Algorithmic challenges in mass spectrometry

Anna Gambin

Institute of Informatics,
University of Warsaw

outline

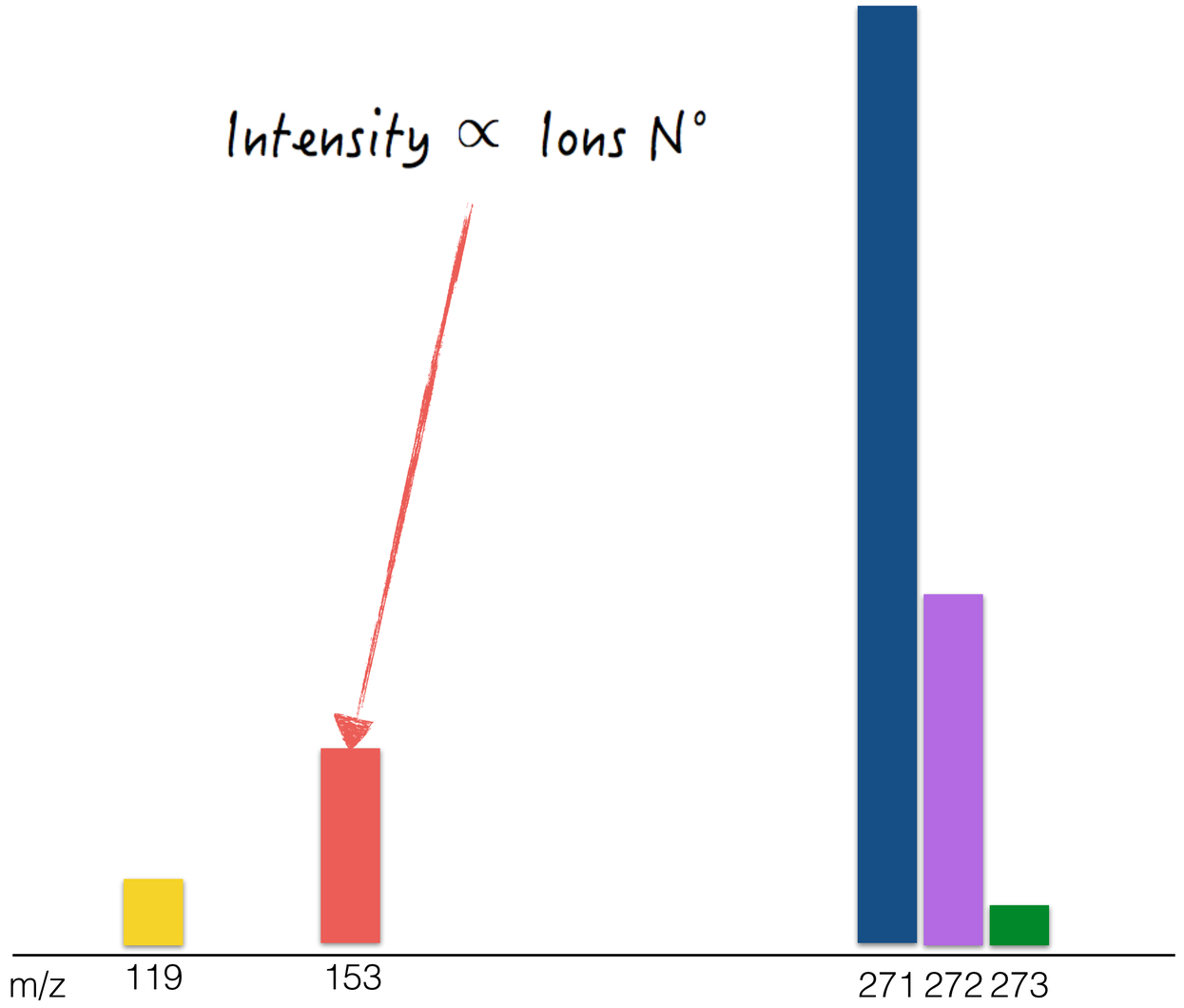
I. **modelling isotopic distribution**
   aggregated structure: BRAIN algorithm
   fine structure: ISOSPEC algorithm

II. **Markov processes:** modelling **fragmentation**

III. **optimal transport in spectroscopy**

Mass spectra

- mass spectrometer =
  = highly precise
  weighing scale
- compounds are
  frequently
  fragmented
- even a single
  compound has a
  complicated signal



Chemical compounds are made of different isotopes

13.0033 - 12 = 1.0033 [Da]

Element	Isotope	Extra Neutrons	Mass [Da]	Probability
Carbon	^{12}C	0	12	0.9893
	^{13}C	1	13.0033	0.0107
Hydrogen	^{1}H	0	1.0078	0.999885
	^{2}H	1	2.0141	0.000115
Nitrogen	^{14}N	0	14.0031	0.99632
	^{15}N	1	15.0001	0.00368
Oxygen	^{16}O	0	15.9949	0.99757
	^{17}O	1	16.9991	0.00038
	^{18}O	2	17.9992	0.00205
Sulfur	^{32}S	0	31.9721	0.9493
	^{33}S	1	32.9714	0.0076
	^{34}S	2	33.9679	0.0429
	^{36}S	4	35.9671	0.0002

2

3

4

elements have different numbers of stable isotopes

differences in frequencies of observation

isotopes of *different* elements differ in *mass differences*

32.9714 - 31.9721 = 0,9993 [Da]

mathematical model of mass spectra

product of multinomial distributions

Assume

1) variants of isotopes of atoms are **independent**

2) elements **vary in abundances** of isotopes

$$P(^{12}C_{c_0}\ ^{13}C_{c_1}\ ^{1}H_{h_0}\ ^{2}H_{h_1}\ ^{14}N_{n_0}\ ^{15}N_{n_1}\ ^{16}O_{o_0}\ ^{17}O_{o_1}\ ^{18}O_{o_2}\ ^{32}S_{s_0}\ ^{33}S_{s_1}\ ^{34}S_{s_2}\ ^{36}S_{s_4}) =$$

$$\binom{c}{c_0, c_1}\mathcal{P}(^{12}C)^{c_0}\mathcal{P}(^{13}C)^{c_1}\binom{h}{h_0, h_1}\mathcal{P}(^{1}H)^{h_0}\mathcal{P}(^{2}H)^{h_1}\binom{n}{n_0, n_1}\mathcal{P}(^{14}N)^{n_0}\mathcal{P}(^{15}N)^{n_1} \times$$

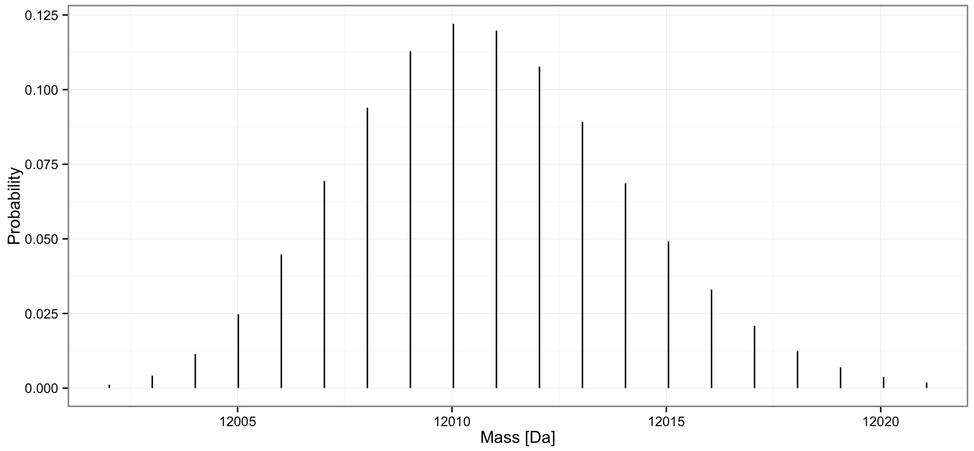
$$\binom{n}{o_0, o_1, o_2}\mathcal{P}(^{16}O)^{o_0}\mathcal{P}(^{17}O)^{o_1}\mathcal{P}(^{18}O)^{o_2}\binom{s}{s_0, s_1, s_2, s_4}\mathcal{P}(^{32}S)^{s_0}\mathcal{P}(^{33}S)^{s_1}\mathcal{P}(^{34}S)^{s_2}\mathcal{P}(^{36}S)^{s_4}$$

frequencies of isotopes

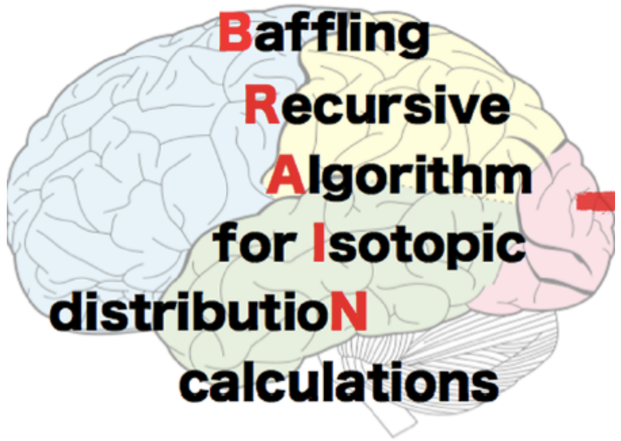
IUPAC

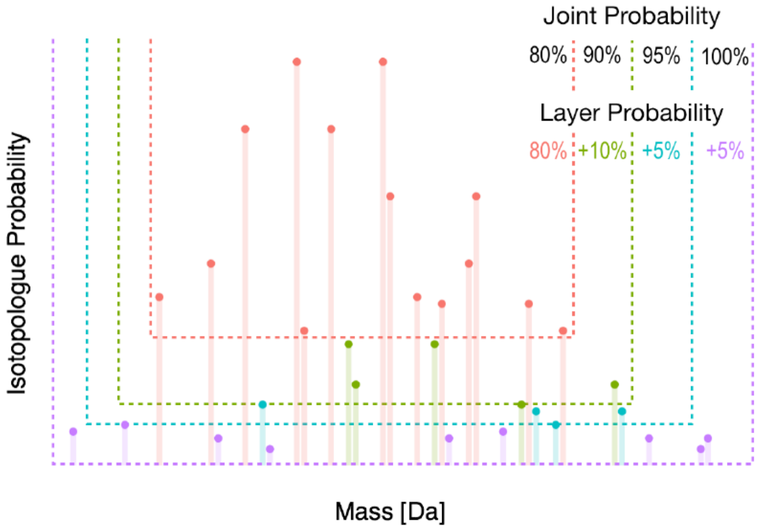
INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

mathematical model of mass spectra



low resolution problem:
aggregated isotopic structure





high resolution problem:
fine isotopic structure

UNIVERSITAS VARSOVIENSIS

ISOSPEC

The Fine Isotopic Structure Calculator

aggregated isotopic distribution

we **group** together variants with **the same number of additional neutrons**

for molecular formula $C_v H_w N_x O_y S_z$ consider polynomial:

$$Q(I;\, v,\, w,\, x,\, y,\, z) = (P_{C_{12}} I^0 + P_{C_{13}} I^1)^v \times (P_{H_1} I^0 + P_{H_2} I^1)^w$$
$$\times (P_{N_{14}} I^0 + P_{N_{15}} I^1)^v \times (P_{O_{16}} I^0 + P_{O_{17}} I^1 + P_{O_{18}} I^2)^y$$
$$\times (P_{S_{32}} I^0 + P_{S_{33}} I^1 + P_{S_{34}} I^2 + P_{S_{36}} I^4)^z$$

$$n = v + w + x + 2y + 4z$$

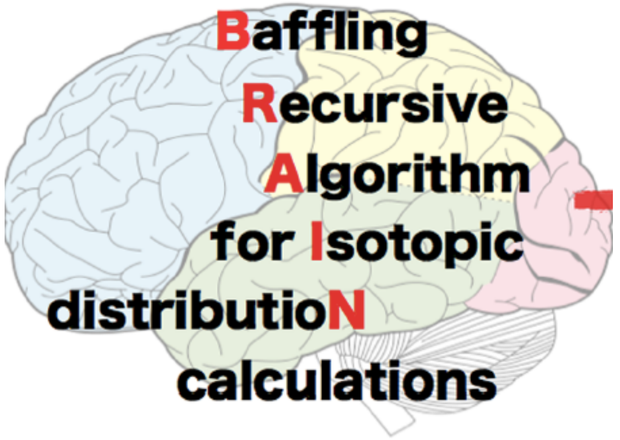
$$Q(I; v, w, x, y, z) \equiv \sum_{j=0}^{n} q_j I^j$$

frequencies of isotopes, e.g.

$P_{C_{12}} = 98.93\%$ and $P_{C_{13}} = 1.07\%$

probability of peak with j additional neutrons

algorithm



to compute coefficients of polynomial:

$$Q(I; v, w, x, y, z) \equiv \sum_{j=0}^{n} q_j I^j$$

determine its roots: $r_C = -\frac{P_{C_{12}}}{P_{C_{13}}}, \quad r_H = -\frac{P_{H_1}}{P_{H_2}}, \quad \text{and} \quad r_N = -\frac{P_{N_{14}}}{P_{N_{15}}}\,.$ $\qquad r_O, \bar{r}_O = \frac{-P_{O_{17}} \pm \sqrt{P_{O_{17}}^2 - 4P_{O_{16}}P_{O_{18}}}}{2P_{O_{18}}}$

apply the recurrent formula (follows from Newton-Girard theorem and Viete's formulae)

$$q_j = -\frac{1}{j}\sum_{l=1}^{j} q_{j-l}\psi_l$$

where

$$\psi_l = v(r_C)^{-l} + w(r_H)^{-l} + x(r_N)^{-l} + \quad (r_O)^{-j} + (\bar{r}_O)^{-j}$$
$$(r_{S,1})^{-j} + (\bar{r}_{S,1})^{-j} + (r_{S,2})^{-j} + (\bar{r}_{S,2})^{-j}$$

complexity: quadratic; **exact** values calculated

fine isotopic distribution

$$\mathcal{P}(^{16}\mathrm{O}) = \frac{4}{9} \quad \mathcal{P}(^{17}\mathrm{O}) = \frac{3}{9} \quad \mathcal{P}(^{18}\mathrm{O}) = \frac{2}{9}$$

100 oxygen atoms

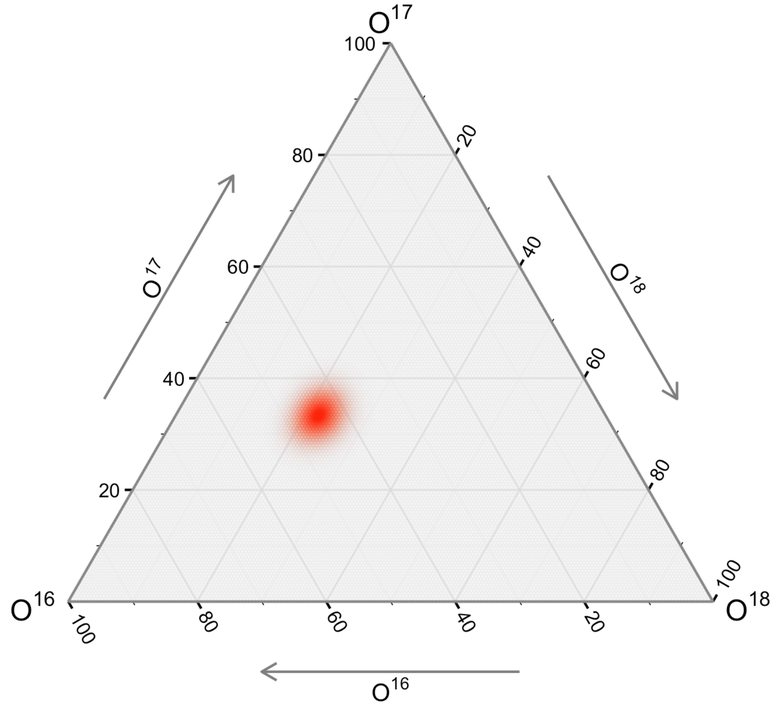
$$o_0 + o_1 + o_2 = 200$$

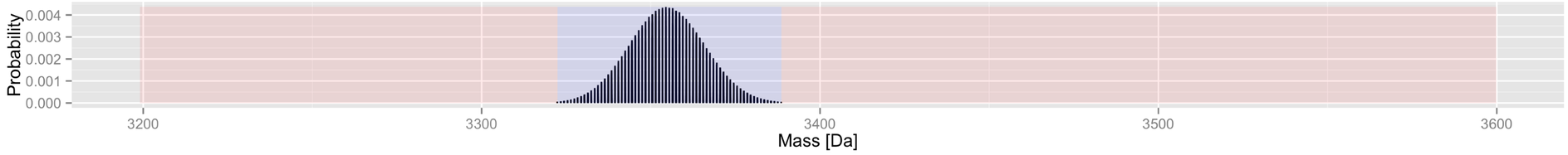
(not real world values!)

20301 variants

whereas 1043 bear 99% prob.

→ **huge number of isotopologues**

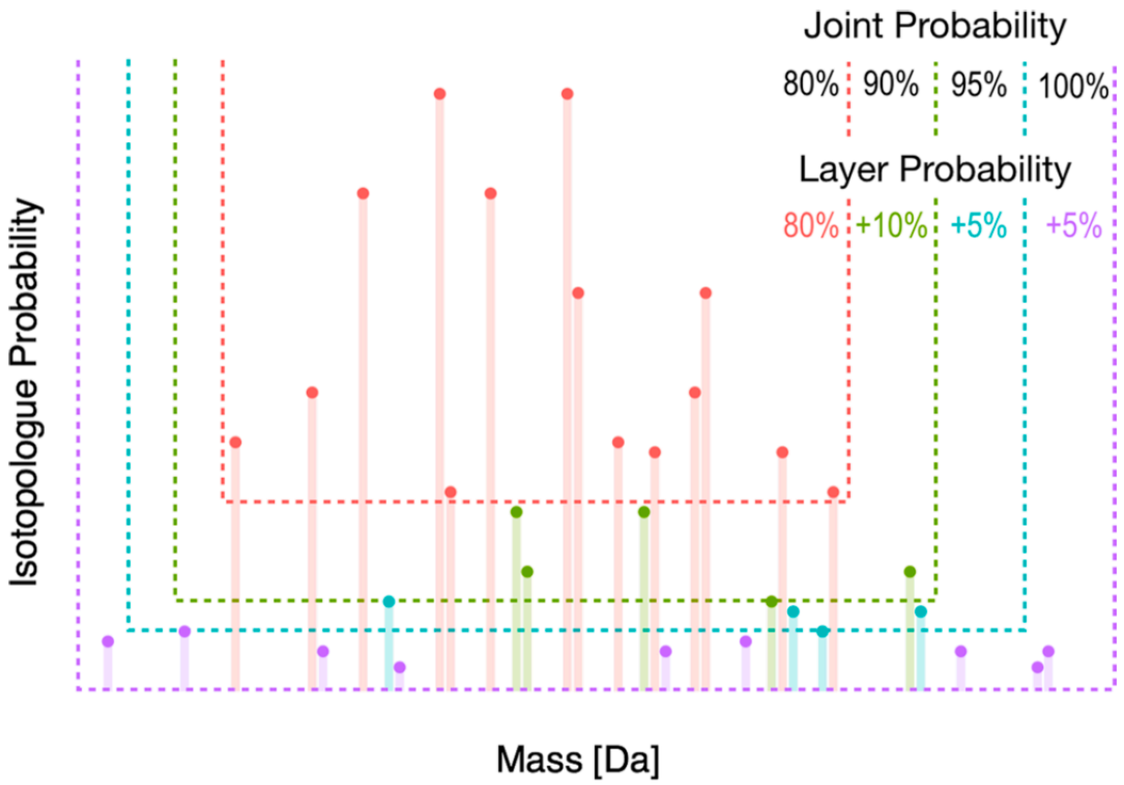




fine isotopic distribution

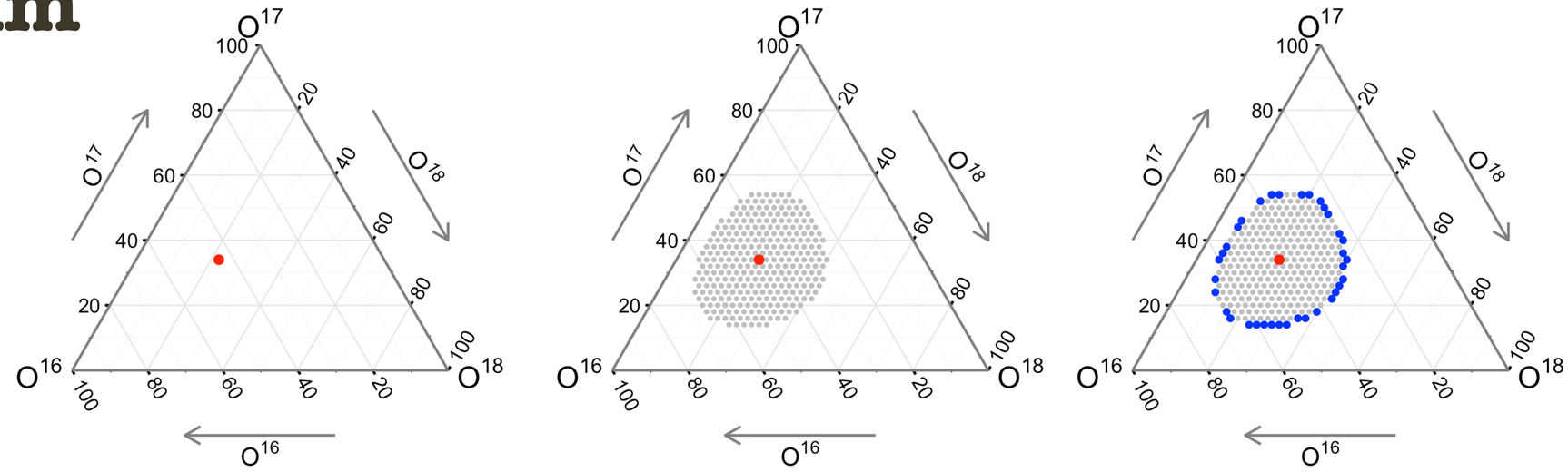
optimal p-set

smallest set of isotopologues that jointly surpass probability p



division of isotopic distribution into optimal p-sets: 80%, 90%, 95%, 100%

algorithm



To get the **optimal P-set:**

Find the **most probable variant**

while **Total Probability < P** :

Get **layer** of v so that $p > P(v) > 0{,}5\ p$ where

$p = P(v_{min\ previous\ layer})$

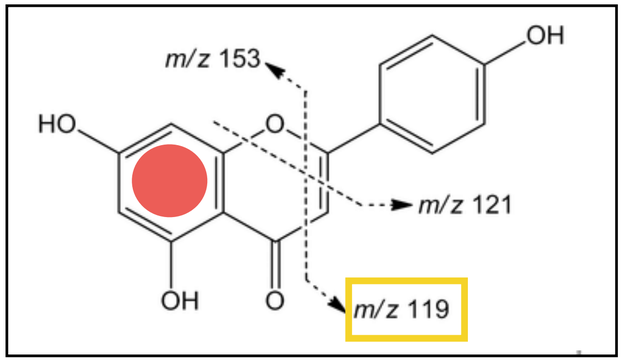
Trim the **least probable variants** from the last layer so that **Total Probability = P**

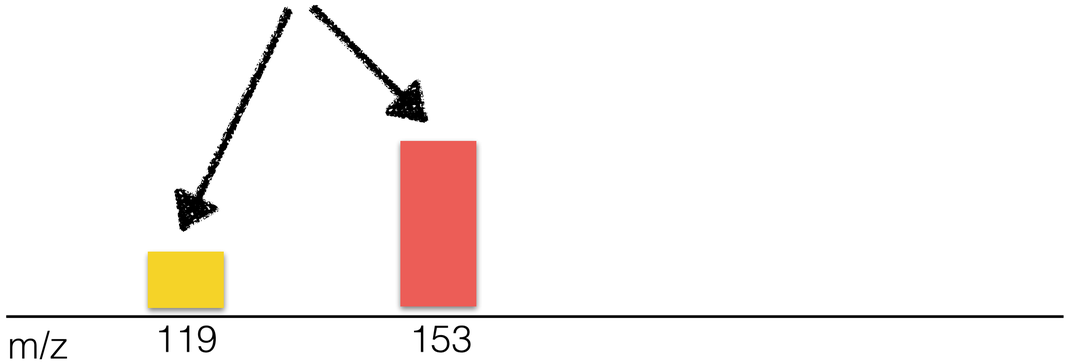
complexity:
O(n) in the total number of configurations

ISOSPEC

The Fine Isotopic Structure Calculator

II. **Markov processes:** modelling **fragmentation**



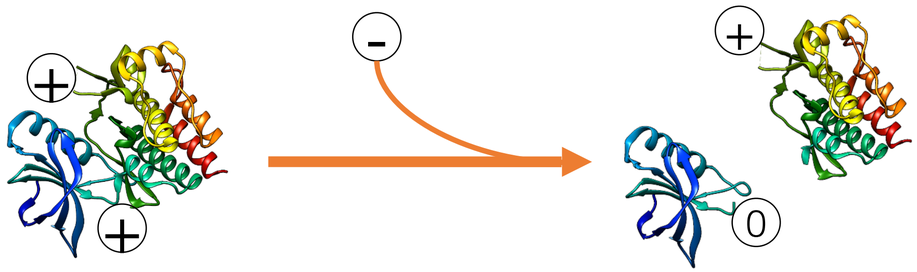


some **bonds** get easily broken

electron is transferred to the positively-charged protein or peptide, causing fragmentation along the peptide backbone

.. others **not**

modelling fragmentation: problem



inside MassSpec

Cleavage of protein backbone by a rapid neutralization of charge

ETD — main reaction, others = side reactions

PTR	$[M + nH]^{n+} + A^{\bullet -}$	$\longrightarrow$	$[M + (n\text{-}1)\,H]^{(n-1)+} + AH$
ETnoD	$[M + nH]^{n+} + A^{\bullet -}$	$\longrightarrow$	$[M + nH]^{(n-1)+\bullet} + A$
ETD	$[M + nH]^{n+} + A^{\bullet -}$	$\longrightarrow$	$[c + xH]^{x+} + [z + (n - x)H]^{(n-x-1)+\bullet} + A$

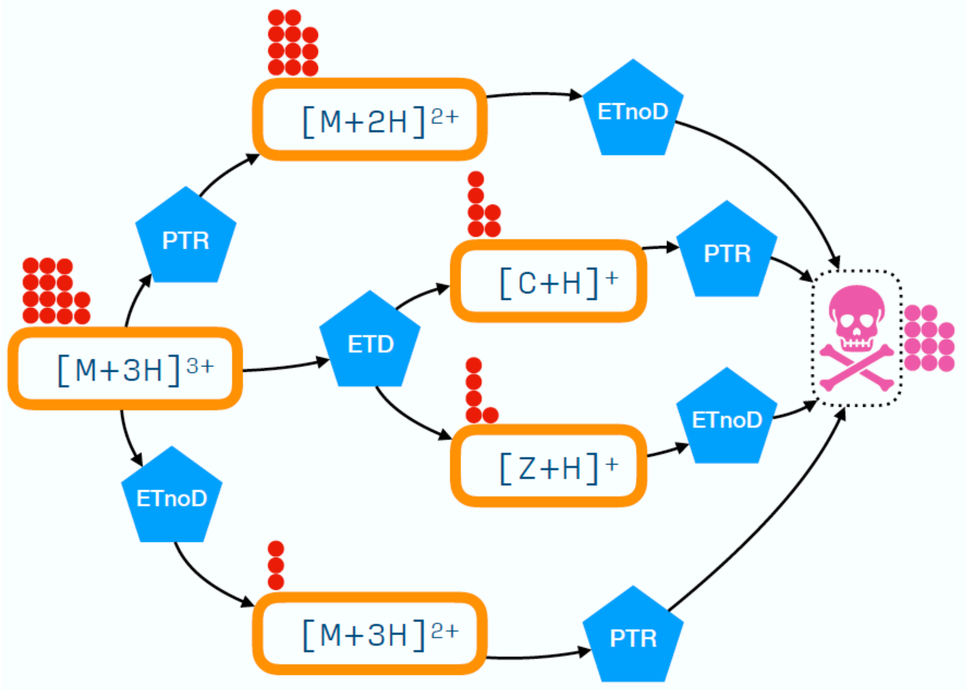
problem: for the set of biochemical reactions
determine their intensitie having observed the substrates

modelling fragmentation: solution

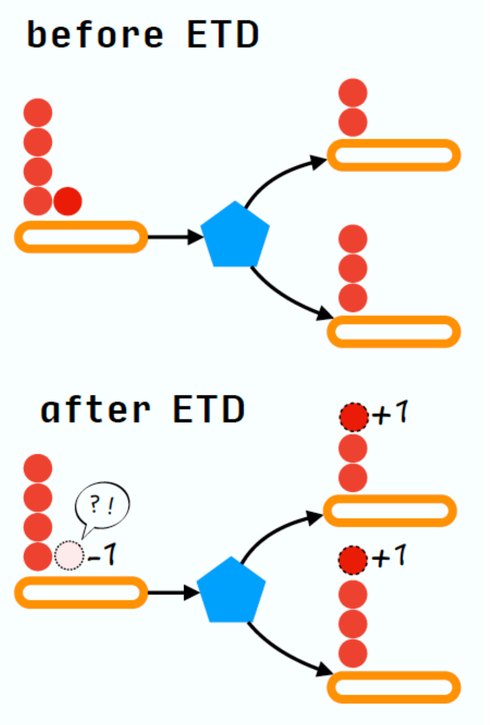
model the phenomena as Markov process describing the flow of particles through the fragmentation graph

calculate expectance in the model:
use ODE description for big population of particles

compare to observed data:
find intensities that best predict the observed data
by minimising the discrepancy (nonlinear optimisation)

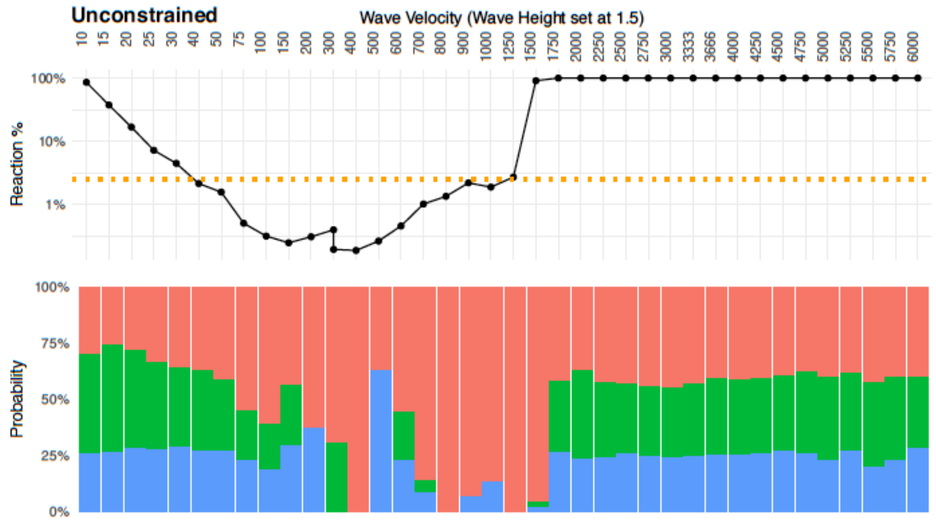


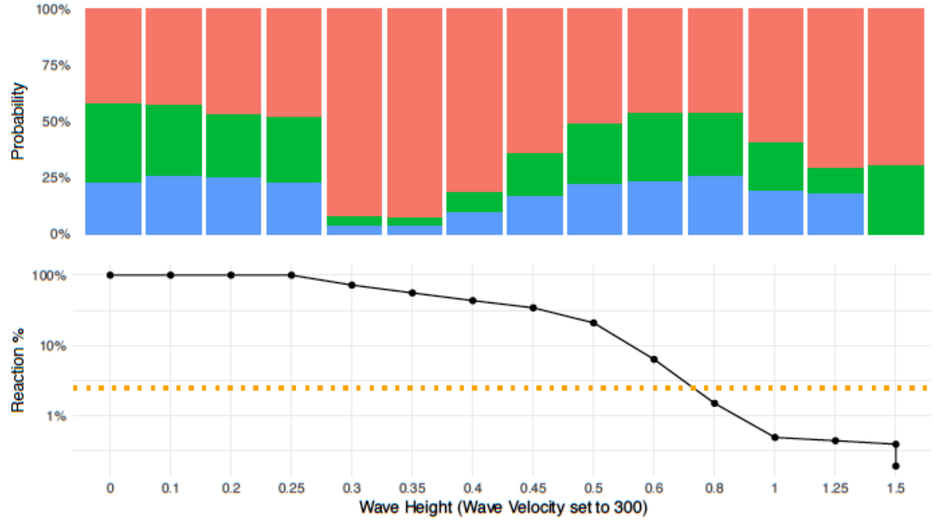
(a) A Petri Net



(b) ETD reaction

modelling fragmentation: results





proportions of PTR, ETD, ETnoD for different MS setup

PTR $[M + nH]^{n+} + A^{\bullet -} \longrightarrow [M + (n\text{-}1)\,H]^{(n-1)+} + AH$

ETnoD $[M + nH]^{n+} + A^{\bullet -} \longrightarrow [M + nH]^{(n-1)+\bullet} + A$

ETD $[M + nH]^{n+} + A^{\bullet -} \longrightarrow [c + xH]^{x+} + [z + (n - x)H]^{(n-x-1)+\bullet} + A$

III. optimal transport in spectroscopy

Eight tons of hope: world's strongest persistent magnet for NMR at ETH

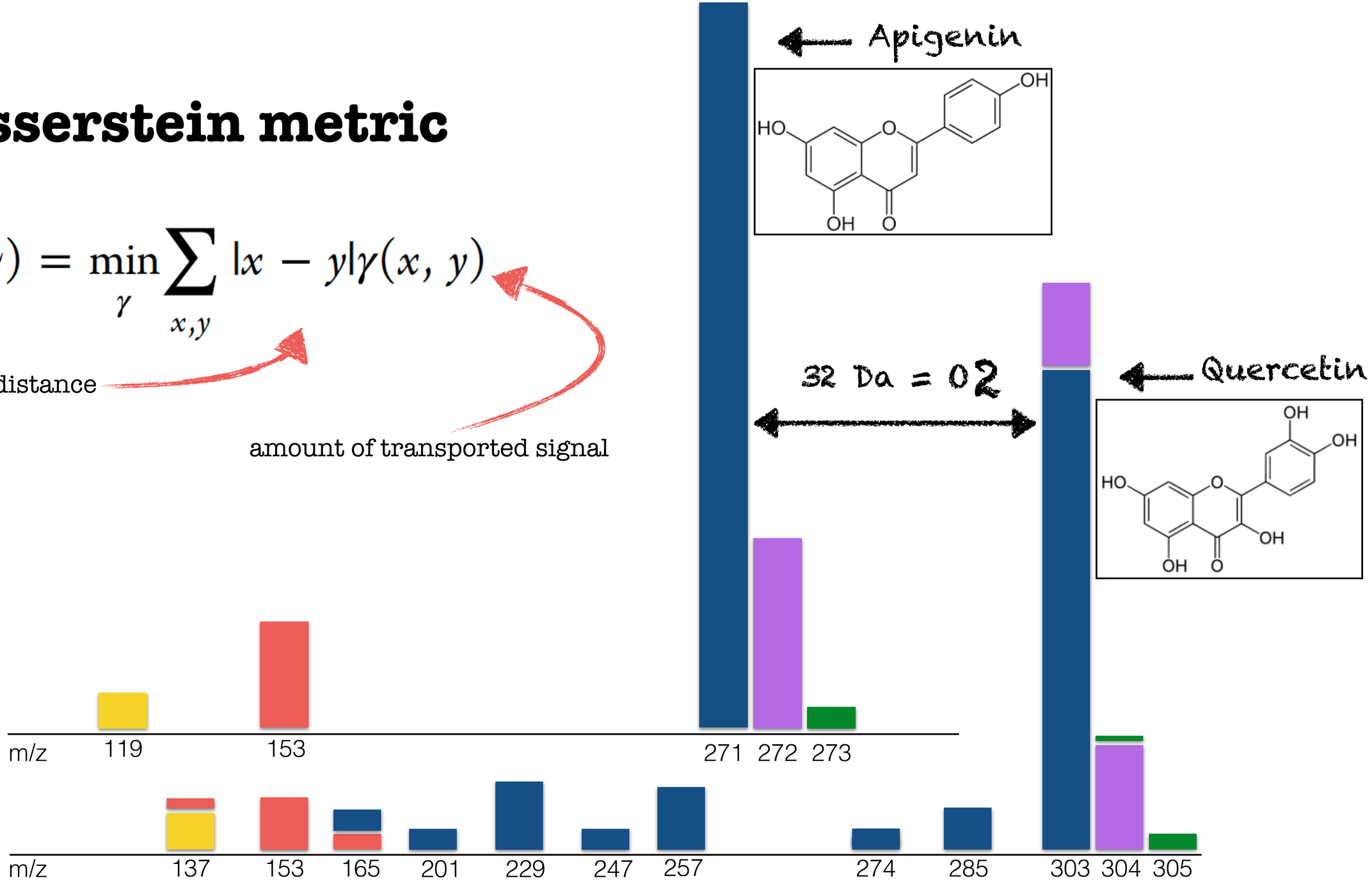
08.06.2020 by Julia Ecker

Wasserstein metric

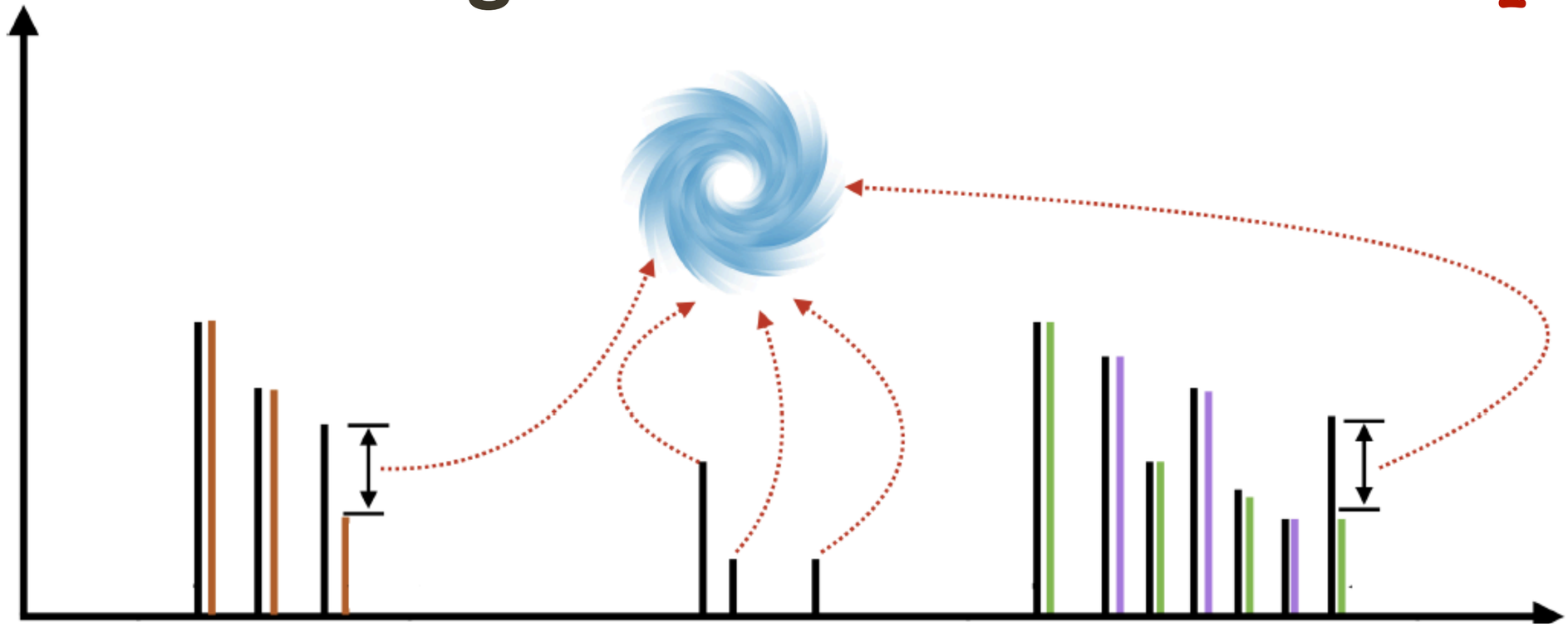
$$W(\mu, \nu) = \min_{\gamma} \sum_{x,y} |x - y| \gamma(x, y)$$

distance

amount of transported signal



Wasserstein regression: deconvolution of spectra



linear combination of simulated = model spectra → $\nu_p = p_1\nu_1 + \cdots + p_k\nu_k$

observed spectrum = model + noise → $\mu(x) = \nu_p(x) + \varepsilon(x)$

Wasserstein regression: find proportions → $p^* = \arg\min_p W(\mu, \nu_p)$

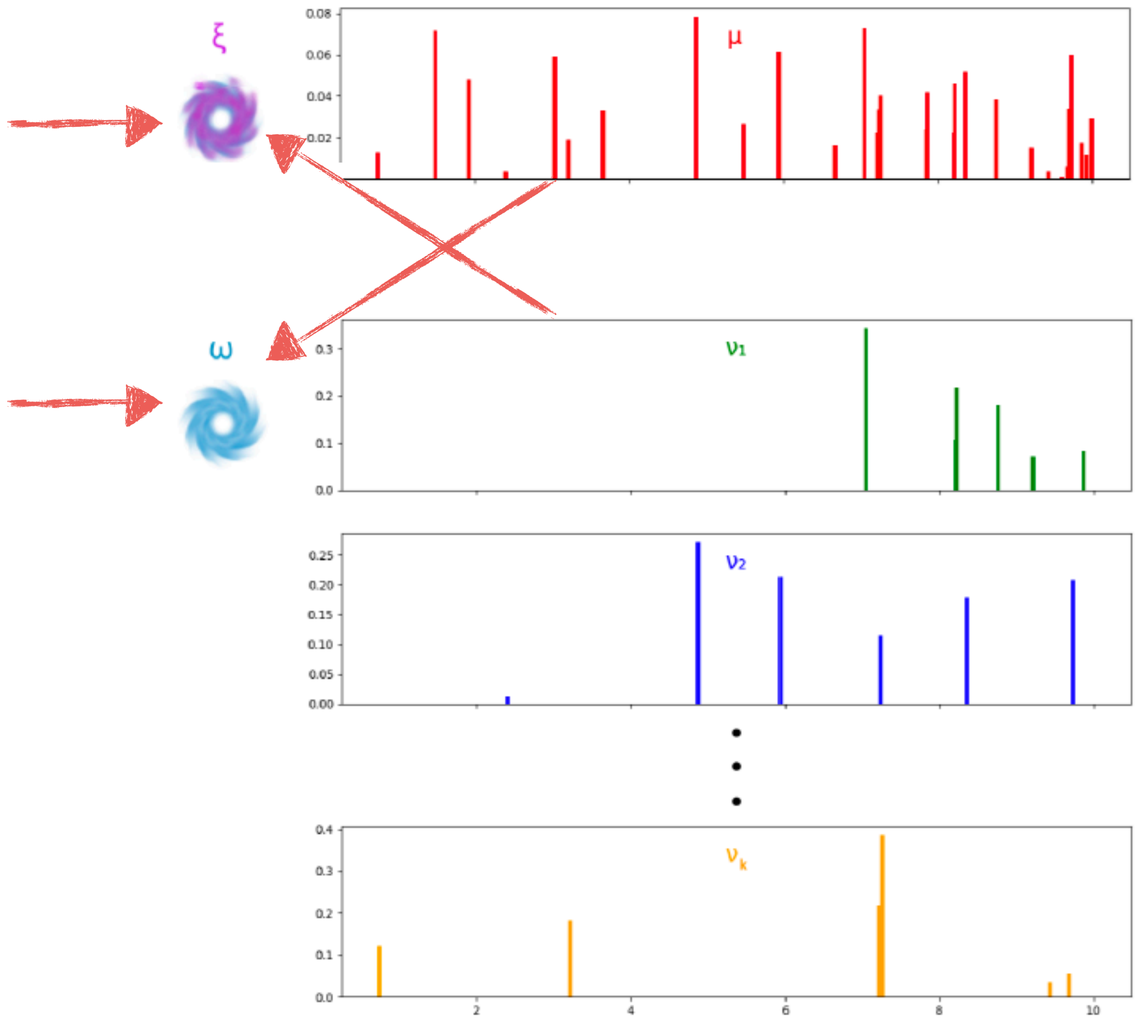
finally add vortex to transport noise → $p^* = \arg\min_p W(\mu, p_0\omega + \nu_p)$

working with two vortexes

$$p^* = \min_{p=(p_1,p_2,\ldots,p_k)} W(p_0\omega + p_1\nu_1 + p_2\nu_2 + \ldots + p_k\nu_k, (1-p_0')\mu + p_0'\xi),$$

to remove excess of hypothetical spectra in the model

to remove noise from experimental data



1 dimensional case is easy to calculate:

$$W(\mu,\nu) = \int_{\mathbb{R}} |M(t) - N(t)|dt$$

using

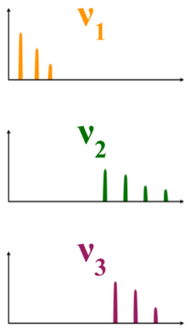
$$N(t) = \sum_{j=1}^{k} p_j N_j(t)$$

$$M(t)$$

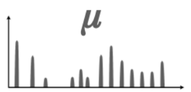
cumulative distribution functions

Magnetstein for NMR analysis

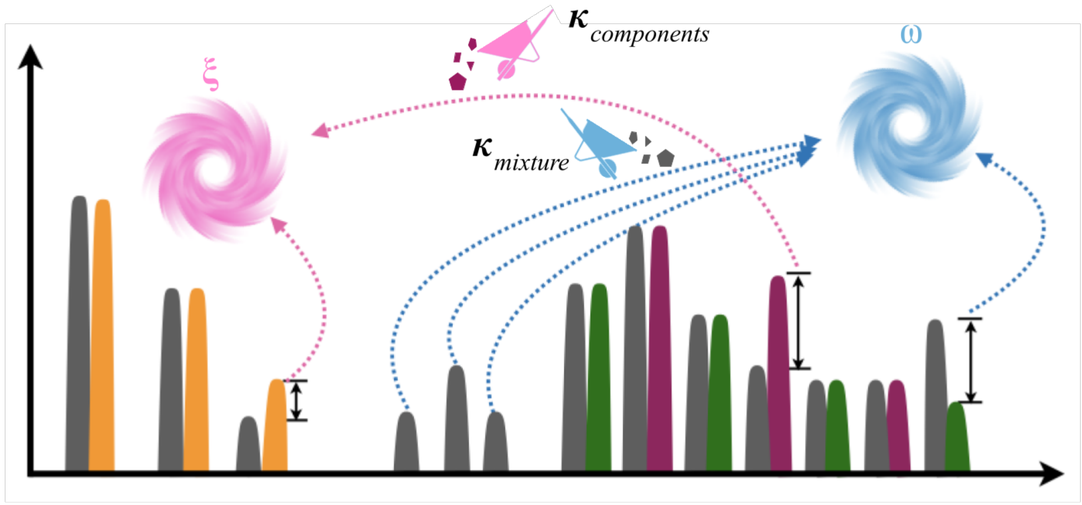
INPUT:
- components



- mixture

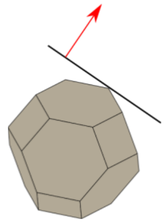


normalization



Regression with the Wasserstein distance:

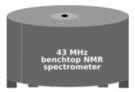
$$\min_{(p_0, p_0', p_1, p_2, p_3)} W_\kappa(p_0\omega + p_1\nu_1 + p_2\nu_2 + p_3\nu_3 , (1 - p_0')\,\mu + p_0'\xi)$$



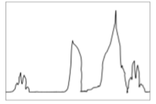
linear program

OUTPUT:
- proportions of components:
  p_1^*, p_2^*, p_3^*
- proportion of noise in mixture:
  p_0^*
- proportion of noise in components:
  $p_0'^*$

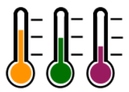
low resolution



lineshape distortion

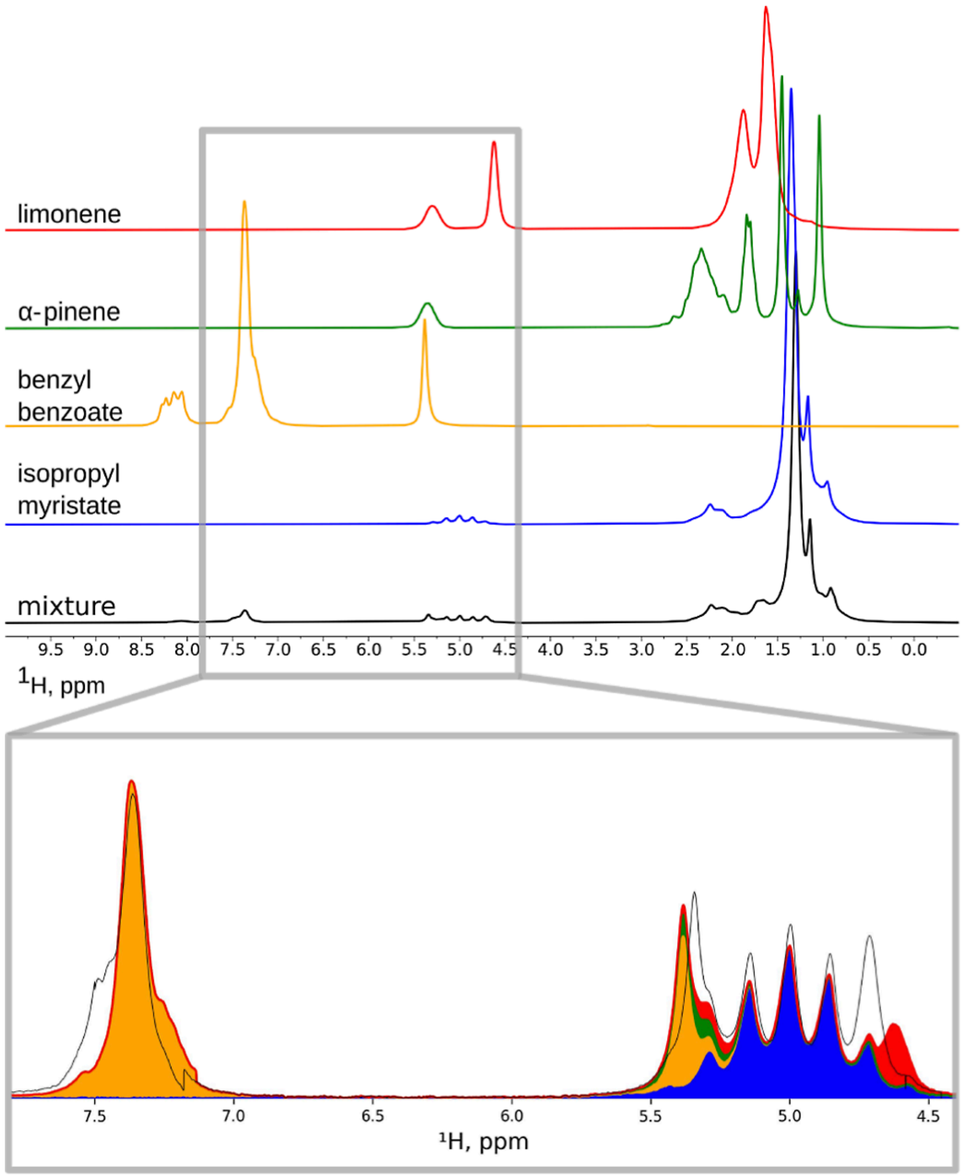


different temperatures

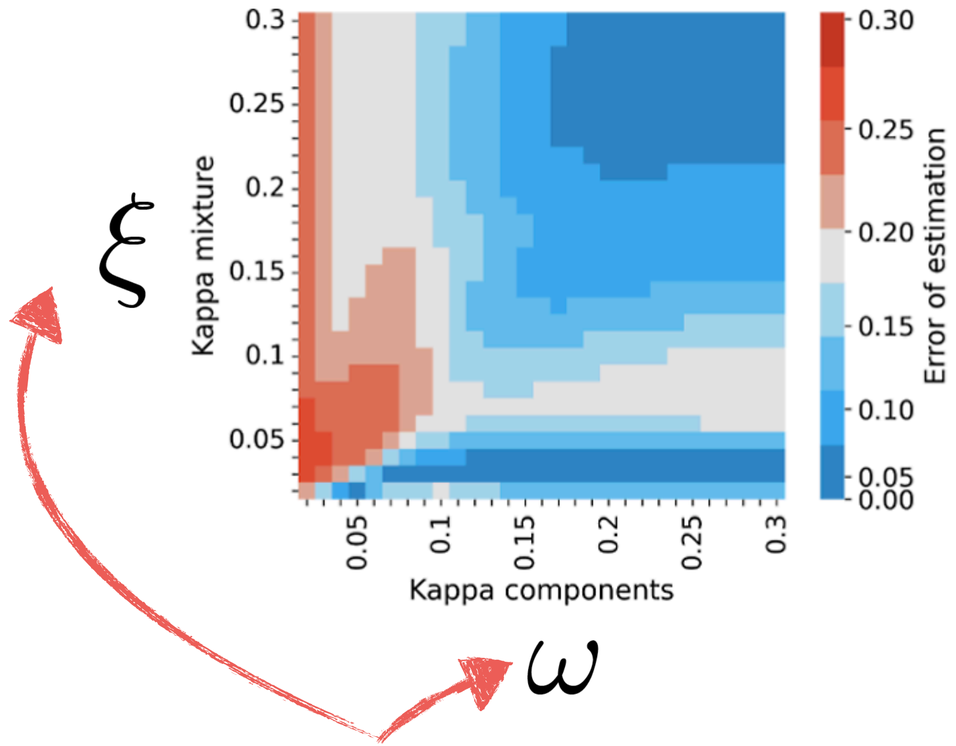


regression with Wasserstein distance can be formulated as linear program and solved efficiently

Magnetstein in action

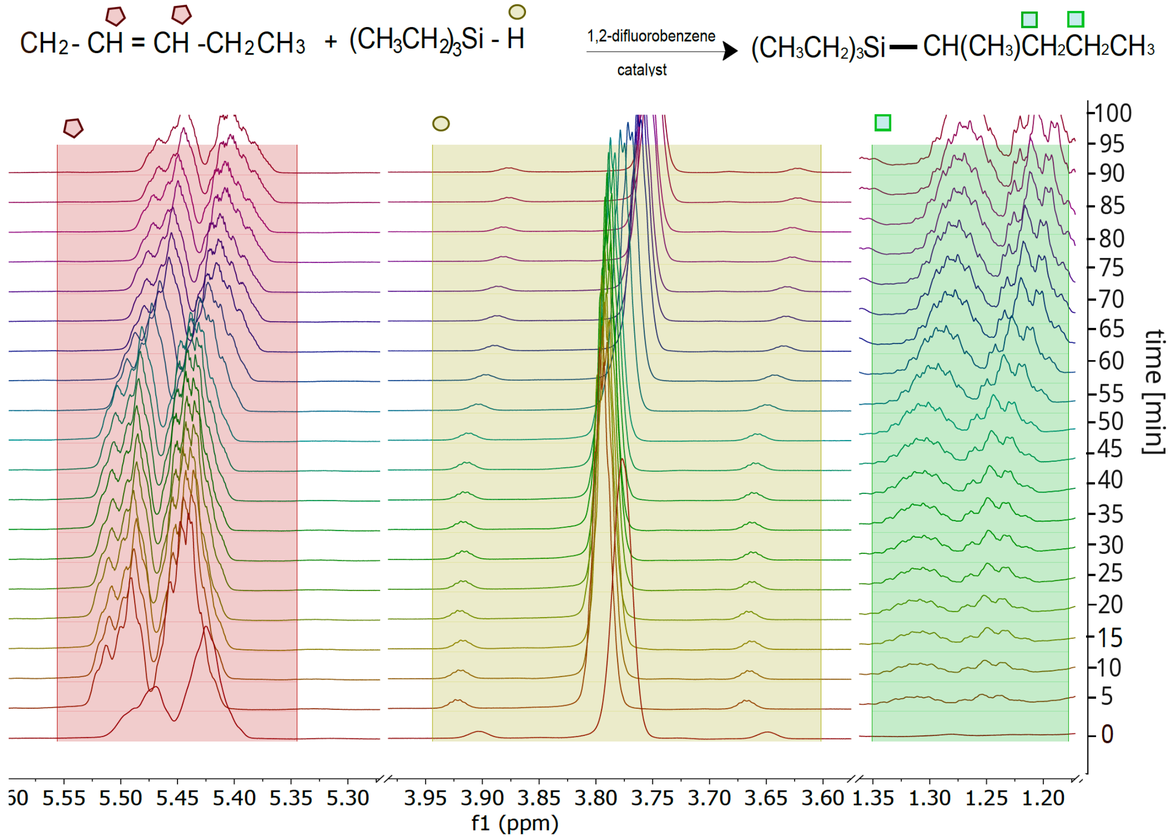


Magnetstein can quantitatively analyze difficult spectra with the estimation trueness an order of magnitude higher than that of commercial tools...



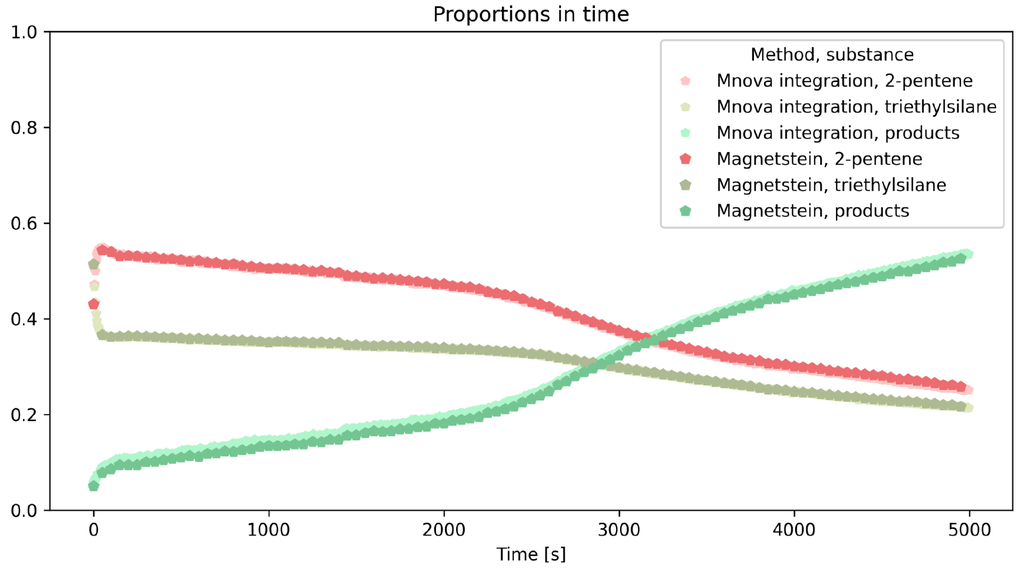
... having only two parameters with default values applicable to a broad range of experiments...

chemical reactions revisited

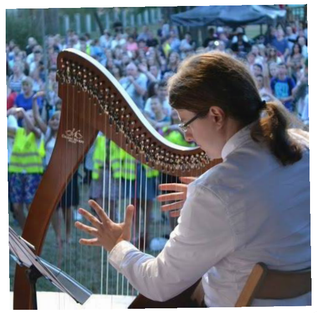


we can effectively quantify the components of a reacting mixture without a need for peak-picking

1. **solve Wasserstein regression**
2. **get a sequence of proportions in consecutive timepoints**
3. **infer about kinetics of the monitored reaction**



Many thanks to collaborators



Michał Startek

Dirk Valkenborg

Grzegorz Skoraczyński

Błażej Miasojedow

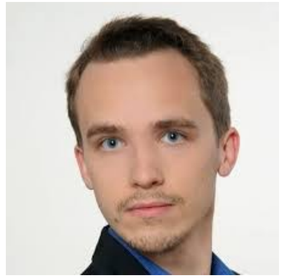
Frederik Lermyte

Mateusz Łącki

Krzysztof Kazimierczuk

Piotr Dittwald

Frank Sobott



Michał Ciach

Barbara Domżał

Alan Rockwood