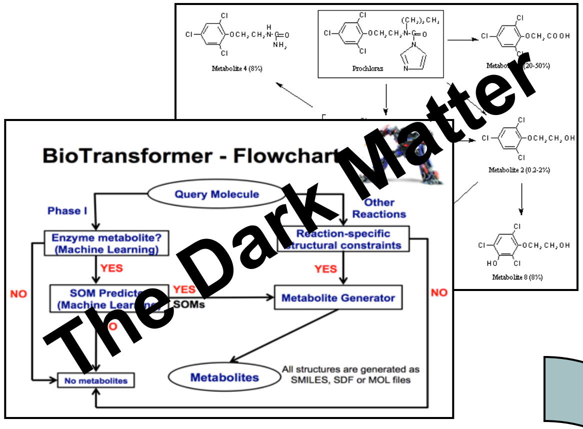
In Silico Metabolomics

HMDB/DrugBank/T3DB



Known compounds (250,000)

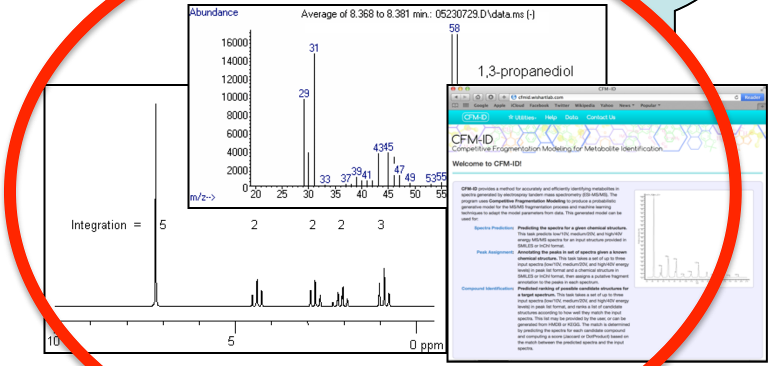
BioTransformer


Prochloraz
Metabolite 4 (8%)
Metabolite 2 (0.2-2%)
Metabolite 8 (8%)
BioTransformer - Flowchart
Query Molecule
Phase I
Other Reactions
Enzyme metabolite? (Machine Learning)
Reaction-specific structural constraints
YES
YES
NO
NO
SOM Predict (Machine Lear
YES
SOMs
Metabolite Generator
No metabolites
Metabolites
All structures are generated as SMILES, SDF or MOL files

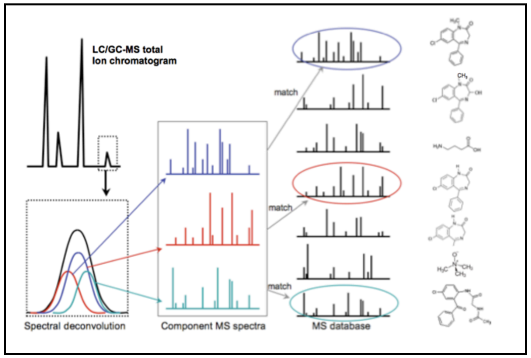
The Dark Matter

Predicted biotransformations
(250,000 --> 2,500,000)

CFM-ID/NMRPred

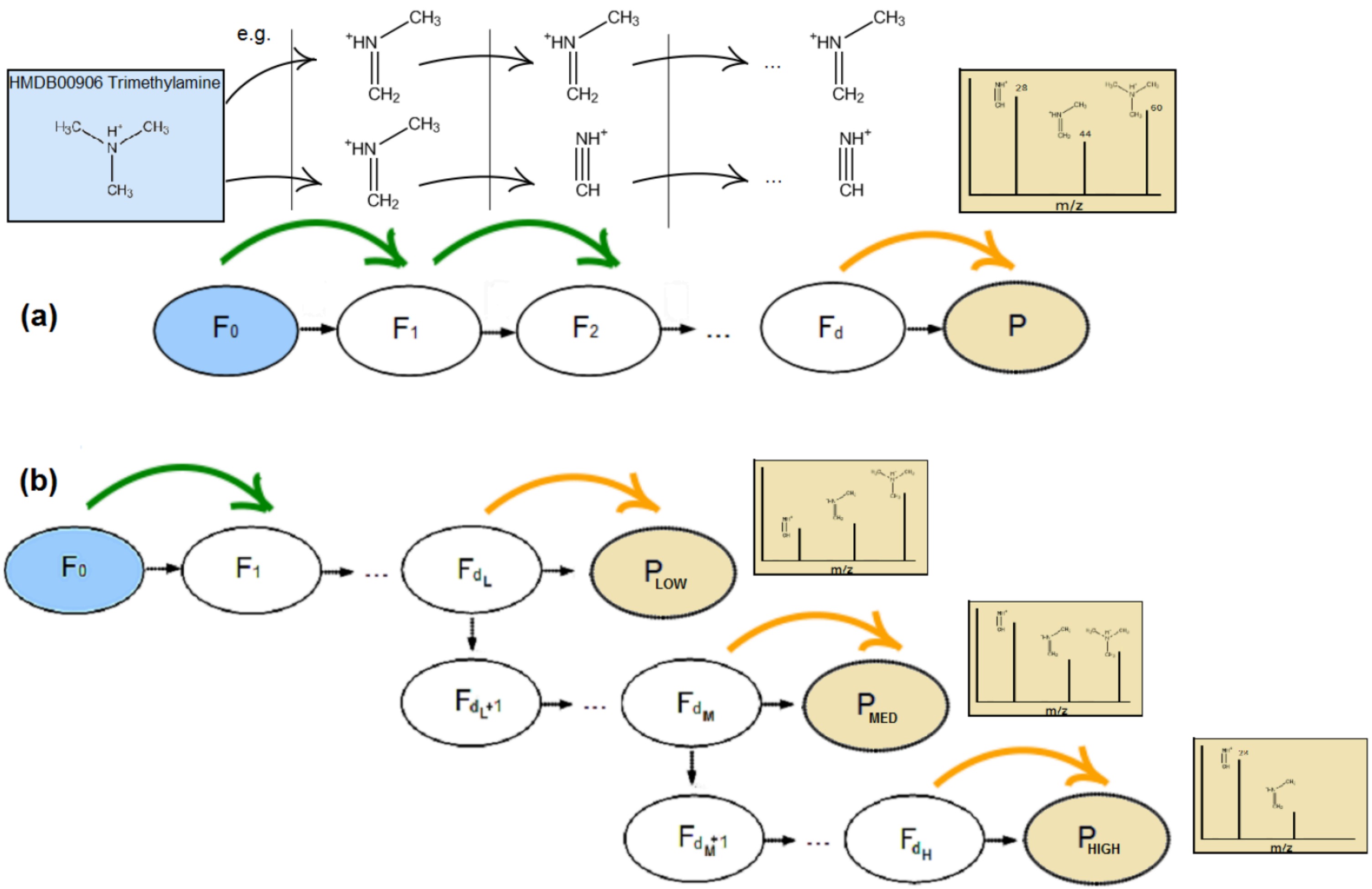


Predicted MS/MS, NMR, GC-MS
Spectra of knowns + biotransformed



Match observed spectra
to predicted spectra to ID compounds

Competitive Fragment Modeling (CFM-ID)

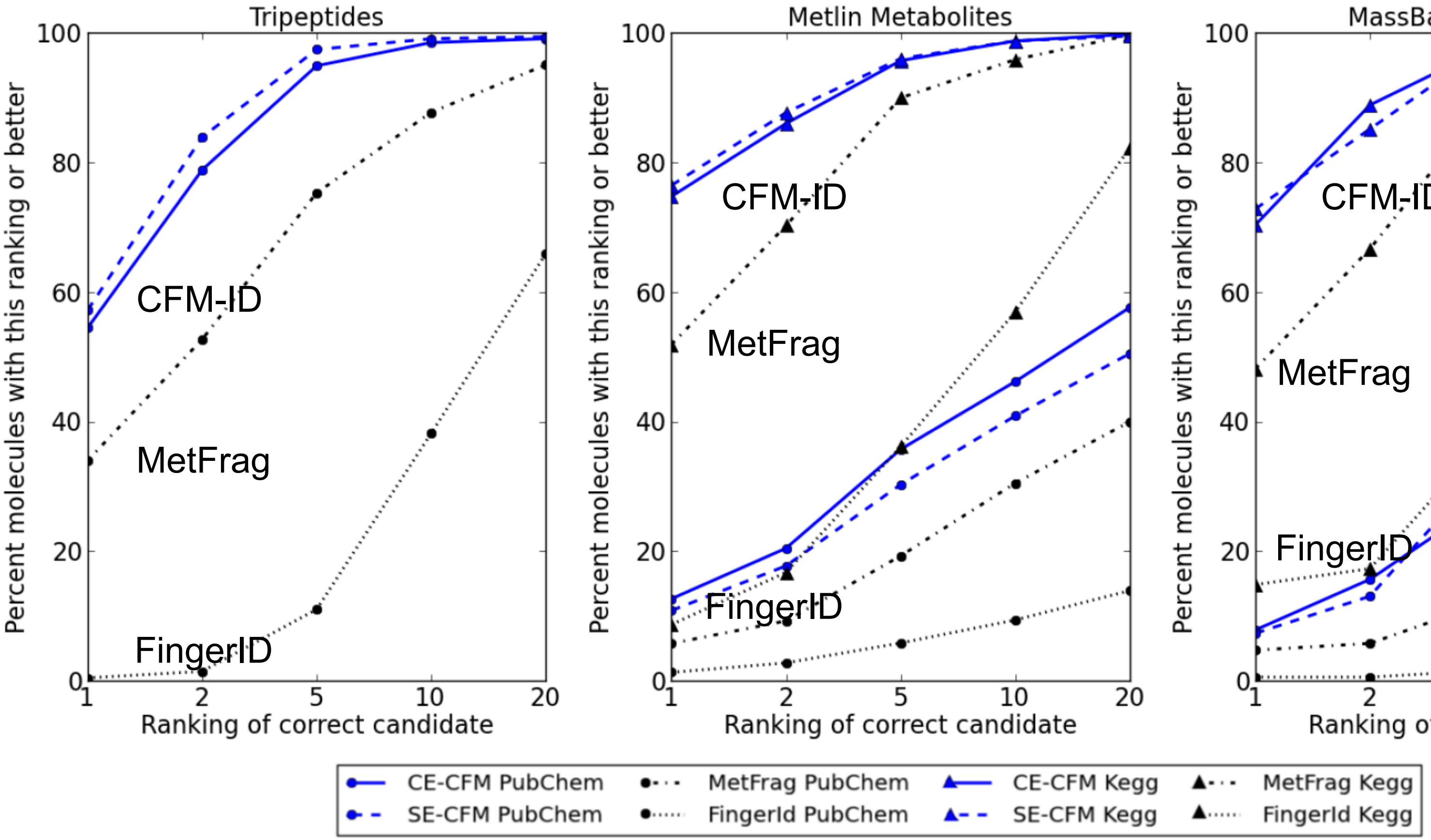


F. Allen et al. Metabolomics 11: 98-110 (2015).

CFM-ID

- Uses a large training set of high resolution MS/MS data of known compounds at low (10 eV) medium (20 eV) and high (40 eV) collision energies
- Employs an initially naïve chemical fragmenter that generates potential fragments and the coresponding MS/MS spectra
- The fragmenter slowly learns from its training data (via an HMM)
- *The more training data, the better the overall performance*

CFM-ID Performance

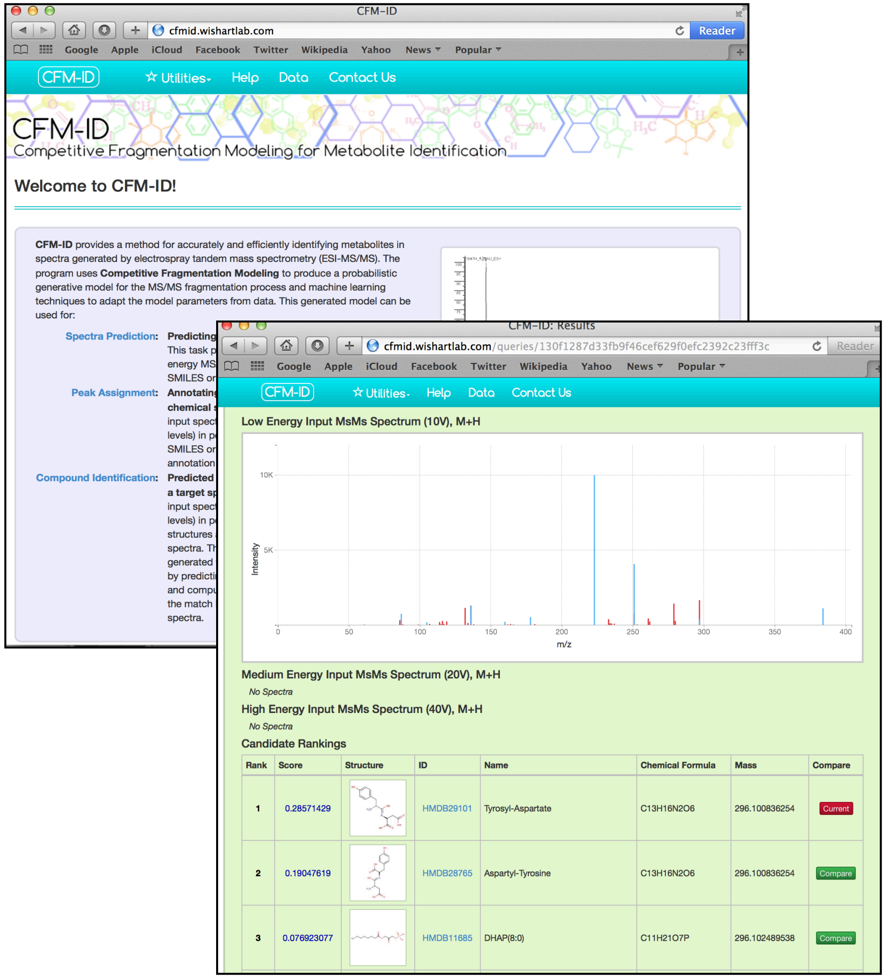


F. Allen et al. Metabolomics 11: 98-110 (2015).

Performance

- Significant performance improvement in CFM-ID and all other fragment or structure predictors if the database being searched is smaller or more targeted
- Significant improvement if multiple collision energies (10, 20, 40 eV) are used rather than a single collision energy
- 80% correct for DB ~30,000 cmpds
- 50% correct for DB ~1,000,000 cmpds
- 20% correct for DB ~50,000,000 cmpds

The CFM-ID Server



CFM-ID

Competitive Fragmentation Modeling for Metabolite Identification

Welcome to CFM-ID!

CFM-ID provides a method for accurately and efficiently identifying metabolites in spectra generated by electrospray tandem mass spectrometry (ESI-MS/MS). The program uses **Competitive Fragmentation Modeling** to produce a probabilistic generative model for the MS/MS fragmentation process and machine learning techniques to adapt the model parameters from data. This generated model can be used for:

Spectra Prediction: **Predicting**
This task p
energy MS
SMILES or

Peak Assignment: **Annotating**
chemical s
input spect
levels) in p
SMILES or
annotation

Compound Identification: **Predicted**
a target sp
input spect
levels) in p
structures
spectra. Th
generated
by predicti
and compu
the match
spectra.

Low Energy Input MsMs Spectrum (10V), M+H

Medium Energy Input MsMs Spectrum (20V), M+H

No Spectra

High Energy Input MsMs Spectrum (40V), M+H

No Spectra

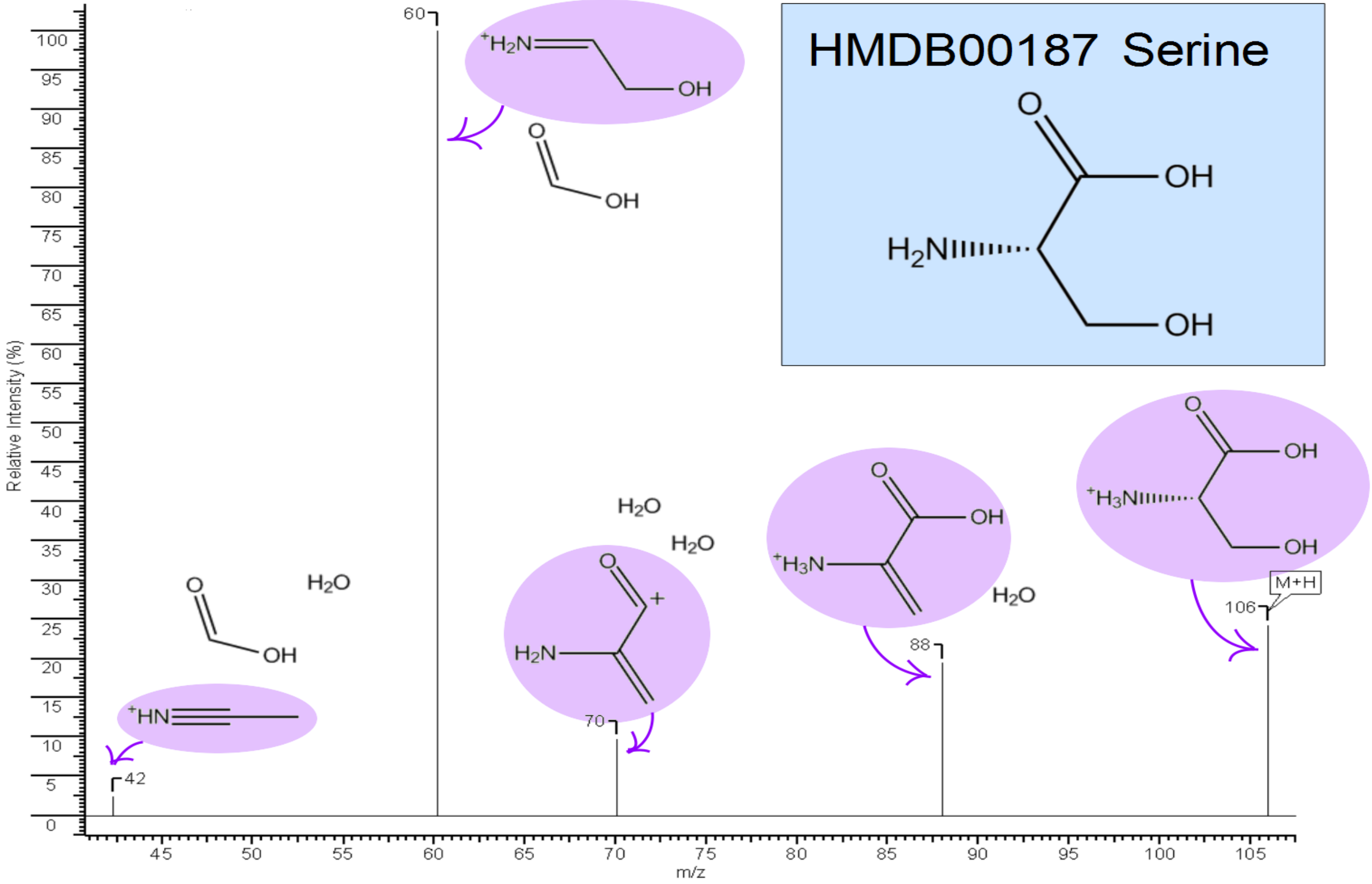
Candidate Rankings

Rank	Score	Structure	ID	Name	Chemical Formula	Mass	Compare
1	0.28571429		HMDB29101	Tyrosyl-Aspartate	C13H16N2O6	296.100836254	Current
2	0.19047619		HMDB28765	Aspartyl-Tyrosine	C13H16N2O6	296.100836254	Compare
3	0.076923077		HMDB11685	DHAP(8:0)	C11H21O7P	296.102489538	Compare

http://cfmid.wishartlab.com

- **A web server that predicts MS/MS spectra, annotates input MS/MS spectra and permits compound identification from input MS/MS spectra**
- **Matches predicted MS/MS spectra (from HMDB or KEGG) to input MS/MS spectra**
- **1st and 2nd in the 2014 CASMI competition, used by winners of 2016 CASMI competition**

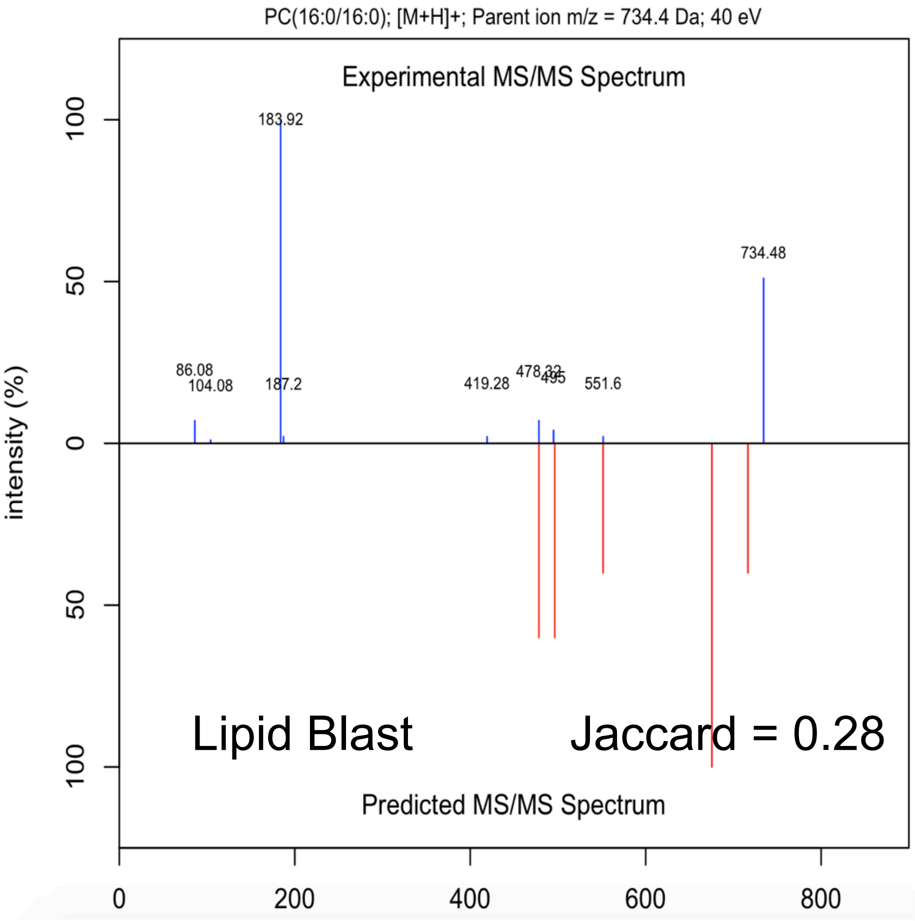
CFM-ID Example Output



CFM-ID Updates

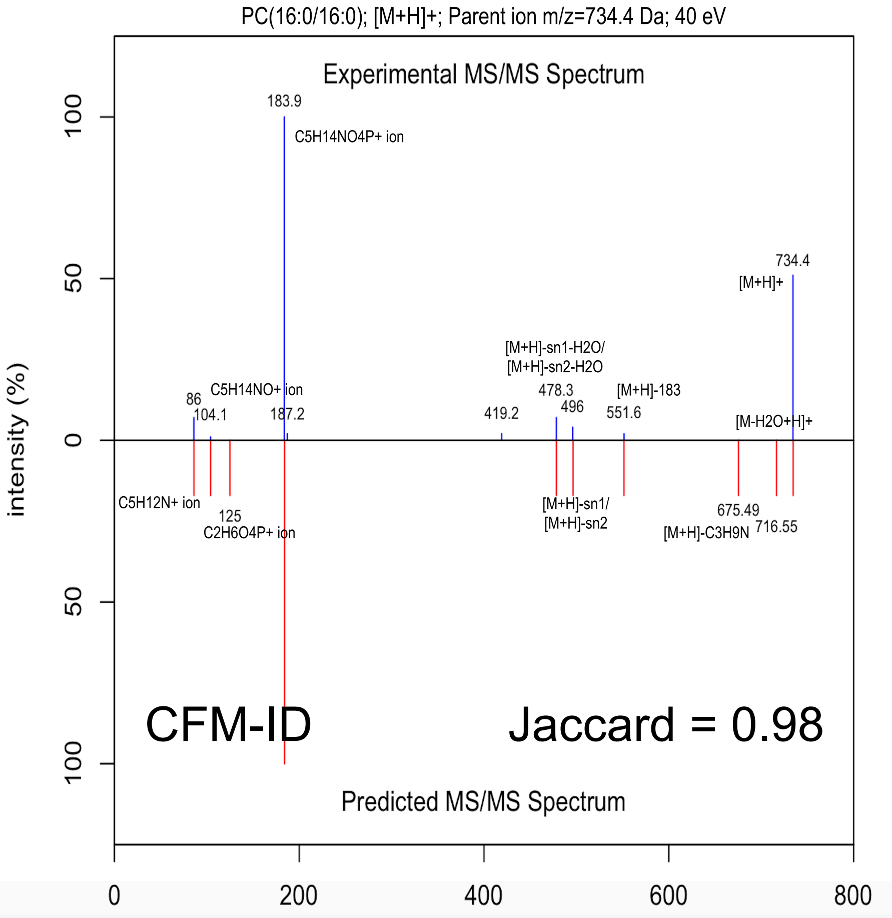
- **Version 3.0 completed (to be released in Dec):**
  - Significantly improved (5X) performance with respect to lipid MS/MS spectral prediction
  - Supports matches to known MS/MS spectra for better compound ID, improved scoring based on citations
- **Version 4.0 in progress (to be finished by 2019)**
  - Much larger training data set (4X larger) covering QTOF and OrbiTrap MS/MS spectra at multiple collision energies
  - Improved generative function, improved chemical and bond descriptors boosts spectral prediction performance by 30-40% over previous CFM-ID version
  - Combined improvements should increase overall performance by at least 50% (still not perfect)

CFM-ID 3.0 for Predicting 70,000 Lipid Spectra in the HMDB



Lipid Blast

Jaccard = 0.28



CFM-ID

Jaccard = 0.98

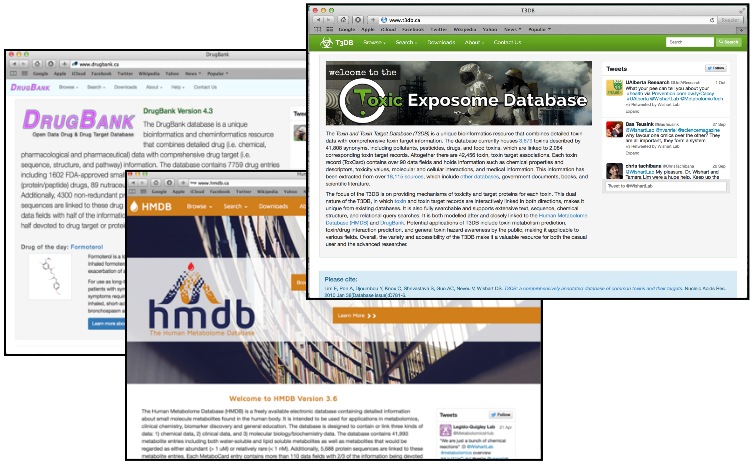
“Observables” Prediction

- MS/MS spectral prediction alone will not be sufficient to ID all unknown compounds
- Other observables need to be included for confirmation such as RT (retention time), RI (retention index), CCS (collisional cross section), and gas phase IR or IR ion spectra

Martens et al., *J. Inherit Metabol. Dis.* 2018 41(3):367-377

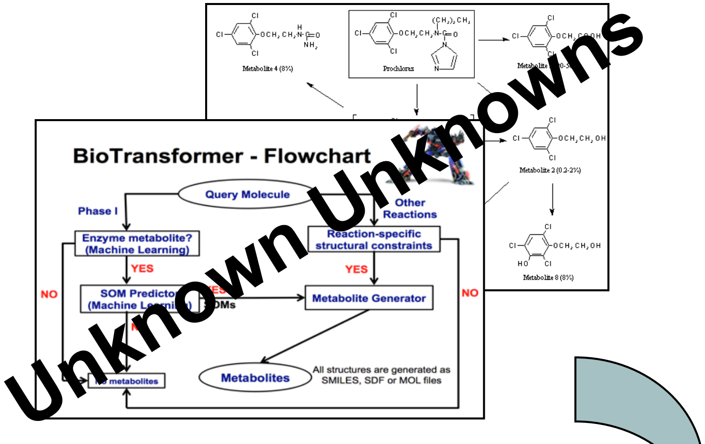
In Silico Metabolomics

HMDB/DrugBank/T3DB



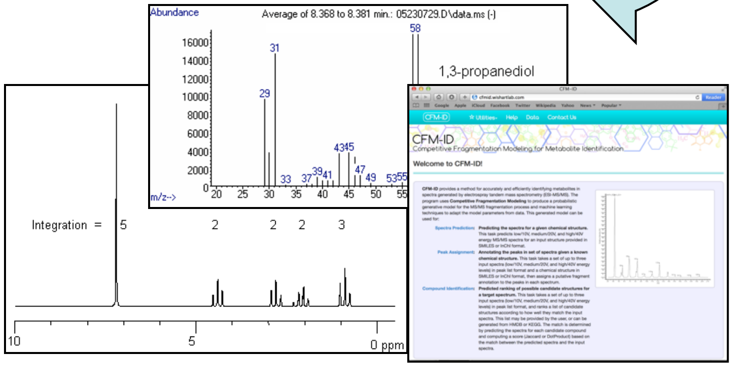
Known compounds (250,000)

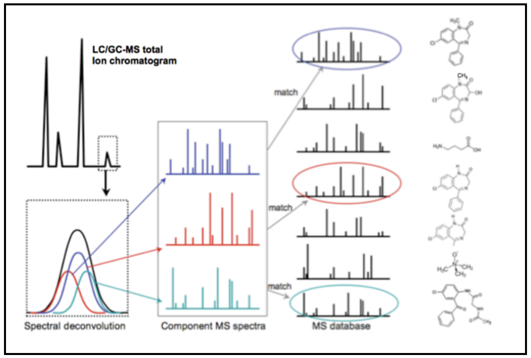
BioTransformer


BioTransformer - Flowchart
Query Molecule
Phase I
Other Reactions
Enzyme metabolite? (Machine Learning)
Reaction-specific structural constraints
YES
YES
NO
NO
SOM Predictor (Machine Learning)
Metabolite Generator
metabolites
Metabolites
All structures are generated as SMILES, SDF or MOL files
Prochloraz
Metabolite 4 (8%)
Metabolite 2 (0.2-2%)
Metabolite 8 (8%)
Unknown Unknowns

Predicted biotransformations
(250,000 --> 5,000,000)

CFM-ID/NMRPred

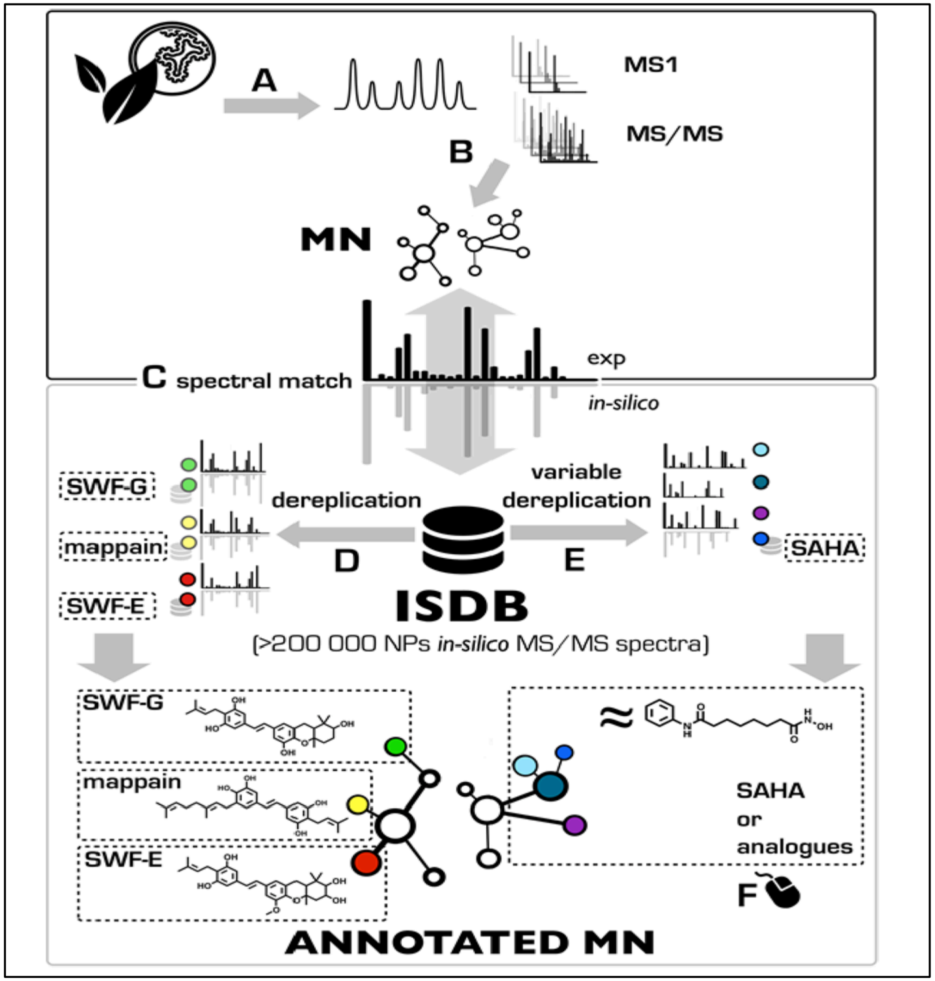


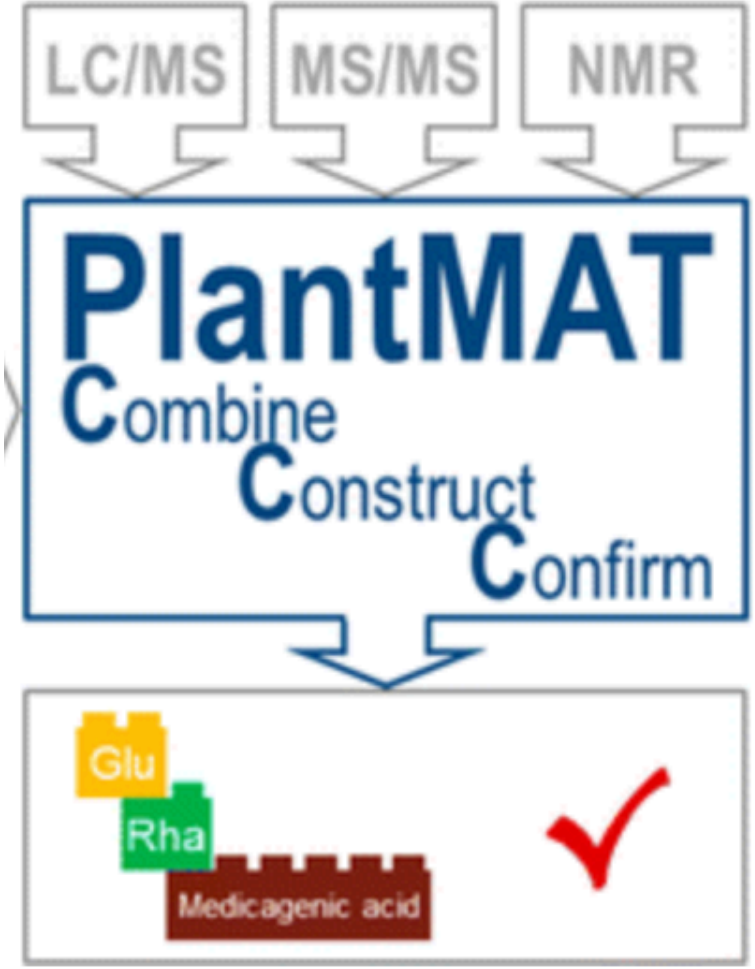


Match observed spectra
to predicted spectra to ID compounds

Predicted MS/MS, NMR, GC-MS
Spectra of knowns + biotransformed

Who Is Doing This?





PM Allard et al. *Anal. Chem*. 2016 88(6) 3317-3323

F Qiu et al. *Anal. Chem*. 2016 88(23) 11373-11383

Conclusions

- Compound identification of unknown compounds by MS/MS analysis is hard
- We have insufficient MS/MS and compound resources now and the foreseeable future
- *In silico* methods offer a possible solution to the problem of inadequate spectral libraries and inadequate collections of compounds
- Predicted compound libraries and and predicted MS/MS spectra are still imperfect, but they are getting better every year
- These *in silico* methods are already being used by several groups in natural product analysis

Thanks To…

- Yannick Djoumbou Feunang
- Ana Marcu
- AnChi Guo
- Kevin Liang
- Rosa Vazquez-Fresno
- Tanvir Sajed
- Daniel Johnson
- Carin Li
- Naama Karu
- Zinat Sayeeda
- Elvis Lo
- Nazanin Assempour
- Augustin Scalbert
- Sandeep Singhal
- David Arndt
- Yongjie Liang
- Hasan Badran
- Jason Grant
- Arnau Serra-Cayuela
- Yifeng Liu
- Rupasri Mandal
- Vaness Neveu
- Allison Pon
- Craig Knox
- Mike Wilson
- Claudine Manach

CIHR IRSC

GenomeAlberta

GenomeCanada

Canada Foundation for Innovation
Fondation canadienne pour l'innovation

NSERC
CRSNG

TMIC THE METABOLOMICS INNOVATION CENTRE