

ACTIONS
MARIE CURIE

EDA EMERGE

eawag
aquatic research

Structure Elucidation Beyond Compound Databases *The Trickier Cases*

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

The trickier cases

- What if the answer isn't in the database? How can we be sure?
 - Structure Generation, Spectral Classifiers and MOLGEN-MS
 - Uploading generated structures to MetFrag / MetFusion
- Candidate filtering *versus* concept of "consensus scoring"
 - Combining all criteria for candidate selection
- Examples
 - GC-MS Unknown from Blue Rayon sample
 - Transformation Products: special cases

eawag
aquatic research

Cases where the compound isn't in a database...

Ideas from the audience?

- Database too small / incorrect "focus"
 - Looking for pesticides using a pharmaceutical database?
 - Looking for LC-MS compounds using NIST GC-MS database
- Database incomplete (yes, even 25,000,000 isn't enough!)
- You have the wrong mass? Wrong formula?
 - Have you done adduct detection? Isotope peaks? Multiple charge?
 - Included enough elements in the formula?
 - Are you searching the neutral or charged species (which is needed)?
- Is your compound a transformation product?
 - Some common TPs are in the databases; many aren't
- Do you have a truly "unknown unknown"?

eawag
aquatic research

Size of Databases vs. Compound Spaces

Structures:

- elements C, H, N, O
- at least 1 C atom
- standard valences
- no charges
- no radicals
- no stereoisomers
- only connected structures

Source: A. Kerber, R. Laue, M. Meringer, C. Rücker: Molecules in Silico: Potential versus Known Organic Compounds. MATCH 54 (2), 301-312, 2005.

eawag
aquatic research

Size of Compound Space

Number of compounds is huge: need to select candidates

600,000,000 highly probable FORMULAS (7 Golden Rules)

700,000 formulas in PubChem
...and 25,000,000 compounds...

T. Kind & O. Fiehn, Bioanalytical Reviews, 2010, 2:23-60.
C. Hildebrandt, S. Wolf & S. Neumann, Journal of Integrative Bioinformatics, 2011, 8(2):157.

eawag
aquatic research

Spectral Interpretation – Substructure "Classifiers"

Fragments present/absent, losses → Substructure Classifiers^{1,2}

'M' peak & isotope pattern

¹Database Independent EI-MS Classifiers – Varmuza, K. et al. (1996) J. Chem. Inf. Comp. Sci. 36, 323-333
²NIST/EPA/NIH (2008). NIST Mass Spectral Library, NIST, USA; Stein, S. (1995) J. Am. Soc. Mass. Spec. 6, 644

CASE via Structure Generation and MOLGEN-MS

MOLGEN-MS*:

- EI-MS Interpretation
 - Molecular formula
 - Structures absent/present
- Structure Generation
 - All possible structures matching spectrum
- Fragment Prediction
 - Predict MS for generated structures
 - Compare with experimental spectrum

**Developed by Prof. A. Kerber, M. Meringer and co-workers, Maths Department, University of Bayreuth*

Kerber, A., Laue, R., Meringer, M. and Varmuza, K. (2001) *Adv. Mass Spec.* 15, 939-940. www.molgen.de

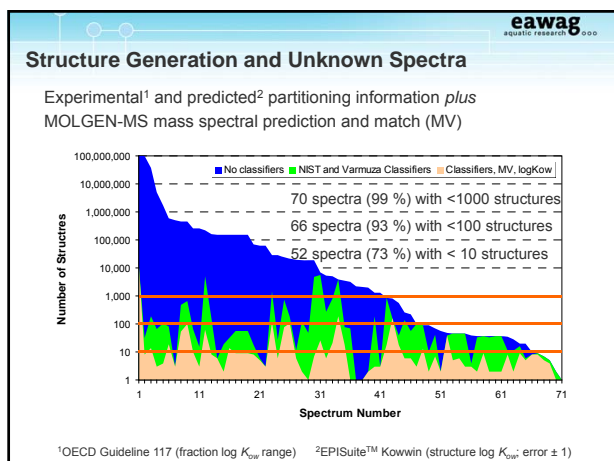
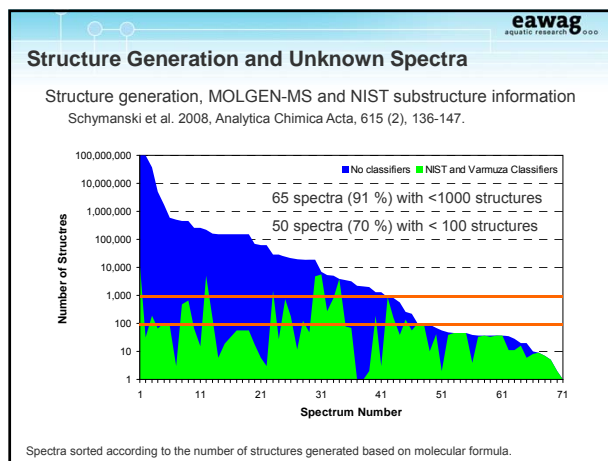
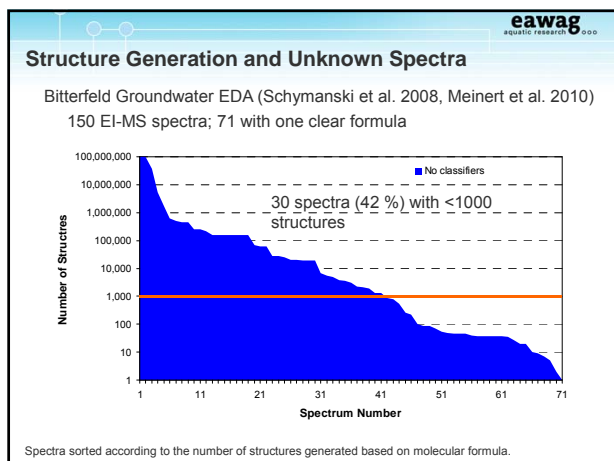
Candidate Reduction via Substructure Information

$C_4H_2Cl_4$
(40 molecules)

'no C=C in ring'
(25 molecules)¹

'C=C-C=C', 'no CH₂',
'no C=C in ring'
(4 molecules)²

¹MOLGEN-MS substructure information ²MOLGEN-MS + NIST substructure information



Additional Strategies for Structure Elucidation - RI

Retention behaviour used to confirmation identifications¹

- Kovat's RI (KRI) – C₆ – C₃₆ alkanes
- Lee RI (LRI) – 2-5 ring PAHs

General Equation¹

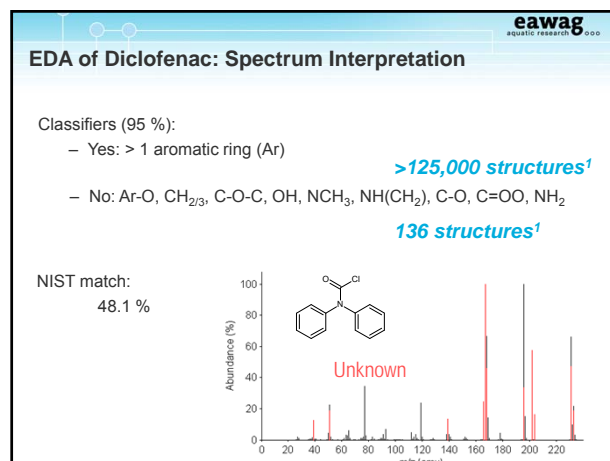
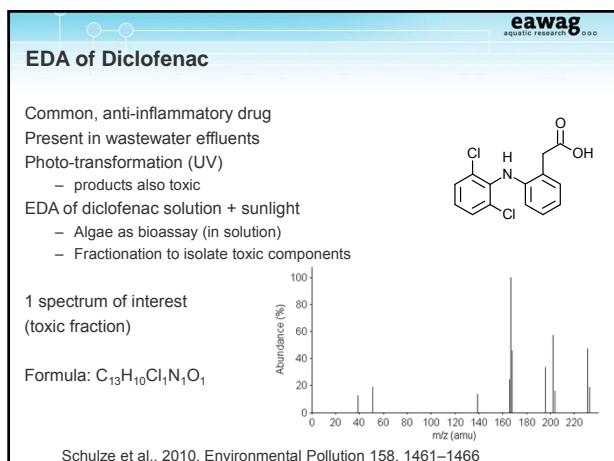
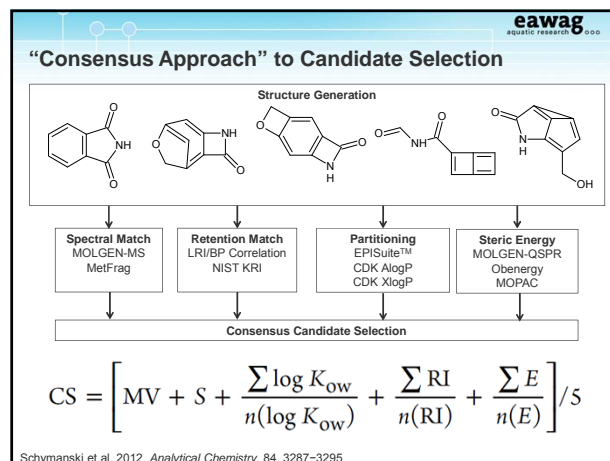
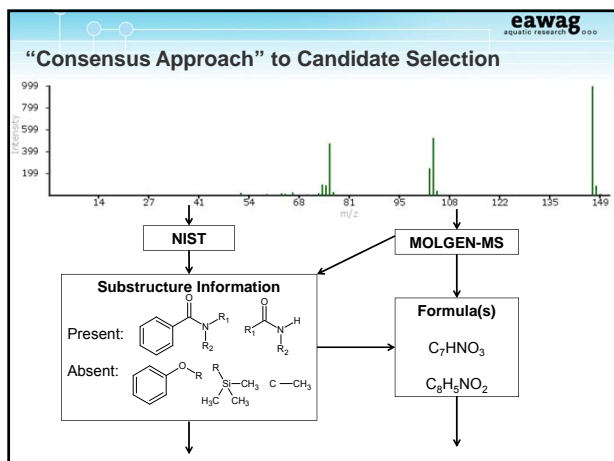
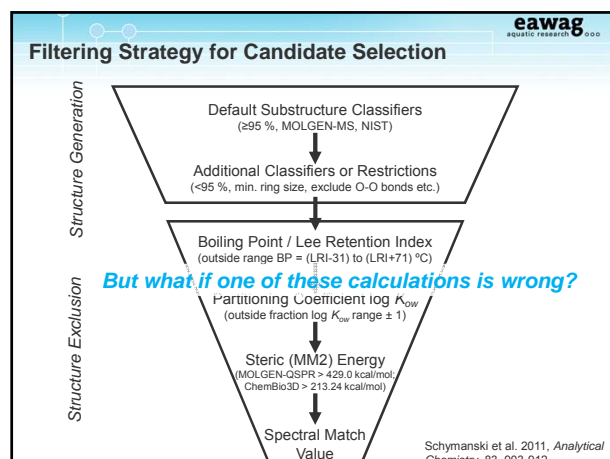
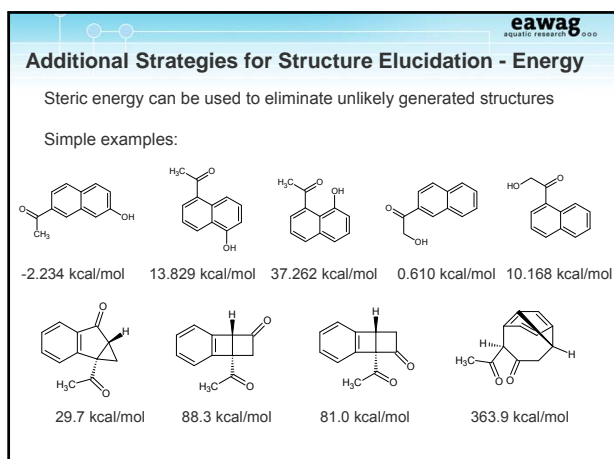
$$RI_x = 100 \left(n + \frac{T_n - T_x}{T_{n+1} - T_n} \right)$$

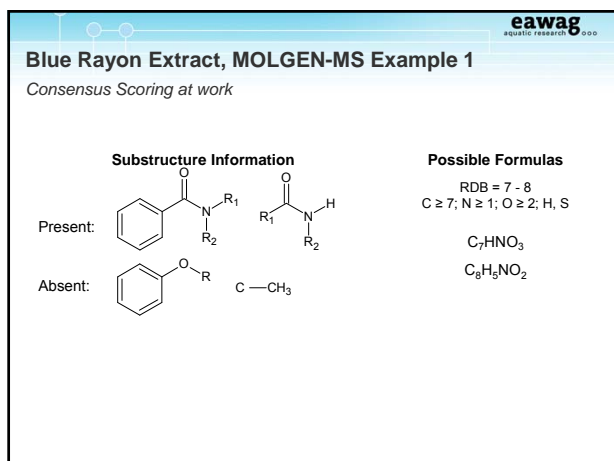
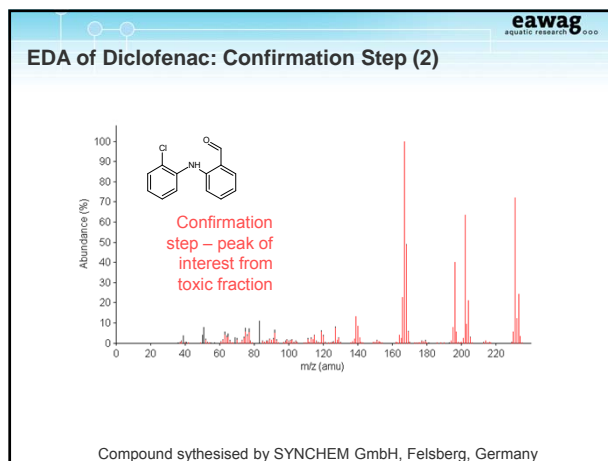
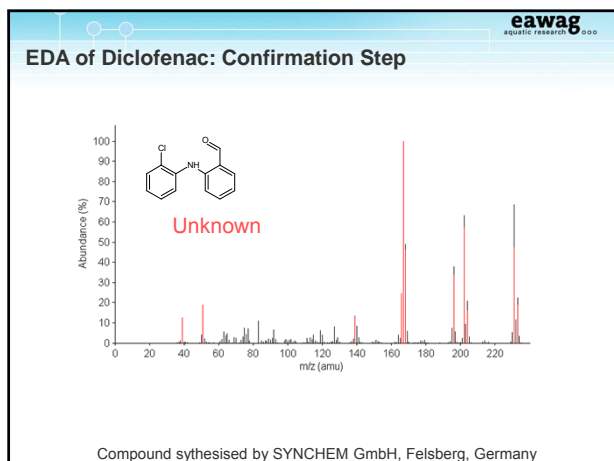
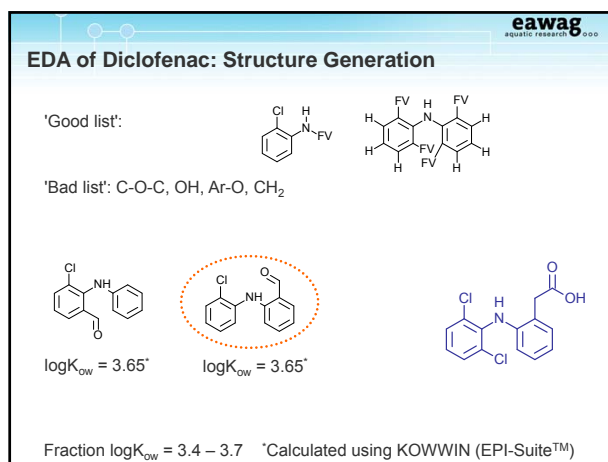
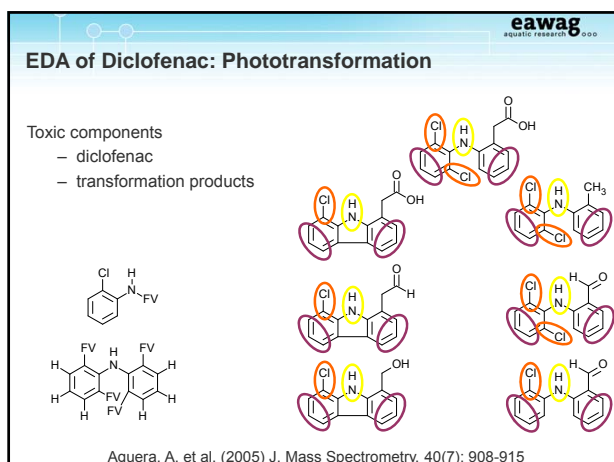
LRI/Boiling Point (BP) correlation for structures²

Original LRI/BP² $BP(^{\circ}C) = [(LRI-10), (LRI+50)]$

Modified LRI/BP³ $BP(^{\circ}C) = [(LRI-31), (LRI+71)]$
(including prediction errors)

¹Rostad and Pereira (1986) *J. High Res. Chrom. & Chrom. Commun.* 9 (6) 328
²Eckel & Kind (2003) *Anal. Chim. Acta* 494 (1-2) 235
³Schymanski, Meringer and Brack (2011) *Anal. Chem.*, 83, 903-912





Blue Rayon Extract, MOLGEN-MS Example 1

Consensus Scoring at work...

Structure	14	43	44	274
MV (%)	62.15	62.06	62.06	60.88
MetFrag S	1	0.695	1	0.932
CS	0.744	0.723	0.744	0.848

Blue Rayon Extract, MetFrag Example

MetFrag – ChemSpider Search

MetFrag Score and MS/MS Matching
<http://msbi.ipb-halle.de/MetFrag>

Partitioning Coefficient log K_{ow}
 (outside fraction log K_{ow} range ± 1)
 EPISuite, ACD Labs via www.chemspider.com

Acid/Base Behaviour
<http://sparc.chem.uga.edu/sparc>

Mutagenicity
 J. Kazius *et al.* J. Med. Chem. 2005, 48(1): 31

Blue Rayon Extract, MetFrag Example

Signal: $[M+H]^+ = 273.1236$, retention time 19.4 min

- $C_{15}H_{16}N_2O_3$ identified using MOLGEN-MSMS¹
 - 2902 candidates in MetFrag with ChemSpider
 - Present in positive mode ionisation, not in negative mode
 - Neutral fraction only, log $K_{ow} = [1.2 - 3.2]$

Gallampoio *et al.* in prep. ¹M. Meringer *et al.* MATCH (2011) 65:259. Images from <http://msbi.ipb-halle.de/MetFrag/>

Blue Rayon Extract, MetFrag Example

Possible Candidates: Predicted Properties

- Very high MetFrag Score (1, 0.987 – ranks 1 & 2)
- Predicted log K_{ow} within fraction range
- All potentially present in neutral fraction
- Structures support non-detection in negative mode

Gallampoio *et al.* in prep.

Blue Rayon Extract, MetFrag Example

Additional Evidence:

- $m/z[M+H]^+ = 273.1236$ NOT present in basic or acidic fractions
 - Can only elute in neutral fraction
- MS/MS only two peaks => symmetrical fragmentation

... plus many potential isomers
 ... unlikely to be mutagenic

Gallampoio *et al.* in prep.

Summary: The Trickier Cases
 Identification Using Mass Spectrometry and Calculated Properties

GC-EI-MS and Structure Generation

- Method applied successfully to identify two unknown compounds
 - Improvements to steric energy criteria needed
- Challenge to purchase top candidates of structure generation

LC-MS/MS and Database Searching

- Databases do not contain all compounds
- Predicted properties essential for candidate selection
 - ...but identification is still hit and miss
 - Complex prediction for multi-functional groups

Finally...

- We urgently need more standard spectra in databases and for training sets
 - NORMAN MassBank: <http://massbank.normandata.eu/MassBank/>

ChemSpider The free chemical database

PubChem Structure Search

MetFrag

MOLGEN

Practice Session
 The Tricky Cases

MetFusion

MassBank

MassBank *norman*