

ACTIONS
MARIE CURIE

EDA EMERGE

eawag
aquatic research

Databases:

Compound Databases

Spectral Databases

Eawag: Swiss Federal Institute of Aquatic Science and Technology

Emma Schymanski
Marie Curie Inter-European Postdoctoral Fellow

Plus many others who I have worked with...

eawag
aquatic research

An overview of databases

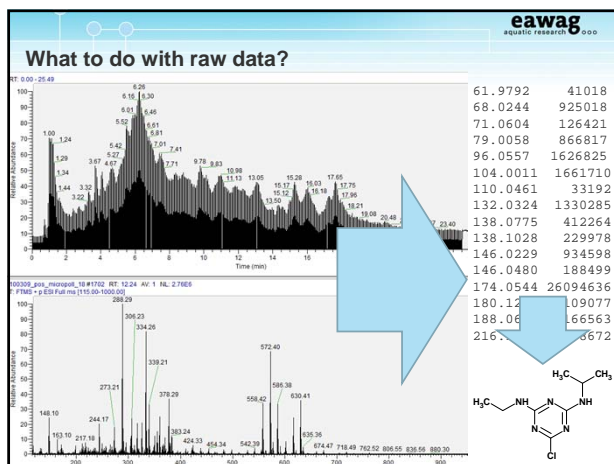
- Compound databases (at least the free ones!)
 - ChemSpider
 - PubChem
- GC-EI-MS Databases
 - Demonstration using NIST
- Online MS and MS/MS databases
 - MassBank
 - METLIN

ChemSpider
The free chemical database

PubChem
Structure Search

MassBank

METLIN



eawag
aquatic research

Database Searching

Some Definitions

- Compound Database:
 - A collection of structures, basic properties and associated information¹
 - Generally, no spectral data – just structures
 - ChemSpider/PubChem have >25,000,000 entries
 - But don't assume that everything is in here, it's not!
- Mass Spectral Databases (or Libraries)
 - A collection of structures, their mass spectra and associated information
 - NIST and Wiley Mass Spectral Libraries are the very widely accepted resources for GC-EI-MS
 - NIST11: >200,000 spectra; Wiley 9th: >660,000 spectra²
 - MS/MS databases are growing; none are yet "established"
 - MassBank: 31,152 spectra;
 - METLIN: 52,904 HR-ESI-MS/MS

¹<http://www.chemspider.com/About.aspx>. ²<http://www.sisweb.com/software/ms/wiley.htm>

eawag
aquatic research

Database Searching

What do you really want to achieve?

- Compound database:
 - I want to find out more about compound X
 - I have a mass and error – what could it be?
 - I have a molecular formula – what could it be?
- Mass Spectral database
 - I have a mass and error or a molecular formula
 - Are there any spectra? What do they look like?
 - I have an EI-MS spectrum and NIST
 - How to interpret NIST results - %, match, reverse match, ...
 - I have a MS/MS peak list
 - Spectrum search with MassBank, METLIN

eawag
aquatic research

Compound Database 1: ChemSpider

<http://www.chemspider.com/>

ChemSpider
The free chemical database

RSC | Advancing the Chemical Sciences

About | More Searches | Web APIs | Help

Simple search | Structure search | Advanced search

eg. Aspirin

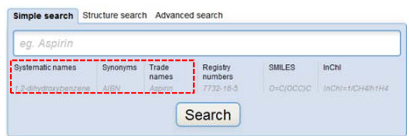
Systematic names	Synonyms	Trade names	Registry numbers	SMILES	InChI
2-dihydroxybenzoic acid	Aspirin	Aspirin	7722-83-5	O=C(O)C1=CC=CC=C1	InChI=1C1=CC=C(C=C1)C(=O)O

Search

YOU COULD ADVERTISE HERE

Compound Database 1: ChemSpider
Simple Search - Name

o "Simple Search" – is already quite complicated!

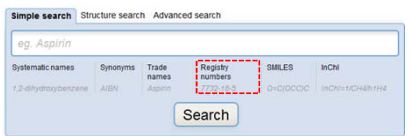


o Names are quite straightforward – type it in and see

o Many compounds even have local language names now

Compound Database 1: ChemSpider
Simple Search – CAS Number

o "Simple Search" – is already quite complicated!



o Registry numbers – e.g. CAS Number

o CAS = Chemical Abstract Services = all registered chemicals

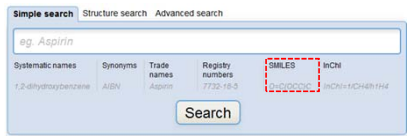
o Generally the form XXX-YY-ZZ – Xs can be 5 or 6 digits long.

o CAS have their own database

o <https://scifinder.cas.org> – but not for free!

Compound Database 1: ChemSpider
Simple Search - SMILES

o "Simple Search" – is already quite complicated!



o SMILES – not to be confused with smileys ☺

o Simplified Molecular Input Line Entry System

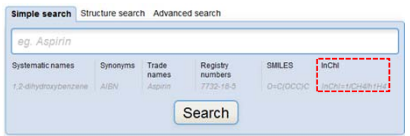
o Basically, summarises your molecule into one line of text

CCCCC Oc1ccccc1 c1(cc(nc(n1)Cl)NCC)N(C)C Cc1nc2c(ncn2C)nc1

o <http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>

Compound Database 1: ChemSpider
Simple Search - InChIs

o "Simple Search" – is already quite complicated!



o InChI = IUPAC *I*nternational *C*hemical *I*dentifier

o <http://www.iupac.org/inchi/> - the "official" website

InChI=1S/C6H4ClN5/c1-4-10-7-12-6(9)13-8(14-7)11-5(2)3/h5H,4H2,1-3H3,(H2,10,11,12,13,14)

InChI Key: MXWVACNOCSCGKHUHF84D6SA-2H3-6H2,1-2H3


InChI Key: VLKZOEYAKHREP-UHFFFAOYSA-N

InChI=1S/C6H6O/c7-6-4-2-1-3-5-6/h1-5,7H

InChI Key: ISWSIDIOOBJBQZ-UHFFFAOYSA-N

Compound Database 1: ChemSpider
Advanced Search

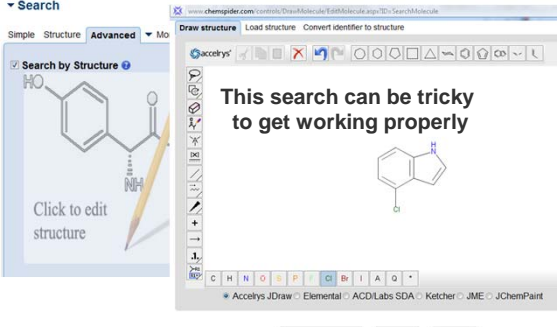
o "Simple Search" – is already quite complicated!



Be careful!

Compound Database 1: ChemSpider
Advanced Search - Structure

o "Simple Search" – is already quite complicated!



This search can be tricky to get working properly

Compound Database 1: ChemSpider
Advanced Search - Elements

Search by Elements

The elements below, if chosen, ☐ may ☒ must be present in compounds being found

CHNOPS FClBrI LiNaKCaZnFeCuCoNi Select All Deselect All Invert Selection

☒ H ☒ Be ☒ D ☒ T ☒ B ☒ C ☒ N ☒ O ☒ F ☒ Ne

The elements below, if chosen, ☐ may ☒ must not be present in compounds being found

CHNOPS FClBrI LiNaKCaZnFeCuCoNi Select All Deselect All Invert Selection

☒ H ☒ Be ☒ D ☒ T ☒ B ☒ C ☒ N ☒ O ☒ F ☒ Ne

☒ Li ☒ Be ☒ D ☒ T ☒ B ☒ C ☒ N ☒ O ☒ F ☒ Ne

☒ Na ☒ Mg ☒ Al ☒ Si ☒ P ☒ S ☒ Cl ☒ Ar

☒ K ☒ Ca ☒ Sc ☒ Ti ☒ V ☒ Cr ☒ Mn ☒ Fe ☒ Co ☒ Ni ☒ Cu ☒ Zn ☒ Ga ☒ Ge ☒ As ☒ Se ☒ Br ☒ Kr

☒ Rb ☒ Sr ☒ Y ☒ Zr ☒ Nb ☒ Mo ☒ Tc ☒ Ru ☒ Rh ☒ Pd ☒ Ag ☒ Cd ☒ In ☒ Sn ☒ Sb ☒ Te ☒ I ☒ Xe

☒ Cs ☒ Ba ☒ La ☒ Ce ☒ Pr ☒ Nd ☒ Pm ☒ Sm ☒ Eu ☒ Gd ☒ Tb ☒ Dy ☒ Ho ☒ Er ☒ Tm ☒ Yb ☒ Lu

☒ Fr ☒ Ra ☒ Ac ☒ Th ☒ Pa ☒ U ☒ Np ☒ Pu ☒ Am ☒ Cm ☒ Bk ☒ Cf ☒ Es ☒ Fm ☒ Md ☒ No ☒ Lr

Compound Database 1: ChemSpider
Advanced Search - Properties

Search by molecular formula, weight or masses

Search by Properties

☒ Molecular Formula: Exact match only
☐ min/max \pm +/-
☒ Molecular Weight: \pm 1.0 (example: 123 \pm 1)
☐ min/max \pm +/-
☒ Nominal Mass: \pm 0.1
☐ min/max \pm +/-
☒ Average Mass: \pm 0.001
☐ min/max \pm +/-
☒ Monoisotopic Mass: \pm 0.001
☐ min/max \pm +/-

Only the exact molecular formula is accepted (not a "fuzzy formula")
 The drop-down menus have an impressive range of options that are very useful for mass spectrometry measurements! (unfortunately they don't appear on screen shots...)

Compound Database 1: ChemSpider
Advanced Search - Calculated Properties

Search by calculated properties – unfortunately only ACD ones...

Search by Calculated Properties

☒ ACD/LogP: to
☒ ACD/LogD (pH 5.5): to
☒ ACD/LogD (pH 7.4): to
☒ Rule Of 5: to
☒ Number of Hydrogen Bond Acceptors: to
☒ Number of Hydrogen Bond Donors: to
☒ Number of Freely Rotatable Bonds: to
☒ Polar Surface Area: to
☒ Molar Volume: to
☒ Refractive Index: to
☒ Boiling Point: to
☒ Flash Point: to
☒ Density: to
☒ Surface Tension: to

The most relevant ones for EDA are likely to be the top 3
 Martin will tell you more about this...

Compound Database 1: ChemSpider
An example record

<http://www.chemspider.com/Chemical-Structure/2169.html> - Atrazine

Records are "compact" by default: if you want to see more, click around

Search term: atrazine (Found by approved synonym)

Atrazine

ChemSpider ID: 2169
 Molecular Formula: C6H9ClN3
 Average mass: 215.68304 Da
 Monoisotopic mass: 215.093781 Da
 Systematic name: 6-Chloro-N-ethyl-N'-isopropyl-1,3,5-triazine-2,4-diamine
 SMILES and InChI
 Cite this record

Names and Identifiers
 ChemSpider Searches
 Properties

Compound Database 1: ChemSpider
An example record

<http://www.chemspider.com/Chemical-Structure/2169.html> - Atrazine

Of most interest for us: typically the properties

Properties

Experimental data | Predicted - ACD/Labs | Predicted - EPI Suite | Predicted - ChemAxon

Data supplied by datasources and users.

Experimental Physchem Properties

Melting Point: 175 °C Tokyo Chemical Industry Ltd
 Boiling Point: 263 °C Tokyo Chemical Industry Ltd
 Decomposes NIOSH
 Specific Gravity: 1.19 NIOSH
 LogP: 2.632 Vitas-M
 Solubility: 0.003% NIOSH
 Vapor Pressure: 0.000003 mmHg NIOSH

Predicted Physchem Properties

Compound Database 1: ChemSpider
An example retrieval – with symmetry

Epicatechin – three entries, different stereochemistries

ID	Structure	Molecular Formula	Molecular Weight	# of Data Sources	# of References	# of PubMed	# of RSC
65230		<chem>C15H11O6</chem>	290.2681	52	104	1919	259
1166		<chem>C15H11O6</chem>	290.2681	33	1603	0	259
158494		<chem>C15H11O6</chem>	290.2681	14	17	0	0

Compound Database 2: PubChem
 Happy Birthday PubChem – 8 years, made a huge difference!
<http://pubchem.ncbi.nlm.nih.gov/>

PubChem can be a little tricky sometimes

Chemical structure search

PTO

More than 5 million structures from SCRPDB are now available in PubChem, aggregating data from more than 300,000 USPTO patents between the years of 2001-2012. [see more...](#)

PubChem databases and services are now HTTPS compatible. [see more...](#)

The PUG REST 1.0 version is now available! Use simple HTTP requests to download customized output for PubChem records. [see more...](#)

PubChem reaches milestones on 8th day! +100M substances and +200M bioactivity outcomes from +200 data submitters! [see more...](#)

Compound Database 2: PubChem
 PubChem Compound (not Substance!)
<http://pubchem.ncbi.nlm.nih.gov/search/>

PubChem Compound
 Limits Advanced search

Search By: Name/Text, Identity, Substructure/Superstructure, Molecular Formula, 3D Conformer, Saved Search

Synonyms/Descriptors/MeSH term etc:

Search

Clear

This advanced search option was broken when I tried it...
 But not the option from the first website:

Compound Database 2: PubChem
 PubChem Compound – Advanced Search
 Similar to ChemSpider without so much clicking...

PubChem Compound
 Compound

Limits

Search Field Tags

Field: All Fields

Date (yyyy, yyyy-mm, or yyyy-mm-dd)

Create Date from to

Chemical Properties

MolecularWeight from to

XLogP from to

HydrogenBondDonorCount from to

HydrogenBondAcceptorCount from to

...and many more options

Compound Database 2: PubChem
 PubChem Search Example

PubChem Compound
 PubChem Compound atrazine

Save search Limits Advanced

Display Settings: Summary, 20 per page, Sorted by Default order

Send to:

Results: 1 to 20 of 23

1. **atrazine, Oligosaprim, Aktikon**
 MW: 215.683269 g/mol MF: C₆H₉ClN₃
 IUPAC name: 6-chloro-4-N-ethyl-2-N-propan-2-yl-1,3,5-triazine-2,4-diamin...
 CID: 2256
[Summary](#) [Similar Compounds](#) [Same Parent Connectivity](#) [Mixture/Component Compounds](#) [Publited](#) [MeSH Keyword](#) [Active in 8 of 736 BioAssays](#)

2. **Proxine, Atrazine-pendimethalin mixture, Atrazine mixture with pendimethalin**
 MW: 498.990920 g/mol MF: C₂₂H₃₄ClN₆O₄
 IUPAC name: 6-chloro-4-N-ethyl-2-N-propan-2-yl-1,3,5-triazine-2,4-diamin...
 CID: 3045434
[Summary](#) [Similar Compounds](#) [Mixture/Component Compounds](#)

3. **Gesagrim, Primagram, Primextra**
 MW: 499.476840 g/mol MF: C₂₃H₃₄Cl₂N₆O₃
 IUPAC name: 2-chloro-N-(2-ethyl-5-methylphenyl)-N-(1-methoxypropan-2-yl)...
 CID: 107844
[Summary](#) [Similar Compounds](#) [Mixture/Component Compounds](#)

Compound Database 2: PubChem
 PubChem Record Example

Atrazine - Compound Summary (CID 2256)

Also known as: Oligosaprim, Aktikon, Argosin, Atrazox, Atrazox, Atrazine, Atrazin, Chlorazin, Panamin

Molecular Formula: C₆H₉ClN₃ Molecular Weight: 215.683269 [Molecular Weight](#) [Molecular Weight](#) [Molecular Weight](#)

A selective triazine herbicide. Irritation hazard is low and there are no apparent skin manifestations or other toxicity in humans. Acutely poisoned sheep and cattle may show muscular spasms, fasciculations, stiff gait, increased respiratory rates, adrenal degeneration, and congestion of the lungs, liver, and kidneys. *From: The Merck Index, 11th ed., Pharm. Major*

Table of Contents

Related Records

Use and Manufacturing

Patents

Environmental Fate and Exposure Routes

Exposure Standards and Regulations

Monitoring and Analysis Methods

Literature

Biological Interactions and Pathways

Biological Test Results

Classification

Chemical and Physical Properties

Expand all sub-sections

Identification

Check out this record for more – it's long!

Compound Database Searching
 What do you really want to achieve?

More chances to have a look and ask questions during the practice session

NIST Database
Additional Features

- Searches
 - Define the library you want to search (main, MS/MS, rep, ...)
 - Names, formula, CAS, exact mass, peaks, ...
 - MS/MS listing for those users – but not as many entries
- Plenty of additional features, but will not go into this today...

MassBank
Mixed spectra, web-based database
www.massbank.jp

Database Service
Statistics
Documents
Download
Manuals
About MassBank
Contact
Consortium Members
Site Map
User Restrictions

News
Sep 25, 2012 [Record Editor 2.1](#) was updated. [new](#)
Sep 21, 2012 [Document](#) page was updated. [new](#)
Sep 20, 2012 [MassBank](#) service will stop on Sep 27 for the server maintenance. [new](#)
Sep 03, 2012 [Manuals](#) page was updated. [new](#)
Aug 24, 2012 [Record Editor 2.1](#) was updated. [new](#)

Database Service
Spectrum Search
Quick Search
Peak Search
Substructure Search
Metabolite Identification
Spectral Browser
Batch Service
Browse Page
Record Index

MassBank is financially supported from [National Bioscience Database Center, Japan Science and Technology Agency \(2011-2013\)](#).
The [Mass Spectrometry Society of Japan](#) officially supports MassBank.
Please cite the article [\[DOI\]](#) when using MassBank.
Horai et al. 2010: DOI 10.1002/jms.1777

NORMAN MassBank
European MassBank – for environmental purposes
http://massbank.normandata.eu/MassBank/

NORMAN MassBank

Spectrum Search
Quick Search
Substructure Search
Browse Page
Peak Search
Spectral Browser
Record Index

NORMAN MassBank
European MassBank – for environmental purposes
http://massbank.normandata.eu/MassBank/

NORMAN MassBank

Spectrum Search
Quick Search
Substructure Search
Browse Page
Peak Search
Spectral Browser
Record Index

Search your spectrum through here
Search many other options here
Browse through the records here

MassBank Quick Search
Search compounds or peaks

massbank.normandata.eu/MassBank/QuickSearch.html

Quick Search

Home | Spectrum | Quick | Peak | Substructure | Browser | Browse | Index | MassBank ID: Go

☒ Search by Keyword ☐ Search by Peak

Compound Name
AND ☐ Exact Mass Tolerance 0.3
AND ☐ Formula (e.g. C8H7N5, C5H7N5, C5*)

Instrument Type
☐ EI
☐ EI-B
☐ EI-EBEB
☐ GC-EIMS
☐ GC-EI-TOF
☒ ESI
☒ CE-ESI-TOF
☒ ESI-IT-MS/MS
☒ ESI-ITFT
☒ ESI-QQ
☒ ESI-QqT-MS/MS
☒ ESI-QqQ-MS/MS
☒ FSI-QqTOF-MS/MS

Ionization Mode
☒ Positive ☐ Negative ☐ Both

MassBank Quick Search
Search compounds or peaks

Quick Search

Home | Spectrum | Quick | Peak | Substructure | Browser | Browse | Index | MassBank ID: Go

☐ Search by Keyword ☒ Search by Peak

Peak Data
279.086 22
289.086 107
290.118 14
293.086 959
292.113 162
293.054 34
579.169 37
580.179 15

m/z and relative intensities(0-999), delimited by a space.
[Example1](#) [Example2](#)

Cutoff threshold of relative intensities 5
Number of Results 20

Instrument Type
☐ EI
☐ EI-B
☐ EI-EBEB
☐ GC-EIMS
☐ GC-EI-TOF
☒ ESI
☒ CE-ESI-TOF
☒ ESI-IT-MS/MS
☒ ESI-ITFT
☒ ESI-QQ
☒ ESI-QqT-MS/MS
☒ ESI-QqQ-MS/MS
☒ FSI-QqTOF-MS/MS

Ionization Mode
☒ Positive ☐ Negative ☐ Both

MassBank Record Index

Have a look at what is in MassBank:

Contributor:

Instrument Type:
(Orbitrap = ITFT)

Compound Name: watch out for numbers...

Contributor	Instrument Type	Compound Name
NORMAN/EMPOMASS (2,081)	GC-ELMS (2,997)	A (1,185)
CASMI (42)	LC-ESI-QTOF (2,750)	B (1,122)
Eawag (944)	LC-APCI/ESI (10)	C (1,372)
UFZ (2,509)	LC-ESI-QTOF (1,034)	D (1,559)
Waters (2,993)	LC-ESI-QTOF (1,034)	E (406)
Kyoto Univ. (5,629)	LC-ESI-QTOF (1,034)	F (421)
Kyoto Univ. (185)	LC-ESI-QTOF (1,034)	G (783)
Univ. Toronto (2,028)	LC-ESI-QTOF (1,034)	H (429)
Univ. Toronto (253)	LC-ESI-QTOF (1,034)	I (657)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	J (12)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	K (223)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	L (1,254)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	M (996)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	N (1,005)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	O (388)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	P (3,687)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	Q (184)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	R (287)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	S (941)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	T (1,249)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	U (887)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	V (128)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	W (3)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	X (51)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	Y (5)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	Z (72)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	1.9 (3,844)
Univ. Toronto (340)	LC-ESI-QTOF (1,034)	Others (191)

Searching MassBank

What do you want to achieve?

- MassBank is a mixed database
 - Be careful to pick and mix your settings
 - Adjust the thresholds to the data you have
 - ...but also to the data in MassBank
 - Tolerance is in Da (i.e. one mass unit)
- Quality of spectra is quite mixed
 - ...as is the information provided in them

More in the practice session!

Lucky Last Database - METLIN

A very brief introduction

Make sure you try this in the practice session

METLIN

Statistics

- # Metabolites: 64,750
- # High Resolution MS/MS Spectra: 52,904
- # Metabolites w/ High Resolution MS/MS: 10,657

Functionality

- Single & Batch Precursor ion (m/z) searching
- Single & Multiple Fragment ion (m/z) searching
- Neutral Loss searching
- De Novo Fragment Characterization

METLIN – Simple Search

[http://metlin.scripps.edu/...](http://metlin.scripps.edu/)

METLIN: Metabolite and Tandem MS Database

Simple | Advanced | Batch | Fragment | Multiple Fragment | Neutral Loss | MS/MS Spectrum Match | Unknowns

Mass: Tolerance (s): 30 ppm

Charge:

Find Metabolites | Reset

Visual summary of Databases

- Compound databases (at least the free ones!)
 - ChemSpider
 - PubChem
- GC-ESI-MS Databases
 - Demonstration using NIST
- Online MS and MS/MS databases
 - MassBank
 - METLIN

Practice Session

Compound Databases
Mass Spectral Databases

MassBank

MassBank normal