COMPUTATIONAL PROTEOMICS AND METABOLOMICS

Oliver Kohlbacher, Sven Nahnsen, Knut Reinert

11. Metabolite Identification

LEARNING UNIT 11A METABOLITES AND METABOLOMICS

- Differences between proteomics and metabolomics
- Targeted vs. nontargeted metabolomics

Metabolism

- **Metabolism** = sum of all the chemical processes occurring in an organism at one time
- Concerned with the management of material and energy resources within the cell
- Two types of metabolic processes
  - **Anabolic processes** – processes constructing larger molecules from smaller units (building up)
  - **Catabolic processes** – processes breaking down larger units (degradation or energy generation)
- Metabolites are both educts and products of metabolic processes
- Enzymes (proteins) usually catalyze these metabolic processes (reactions)
- A sequence of several coupled metabolic processes is called a **metabolic pathway**

Metabolomics vs. Proteomics

Proteomics	Metabolomics

FGGTSVANAER

VADILESNAR

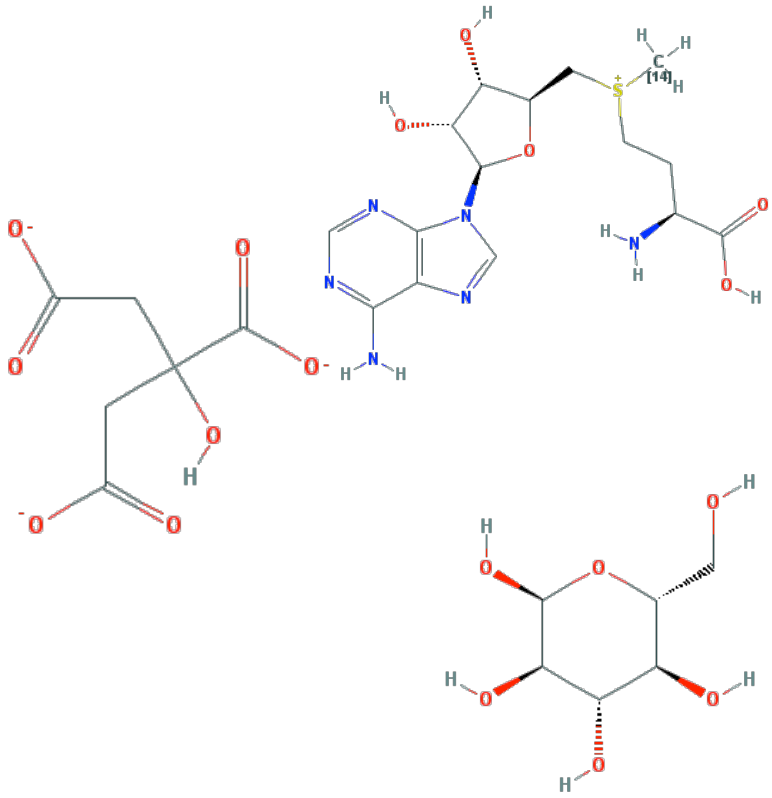
MSIAIMAGVLEAR

TISGQDALPNISDAER

TFVDQEFAQIK

ITNHLVAMIEK

QGQVATVLSAPAK



4

Metabolome vs. Proteome

- **Size and complexity** of the metabolome still largely unknown
- Similar to protein sequence databases, there are also **metabolite databases** listing all known metabolites (usually contains **tens of thousands** of metabolites)
- Differences between **proteome and metabolome**
  - Metabolites belong to wider ranger of chemical compound classes (lipids, sugars, amino acids)
  - Proteins have a more homogenous chemistry (20 proteinogenic amino acids)
  - Metabolites can have complex structures that require a structural formula for a comprehensive description
  - Proteins have a simple, linear structure that can be represented by a sequence
  - Metabolites are **light**: average metabolite mass a 100-300 Da
  - Proteins are **heavy**: median protein length around 300-500 aa, about 40,000 Da molecular weight

Metabolomics Techniques

- Fundamentally two types of approaches
  - **Targeted metabolomics**
    - Identify only a well-defined subset of metabolites, but those with higher accuracy (hundreds?)
    - All of these metabolites can then be identified
  - **Non-targeted metabolomics (metabolic profiling)**
    - Try to see as much of the metabolome as possible (thousands and more)
    - Majority of metabolites can be seen
    - Only a small fraction will be identified
- Similarly, there is also targeted and non-targeted proteomics
- In proteomics, the identification problem is less difficult, though, which is why this distinction is more relevant in metabolomics (where identification is much harder)

Metabolomics vs. Proteomics

- We will restrict ourselves to LC-MS(/MS)-based metabolomics in this course
- LC-MS can be used for quantification of metabolites in the same fashion as for peptides/proteins
- Labeled and unlabeled quantification approaches can be used
- Identification differs significantly
  - Metabolites (in the general case) are not linear polymers, thus there is no sequence
  - Their chemistry is much more diverse and their fragmentation behavior as well

    no search against sequence database, no ion series

Metabolomics vs. Proteomics

- Quantification
  - Wide range of chemical structures does not permit a simple in vitro labeling for arbitrary metabolites (as, for example, iTRAQ)
  - Labeling has to be done in vivo (metabolic labeling)
  - Feature detection is complicated
    - Frequent occurrence of S, Cl, Br leads to complex isotope patterns
    - Averagine hypothesis does not hold

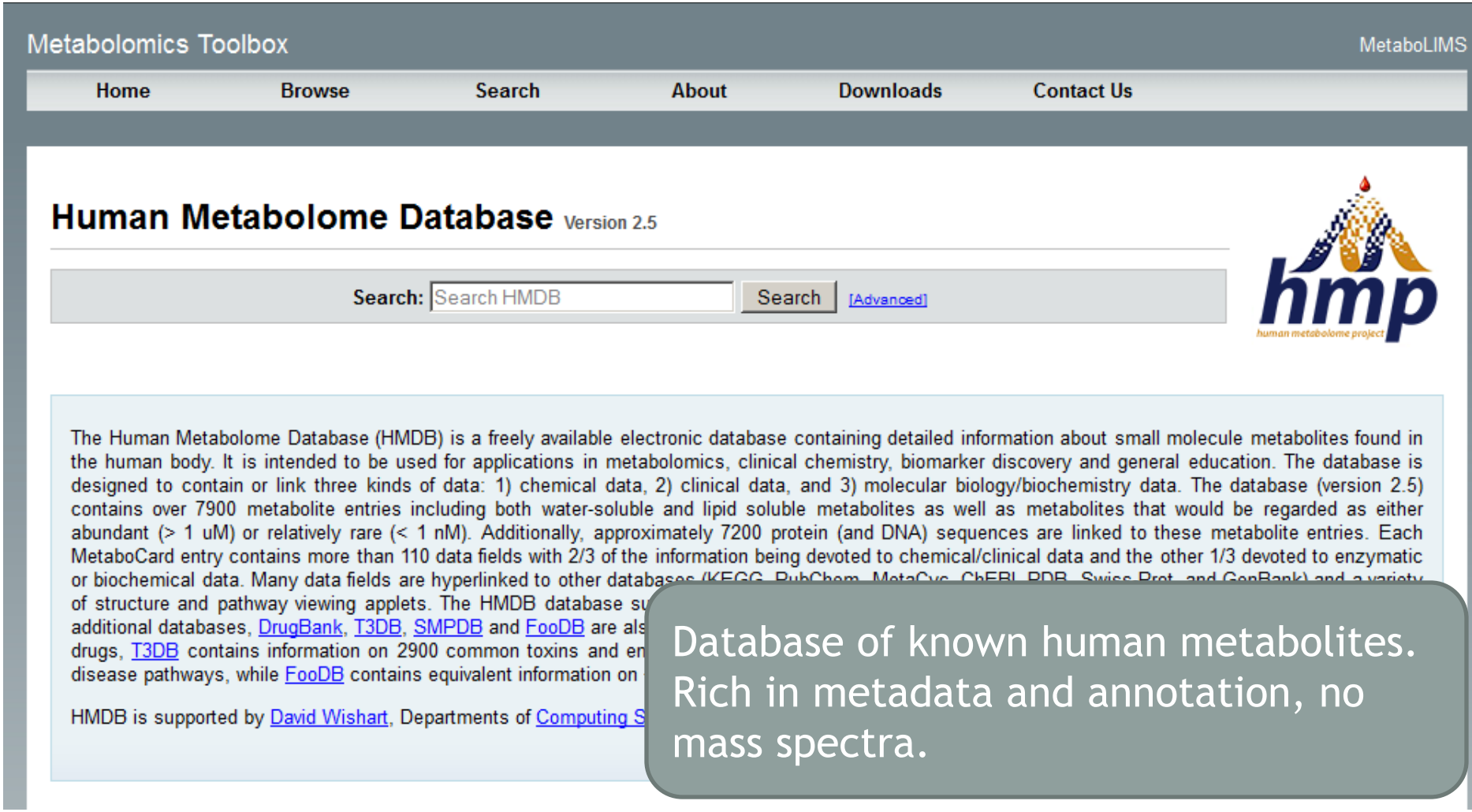
LEARNING UNIT 11B
METABOLITE DATABASES

- Databases for mass spectra
- Databases for compounds
- Databases for pathways
- Comparison of databases

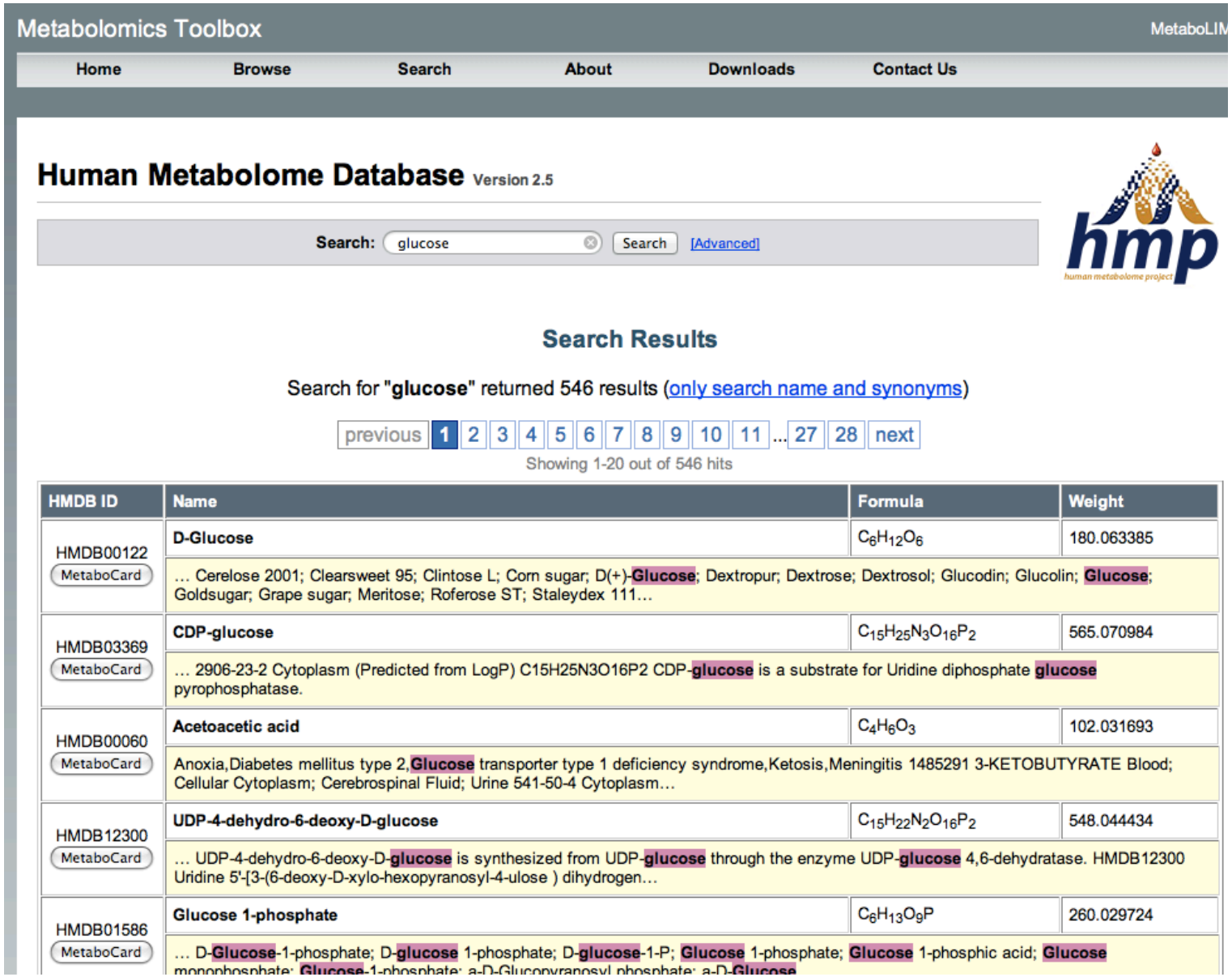
Databases

- Metabolomics has to rely on different types of databases
  - **Structural databases** – contain structures of metabolites
  - **Pathway databases** – contain metabolic networks
  - **Spectrum databases** – contain information on mass spectra
- In contrast to sequencing data, many of the databases are commercial
- Freely accessible spectrum databases for metabolites are still rare

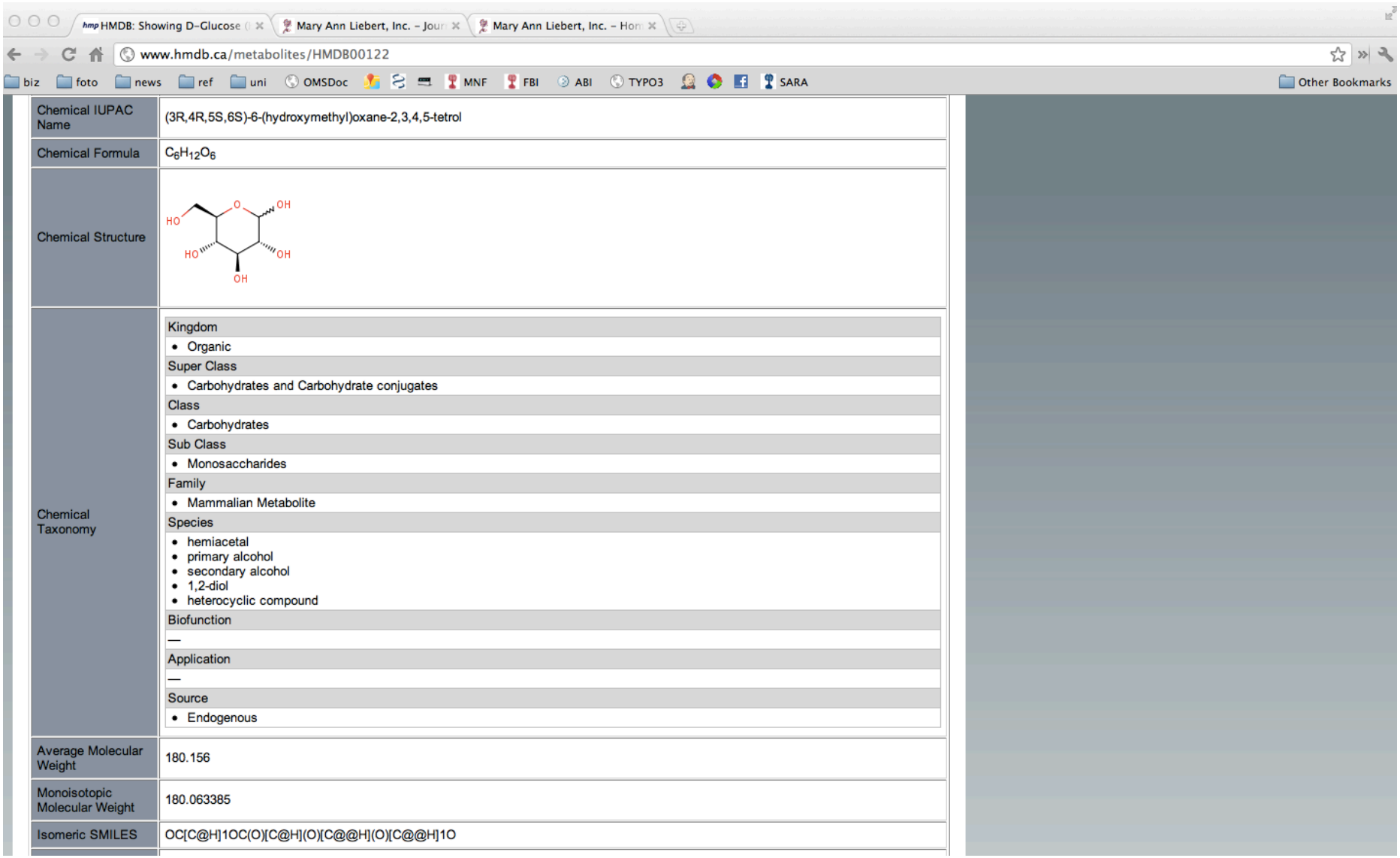
HMDB


Metabolomics Toolbox
MetaboLIMS
Home
Browse
Search
About
Downloads
Contact Us
Human Metabolome Database Version 2.5
Search: Search HMDB
Search
[Advanced]
hmp
human metabolome project
The Human Metabolome Database (HMDB) is a freely available electronic database containing detailed information about small molecule metabolites found in the human body. It is intended to be used for applications in metabolomics, clinical chemistry, biomarker discovery and general education. The database is designed to contain or link three kinds of data: 1) chemical data, 2) clinical data, and 3) molecular biology/biochemistry data. The database (version 2.5) contains over 7900 metabolite entries including both water-soluble and lipid soluble metabolites as well as metabolites that would be regarded as either abundant (> 1 uM) or relatively rare (< 1 nM). Additionally, approximately 7200 protein (and DNA) sequences are linked to these metabolite entries. Each MetaboCard entry contains more than 110 data fields with 2/3 of the information being devoted to chemical/clinical data and the other 1/3 devoted to enzymatic or biochemical data. Many data fields are hyperlinked to other databases
of structure and pathway viewing applets. The HMDB database su
additional databases, DrugBank, T3DB, SMPDB and FooDB are als
drugs, T3DB contains information on 2900 common toxins and en
disease pathways, while FooDB contains equivalent information on
HMDB is supported by David Wishart, Departments of Computing S
Database of known human metabolites. Rich in metadata and annotation, no mass spectra.

Human Metabolome Database

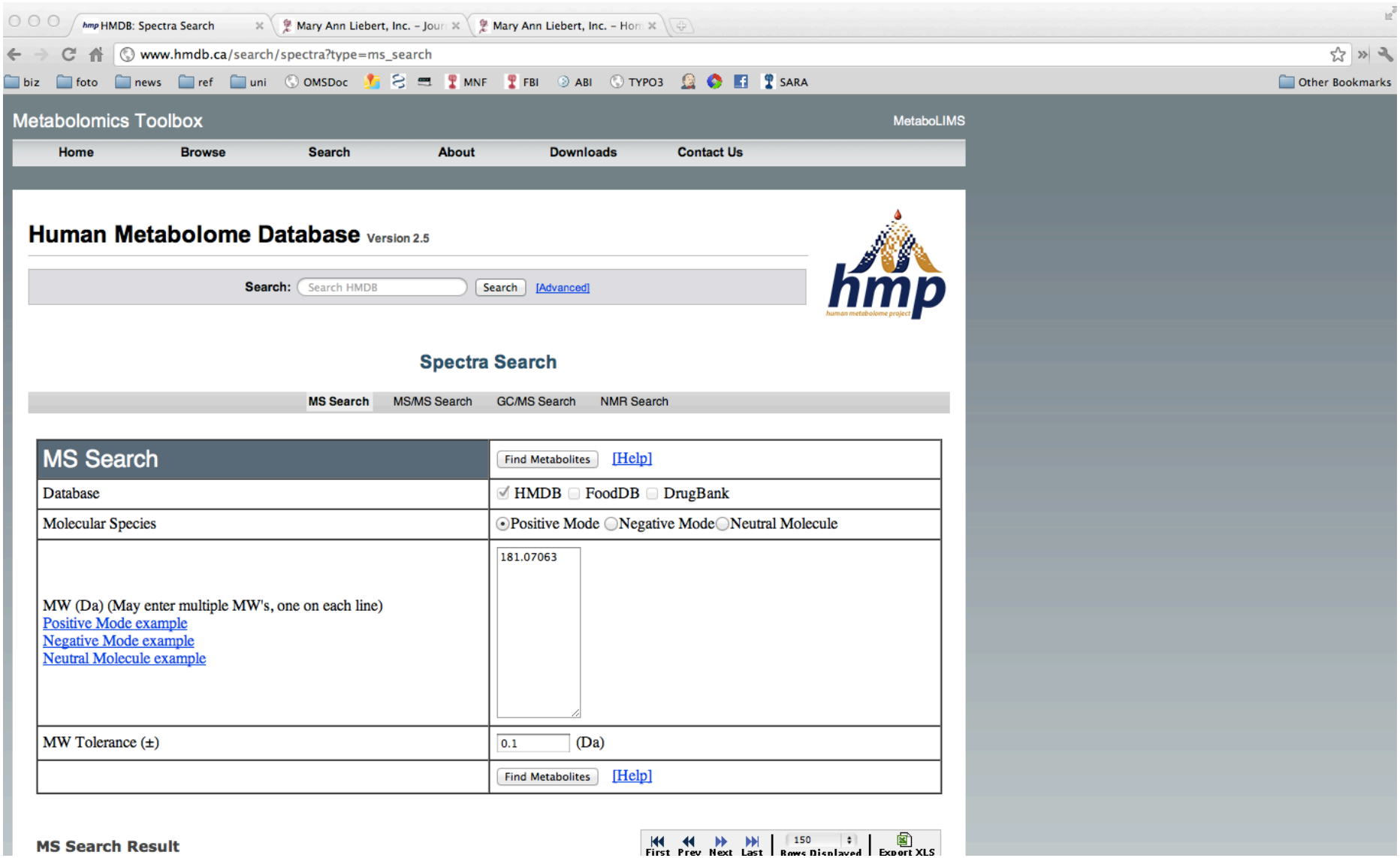


Human Metabolome Database

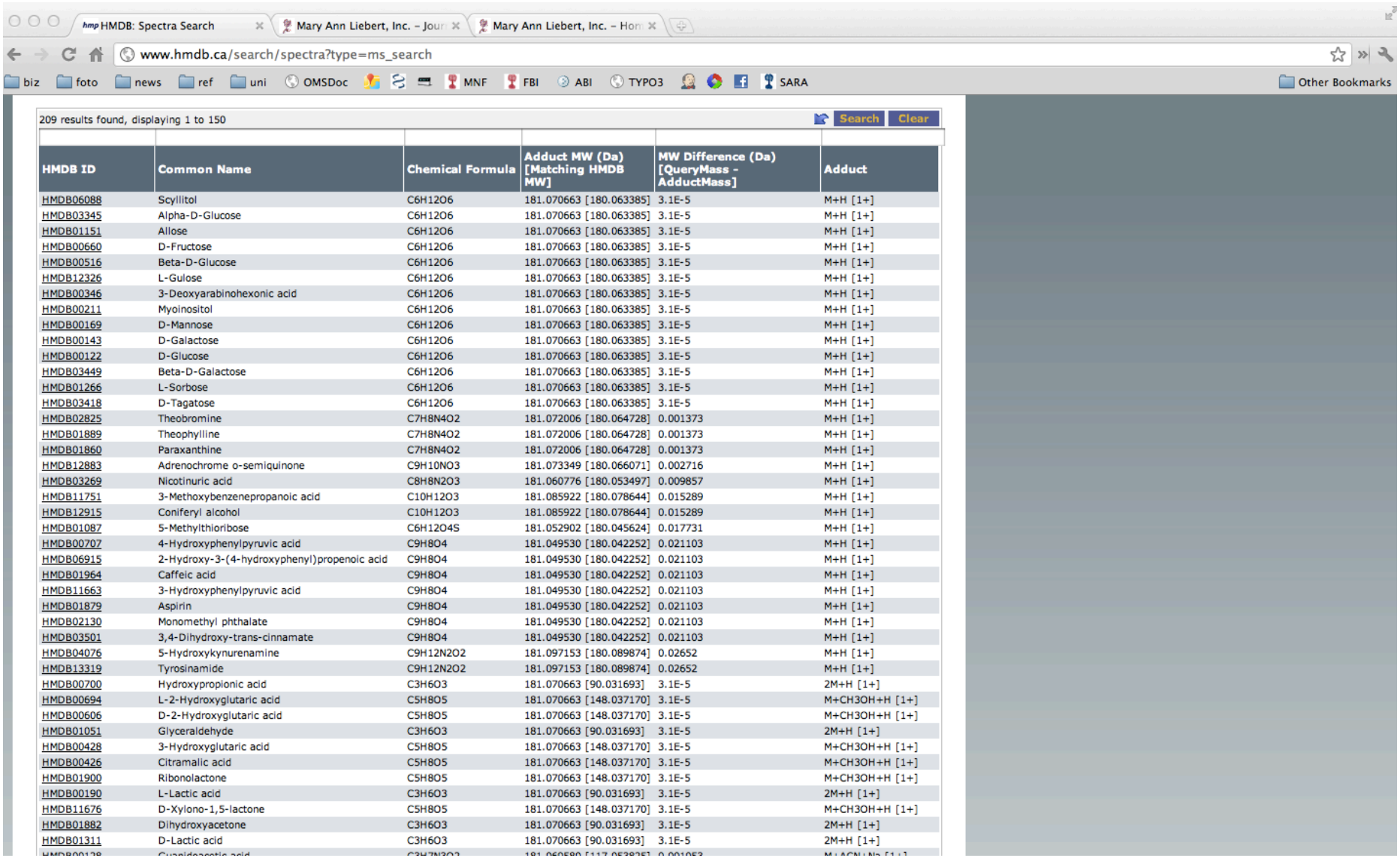


Chemical IUPAC Name	(3R,4R,5S,6S)-6-(hydroxymethyl)oxane-2,3,4,5-tetrol
Chemical Formula	$C_6H_{12}O_6$
Chemical Structure	
Chemical Taxonomy	Kingdom • Organic Super Class • Carbohydrates and Carbohydrate conjugates Class • Carbohydrates Sub Class • Monosaccharides Family • Mammalian Metabolite Species • hemiacetal • primary alcohol • secondary alcohol • 1,2-diol • heterocyclic compound Biofunction — Application — Source • Endogenous
Average Molecular Weight	180.156
Monoisotopic Molecular Weight	180.063385
Isomeric SMILES	OC[C@H]1OC(O)[C@H](O)[C@@H](O)[C@@H]1O

Human Metabolome Database



Human Metabolome Database

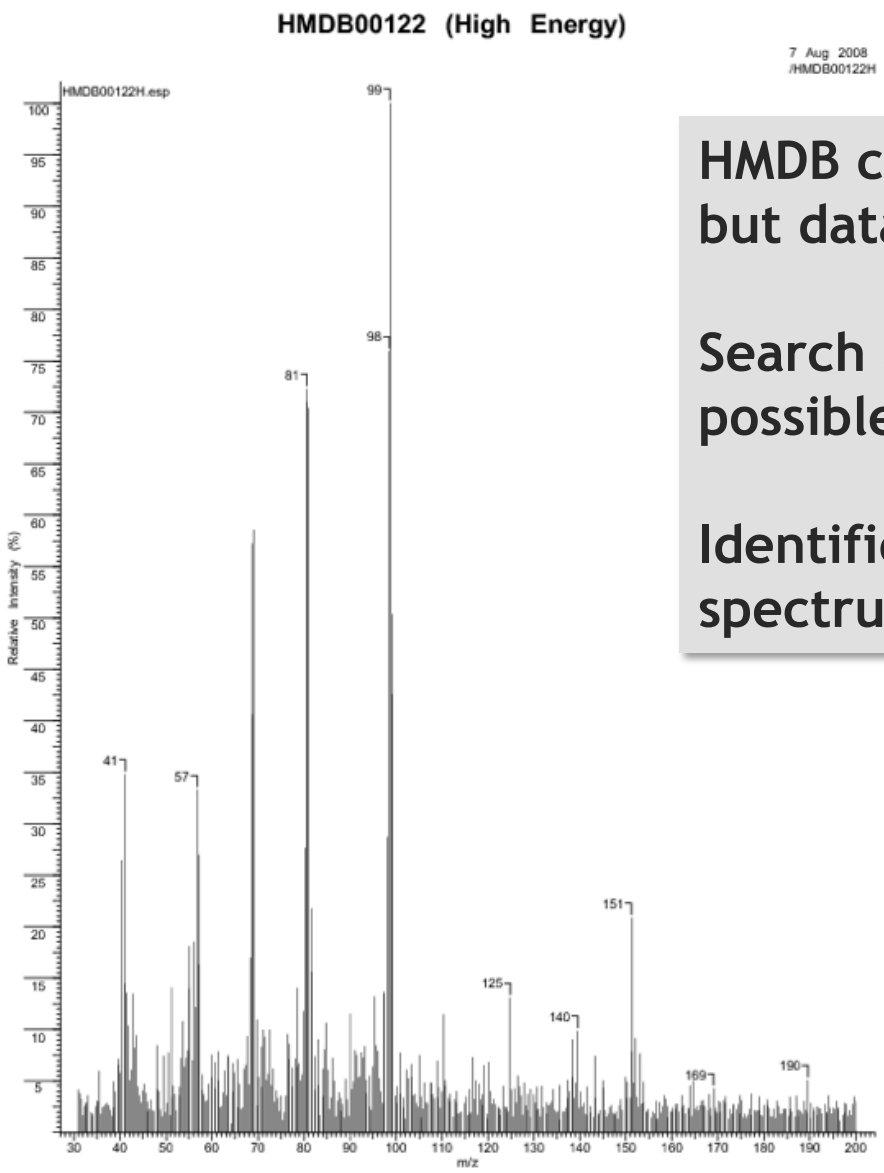


HMDB ID	Common Name	Chemical Formula	Adduct MW (Da) [Matching HMDB MW]	MW Difference (Da) [QueryMass - AdductMass]	Adduct
HMDB06088	Scyllitol	C6H12O6	181.070663 [180.063385]	3.1E-5	M+H [1+]
HMDB03345	Alpha-D-Glucose	C6H12O6	181.070663 [180.063385]	3.1E-5	M+H [1+]
HMDB01151	Allose	C6H12O6	181.070663 [180.063385]	3.1E-5	M+H [1+]
HMDB00660	D-Fructose	C6H12O6	181.070663 [180.063385]	3.1E-5	M+H [1+]
HMDB00516	Beta-D-Glucose	C6H12O6	181.070663 [180.063385]	3.1E-5	M+H [1+]
HMDB12326	L-Gulose	C6H12O6	181.070663 [180.063385]	3.1E-5	M+H [1+]
HMDB00346	3-Deoxyarabinohexonic acid	C6H12O6	181.070663 [180.063385]	3.1E-5	M+H [1+]
HMDB00211	Myoinositol	C6H12O6	181.070663 [180.063385]	3.1E-5	M+H [1+]
HMDB00169	D-Mannose	C6H12O6	181.070663 [180.063385]	3.1E-5	M+H [1+]
HMDB00143	D-Galactose	C6H12O6	181.070663 [180.063385]	3.1E-5	M+H [1+]
HMDB00122	D-Glucose	C6H12O6	181.070663 [180.063385]	3.1E-5	M+H [1+]
HMDB03449	Beta-D-Galactose	C6H12O6	181.070663 [180.063385]	3.1E-5	M+H [1+]
HMDB01266	L-Sorbose	C6H12O6	181.070663 [180.063385]	3.1E-5	M+H [1+]
HMDB03418	D-Tagatose	C6H12O6	181.070663 [180.063385]	3.1E-5	M+H [1+]
HMDB02825	Theobromine	C7H8N4O2	181.072006 [180.064728]	0.001373	M+H [1+]
HMDB01889	Theophylline	C7H8N4O2	181.072006 [180.064728]	0.001373	M+H [1+]
HMDB01860	Paraxanthine	C7H8N4O2	181.072006 [180.064728]	0.001373	M+H [1+]
HMDB12883	Adrenochrome o-semiquinone	C9H10NO3	181.073349 [180.066071]	0.002716	M+H [1+]
HMDB03269	Nicotinuric acid	C8H8N2O3	181.060776 [180.053497]	0.009857	M+H [1+]
HMDB11751	3-Methoxybenzenepropanoic acid	C10H12O3	181.085922 [180.078644]	0.015289	M+H [1+]
HMDB12915	Coniferyl alcohol	C10H12O3	181.085922 [180.078644]	0.015289	M+H [1+]
HMDB01087	5-Methylthioribose	C6H12O4S	181.052902 [180.045624]	0.017731	M+H [1+]
HMDB00707	4-Hydroxyphenylpyruvic acid	C9H8O4	181.049530 [180.042252]	0.021103	M+H [1+]
HMDB06915	2-Hydroxy-3-(4-hydroxyphenyl)propenoic acid	C9H8O4	181.049530 [180.042252]	0.021103	M+H [1+]
HMDB01964	Caffeic acid	C9H8O4	181.049530 [180.042252]	0.021103	M+H [1+]
HMDB11663	3-Hydroxyphenylpyruvic acid	C9H8O4	181.049530 [180.042252]	0.021103	M+H [1+]
HMDB01879	Aspirin	C9H8O4	181.049530 [180.042252]	0.021103	M+H [1+]
HMDB02130	Monomethyl phthalate	C9H8O4	181.049530 [180.042252]	0.021103	M+H [1+]
HMDB03501	3,4-Dihydroxy-trans-cinnamate	C9H8O4	181.049530 [180.042252]	0.021103	M+H [1+]
HMDB04076	5-Hydroxykynurenamine	C9H12N2O2	181.097153 [180.089874]	0.02652	M+H [1+]
HMDB13319	Tyrosinamide	C9H12N2O2	181.097153 [180.089874]	0.02652	M+H [1+]
HMDB00700	Hydroxypropionic acid	C3H6O3	181.070663 [90.031693]	3.1E-5	2M+H [1+]
HMDB00694	L-2-Hydroxyglutaric acid	C5H8O5	181.070663 [148.037170]	3.1E-5	M+CH3OH+H [1+]
HMDB00606	D-2-Hydroxyglutaric acid	C5H8O5	181.070663 [148.037170]	3.1E-5	M+CH3OH+H [1+]
HMDB01051	Glyceraldehyde	C3H6O3	181.070663 [90.031693]	3.1E-5	2M+H [1+]
HMDB00428	3-Hydroxyglutaric acid	C5H8O5	181.070663 [148.037170]	3.1E-5	M+CH3OH+H [1+]
HMDB00426	Citramalic acid	C5H8O5	181.070663 [148.037170]	3.1E-5	M+CH3OH+H [1+]
HMDB01900	Ribonolactone	C5H8O5	181.070663 [148.037170]	3.1E-5	M+CH3OH+H [1+]
HMDB00190	L-Lactic acid	C3H6O3	181.070663 [90.031693]	3.1E-5	2M+H [1+]
HMDB11676	D-Xylono-1,5-lactone	C5H8O5	181.070663 [148.037170]	3.1E-5	M+CH3OH+H [1+]
HMDB01882	Dihydroxyacetone	C3H6O3	181.070663 [90.031693]	3.1E-5	2M+H [1+]
HMDB01311	D-Lactic acid	C3H6O3	181.070663 [90.031693]	3.1E-5	2M+H [1+]

Human Metabolome Database

Material Safety Data Sheet (MSDS)	• Show PDF/HTML
MOL File	Show
SDF File	Show
PDB File	Show
2D Structure	View 2D Structure
3D Structure	View 3D Structure
Experimental PDB ID	1A47
Experimental PDB File	Show
Experimental PDB Structure	View 3D Structure
Experimental ^{1}H NMR Spectrum	Download Spectrum Download FID (Varian) Show Experimental Conditions
Experimental ^{13}C NMR Spectrum	Download Spectrum Download FID (Bruker) Show Experimental Conditions
Experimental ^{13}C HSQC Spectrum	Download Spectrum Download FID (Bruker) Show Experimental Conditions
Predicted ^{1}H NMR Spectrum	Show Image Show Peaklist
Predicted ^{13}C NMR Spectrum	Show Image Show Peaklist
Mass Spectrum	Low Energy Download File Show Experimental Conditions Medium Energy Download File Show Experimental Conditions High Energy Download File Show Experimental Conditions

Human Metabolome Database



HMDB contains mass spectra of metabolites, but data is not downloadable in bulk.

Search against the spectra database is only possible through the web interface.

Identification by accurate mass or MS/MS spectrum comparison.

Human Metabolome Database

Spectra Search

MS Search | **MS/MS Search** | GC/MS Search | NMR Search

NOTES:
To query the database using spectral pattern matching, upload the MS/MS data file for the metabolite OR paste its content in the textarea box below.
** Fields are mandatory

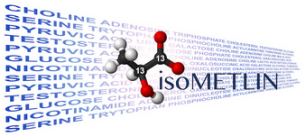
MS/MS Search	Find Metabolites [Help]
Search By	MS/MS Peaklist Data
**MW of Parent Ion	174 (Da)
**MW Tolerance (±)	0.1 (Da)
Instrument Type	Triple_Quad
**Fragment Ion Tolerance (±)	0.5 (Da)
CID Energy Level	Low Energy
Ionization Mode	Positive
MS/MS Data File	Choose File No file chosen
OR	
Content of MS/MS Data File Aconitic Acid (Low Energy, Positive Ion Mode) example Xanthine (Medium Energy, Positive Ion Mode) example Hippuric Acid (Medium Energy, Negative Ion Mode) example Citric Acid (Low Energy, Negative Ion Mode) example	41.400 9.926 85.030 100.000 111.118 24.412 128.847 30.000 172.851 11.912 **NOTES:** 1. m/z (Da) and relative intensities, RInt (%), delimited by a space (" "). 2. m/z and RI MUST contain a decimal. 3. m/z MUST be less than m/z of the parent ion minus 10 Da.
	Find Metabolites [Help]

METLIN

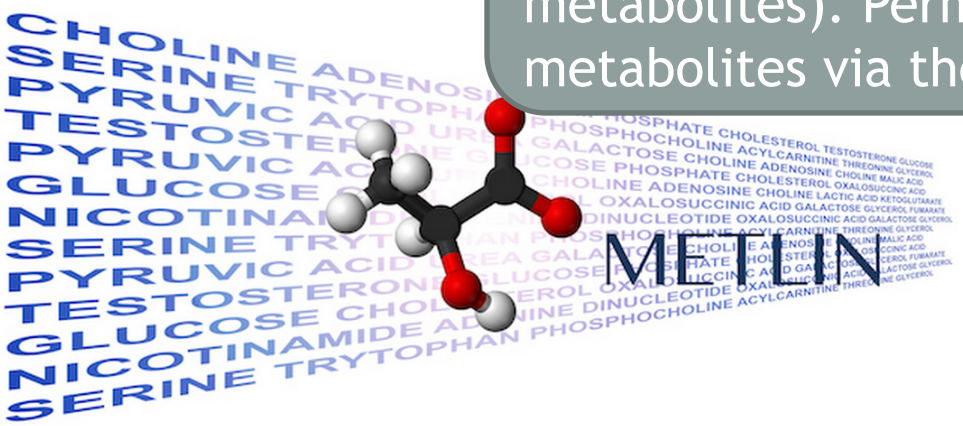
Scripps Center For Metabolomics

METLIN: Metabolite and Tandem MS Database

MS HOME | Overview | Search | XCMSOnline | Software/Services | Metabolomics Science | Publications



Database containing a large number of metabolites (240,000+) and spectra for those (12,000 metabolites). Permits search of metabolites via their mass spectra.



Statistics

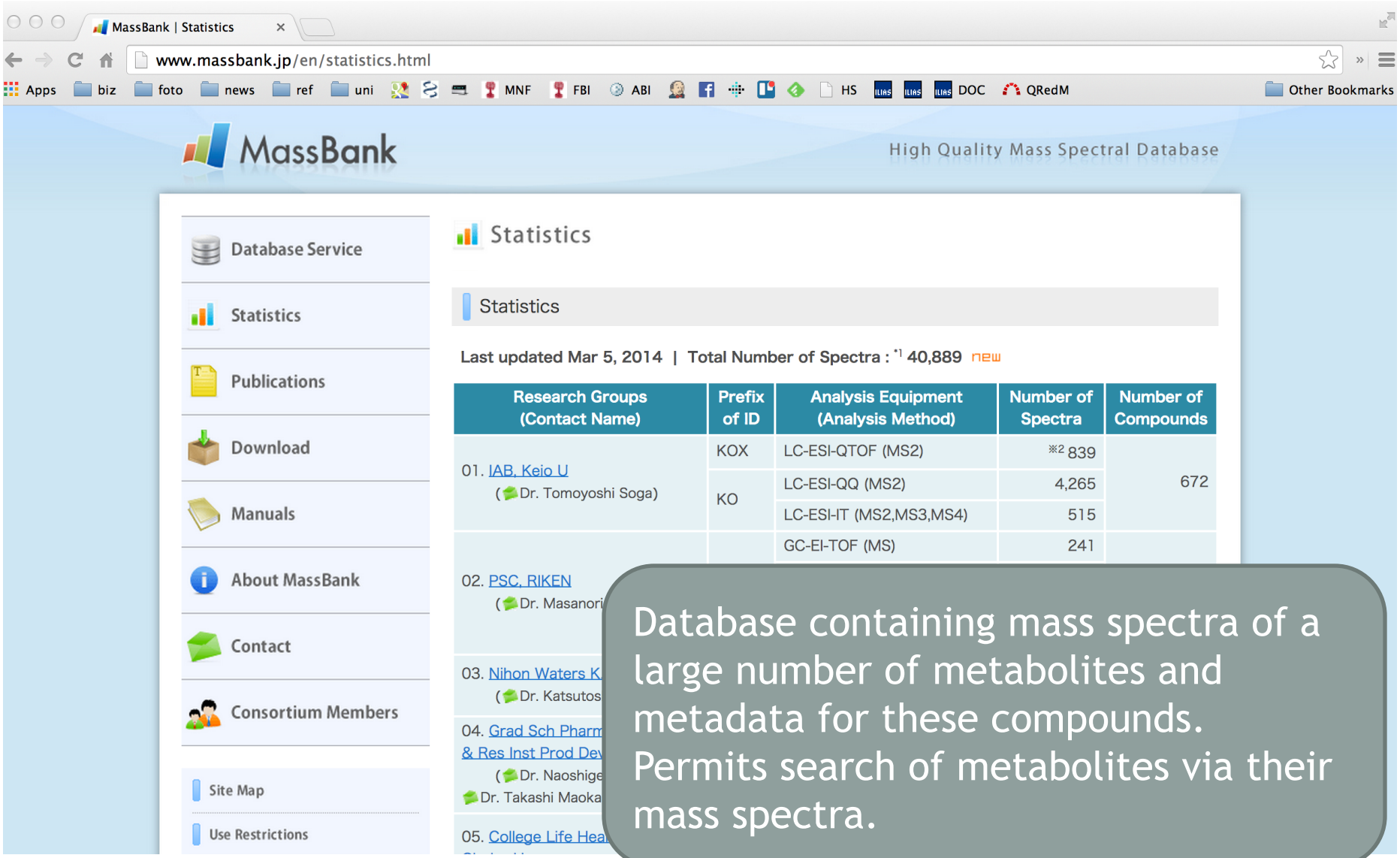
- # Metabolites: 240,516
- # High Resolution MS/MS Spectra: 61,872
- # Metabolites w/ High Resolution MS/MS: 12,057

Functionality

- **Single** & **Batch**
  Precursor Ion *(m/z)* searching
- **Single** & **Multiple**
  Fragment Ion *(m/z)* searching

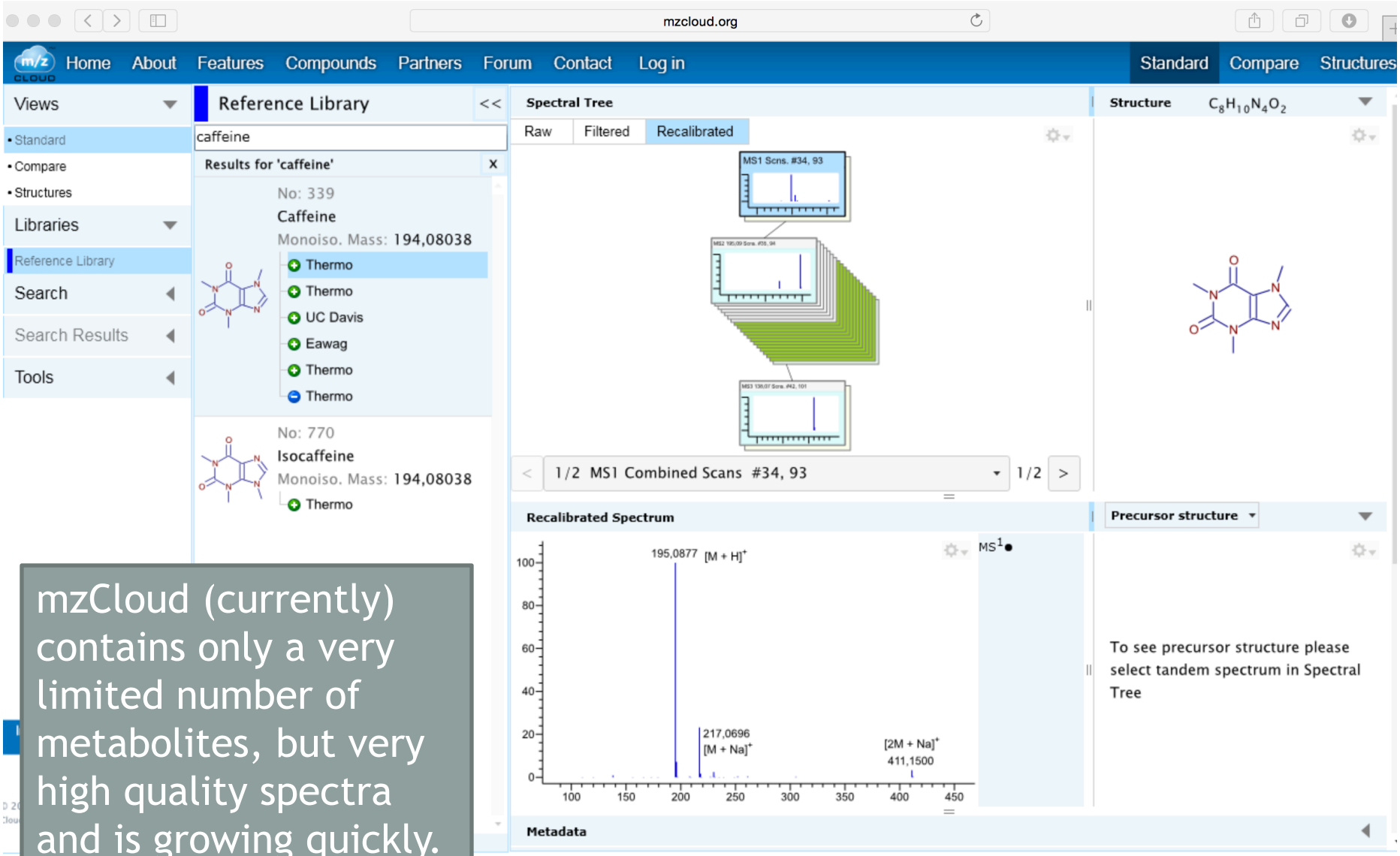
http://metlin.scripps.edu/

MassBank



http://www.massbank.jp/

mzCloud



mzCloud (currently) contains only a very limited number of metabolites, but very high quality spectra and is growing quickly.

KEGG

KEGG - Table of Contents

KEGG2 PATHWAY GENES LIGAND KO SSDB EXP

KEGG (Kyoto Encyclopedia of Genes and Genomes) contains structural information of metabolites as well as pathway information.

1. KEGG Databases

Category	Database		Search & Compute	DBGET Search	
Pathway information	**KEGG PATHWAY Database**	XML	Search objects in KEGG pathways Color objects in KEGG pathways	PATHWAY	
Genomic information	**KEGG GENES Database**	KO	Search orthologs or gene clusters in SSDB Search similar GENES sequences Search similar GENOME sequences	KO	
				GENES	
				GENOME	
Chemical information	**KEGG LIGAND Database**	RC	Search similar compound structures Search similar glycan structures Predict reactions and assign EC numbers Generate possible reaction paths	COMPOUND	LIGAND
				GLYCAN	
				REACTION	
				RPAIR	
				ENZYME	

22

http://www.genome.jp/kegg/

HumanCyc

HumanCyc

Encyclopedia of Human Genes and Metabolism

HumanCyc contains a very comprehensive set of metabolic pathways.

BioCyc Home

Search

Database Search
Advanced Database Search
Help

News

Nov 09 BioCyc 8.6 released
Sep 17 BioCyc 8.5 released
Sep 17 Online Licensing

Services

Software/Data Download
User Support
Subscribe to Mailing List
EcoCyc T-shirts

Information

Introduction to BioCyc
Guided Tour

HumanCyc: Encycloped
Genes and Metabolism

- Query the HumanCyc database

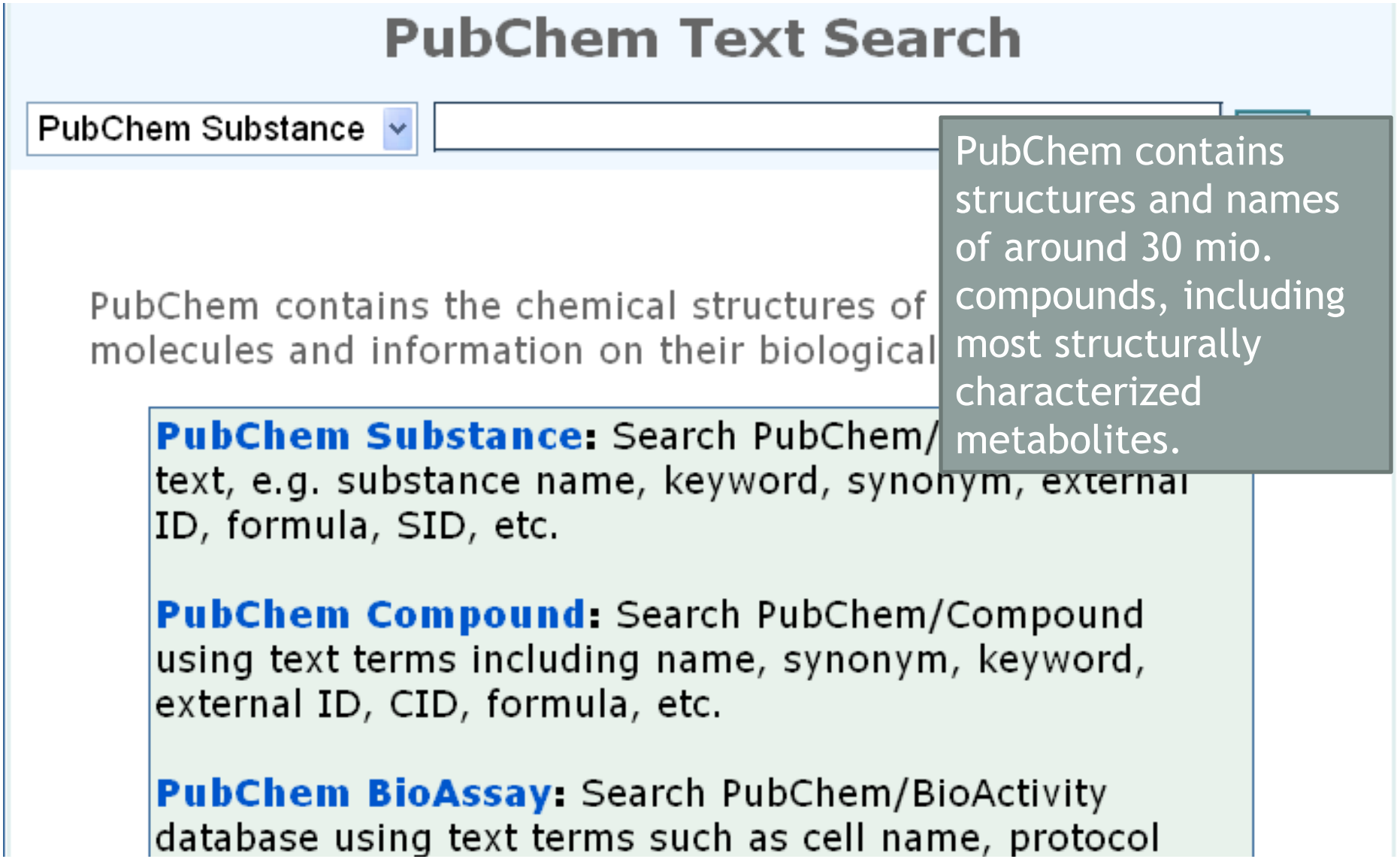
Authors

Pedro Romero, Markus Krummenacker and Peter D. Karp, SRI International.

Project Overview

HumanCyc is a bioinformatics database that describes the human metabolic pathways and the human genome. By presenting metabolic pathways as an organizing framework for the human genome, HumanC

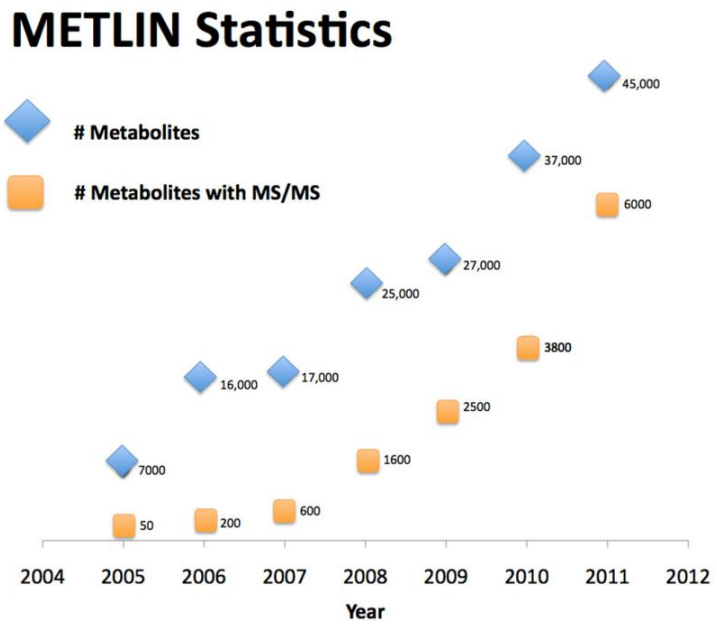
PubChem



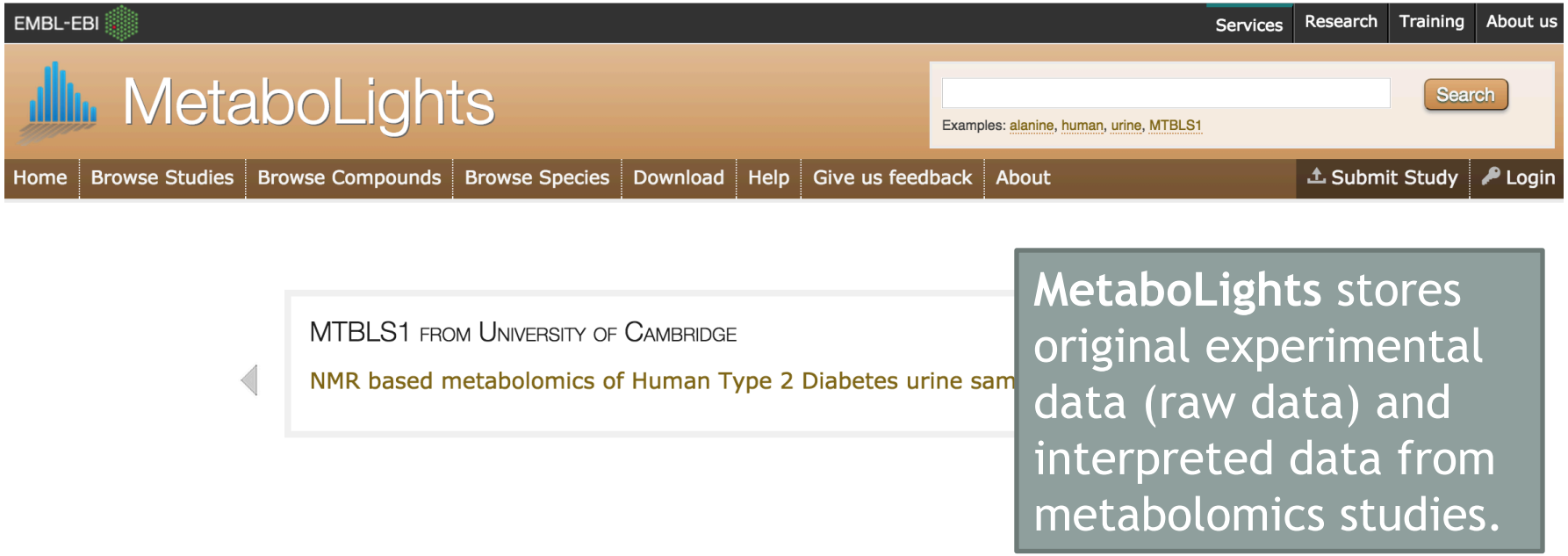
PubChem contains structures and names of around 30 mio. compounds, including most structurally characterized metabolites.

Database Comparison

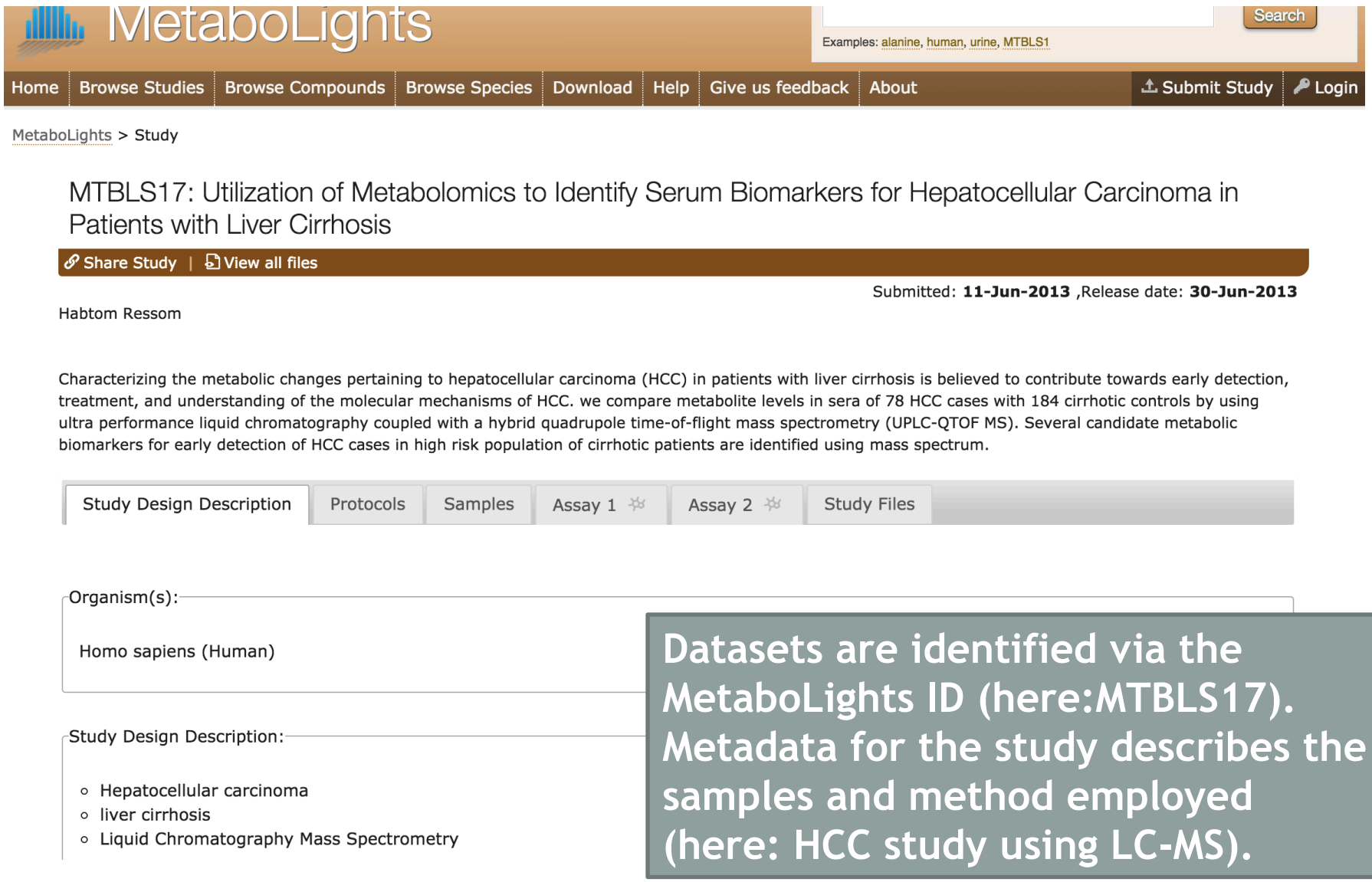
DB	# cpds	# MS^2 spectra
HMDB (2.5)	41,818	10,869
METLIN (2015)	240,000	62,204
MassBank (2015)	?	40,889
NIST (commercial, 2011)	4,694	~100,000
mzCloud (2015)	2,091	142,690
PubChem (2015, cpds only!)	63,156,000	-



Raw Data Repositories



Raw Data Repositories



Datasets are identified via the MetaboLights ID (here:MTBLS17). Metadata for the study describes the samples and method employed (here: HCC study using LC-MS).

Raw Data Repositories

Data

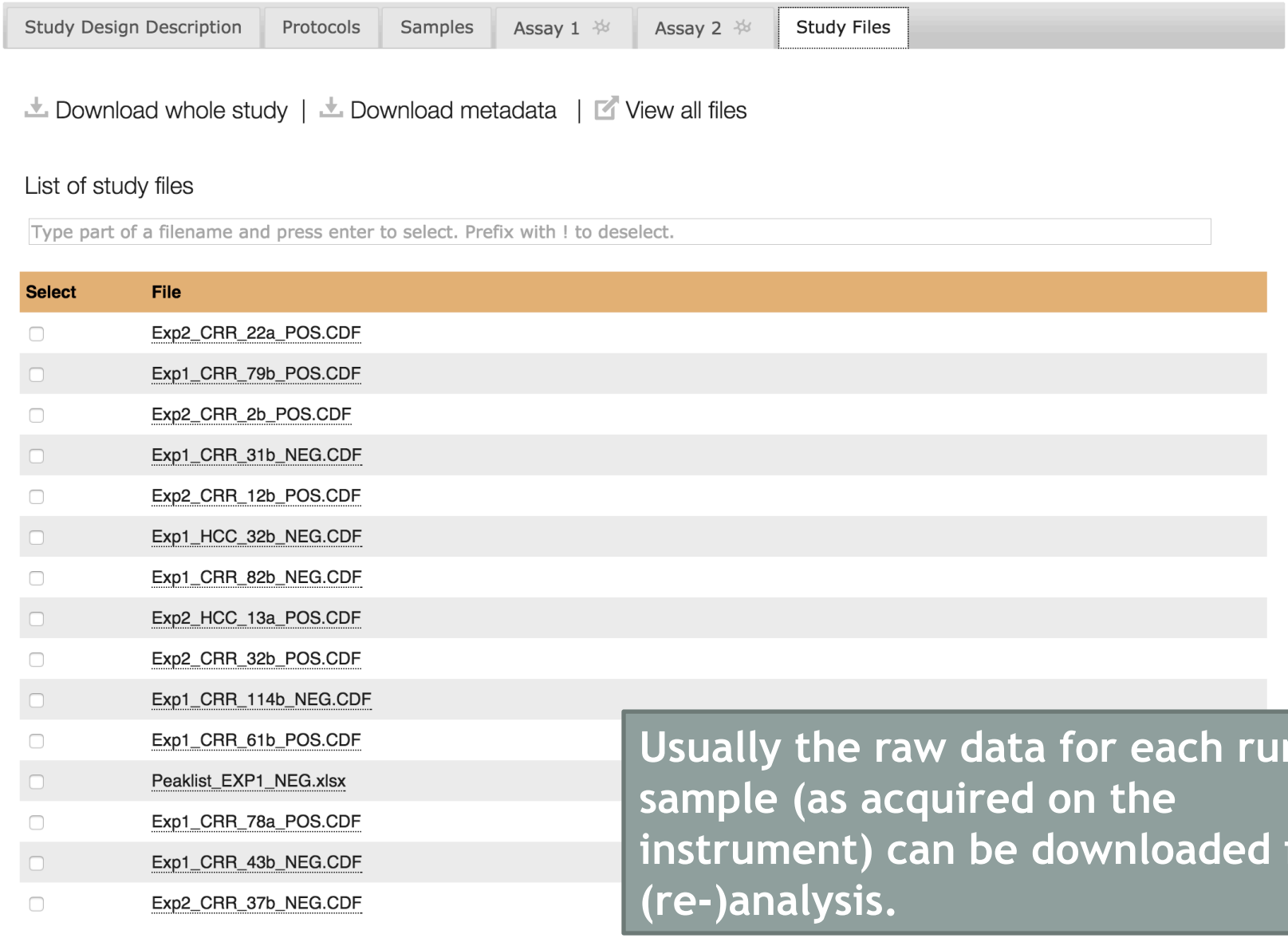
Show 10 entries

Filter:

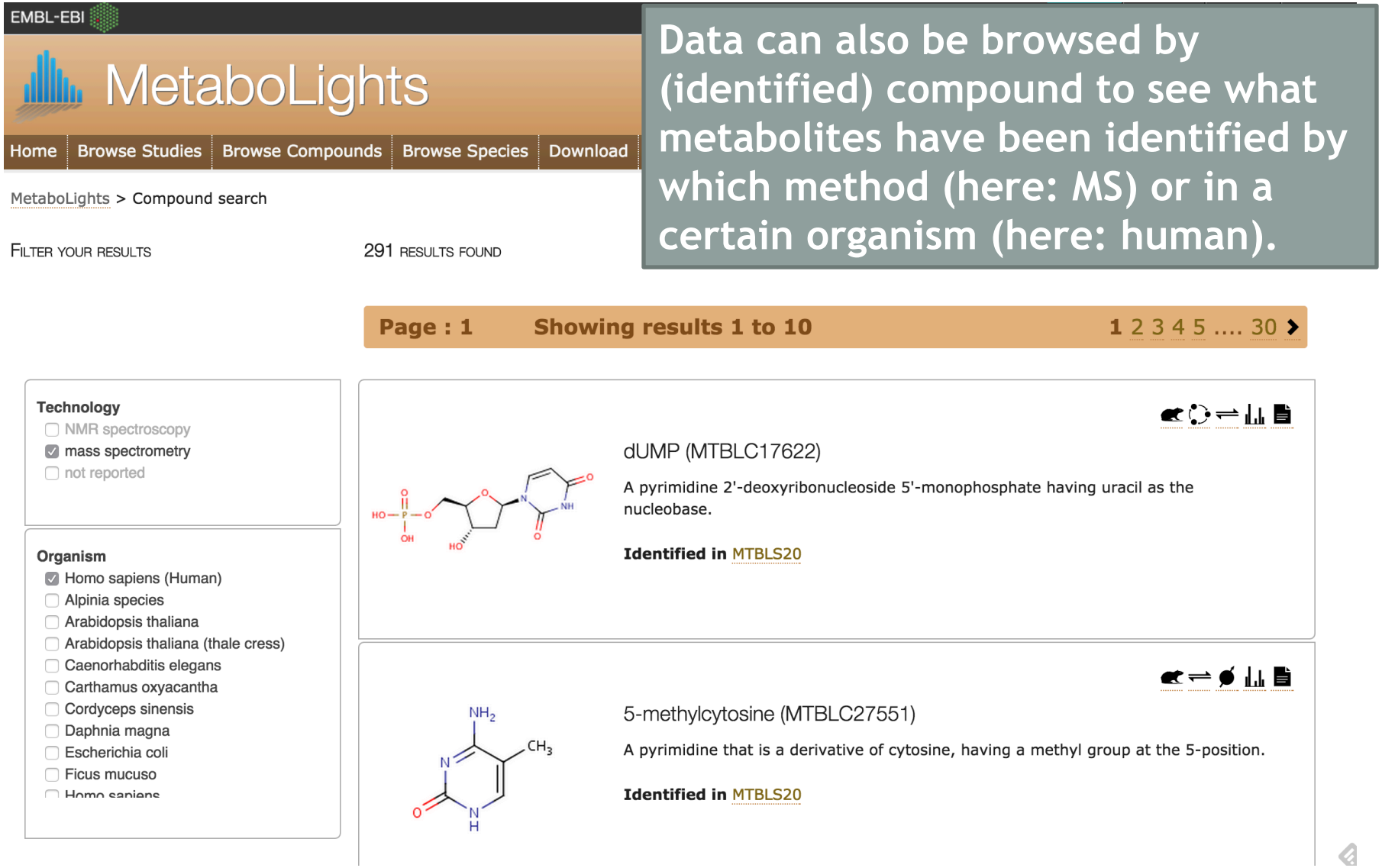
Sample Name	Protocol REF	Post Extraction	Derivatization	Extract Name	Protocol REF	Chromatography Instrument	Column model	Colum type
Exp1_CRR_1a_NEG	Extraction		none		Chromatography	ACQUITY UPLC Systems with 2D Technology	50 × 2.1 mm ACQUITY 1.7-µm C18 column (Waters)	reverse phase
Exp1_CRR_1b_NEG	Extraction		none		Chromatography	ACQUITY UPLC Systems with 2D Technology	50 × 2.1 mm ACQUITY 1.7-µm C18 column (Waters)	reverse phase
Exp1_CRR_3a_NEG	Extraction		none		Chromatography	ACQUITY UPLC Systems with 2D Technology	50 × 2.1 mm ACQUITY 1.7-µm C18 column (Waters)	reverse phase
					Chromatography	ACQUITY UPLC Systems with 2D Technology	50 × 2.1 mm ACQUITY 1.7-µm C18 column (Waters)	reverse phase

For every study there is also information available on each of the individual samples/runs (here: instrument, separation, sample description).

Raw Data Repositories



Raw Data Repositories



Data can also be browsed by (identified) compound to see what metabolites have been identified by which method (here: MS) or in a certain organism (here: human).

Growth of Metabolomics Data

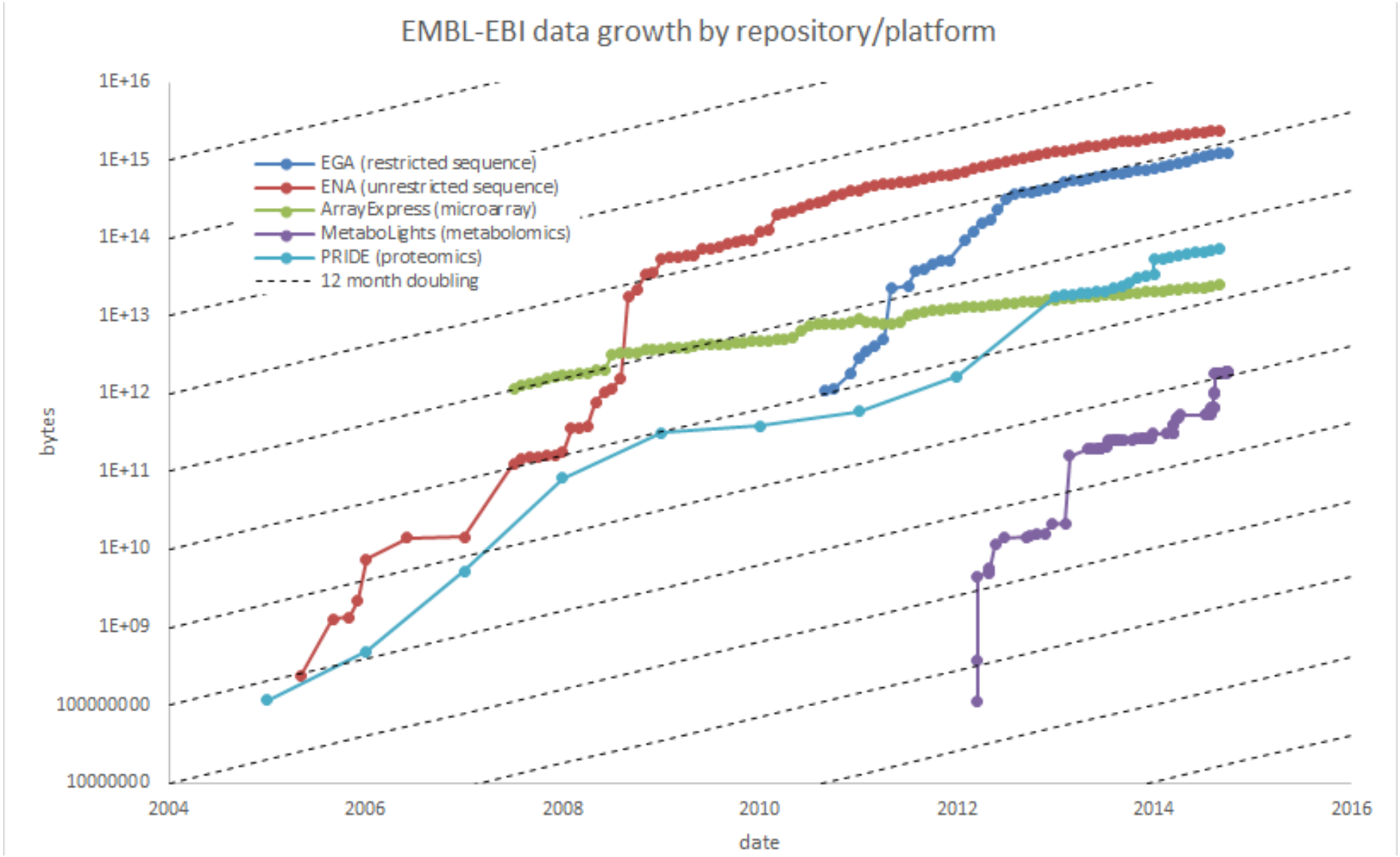


Illustration: Christoph Steinbeck, EBI

LEARNING UNIT 11C
METABOLITE ID VIA SPECTRAL MATCHING AND ACCURATE MASS

- Distribution of metabolite masses
- Matching MS/MS spectra against databases
- Accurate mass search

32

Identification Algorithms

- Identification by **accurate mass**
  - Given a sufficient mass resolution, the mass alone contains valuable information on the metabolite
  - Usually, these masses are not unique
  - Problem: structural isomers are isobaric!
- Identification through **spectrum comparison**
  - **Library search**
    - Compare the experimental spectrum against other experimental spectra
    - Requires library of experimental spectra (difficult)
  - **Fragment tree approaches**
    - Try to predict fragmentation patterns
    - Compare spectrum against theoretical fragmentation pattern

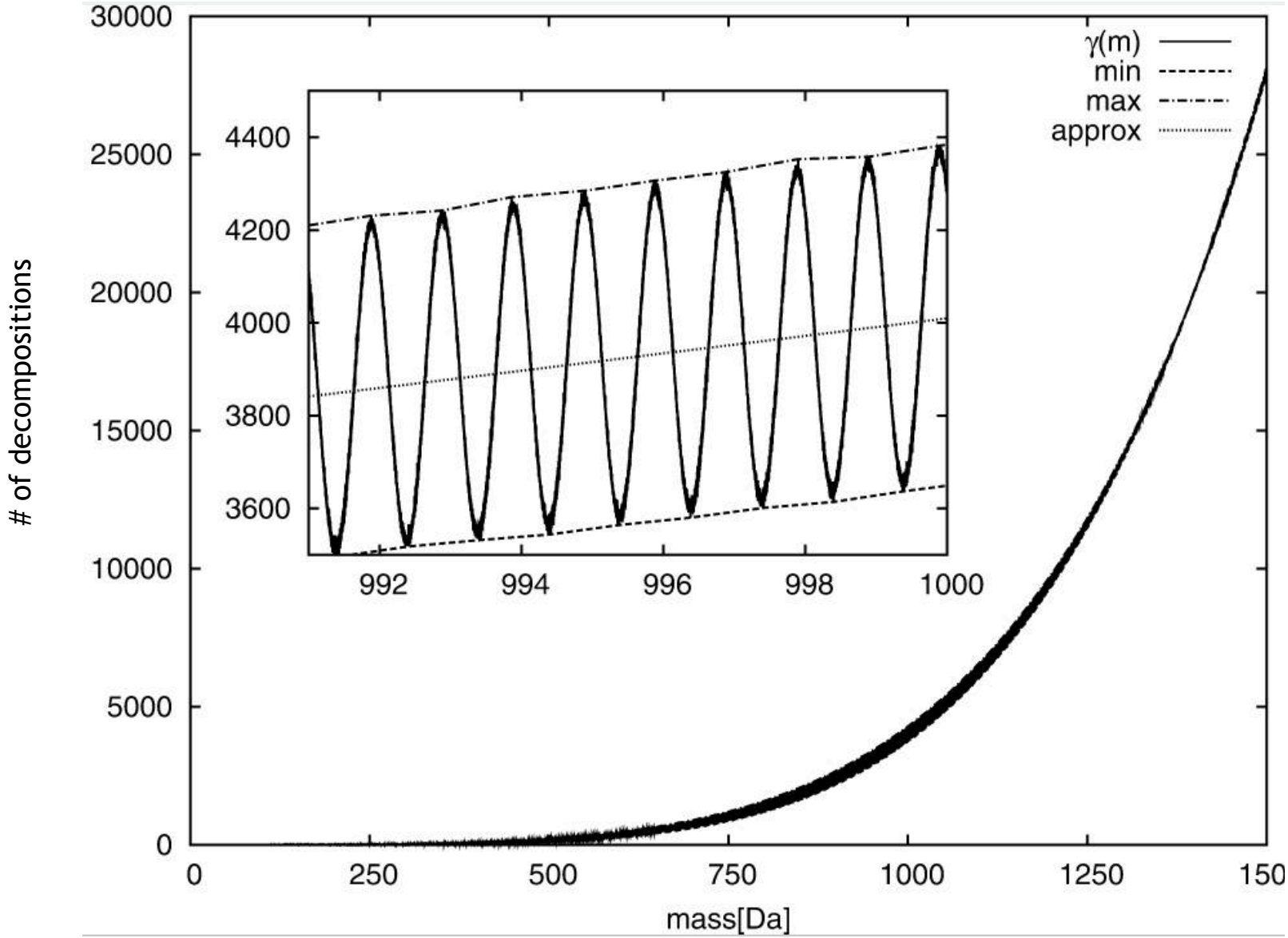
Mass Distribution

Mass Range	# of PubChem Hits
50-51 Da	71
100-101 Da	2,180
150-151 Da	9,163
200-201 Da	23,867
250-251 Da	48,909
300-301 Da	78,577
350-351 Da	112,566
400-401 Da	130,737

Mass Range	# of PubChem Hits
100.0-100.1 Da	367
100.1-100.2 Da	1,628
100.2-100.3 Da	60
100.3-100.4 Da	10
100.4-100.5 Da	21
100.5-100.6 Da	28
100.6-100.7 Da	3
100.7-100.8 Da	3
100.8-100.9 Da	21
100.9-101.0 Da	39

Limited number of different isotope masses gives rise to ‘lumpy’ distribution across the mass range.

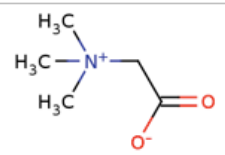
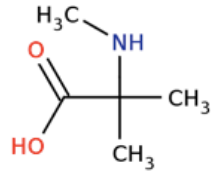
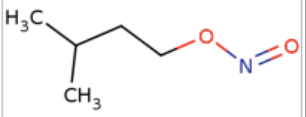
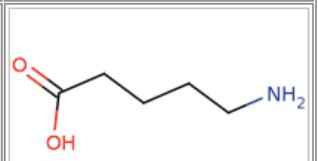
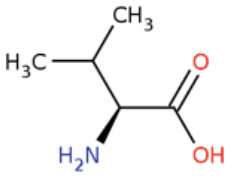
Mass Distribution

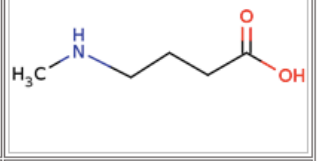
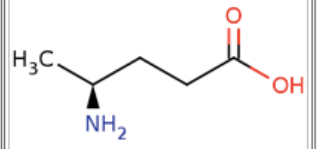
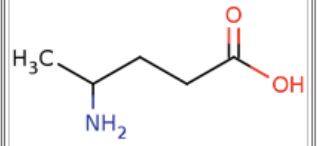
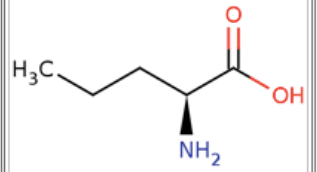
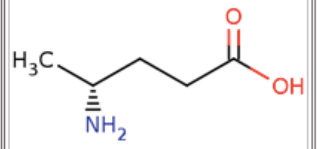


Böcker et al., Bioinformatics (2009); 25(2):218–224.

Isomers

Example: Structural isomers with mass 117.0790 Da, formula: $C_5H_{11}NO_2$

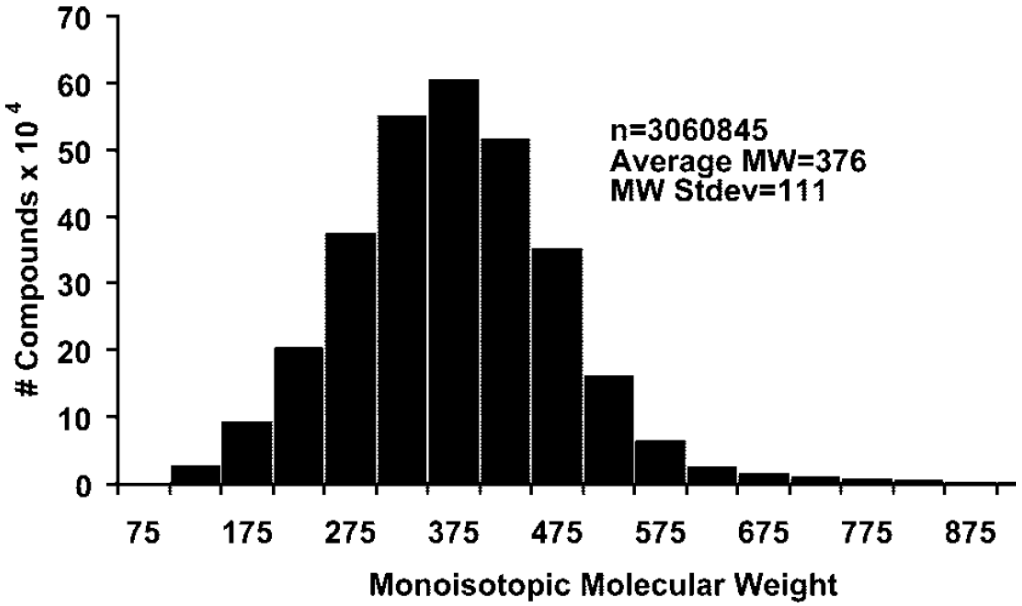
METLIN ID	MASS	Δppm	NAME	MS/MS	STRUCTURE
287	117.0790	0	**Betaine** *Formula:* C5H11NO2 *CAS:* 107-43-7	View	
6508	117.0790	0	**N-Methyl-a-aminoisobutyric acid** *Formula:* C5H11NO2 *CAS:*	View	
6762	117.0790	0	**isoamyl nitrite** *Formula:* C5H11NO2 *CAS:*	NO	
6902	117.0790	0	**5-Aminopentanoic acid** *Formula:* C5H11NO2 *CAS:*	View	
35	117.0790	0	**L-Valine** *Formula:* C5H11NO2 *CAS:* 72-18-4	View	

63884	117.0790	0	**4-Methylaminobutyrate** *Formula:* C5H11NO2 *CAS:* 1119-48-8	NO	
35941	117.0790	0	**4S-aminopentanoic acid** *Formula:* C5H11NO2 *CAS:*	NO	
35942	117.0790	0	**4-amino-pentanoic acid** *Formula:* C5H11NO2 *CAS:*	NO	
35949	117.0790	0	**2S-amino-pentanoic acid** *Formula:* C5H11NO2 *CAS:*	NO	
35940	117.0790	0	**4R-aminopentanoic acid** *Formula:* C5H11NO2 *CAS:*	NO	

36

Metabolite Masses

- Metabolites have a rather limited mass range (compared to peptides)
- They cluster mostly in a mass range usually not relevant to proteomics (proteomics: 500+ Da)
- Most metabolites have between 50-100 atoms and a mass between 300-500 Da



Hill et al., Anal. Chem., 2008, 80:5574

METLIN MS/MS Search

- METLIN provides search options based on
  - Accurate mass
    - Return all metabolites matching a specific mass
  - Tandem spectra
    - Return all metabolites with a matching tandem spectrum
- As with HMDB, spectra cannot be downloaded in bulk, image display only

Scripps Center For Metabolomics

METLIN: Metabolite and Tandem MS Database

MS/MS Spectrum Match

m/z, intensity

Peaks:
(MAX: 30 peaks)

EXAMPLE DATA
POSITIVE NEGATIVE

Mode: Positive Negative

Collision Energy (eV): 20eV

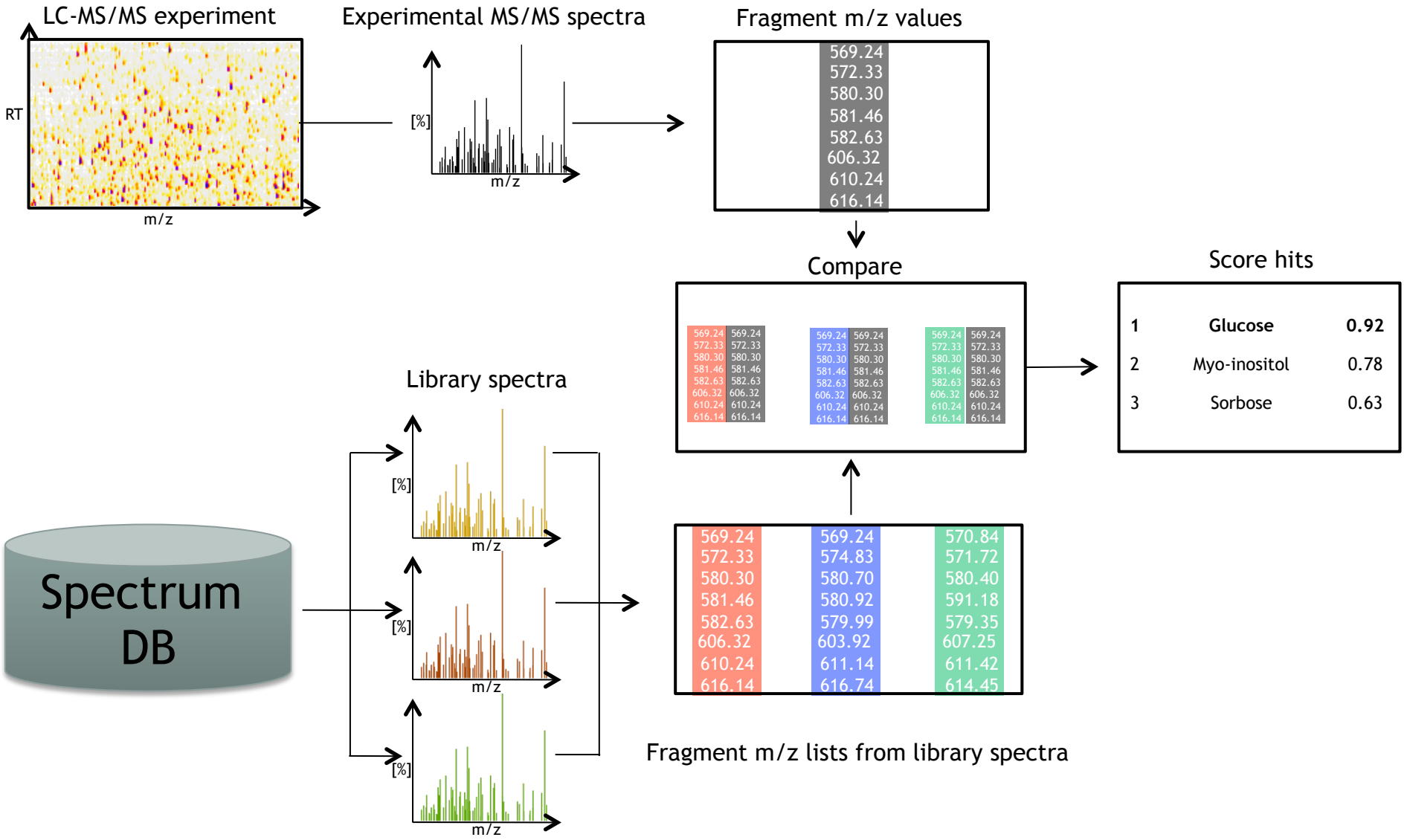
Tolerance MS/MS (Da): 0.01

Tolerance Precursor (ppm) 20

Precursor m/z 195.0877

Find Metabolites Reset

Spectral Library Search



40

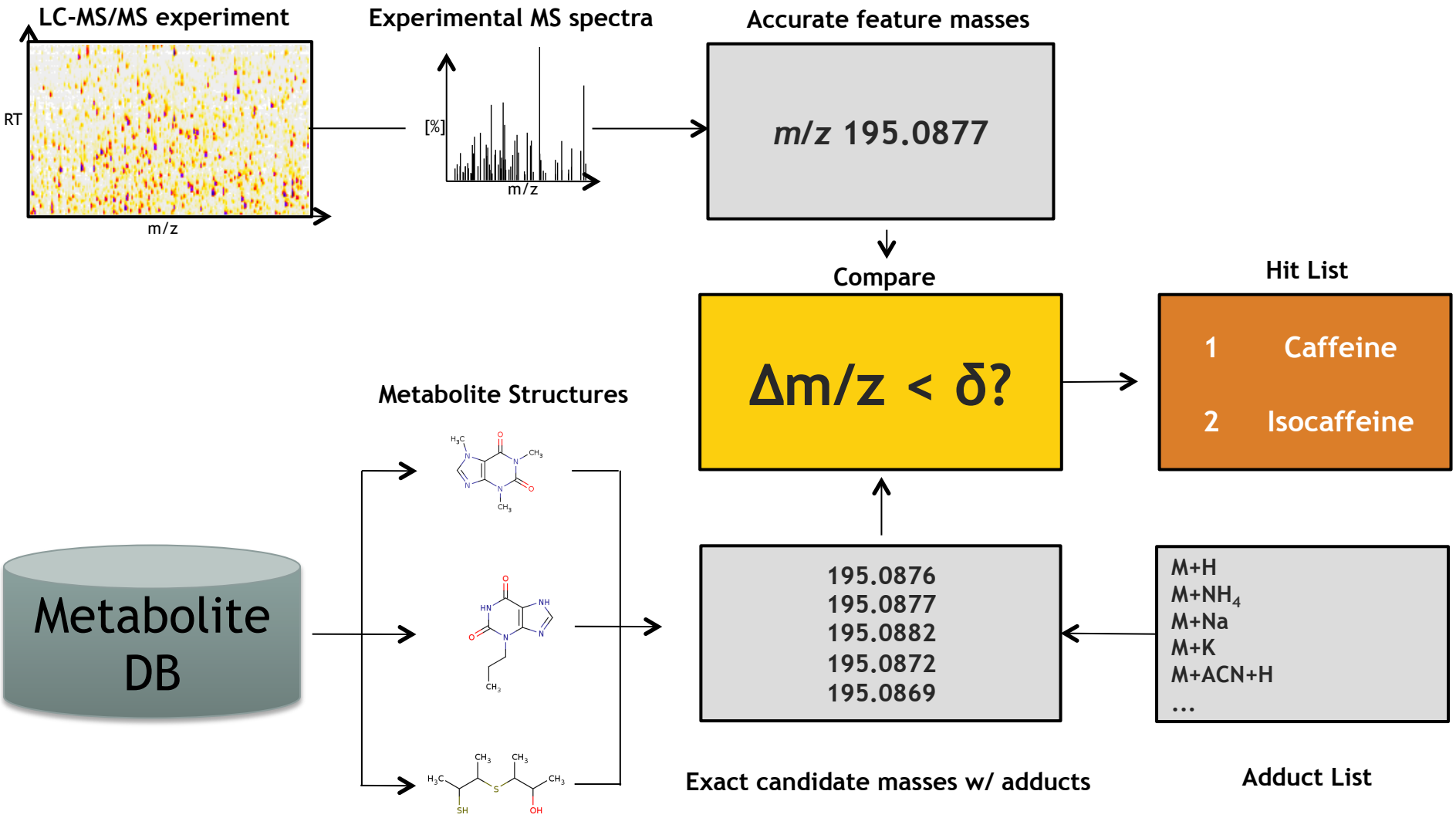
Spectral Library Search

- Searching spectra against a library is done with similarity functions similar to those used in peptide ID (e.g., correlation, shared peak count)
- Success of the spectrum depends on similar experimental conditions:
  - Instrument type
  - Instrument resolution
  - Fragmentation method and parameters (e.g., collision energy)
- Acquiring a library of spectra requires acquisition of spectra, which in turn requires the physical ownership of reference metabolites
- Not all metabolites are readily commercially available

Spectral Library Search

- Key issues
  - Metabolite spectra often less information-rich than peptide spectra
  - Few readily available (downloadable!) databases of metabolite spectra
  - Coverage of spectra thus much worse than in proteomics – we usually can identify only a negligible fraction of the metabolome
  - Fragmentation differs drastically between instruments, depends on collision energy

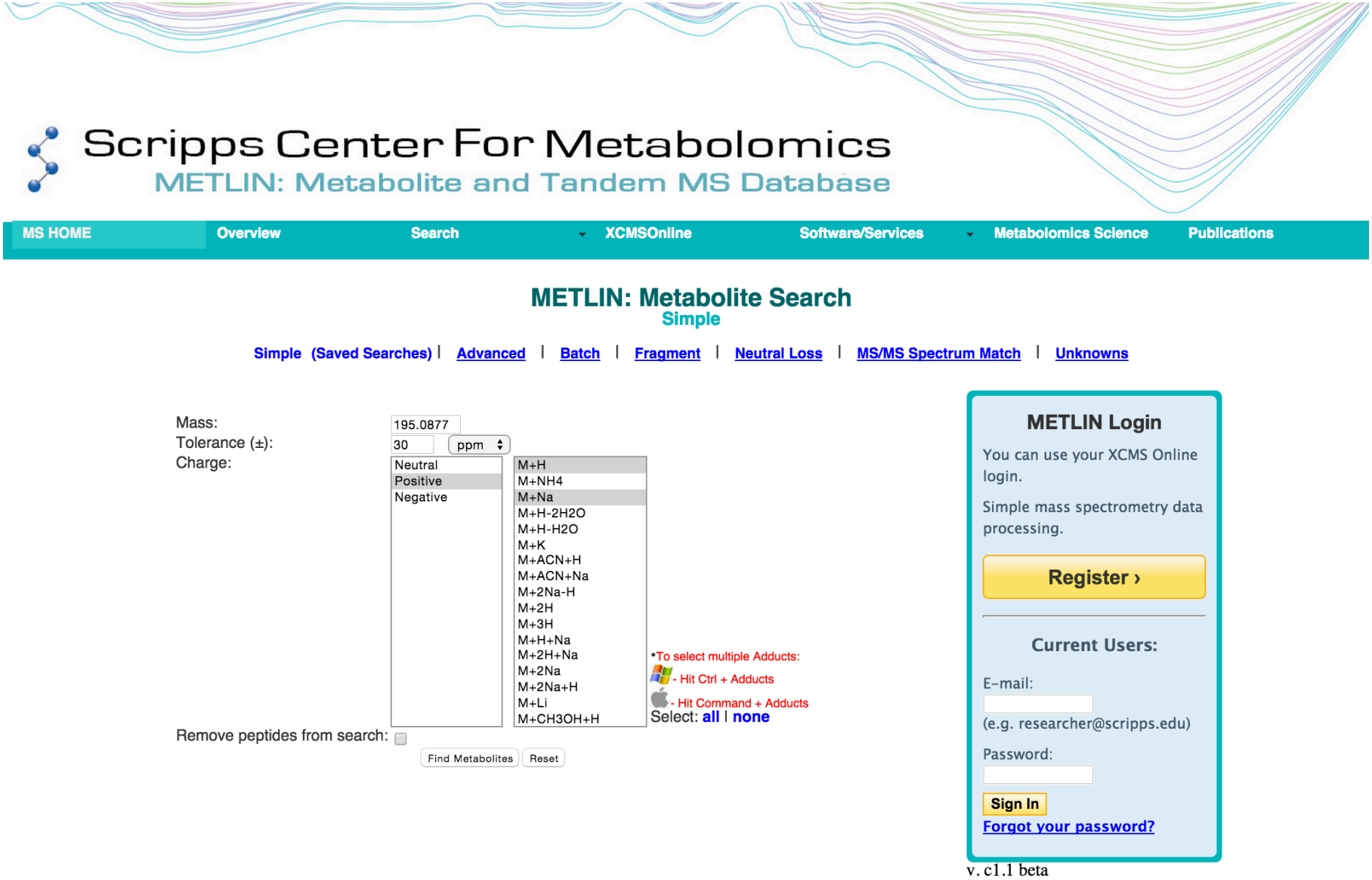
Accurate Mass Search



Accurate Mass Search

- If spectra are not available for a compound, search can be done
- The metabolite m/z differs from the metabolite monoisotopic mass:
  - Ionization adds or removes proton(s)/electron(s)
  - Metabolites often appear as adducts
    - Addition of counter ions (sodium, ammonium)
    - Solvent adducts (acetonitril, methanol)
- Searching each for each compound with multiple possible adducts increases search space
- Search critically depends on instrument mass accuracy (recalibration helps)

Accurate Mass Search w/ METLIN



Accurate Mass Search w/ METLIN

Scripps Center For Metabolomics

METLIN: Metabolite and Tandem MS Database

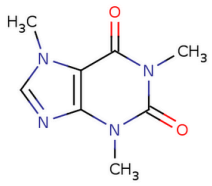
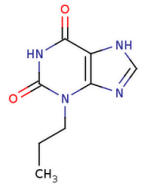
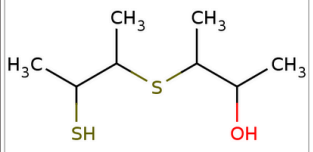
MS HOME | Overview | Search | XCMSOnline | Software/Services | Metabolomics Science | Publications

METLIN
Metabolites

Mass 195.0877 with 30 ppm mass accuracy

Change Query

Total: 18 Metabolites

METLIN ID	MASS	Δppm	NAME	MS/MS	STRUCTURE
1455	[M+H]+ **m/z 195.0877** M 194.0804	0	**Caffeine** *Formula:* C8H10N4O2 *CAS:* 58-08-2	View	
85437	[M+H]+ **m/z 195.0877** M 194.0804	0	**Enprofylline** *Formula:* C8H10N4O2 *CAS:* 41078-02-8	NO	
92244	[M+H]+ **m/z 195.0872** M 194.0799	2	**3-[(2-Mercapto-1-methylpropyl)thio]-2-butanol** *Formula:* C8H18OS2 *CAS:* 54957-02-7	NO	

Accurate Mass Search w/ METLIN

Scripps Center For Metabolomics

METLIN: Metabolite and Tandem MS Database

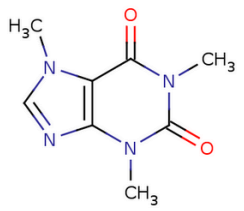
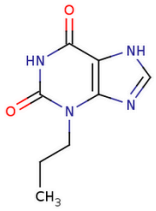
MS HOME | Overview | Search | XCMSOnline | Software/Services | Metabolomics Science | Publications

METLIN
Metabolites

Mass 195.0877 with 1 ppm mass accuracy

Change Query

Total: 2 Metabolites

METLIN ID	MASS	Δppm	NAME	MS/MS	STRUCTURE
1455	[M+H]+ m/z 195.0877 M 194.0804	0	**Caffeine** *Formula:* C8H10N4O2 *CAS:* 58-08-2	View	
85437	[M+H]+ m/z 195.0877 M 194.0804	0	**Enprofylline** *Formula:* C8H10N4O2 *CAS:* 41078-02-8	NO	

Accurate Mass Search - Caveats

- Limiting factors are
  - **Instrument accuracy**: big mass error implies huge search space
  - **Database completeness**: we still search against a (metabolite structure) database – compounds need to be known to be identifiable
  - Isobaric ambiguities: isobaric metabolites cannot be distinguished
- Consequence: identified metabolites are validated again experimentally (standard compounds, reference spectra) whenever possible

LEARNING UNIT 11D
FRAGMENTATION TREES

- Fragmentation tree concepts
- Algorithmic approaches
- Software packages

Fragmentation Trees

- Fragmentation trees are directed acyclic graphs relating the molecular structure of a metabolite to potential fragment masses
- **Idea**
  - Fragmentation means breaking bonds in the metabolite
  - By systematically and repeatedly breaking all bonds in a structure and its fragments, one can construct the set of all possible fragments
  - Fragmentation occurs preferentially along low-energy dissociation pathways
- **Problems**
  - Not all fragments will be observed
  - Gas phase fragmentation chemistry is more complex
    - Rearrangement reactions lead to different structures
    - Neutral losses are not covered by this
    - Cycles in molecular graphs pose problems

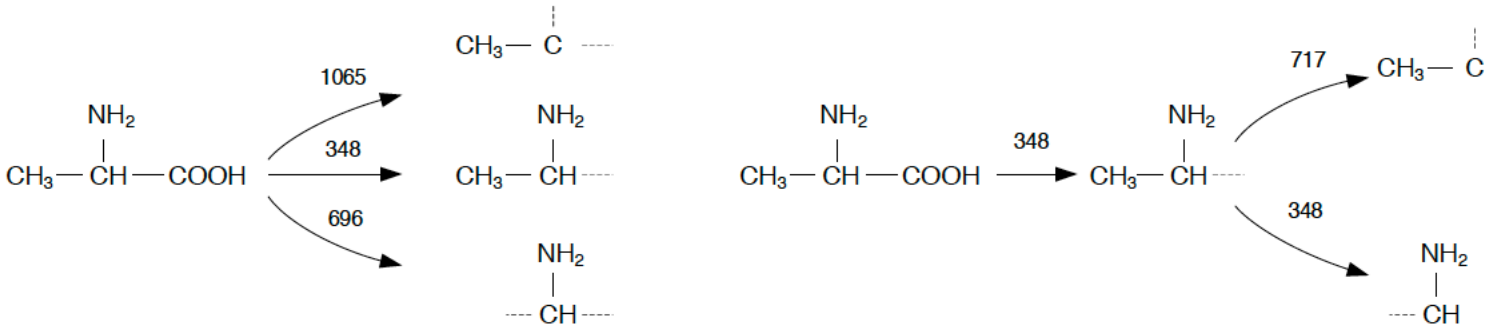
Fragmentation Trees

- There are different ways for constructing fragmentation trees
  - Decomposition of molecular structures
  - Decomposition of molecular formulas
- The first method requires – and leverages – information from available molecular structures
- The second approach relies solely on sufficiently accurate masses
- We will first discuss the approach based on structures as suggested by Heinonen (2006) and then discuss the algorithmic strategies of the second approach by Böcker & Rasche (2008)

Heinonen et al., Proc. German Conf. Bioinfo (GCB 2006)
Böcker & Rasche, Bioinformatics (2008), 24:i49-i55.

Structural Fragmentation Trees

- Heinonen et al. suggest to start with a molecular graph *G = (V, E)* where *V* represent atoms and *E* bonds between atoms and decompose this graph by breaking bonds, i.e. removing edges from *G*



- Any connected subgraph of *G = (V, E)* is a possible fragment that can be formed from *G*
- More formally:

  The fragmentation tree $G_F = (F, E_F, c)$ of *G* is an acyclic directed graph where
  - *F* is the set of nodes corresponding to all fragments that can be formed of *G* (i.e., all possible connected subsets $E' \subseteq E$)
  - E_F is the set of directed edges from each fragment in *F* to its subfragments (i.e., fragments that are themselves connected subsets of a fragment or the original molecule)
  - c: $E_F \rightarrow \mathrm{R}$ associates each edge with a cost for forming this fragment

Structural Fragmentation Trees

- Source molecules can be fragmented in a single step (yielding MS2 spectra) or iteratively until no further fragments can be found (yielding multiple fragmentations as in MS^n spectra)
- The cost function models the likelihood for forming a particular fragment
- Cost function can be modeled by bond energies (more stable bonds are more 'expensive', less likely to break)
- The best explanation for a tandem spectrum (or for fragment masses observed across multiple fragmentation experiments) should thus be the **most likely fragmentation pattern**

Heinonen et al., Proc. German Conf. Bioinfo (GCB 2006)

Structural Fragmentation Trees

- **Identification**
  - Given a spectrum $S = (s_1, \ldots, s_k)$ with distinct peak masses s_i
  - Find the lightest connected subtree G_F^* of G_F of the fragmentation tree that explains all masses s_i in the spectrum
- This can be repeated for a set of possible structures and results can be scored
- The lightest subtree needs to break the fewest/weakest bonds and still explains all fragments
- Masses are represented by internal nodes as well as leaves of G_F^*
- It has to be connected, because the ions represented by leaves cannot be formed if the fragments represented by internal nodes are not formed

Structural Fragmentation Trees

- Heinonen *et al.* suggest that lightest fragmentation trees can be found using mixed integer linear programming (MILP) – for details see their paper
- They tested the method on a small set of metabolite spectra
- For the majority of metabolites, their method could explain 80-90% of the fragments
- Their method was not used to perform a thorough identification benchmark, though

Compositional Fragmentation Trees

- Heinonen's structural fragmentation trees are the metabolomics equivalent of protein database search: a set of structures is fragmented and theoretical fragmentation trees can then be compared to an experimental spectrum
- Compositional fragmentation trees can be used in a way similar to de novo identification in proteomics
- The idea is based on the concept of mass decomposition
  - A fragment/molecule mass has to correspond to the molecular mass of an existing chemical structure
  - It thus has to correspond to the sum of integer multiples of atomic masses
  - Given a mass measurement accuracy δ, each measured fragment mass m_i has to satisfy

    $$m_i - \delta \leq \sum_\iota a_i c_i \leq m_i + \delta$$

    for monoisotopic atomic masses a_i of all elements involved and $c_i \in N_0$
  - Coefficients c_i define the composition of the fragment/molecule, i.e. its chemical formula (number of atoms of each element)

Rapid Mass Decomposition

- Given a fragment mass m_i and monoisotopic atomic masses corresponds to an integer knapsack problem ('find all possible ways to pack the atoms into a knapsack such that it has the required fragment mass')
- Given a limited set of elements (CHNOPS), one can easily enumerate all possible compositions for a given mass (e.g., using dynamic programming, but more efficient solutions exist)
- For CHNOPS, there are about 700 mio. different compositions with a mass below 1,000 Da (Böcker et al., 2006)
- Given a sufficient mass accuracy (1-2 ppm), it is usually possible to uniquely identify the composition of a molecule
- Accordingly, it is possible to reconstruct the compositions for fragment ions

Böcker et al., WABI 2006, LNBI 4175, pp. 12-23, Springer, 2006

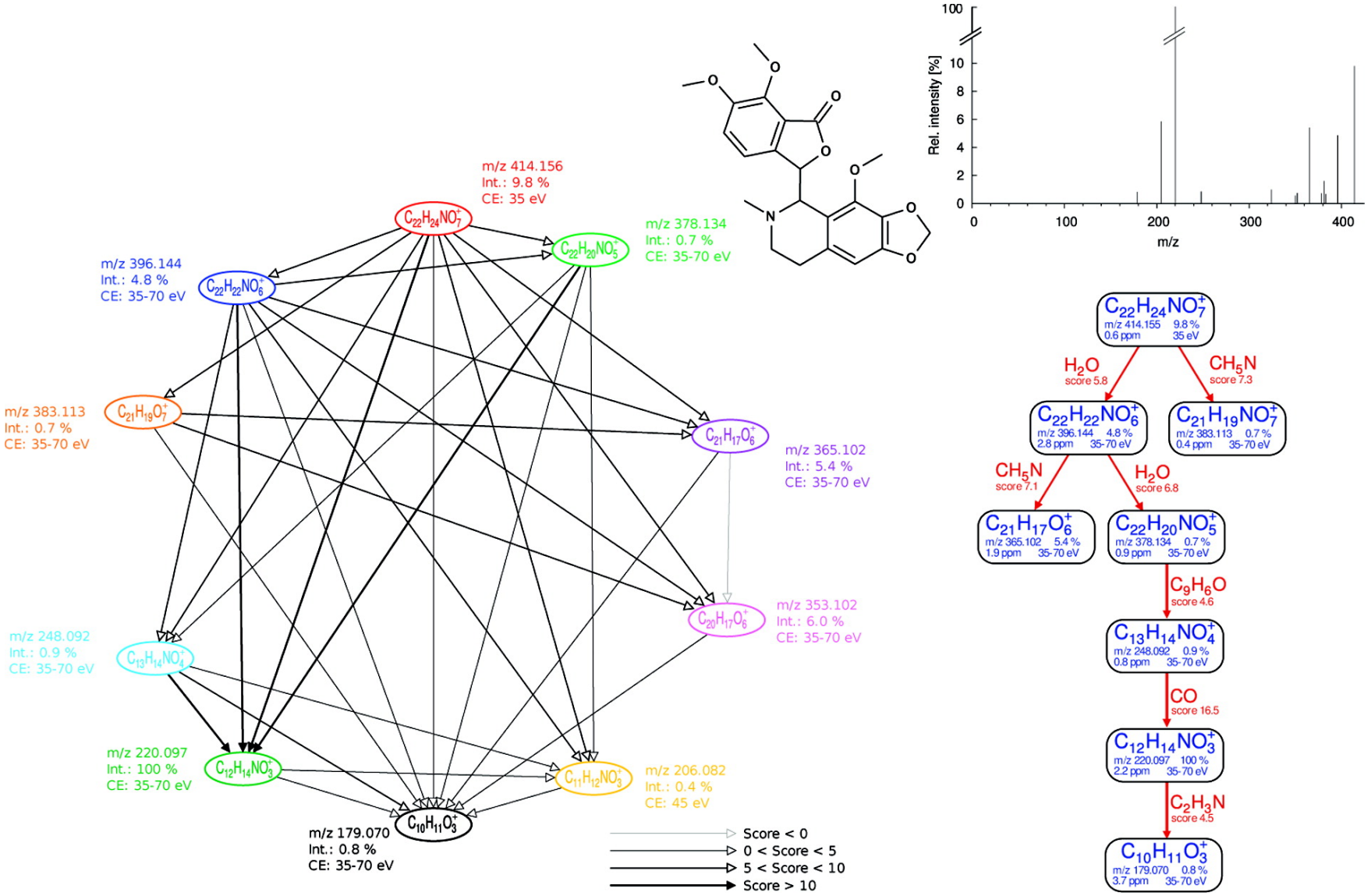
Compositional Fragmentation Trees

- A **compositional fragmentation graph** is then constructed from the observed masses m_i of fragment ions and the molecular mass M obtained in a set of experiments
- A composition $c = (c_1, \ldots, c_k)$ (molecular formula) of a fragment is kept for a mass, if it is a **submolecule** of the metabolite composition $C = (C_1, \ldots, C_k)$, i.e. if

$$c_i \leq C_i, \forall 1 \leq i \leq k$$

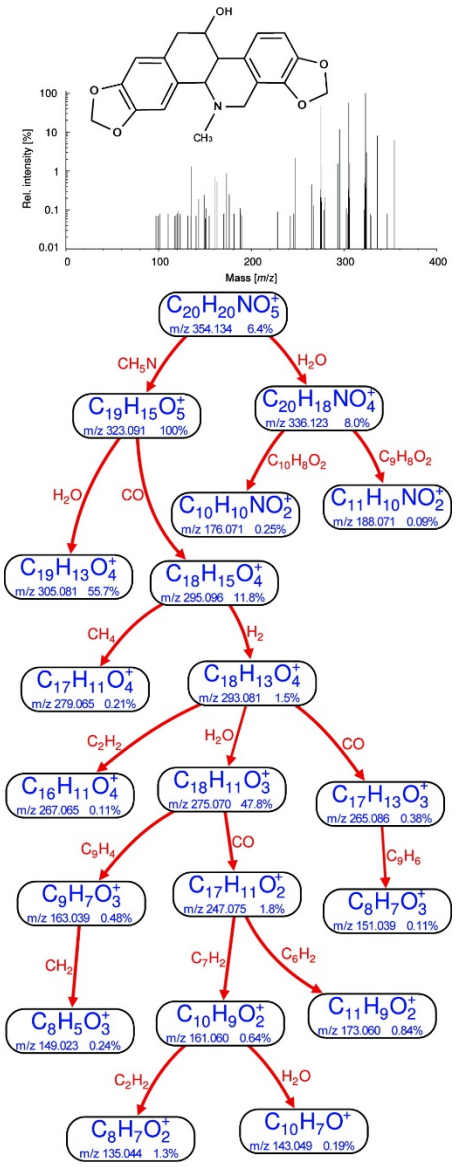
- The compositional fragmentation graph $G_C(V, E)$ is then a **directed acyclic graph** with **nodes representing each possible composition** of any of the fragment masses in the spectrum and the mass of the original metabolite
- A directed edge (u, v) is inserted if node v represents a submolecule of the fragment represented by node u
- Each node is assigned a **color based on the mass** its composition was derived from

Compositional Fragmentation Trees

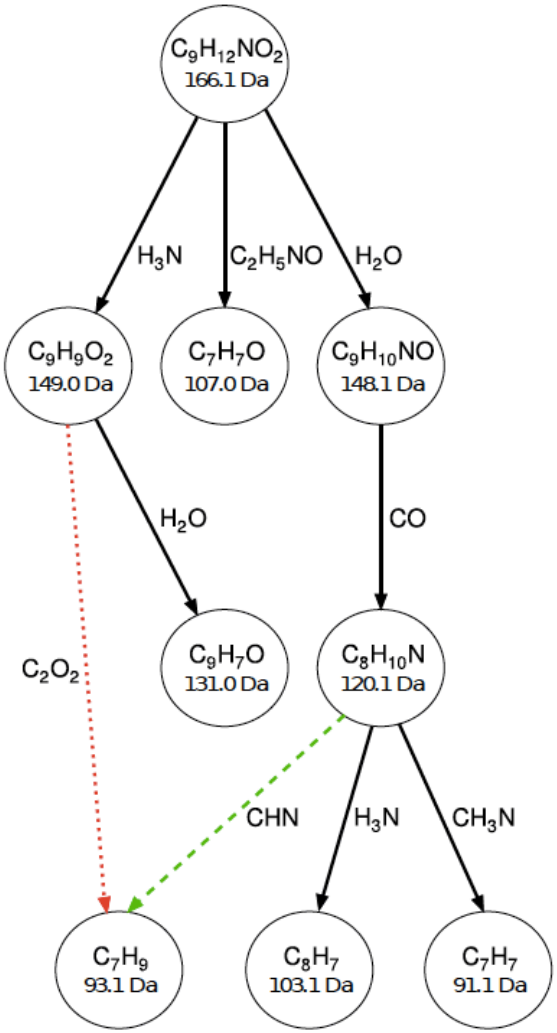


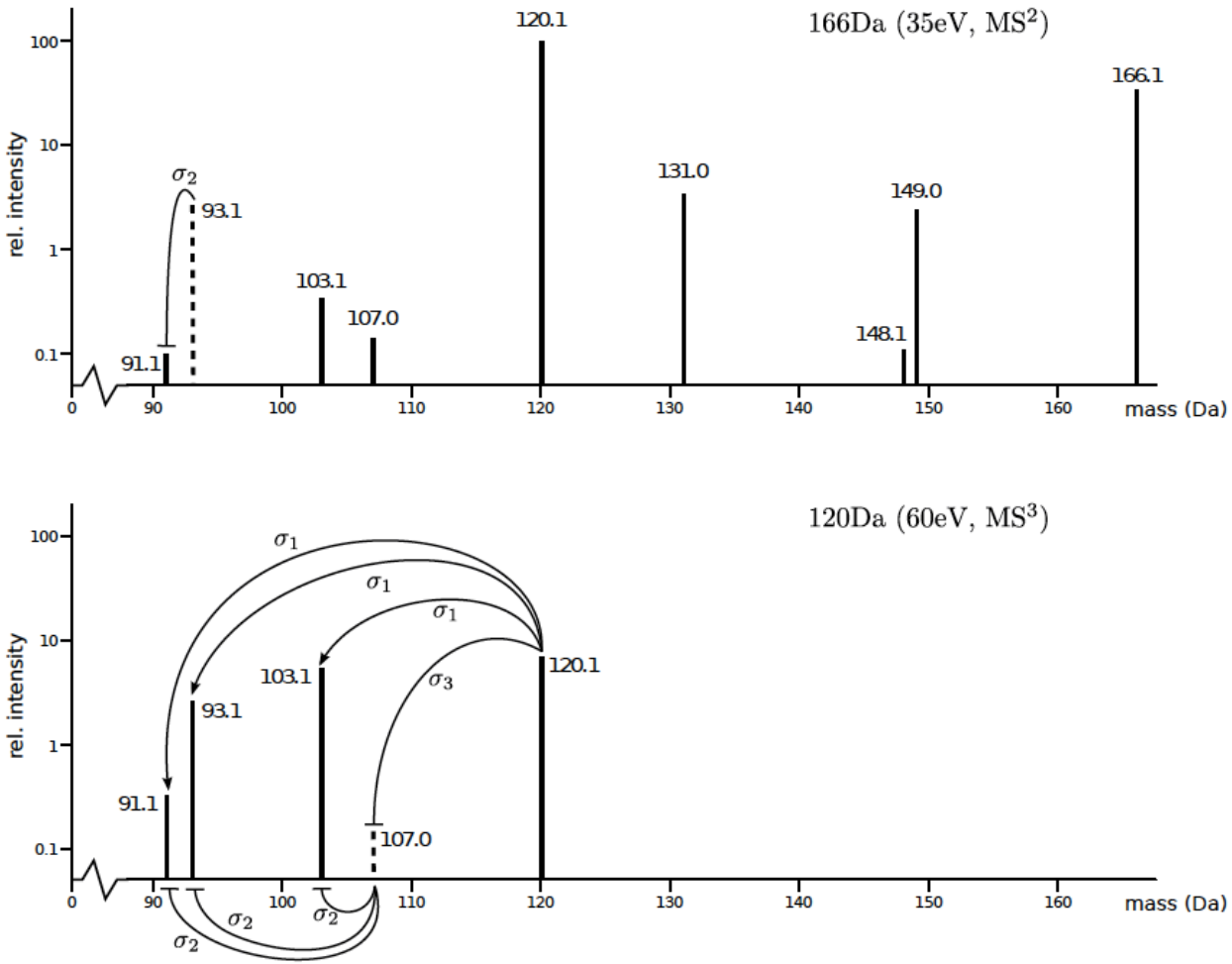
Compositional Fragmentation Trees

- Notes:
  - Nodes represent compositions and thus many nodes may represent the same peak
  - Multiple nodes can thus have the same color
  - The graph is transitive: (u, v), (v, w) $\in$ E also implies (u, w) $\in$ E
  - Information from multiple spectra (MS^n or MS^2 at varying collision energies) can be freely combined



Compositional Fragmentation Trees





Maximum Colorful Subtree

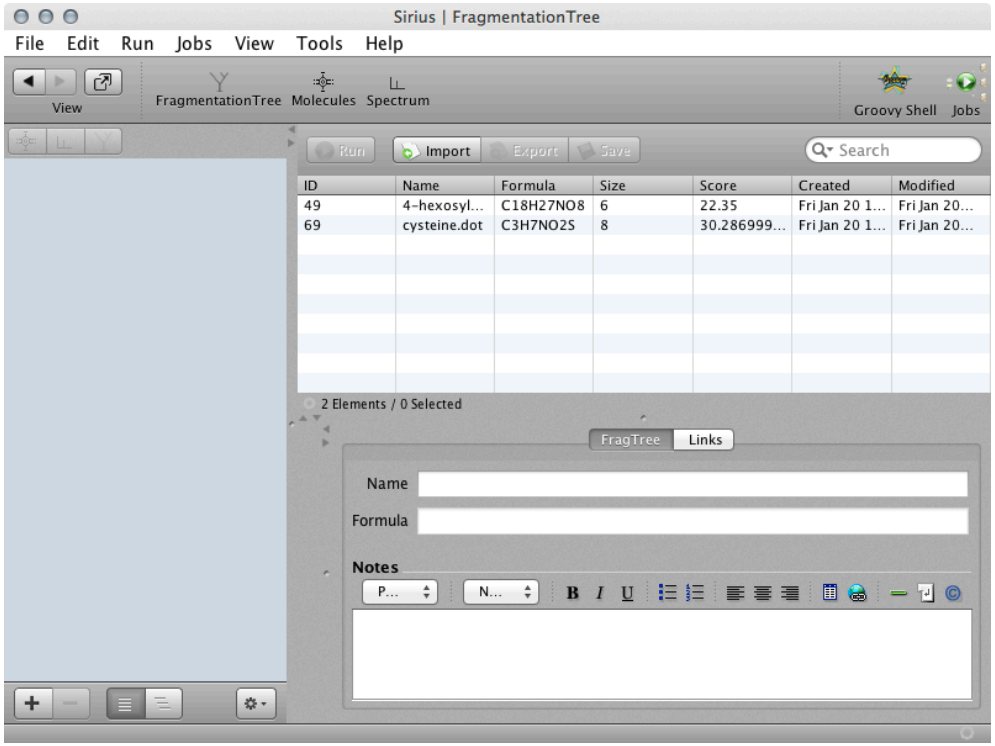
- **Assumption**
  - Every peak is explained by a single molecular fragment
- This implies, that a valid fragmentation tree
  - is a subtree of the fragmentation graph and
  - contains at most one node of each color
- We can define a **colorful subtree** T = (V_T, E_T) of a vertex-colored DAG *G* as a subtree of *G* which uses each color at most once
- If we are given edge weights that correspond to fragmentation probabilities, we can try to identify the most likely fragmentation tree
- This corresponds to finding the **maximum colorful subtree**, i.e. the colorful subtree of G with the maximal weight

Maximum Colorful Subtree

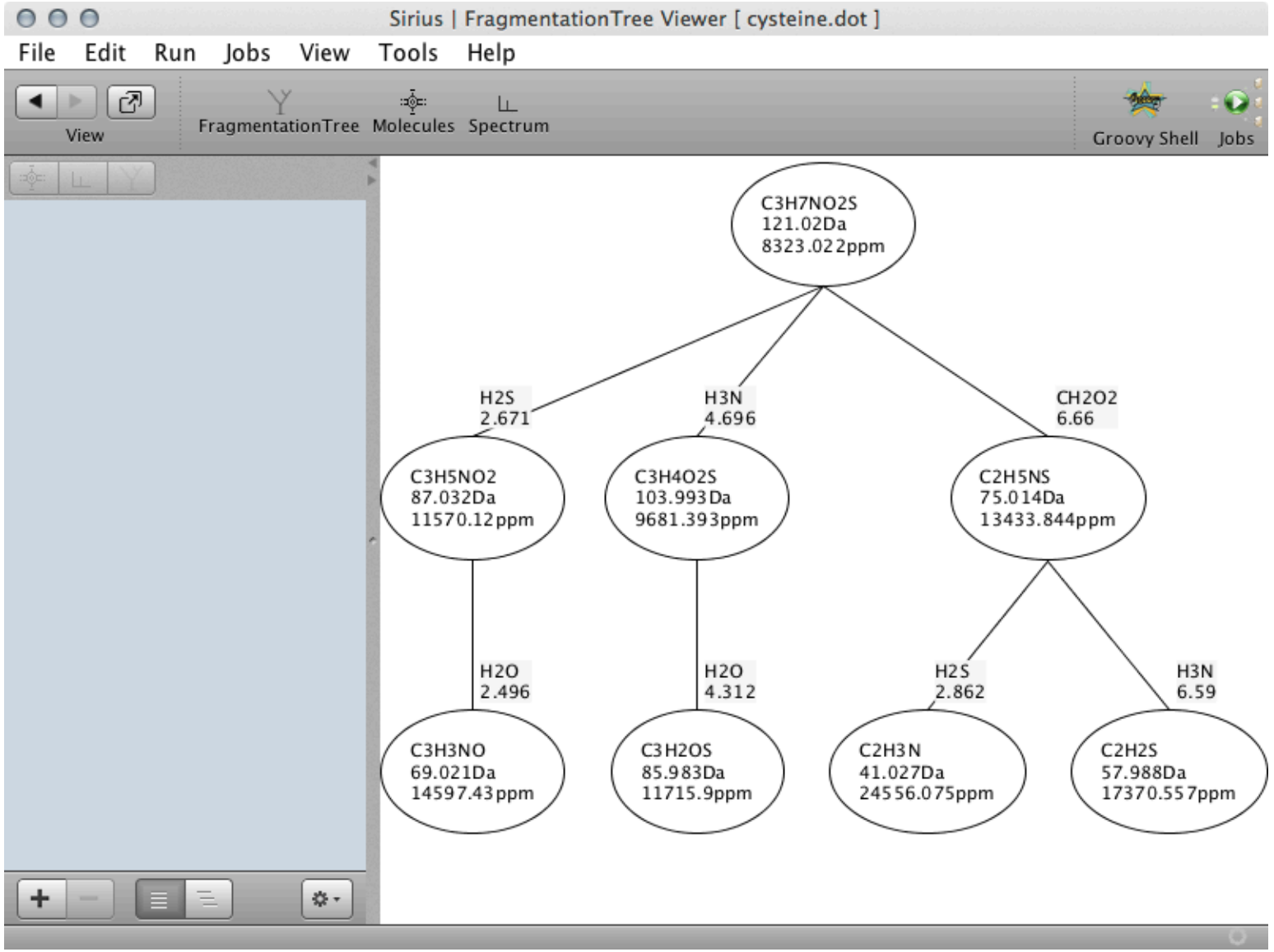
- Finding a maximum colorful subtree is in general NP-hard even for unit weights
- Böcker & Rasche show this by a reduction from SAT
- Accordingly, standard methods for solving NP-hard problems have been applied
  - Branch & bound
  - FPT algorithms (fixed-parameter tractable)
  - Dynamic programming
  - Heuristic solutions
- All these methods yield good solutions for the identification of fragmentation trees and can be used to identify metabolites
- Notes
  - **Identification of the chemical formula does not yet uniquely identify the metabolite**!
  - Structural isomers still have the same chemical formula

Sirius²

- Sirius² is a software solution implementing fragmentation tree approaches for metabolite identification developed by Sebastian Böcker's group in Jena
- It permits the calculation of theoretical fragmentation trees from sets of spectra and a scoring of fragmentation trees against spectra and structures
- The scoring is based an a modified Bayesian formulation for fragmentation trees



Sirius^2



Rasche et al., Anal. Chem., 2011, 83:1243–1251

References

- **Databases**
  - **HMDB**
    http://hmdb.ca
  - **METLIN**
    http://metlin.scripps.edu
  - **mzCloud**
    http://mzcloud.org
  - **MassBank**
    http://massbank.org
  - **MetaboLights**
    http://www.ebi.ac.uk/metabolights
- **Papers**

  - Hill DW, Kertesz TM, Fontaine D, Friedman R, Grant DF. Mass spectral metabonomics beyond elemental formula: chemical database querying by matching experimental with computational fragmentation spectra. Anal Chem. 2008, 80:5574-82.
  - Heinenen M, Rantanen A, Mielikäinen, Pitkänen E, Kokkonen J, Rousu J, Ab initio prediction of molecular fragments from tandem mass spectrometry data, Proc. German Conference on Bioinformatics (GCB 2006), LNI P-83, 2006, Gesellschaft für Informatik
  - Scheubert K, Hufsky F, Rasche F, Böcker S. Computing Fragmentation Trees from Metabolite Multiple Mass Spectrometry Data, RECOMB 2011 (2011), LNBI 6577, pp. 377-391, Springer, Heidelberg, 2011.
  - Böcker S, Letzel MC, Liptak Z, Pervukhin A. Decomposing Metabolomic Isotope Patterns, WABI 2006, LNBI 4175, pp. 12-23 , 2006, Springer, Heidelberg
  - Böcker S and Rasche F. Towards de novo identification of metabolites by analyzing tandem mass spectra. Bioinformatics, 2008, 24:i49-i55.
  - Florian Rasche, Aleš Svatoš, Ravi Kumar Maddula, Christoph Böttcher, and Sebastian Böcker. Computing Fragmentation Trees from Tandem Mass Spectrometry Data. Analytical Chemistry (2011) 83 (4): 1243–1251

Materials

- Learning Units 11A-C