University of Warsaw

Faculty of Mathematics, Informatics and Mechanics

Mateusz Krzysztof Łącki

Student no. 234189

Computational and Statistical Methods for Mass Spectrometry Data Analysis

PhD's dissertation

in COMPUTER SCIENCE

Supervisors:

Prof. Anna Gambin

Institute of Informatics, University of Warsaw

Dr Błażej Miasojedow

Institute of Applied Mathematics, University of Warsaw

September 2017

Supervisor's statement

Hereby I confirm that the presented thesis was prepared under my supervision and that it fulfils the requirements for the degree of PhD of Computer Science.

Date Supervisor's signature

Author's statement

Hereby I declare that the presented thesis was prepared by me and none of its contents was obtained by means that are against the law.

The thesis has never before been a subject of any procedure of obtaining an academic degree.

Moreover, I declare that the present version of the thesis is identical to the attached electronic version.

Date Author's signature

Abstract

Computational and Statistical Methods for Mass Spectrometry Data Analysis

This dissertation covers a series of related topics in the mathematical modelling of mass spectrometry data. The dissertation opens by a presentation of an optimal algorithm for the generation of the fine isotopic structure. We further show the applications of that algorithm to the problem of deconvoluting mixed isotopic signals, in two different ways. We also approach the problem of estimating the deep parameters of mass detectors, estimating the parameters of a function that relates the instrument-generated intensities to the numbers of ions. These solutions are applied to the problem of understanding Electron Driven reactions, whose principal aim is to induce ion fragmentation and, in that way, enhance the instrument's identification capabilities. Finally, we show how to apply the mathematical theory of reaction kinetics to estimate the reaction rates of the electron transfer reactions.

Metody obliczeniowe i statystyczne analizy danych ze spektrometrów masowych

Niniejsza rozprawa doktorska dotyczy szeregu tematyk z zakresu matematycznego modelowanie widm masowych. W pracy przedstawiam algorytm służący obliczeniom związanym z rozkładami izotopowymi cząsteczek. Algorytm ów wykorzystuję w problemie dekonwolucji mieszanek sygnałów ze znanych źródeł molekularnych, na dwa różne sposoby. Przedstawiam również sposób na wyznaczenie zależności pomiędzy zarejestrowanym sygnałem a liczbą jonów dla różnych detektorów jonów. Powyższe rozwiązania zostają również wykorzystane w celu dokładniejszego zrozumienia zasad działania fragmentacji jonów za pomocą transferu elektronu, która znacząco poszerza możliwości identyfikacji substancji. Pokazuję również sposób na wyestymowanie parametrów tych reakcji, wykorzystując w tym celu matematyczny model kinetyki reakcji.

Keywords

Mass Spectrometry, Isotopic Fine Structure, Estimation of Chemical Rates, Electron Transfer Dissociation, Deconvolution of Fine Isotopic Structures

Thesis domain (Socrates-Erasmus subject area codes)

11.3 Informatyka

Subject classification

I.6. Simulation and Modelling
J.2. Physical Sciences and Engineering

Tytuł pracy w języku polskim

Metody obliczeniowe i statystyczne analizy danych ze spektrometrów masowych

Contents

List of Figures

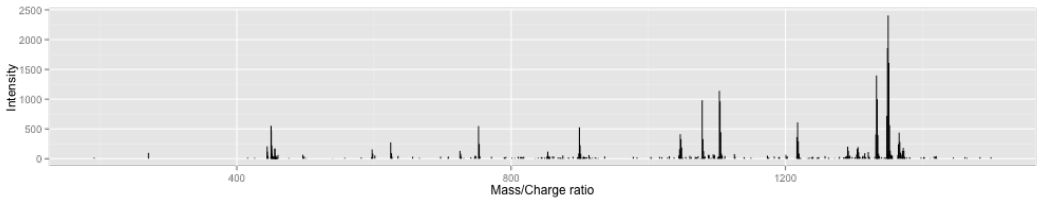
List of Tables

1 Introduction

"Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning."

— Winston Churchill

Mass Spectrometry is a subfield of the Analytical Chemistry that studies and develops instruments useful for analysing the molecular content of samples. The instruments, that are called mass spectrometers, have been developed by Joseph J. Thomson just before the First World War and used to study the presence of the isotopes of natural elements (Thomson, 1913). The output of a mass spectrometer – a *mass spectrum* – is a histogram: each bar has its own specific position in the mass-to-charge domain and height equal to that of the observed intensity. The intensities are usually assumed to be proportional to the number of ions. Below, we show how a mass spectrum might look like.



Let us study the case of human insulin, $C_{520}H_{817}N_{139}O_{147}S_8$, to see how complex can a signal be, even in case of one single source. the signal even generated by one source

Table 1.1: Masses and Frequencies of isotopes of elements that build up the proteins (Brand *et al.*, 2014).

Isotope	Mass	Frequency
^{1}H	1.0078	0.9999
^{2}H (D)	2.0141	0.0001
^{12}C	12.0000	0.9892
^{13}C	13.0034	0.0108
^{14}N	14.0031	0.9964
^{15}N	15.0001	0.0036
^{16}O	15.9949	0.9976
^{17}O	16.9991	0.0004
^{18}O	17.9992	0.0021
^{32}S	31.9721	0.9499
^{33}S	32.9715	0.0075
^{34}S	33.9679	0.0425
^{36}S	35.9671	0.0001

might be complex. To start with, all atoms in the above formula can assume one of multiple isotopic variants, from a set that is different for each element. Finding a particular isotope in nature is largely a random event. This does not mean that it is not predictable: when together in large numbers, they do follow many well studied patterns. The International Union for Pure and Applied Chemistry (IUPAC) is performing continuous measurements of the frequencies of natural isotopes. Table 1.1 summarizes a small fraction of their findings up till year 2014.

From the viewpoint of statistical modelling, the abundances reported in Table 1.1 represent the probabilities of finding a particular isotope for one given atom of a given element. By far the easiest way to construct a joint probability measure out of these marginals is to assume that isotopic variants of different atoms are mutually independent. This assumption dates back to the '60-ies (Beynon, 1960). Of course, in certain particular situations, such as isotopic labelling, one would certainly have to modify that asumption, as these could in introduce some non-trivial dependence between the isotopic variants. In general, however, it would be difficult to come up with a theoretical mechanism that could result in a significant departure from the independence asumption. Experimental findings does not reject that hypothesis too.

Mass spectrometer does not distinguish compounds with the same number of isotopic variants. For instance, if one of the 817 hydrogen atoms of human insuline is deuterium, then, based on mass spectrum alone, there is no way of telling which particular atom was

the heavier one. The observed signal depends only upon counts of different isotopic variants. Similarly to how chemical formulas such as $\mathrm{C_c\,H_h\,N_n\,O_o\,S_s}$ abstract from spatial composition, we may introduce a more detailed description of the isotopic content of a molecule, such as

$$\texttt{iso} = {}^{12}\mathrm{C}_{c_0}\,{}^{13}\mathrm{C}_{c_1}\,{}^{1}\mathrm{H}_{h_0}\,{}^{2}\mathrm{H}_{h_1}\,{}^{14}\mathrm{N}_{n_0}\,{}^{15}\mathrm{N}_{n_1}\,{}^{16}\mathrm{O}_{o_0}\,{}^{17}\mathrm{O}_{o_1}\,{}^{18}\mathrm{O}_{o_2}\,{}^{32}\mathrm{S}_{s_0}\,{}^{33}\mathrm{S}_{s_1}\,{}^{34}\mathrm{S}_{s_2}\,{}^{36}\mathrm{S}_{s_3}.$$

Above, $\mathrm{c_0}$ stands for the number of ^{12}C isotopes within the molecule, $\mathrm{c_1}$ – number of ^{13}C isotopes, and so on. This is essentially what we call an *isotopologue*. This definition coincides with that provided by the International Union of Pure and Applied Chemistry (McNaught and Wilkinson, 1997). The independence assumptions suggest that the probability of observing an isotopologue is that of a product of multinomial distributions, each for one element, or

$$p_{\texttt{iso}} = \binom{c}{c_0, c_1}\mathbb{P}\left({}^{12}\mathrm{C}\right)^{c_0}\mathbb{P}\left({}^{13}\mathrm{C}\right)^{c_1}\binom{h}{h_0, h_1}\mathbb{P}\left({}^{1}\mathrm{H}\right)^{h_0}\mathbb{P}\left(\mathrm{D}\right)^{h_1}\binom{n}{n_0, n_1}\mathbb{P}\left({}^{14}\mathrm{N}\right)^{n_0}\mathbb{P}\left({}^{15}\mathrm{N}\right)^{n_1}$$
$$\times\binom{o}{o_0, o_1, o_2}\mathbb{P}\left({}^{16}\mathrm{O}\right)^{o_0}\mathbb{P}\left({}^{17}\mathrm{O}\right)^{o_1}\mathbb{P}\left({}^{18}\mathrm{O}\right)^{o_2}\binom{o}{o_0, o_1, o_2}\mathbb{P}\left({}^{32}\mathrm{S}\right)^{s_0}\mathbb{P}\left({}^{33}\mathrm{S}\right)^{s_1}\mathbb{P}\left({}^{34}\mathrm{S}\right)^{s_2}\mathbb{P}\left({}^{36}\mathrm{S}\right)^{s_3}.$$

The mass of `iso` is given by mutliplying counts of different isotopes times their masses (Table 1.1), $m_{\texttt{iso}} = m({}^{12}\mathrm{C})c_0 + m({}^{13}\mathrm{C})c_1 + \cdots + m({}^{36}\mathrm{S})s_3$. The set of all pairs $(p_{\texttt{iso}}, m_{\texttt{iso}})$ that corresponds to one chemical formula $\mathrm{C_c\,H_h\,N_n\,O_o\,S_s}$ makes up the *isotopic fine structure*. The isotopic fine structure is typically used directly to model the signal in the instrument.

Isotopic Calculations

Assume that a chemical compound is made up of elements $\mathcal{E}$, each occuring as n_e atoms with possible i_e isotopic variants. Then, especially for bigger molecules with $n_e \gg 0$, it does not make any sense to generate the set of all isotopologues, as it comprises $\prod_{e\in\mathcal{E}}\binom{n_e+i_e-1}{n_e}$ elements. Using the Stirling's formula, we note that this is an expression of order $\mathcal{O}(\prod_{e\in\mathcal{E}} n_e^{i_e-1})$. For example, the isotopic fine structure of human insulin comprises more than 10^{14} different elements, requiring terabytes of storage. By far, it is also by far not the biggest known chemical compound. We are also bounded by the instrumental physics, such as detection thresholds, finite resolution, and limitiations in terms of numbers of ions inside the spectrometer[1]. All these factors severly limit the number of observed isotopologues, questioning the need to perform the above calculations. However, if take into considerations also the probability distribution, then only 1 716 most probable isotopologues represent 99% of all the probability mass, 5 403 represent 99.9% of probability, and 13 101 – 99.99%. This means, that it is beneficial to search for a smaller set of configurations with a probability coverage we could control.

[1] Ions have the same charge and repell each other, diverging from their predictable trajectories inside the mass spectrometer.

Definition. *For a given compound, the optimal P-set is the smallest set consisting of the most probable peaks of the fine isotopic distribution whose joint probability surpasses P. In case of more than one such set, we choose any representative of that class.*

Chapter 2 describes a particularly efficient and elegant way to quickly generate high coverage subsets of the isotopic fine structure – the `IsoSpec` algorithm (Łącki *et al.*, 2017b). The algorithm makes use of two fundamental features of the multinomial distibution: (1) measure concentration around its mean (Giannopoulos and Milman, 2000), and (2) unimodality (Finucan, 1964). Generally speaking, measure concentration implies that relative few configurations bear most of the probability mass. To define unimodality in the context of a discrete distribution, we must first define the relationship of neighbourhood between its configurations. With that at hand, it can be restated in terms of connectedness of the set of local probability maxima. The unimodality is crucial for the algorithm to work fast with minimal additional data structures: enumerating configurations of the multinomial distribution can be carried out by a *hill descent*. The two above properties do *tensorize*, i.e. are retained while considering products of distributions.

In Chapter 2 we prove that `IsoSpec` has the optimal, linear time complexity of isotopologues generation. In the proof, we apply the *Central Limit Theorem* to approximate the number of elements inside an optimal P-set by

$$M = \frac{q_{\chi^2(k)}(P)^{\frac{k}{2}}}{C} \frac{\pi^{k/2}}{\Gamma(k/2+1)} \sqrt{\prod_{e\in\mathcal{E}} \left(n_e^{i_e-1} \prod_{j=0}^{i_e-1} \tilde{d}_{ej} \right)}.$$

It results, that the number of elements inside an optimal P-set is of order $\mathcal{O}(\sqrt{\prod_{e\in\mathcal{E}} n_e^{i_e-1}})$. This is half the degree of number of all isotopologues in the fine isotopic structure. What is more, the implementation of the algorithm significantly outperforms other existing isotopic calculators. The applications of the ideas that resulted in this algorithm go beyond mass spectrometry, and their use is now investigated in statistics and stochastic simulation.

Deconvolution of Mass Spectra & Ion Statistics

The isotopic fine structure only describes the isotopic variants of roughly one molecule, while the intensity observed in a mass spectrometer is a function of a relatively high number of ions. If we assume, that ions reach the detector independently and in large numbers, than the signal of one substance, normalized by the sum of intensities, should be approximately proportional to the isotopic distribution we describe, which follows from the *law of large numbers*. The above fact is used in many algorithms that perform signal *deisotopisation* – a procedure that aims at tracing all isotopologues of one substance in a given mass spectrum. When the potential sources of signal are known in advance, as while performing a database

search, one can use methods of nonnegative regression (Slawski *et al.*, 2012) that we will describe in Chapter 3.

However, a more detailed approach to the problem, i.e. one that takes into account randomness in ion statistics, can provide interesting insights into the number of observed ions. It has been theoretically argued that the distribution of the number of ions reaching the detector should follow a Poisson distribution (Ipsen and Ebbels, 2012; Ipsen, 2015). This argument goes as follows: assumes that ions move independently throughout the instrument with a limited chance of reaching their final destination. Then, the number of successful detections is binomial. The probability that a sample ion will ever reach the detector is very small, so the binomial distribution is well approximated by the Poisson distribution, which is sometimes reffered to as the *law of rare events*. Chapter 5 describes our attempt at merging the concepts of the isotopic distribution with the Poissonian ion statistics – a tool we call `MassOn`.

The model we propose in `MassOn` also tackles two other important problems in signal processing: (1) the deconvolution of a compound signal and (2) the estimation of the number of observed ions. The nature of that first problem lies in the limited capability of a mass spectrometer to resolve close mass-to-charge ratios. In particular, more than one group of isotopologues can be represented by one peak. This is schematically visualized in Figure 1.1.

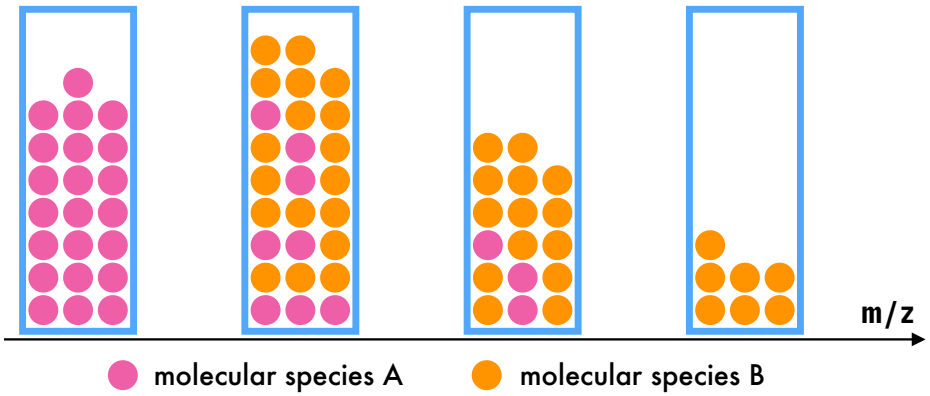


Figure 1.1: A schematic representation of the convolution of two isotopic distributions. Each ball represents one ion, either of kind A or kind B. The above pattern is typically found in problems where two formulas differ by exactly one hydrogen atom, as that difference shifts the spectrum by around 1 dalton. If the $\frac{m}{z}$ is the ratio of the lightest isotope, then other isotopologues tend to cluster around $\frac{m+k}{z}$, where $k \in \mathbb{N}$. This can be attributed to the number of protons inside the nucleai of atoms that make up the molecule.

The second problem stems from the fact that most of the instruments record the ion current that is deemed proportional to the number of passing ions, at least within their trusted dynamic range. The problem of estimating the above proportionality factor is of great relevance, as it appears in most of expressions involving the standard deviations of statistics derived from the theoretical mass spectrum. In particular, if one assumes that the recorded peak heights truly results from an independent motion of ions close to the

detector, then the standard devation of that peak is a function of the square root of the overall number of ions. Both Chapters 3 and 4 describe other important statistics that rely on the specification of the recorded number of ions [2]. `MassOn` tackles both these problems in a fully Bayesian setting relying on a *data augmentated* Gibbs sampling scheme.

Understanding Reaction Pathways

Another limitation of any mass spectrometer is the inability to tell apart substances with the same chemical formula but differing in their 3D structure. In particular, this is the case of two post-translationally modified proteins that have the same modification that could be found on more than one residue. The spatial positioning of a modification is critical for the folding of the protein, and thus – its function. To position a PTM, one has to use more specific techniques, fragmentation being one of them. Ions can be fragmented either outside the instrument, via proteolytic digestion, or inside the instrument. Two prominent ways of inducing fragmentation inside the instrument are the Collisional Induced Dissociation (CID) and the Electron Transfer Dissociation (ETD). The first one consists in heating up the sample cations by exposing them to collisions with some inert gas. This method produces more noisy spectra, as different parts of the molecule detach due to their increased internal motion. ETD is much more subtle technique: it consists of an ion-ion reaction between the sample cations and anions, each carrying a radicals – an electron in a higher energy state. The meeting between these ions is deemed to result in four possible outcomes:

- the transfer of electron from the anion to the cation resulting in the dissociation of the cation – the proper ETD
- the transfer of electron that does not result in any dissociation – ETnoD
- the ETD dissociation followed by a subsequent hydrogen transport – HTR
- the proton transfer reaction – PTR

To study the products of these fragmentation, we devised an approach named `MassTodon`. Chapter 3 provides a detailed explanation of the approach we take to study these reactions. The presented workflow can find the products of these reactions in the spectrum. Furthermore, `MassTodon` can deconvolute the isotopic distributions of different products using constrained quadratic programming[3]. It outputs the estimates of joint intensities of each chemical formula it found from the set of potential reaction products and substrates.

[2] The problem of estimating that number seems also to be a preliminary step to the much more complex problem of the estimation of the molar content of the molecular species within the sample.

[3] The deconvolution performed by `MassTodon` is simpler than that presented in `MassOn`. For this reason, `MassOn` will be described after `MassTodon`.

It can also estimate the probabilities with which different reactions occured in one experiment. This simplifies the comparison of different mass spectra, offering a possibility to better study the influence of different instrumental settings upon the sample; finally, it also simplifies the comparison of different instruments. In particular, `MassTodon` has already found its use to study the unfolding of proteins inside a mass spectrometer (Lermyte *et al.*, 2017), as one can consider the odds ratio between the probabilities of two reactions taken into consideration (ETnoD and PTR).

Reaction Kinetics of Electron Transfer Reactions

With the estimates of the intensity of particular molecular species at hand, as provided by `MassTodon`, it seems natural to pose more specific questions about the nature of the chemical process that could result in a similar mass spectrum. In Chapter 4 (Ciach *et al.*, 2017), we follow a natural approach in this context, which is to apply the well developed mathematical apparatus provided by the theory of reaction kinetics.

We have adapted an approach based on a dynamic stochastic Petri net proposed by Gambin and Kluge (2010). In the particular setting we study, the structure of that net reduces to a directed acyclic graph. This fact significantly increases the theoretical tractability of the problem, as the chemical master equations can be directly applied to establish recursive formulas for the average numbers of ions across the net at a given time. The solution to the above equations depends on a set of reaction rates, each specific for a different reaction. By manipulating these parameters, we can thus compute theoretical numbers of ions and compare them with results obtained by `MassTodon`. We try to minimize the resulting error using a gradient-free L-BFGS-B algorithm.

The developed tool, called `ETDetective` is fully integrated with `MassTodon`. Both algorithms are available for download for free. We are also completing works on a web-service that will make the two algorithms available to a larger public.

Publications in Mass Spectrometry

Lermyte, F., Łącki, M. K., Valkenborg, D., Gambin, A., & Sobott, F. (2017). Conformational space and stability of ETD charge reduction products of ubiquitin. Journal of The American Society for Mass Spectrometry, 28(1), 69-76.

Lermyte, F., Łącki, M. K., Valkenborg, D., Baggerman, G., Gambin, A., & Sobott, F. (2015). Understanding reaction pathways in top-down ETD by dissecting isotope distributions: A mammoth task. International Journal of Mass Spectrometry, 390, 146-154.

Ciach, M. A., Łącki, M. K., Miasojedow, B., Lermyte, F., Valkenborg, D., Sobott, F., & Gambin, A. (2017, May). Estimation of Rates of Reactions Triggered by Electron Transfer in Top-Down Mass Spectrometry. In International Symposium on Bioinformatics Research and Applications (pp. 96-107). Accepted in the Journal of Computational Biology, doi: 10.1089/cmb.2017.0156.

Łącki, M. K., Lermyte, F., Miasojedow, B., Startek, M., Sobott, F., & Gambin, A. (2017). Assigning peaks and modeling ETD in top-down mass spectrometry. arXiv preprint arXiv:1708.00234. Submitted to the American Journal of Mass Spectrometry.

Łącki, M. K., Startek, M., Valkenborg, D., & Gambin, A. (2017). IsoSpec: Hyperfast Fine Structure Calculator. Analytical Chemistry, 89(6), 3272-3277.

Other Publications

Bielczyński, L.W., Łącki, M.K., Hoefnagels, I., Gambin, A., & Croce, R. (2017). Leaf and plant age affects photosynthetic performance and photoprotective capacity. Conditionally accepted in Plant Physiology.

Łącki, M. K., & Miasojedow, B. (2016). State-dependent swap strategies and automatic reduction of number of temperatures in adaptive parallel tempering algorithm. Statistics and Computing, 26(5), 951-964.

Acknowledgements

I would like to thank all the people that contributed to the creation of this thesis.
I thank both my tutors: prof. Anna Gambin and dr Błażej Miasojedow.
I thank my dear collaborators: Michał Ciach, Michał Startek, Frederik Lermyte, Frank Sobott, Dirk Valkenborg, Mikołaj Olszański, Ludwik Bielczyński, and, of course, Piotr Dittwald.
I would also like to thank important institutions that supported me with their funding:

- the National Science Centre for their financial support and entrusting me with grant number 2015/17/N/ST6/03565
- the Vlaamse Instelling voor Technologisch Onderzoek for funding the research of my collaborators in Belgium, without whose data we could not have done anything worthwhile, apart from `IsoSpec`

Let me also thank my dearest parents, Agnieszka and Krzysztof, and my grandmother, Danuta, for their continuous support and understanding.
Finally, I would love to thank Yani Zhao, who must have gone through hell while I was writing all this. I am really sorry, honey! I will make it up to you!

2

Isotopic Distribution Calculations

"Computers are useless. They can only give you answers."

— Pablo Picasso

Until fairly recently, detection of the fine structure isotopic distribution was generally beyond the capability of any mass spectrometer. However, as both FT-ICR MS and Orbitrap instruments continue to be improved, obtaining higher resolution and sensitivity, the detection of fine structure is becoming routine (Nikolaev *et al.*, 2012; G. Marshall *et al.*, 2013; Michalski *et al.*, 2012). As much as 20M FWHM has already been recorded (Hendrickson *et al.*, 2015). The rise of high-resolution (HRMS) and high-throughput mass spectrometry leads to more informative data providing valuable insights into, e.g., molecular identity. Experiments confirm superior identification powers of HRMS, enabling, for instance, correct recognition of metabolites (Nagao *et al.*, 2014) and lipids (Schwudke *et al.*, 2011).

However, more information is more data to analyze: a low resolution full scan mass spectrum of a single molecule consists of only a few peaks, where each peak counts ions that have roughly the same nominal mass. HRMS can resolve these clusters of ions into finer ones. Ideally, with high enough resolution, one could resolve individual isotopologues (McNaught and Wilkinson, 1997), i.e. molecules with the same isotopic composition. For instance, using HRMS one can discern water isotopologues $HD^{16}O$ and $H_2\ ^{17}O$, both with a nominal mass equal to 19 Da. In consequence, more peaks need to be interpreted.

Regardless of the resolution reached by modern instruments and its theoretical limits resulting from thermodynamics (Dittwald *et al.*, 2015), it is instructive to consider the unrealizable case of infinite resolution. In such a setting, the full isotopic distribution of Bovine Insulin, $C_{254}\,H_{377}\,N_{65}\,O_{75}\,S_6$, would be composed of more than 1.5 trillion different isotopologues. This number can be massively reduced if one introduces the probabilistic concept of the chance of finding a given type of isotopologue. Assuming statistical independence of the isotopic variants of atoms (Kienitz, 1961), 414 configurations are enough to represent around 99% of the overall probability. This phenomenon is known as probability *measure concentration* (Talagrand, 1996).

Related Research

To bypass the problem of the rapid increase in the number of isotopologues traditional approaches to isotope calculations have mostly assumed nominal mass approximation (Rockwood, 1995; Dittwald *et al.*, 2013; Snider, 2007; Böcker *et al.*, 2009), binning isotopologues with the same mass number; see Valkenborg *et al* (Valkenborg *et al.*, 2012). In this approach isotopologues with the same nominal mass are indistinguishable: the theoretical distribution is centroided so that highly resolved peaks are represented together with their mass averaged out. The Fourier transform method proposed by Rockwood *et al.* (1996) exempts this rule: it relies on probing the Fourier transform of the mass distribution and offers, in principle, extremely high levels of resolutions. Still, one cannot expect to know *a priori* where to probe the transform and has to resolve to a meticulous search over a grid of mass values, which raises the task's computational complexity.

Recently, the interest shifted towards direct calculation of fine isotopic peaks, giving rise to elegant algorithms, such as `ecipex` (Ipsen, 2014) or `enviPat` (Loos *et al.*, 2015). `ecipex` generalizes the Fourier transform approach investigated by Rockwood to higher dimension. `enviPat` has recently bested `ecipex` in terms of runtime, which can be attributed to direct inspection of the problem on the level of counts of isotopes and by performing pruning of the so called *transition trees*. Both approaches do harness the probability *measure concentration* we exposed on the Bovine Insulin example. However, they specify their outcome in terms of heights of the reported peaks. For instance, they let one neglect all peaks below a given percentage of the highest peak, which is a heuristics first developed by Yergey (1983). A different approach to fine structure calculations, presented by Li *et al.* (2010a), does not present such a disadvantage and the user can specify some joint probability p of the fine structure to be revealed. However, the output of that approach might not be the smallest possible set of isotopologues that is p probable. Together these peaks might be jointly p probable, but there are smaller sets of peaks with this quality.

To our best knowledge, the question of how the choice of the peak-height threshold re-

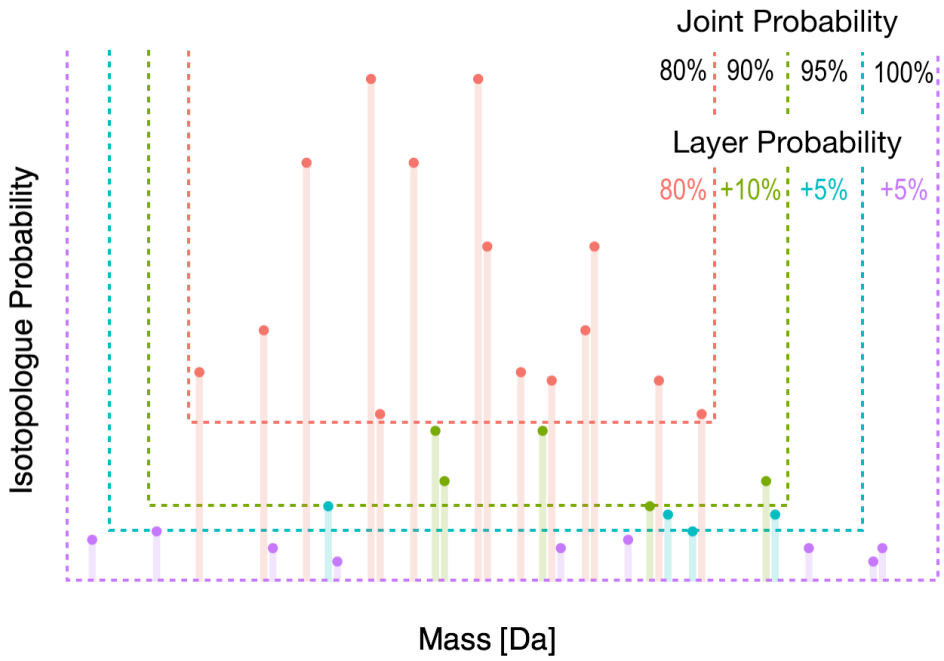


Figure 2.1: Division of isotopic envelope into optimal p-sets, $p \in \{80\%, 90\%, 95\%, 100\%\}$, for a toy molecule. Red peaks correspond to the smallest set of isotopologues that is at least 80% probable; in green we show the minimal additional *layer* of peaks that together with all previous ones are at least 90% probable; in cyan – 95%, in violet - 100%. `IsoSpec` finds minimal sets with a given joint probability without requiring a threshold on peak height, i.e. without a cut-off on the y-axis.

lates to the joint probability of the envelope has not yet been investigated. As demonstrated in Fig. 2.2, this relation is far from trivial, potentially leading to calculations involving isotopologues that are altogether not so important for the analysis. In the case of Bovine Insulin, the smallest set that is 99.99% probable contains 6196 isotopologues in addition to the 414 contained in the smallest 99.9% probable set. On average, these 6196 isotopologues will amount to one per mille of all of the observed ions, making it impractical to consider them. The effect of *overrepresenting an improbable set* is more pronounced for bigger compounds, especially with many atoms of elements that have more than one abundant isotope, such as selenium or sulfur. This underlines the role of precision in the choice of proper pruning threshold.

Peak-height threshold versus joint-probability threshold

The algorithm presented by Li *et al.* (2010a) is the only one that can calculate an isotopic distribution given a joint-probability coverage. Other isotopic calculators usually require a simpler peak-height-based threshold, and stop calculations after finding all peaks that are higher than that value. This threshold can be precised either as an absolute value, or as (small) percentage of the height of the heighest peak. Compared to the joint-probability thresholding, the peak-height thresholds are impractical. First of all, the joint-probability is a direct metric of how much of the theoretical spectrum is revealed. Secondly, it is extremely difficult to predict the joint probability of peaks higher than some peak-height threshold.

If that coverage is too small, then the calculations have to be continued or, more likely, redone; if it is too high, then a lot of computational time has been wasted. What is more, a joint probability threshold guarantees that a lower bound for the actual coverage holds consistently for different chemical formulas. This might not be the case when one fixes the peak-height threshold to some default value (which is usually the case, as defaults are seldom changed by anyone).

Fig. 2.2 explores that situation, showing that the same *peak height* thresholds may lead to large deviations in the joint probability of the revealed spectra. The figure presents result obtained for chemical formulas of almost 15 thousand different human proteins. If the lowest peak is set to be at most 10 times smaller than the heighest one (red row in Fig. 2.2), then the *spread* in the joint probability of the simulated isotopologues might be as big as 69%. While studying this figure, one has to bear in mind that proteins are long polymeric chains. As a result, their atomic content follows an almost linear pattern (Senko *et al.*, 1995). In general, when taking into account a bigger range of molecules, such as lipids or metabolites, the spread might be much wider. Therefore, in addition to not knowing in advance how much probability will be revealed by an arbitrary choice of a *peak height* threshold, one cannot be sure that the selected value will provide consistent coverages for different molecules. As a result, the preparation of data for a consistent statistical analysis of e.g. database driven compound identification becomes unnecessarily cumbersome, as it should involve a procedure that taking under scrutiny each chemical formula and assuring a minimal coverage is attained. One could argue that setting one low peak-heigh threshold might solve all the problems with coverage in most cases. Fig. 2.2 suggest this is true, as ultimately points tend to diverge into regions of high coverage (top-right corner). However, one cannot forget that for these thresholds any calculator will be operating orders of magnitude longer, as the number of output peak greatly increases.

One might argue that a sensible selection of a peak-height threshold could be carried out based on the experimental spectrum. Simply: find both the tallest peak and the smallest peak in an isotopic cluster and divide the height of the first by the height of the latter to estimate the relative threshold. However, such estimate might be highly erroneous because of the following reasons:

1. the smallest peak can be mistaken for a noise peak – data would have to be deconvoluted from noise, which is difficult.

2. peaks are not infinitely resolved and what one believes to be one isotopologue peak in data might be in reality a cluster of many peaks.

3. smaller peaks have higher height variability due to low ion presence.

Avoiding (1) is difficult, as theoretical envelopes do not follow any specific pattern in the

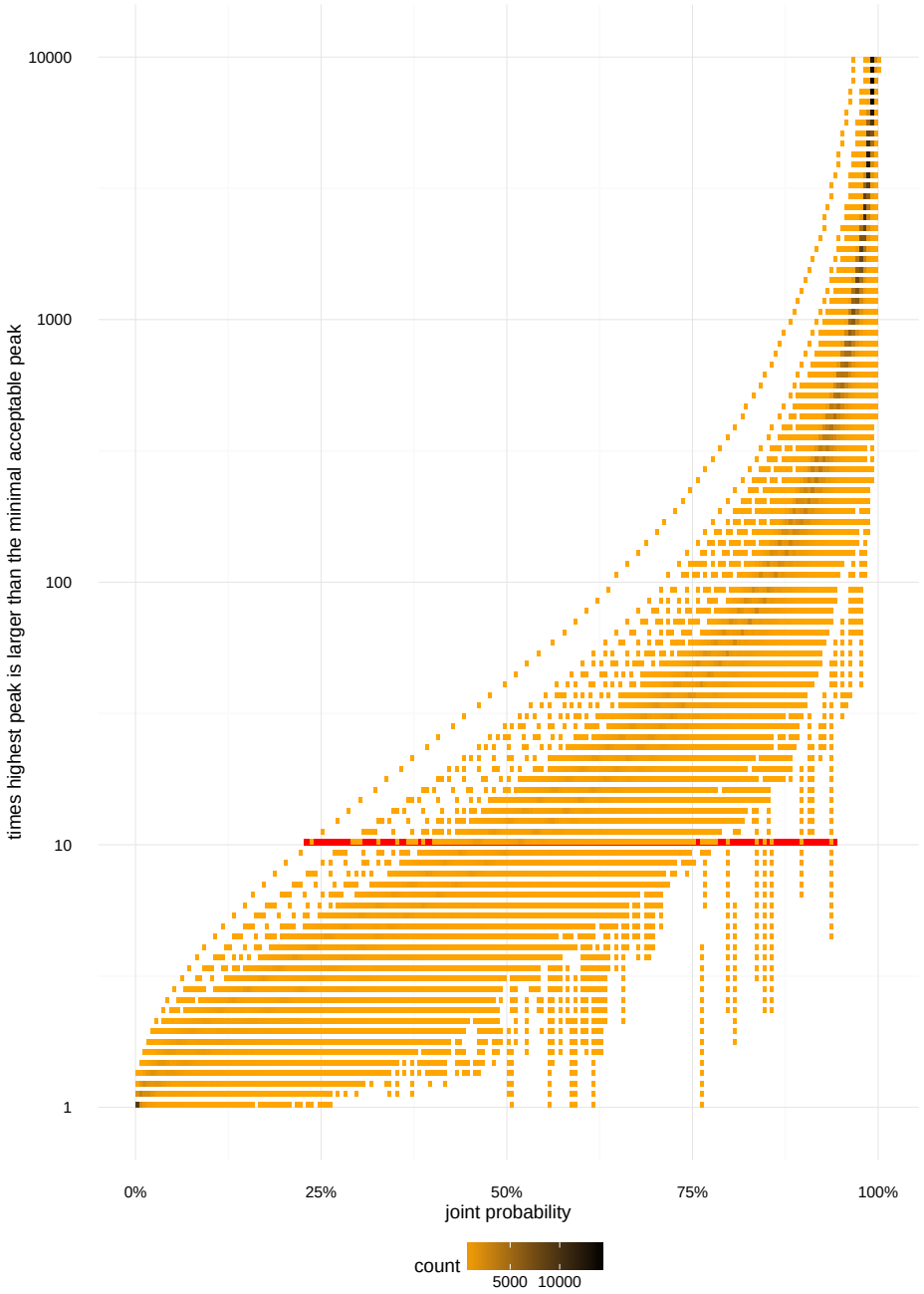


Figure 2.2: Problems resulting from fixing relative peak height threshold at a given value (y-axis) might lead to very different outcomes in terms of joint probability (x-axis) for different chemical molecules. The above plot summarizes results obtained on a set of 14897 human protein chemical formulas from UNIPROT, preselected to contain at most 30 atoms of sulfur. Each distinct row depicts with changing color (yellow to black) the concentration obtained for the 14897 values. Only when the highest peaks is much taller than the smallest one can one notice some form of concentration of results. The red row contains results obtained when the top probable isotopologue is around 10 times more probable than the lowest acceptable. This threshold could be chosen while trying to generate a spectrum of a limited coverage, for quickly disqualifying a molecule as potentially identified in an experimental spectrum.

mass domain. The second case might in turn lead to a massive overestimation of the peak height. Finally, let us present a simplistic statistical argument that motivates (3).

Suppose that the total ion count of a given molecular species N follows the Poisson distribution Ipsen and Ebbels (2012). The molecules of that species can be further divided into groups defined by their isotopic variants. The probabilities of these groups can be obtained with `IsoSpec`. A subisotopologue group with probability p will be populated by a molecules whose count also follow the Poisson law, albeit with intensity equal to λp[1]. The coefficient of variation of the number of ions in that group equals then

$$CV = \frac{\text{standard deviation}}{\text{expected value}} = \frac{\sqrt{\lambda p}}{\lambda p} = (\lambda p)^{-0.5}.$$

Suppose we investigate two groups of isotopologues: a highly probable one, H, and an unlikely one, L. Then, if we compare their coefficients of variation we see that

$$\frac{CV_L}{CV_H} = \sqrt{\frac{p_H}{p_L}},$$

showing that the variation of low probable peaks is higher than that of high probable peaks as p_H should be orders of magnitude higher than p_L. However, this is attenuated by the square root function.

Our approach

In this chapter, we will present an algorithm for retrieving the smallest possible set of isotopologues with a given probability that the user wishes to unveil. Our algorithm bridges the apparent gap between algorithms such as `enviPat` or `ecipex` and the recursive approach developed by Li *et al.* (2010a). In contrast to many other approaches, we also analyze the computational complexity of the presented solutions. We prove that our algorithm is optimal in terms of time complexity. Finally, we present an implementation of `IsoSpec` that is superior to the fastest fine structure calculator to date, `enviPat`, as tested on a set of more than 800,000 chemical formulas obtained by *in silico* fragmentation of 1,000 human proteins.

The infinitely resolved spectrum can comprise thousands of peaks for just one molecule. One could doubt the usefulness of this concept arguing that this is experimentally unachievable. However, isotopologues can be aggregated based on the similarity of their masses so as to match the resolution of the used instrument, see Li *et al.* (2010a). Our approach guarantees that this can be achieved quickly and with control over the error of the approximation.

In the rest of this chapter we describe the theoretical gains from any strategy resulting in optimal pruning. Then, we describe the `IsoSpec` algorithm. Finally, we compare its

[1] This remark will be heavily used in Chapter 5.

runtime with the `enviPat` algorithm. In our presentation we focus on proteins; however, the implementation and the analysis both apply to any known compounds, even those containing other elements than carbon, hydrogen, nitrogen, oxygen, and sulfur.

The Complexity of Pruning

Consider a protein with a formula $\mathrm{C_c\,H_h\,O_o\,N_{\mathit{n}}\,S_s}$, i.e. with c atoms of carbon, h hydrogen, *n* nitrogen, o oxygen, and s sulfur. Denote by $\mathcal{E}$ the set of the chemical elements the protein is composed of and by n_e the number of atoms of a given element e composing the protein, i.e. $n_e \in \{\mathrm{c,h,n,o,s}\}$. Finally, denote by i_e the number of stable isotopes of that element. The total number of different isotopic variants of $\mathrm{C_c\,H_h\,O_o\,N_{\mathit{n}}\,S_s}$, i.e. the total number of its *isotopologues* (McNaught and Wilkinson, 1997), equals $\prod_{e\in\mathcal{E}} \binom{n_e+i_e-1}{n_e}$. Using Stirling's approximation of the factorial (Feller, 1968), one finds that the above is approximately

$$\prod_{e\in\mathcal{E}} \frac{e^{i_e-1}}{\sqrt{2\pi(i_e-1)}}\left(\frac{n_e}{i_e-1}+1\right)^{i_e-1}.$$

Extracting n_e we conclude that the total number of isotopologues is asymptotically polynomial in the numbers of atoms, $\mathcal{O}(\prod_{e\in\mathcal{E}} n_e^{i_e-1})$.

Carbon, nitrogen and hydrogen have two stable isotopes each, resulting roughly in a linear increase in isotopologues with the number of atoms of these elements. With respectively three and four stable isotopes the relation for oxygen becomes quadratic, and cubic for sulfur. This quantifies the extent of *combinatorial explosion* of the direct enumeration of all isotopologues. We want to avoid calculating unlikely isotopologues. Assuming that the isotopic variants of atoms composing $\mathrm{C_c\,H_h\,O_o\,N_{\mathit{n}}\,S_s}$ are independent and drawn with the same abundances across elements (Kienitz, 1961), one pinpoints the probability of an isotopologue to be a product of multinomial distributions, equal to

$$\prod_{e\in\mathcal{E}} \binom{n_e}{n_{e0},\ldots,n_{e,i_e-1}} p_{e,0}^{n_{e,0}}\cdots p_{e,i_e-1}^{n_{e,i_e-1}}, \tag{2.1}$$

and mass to $\sum_{e\in\mathcal{E}}\sum_{i=0}^{i_e-1} m_{e,i}n_{e,i}$, where n_{ej} is the count of element e's j^{th} isotope, and p_{ej} and m_{ej} are respectively its abundance and mass in daltons, both reported by IUPAC (Brand *et al.*, 2014). With Eq. (2.1) at hand, it is natural to search for sets of isotopologues that jointly surpass some limiting value of probability that is close to 100%, say p. Many such sets exist, so it seems reasonable to limit one's attention to the smallest one. That set must include the heighest peaks. We call such a set an optimal p-set – Fig. 2.1 explores that concept. The p-set is not necessarily unique. For instance, consider a fictitious monatomic compound with two equally possible isotopes, I_1 and I_2. Then there are two possible optimal 50%-sets: that composed of I_1 and that composed of I_2. In general, meeting a multitude of optimal p-sets is highly unlikely.

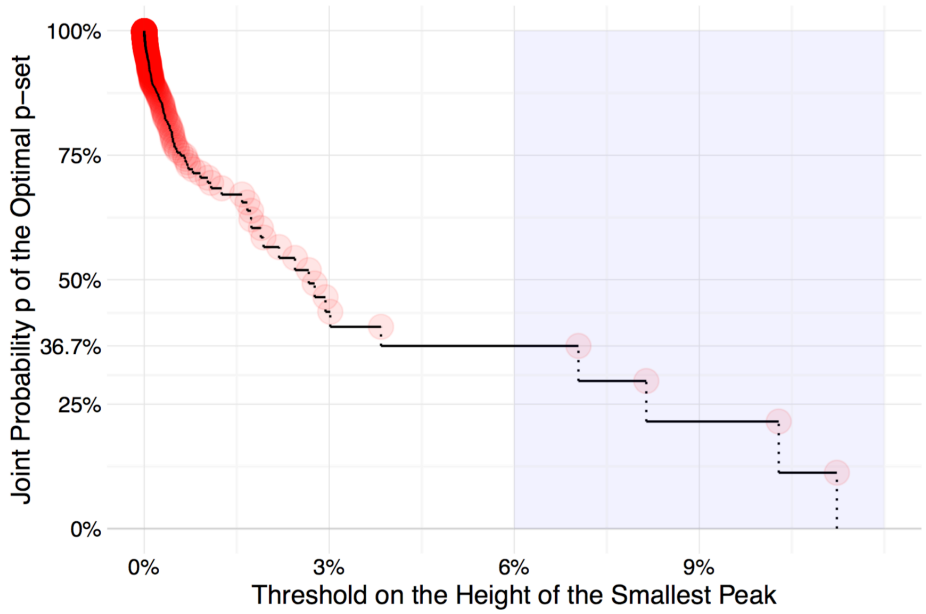


Figure 2.3: The *threshold function* obtained for Bovine Insulin. The function relates the choice of peak height threshold τ with the joint probability p of the resulting set of isotopologues, i.e. the ones with peak height at least τ. It usually happens that there is no peak with height exactly τ: the *effective* configuration (in red) is then to be found to the right on the same level. Trimming peaks less than 6%-probable (height below 0.06) one gets a set of 4 isotopologues (red dots on the blue background) with joint probability 36.7%. Higher intensity of red in top-left corner indicates that lower thresholds rapidly increase the number of resulting isotopologues.

Observe that the optimal p-sets in Fig. 2.1 are separated by horizontal dashed lines up to configurations with the same probability. To obtain an optimal p-set one can choose a threshold on peak height and then discard some of the low probable peaks of the same height. Usually there is only one peak with minimal height, so that the output of both the `enviPat` and `ecipex` algorithms coincides with an optimal p-set, for some joint probability p. However, to get p one has to establish a set of isotopologues first.

The relationship between the input threshold and the joint probability of the output p is presented in Fig. 2.3 on the example of Bovine Insulin. The resulting *threshold function* is locally flat, non-increasing, and right-continuous. The input threshold will usually be smaller than the actual minimal probability observed in the output p-set: we call isotopologues with that probability *effective*. They are depicted as red, semitransparent circles in Fig. 2.3, and correspond to right ends of the intervals that make up the curve. High concentration of the *effective isotopologues* in the top left region suggests high sensitivity of the number of configurations in the optimal p-set to the choice of the input threshold. The idea behind the `IsoSpec` algorithm is to reach the input joint probability p by moving along the graph of the *threshold function*, from bottom-right to upper-left.

Before describing in detail the `IsoSpec` algorithm, let us briefly elaborate on the potential gains resulting from either peak-height thresholding or joint-probability thresholding. An isotopologue of $C_c\,H_h\,O_o\,N_n\,S_s$ can be fully described by the numbers of isotopes of dif-

ferent elements that compose it, called *subisotopologues* (Loos *et al.*, 2015), as in

$$\underbrace{{}^{12}\mathrm{C}_{c_0}\ {}^{13}\mathrm{C}_{c_1}}_{\text{subisotopologue}}\ \underbrace{{}^{1}\mathrm{H}_{h_0}\ {}^{2}\mathrm{H}_{h_1}}_{\text{subisotopologue}}\ \underbrace{{}^{14}\mathrm{N}_{n_0}\ {}^{15}\mathrm{N}_{n_1}}_{\text{subisotopologue}}\ \underbrace{{}^{16}\mathrm{O}_{o_0}\ {}^{17}\mathrm{O}_{o_1}\ {}^{18}\mathrm{O}_{o_2}}_{\text{subisotopologue}}\ \underbrace{{}^{32}\mathrm{S}_{s_0}\ {}^{33}\mathrm{S}_{s_1}\ {}^{34}\mathrm{S}_{s_2}\ {}^{36}\mathrm{S}_{s_3}}_{\text{subisotopologue}}$$

A subisotopologue corresponding to element e can be thus represented as a tuple

$$\boldsymbol{n}_e = (n_{e,0}, \ldots, n_{e,i_e-1})$$

of specific isotope counts, where $\sum_{j=0}^{i_e-1} n_{e,j} = n_e$. The inspection of the probability of an isotopologue described by equation (2.1) further reveals that each multinomial distribution present in the product corresponds to the probability of exactly one subisotopologue. If e has three isotopes, then one can depict subisotopologues on a ternary plot, as in any subplot of Fig. 2.7. In general, subisotopologues constitute a discrete grid on the simplex. With a growing number of atoms of each element in a chemical compound, the multinomial distributions in Equation (2.1) can be individually approximated by multivariate Gaussian distributions with the same mean and covariance matrix.

The Gaussian approximation

It is well known that the mean of a multinomial distribution is equal to $\mu_e = n_e p_e$, where p_e is the vector of probabilities of individual outcomes and n_e is the number of trials. It is also easy to notice, that its covariance matrix equals $\Sigma_e = n_e(d(p_e) - p_e p_e^t)$, where by $d(p)$ we understand a matrix with vector p on the diagonal and zeros elsewhere, and by $p_e p_e^t$ – a matrix of a projection on vector p_e. Matrix Σ_e is degenerate and one cannot use the standard formula for the normal density[2]. This is because the multinomial distribution is itself well defined in a $(i_e - 1)$-dimensional simplex embedded in a i_e-dimensional space of possible outcomes: any approximation must lie in the same subspace. It is however well defined an invertible on the space perpendicular to p_e. Let us perform the SVD decomposition of Σ_e, which by its self-adjointness equals $U\Delta_e U^t$, where U is unitary and Δ_e is diagonal, with exactly one entry on the diagonal equal to o,

$$\Delta_e = \begin{bmatrix} d_{e1} & 0 & \cdots & 0 & 0 \\ 0 & d_{e2} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & d_{e,i_e-1} & 0 \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

[2] A similar procedure can be applied to the case described by Kaur and O'Connor (2004) in the context of the estimation of the number of observed ions. There, the condition $\det \Sigma = 0$ is neglected, which is truly appauling.

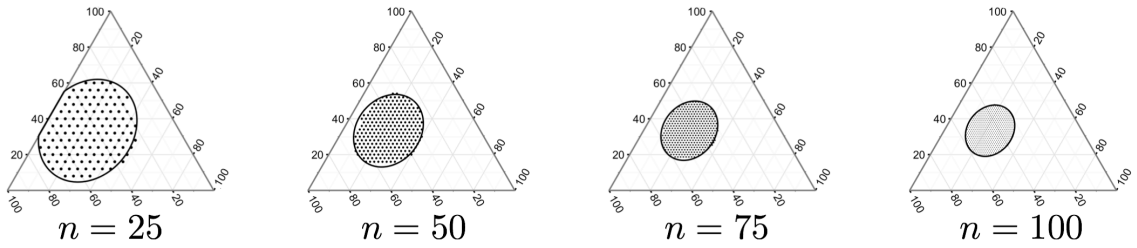


Figure 2.4: The quality of the Gaussian approximation to the optimal p-set for a toy example one element compound with two isotopes. As predicted by the *Central Limit Theorem*, the shape of the optimal p-set for a one element compound can be well approximated by an ellipsoid defined by the mean and covariance matrix of the multinomial distribution. The simplices are normalized to the number of atoms of the toy compound. Notice sublinear growth of the volume of the ellipse: according to approximations, its area should behave approximately like a square root of n - the number of atoms.

Only one of the columns of U can generate a linear space where Σ_e degenerate. All the other columns of U constitute a new coordinate system. Shift that system to μ_e and there can one properly define the Gaussian distribution with the covariance matrix equal to the largest nondegenerate minor of Δ and zero mean. This procedure can be performed for all elements that have isotopes.

The *Central Limit Theorem* (Kallenberg, 1997) assures that with the growing number of atoms, the multinomial distribution converges to the Gaussian distribution with the same mean vector and covariance matrix. The concept of the optimal p-set in case of a continuous distribution naturally reduces to the notion of a smallest set with a fixed probability p. For normal distribution, the ellipsoids of confidence match exactly that notion, so we approximate the original optimal p-sets with ellipsoids containing p probability, see Fig. 2.4. This figure also shows that the relative quality of approximation increases with the number of atoms.

To approximate the whole product of multinomial distributions it is enough to approximate each element of the product by the appropriate normal distribution. This leads to a product of normal distributions. A product of multivariate normal distributions is again a normal distribution, yet higher dimensional. To be more specific, if the individual normal distributions had means μ_e and covariance-matrices Σ_e, then the *joint* normal distribution has mean $\mu = (\mu_1, \ldots, \mu_{|\mathcal{E}|})$ and covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 & 0 \\ 0 & \Sigma_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \Sigma_{|\mathcal{E}|-1} & 0 \\ 0 & 0 & \cdots & 0 & \Sigma_{|\mathcal{E}|} \end{bmatrix},$$

where 0 are block matrices of zeros. Similarly to what was described above, we can perform the SVD decomposition of Σ and perform calculations in the new space, spanned by all the eigenvectors other than vectors of ones. In that space, one would typically consider an

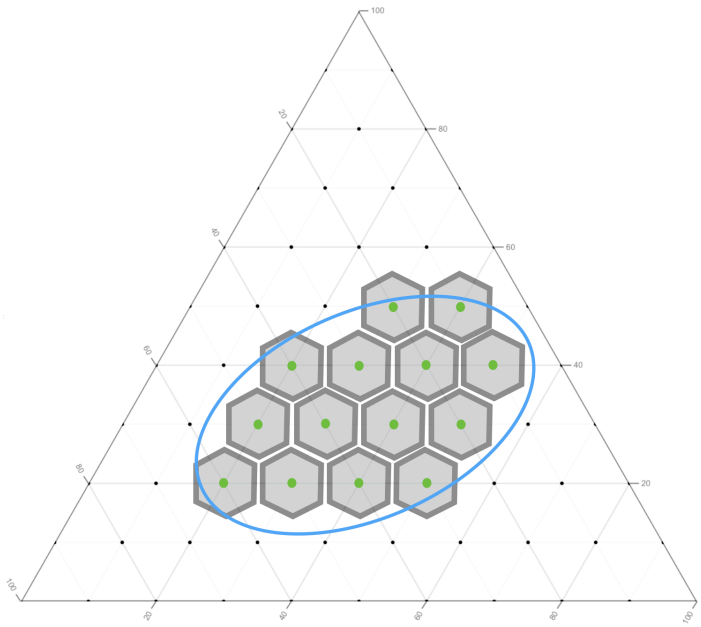


Figure 2.5: Idea behind the proof of proportionality of the ellipsoid volume to the number of subisotopologues on the simplex. The 2D ellipsoid (in blue) contains 14 subisotopologues (in green). Each subisotopologue is surrounded by a grey area resulting from a Voronoi diagram partition of the simplex. The more atoms there are, the relatively finer the Voronoi tessellation, and the better gets the Gaussian approximation.

ellipsoid $(x-\mu)^t\Sigma^{-1}(x-\mu) \leq R^2$. In our case, however, Σ is degenerate, and Σ^{-1} is meaningless, and should be replaced by a pseudoinverse. We restrict the bilinear form Σ to a linear subspace that contains all the considered simplices. In the new coordinates, obtained through SVD, the ellipsoid with radius R equation is simply

$$E_R = \left\{ x \in \mathbb{R}^k : \sum_{e\in\mathcal{E}} \sum_{j=0}^{i_e-1} \frac{x_{ej}^2}{d_{ej}} \leq R^2 \right\},$$

where $k = \sum_{e\in\mathcal{E}} i_e - |\mathcal{E}|$. The volume of E_R equals

$$\textsc{Vol}(E_R) = R^k \frac{\pi^{k/2}}{\Gamma(k/2+1)} \sqrt{\prod_{e\in\mathcal{E}} \prod_{j=0}^{i_e-1} d_{ej}},$$

where Γ is the gamma function, $\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}\mathrm{d}\,t$. The choice of R has to be performed so that the ellipsoid contains exactly p probability. It is widely known, that the probability of E_R equals to the value of the chi-square cumulative distribution function (Izenman, 2008) evaluated at R^2. Therefore, to get the appropriate value of R one has to consider the p^{th} quantile of that distribution, so that $R^2 = q_{\chi^2(k)}(p)$.

The volume of an ellipsoid is proportional to the number of isotopologues contained within it. This can be proved for individual simplices and then the argument extends by *tensorization*. For one simplex derived for subisotopologues with 3 isotopes, the situation is depicted in Fig. 2.5. It is enough to consider a Voronoi (Okabe *et al.*, 2000) diagram partitioning of the simplex: namely, consider a partition into sectors closest to individual

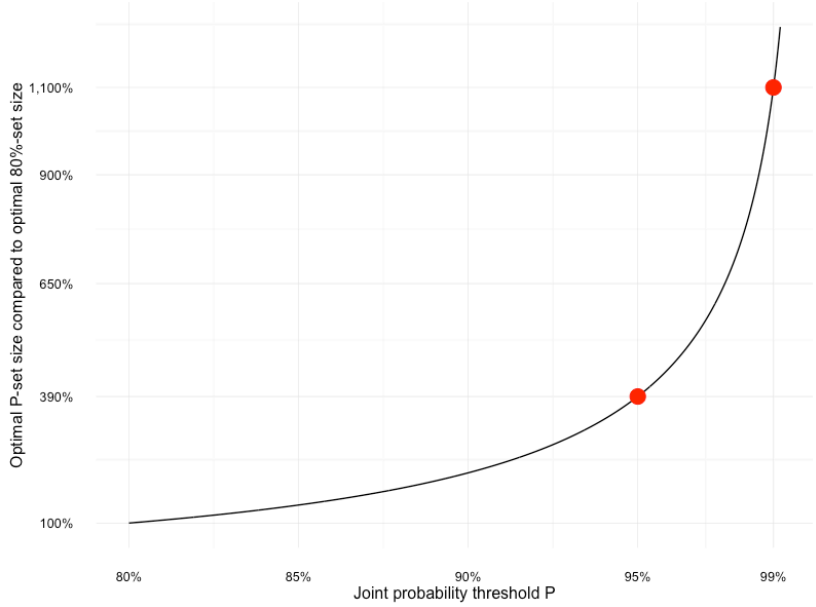


Figure 2.6: Approximate size of the optimal P-set in terms of the size of the optimal 80%-set (y axis, logarithmic scale) for different joint thresholds P (x axis).

subisotopologues in the Euclidean distance. The part of this *honeycomb* structure obtained by restricting attention only to points inside the ellipse is an approximation to the overall shape of the ellipse. A limiting argument shows that if one normalizes the simplex to unit size, then the bigger the number of atoms, the better the approximation.

Let us express the volume of an ellipsoid in terms of the volume of the *honeycomb*. To this end, we have to know the volume of one basic unit of the Voronoi tessellation. For a 2D simplex, this unit corresponds to an individual hexagon in Fig. 2.5. The idea is simple: each hexagon is centered at exactly one subisotopologue, so if we knew the volume of the hexagon then we could reexpress the volume of the ellipsoid in terms of isotopologues.

A 2D hexagon can be decomposed into l regular simplices with the edge equal to half the distance between two subisotopologues b. In general, for a simplex of dimension d, l should be equal to the number of neighbouring subisotopologues, equal to $2\binom{d+1}{2}$. This result comes from the following reasoning: two neighbouring subisotopologues differ on exactly two coordinates out of $d+1$ coordinates. This gives $\binom{d+1}{2}$ possible pairs of coordinates to change. If we know which coordinates should change, there are only two ways to change them: by adding one to the first and substracting one to the latter or *vice versa*. We can neglect the case where the subisotopologue is to be found on the border of the studied simplex: it is irrelevant for the study of asymptotics. Also, the euclidean distance b between two subisotopologues equals $\sqrt{2}$ and does not depend on the dimension.

The overall volume C is a product of the ones obtained for individual subisotopologues, $\prod_{e\in\mathcal{E}} C_e$. For instance, in case of bovine insuline considered above, composed out of three

elements with two stable isotopes, one with 3 stable isotopes and one with four, we obtain

$$C = \left(2\frac{\sqrt{2}}{1!}\binom{2}{2}\right)^3 2\frac{\sqrt{3}}{2!}\binom{3}{2}2\frac{\sqrt{4}}{3!}\binom{4}{2} = 48\sqrt{6}.$$

Otherwise said, dividing the volume of the ellipsoid by this number will result in an approximate number of covered isotopologues.

Finally, note that since $\Sigma_e = n_e(d(p_e) - p_e p_e^t)$, then we can extract the n_e factor from d_{ei} in Δ_e, $d_{ei} = n_e\tilde{d}_{ei}$. Therefore, the total number of isotopologues M is approximately

$$M = \frac{q_{\chi^2(k)}(p)^{\frac{k}{2}}}{C}\frac{\pi^{k/2}}{\Gamma(k/2+1)}\sqrt{\prod_{e\in\mathcal{E}}\left(n_e^{i_e-1}\prod_{j=0}^{i_e-1}\tilde{d}_{ej}\right)}. \tag{2.2}$$

The above formula can be innaccurate for small ellipses: in such cases the subisotopologues close to the edge of ellipse can distort the calculation behind the proportionality constant C.

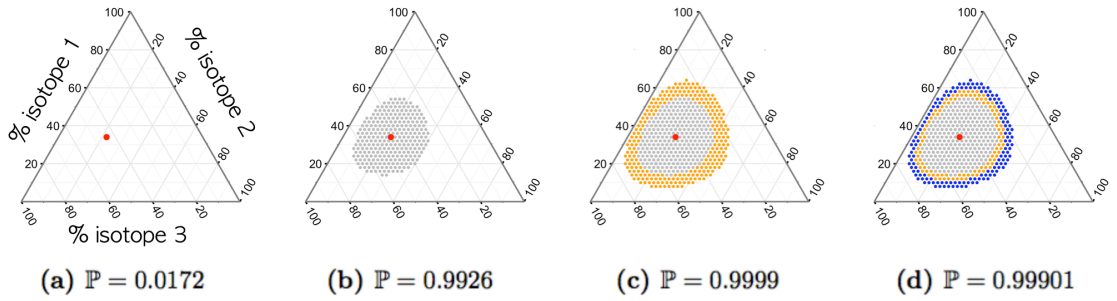
Observe that all values in the above formula can be easily calculated either by a closed formula, like C, or numerically. Note also, that if one assumes a computational model prescribing the same amount of time to the calculation of each configuration, then the above formula approximately quantifies the runtime-probability trade-off: namely, it should behave as the quantile function of the chi square distribution raised to $k/2$. Moreover, given two different joint probability thresholds their relative runtime can be expressed as simple as $(q_{\chi^2(k)}(P)/q_{\chi^2(k)}(P'))^{k/2}$. For instance, if $k = 8$, which is the case for any compound composed out of carbon, hydrogen, nitrogen, oxygen and sulphur (we take into account only the stable isotopes) one can plot Fig. 2.6.

Observe, that the optimal 95% and 99% sets are respectively approximately 3.9 and 11 times larger than the optimal 80% set. By transitivity, the 99%-set should be approximately 2.82 times larger than the 95%-set.

It results from Eq. (2) that the overall number of isotopologues above a given probability threshold behaves asymptotically like $\mathcal{O}\left(\sqrt{\prod_{e\in\mathcal{E}} n_e^{i_e-1}}\right)$. This is roughly a square root of the order of the total number of isotopologues. Therefore, trimming truly effectively averts the *combinatorial explosion*.

The IsoSpec Algorithm

The `IsoSpec` algorithm consists of four procedures: (1) the generation of subisotopologues, (2) the merger of subisotopologues into sets of isotopologues above a given threshold, (3) the generation of a sequence of consecutive thresholds, and finally (4) the trimming of the output into the final shape. The first two steps are interwoven and describe a fully operational *peak height trimming* algorithm that we call `IsoSpec` Threshold. Using these four procedures, `IsoSpec` works as follows: first it generates the top probable subisotopologues.



(a) $\mathbb{P} = 0.0172$ (b) $\mathbb{P} = 0.9926$ (c) $\mathbb{P} = 0.9999$ (d) $\mathbb{P} = 0.99901$

Figure 2.7: The principle behind the `IsoSpec` algorithm. Consider a $n_e = 50$ atoms molecule made up entirely out of one fictitious element with three isotopes. The concepts of subistopologue and isotopologue coincide. Isotope content of isotopologues is represented as points in the above ternary plots. In general, isotopologues correspond to tuples of points on different simplices. To find the optimal 99.9%-set, one first establishes the most probable isotopologue, like in **(a)** in red. Then, one finds the first optimal p_1-set, as in **(b)** by choosing some threshold τ_1, like in **(b)** in grey. One then sums all peaks heights to see that $p_1 = 99.26\%$, smaller than 99.9%. One gets another threshold τ_2, establishes new layer of isotopologues, **(c)** in orange, and finds that $p_2 = 99.99\%$. This set is too big and one trims out the isotopologues in blue in **(d)**. Then, $p > 99.99\%$, but removing more isotopologues would bring joint probability below 99.99%.

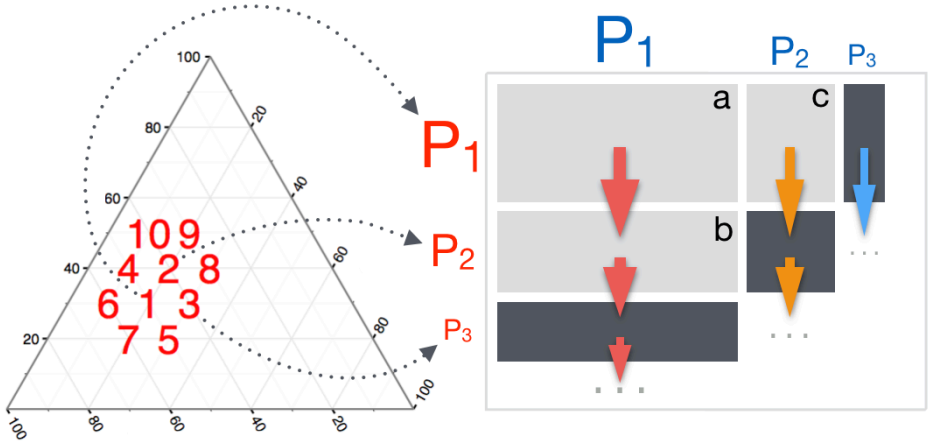


Figure 2.8: Merging subisotopologues into isotopologues on a toy example of a two element molecule. The lengths of the edges of rectangles correspond to probabilities of subisotopologues: these are decreasing for both the **red** and the **blue** element, and correspond to subisotopologues that concentrate around the most probable subisotopologue, as in the ternary plot. Isotopologues are visited lexicographically: first, one travels down the red pathway (column with rectangles **a** and **b**) till reaching a dark rectangle with area below threshold τ. Then, one travels down the orange pathway (column with rectangle **c**); and so on. Dark rectangles form the *fringe*: a set of neighbors of isotopologues more probable than τ. Having obtained another threshold $\upsilon < \tau$, one continues the lexicographic descent starting from the fringe until first isotopologues less probable than υ are reached, forming a new *fringe*.

Eq. (2.1) indicates that together they form the top probable isotopologue. Then, `IsoSpec` iteratively produces optimal p-sets of isotopologues, each corresponding to some threshold τ from the sequence of thresholds. Every time a p-set is obtained, its joint probability p is established and compared with the target value $\mathbb{P}$. This is repeated until p gets larger than $\mathbb{P}$. Finally, the last *layer* of peaks is trimmed leaving the required optimal $\mathbb{P}$-set. Fig. 2.7 visualizes this approach on a simplified molecule composed of exactly one element.

Calculating subsequent subisotopologues corresponds to reporting configurations of a given multinomial distribution with decreasing probability. This is easy thanks to its unimodality. To define what we mean by unimodality, we first relate subisotopologues of element e spatially: let two subisotopologues $\boldsymbol{n}_e^1$ and $\boldsymbol{n}_e^2$ be neighbors, $\boldsymbol{n}_e^1 \sim \boldsymbol{n}_e^2$, iff one is obtainable from the other by changing the isotopic variant of exactly one atom. For

instance, $^{16}O_3 \sim {}^{17}O^{16}O_2$ as one atom changed from ^{16}O to ^{17}O. However, $^{16}O_3 \not\sim {}^{17}O_2{}^{16}O$, as two atoms would have to change from ^{16}O to ^{17}O. Two neighbors are also close on the simplex in the geometric sense, like dots in Fig. 2.7. A discrete distribution is unimodal, if and only if the set of global maxima is connected. Consequently, every configuration not top probable has an equally or more probable a neighbor. The multinomial distribution is unimodal in that sense (Finucan, 1964).

Unimodality simplifies the task of reporting subisotopologues sorted by decreasing probability for a given element e. Algorithm 1 (by the end of the chapter) precises how to carry out this task. We call such procedure a *subgenerator*. A *subgenerator* starts from top probable subisotopologue. It gets there by a simple *hill climbing* algorithm: it starts with a subisotopologue close to the mean of the multinomial distribution and follows the direction of increasing probability until the maximum is reached. By unimodality, it must be a global one. It then enlists it in an empty priority queue PQ, with priorities set to probabilities of subisotopologues. Then, it iteratively extracts the top probable element from PQ and inserts its yet unvisited neighbors. By unimodality one can only insert subisotopologues less probable than those popped out. Each configuration has a limited number of neighbors, so the size of PQ is of the order of the number of already visited subisotopologues, n. Using the standard heap implementation of the PQ, calculations involving n configurations take up $O(n \log(n))$ time.

We store the results of previous calls as well as the state of the subgenerator to avoid unnecessary recomputations. This way the retrieval of the already calculated probability, e.g. while passing from red pathway to orange pathway in Fig. 2.8, can be done faster. Multiple visits to subisotopologues can be avoided through hashing. The computational complexity of operations on subisotopologues is negligible compared to subisotopologue merger. A *subgenerator* provides the k-th most probable subisotopologue and its probability. To get an isotopologue, one considers a tuple of $|\mathcal{E}|$ different subisotopologues, each obtained with a different *subgenerator*. The probability of an isotopologue is the product of probabilities of its constituent subisotopologues. `IsoSpec` uses a series of thresholds to obtain *layers* of isotopologues. It starts by merging the top probable subisotopologues. Given any isotopologue γ, it uses *subgenerators* to establish its less probable neighbors – the *successors*. A *successor* of γ has precisely one subisotopologue changed to the next one in line. For instance, isotopologues **b** and **c** are successors of **a** in Fig. 2.8. To generate isotopologues above a threshold τ consists in inserting and popping elements from a queue. In comparison to *subgenerator*, sorting elements is redundant, and so a priority queue can be replaced with a simple FIFO queue (Cormen, 2009). To avoid repeated visits to the same configurations, `IsoSpec` follows a lexicographic visiting schedule, as shown by colored arrows in Fig. 2.8. Each *popped out* isotopologue qualifies to a given *layer* if its probability is

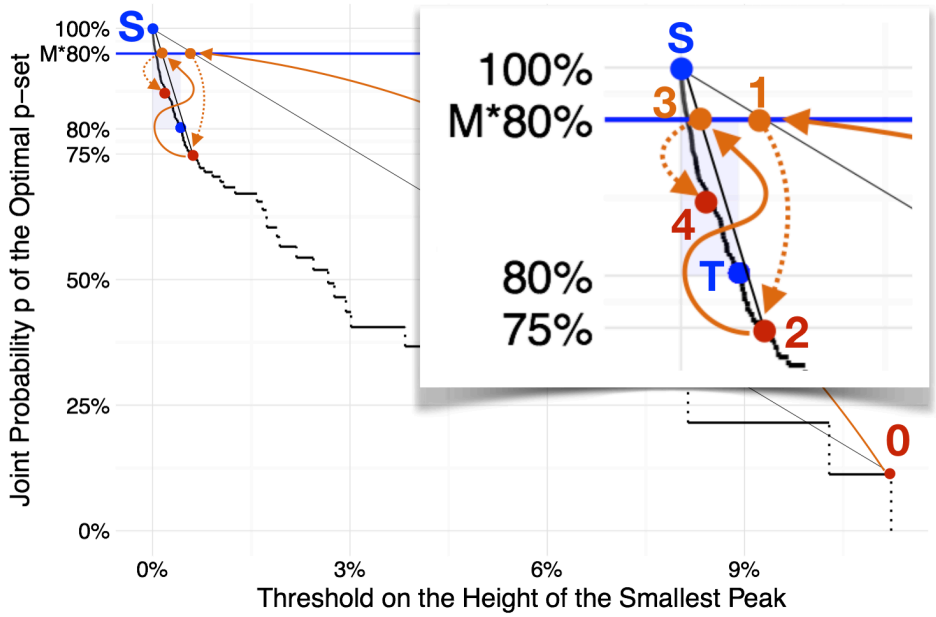


Figure 2.9: Adaptive linear approximation to the *threshold* function. It starts at point (P,P) – the top probable isotopologue, **0**, and aims at finding the optimal 80%-set, point **T**. Point **1** on line **0-S** is where we would get if our approximation using multiplier M was perfect. Instead, it leads to only 75% of the joint probability, as indicated by point **2**. Line **2-S** provides another approximation, and suggests point **3**. In reality, we move to **4** – already above the target 80%. The *effective isotopologues* on the *threshold function* between points **4** and **T** can to be trimmed.

above τ. Otherwise, it is stored in a so-called *fringe*, and used in the next iteration with a new threshold. The procedure is repeated until the joint probability exceeds that required by the user.

Successive threshold values result from an adaptive linear approximation to the *threshold function*, see Fig. 2.9. Given the top probable isotopologue with probability P we can draw a line between point (P, P) and point $S = (0, 1)$. Point S lies on the *threshold function*, as the choice of a 0 threshold on peak height results in a full set of isotopologues, i.e. a 100% probable set. On that line we find a point slightly above the required value $\mathbb{P}$, say $M\mathbb{P}$ where $M > 1$ is chosen heuristically. The x coordinate of that point provides the first threshold, τ_1. Applying the previous procedure on τ_1 we get the optimal probability p_1. A new line is drawn between point S and (τ_1, p_1) and the procedure is iteratively repeated until $p_k > \mathbb{P}$, where k is the number of the last iteration. A slight overestimate is needed for the algorithm to converge.

Finally, the trimming of the last *layer* of isotopologues can be performed in a linear time with its size using the QuickTrim algorithm. The QuickTrim algorithm, as specified by pseudocode in Algorithm 2 (by the end of the chapter), is a modified version of the classic QuickSelect algorithm. Denote the number of isotopologues in the last layer to be trimmed by n_{LL}. Then, the algorithm achieves an $O(n_{\mathrm{LL}})$ pessimistic runtime if the Magic Fives Blum *et al.* (1973) algorithm is used for pivot selection. In practice, the pivot is selected at random, resulting in $O(n)$ average runtime, and $O(n^2)$ pessimistic runtime.

Before providing results on the overal time complexity of the algorithm, we shall give brief account on the numerical questions regarding both the calculations of the probabilities and masses of isotopologues. First of all, we prefer to calculate the logarithms of probabilities over probabilities. The way we perform the calculation of the logarithms of probabilities of individual isotopologues differs from that presented in the literature (Yergey, 1983; Li *et al.*, 2008, 2010a). In particular, we do not calculate them recursively, which is prone to numerical error propagation. Instead, we calculate them seperately for every configuration. The cost of calculating masses is minimal (i_e multiplications and $i_e - 1$ additions for each isotopologue - a fixed costs). To calculate probabilities we have to quickly calculate the logarithm of Eq. (2.1),

$$\log\left(\prod_{e\in\mathcal{E}}\binom{n_e}{n_{e0},\ldots,n_{e,i_e-1}}p_{e,0}^{n_{e,0}}\cdots p_{e,i_e-1}^{n_{e,i_e-1}}\right).$$

The logarithm of probabilities can be precomputed. What remains is the logarithm of the generalised Newton symbol

$$\log(n_e!) - \sum_{j=0}^{i_e-1}\log n_{ej}!.$$

Each log-factorial can be calculated using the Stirling approximation, as exposed in `http://www.johndcook.com/blog/2010/08/16/how-to-compute-log-factorial/`. This way, the calculations are exact up to 14 significant numbers. Since we use double precision standard to represent real numbers, which has a precision of around 16 significant numbers, we are making an extremely small error. In the approximation we use the approximation

$$\log(n!) \approx (n-1/2)\log(n) - n + (1/2)\log(2\pi) + 1/(12n),$$

for $n > 256$, and use a precomputed value for smaller n. We can further reduce the error by adding the $1/(360x^3)$ term.

The overall time complexity of the `IsoSpec` algorithm, as specified by the pseudocode in Algorithm 3, is

$$\mathcal{O}\Big(M + \sum_{e\in\mathcal{E}} m_e\log(m_e)\Big), \tag{2.3}$$

where M is the number of isotopologues in the optimal p-set and m_e is the number of subisotopologues involved in the calculations. Recall that following Eq. (2), asymptotically $\mathcal{O}\Big(\sqrt{\prod_{e\in\mathcal{E}} n_e^{i_e-1}}\Big)$. It also follows that $m_e = \mathcal{O}(\sqrt{n_e^{i_e-1}})$. The pending question is, which of the two terms in Eq. (2.3) dominates the calculations. In particular, we are interested when the first term dominates,

$$\mathcal{O}\Big(M + \sum_{e\in\mathcal{E}} m_e\log(m_e)\Big) = \mathcal{O}(M),$$

as this assures that the complexity if optimal, being linear.

The precise answer to this question requires some assumptions on the numbers of atoms of different elements that can appear in the compound. Observe, that monoisotopic elements do not influence at all the complexity of the problem, as $i_e - 1 = 1 - 1 = 0$. Thus, we should consider a chemical compound composed out of $|\mathcal{E}| \geq 2$ polyisotopic elements. Let us also focus first on chemical formulas containing more than one element. Let us assume that the considered counts of atoms n_e all grow polynomially with some hidden parameter n, possibly with different degrees, $n_e = \mathcal{O}(n^{\mathtt{deg}_e})$, leading to $M = \mathcal{O}(n^{\sum_{e\in\mathcal{E}} \mathtt{deg}_e \frac{i_e-1}{2}})$ and $m_e \log(m_e) = \mathcal{O}(n^{\mathtt{deg}_e \frac{i_e-1}{2}} \log n)$. Clearly, the degree of M must then be greater than that of $m_e log(m_e)$, as the logarithm grows slower than any function n^ϵ, with $\epsilon > 0$.

The assumption on the growth of the atomic content of molecules we make is true in most realistic cases. For instance, consider the *averagine* – a model of an averaged protein (Senko *et al.*, 1995). The averagine is known to well approximate the atomic composition of any protein, and works reasonably well due to the fact that proteins are just long chains of amino acids. The chemical formula of an averagine composed of n proteins in chain is given by

$$\mathrm{C}_{\lfloor n\times 4.9384\rfloor}\,\mathrm{H}_{\lfloor n\times 7.7583\rfloor}\,\mathrm{N}_{\lfloor n\times 1.3577\rfloor}\,\mathrm{O}_{\lfloor n\times 1.4773\rfloor}\,\mathrm{S}_{\lfloor n\times 0.0417\rfloor}.$$

Therefore, at least for proteins, $\mathtt{deg}_e = 1$.

Now, let us consider the case of a molecule built up from atoms of only one polyisotopic element. As pointed out in the main manuscript, in that case the notions of subisotopologue and isotopologue coincide and there is no merger step. However in that case too we can apply the layered concept exposed in Algorithm 3, this time directly to the multinomial distribution. Again, if the number of atoms was m, then generating them would take $\mathcal{O}(m)$. The layered complex cannot be applied in the general context and the estimate given by Eq. (2.3). Note that it is crucial for the subisotopologue generator to return subisotopologues in a decreasing order of their probability. The `IsoSpec` algorithm thus offers a huge difference compared with the theoretical results obtained for the `ecipex` algorithm (Ipsen, 2014).

Theoretical questions aside, the implementation of `IsoSpec` offers huge time savings compared to other available software, as shown in the next section.

Experimental Results

We perform runtime analysis on a set of more than 800 000 ions' formulas generated from a list of 1000 human proteins from Uniprot. This set of formulas contains 1000 precursors and all derived b and y ions. This computational experiment therefore simulates the spectra preparation step for a tandem MS database driven identification procedure.

Both `enviPat` and `IsoSpec` are implemented in C++. That said, `enviPat` can only be

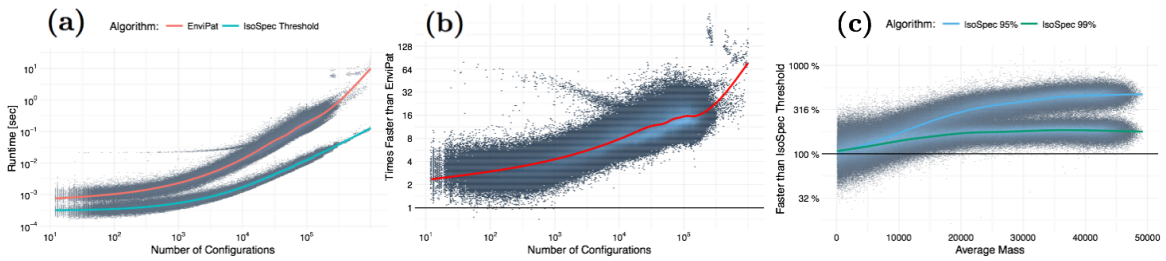


Figure 2.10: Comparison of `enviPat` and `IsoSpec` Threshold **(a,b)** and of `IsoSpec` Threshold with `IsoSpec` calculating the optimal 99% and 95% sets. Absolute peak height threshold was set to equal one ten-thousandth of the highest peak height (enviPat default). Fig. **(a)** shows the absolute runtime as a function of the overall number of calculated configuration. In Fig. **(b)**, we express the relative runtime of `enviPat` in the runtime of `IsoSpec` Threshold, showing how much faster is our approach. Both axis in **(a,b)** are in logarithmic scales. Fig. **(c)** shows how much faster is the calculation of the optimal 99% and 95% sets (with `IsoSpec`) than obtaining the set of isotopologues more probable than 10^{-4} HP (with IsoSpec-threshold). In contrast to **(a)** and **(b)**, the abscissa states the average mass of a compound, as the number of configurations (isotopologues) is variable for the different sets. Smooth lines represent fitted polynomial trend lines in all plots. The analysis is based on 805,367 compounds.

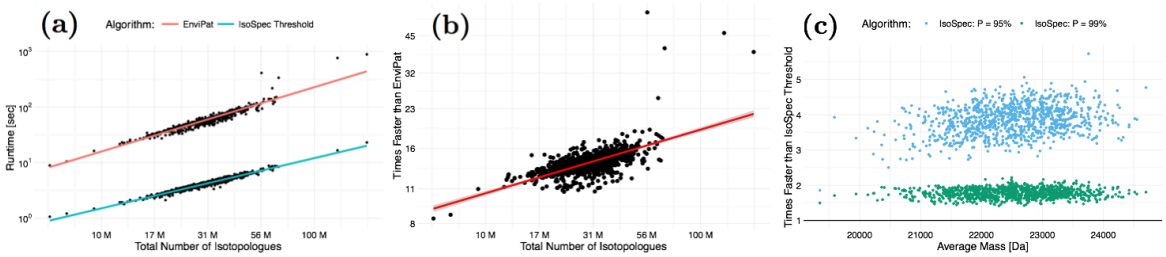


Figure 2.11: Comparison of `enviPat` with `IsoSpec` Threshold **(a,b)** and `IsoSpec` Threshold with `IsoSpec` aiming at joint probability equal to 99% and 95% **(c)** on *fragment identification* problem (1000 compounds). In **(a)** we see the absolute runtimes of `enviPat` and IsoSpec Threshold: **(b)** specifies how much faster is the second approach in terms of the runtimes of the first one. On the x-axis of the **(a,b)** plots we show the total number of configurations generated in the tandem MS theoretical simulation for a given protein. Both axes are in logarithmic scales. In **(c)** one notices speedup resulting from a search for the optimal 95% and 99% sets (`IsoSpec`) instead of the set of isotopologues more probable than 10^{-4} HP (IsoSpec-threshold).

called from R and `IsoSpec` can be called from C++, C, R and Python. We have used the Python interface in our simulations.

In Fig. 2.10 **(a,b)** we compare runtimes of `enviPat` and IsoSpec Threshold on individual fragments. Both tools aim at calculating the same set of isotopologues defined by a common threshold on peak height, equal to one ten-thousandth of the highest peak, 10^{-4}HP for short. Fig. 2.10 **(a)** reports absolute runtimes in seconds. Fig. 2.10 **(b)** expresses `enviPat`'s runtime in that of IsoSpec Threshold to show directly how much faster is the latter, which is roughly 2 to more than 100 fold, the gap widening with the size of a molecule. The optimal 99% and 95% sets are always smaller than the set of isotopologues more probable than 10^{-4}HP and can be usually obtained faster using `IsoSpec`, as can be seen in Fig. 2.10 **(c)**. The advantage clearly increases with compound size. This opens way for various *rapid scan* procedures that could compare the actual spectrum with a relatively small optimal p-set to rule out that a given compound is there.

The overall time to compute spectra for a CID identification procedure for a given substance is the sum of runtimes needed to obtain the spectra of the precursor and all

fragments. We report these total runtimes in Fig. 2.11, which simply aggregates information conveyed in Fig. 2.10. In particular, subfigure **(a)** confirms that a procedure based on `IsoSpec` will be at least an order of magnitude faster as compared to `enviPat`.

`IsoSpec` can be freely downloaded under a 2-clause BSD license from `http://matteolacki.github.io/IsoSpec/`. It can be also downloaded from Python Package Index.

Further Applications

The concept of the isotopic distribution is used in all of the projects described in the next chapters. In particular, it is highly useful for signal deconvolution. Potentially, it can be also used to estimate the natural frequencies of isotopes appearing in the samples. Finally, we will be using it to estimate the number of observed ions.

The `IsoSpec` project is continuously maintained and developed. The ideas that are being now implement include:

1. getting rid of the queue data structure
2. memoization of values of critical mathematical functions
3. parallelizing the code
4. simplification of the interface

Initial tests show that 1 and 2 considerably speed up the calculations. Finally, let us mention that the project will be included in the `OpenMS` platform (`https://www.openms.de/`) as a low level function.

The `IsoSpec` has been recently used in the context of drawing identically distributed random samples. The algorithms offers both theoretical and practical speed ups. For details, please see Startek (2016).

Other work in progress include the calcution of the precise value of a hypothesis test, the Tanimoto index, whose goal is to test for the independence of two binary vectors. This sort of problem appears naturally in the context of the classification of chemical reactions described by a set of binary molecular descriptors.

Algorithm 1 The subgenerator

INPUT:

Multinomial distribution parameters.

OUTPUT:

A sequence of subisotopologues in decreasing order of probability.

PQ = empty max-priority queue
α = the most probable subisotopologue found in hill climb
V = $\{\alpha\}$ (set implemented as a hash table)
PQ.push(α, priority = $\mathbb{P}(\alpha)$)
while PQ is not empty **do**
    $\beta = PQ$.pop()
    for all neighbors n of β **do**
        if $n \notin V$ **then**
            V.add(n)
            PQ.push(n, priority = $\mathbb{P}(n)$)
        end if
    end for
    yield β
end while

Algorithm 2 The QuickTrim algorithm.

INPUT:

A - array of isotopologues

p - the desired total probability of selected isotopologues

OUTPUT:

A - permuted in-place by the algorithm to minimize k

k - lowest integer such that $\sum_{i=0}^{k} \mathbb{P}(A[i]) \geq p$ and $\forall i \leq k \; \forall j > k \;\; \mathbb{P}(A[i]) \geq \mathbb{P}(A[j])$

$sum = 0$
$start = 0$
$end = \text{length}(A) + 1$
while $start \neq end$ **do**
  $idx_{\text{pivot}} = \text{SelectPivot}(start, end, A)$
  $A[idx_{\text{pivot}}] \leftrightarrow A[end - 1]$
  $idx_{\text{lower}} = start$
  for $i \in \{start, .., end - 2\}$ **do**
    if $\mathbb{P}(A[i]) > \mathbb{P}(A[end - 1])$ **then**
      $A[i] \leftrightarrow A[idx_{\text{lower}}]$
      $idx_{\text{lower}} + +$
    end if
  end for
  $A[idx_{\text{lower}}] \leftrightarrow A[end - 1]$
  $psum = \sum_{i=start}^{idx_{\text{lower}}} \mathbb{P}(A[i])$
  if $psum < p$ **then**
    $start = idx_{\text{lower}} + 1$
    $p = p - psum$
  else
    $end = idx_{\text{lower}}$
  end if
end while
Return last value of $start$ as k

Algorithm 3 The `IsoSpec` algorithm.

INPUT:

A molecule consisting of n different elements

Table of isotopic frequencies of elements

p - the target probability

OUTPUT: Optimal P-set.

$sum = 0$

$S_i =$ Initialize subgenerators for each element

$\alpha =$ the most probable isotopologue

$layer_{\text{next}} =$ an empty FIFO (or FILO) queue

$layer_{\text{next}}.append(\alpha)$

$ret =$ an empty list

while $sum < P$ **do**

$accepted =$ an empty list

$sum_{\text{accept}} = 0.0$

$layer_{\text{curr}} = layer_{\text{next}}$

$layer_{\text{next}} =$ an empty FIFO/FILO queue

$threshold =$ subsequent threshold

while $layer_{\text{curr}}$ is nonempty **do**

$I_{\text{curr}} = layer_{\text{curr}}.pop()$

if $\mathbb{P}(curr) < threshold$ **then**

$layer_{\text{next}}.push(I_{\text{curr}})$

else

$accepted.append(I_{\text{curr}})$

$sum_{\text{accept}} = sum_{\text{accept}} + \mathbb{P}(I_{\text{curr}})$

for $next \in S.successors(I_{\text{curr}})$ **do**

$layer_{\text{curr}}.push(I_{\text{curr}})$

end for

end if

end while

$sum = sum + sum_{\text{accept}}$

if $sum \geq p$ **then**

Terminate algorithm.

Return $ret +$ QuickTrim($accepted$).

end if

$ret.extend(accepted)$

end while

3

Quantifying Electron Transfer Reactions

"Science never solves a problem without creating ten more."

— George Bernard Shaw

In recent years, there has been growing interest in electron-based dissociation (ExD) – primarily electron capture (ECD) (Zubarev *et al.*, 1998) and electron transfer dissociation (ETD) (Syka *et al.*, 2004) in protein mass spectrometry. These fragmentation methods allow the cleavage of the backbone of a protein or peptide without significantly disrupting other bonds (even preserving noncovalent interactions) and as such, much effort has gone into the use of ExD methods for top-down sequencing, as well as the study of labile post-translational modifications and even binding sites of non-covalent ligands (Garcia *et al.*, 2007; Håkansson *et al.*, 2001; Ayaz-Guner *et al.*, 2009; Ge *et al.*, 2009; Tsybin *et al.*, 2011; Fornelli *et al.*, 2012; Cournoyer *et al.*, 2005; Li *et al.*, 2010b; Xie *et al.*, 2006; Jackson *et al.*, 2009; Yin and Loo, 2010; Göth *et al.*, 2016). Additionally, considerable efforts have been made to determine preferential reaction pathways and cleavage sites in ExD of known precursors, to obtain insight into gas-phase protein/peptide conformation (Breuker *et al.*, 2002; Oh *et al.*, 2002; Skinner *et al.*, 2012, 2013; Zhang *et al.*, 2011, 2013, 2014; Lermyte *et al.*, 2014; Lermyte and Sobott, 2015; Zhang *et al.*, 2016; Lermyte *et al.*, 2017) as well as to investigate the reaction mechanism (Tureček, 2003; Tureček and Syrstad, 2003; Chung and Tureček, 2010). Ideally, reaction products are not only identified, but also quan-

tified in these efforts. Because of the information-rich nature of top-down ExD spectra, data processing is usually performed with the help of specialized software.

The first, and arguably most critical step in this data processing is usually spectral deisotopisation, i.e. reducing the multitude of signals observed in the m/z dimension due to various charge states and isotopologues to a minimal set of components and abundances. Most of the readily available software tools for this – e.g. `THRASH` (Horn *et al.*, 2000), `MASH` (Guner *et al.*, 2014; Cai *et al.*, 2016), `DeconMSn` (Mayampurath *et al.*, 2008), `Decon2LS` (Jaitly *et al.*, 2009) – utilize an averagine-scaling approach (Senko *et al.*, 1995) to determine charge states, monoisotopic masses, and ion intensities. As this requires resolution of the (aggregated) isotope peaks, these tools are mostly used to process FTICR or Orbitrap data, particularly as they can natively process Bruker and/or Thermo data files (in fact, a modified `THRASH` algorithm, called `SNAP`, is built into the Bruker DataAnalysis software).

Observed isotope clusters are often composed of multiple overlapping isotope distributions (envelopes), each generated by ions whose chemical formulas differ by one (or a few) hydrogen atoms. These shifts (by an integer number of hydrogen masses) are commonly observed in ExD spectra and provide information on reaction pathways (Lermyte *et al.*, 2017; O'Connor *et al.*, 2006; Tsybin *et al.*, 2007). As such, it is desirable to preserve the information contained in observed isotope distributions during and after the deconvolution procedure.

Thus, there is a need for software tools which are able to process high-resolution tandem MS data from a variety of instruments, utilize the high-resolution information (e.g. assign highly resolved peaks) to perform thorough data analysis, and provide the user with information regarding preferred cleavage sites and relative probabilities of competing reaction pathways. Ideally, this should not require the user to possess extensive expertise regarding statistics and/or gas-phase ion/ion chemistry. Recently, we have demonstrated the use of an in-house developed software for deconvoluting complex isotope clusters occurring in top-down ETD spectra acquired on a Waters Synapt G2 Q-IM-TOF instrument (Lermyte *et al.*, 2015a). Furthermore, we have shown how this allows us to infer branching ratios and how this correlates to collision cross-sections and gas-phase conformations of ubiquitin (Lermyte *et al.*, 2017). Here, we present in detail the above computational workflow, together with extensions that shed further light onto the electron transfer driven reactions. The Python implementation of that workflow, called `MassTodonPy`, is made publicly available for download via the Python Package Index.

In the remaining part of this chapter, we shall describe the stages of the proposed workflow: (1) the preprocessing of the spectrum, (2) the generation of potentially observable chemical formulas, (3) the deconvolution of spectra, which involves the estimation of the intensities of the potential products of the considered set of reactions, (4) the pairing of

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

Table 3.1: Considered chemical reactions. M stands for either a precursor ion or a fragment ion. The HTR reaction can happen only after ETD and consists in the transfer of a hydrogen atom from the c to the z fragment.

fragment ions, resulting in estimates of the probabilities of the considered reactions and fragmentations. The workflow was tested *in silico* and on around 200 mass spectra. Finally, we mention some possible extensions to the workflow.

Materials & Methods

Experimental mass spectra.

Two sets of data were acquired under varying experimental conditions using the Synapt G2 and LTQ Orbitrap Velos instruments. In case of the Synapt G2, a subset of experimental settings included 38 different wave velocities in a range of values between 10 to 6000 m/s, while keeping the wave height fixed at 1.5 V; another subset consisted of 14 values of wave height ranging from 0 to 1.5 V, with wave velocity fixed to 300 m/s. In case of the LTQ Orbitrap Velos mass spectrometer, we have collected spectra from a range of different reaction times (from 0.03 to 100 ms), for two different isolation windows in MS1 (selecting 6+ and 9+ precursor ions), and applying different levels of preactivation and supplemental activation. For more details, please refer to our previous publication (Lermyte *et al.*, 2015a).

Data Preprocessing

We assume that the input spectrum was already calibrated. The spectrum should not be centroided, as `MassTodon` does its own centroiding, as described later in the peak picking section.

To mitigate the possibility of fitting to noise peaks, some parts of the mass spectrum need to be trimmed out. We offer two simple ways to do this. The first way focuses on the intensity of individual peaks and amounts to trimming out peaks with intensity below a user-provided threshold. The second way retains only the highest peaks whose joint intensity covers the user-specified percentage of the total intensity in the spectrum. To make that idea more clear, consider a spectrum comprised of three peaks with intensities equal to 1000, 990, and 10. Also, set the joint threshold at 99%. The intensity of the first peak

amounts to $\frac{1000}{1000+990+10} = 50\%$ of the entire intensity in the spectrum. The joint intensity of the two highest peaks amounts to $\frac{1000+990}{1000+990+10} = 99.5\%$ of the overall intensity. It is the smallest set of highest peaks that jointly surpass the required threshold of 99% and so only these peaks are left, and the third one is trimmed out. Observe that the same effect would be achieved if we were trimming out peaks with intensities higher than any number higher than 10 and smaller or equal to 990. For each run of the second trimming spectrum we calculate that implicit cut-off and store it for inspection by the user.

Finally, the mass to charge ratios are adjusted to better match the theoretical spectra, as described later on.

Generating chemical formulas.

`MassTodon` exhaustively finds the formulas of all molecular species that might be present in the set of considered reactions. The theoretical envelopes of these molecules are then fitted to the spectral data at a later stage.

The presented workflow considers a set of known chemical reactions occurring under ETD conditions, c.f. Table 3.1. The Proton Transfer Reaction (PTR) and the non-dissociative Electron Transfer Dissociation (ETnoD) do not result in any fragments; they affect the charge state and the mass of the cation alone. The Electron Transfer Dissociation (ETD), potentially followed by the transfer of a hydrogen between fragments (HTR), result in c and z fragments (Roepstorff and Fohlman, 1984). We assume that PTR and ETnoD may occur multiple times on the same ions, including the c and z fragments. We assume that fragments cannot further fragment, as the internal fragments are scarcely ever observed experimentally in ETD. The number of fragments depends on the charge of the precursor isolated during MS_1, denoted Q, the amino acid sequence and the existing modifications. We neglect the ordering of reactions within one pathway. Thus, the product of the PTR reaction followed by the ETnoD reaction is the same as the product of the ETnoD reaction followed by the PTR reaction. In general, reaction pathways leading to the same product are indiscernible until the last stage of the algorithm.

Every molecular species is described by its elemental composition and charge q. Each reaction (except HTR) consumes one charge. During ETnoD, the radical passes from anion to cation reducing its charge without significantly changing its mass (we neglect the mass of the electron). This motivates the introduction of an additional quantity, the *quenched charge* g, that describes the number of extra hydrogen masses with respect to precursor's hydrogen content, see Lermyte *et al.* (2015a). An increase in g corresponds to an increase in one atomic mass unit and does not change the charge state.

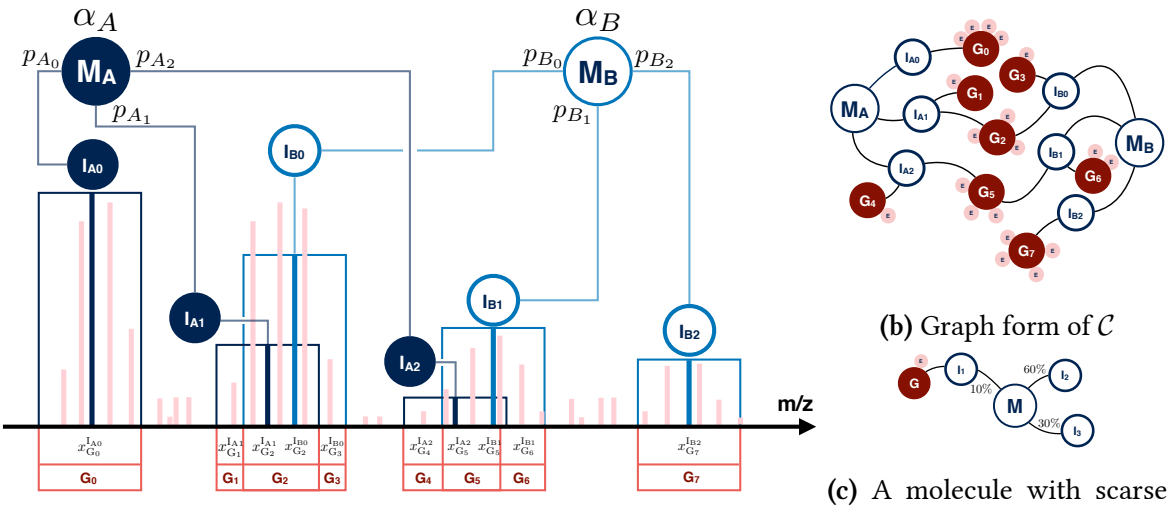
To exemplify the above concept, consider triply charged Substance P, with amino acid sequence $\texttt{RPKPQQFFGLM}^{3+}$. The mass of its monoisotopic isotopologue equals 1347.712 u,

when rounded to the third decimal place. Add the mass of two protons and one quenched charge and divide it by the two present charges to get $\frac{1347.712+3\times1.008}{2} = 675.368$ Th. Thus, the regions of the mass spectrum close to that value can contain ions belonging to that molecular species. Consider the case of isolating only triply charged precursors during the MS1, i.e. selecting ions with m/z around 450.245 Th. One then knows that these ions must have undergone exactly one ETnoD step. This is because ETnoD reduces their charge by one, $3^+ \longrightarrow 2^+$, and increases the number of quenched charges by one, see Table 3.1. Further on we show how to infer the number of reactions both from precursor ions and fragments.

While studying the above example, it is important to note that other sources of ions can explain the same peak. In particular, consider the second most probable isotopologue of the precursor ion that underwent the PTR reaction. One of the ^{12}C carbon atoms in this isotopologues is exchanged for a heavier isotopic variant, ^{13}C. These ions are only slightly less likely to be found in the sample than the monoisotopic ions: on average in 29.6% of cases for this ion versus 43.1% for the monoisotopic peak. Their mass is 1348.716 u. When equiped with two charges, their m/z equals $\frac{1348.716+2\times1.008}{2} = 675.366$ Th. Most instruments would not resolve the 0.002 Th difference between the two molecular species. However, confusing the two ions sources leads to a poor estimate of the relative extent of PTR versus ETnoD. Based on one peak alone it is impossible to correctly identify the relative proportions of different molecular species. In most cases, it is possible to differentiate between various molecular species by looking at their isotopic distributions as a whole. This opens the possibility to evaluate how much of observed intensity can be attributed to particular ions. Further on we show how this can be achieved.

Observe that the quenched charge may also be used to record information on a hydrogen radical transferred during HTR. This is convenient, as there is no real difference between a quenched charge and a *regular* hydrogen atom within one molecule. Consider then a precursor that undergoes a direct HTR reaction: the c fragment must then have a number of quenched charges equal to -1, which we consider a valid possibility. It is also the only case when this quantity assumes a negative value.

During the fragmentation, the remaining charge and quenched charge (if positive) are distributed among the fragments. One might expect the charge state of smaller fragments to be limited, due to Coulomb repulsion. For this reason, `MassTodon` omits formulas with too many charges per a given number of amino acids. By default, we assume that each two charges must be four residues apart. In case of Substance P, this means that we could not observe a c_3 fragment with two charges simply because it is composed out of only three residues. On the other hand, we assume it might be possible to observe a c_5 fragment, presumingly with charges placed on its first and last residue. The charge distance parameter



(a) $\mathcal{C}$, a connected component of $\mathcal{G}$

(b) Graph form of $\mathcal{C}$

(c) A molecule with scarse experimental support

Figure 3.1: A connected component $\mathcal{C}$ of the *deconvolution graph* $\mathcal{G}$. Experimental peaks are shown in pink. Among the nodes of $\mathcal{G}$ we find the molecules M, their isotopologues I, and experimental groups G. The probability p of meeting I among the M ions decorates the edge between I and M. Edges between I and G are not plotted for clarity in (a); we do mark however their corresponding flow variables, x. They denote the amount of experimental intensity attributed to a given isotopologue. The aim of the deconvolution is to establish total intensities of M_1 and M_2, denoted respectively as α_A and α_B. In (b) we show $\mathcal{C}$ as a graph. The experimental peaks (in pink) are depicted only for clarity of the representation and are not actually in $\mathcal{G}$. In (c) we show a molecule M with scarse experimental support: meeting an isotopologue paired with an experimental group would occur in one in ten cases only.

can be adjusted by the user.

If one considered only the PTR and ETnoD reactions, the precursor molecule could result exactly in $\frac{Q(Q+1)}{2}$ different molecular species. Each product can be further fragmented into pairs of different c and z fragments. The number of such pairs is $K - K_P$, i.e. the number of amino acids in the provided sequence, minus the number of prolines, that cannot be fragmented easily by electron transfer due to their ring structure. Then, each fragment can again undergo several PTR and ETnoD reactions. The number of all fragments is thus of the order of $\mathcal{O}(KQ^4)$.

Generating the isotopic distributions. The isotopic distribution of a given molecular species models the expected signal one could register in the mass spectrometer. One would expect that peaks assigned to one molecular species would follow a pattern similar to the isotopic distribution, which would imply that observed intensities should follow certain proportions.

Each reaction product is described by its elemental composition, charge q, and quenched charge g. This information is sufficient to generate the theoretic isotopic distribution using any isotopic calculator. To perform calculations here, we use the `IsoSpec` algorithm (Łącki *et al.*, 2017b), described in Chapter 2. Given the elemental composition, `IsoSpec` produces a series of infinitely resolved isotopologues, represented as tuples (mass, probability). Racall, that to avoid the combinatorial explosion in their number (Valkenborg *et al.*, 2012), `IsoSpec` reports only the smallest possible set of peaks, such that their cumulative probability does

not fall under some user specified threshold, e.g. 99.9%. The masses of the envelopes are adjusted according to formula $\frac{m+q+g}{q}$ to obtain valid mass-over-charge ratios.

Because of the use of infinitely resolved peaks, our workflow can be adapted to model outcomes of ETD on instruments that offer different degrees of resolution. In particular, to model low-resolution spectra, one does not need infinitely resolved theoretical envelopes. Whenever small differences between the m/z ratios cannot be discerned, one can safely aggregate peaks with similar m/z ratios. This can be advantageous, as a smaller number of peaks makes the deconvolution problems smaller (as shown later on) and quicker to solve. On the other hand, by lowering the resolution one introduces additional variance to the estimates of the total intensities with which the molecular species appear in the spectrum. Also, some highly resolved peaks may be specific to a smaller number of substances. Losing that information by unnecessary aggregation would render the deconvolution considerably more difficult. In the current workflow, we ask the user to provide a measure of the instrument's resolution in terms of one parameter alone – the peak's m/z tolerance tol. Experimental peaks are deemed to potentially originate from a molecule M if their m/z ratios are within the tol distance from a theoretical isotopologue I of that molecule. This is shown in Figure 3.1a. By default, we assume that differences between m/z ratios an order of magnitude smaller than tol cannot be discerned. This implies a finite granularity of the spectrum: if tol amounted to 0.05 u, then the smallest difference between peaks would be that of 0.001 u. To obtain such spectrum, peaks with the same first three significant digits are aggregated, i.e. they are represented by one peak with the same rounded m/z and intensity equal to the total intensity of these peaks. In general, given tolerance tol, we round the spectrum to the significant digit given by $\lceil -\log_{10}(tol) \rceil$ and then aggregate it. By convention, we call the so obtained cluster of isotopologues an isotopologue. The same operations are performed on the experimental spectrum.

Peak picking. The aim of the peak picking is to assign peaks in the mass spectrum to the potential molecular species. This is done by comparing the m/z ratios of the experimental peaks with those of the peaks in the theoretical isotopic envelopes, as described in the previous section and visualized in Figure 3.1a. Figure 3.1a also shows that finding potential explanations for a given experimental peaks corresponds to finding all intervals of the form $[\frac{m}{z} - tol, \frac{m}{z} + tol]$ to which its m/z value belongs. To find these intervals effectively, we make use of the interval trees data structure (Cormen, 2009).

Different intervals might overlap, as is the case for isotopologues I_{A1} and I_{B0} in Figure 3.1a. The intersections of these intervals partition the m/z axis into regions that can be traced back to originate from different sets of molecules and regions that cannot be explained by any of the products of the considered reactions. Experimental peaks inside such intersections (there might be more than one) form experimental groupings G. The

total intensity within one such grouping is stored and denoted by $G_{\texttt{intensity}}$. After these operations, the experimental peaks do not play any more role in calculations and can be deleted.

Considered together, molecules M, their isotopologues I, and the experimental groupings G form nodes of the *deconvolution graph*, $\mathcal{G}$, as shown in Figures 3.1a and 3.1b. In $\mathcal{G}$, molecule nodes M are naturally joined with their isotopologue nodes I, that are in their turn joined with experimental groupings G they could explain. The graph $\mathcal{G}$ is usually composed of several connected components, like the one presented in Figure 3.1a. Note that higher tol parameter (check previous section) results in a lower number of both the G and I nodes.

While picking the peaks, one can easily spot molecules M with poor experimental support. More precisely, if the sum of probabilities of isotopologues of M connected to some G does not exceed some percentual threshold P (by default, 70%), then we can discard it. For instance, we would discard the molecular species shown in Figure 3.1c, as one would expect that only in one case in ten would one of its isotopologues ever appear close to an experimental group. This additional preprocessing eliminates substances that alone could not explain more than the P percent of the total experimental intensity within the considered subproblem, and thus makes part of the overall variable selection procedure we consider.

Each connected component of $\mathcal{G}$ gives rise to some deconvolution problem, as several molecules might compete for the explanation of the given range of the mass spectrum. These problems might be solved independently and simultaneously rather than sequentially. `MassTodonPy` offers both ways of performing these calculations.

Deconvolution. The problem of deconvoluting the intensities within one connected component of graph $\mathcal{G}$ is reminiscent of linear regression. Indeed, the goal is to express the observed signal as a weighted sum of the isotopic envelopes. One weight, denoted by α as in Figure 3.1a, can be interpreted as the total intensity of a given molecular species in the entire mass spectrum. In particular, α cannot be negative. This restriction also partially alleviates some problems with which the ordinary least squares regression would struggle, such as the collinearity of the predictors. In our setting, the collinearity would correspond to a high (positive) correlation between the shapes of different isotopic distributions. This problem is mitigated, because under the nonnegativity constraints there is simply much less space for linear dependence (Davis, 1954).

The approach we take is similar to the one proposed by Slawski *et al.* (2012). That approach also relies on non-negative least squares. However, we use a different approach to model mass inaccuracy itself: instead of assuming that mass inaccuracy follows the gaussian distribution, we assume that mass can be spread around the mass tolerance regions. Our approach should be better fitted to cases of spectra that are not perfectly calibrated.

In advance, one does not know how to redistribute the intensity of I among the neighboring experimental groupings G. This motivates the introduction of the *flows* between G and I, denoted by x_G^I, For instance, in Figure 3.1a isotopologue I_{B0} is linked with experimental intentensities G_2 and G_3. It absorbs $x_{G_2}^{I_{B0}}$ of the intensity of G_2, and $x_{G_3}^{I_{B0}}$ of the intensity of G_3. I_{B0} should contribute $x_{G_2}^{I_{B0}} + x_{G_3}^{I_{B0}}$ to M_B. On the other hand, this should be equal to a fraction p_{B_0} of the total intensity of M_B, denoted by α_B. In other words, $p_{B_0}\alpha_B = x_{G_2}^{I_{B0}} + x_{G_3}^{I_{B0}}$. In general, the intensities of isotopologues I and molecules M are related via a set of linear restrictions $\alpha_M p_M^I = \sum_{G:G\leftrightarrow I} x_G^I$, where under the sum we iterate over all experimental groups G that neighbor isotopologue I.

It is sensible to choose molecular intensities α and isotopologue intensities x to assure a minimal divergence between the observed group intensities $G_{\texttt{intensity}}$ and the total outflows of intensity from these nodes towards the isotopologue nodes. The overall deconvolution problem can thus be formalized as

$$\min_{x,\alpha} \sum_G \left(G_{\text{intensity}} - \sum_{I:G\leftrightarrow I} x_G^I\right)^2 \quad \text{so that}$$
$$\alpha_M p_M^I = \sum_{G:G\leftrightarrow I} x_G^I, \quad x_G^I \geq 0$$

To minimize the risk of numerical instability and perform model selection one can include in the cost function additional penalty terms (James *et al.*, 2013),

$$L_1^x \sum_{G\leftrightarrow I} x_G^I + L_1^\alpha \sum_M \alpha_M + L_2^x \sum_{G\leftrightarrow I} (x_G^I)^2 + L_2^\alpha \sum_M \alpha_M^2.$$

By default, we set L_1^α, L_2^α, L_1^x, and L_2^x to 0.001. The penalty terms after L_1^α and L_1^x should round small estimates to zero, as in the lasso model selection approach (James *et al.*, 2013). The above problem can be efficiently solved with quadratic programming. MassTodon relies on the `CVXOPT` Python module (Andersen *et al.*, 2013) that solves quadratic programs with a path following algorithm.

After each deconvolution, we calculate and report various error statistics. These include the sum of the absolute values of the errors, the sum of overestimated values, and the sum of the underestimated values. The above quantities are also divided by the total ion current or the total intensity within the tolerance regions of any of the theoretically molecular species.

The cost function is minimized simultaneously in xs and αs. Only αs are analyzed in the next, final stage of the algorithm.

Pairing of the observed ions. Up to this step, the algorithm obtained estimates of intensities of each considered product molecule, uniquely defined by its type (precursor, c or z fragment), charge q, quenched charge g. It is relatively easy to estimate the probabilities of PTR and ETnoD reactions alone based solely on the estimates of intensities of precursor ions, without taking into acount the fragments (Lermyte *et al.*, 2017). These can then be used to calculate their odds ratio, which is known to chemists as the branching ratios.

Given a non-fragmented molecular species with charge q and quenched charge g, one can retrieve the numbers of the PTR and ETnoD reactions by solving

$$q = Q - N_{\text{PTR}} - N_{\text{ETnoD}}, \tag{3.1}$$

$$g = N_{\text{ETnoD}}, \tag{3.2}$$

for N_{PTR} and N_{ETnoD}. Eq. (3.1) states that each reaction reduces the observed charge by one. Eq. (3.2) traces the origin of all quenched charges on the precursor molecules solely to the ETnoD reaction.

To estimate the probabilities of ETnoD and PTR we make use of a simple stochastic model. We make use of a simple branching process: we assume that each ion can undergo several PTR and ETnoD events, as shown in Fig. 3.2 Each reaction happens independently and each ion is also treated independently. Denote by I_i the intensity of an i-th group of precursor ions. Let the number of PTR and ETnoD events for that group equal N^i_{PTR} and N^i_{ETnoD} respectively. The intensity is known to relate to the number of observed ions linearly within the dynamic range of the instrument, so that if the actual number of ions is denoted by N_i, then $I_i = CN_i$, where C is what we call the *ion-intensity* exchange rate. We do not have direct access to C: an attempt to estimate it will be described in Chapter 5. We shall now show, that even not knowing C, we can still obtain point estimates of the correct branching ratio.

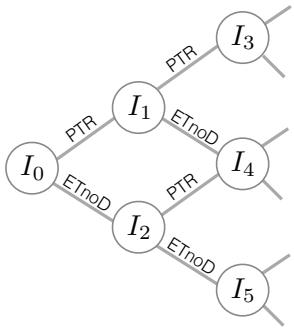


Figure 3.2: Simple branching model. I denotes the estimated intensities.

Observe that given the numbers N_i, the assumptions we made result in a likelihood given by

$$L = \prod_i \left(p_{\text{PTR}}^{N^i_{\text{PTR}}} p_{\text{ETnoD}}^{N^i_{\text{ETnoD}}} \right)^{N_i}$$

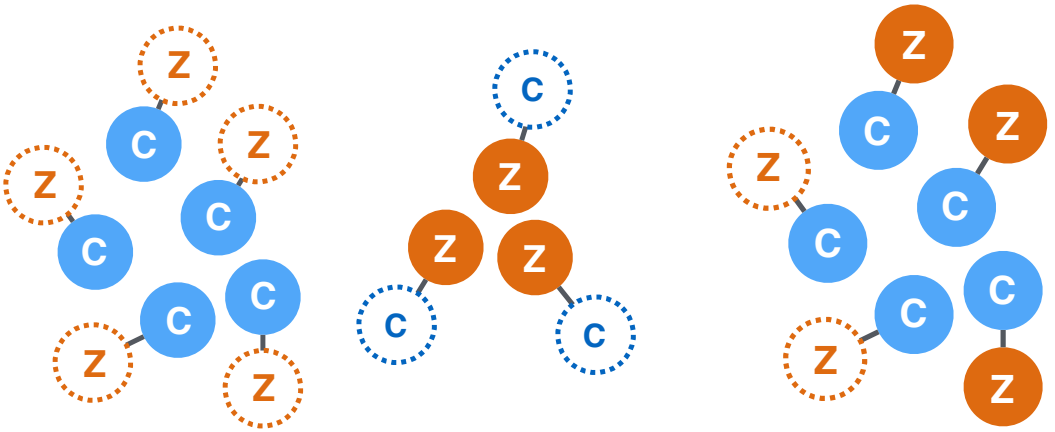
The maximum likelihood estimates then equal

$$\check{p}_{\text{ETnoD}} = \frac{\sum_i N^i_{\text{ETnoD}} N_i}{\sum_i (N^i_{\text{ETnoD}} + N^i_{\text{PTR}}) N_i} \quad \text{and} \quad \check{p}_{\text{PTR}} = \frac{\sum_i N^i_{\text{PTR}} N_i}{\sum_i (N^i_{\text{ETnoD}} + N^i_{\text{PTR}}) N_i}$$

The numerator counts ions that underwent ETnoD, I_i is the estimated intensity of the precursor with charges (q_i, g_i). The denominator additionally contains the count of ions undergoing PTR. Given the relationship $I_i = CN_i$, we could plug in I_i instead of N_i and the result would be the same. This heuristical argument can be made precise assuming independence of the random quantities N_i and C. This leads finally to

$$\hat{p}_{\text{ETnoD}} = \frac{\sum_i N^i_{\text{ETnoD}} I_i}{\sum_i (N^i_{\text{ETnoD}} + N^i_{\text{PTR}}) I_i} \quad \text{and} \quad \hat{p}_{\text{PTR}} = \frac{\sum_i N^i_{\text{PTR}} I_i}{\sum_i (N^i_{\text{ETnoD}} + N^i_{\text{PTR}}) I_i},$$

and so the branching ratio can be estimated as $\hat{BR} = \sum_i N^i_{\text{ETnoD}} I_i / \sum_i N^i_{\text{PTR}} I_i$. The problem with the above method is that the precise expressions for the standard deviations are not



(a) Lavish **(b)** Parsimonious

Figure 3.3: Two interpretations of observing 5 c and 3 z matching fragments: lavish (a) and parsimonious (b). Nodes with dashed edges symbolize cations that never reach the detector. (a) maximizes the number of missing cations needed to explain the spectrum, while (b) minimizes that number.

independent of the *ion-intensity* exchange rate C, and one cannot deal with it without a more elaborate mathematical apparatus.

The above method cannot be directly generalized to include fragments. This is because counts of reactions are not directly accessible and only estimates of the overall intensity of c and z fragments are at hand. To determine the number of fragmentation events, one has to pair back the matching c and z fragments. Pairing should occur only between matching ions: a c_k fragment should be matched only with a z_{K-k} fragment, where K is the total number of amino acids in a given sequence. Moreover, pairing should include natural restrictions on the charge states (q_c, q_z) and quenched charges (g_c, g_z) of both fragments.

There exists a whole range of possible pairing strategies. The two extremes are: (1) to assume that ions come from entirely separate groups of precursors, and (2) that the observed fragments are generated by a minimal number of precursors. For instance, Figure 3.3 shows a situation where 5 c and 3 z matching fragments were observed (filled circles). It might be possible that at the beginning of the experiment there were 8 precursor ions and each out of them undergone a fragmentation and that for each pair of fragments only one of them made it to the detector. This is the *lavish* interpretation, as shown in Figure 3.3a. The question would remain though, why the other fragments were missing. Assuming that this was only because they undergone ETnoD and PTR so many times that their entire charge depleted is possible, but it would also inflate the total number of reactions that must have occured to produce the observed output. Another interpretation could assume that only a minimal number of reactions is necessary to explain the observed output, as shown in Figure 3.3b. Here, a maximal pairing is performed, and only two c fragments have to be paired with z fragments with a depleted charge (dashed circles). This approach is by definition *parsimonious* in terms of reactions needed to explain the experimental results.

The above *principle of parsimony* is implemented in three different algorithms. The key differences between them lie in the definition of a molecular species and the applied optimization scheme, see Figure 3.4. For mathematical and implementational simplicity, all algorithms do not discern between ETD and an ETD followed by HTR. Including HTR would complicate the structure of the pairing graph that we introduce below. The two reactions are considered *jointly*, as if they were one fragmentation reaction. Practical remarks on the use of algorithms will be provided in the next section.

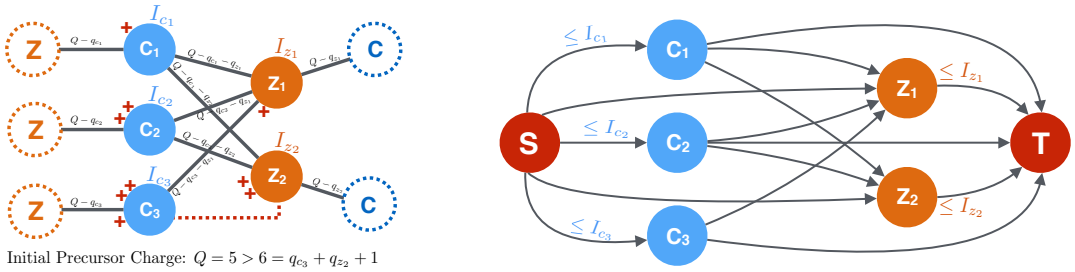
The *basic* algorithm assumes that one can safely disregard the differences in molecular species due to the number of quenched charges they bear. In other words, the estimates of intensities of the c_k fragments with with the same charge q_c but different quenched charges g_c are summed. Similarly, we merge the intensities of z_{K-k} fragments with the same charge q_z but different quenched charges g_z. We then construct the *pairing graph*, as shown in Figure 3.5a. The nodes of the *pairing graph* correspond to different observed molecular species and store information on their total estimated intensity. Special dummy nodes are added to denote the matching co-fragments that had lost all their charge. Our approach assumes that the only way ions can end up being undetected is solely through the total loss of charge. Edges are drawn between c and z nodes with complementary sequences if their total charge plus one (the ETD event neutralizes one charge) does not exceed that of the precursor selected in the MS1 stage of the experiment, $q_c + q_z + 1 \leq Q$.

The pairing of fragments corresponds to the redistribution of the estimated intensities I in the nodes along the edges of the pairing graph. Assigning intensity to an edge diminishes the intensities in both end nodes by the same amount. All intensity must be assigned to some edges. Assigning intensity comes at a cost reflecting the number of reactions the pair of ions underwent during the whole experiment. In the basic approach, fragments with charges (q_c, q_z) together underwent $Q - 1 - q_c - q_z$ reactions. The optimization task we are about to set up lets us ignore the extra fragmentation count, fixing these costs at $Q - q_c - q_z$, equal to the total number of ETnoD and PTR reactions that both fragments underwent, $N^{cz}_{\text{ETnoD}} + N^{cz}_{\text{PTR}}$. Note that this equation holds also for pairings involving co-fragments that entirely disappeared due to the loss of all charge.

The pairing problem turns into an optimization problem where one wants to minimize

Algorithm	charge	quenched charge	network algorithm	rounding small estimates to 0
Basic	✓		✓	
Intemediate	✓	✓	✓	
Advanced	✓	✓		✓

Figure 3.4: Summary of the proposed pairing algorithms.



(a) Pairing Graph **(b)** Max Flow Problem

Figure 3.5: A *pairing graph* (a) and it representation as a *max flow* optimization problem (b). Nodes with dashed edges correspond to ions that lost their charge; other nodes correspond to observed fragments. In (a), charges are shown as red plus signs. Gray edges mark possible pairings. Red dashed line between c_3 and z_2 marks an impossible pairing: if combined, both fragments must have originated from a 6+ precursor, which was not possible. The task is to redistribute the intensity in nodes along the edges. This comes at a cost $Q - q_c - q_z$. To turn (a) into (b), one has to: (1) remove unobserved ion nodes (2) direct remaining edges from c to z (3) add sink S and terminal T (4) add edges directed from S to c nodes and from z nodes to T and add capacities equal to observed ion intensities (5) add edges from S to z fragments and edges from c fragments to T: these correspond to pairings with unobserved ions. This representation is possible for *basic* and *intermediate* pairing algorithms.

the total number of reactions that could have produced the observed c and z fragments. More specifically, we face a constrained linear optimization task:

$$\min_{I_{cz}:\, c\in\mathcal{A}_C,\, z\in\mathcal{A}_Z} \sum_{\substack{c\,\in\,\mathcal{A}_C \\ z\,\in\,\mathcal{A}_Z}} (N^{cz}_{\text{ETnoD}} + N^{cz}_{\text{PTR}}) I_{cz} \tag{3.3}$$

$$\forall_{c\in\mathcal{O}_C}\ I_c = \sum_{z\in\mathcal{A}_Z} I_{cz},\ \forall_{z\in\mathcal{O}_Z}\ I_z = \sum_{c\in\mathcal{A}_C} I_{cz}. \tag{3.4}$$

Above, $\mathcal{O}_C$ and $\mathcal{O}_Z$ denote sets of observed c and z nodes, and $\mathcal{A}_C$ and $\mathcal{A}_Z$ additionally contain the unobserved co-fragments.

The above simplifies to a *max flow* problem: subtract flows between observed fragments from both sides of equalities in (3.4) and what results are the expressions for flows between observed fragments and their unobserved co-fragments. Plugging these into Eq. (3.3) and some simple algebra results in

$$\max_{I_{cz}:\, c\in\mathcal{A}_C,\, z\in\mathcal{A}_Z} \sum_{\substack{c\,\in\,\mathcal{O}_C \\ z\,\in\,\mathcal{O}_Z}} I_{cz} \quad \text{s.t.}$$

$$\forall_{c\in\mathcal{O}_C} I_c \geq \sum_{z\in\mathcal{O}_Z} I_{cz}, \quad \forall_{z\in\mathcal{O}_Z} I_z \geq \sum_{c\in\mathcal{O}_C} I_{cz}.$$

Of course, all flows I_{cz} are non-negative. To solve the max flow problem we use the Edmonds-Karp algorithm (Edmonds and Karp, 1972) as implemented in the `NetworkX` Python module (Hagberg *et al.*, 2008).

The solution to the above problem provides us with estimates of the total intensities of ions undergoing a specific type of fragmentation. In particular, this lets us estimate the probabilities of fragmentation along the protein. It also lets us estimate the probability with which the precursor will fragment. However, this setting does not offer any possibility to

estimate the number of ETnoD and PTR reactions undergone by fragments. These might become important in case of experiments where bigger and more charged substances are studied, or when much of the precursor ions reacted away, mostly through fragmentation.

To provide a solution to the above problems, we have developed another algorithm – the *intermediate* approach. In this approach, we do not aggregate the intensities of observed ions with different quenched charges. As a result, the *pairing graph* contains more nodes, both observed and dummy ones. Had we followed the previous approach, then each observed fragment could match several unobserved co-fragments, all amounting to the same overall number of reactions but differing in specific numbers of ETnoD and PTR. Unfortunately, the existence of many unobservable co-fragments would prevent us from reducing the problem to a *max flow* optimization, making it impossible to derive equations for all flows between observed and unobserved fragments. To solve this problem, we reduce the number of potential dummy nodes by combining them together.

The edges between existing fragments now convey information necessary to tell how many PTR and ETnoD reactions happened on both fragments throughout their history, including the period before any fragmentation occurred. Similarly to equations (3.1) and (3.2), the numbers of PTR and ETnoD reactions on a given pair of fragments characterized by charges (q_c, q_z) and quenched charges (g_c, g_z) follow equations

$$\begin{aligned} N_{\text{PTR}} &= Q - 1 - q_c - q_z - g_c - g_z \\ N_{\text{ETnoD}} &= q_c + q_z. \end{aligned}$$

Note that due to aggregation, the same cannot be said about edges between the observed and unobserved ions. Otherwise said, if a mass spectrum does not contain *pairable* fragments, then the only source of information on the numbers of ETnoD and PTR reactions can be obtained solely from the precursor products.

Finally, we investigated a third solution to the *pairing problem*, the *advanced* approach. It includes the introduction of additional penalty terms to the cost function,

$$\lambda_1 \sum_{\substack{c \in \mathcal{A}_C \\ z \in \mathcal{A}_Z}} I_{cz} + \lambda_2 \sum_{\substack{c \in \mathcal{A}_C \\ z \in \mathcal{A}_Z}} I_{cz}^2.$$

Above, λ_1 corresponds to a lasso-type penalty and λ_2 - a ridge penalty. This approach was investigated mainly for its ability to automatically round the estimates of small flows to zero. The above problem cannot be cast into the *max flow* setting because of the quadratic terms in the cost function. For this reason, we use yet again the general purpose `CVXOPT` solver.

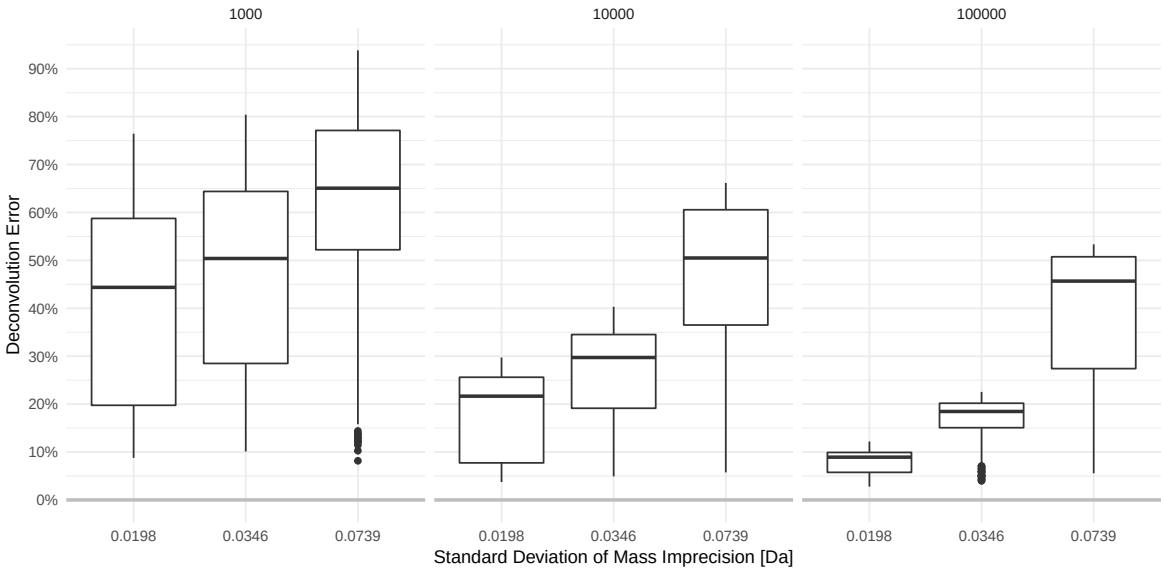


Figure 3.6: Error rates of the deconvolution procedure on *in silico* data for different numbers of initial precursor ions (N = 1 000, 10 000, 100 000) and under three different levels of noise in the mass/charge values (as measured by the standard deviation σ plotted on the x axis). The tolerance interval in `MassTodon` was set to 0.05 Th. To measure error we sum the absolute differences of peak heights and normalize the result to the number of the precursor ions (the result does not need to sum to 100%).

Results & Discussion

In Silico results. In order to test the entire workflow, we conducted *in silico* experiments. A chemical process was simulated using a tailored Gillespie algorithm (Gillespie, 1977a), as described in Algorithm 4. Briefly, the process generates a random series of three chemical reactions (PTR, ETnoD, and ETD; HTR is neglected) occurring in particular moments of time. The length of time intervals between reaction events is random and depends upon the number of charged ions at particular charge state, following McLuckey and Stephenson (1999).

`MassTodon` was tested in various conditions: we checked all the combinations of settings of different initial numbers of precursors, $N = 1000, 10000$, or 100000 ions, initial precursor charges $Q = 3, 6, 9$, and 12, three levels of noise in the mass/charge values (as measured by standard deviation σ), and 12 different sets of probabilities of reactions.

The deconvolution procedures implemented in MassTodon fail in case of extremely noisy spectra, by which we understand either spectra with extremely low ion content or poor resolution, as seen in Figure 3.6. The algorithm works best when there are enough ions to form a well sampled isotopic distribution (in case of our simulations – 100 000 ions). In case of high-resolution mass data, when thousands of isotopologue peaks are present in the mass spectrum, it is advisable to combine spectra from several runs of the instrument to assure that there are enough ions to correctly identify the relative proportions of peaks. It is also vital not to underestimate the size of the tolerance interval. Of course, the above remarks are intrinsic to any peak assigning procedure that uses peak intensities, rather than

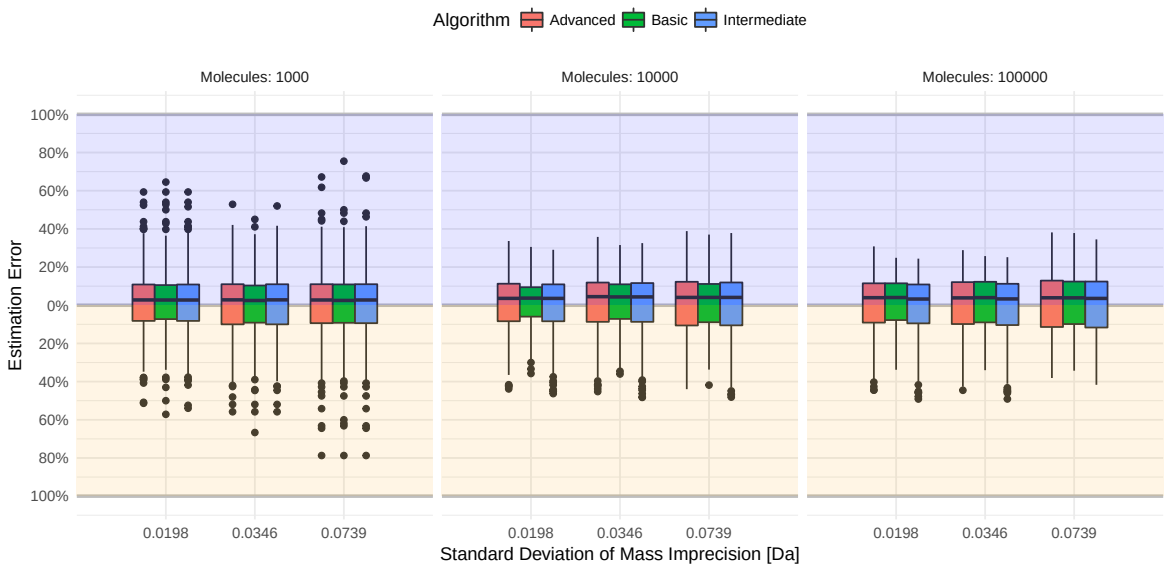


Figure 3.7: The distribution of distance between the estimates $(\hat{p}_{\mathrm{ETnoD}}, \hat{p}_{\mathrm{PTR}})$ and the true values $(p_{\mathrm{ETnoD}}, p_{\mathrm{PTR}})$ for different approaches we take, measured by the euclidean distance normalized to the maximal distance $\sqrt{2}$. Estimates in the blue regions favor PTR, while those in the yellow - ETnoD. The distributions are conditional on the number of initial precursor ions (N = 1 000, 10 000, 100 000) and different level of mass inaccuracy σ (on x axis).

relying solely on their mass-over-charge ratios.

While running simulation described by Algorithm 4 we store the numbers of each molecule M drawn in the process. We have compared these numbers with the estimates of MassTodon to check the quality of the applied deconvolution procedures. Figure 3.6 reports the obtained error rates.

Interestingly, the number of ions in the sample is of limited importance if one is interested in the estimation of the probabilities of ETnoD and PTR reactions, as shown in Figure 3.7. We note, that the parsimonious approach we have taken on average only slightly overestimates values of the true parameters, showing a preference towards the PTR reaction. Note also, that the *basic* approach to the pairing problem seems to offer estimates with the smallest variance.

The above results indicate that it is best to use the *basic* approach to obtain estimates of PTR and ETnoD reactions, and to use either the *intermediate* or *advanced* approaches in cases when the joint intensity of fragments greatly surpasses that of the non-fragmented ions. We would like to mention also, that running all of the above algorithms is fast and takes only a small fraction of the entire workflow's runtime.

Experimental results. Mass spectra have been acquired for purified Substance P and ubiquitin as described in detail in the previous publications (Lermyte *et al.*, 2015a,b).

The outcomes of `MassTodon` can be used to more easily compare mass spectra gathered at different instrumental settings. Figures 3.9 and 3.10 explore the differences and similarities of the information conveyed in different mass spectra, including their percentual content of products of all studied reactions, the probabilities of fragmentation, and intensi-

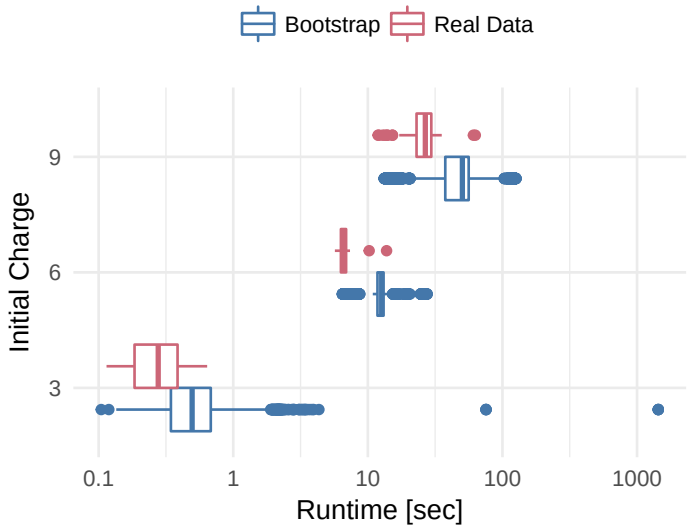


Figure 3.8: `MassTodonPy` runtime distribution. The analysis contains all the stages of the algorithm, including running all three *pairing algorithms*. The 3+ precursors correspond to Substance P spectra; other results are obtained for ubiquitin. Usually, it takes more time to process a spectrum randomly reshuffled by bootstrap than the original version. Runtimes were obtained using the sequential version of the algorithm, which solves the *deconvolution problems* one after another. It is possible to reduce this time for larger problems using the multiprocessing option.

ties and probabilities of the ETnoD and PTR reactions.

`MassTodon` provides point estimates of the above parameters. Given that the analysis of one spectrum is reasonably fast (see Figure 3.8) we decided to rely on bootstrap procedures (Efron and Tibshirani, 1994; Wasserman, 2013) to estimate the standard deviations of the above parameters. In particular, each mass spectrum was randomly reshuffled multiple times. We assume that each bootstrap spectrum is composed out of N ions. The m/z ratios of these ions were then independently drawn among the original ratios, with probabilities equal to the heights of the corresponding peaks, normalized to the total ion current. The number of observed ions in the spectrum N is not truly known in advance. In our simulations, we assumed that the whole spectrum consists of around 100 000 ions. We draw 250 random spectra for each real one and run `MassTodon` on each one of them.

Figure 3.9a shows the overall fitting quality in case of the Substance P spectra. On average, the products of the considered reactions on average account for all but 30 to 40% of the mass spectrum. Shifting our attention only to those regions of the mass spectrum which fall within the range of any potential product, the error estimates drops in a range between 10 to 20%. Note that for spectra gather at wave height fixed at 1.5 V and wave velocity between 700 to 1500 m/s the errors grow significantly.

Figure 3.9b presents the estimates of probabilities of fragmentation for Substance P. Interestingly, the probabilities are almost constant across different experimental settings. They are also almost uniformly distributed along the possible fragmentation sites (proline not being one of them). This is what would be expected of a small molecule, like Substance

P, with a trivial tertiary structure. Again, significant departures from this pattern emerge in the same region of wave velocity.

Figure 3.9c seems to shed some light on the nature of these anomalies. It presents the estimates of the intensity of ions that underwent ETnoD and PTR, which is a proxy for the number of these events to happen on the ions of Substance P within the sample. In particular, it can be noted that the range of wave velocity between 700 to 1500 m/s contains a particularly small amount of ions that could have been assigned to ETnoD or PTR. By comparison, all estimates where these intensities were above 40 000 show a much smaller amount of variance. Note also, that Figure 3.9c suggests that the relative ratios of ETnoD and PTR remain stable under most experimental settings, with the exception of small wave velocities. These ratios can be interpreted as relative probabilities of the ETnoD and PTR reactions, conditional on one of the reactions happening.

Interestingly, a similar pattern re-emerges in mass spectra of ubiquitin, as shown in Figure 3.10. In spectra where the isolated precursor ion was bearing 6 charges, the ETnoD vastly dominates over PTR. In one of our previous papers (Lermyte *et al.*, 2015a) we show, that this might be related to the relatively compact gas-phase conformation of the 6+ protein. In other words, the fragmentation cannot happen because the two fragments remain bound by non-covalent interactions, giving rise to a higher percentage of the ETnoD products.

Conclusions

As high-performance mass spectrometers and the use of ExD methods become more prevalent, there will be an increasing demand for software methods to assist in processing the resulting, considerable amounts of data. Here, we have presented a user-friendly software package to analyze high-resolution ETD data, deconvolute isotope distributions, and infer information about various competing reaction pathways occurring under ETD conditions.

Chapter 5 casts more light on how to fit parts of the entire framework into a Bayesian setting, in order to provide the user with a better understanding of the uncertainties of the estimates and potential correlations of results. In particular, the user might be interested in the some ranges of the spectrum could be alternatively explained by other substances. Obtaining such information could be done by looking at the joint distribution of the counts of molecules that compete for the explanation of a given part of the spectrum.

Moreover, it would be interesting to free the user from the need to specify the tolerance parameter. This should be obtained automatically and potentially vary for different mass-to-charge ratios.

The implementation of the `MassTodon` algorithm is freely available for downloads from

the Python Package Index. Installation instructions and documentation can be found at readthedocs. Sourceh code is available for download from github. The software is distributed under the terms of the GNU AGPL V3 public license.

Algorithm 4 *In silico* spectra generator

INPUT:

A list $\mathcal{I}$ comprising N precursor ions with a given charge Q and sequence F.

Probabilities of reactions $p_{\text{PTR}}, p_{\text{ETnoD}}, p_{\text{ETD}}$.

Overall intensity I of the process.

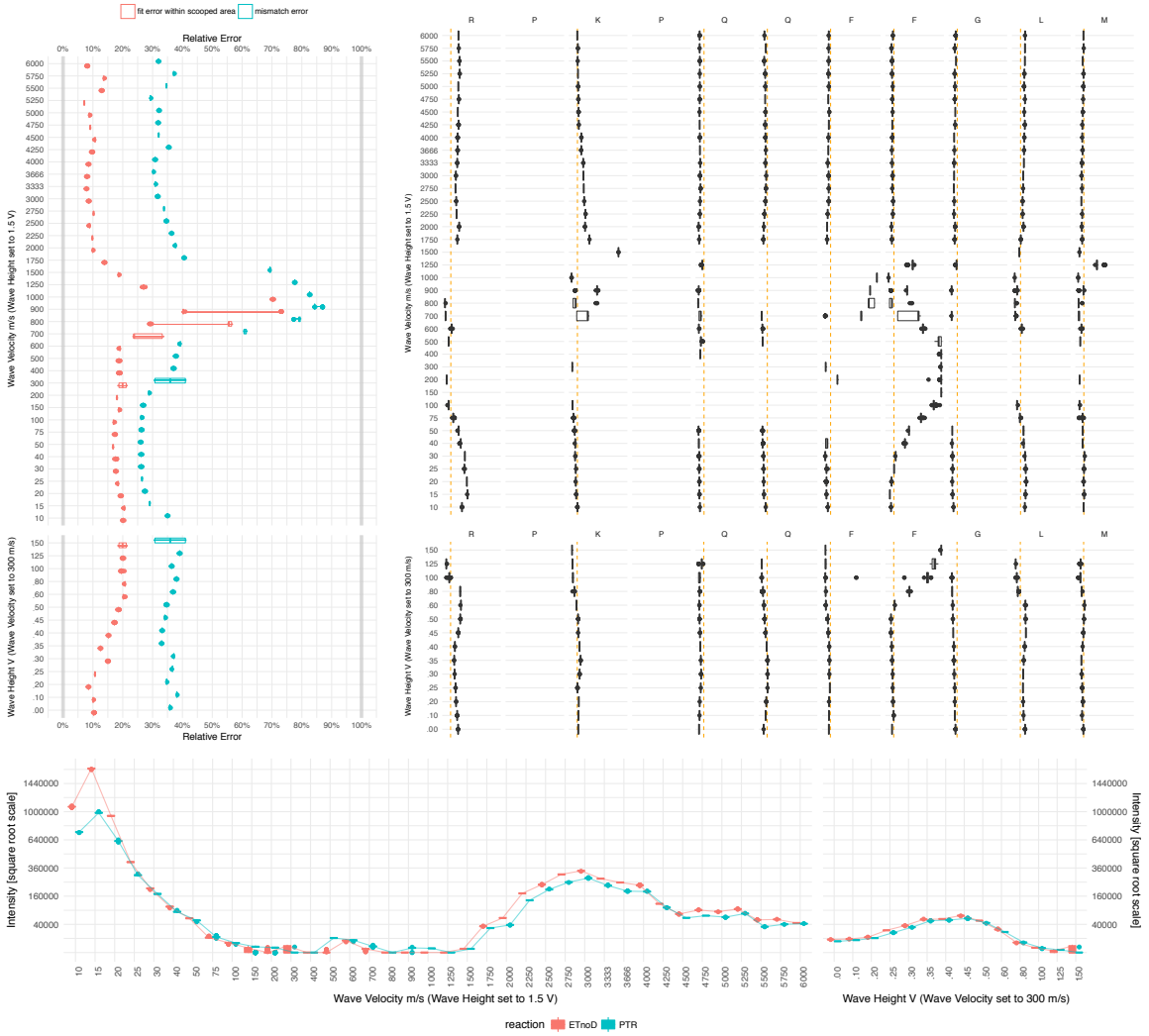
Standard deviation of mass inaccuracy σ.

OUTPUT: A mass spectrum.

```
Draw the placements of charges q along the fasta sequence.
Set experiment time to zero, T = 0.
while T < 1 do
    Increase T by a random time interval sampled from
      the exponential distribution with intensity I ∑_i N_i q_i^2.
    Extract ion M from I with probability prop. to N_i q_i^2.
    Draw R from PTR, ETnoD, and PTR,
      with probabilities p_PTR, p_ETnoD, p_ETD.
    if R = ETD then
        if fragmentation occurred twice then
            Discard ion M.
        else
            Draw the fragmentation spot.
            Add fragments with q > 0 to I.
        end if
    else
        Reduce charge by one.
        Adjust the quenched charge.
        Add M to list I.
    end if
end while
for all M in I do
    Randomly choose the isotopic variant of M.
    Blur its mass with Gaussian noise.
end for
Bin the spectrum
return spectrum
```

(a) Mismatch and Fitting Errors

(b) Probabilities of Fragmentation



(c) Intensities of ETnoD and PTR

Figure 3.9: Selected results of the `MassTodon` as run on Substance P spectra. Data were acquired on the Synapt G2 mass spectrometer. The instrumental settings were obtained for two strips of settings in the two-dimensional space comprising wave height and velocity: one strip was obtained by fixing wave height to 1.5 V, the other by fixing the wave velocity to 300 m/s. Results in (a) and (b) show bootstrap estimates (250 repetitions). Results in (c) contain additional lines linking together the estimates obtained for the actual mass spectra. Figure (a) shows estimates of the mismatch error and the fitting error. Both are calculated using the normalized l_1 distance, $E(p,q) = \frac{\sum_k |p_k - q_k|}{\sum_k p_k + \sum_k q_k}$, where p and q are maps with keys k (different m/z ranges) and values p_k and q_k (i.e. real intensities and their estimates). In case of the mismatch error, we compare in this way the estimated spectrum versus the whole experimental mass spectruversus the whole experimental mass spectrumm, which includes peaks that are not among the studied reaction products. The fit error restricts this comparison to the regions of the mass spectrum that actually could be explained by a theoretical reaction product. Figure (b) shows estimates of the probabilities of fragmentation along the backbone of Substance P, whose amino sequence is RPKPQQFFGLM. Fragmentation on prolines (P) is impossible due to the ring structure of this amino acid. The vertical orange dashed lines correspond to probability equal to $1/9$, which would be attained assuming a fully uniform probability of fragmentation. Figure (c) shows the estimates of the intensity of the ETnoD and PTR reactions. Values of intensities in the y axis have been transformed by a square root scaling in order to expose the behavior of the lower estimates.

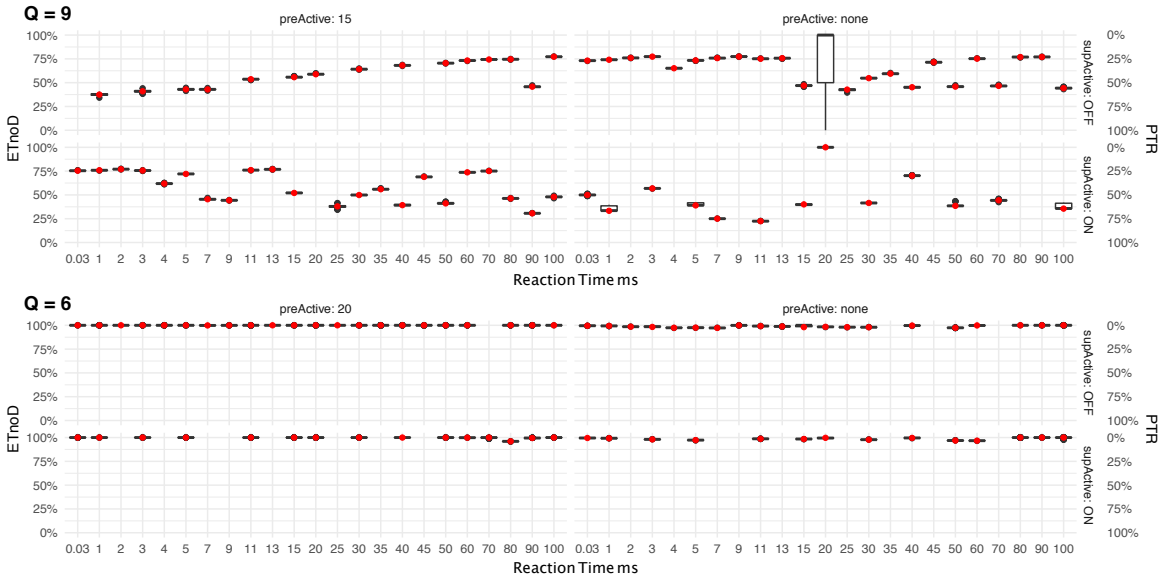


Figure 3.10: Estimates of the probabilities of ETnoD and PTR conditional on one of these events happening obtained for ubiquitin. Data were collected using LTQ Orbitrap Velos. Red dots correspond to estimates performed on real data. The black box plots, mostly extremely narrow, correspond to 250 sample bootstrap estimates. Precursor charge Q is shown in top-left parts of the panels. Each panel is subdivided into subpanels corresponding to different experimental settings. *Nota bene:* left panels correspond to different levels of pre-activations. For $Q = 9$ the energy of preactivation was set to 15, while for $Q = 6$ to 20. The x axis shows the retention time RT, while the y axis shows the percentual content of the ETnoD and PTR reactions. For the spectrum gathered at $Q = 9$ and RT = 20, without pre-activation and without the supplementary activation, there were no ions found that could undergo ETnoD or PTR in the real spectrum under the given threshold on the intensity (results contain the 95% of the highest peak in that spectrum), so the red dot is missing.

4

Estimating Reaction Kinetics of Electron Transfer Reactions

""Excellent! I cried. "Elementary," said he."

— Dr Watson

Mass spectrometry is an analytical technique of measuring the ratio of mass to charge (m/z) of molecular compounds. Ionized molecules are separated in an electromagnetic field. The intensity of the detected signal is plotted against the corresponding m/z values on a mass spectrum. In most of its range, the signal intensity is proportional to the number of the detected particles (Housecroft and Constable, 2010).

Among many of its applications, mass spectrometry can be used for identifying compounds in biological samples. In the case of proteins, however, the mass of the whole molecule provides little information about its amino acidic sequence, and even less so on its tertiary structure. In particular, any permutation of amino acids in the sequence results in the same signal in the spectrum. One can gain much more insight into the structure of sample molecules by inducing their fragmentation and recording the resulting signal. In particular, knowing the masses of all consecutive fragments can reveal the protein's sequence.

There are two main approaches to protein fragmentation: bottom-up and top-down. In

bottom-up proteomics the protein is partially digested by a proteolytic enzyme and mass spectrometry is used to measure the m/z ratios of the fragments. In the top-down approach, sample proteins are subject to fragmentation only inside the mass spectrometer, without the use of any proteases.

One of the fragmentation methods used in top-down mass spectrometry is Electron Transfer Dissociation (ETD). This ion-ion technique exploits the naturally occurring interaction between the multi charged, non-radical protein/peptide cation on one side, and the radical reagent anion on the other (Syka *et al.*, 2004; Zhurov *et al.*, 2013). However, while this method is becoming ever more ubiquitous in the MS-based proteomics analyses, important questions remain regarding the precise reaction mechanism, fragmentation patterns, and the level(s) of protein structure that can be probed using ETD (Sohn *et al.*, 2009, 2015). Shedding more light on the nature of ETD can thus lead to optimization of the instrumental settings and the overall improvement of the identification of peptide sequences and the post-translational modifications.

There are several other fragmentation techniques used in the top-down approach, most importantly the Collision-Induced Dissociation (CID), where the cleavage is induced by colliding ions with nonreactive gas molecules (Mitchell Wells and McLuckey, 2005). A major disadvantage of the CID compared to ETD is that it often leads to loss of posttranslational modifications, particularly phosphorylation (Kim and Pandey, 2012). Electron Transfer Dissociation has also been found to provide more uniform fragmentation than CID, which preferentially cleaves the weakest bonds (Kim and Pandey, 2012; Zhurov *et al.*, 2013). However, a notable amount of work has been devoted to analyzing and mathematically modeling the CID process (Zhang, 2004, 2005; Wysocki *et al.*, 2000), while ETD has received less attention.

The fragmentation in ETD is induced by the transfer of an electron from a radical anion to the sample peptide/protein cation the after a series of electron rearrangements results in a cleavage of one of the peptides ($N-C_\alpha$) bonds. The sample cations are positively charged during the electrospray ionization (ESI) step (Fenn *et al.*, 1989), leading to the formation of $[M+nH]^{n+}$ ions, i.e. adding both charge and mass to the analyte molecule M.

Apart from ETD, other reactions occur concurrently adding their products to the signal observed in the mass spectrometer. Figure 4.1 presents the considered set of reactions. Unlike in ETD, during PTR the proton gets transferred from the protein's backbone to the anion. The mechanism of ETnoD closely resembles that of ETD, with the difference that the protein fails to fragment into the c and z. The appearance of the ETnoD fragments in the experimental data can be traced to the folding of proteins: although backbone cleavage occurs, noncovalent interactions keep the resulting fragments from separating. The ETnoD can also be caused by accommodation of an electron, e.g. in an aromatic side chain (Lermyte *et al.*, 2014; Lermyte and Sobott, 2015). It is assumed that, regardless of the precise reaction

mechanism, the electron obtained by ETnoD causes neutralization of one ESI-generated proton (Lermyte *et al.*, 2015a), referred to as the *quenched proton* further on. In all of the reactions described above, one charge is neutralized.

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	$[M + nH]^{n+} + A^{\bullet -}$	$\longrightarrow$	$[C^{+}xH]^{x+} + [Z^{+}(n\text{-}x)H]^{(n-x-1)+\bullet} + A$

Table 4.1: Chemical reactions considered by the `ETDetective`. M stands for a precursor or a fragment ion, C and Z stand for fragment ions. Observe that compared to the table presented in Chapter 3, we do not consider the HTR reaction.

A single cation can undergo several reaction events, being approached multiple times by different anions. However, the so-called internal fragments of proteins, i.e. resulting from two backbone cleavage events, are usually not observed, suggesting that double ETD scarcely ever occurs. On the other hand, there is a lot of evidence that one analyte molecule can undergo multiple ETnoD and PTR (Lermyte *et al.*, 2015c). Note that only molecules with non-zero charge are observed in the mass spectrometer: after a sufficiently large number of reactions molecules simply disappear.

The isotope distributions of reaction products show considerable overlap, especially for large molecules, as illustrated in Fig. 4.1. In particular, the products of PTR and ETnoD reactions on the same substrate differ only by 1Da mass (the mass of the electron can be neglected, falling beyond the resolving power of most modern instruments).

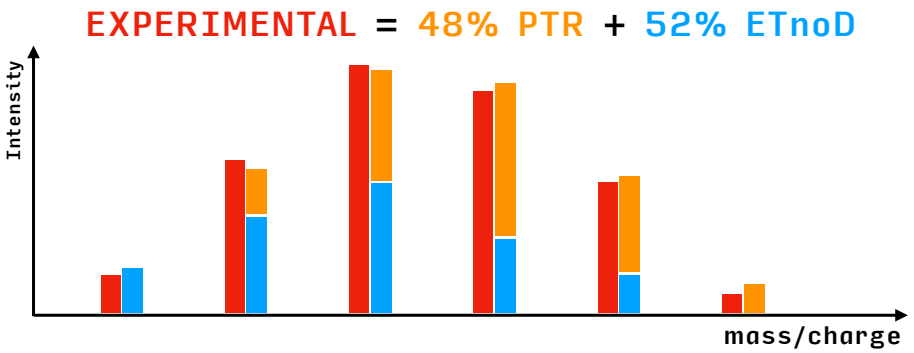


Figure 4.1: The deconvolution of the observed isotopic envelopes performed by MassTodon. The observed signal (in red) is represented as a combination of two theoretical isotopic patterns (orange and blue).

The peptide bond cleavage induced by ETD is believed to be fairly uniform (Li *et al.*, 2011). A notable exception from this rule is the peptide bond of proline: due to the ring structure of this amino acid, the c- and z-ions are held together even after the $N - C_\alpha$ bond cleavage.

A specific type of $N - C_\alpha$ bond cleavage occurs on the N-terminus, leading to a loss of one ammonia molecule. The precise mechanism of this reaction is not yet known. Here, we assume this reaction to be an instance of ETD and treat the ammonia molecule as a c

fragment. Therefore, the number of considered ETD cleavage sites is equal to the number of amino acids other than proline in the protein/peptide sequence.

Our contribution. We propose a formal model of the electron-driven reactions occurring inside the mass spectrometer. We follow a modeling strategy first developed by Gambin and Kluge (2010) to study the degradation of proteins by proteolytic enzymes. The model of ETD reaction can be obtained conceptually in the same way: the stochastic description of the reaction, based on a Markov Jump Process (MJP), is transformed to a populational description of a large number of molecules based on a system of Ordinary Differential Equations (ODEs). Given the intensities of transitions in the process, we solve the ODEs numerically with a recursive algorithm to obtain the expected number of molecules. The space of possible intensities is then searched for the best possible set of parameters by solving an optimization problem.

The model we propose lets us express the mass spectrum in terms of parameters such as the total intensity of reactions and the probabilities of the three studied reactions: ETD, PTR, and ETnoD. A process described by a handful of parameters can be easily visualized and thus easily understood. Also, the comparison of different spectra, e.g. coming from different instrument settings, is highly simplified.

We apply our method to mass spectra gathered in controlled experiments, obtained for highly purified compounds. The identity of the precursor ion and all fragments obtained given a set of possible reactions is known and the quantities of these fragments can be established using our in-house developed identification tool called MassTodon (Lermyte *et al.*, 2015a, 2017; Łącki *et al.*, 2017a). Given a mass spectrum and a precursor molecule, MassTodon outputs a list of reaction products together with their estimated intensities (that are usually assumed to be proportional to the actual number of ions). It performs deisotopisation and deconvolution of the spectrum, i.e. reports total intensities of chemical compounds in possibly overlapping isotope clusters (see Figure 4.1).

The model and the fitting procedure have been implemented in Python. The software tool, called ETDetective, is designed as an extension to MassTodon workflow, see `https://matteolacki.github.io/MassTodonPy/`. The control flow of the whole process from obtaining a spectrum to obtaining the reaction rates and fragmentation patterns has been depicted on Figure 4.2. ETDetective together with example data is available to download at `https://github.com/mciach/ETDetective` under the 2-clause BSD license.

Related research. Various approaches have been taken to model different protein fragmentation techniques (Breuker *et al.*, 2004; Simons, 2010; Zhurov *et al.*, 2013; Tureček and Julian, 2013). A somewhat similar approach to the one taken by us was presented by Zhang (2004, 2005) to study CID fragmentation, who uses a kinetic model to study fragmentation. Zhang (2010) adapts the model to model mass spectra obtained with the use of ETD.

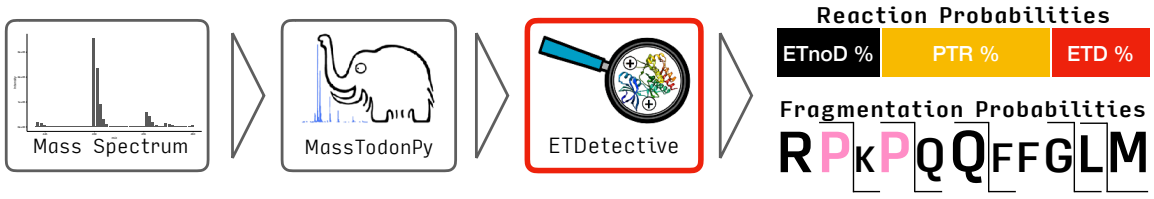


Figure 4.2: The process of mass spectrum interpretation with MassTodon and ETDetective.

The model uses 280 parameters and its derivation is grounded in the theory of statistical mechanics. The model was fitted to a training data set consisting of more than 7000 ETD spectra simultaneously.

There are important differences between that approach and ours. Zhang's model is derived from the first principles of statistical physics, whereas the one we propose is more phenomenological. In our approach, the physics of the phenomenon dictates only the potential states and the transitions between them. We then cast the problem into the well-studied setting of continuous time Markov Jump Processes. Our current approach also builds upon the approach for parameter estimation introduced previously in the MassTodon paper. MassTodon used a heuristical approach to estimate some of the deep parameters of the process, relying on the idea of parsimony. The approach we present here is theory driven. That said, ETDetective can use some of the estimates provided by MassTodon and not optimize them. This can greatly reduce the number of existing parameters, as one can skip the estimation of the fragmentation probabilities. In contrast, parameters described by Zhang are fairly complex, making it more difficult to limit their number. Limiting the number of parameters also reduces the risk of model's unidentifibility. Finally, one can use the results obtained using our model as an input for another model that, similarly to Zhang, includes more of the underlying physical principles. For instance, the reaction rates we provide appear in the Arrhenius equations.

Apart from these mostly theoretical considerations, the ability to fit to individual mass spectra also simplifies the process of comparing results obtained with different instruments. This is an important step in experiment design, see (Lermyte *et al.*, 2015a).

A notable amount of literature has been built up around the idea of purely data-driven prediction of the intensity of peptides in tandem MS experiments (Elias *et al.*, 2004; Arnold *et al.*, 2006; Degroeve *et al.*, 2013). A more exploratory approach targeted at studying fragmentation patterns was taken by Li *et al.* (2011). That said, the above approaches have been applied mainly to study CID.

Organization of the chapter. First, we introduce the theoretical considerations behind our model. Then, we describe the procedures used to obtain our data sets (experimental and *in silico*). Then, we assess the performance of the model. Finally, we discuss existing problems and possible extensions.

Kinetic Model of ETD

Statement of the model

Following the ideas outlined in Gambin and Kluge (2010), we model ETD and its side reactions as a continuous time Markov Jump Process (MJP), which is a well-established approach to modeling chemical reactions. Below, we describe the state space of our model and provide elementary lemmas on its size and properties. Next, we define the transition intensities of our MJP.

Our model can be described by a Petri net, in which places correspond to molecular species, transitions to reactions, and tokens to molecules of a given species (Figure 4.3).

All molecules that cannot be observed, e.g. the internal fragments or ions in which all charges have been neutralized, are merged into the *cemetery*—a unique place without any outgoing transitions. Note, however, that the reactions which yield such molecules are still present in the graph. We will refer to this net as the *reaction graph*.

Definition 1. *A reaction graph is a bipartite, directed, connected graph $\langle \mathcal{M}, \mathcal{R}, \mathcal{F} \rangle$, in which*

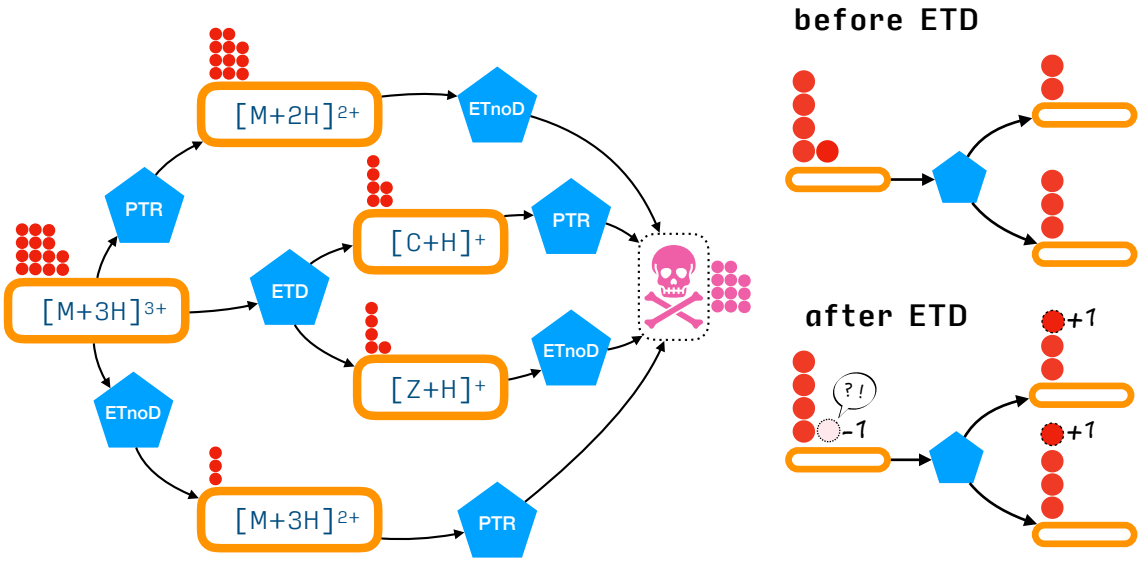
- *$\mathcal{M}$ is a set of vertices called molecular species or places,*
- *$\mathcal{R}$ is a set of vertices called reactions or transitions,*
- *$\mathcal{F} \subset (\mathcal{M} \times \mathcal{R}) \cup (\mathcal{R} \times \mathcal{M})$ is a set of edges connecting species and reactions, and*
- *$W : \mathcal{M} \to \mathbb{N}$ is a function denoting the number of molecules or tokens of a molecular species.*

Each molecular species $u \in \mathcal{M}$ is described by the sequence of amino acids s, the charge of the cation q, and the number of quenched protons g, so that $u = (s, q, g)$. Note that we do not model the positions of the charges, i.e. we assume to know only the numbers of protons on the backbone. We denote the charge of u as q_u. The sequence and number of quenched protons are denoted accordingly as s_u and g_u.

The *precursor* or *root* of the reaction graph, denoted $r = (s, q_0, 0)$, is the unique molecular species with no incoming transitions (i.e. the root of the reaction graph).Based on the description of the set of molecular species, we can approximate the size of this set as follows:

Lemma 1. *The number of the places in a reaction graph corresponding to a precursor molecule $r = (s, q_0, 0)$ is $O(Lq_0^2)$, where L is the length of s.*

Proof. Since in the reaction graph we do not include the internal fragments (i.e. infixes of the amino acid sequence), there are $O(L)$ possible sequences of molecular species. Furthermore, for each molecular species $u = (s_u, q_u, g_u)$, we have $q_u + g_u \leq q$. ■



(a) A Petri Net **(b)** ETD reaction

Figure 4.3: A model of the ETD reaction. (a) A fragment of the reaction graph for a triply charged precursor. The *molecular species* are depicted in orange the *reactions* in blue. The pink skull represents the *cemetery*. The reaction graph serves as a board for *tokens* that represent the numbers of molecules of a given species, depicted as red circles. Only one ETD transition has been shown for clarity of the image. Tokens reach cemetary when they lose all their charge and are depicted in pink (for eternity). (b) During each reaction, a token disappears on the substrate side and product tokens appear: one in the case of ETnoD and PTR, two in the case of ETD.

For two molecular species u and v, we write $u \to v$ if v can be reached from u by a single reaction. We write $u \geq v$ if there exist molecular species $m_1, m_2, \ldots, m_n$ such that $u = m_1 \to m_2 \to \cdots \to m_n = v$. Note that $u \geq u$. We also write $u > v$ if $u \geq v$ and $u \neq v$. In this case, u is referred to as the *ancestor* or *ancestral molecule* of v.

For a reaction $R \in \mathcal{R}$, all molecules u such that $(u, R) \in \mathcal{F}$ are called *substrates* of R. Similarly, all molecules v such that $(R, v) \in \mathcal{F}$ are called *products* of R. If u is the substrate of reaction $R \in \mathcal{R}$ and $v_1, v_2, \ldots, v_m$ are its products, then we denote R as $u \to v_1 + v_2 + \cdots + v_m$. Species v_i are referred to as the *daughter* species of u_i's, and u_i's are called *parent* species of v_i's.

Note that in our model, any reaction can be uniquely identified by its substrate and one of the products. Therefore, we will write $u \to v_1$ or $u \to v_2$ to denote a reaction $u \to v_1 + v_2$. We will also write $u \to v$ to indicate the existence of a reaction for which u is a substrate and v is a product.

We assume that at the onset, before any reaction occurred, positive charges are attached randomly to basic amino acids of the molecules, i.e. on lysines, arginines, and histidines, at most one charge per site. This restricts the number of protons on a molecular species: for any molecule, m, $q_m + g_m \leq B_m$ must hold, where B_m is the number of basic amino acids in its sequence.

If one does not know the position of charges before ETD than one cannot know how many protons should appear on the fragment ions. Therefore, a single fragmentation reac-

tion at a given residue gives rise to several different outcomes. This leads to the following lemma. We have the following lemma.

Lemma 2. *Assume a random placement of charges and quenched protons on basic amino acids of a molecule $m = (s, q, g)$. Let c_l be the l-th prefix of the sequence, and let z_{L-l} be the l-th suffix. Let B_c be the number of basic amino acids in the backbone of c_l, and B_z be the number of basic amino acids on the backbone of the corresponding z_{L-l} fragment. Then, the probability of observing q_c charges and g_c quenched protons on c_l after ETD cleavage on l-th amino acid is equal to*

$$P_l(q_c, g_c) = \frac{\binom{B_c}{q_c}\binom{B_z}{q-1-q_c}}{\binom{B_c+B_z}{q-1}} \frac{\binom{B_c-q_c}{g_c}\binom{B_z-q+q_c+1}{g-g_c}}{\binom{B_c+B_z-q+1}{g}},$$

and also equal to the probability of observing $q_z = q-1-q_c$ charges and $g_z = g-g_c$ quenched protons on z_{L-l}.

Proof. Since one charge gets neutralized during the reaction, both fragments have $q - 1$ charges and g quenched protons in total. As each charge is placed randomly and independently of other charges on the unoccupied basic sites, the probability of observing q_c charges on c_l is equal to the probability of choosing q_c out of B_c basic amino acids and $q - 1 - q_c$ out of B_z basic amino acids randomly and without replacement. After placing the charges on the sequence, there are B_c+B_z-q+1 unoccupied basic sites. The probability of observing g_c quenched protons on c_l, given q_c charges, is then equal to the probability of choosing g_c out of $B_c - q_c$ basic amino acids and $g - g_c$ out of $B_z - (q - 1 - q_c)$ basic amino acids. ■

The outcomes of the PTR and ETnoD reactions are unique. It follows that the number of outgoing transitions for a molecular species other than the cemetery is equal to the number of ETD transitions plus two side reactions:

$$2 + \sum_{l=1}^{L} \binom{B_{c_l} + B_{z_{L-l}}}{q-1} \binom{B_{c_l} + B_{z_{L-l}} - q + 1}{g}.$$

However, many transitions lead directly to the cemetery. This is especially the case for any molecule with a single charge or any ETD reaction of a molecular species which has already undergone an ETD.

The *rate* of a reaction $R = u \to v$ is denoted λ_{uv}. We assume that this rate can be factorized into a product of base reaction intensity, I, squared charge of the substrate, q_u, and reaction probability P_R, so that

$$\lambda_{uv} = I q_u^2 P_R \text{ for } R = u \to v,$$

where

$$P_R = \begin{cases} P_{PTR} & \text{for} \quad R = (s,q,g) \to (s,q-1,g), \\ P_{ETnoD} & \text{for} \quad R = (s,q,g) \to (s,q-1,g+1), \\ P_{ETD_l} P_l(q_c,g_c) & \text{for} \quad R = (s,q,g) \to (c_l,q_c,g_c) + (z_{L-l},q_z,g_z) \\ & \qquad \text{for } q_z = q-1-q_c,\ g_z = g-g_c. \end{cases}$$

In the above definition, P_{ETD_l} is the probability of ETD reaction on the l-th amino acid, regardless of the distribution of charge among product fragments. Note that the rates $u \to c_l$ and $u \to z_{L-l}$ are equal, as they correspond to the same reaction. The assumption that the microscopic intensity of a given reaction is proportional to squared substrate charge is motivated by the kinetics of ion reactions (McLuckey and Stephenson, 1999).

We further define the *outflow rate*, λ_{uu}, as $\lambda_{uu} = -\sum_{v:u\to v} \lambda_{uv}$. Since the probabilities of reactions sum to 1, λ_{uu} can be expressed by a simple closed formula:

$$\lambda_{uu} = -I q_u^2.$$

We then construct a Markov Jump Process (MJP) to describe the flow of molecules across the reaction graph. Denote the number of tokens at place m in time t by $X_m(t)$. The state of the MJP, denoted as $X(t)$, is defined as a collection of all token counts at a given moment in time, so that $X(t) = (X_m(t))_{m\in\mathcal{M}}$. We assume that at time 0, only the precursor molecules are observed. Throughout this work, we assume the state $X(0)$ to be fixed. It follows that the state space of the process, say E, is a finite subset of $\mathbb{N}^{\mathcal{M}} = \{x = (x_m)_{m\in\mathcal{M}} : \forall_{m\in\mathcal{M}} x_m \in \mathbb{N}\}$.

From a given state $x \in \mathbb{N}^{\mathcal{M}}$, the system can evolve to another state following one of the reactions in Figure 4.3. We denote the change in token numbers induced by the transition $R \in \mathcal{R}$ as a vector $\delta^R = (\delta^R_m)_{m\in\mathcal{M}}$, so that

$$\delta^R_m = \begin{cases} -1 & \text{if} \quad (m,R) \in \mathcal{F} \\ 1 & \text{if} \quad (R,m) \in \mathcal{F} \\ 0 & \text{otherwise.} \end{cases}$$

We assume that the anion radicals do not deplete in time, and the spatial interactions are negligible, so that each molecule (i.e. each token) reacts independently of the other ones. This shows that process $X(t)$ is in fact a sum of independent, time-uniform Markov processes describing individual molecules. Consider two neighbouring states, x and $y = x+\delta_R$. Let u be the substrate molecular species of R and v be one of it's products. With the aforementioned assumptions, the intensity of transition from x to y is the sum of reaction rates λ_{uv} of molecules on u. The *transition intensity* Q_{xy} for $x \neq y$ then equals

$$Q_{xy} = \begin{cases} x_u \lambda_{uv} & \text{if } y = x + \delta^{u\to v}, \\ 0 & \text{otherwise.} \end{cases}$$

Such form of Q_{xy} results from an assumption that each molecule (i.e. each token) reacts independently of the other molecules with rate λ_{uv}. We also define the *outflow intensity*, Q_{xx}, as $Q_{xx} = -\sum_{y \in \mathbb{N}^{\mathcal{M}}} Q_{xy}$. Similarly to λ_{uu}, Q_{xx} can be expressed in a simple form:

$$Q_{xx}(t) = \sum_{u \in \mathcal{M}} x_u \lambda_{uu} = -\sum_{u \in \mathcal{M}} x_u I q_u^2.$$

The above equations fully describe our model. The model has $L + 3$ parameters: L probabilities of ETD (including cleavage of the N-terminal amino group), 2 probabilities of side reactions, and the base intensity.

Analytical results

We now describe theoretical results concerning the dynamics of the substrates and products of some of the molecular species. In particular, we provide a full description of the initial precursor's dynamics, the description of the dynamics of the expected evolution of all molecular species and results on the dynamics of some of the second moments. Finally, we show when one should expect the reaction to get totally depleted. The above results are vital for narrowing down the space of parameters for the fitting procedure.

The following lemma will be used in proofs:

Lemma 3. *If $u > v$, then $\lambda_{uu} < \lambda_{vv}$.*

Proof. Since $u > v$, there exists a set of transitions by which v can be obtained from u. As each transition leads to a loss of at least one charge (exactly one in case of PTR and ETnoD), we have $q_u > q_v$; Since by definition $I > 0$, it follows that $-Iq_u^2 < -Iq_v^2$. ■

The following theorem fully describes the dynamics of the initial precursor.

Theorem 1. *Let $X_r(t)$ be the number of precursor molecules, $r = (s, q_0, 0)$, at time t, and let $N = X_r(0)$. Then, $X_r(t)$ has a binomial distribution with N trials and probability of success equal* $\exp(-Iq_o^2 t)$*:*

$$\mathbb{P}(X_r(t) = n) = \binom{N}{n} \exp(-nIq_0^2 t)(1 - \exp(-Iq_0^2 t))^{N-n}.$$

Corollary 1. *Let $X_r(t)$ be the number of precursor molecules $r = (s, q_0, 0)$ at time t, and let $N = X_r(0)$. Then,*

$$\mathbb{E}X_r(t) = N \exp(-Iq_0^2 t),$$
$$\mathrm{Var}X_r(t) = N \exp(-Iq_0^2 t) - N \exp(-2Iq_0^2 t).$$

Proof. Consider a single token of the precursor molecular species. Let τ be the first time of any reaction of such token. By construction of the process, τ has an exponential distribution with parameter Iq_0^2. It follows that

$$\mathbb{P}(\tau < t) = 1 - \exp(\lambda_{rr} t) = 1 - \exp(-Iq_0^2 t).$$

The probability that the considered token is on the precursor molecular species at time t is equal to the probability that the first reaction occured after time t. Since the tokens react independently, the total number of precursor molecules realizes a binomial scheme with N trials and the probability of success equal to $\exp(-Iq_0^2 t)$. ■

In general, due to the complicated structure of the reaction graph and the fact that the ETD reactions have more than one product, it is difficult to obtain distributions of all molecular species. However, we can obtain a relatively simple system of ordinary differential equations for the expected number and variance of molecules, and solve them recursively by a numerical procedure:

Theorem 2. *Let $u, v \in \mathcal{M}$ be two neighbouring molecular species (i.e. $u \to v$ or $v \to u$). Let $\mathbb{E}X_u(t)$ and $\mathrm{Var}X_u(t)$ denote the expected number and variance of the number of u molecules, and let $\mathrm{Cov}(X_u(t), X_v(t))$ denote the covariance between the numbers of u and v molecules. Then, we have*

$$\frac{\partial}{\partial t}\mathbb{E}X_u(t) = \sum_{w:\, w\to u} \lambda_{wu}\mathbb{E}X_w(t) + \lambda_{uu}\mathbb{E}X_u(t) \tag{4.1}$$

$$\begin{aligned}\frac{\partial}{\partial t}\mathrm{Var}X_u(t) = &\sum_{w:\, w\to u} 2\lambda_{wu}\mathrm{Cov}(X_u(t), X_w(t)) + 2\lambda_{uu}\mathrm{Var}X_u(t) \\ &+ \sum_{w:\, w\to u} \lambda_{wu}\mathbb{E}X_w(t) - \lambda_{uu}\mathbb{E}X_u(t).\end{aligned} \tag{4.2}$$

$$\begin{aligned}\frac{\partial}{\partial t}\mathrm{Cov}(X_u(t), X_v(t)) = &\sum_{w:\, w\to u} \lambda_{wu}\mathrm{Cov}(X_w(t), X_u(t)) \\ &+ \sum_{w:\, w\to v} \lambda_{wv}\mathrm{Cov}(X_w(t), X_v(t)) \\ &+ (\lambda_{uu} + \lambda_{vv})\mathrm{Cov}(X_u(t), X_v(t)) \\ &- \lambda_{uv}\mathbb{E}X_u - \lambda_{vu}\mathbb{E}X_u.\end{aligned} \tag{4.3}$$

Proof. Let $[t, t+h]$ be a time interval short enough that only one reaction can occur. In such interval, the number of u molecules can either increase by 1, decrease by 1, or stay unchanged. Consider the expected number of u molecules at time $t+h$ conditioned on the state of the process at time t. From the definition of the expected value and construction of

the reaction graph, we have

$$\begin{aligned}\mathbb{E}X_u(t+h)|X(t) = (X_u(t)+1)\mathbb{P}(X_u(t+h) = X_u(t)+1|X(t)) \\ + (X_u(t)-1)\mathbb{P}(X_u(t+h) = X_u(t)-1|X(t)) \\ + X_u(t)\mathbb{P}(X_u(t+h) = X_u(t)|X(t)).\end{aligned}$$

Consider $X(t) = x$. From the definition of transition intensity, we have

$$\mathbb{P}(X_u(t+h) = X_u(t)+1|X(t) = x) = \sum_{y: y_u = x_u+1} (Q_{xy}h + o(h)) = \sum_{w:\ w\to u} (x_w\lambda_{wu}h + o(h)).$$

Since the state space is finite, we have $\sum_{w:\ w\to u}(x_w\lambda_{wu}h+o(h)) = \sum_{w:\ w\to u}(x_w\lambda_{wu}h)+ o(h)$. By similar reasoning for the other terms, we get

$$\begin{aligned}\mathbb{E}X_u(t+h)|X(t) = (X_u(t)+1)\sum_{w:\ w\to u} X_w(t)\lambda_{wu}h + (X_u(t)-1)\sum_{w:\ u\to w} X_u(t)\lambda_{uw}h \\ + X_u(t)\left(1 - \sum_{w:\ w\to u} X_w(t)\lambda_{wu}h - \sum_{w:\ u\to w} X_u(t)\lambda_{uw}h\right) + o(h).\end{aligned}$$

After basic algebraic manipulations, we get

$$\mathbb{E}X_u(t+h)|X(t) = \sum_{w:\ w\to u} X_w(t)\lambda_{wu}h - \sum_{w:\ u\to w} X_u(t)\lambda_{uw}h + X_u(t) + o(h).$$

By taking expectation with respect to $X(t)$, we obtain

$$\mathbb{E}X_u(t+h) = \sum_{w:\ w\to u} \mathbb{E}X_w(t)\lambda_{wu}h - \sum_{w:\ u\to w} \mathbb{E}X_u(t)\lambda_{uw}h + \mathbb{E}X_u(t) + o(h).$$

Now, after subtracting $\mathbb{E}X_u(t)$ from both sides, dividing by h and taking a limit $h\to 0$, we arrive at

$$\frac{\partial}{\partial t}\mathbb{E}X_u(t) = \sum_{w:\ w\to u} \mathbb{E}X_w(t)\lambda_{wu} - \sum_{w:\ u\to w} \mathbb{E}X_u(t)\lambda_{uw} = \sum_{w:\ w\to u} \mathbb{E}X_w(t)\lambda_{wu} + \lambda_{uu}\mathbb{E}X_u(t),$$

which proves Equation (4.1).

Now, consider the second moment of the number of molecules of species u, $\mathbb{E}X_u^2(t)$. We have

$$\begin{aligned}\mathbb{E}X_u^2(t+h)|X(t) = X_u^2(t)\mathbb{P}(X_u(t+h) = X_u(t)|X(t)) \\ + (X_u(t)+1)^2\mathbb{P}(X_u(t+h) = X_u(t)+1|X(t)) \\ + (X_u(t)-1)^2\mathbb{P}(X_u(t+h) = X_u(t)-1|X(t)).\end{aligned}$$

Substituting for the probabilities, we get

$$\mathbb{E}X_u^2(t+h)|X(t) = X_u^2(t)\left(1 - \sum_{w:\, w\to u}\lambda_{wu}X_w(t)h - \sum_{w:\, u\to w}\lambda_{uw}X_u(t)h\right) + (X_u(t)+1)^2\sum_{w:\, w\to u}\lambda_{wu}X_w(t)h + (X_u(t)-1)^2\sum_{w:\, u\to w}\lambda_{uw}X_u(t)h + o(h).$$

After grouping terms and averaging over $X(t)$, we get

$$\mathbb{E}X_u^2(t+h) = \sum_{w:\, w\to u}2\lambda_{wu}\mathbb{E}X_u(t)X_w(t)h + \sum_{w:\, w\to u}\lambda_{wu}\mathbb{E}X_w(t)h + \sum_{w:\, u\to w}\lambda_{uw}\mathbb{E}X_u(t)h - \sum_{w:\, u\to w}2\lambda_{uw}\mathbb{E}X_u^2(t)h + \mathbb{E}X_u^2(t),$$

which, after performing simple algebraic manipulations and taking a limit $h \to 0$, yields

$$\frac{\partial}{\partial t}\mathbb{E}X_u^2 = \sum_{w:\, w\to u}2\lambda_{wu}\mathbb{E}X_u(t)X_w(t) + \sum_{w:\, w\to u}\lambda_{wu}\mathbb{E}X_w(t) + \sum_{w:\, u\to w}\lambda_{uw}\mathbb{E}X_u(t) - \sum_{w:\, u\to w}2\lambda_{uw}\mathbb{E}X_u^2(t).$$

Now, from the fact that $\mathrm{Var}X_u(t) = \mathbb{E}X_u^2(t) - \mathbb{E}^2X_u(t)$, we have

$$\frac{\partial}{\partial t}\mathrm{Var}X_u(t) = \frac{\partial}{\partial t}\mathbb{E}X_u^2(t) - 2\mathbb{E}X_u(t)\frac{\partial}{\partial t}\mathbb{E}X_u(t).$$

Substituting for the time derivative of the expected value, we get Equation (4.2).

Now, assume that $u \to v$, and consider the mixed moment, $\mathbb{E}(X_u(t)X_v(t))$. In the time interval $[t, t+h]$, we have the following possibilities:

- The number of u molecules increases,
- The number of v molecules increases due to reaction other than $u \to v$,
- The number of u molecules decreases due to reaction other than $u \to v$,
- The number of v molecules decreases,
- The number of u molecules decreases by 1, and the number of v molecules increases by 1, due to reaction $u \to v$,
- Their numbers stay unchanged.

$$\begin{aligned}\mathbb{E}(X_u(t+h)X_v(t+h)|X(t)) = (X_u(t)+1)X_v(t) &\sum_{w\colon w\to u}\lambda_{wu}X_w h\\ + X_u(t)(X_v(t)+1) &\sum_{w\colon w\to u, w\neq u}\lambda_{wu}X_w h\\ + (X_u(t)-1)X_v(t) &\sum_{w\colon u\to w, w\neq v}\lambda_{uw}X_u h\\ + X_u(t)(X_v(t)-1) &\sum_{w\colon v\to w}\lambda_{wu}X_v h\\ + (X_u(t)-1)(X_v(t)+1)&\lambda_{uv}X_u h\\ + X_u(t)X_v(t)(1-c) &+ o(h),\end{aligned}$$

where $c = 1 - \mathbb{P}(X_u(t+h) = X_u(t), X_v(t+h) = X_v(t)|X(t))$, equal to

$$\begin{aligned}c = &\sum_{w\colon w\to u}\lambda_{wu}X_w h + \sum_{w\colon w\to u, w\neq u}\lambda_{wu}X_w h + \sum_{w\colon u\to w, w\neq v}\lambda_{uw}X_u h\\ &+ \sum_{w\colon v\to w}\lambda_{wu}X_v h + \lambda_{uv}X_u h.\end{aligned}$$

By proceeding as before and using the identity $\mathrm{Cov}(X_u(t)X_v(t)) = \mathbb{E}X_u(t)X_v(t) - \mathbb{E}X_u(t)\mathbb{E}X_v(t)$, we obtain

$$\begin{aligned}\frac{\partial}{\partial t}\mathrm{Cov}(X_u(t), X_v(t)) = &\sum_{w\colon w\to v}\lambda_{wu}\mathrm{Cov}(X_w(t), X_u(t))\\ &+ \sum_{w\colon w\to u}\lambda_{wv}\mathrm{Cov}(X_w(t), X_v(t))\\ &+ (\lambda_{uu}+\lambda_{vv})\mathrm{Cov}(X_u(t), X_v(t))\\ &- \lambda_{uv}\mathbb{E}X_u.\end{aligned}$$

Finally, note that for any two molecular species u and v, if $\lambda_{uv} \neq 0$, then $\lambda_{vu} = 0$. Therefore, we may freely add the term $-\lambda_{vu}\mathbb{E}X_v$ to make the formula symmetric with respect to X_u and X_v, and obtain Equation (4.3). ■

Since we have defined λ_{uv} to be zero when $u \not\to v$, Equation (4.3) can be also used for most other molecular species. One important caveat is the case when both u and v are products of the same ETD reaction, in which case their numbers can increase simultaneously and the formula requires an additional term to account for that possibility.

Theorem 2 allows us to obtain the analytical equations for mean number and variance of the numbers of molecules of species connected to the precursor by a single reaction.

Lemma 4. *Let $r = (s, q_0, 0)$ be the precursor molecular species, and let $N = X_r(0)$. Let u be a daughter molecular species of r after reaction R (either PTR, ETnoD or an ETD at a given residue with a given distribution of charges and quenched protons among fragments). Then,*

$$\mathbb{E}X_u(t) = NP_R\frac{q_0^2}{q_0^2 - q_u^2}(\exp(-Iq_u^2 t) - \exp(-Iq_0^2 t)) \tag{4.4}$$

$$\mathrm{Var}X_u(t) = \mathbb{E}X_u(t) - (\mathbb{E}X_u(t))^2/N = N\frac{\mathbb{E}X_u(t)}{N}\left(1 - \frac{\mathbb{E}X_u(t)}{N}\right) \tag{4.5}$$

Proof. Since u is a daughter species of r, it has only one incoming reaction, $r \to u$. From Theorem 2, we get a differential equation for the mean value:

$$\frac{\partial}{\partial t}\mathbb{E}X_u(t) = \lambda_{ru}\mathbb{E}X_r(t) + \lambda_{uu}\mathbb{E}X_u(t).$$

The solution to this equation with boundary condition $\mathbb{E}X_u(0) = 0$ is

$$\mathbb{E}X_u(t) = N\frac{\lambda_{ru}}{\lambda_{rr} - \lambda_{uu}}(\exp(\lambda_{rr}) - \exp(\lambda uu)),$$

which, after substituting for $\lambda_{rr}, \lambda_{uu}$ and λ_{ru}, gives Equation (4.4).

The equation for covariance between X_r and X_u from Theorem 2 is

$$\begin{aligned}\frac{\partial}{\partial t}\mathrm{Cov}(X_r(t), X_u(t)) = \lambda_{ru}\mathrm{Cov}(X_r(t), X_r(t)) - \lambda_{ru}\mathbb{E}X_r(t) \\ + (\lambda_{rr} + \lambda_{uu})\mathrm{Cov}(X_r(t), X_u(t)).\end{aligned}$$

By the identity $\mathrm{Cov}(X_r(t), X_r(t)) = \mathrm{Var}X_r(t)$, we can use Corollary 1 to substitute for $\mathrm{Cov}(X_r(t), X_r(t))$ and $\mathbb{E}X_r(t)$. The differential equation for covariance can now be solved to get

$$\mathrm{Cov}(X_r(t), X_u(t)) = N\frac{\lambda_{ru}}{\lambda_{rr} - \lambda_{uu}}\exp(\lambda_{rr}t)(\exp(\lambda_{uu}t) - \exp(\lambda_{rr}t)).$$

From Theorem 2, the equation for variance of $X_r(t)$ is

$$\frac{\partial}{\partial t}\mathrm{Var}X_u(t) = 2\lambda_{ru}\mathrm{Cov}(X_r(t), X_u(t)) + 2\lambda_{uu}\mathrm{Var}X_u(t) + \lambda_{ru}\mathbb{E}X_r(t) - \lambda_{uu}\mathbb{E}X_u(t).$$

After substituting and solving the above equation, we arrive at

$$\begin{aligned}\mathrm{Var}X_u(t) = &-N\frac{\lambda_{ru}^2}{(\lambda_{rr} - \lambda_{uu})^2}\exp(2\lambda_{uu}t) + 2N\frac{\lambda_{ru}^2}{(\lambda_{rr} - \lambda_{uu})^2}\exp((\lambda_{rr} + \lambda_{uu})t) \\ &- N\frac{\lambda_{ru}^2}{(\lambda_{rr} - \lambda_{uu})^2}\exp(2\lambda_{rr}t) + N\frac{\lambda_{ru}}{\lambda_{rr} - 2\lambda_{uu}}\exp(\lambda_{rr}t) \\ &- N\frac{\lambda_{ru}\lambda_{uu}}{(\lambda_{rr} - \lambda_{uu})(\lambda_{rr} - 2\lambda_{uu})}\exp(\lambda_{rr}) - N\frac{\lambda_{ru}}{\lambda_{rr} - \lambda_{uu}}\exp(\lambda_{uu}t),\end{aligned}$$

which, after grouping terms, simplifies to

$$N\frac{\lambda_{ru}}{\lambda_{rr} - \lambda_{uu}}(\exp(\lambda_{rr}t) - \exp(\lambda_{uu}t)) - N\frac{\lambda_{ru}^2}{(\lambda_{rr} - \lambda_{uu})^2}(\exp(\lambda_{uu}t - \lambda_{rr}t)^2,$$

equal to $\mathbb{E}X_u(t) - (\mathbb{E}X_u(t))^2/N$. ∎

We end this section with an interesting result on the boundaries of reasonable reaction times. The result is also useful to specify boundaries in which to search for the base intensity when fitting the model to data.

Proposition 1. *Let T_{END} be the expected reaction time in which all molecules lose all their charges (i.e. become unobservable). Then,*

$$\frac{q_0}{I} \geq T_{END} \geq \frac{1}{I}\sum_{i=1}^{q_0}\frac{1}{i^2}.$$

Proof. Consider a single precursor molecule. Since each reaction leads to a neutralization of one charge, there are exactly q_0 reactions needed to fully neutralize all of it's charges. Let τ_1 be the first reaction time and let τ_i be the time between $i-1$'th and i'th reaction. We have $T_{END} = \tau_1 + \tau_2 + \cdots + \tau_{q_0}$.

From the construction of the process, τ_1 follows an exponential distribution with parameter $-\lambda_{rr} = Iq_0^2$. Therefore,

$$\mathbb{E}\tau_1 = (Iq_0^2)^{-1}.$$

If $q_0 = 1$, then the above equation proves the proposition. Assume that $q_0 > 1$. We now have two scenarios:

- The first reaction was either a PTR or ETnoD. Then, τ_2 follows an exponential distribution with parameter $I(q_0-1)^2$, and it's expected value is $(I(q_0-1)^2)^{-1}$.
- The first reaction was an ETD. Then, since both fragments now react independently, τ_2 follows an exponential distribution with parameter $I(q_c^2+q_z^2)$, where q_c and q_z are the fragment charges, and it's expected value is $(I(q_c^2+q_z^2))^{-1}$

Now, since $q_c^2 + q_z^2 \leq (q_c+q_z)^2 = (q_0-1)^2$, in both scenarios we have

$$\mathbb{E}\tau_2 \geq (I(q_0-1)^2)^{-1}.$$

Note also that since $q_0 - 1 > 0$, we have $\mathbb{E}\tau_i \leq I^{-1}$ for $i = 1, 2$. Iterating the above reasoning, we get that

$$\frac{q_0}{I} \geq \sum_{i=1}^{q_0}\mathbb{E}\tau_i \geq \frac{1}{I}\sum_{i=0}^{q_0-1}\frac{1}{(q_0-i)^2},$$

which, after changing the summation index, proves the result. ■

Fitting the model to data

Here, we describe how to fit our model to the observed data. The input for ETDetective consists of a mass spectrum parsed by the MassTodon software. Given a mass spectrum and the precursor's sequence and charge, MassTodon outputs a list of intensities of observed molecular species $(O_u)_{u\in\mathcal{M}}$. We normalize this list so that the intensities sum to 1 and look for a set of model parameters that will best predict the observed molecule proportions. The homogeneity of the considered MJP implies that reaction time and base reaction intensity

are exchangeable, and therefore only one of them can be identified. We thus set the time of reaction to be equal to 1.

For the purposes of numerical stability, we reparametrize our model by the following transformation of the original parameters:

$$\theta = \Big(\log(IP_{PTR}), \log(IP_{ETnoD}), \log(IP_{ETD_1}), \log(IP_{ETD_2}), \ldots, \log(IP_{ETD_L})\Big),$$

where L is the length of the precursor's sequence, and P_{ETD_l} is the probability of cleavage between $l-1$-th and l-th amino acid, including dissociation of the N-terminal amino group as P_{ETD_1}. The new parameters are therefore in $\mathbb{R}^{L+2}$.

The general scheme of fitting the model is as follows: for a given starting point θ_0 (obtained using the estimates from MassTodon), we calculate the expected number of all molecular species in the reaction graph, normalize it, and compare to the observed molecule proportions. Next, we iteratively update θ to minimize the discrepancy between the prediction and observation and obtain the optimal vector of parameters $\hat{\theta}$.

The loss function is the sum of squared differences between predicted and observed proportions, with an optional penalty term for decharged molecules which are not observed in the spectrum,

$$\sum_{u \in \mathcal{M} \setminus \{c\}} \Big[\mathbb{E}X_u(1) - O_u\Big]^2 + \rho\Big[\mathbb{E}X_c(1)\Big]^2,$$

where c is the cemetery. In our numerical experiments we analyze the cases of $\rho = 0$ and $\rho = 1$. To minimize the loss function, we use the L-BFGS-B algorithm with gradient approximation (Nocedal, 1980).

Obtaining analytical formulas for expected numbers of molecules is complicated because of the complex structure of the reaction graph. However, we can state the general form of a solution, and use it in numerical procedures.

The general form of solutions for Equation (4.1) is

$$\mathbb{E}X_u(t) = \sum_{i=1}^{n_u} A_i^u \exp(B_i^u t), \tag{4.6}$$

where A_i^u and B_i^u are coefficients constant in time, but dependent on the reaction rates. Their overall number, n_u, depends on the position of u in the reaction graph. From Corollary 1 it follows that the coefficients for the precursor molecular species are $n_u = 1$, $A_1^r = X_r(0)$ and $B = -Iq_0^2$.

Proof. Proceed by induction. For the root molecule, the Equation (4.6) follows from Corollary 1. Now, consider a non-precursor molecular species u, and assume that the Equation (4.6) is true for all molecular species v such that $v > u$. From Theorem 2, we have

$$\frac{\partial}{\partial t}\mathbb{E}X_u(t) = \sum_{w:\ w \to u} \lambda_{wu}\mathbb{E}X_w(t) + \lambda_{uu}\mathbb{E}X_u(t).$$

Since in the above equation we have $w > u$, we can use the induction hypothesis to obtain

$$\frac{\partial}{\partial t}\mathbb{E}X_u(t) = \sum_{w\colon\, w\to u}\sum_{i=1}^{n_w}\lambda_{wu}A_i^w\exp(B_i^w t) + \lambda_{uu}\mathbb{E}X_u(t). \tag{4.7}$$

Note that it follows that $B_i^w = \lambda_{vv}$ for some $w \geq v$. The corresponding homogeneous equation is $\frac{\partial}{\partial t}\mathbb{E}X_u(t) = \lambda_{uu}\mathbb{E}X_u(t)$, which implies that the solution to Equation 4.7 is

$$\mathbb{E}X_u(t) = c(t)\exp(\lambda_{uu}t).$$

By differentiating and substituting again into (4.7), we get

$$\frac{\partial c}{\partial t}(t) = \sum_{w\colon\, w\to u}\sum_{i=1}^{n_w}A_i^w\lambda_{wu}\exp((B_i^w - \lambda_{uu})t).$$

Since $w > u$ and $B_i^w = \lambda_{vv}$ for some $v \geq w$, we have $B_i^w \neq \lambda_{uu}$ (Lemma 3). It follows that, for some constant c, we have

$$c(t) = c + \sum_{w\colon\, w\to u}\sum_{i=1}^{n_w}A_i^w\frac{\lambda_{wu}}{B_i^w - \lambda_{uu}}\exp(B_i^w - \lambda_{uu}),$$

$$\mathbb{E}X_u(t) = c\exp(\lambda_{uu}t) + \sum_{w\colon\, w\to u}\sum_{i=1}^{n_w}A_i^w\frac{\lambda_{wu}}{B_i^w - \lambda_{uu}}\exp(B_i^w t).$$

Since u is not the precursor molecule, we have $\mathbb{E}X_u(0) = X_u(0) = 0$, which implies that

$$c = -\sum_{w\colon\, w\to u}\sum_{i=1}^{n_w}A_i^w\frac{\lambda_{wu}}{B_i^w - \lambda_{uu}},$$

and therefore

$$\mathbb{E}X_u(t) = \sum_{w\colon\, w\to u}\sum_{i=1}^{n_w}A_i^w\frac{\lambda_{wu}}{B_i^w - \lambda_{uu}}\left(\exp(B_i^w t) - \exp(\lambda_{uu}t)\right).$$

■

It follows from above, that the coefficients for the other molecules satisfy a recursive dependence,

$$n_u = 1 + \sum_{w\colon\, w\to u} n_w,$$

$$\{(A_i^u, B_i^u) : i = 1, \ldots, n_u - 1\} = \bigcup_{j=1}^{p}\left\{\left(A_k^{w_j}\frac{\lambda_{w_j}}{B_k^{w_j} - \lambda_{uu}}, B_{w_j}^k\right) : k = 1, \ldots, n_{w_j}\right\},$$

$$(A_{n_u}^u, B_{n_u}^u) = \left(\sum_{w\colon\, w\to u}\sum_{i=1}^{n_w}A_i^w\frac{-\lambda_{wu}}{B_i^w - \lambda_{uu}}, \lambda_{uu}\right), \tag{4.8}$$

which allows us to compute them by a numerical procedure. Starting from the precursor molecule, we proceed downwards and compute the coefficients using the above recursive

formulas. The algorithm uses memoization to reduce the computational time by storing coefficients of the already visited nodes. Note that the number n_u grows exponentially with the depth of the reaction graph. However, the number of distinct B_i^u values is bounded by the number of molecules in the graph. Summing A_i^u coefficients corresponding to the same B_i^u values allows to substantially limit the space complexity of the algorithm.

This leads to the following theorem.

Theorem 3. *The time complexity of Algorithm 1 in Figure 4.8 is* $O(L^2 q_0^4)$.

Proof. Observe that the algorithm is a modified DFS search, and over all the recursive calls of the algorithm the loop in Line 11 will run once for each parent-daughter molecular species pair.

Recall from Lemma 1 that there are $O(Lq_0^2)$ molecular species in graph G. Moreover, we have assumed that the secondary fragments are unobserved; therefore, there are only $O(q_0^2)$ species that have $O(L)$ daughters other than the cemetery (the ones corresponding to the non-fragmented species), while all the other species have only two children other than the cemetery. As such, the number of parent-daughter pairs is linear with respect the number of vertices. It follows that the loop in Line 11 in Algorithm 1 will run $O(Lq_0^2)$ times.

Because of the grouping step in Line 15, the size of list L for a given parent of u is bounded by the number of its ancestors, which is $O(Lq_0^2)$. The updating of a single coefficient is performed in constant time. It follows that the time complexity of one run of the loop in Line 11 in Algorithm 1 is $O(Lq_0^2)$, and the time complexity of the whole procedure is $O(L^2 q_0^4)$. ■

Validation & Results

We have applied our model to both *in silico* and on experimental data for Substance P, an 11 amino acid neuropeptide with sequence RPKPQQFFGLM.

Numerical simulations.

Numerical simulations of ETD process were performed to assess the quality of the fitting procedure under fully controlled conditions. The simulation was performed as follows: we start with a given number of Substance P precursor cations. We then simulate the electrospray ionization by placing a given number of protons on randomly chosen basic amino acids. Then, we simulate the Markov Jump Process using standard simulation techniques (Gillespie, 1977b), noting that our process can be simulated as if the cations reacted independently of each other. Ions that find themselves in the same state at the end of the

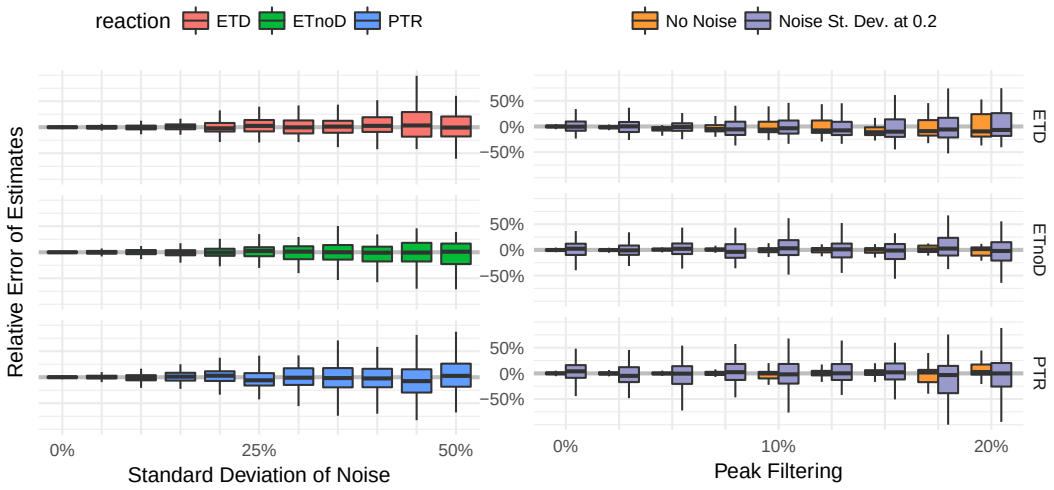


Figure 4.4: Relative errors of the fitting procedure on in silico Substance P data. The known true values of parameters are respectively $P_{\text{ETD}} = 30\%$, $P_{\text{ETnoD}} = 25\%$, $P_{PTR} = 45\%$. Cleavage probabilities were assumed to be uniform (proline being the obvious exception). Each boxplot summarizes the results of 100 independent simulations: whiskers denote the first and ninth decile and the box lids - the first and third quartiles. The left panel presents the response of the relative error of the estimates to the increasing amount of noise in the intensities reported by MassTodon. On the right panel, we study the impact of the random removal of information on the molecular species, both in noiseless conditions and with a modest amount of noise (standard deviation set to 20% of the intensity of the simulated molecule).

simulation are aggregated. The resulting counts of ions simulate results obtainable with MassTodon.

We have also analyzed the robustness of the fitting procedure to noisy or missing data. The random noise is modeled by adding Gaussian noise to the counts, with zero mean and standard deviation expressed as a given percentage of the count. Missing data is modeled by randomly removing a given proportion of the peaks. Finally, the counts obtained in this way are normalized to sum to one. Altogether, the simulation was repeated 100 times for 20 different values of data distortion parameters, see Figure 4.4.

The fitting procedure turned out to be fairly robust toward a moderate noise and missing data, see Figure 4.4. The results of the fitting procedure are unbiased. On noiseless data and data with a moderate amount of noise (up to 50% of variation in simulated intensities), the model was able to predict the reaction intensities with very high accuracy (only after introducing more than 25% of peak variation do the estimates start to surpass the limit of 50% relative error in more than 20 percent of cases).

Application to the experimental data.

Mass spectra have been acquired for purified Substance P. The precise experimental setting is described in detail by Lermyte *et al.* (2015b). The model has been fitted to 53 Substance P spectra, obtained at various travelling-wave height/velocity combinations (the design of the instrument and physical meaning of these parameters are described in detail by Lermyte *et al.* (2015b)). After fitting the model to the data, the validity of the model was further

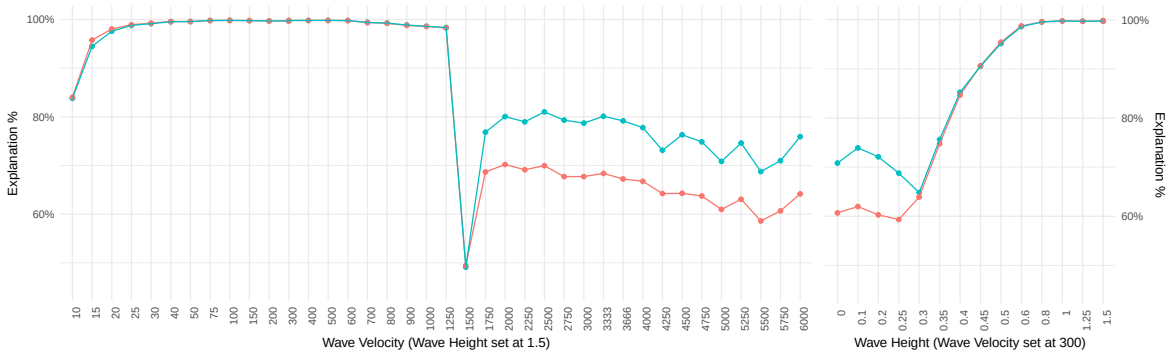


Figure 4.5: Explanation Percentage (EP) for experimental Substance P spectra. Cyan line: EP for model fit without decharging penalty ($\rho = 0$). Red line: EP for model fit with decharging penalty ($\rho = 1$). Left: EP for different values of Wave Velocity with Wave Height set to 1.5. Right: EP for different values of Wave Height with Wave Velocity set to 300.

investigated by computing the percentage of the experimental spectrum accounted for by the theoretically predicted spectrum. We call this value the *Explanation Percentage* (EP) and define it to be the common part of the theoretical and experimental spectrum. Since both spectra are normalized so that they sum to one, the Explanation Percentage can be expressed in a simple formula,

$$EP = \sum_u \min\{y_u, e_u^{\text{norm}}\}.$$

Note that because of normalization of spectra, $0 \leq EP \leq 1$. The Explanation Percentage calculated for considered data sets is presented in Figure 4.5: the values are between 50% and 98%, mostly around 60% for discharged-penalized loss function ($\rho = 1$) and 80% for non-penalized loss function ($\rho = 0$).

The predicted total intensity of all reactions, I, was found between 10^{-3} and 10 in the unconstrained case and between 10^{-3} and 10^{-1} in penalized case (data not shown). However, for reaction intensities above 0.6, the unreacted precursor molecules constitute less than 1% of the predicted spectrum, and most molecules in the spectrum are reaction products; therefore, the loss function becomes flat in this region, as further increase of base intensity causes little change in molecule proportions. This explains the large deviation between the two approaches in this case.

In regions of low reaction intensity, the explanation percentage approaches 100%; however, in these conditions, the mass spectra contain mostly unreacted precursors, and so the fitting is relatively easy to perform. In regions of high reaction intensity (wave height between 0 and 0.3, wave velocity between 10 and 20 or between 1750 and 6000) the spectra are much more informative and even then the model can explain around 70% of the input information. Similar results are obtained for different values of wave velocity. In the regions of high intensity (wave velocity above 1750) the model explains around 75% of the input.

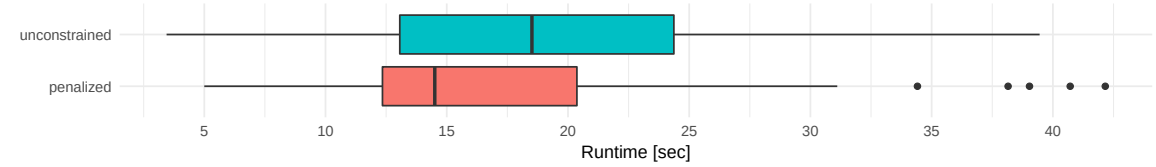


Figure 4.6: The distribution of the runtime of ETDetective, both for the unconstrained ($\rho = 0$, in cyan) and the penalized ($\rho = 1$, in red) versions of the fitting procedure.

A notable source of discrepancy between the observations and our predictions is the absence of doubly-charged precursor (i.e. product of one PTR or ETnoD), which we observe in many mass spectra. This phenomenon of missing products has been described in chemical literature by Schnier *et al.* (1995). However, the reason for this is currently unknown. As for now, our model does not account for such possibility.

In Figure 4.7 we present the results of fitting our model to the data. For different values of wave velocity, in regions of relatively high reaction intensity, we have obtained stable proportions of reaction probabilities. The proportions start to differ considerably in the region between 100 and 1250. However, in this region there are almost no reactions (less than 1% of reaction products), so the spectrum contains very little information. On the contrary, for different values of Wave Height, we have noticed a major change in reaction proportions in the regions of high reaction intensity. For Wave Height between 0.3 and 0.4, ETD is by far the most probable reaction. For higher Wave Heights, the side reactions contribute more to the spectrum. Overall, both parameters influence the reaction intensity, but only the Wave Height seems to influence the proportion of ETD to side reactions.

Finally, Figure 4.6 show that the actual runtime of ETDetective is fairly limited on the considered Substance P results.

Discussion & Conclusions

In this chapter, we have presented a kinetic model of the electron transfer driven reactions. The obtained results are promising for future work, as the model can explain around 80% of the observed intensities of the molecular species. The model is based on stochastic foundations and so the estimated parameters have a probabilistic interpretation, such as the probability of a given cleavage or reaction.

Due to its simplicity, the model described here can be used in further fundamental research into the ETD mechanism, as a discrepancy between experimental observations and the model predictions is expected to have a relatively straightforward physical interpretation. For instance, the underestimation of the asymmetry of corresponding c and z fragment intensity in the current results might indicate that a more sophisticated model of protonation sites should be used (e.g. one that accounts for electrostatic repulsion, see (Morri-

son and Brodbelt, 2016)). Similarly, using the MassTodon software, it has been recently shown (Lermyte *et al.*, 2017) that the observed ratio of PTR to ETnoD depends on protein conformation for intermediate charge states of ubiquitin and, thus, on the reaction history. A more detailed analysis could be easily performed (and similar dependencies thus revealed) using ETDetective.

A natural way for this work to proceed is to explain the influence of the instrumental settings and experimental conditions on the reaction intensity and cleavage preferences. This can be investigated using the statistical methodology, like the generalized linear models, Dirichlet regression in particular.

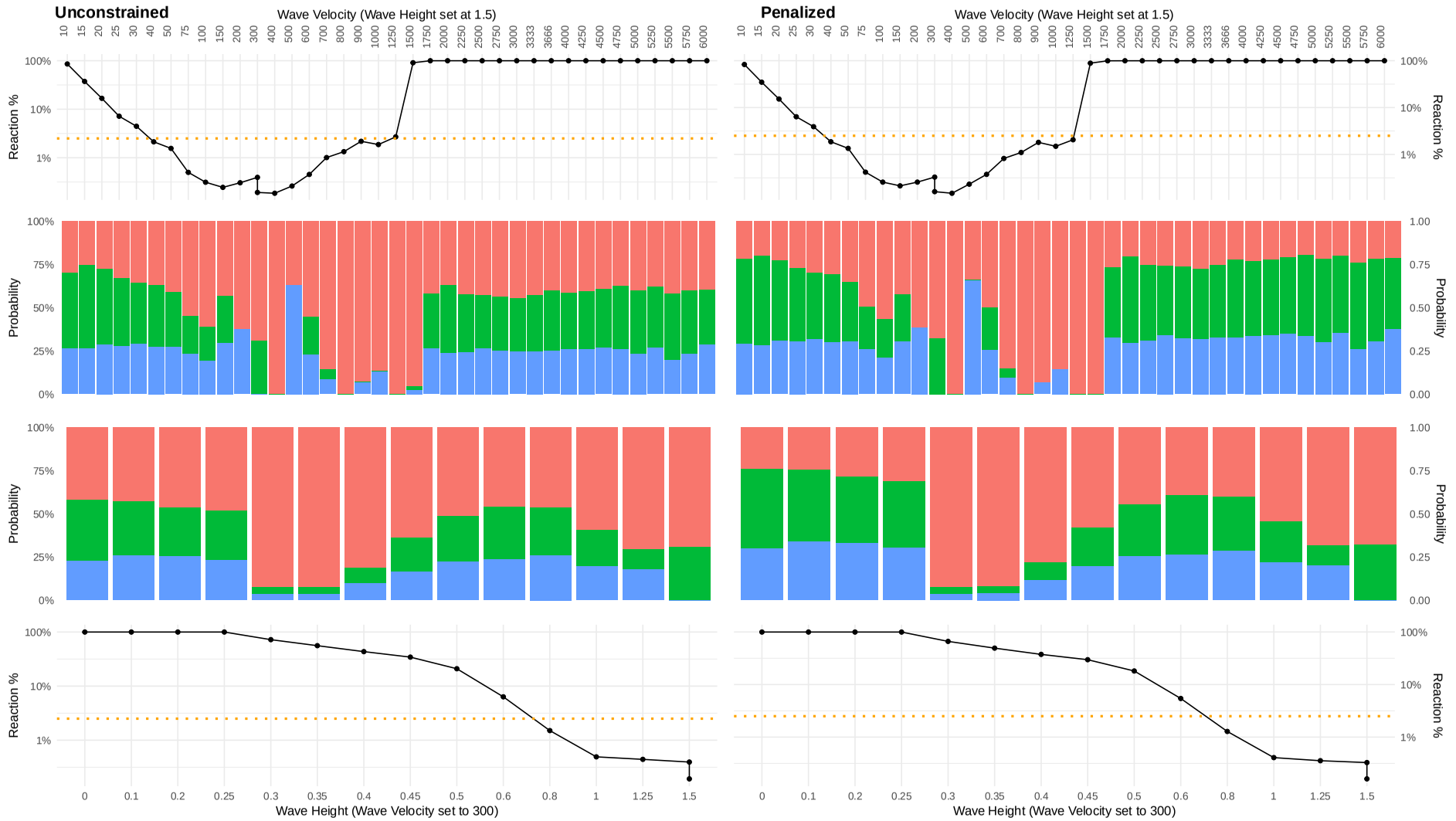


Figure 4.7: Application of ETDetective to experimental data preprocessed by MassTodon. Left: Fitting with penalty for discharging ($\rho = 1$). Right: No penalty for discharging ($\rho = 0$). Top plots show results of model fit for different values of Wave Velocity with Wave Height set to 1.5; Bottom plots show results for different values of Wave Height with Wave Velocity set to 300. Line plots: percentage of reacted molecules in predicted spectrum on the logarithmic scale. Bar plots: Percentages of PTR, ETnoD, and ETD reactions (summed over cleavage sites). The gray dashed line delimits the region in which the reaction products constitute at most 2.5% of the spectrum, and the estimated reaction probabilities are not credible.

Algorithm 1 Computation of expected numbers of molecules

Input: Reaction graph G, time t
Output: Expected numbers of molecules at time t

Procedure get_coefficients(G, u): */decorates G with Eq. (6) coefficients/*
    If u = root(G):
        Let u.coef_list := $[(A_1^r, B_1^r)]$ */list of precursor coefficients/*
        Return u.coef_list
    Else If exists u.coef_list: */if u was already visited, return the result/*
        Return u.coef_list
    Else :
        Initialize empty list C */list to store and update A_i^u, B_i^u coefficients/*
        For w in parents(u):
            Let L := get_coefficients(G, w)
            Update coefficients A_i^w according to Eq. (7)
            Append L to C
        Group and sum A_i coefficients
        Let u.coef_list := C
        Return u.coef_list

Let c := cemetery(G)
get_coefficients(G, c) */compute coefficients for all species in graph/*
For u in G:
    Compute expected number of u molecules using u.coef_list (Eq. 6)

Figure 4.8: Computation of expected numbers of molecules.

5

Deconvolution of Mass Spectra & Ion Statistics

"I am not confused. I'm just well mixed."

— Robert Frost

The deconvolution of signals originating from different molecular species is an important problem in mass spectrometry. As shown in previous sections, it is crucial for the proper understanding of what is the content of the introduced samples. In Chapter 3 we have presented an approach to deconvolution that is using constrained quadratic programming. This approach directly generalized the approach taken by Slawski *et al.* (2012), making it more robust to small shifts of spectra in the mass-to-charge ratios that can appear due to poor calibration.

Here, we present more general models of signal deconvolution. These models

- can be derived from the first principles.
- offer the possibility to estimate the how different estimates of numbers of molecular species depend between themselves. In particular, it offers means to pinpoint signals that cannot be told apart easily.
- can estimate the *ion-intensity* exchange rate – the constant that quantifies the amount of intensity that can be attributed to only one ion.

In order to achieve these tasks, we are using a *data augmentation* methods to solve the problem.

Deriving a model from the *first principles* means that the model takes into account the existing theory of the mass spec signal that asserts, that the overall number of ions observed in the instrument should approximately follow the Poisson distribution (Ipsen and Ebbels, 2012; Ipsen, 2015). As we shall see, the Poisson distribution neatly inscribes in the overall deconvolution problem, extending the existing theory to a more general setting, naturally and without contradictions.

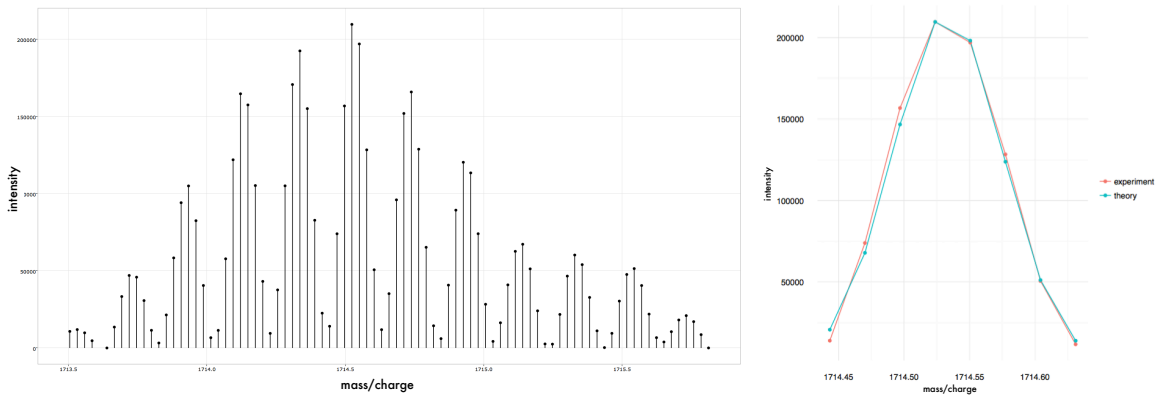
This work is not the first attempt to apply Bayesian reasoning to the problem of the deconvolution of mass spectrometry signals. For instance, Marty *et al.* (2015) tried to apply Bayesian deconvolution to ion mobility spectra. However, the recursive Richardson-Lucy algorithm (Lucy, 1974; Richardson, 1972) applied there is not fit for our needs, as it does not offer enough flexibility to model all the aspects of the problem we face.

Also, the problem of the estimation of the *ion-intensity* exchange rate has been approached, in a maximum-likelihood setting (Kaur and O'Connor, 2004). The model considers only the case of pure isotopic distributions, excluding any possibilities of a convoluted signal coming from several ion sources. The model assumes that ions are independent is based on the multinomial distribution. The model does not formally introduce the notion of the *ion-intensity* exchange rate; however, it is easy to obtain an equivalent model that would actually use that notion.

Data Preprocessing

Although we do believe that our model can be easily adjusted to model any type of mass spectrometry data, we do think it is necessary to show how such a preprocessing could be achieved on one example of data. Let us, therefore, consider data acquired with LTQ Orbitrap Velos instrument (Thermo Fisher Scientific, Bremen, Germany).

As can be noticed in Figure 5.1, even raw data from the exported `mzXML` files is already initially preprocessed. This is particularly noticeable if we consider the incredible quality of the Gaussian fit in Figure 5.1a. Therefore, it definitely does not make sense to model each particular peak in that mass spectrum. Instead, we process that signal so as to obtain in the end a mapping $A \to I_A$, where A is some range in the mass domain, and I_A denotes the total intensity in that range. Let us enumerate the ranges we look at, $(A_i)_{i=1}^W$. Then, we will also write I_i instead of I_{A_i}, as shown in Figure 5.2.



(a) *raw* data **(b)** Gaussian interpolation

Figure 5.1: LTQ Orbitrap Velos data acquired for Ubiquitin (as described in detail by Lermyte *et al.* (2015a)), as exported by the instrument in the `mzXML` format. Observe the repeating bitonic patterns of the consecutive intensity peaks in 5.1a. A closer examination of the 1714.525 Da centered peak in 5.1b shows, that the observed data very closely resembles Gaussian distribution (naïvely fitted with the method of moments). This shows, that the raw data that Orbitrap outputs must be already subject to some form of preprocessing.

Data Generation Model

We assume that the input for the problem consists of a mapping $A \to I_A$, a function $A \to I_A$, where A are some ranges in the mass-to-charge ratio and I_A is the observed intensity within that range.

It is natural to assume, that the total number of ions generated by the m^{th} molecular species in the spectrum, denoted by $N_{\cdot m}$, follows the Poisson distribution (Ipsen and Ebbels, 2012; Ipsen, 2015). The theoretical argument behind this is as follows: the number of ions that actually make it to the detector is very limited. If one assumes that ions move independently throughout the instrument, with some chance of reaching their final destination, then the number of successful detections is directly modeled by the binomial distribution. That distribution is usually well approximated by the Poisson distribution, which is sometimes referred to as the *distribution of rare events*. Denote by Λ_m the intensity of $N_{\cdot m}$. Otherwise said, it is the expected number of ions of the molecular species m. It is natural to choose that Λ_m follows the gamma distribution, as it is the conjugate distribution to the Poisson distribution (Wasserman, 2013). The gamma distribution has density $\gamma_{a,b}(x)$ proportional to $x^{\alpha-1}e^{-\beta x}$. It depends upon two hyper-parameters that need to be chosen in advance. It is natural to assume $\alpha = 1$, as this results in an *a priori* distribution with a mode at 0, as we should expect *a priori* that a molecular species is absent.

For each molecular species m we consider an independent Poisson process X_m. Each process is characterized by an intensity measure μ_m. The process X_m should spread the Λ_m ions along the mass-to-charge half-line. Otherwise said, conditionally on Λ_m (drawn itself from the gamma distribution), $\mu\big([0,\infty)\big) = \Lambda_m$. A self-imposing way to define how that measure spreads over $[0,\infty)$ is to assume that it is proportional to the fine isotopic

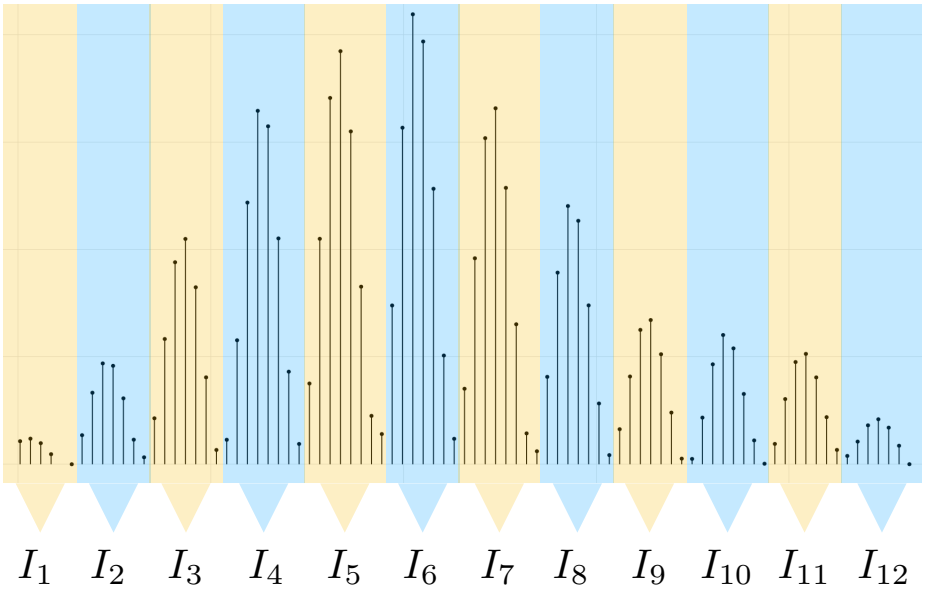


Figure 5.2: Orbitrap Preprocessing Strategy. We divide peaks into clusters so that each cluster contains intensities that form a bitonic sequence, going first up and then down. The precise division into clusters can be achieved in many ways, without any major impact on the overall calculations. The intensities of peaks in each cluster are aggregated.

structure. However, that distribution alone cannot be used, as it is atomic and the intensity measure of a Poisson process has to be absolutely continuous with respect to the Lebesgue measure (Kingman, 1993). In the context of the high-resolution FT-ICR instruments, ions scatter around their expected mass to charge ratios following a mass precision distribution. The usual choices for that distribution are either Gaussian or Lorentzian (Posener, 1974; G. Marshall *et al.*, 2013). For now, we do assume the distribution to be centered at zero. Equivalently, we assume the spectra to be well calibrated. In actual implementation, it is better to use truncated distributions, so as to attack different deconvolution problems independently and – potentially – in parallel. Therefore, the natural intensity measure amounts to a mixture of normal distributions, multiplied by the average number of ions.

$$\mu_m(B|\Lambda_m) = \Lambda_m \sum_{\mathtt{mz}\in\mathtt{MZ}_m} p_{\mathtt{mz}}\mathtt{G}(B-\mathtt{mz}), \tag{5.1}$$

where B is some probing range on the m/z half line, and $\mathtt{G}$ is the chosen mass precision distribution. The peaks are assumed to be identifiable uniquely by their m/z ratio $\mathtt{mz}$ and that all these values form a molecule specific set $\mathtt{MZ}_m$[1]. To make the calculations manageable, instead of using full isotopic distribution, we use an optimal P-set instead, as provided by the `IsoSpec` algorithm. The probabilities in that distribution are divided by $1-P$, so that the new measure remains probabilistic. Observe that by trimming, we do assume to concentrate our attention only to the information contained in ranges $(A_w)_{w=1}^{W}$, so that the

[1]For certain values of the parameters of the multinomial distribution, it might be possible that more than one isotopologue can be identified by precisely the same mass-to-charge ratio. In that cases, we do sum them up to form one peak instead.

resulting variables Λ_m and $N_{m\cdot}$, in fact, describe only quantities of molecular species in that sets. In particular, we do not integrate out with respect to the unobserved data outside these ranges, which would be difficult and would not significantly affect the outcomes.

Let us choose some mass range A. For instance, it is natural to choose a set defining a bin as obtained in the previous section for the Orbitrap instrument. Given that the sum of independent Poisson random variable is also a Poisson variable, the overall number of ions observed in set A should also follow the Poisson law, with intensity equal to the sum of the constituent intensities,

$$N_A \sim \text{Poisson}\Big(\sum_{m=1}^{M} \Lambda_m \sum_{\mathtt{mz}\in \mathtt{MZ}_m} p_{\mathtt{mz}}\mathtt{G}(A-\mathtt{mz})\Big). \tag{5.2}$$

What follows from the properties of the process, if some ranges A and B do not intersect, then variables N_A and N_B are also independent.

Apart from considering different isotopic envelopes, we still need more control over how the intensity is created from the actual ion count N_A. The problem of how the number of ions is reflected in the mass spectrometer is complex and boils down to different technologies being used at the detection step. De Hoffmann and Stroobant (2007) enumerate at least two big classes of detectors: (1) ion counters, and (2) array collectors. The ion counters can detect one ion during some given time lapse (but they fail to record two and more). On the other hand, the array collectors can record multiple mass-to-charge ratios simultaneously. There is little research that aims at quantifying the precise relationship between the actual number of ions N_A and the recorded intensity I_A. It is mostly assumed, that for some of the array collectors that dependence is linear (Koppenaal *et al.*, 2005), at least within the dynamic range of the detector. This is also tacitly assumed by some of the existing statistical research on the topic (Kaur and O'Connor, 2004); however, this research has been carried out on an FT-ICR instrument. The LTQ Orbitrap Velos is paired with a patented dual conversion dynode detector, two off-axis continuous dynode electron multipliers with extended dynamic range, and a digital electronic noise discrimination system[2]. The nature of the *intensity-ion count* relationship is not known in advance here. Here, we will follow a simple approach that assumes that the data has been indeed gathered at the dynamic range of the detector. Therefore, we will assume that

$$I_w = CN_w + \epsilon_w,$$

where $C > 0$ is the *intensity-ion count* exchange rate and ϵ_w is the measurement error. In particular, the above specification abstracts from detector overfill, as $\lim_{N_w\to\infty} I_w = \infty$.

For mathematical convenience, we choose a truncated Gaussian distribution as a prior

[2] As advertised in that company's leaflet.

for C. We truncate it to the positive halfline $(0,\infty)$, and set the mode to 1[3] and set unit variance so that its density is given by

$$f(c) = \frac{\mathbb{1}_{[0,\infty)}}{[1-\Phi(0|1,1)]\sqrt{2\pi}} \exp\Big(-\frac{(c-1)^2}{2}\Big),$$

where $\Phi(0|m,\sigma^2)$ is the cumulative distribution function of the normal distribution with mean m and variance σ^2. The truncation is a natural requirement, as the dependence between the observed intensity and the number of ions should be positive.

Also, we assume that the error of ϵ_w follows a normal distribution with mean ξ and variance ν^2. Also, coordinates of vector $\epsilon = (\epsilon_w)_{w=1}^W$ are assumed mutually independent given (ξ,ν^2). Whenever necessary, we will denote the density of the Gaussian r.v. with mean m and variance σ^2 at point x by $g_{m,\sigma^2}(x)$. We put a Gaussian prior on ξ, with mean zero and unit variance, so that its density is $g_{0,1}(\xi)$. Finally, we assume that $\frac{1}{\nu^2}$ follows the gamma distribution with both parameters set to one.

Notation

From now on, we will denote $e^\alpha \frac{\alpha^k}{k!}$ by $\texttt{poiss}(k|\alpha)$. Also, $\Lambda = (\Lambda_m)_{m=1}^M$, $a = (a_m)_{m=1}^M$, $b = (b_m)_{m=1}^M$, and $I = (I_w)_{w=1}^W$.

By $N_{m\cdot}$ we already denote the total number of ions of the m^{th} molecular species. This number is divided into parts appearing in different mass ranges A_w. We arrange all these counts into a matrix $\mathtt{N}$. We also denote by $\mathtt{N}_{m-}$ the m^{th} row of $\mathtt{N}$, assuming that it is a standing vector. It contains all counts generated by the m^{th} molecular species. By $\mathtt{N}_{|w}$ we denote the w^{th} column of $\mathtt{N}$. It is also a standing vector and contains different components of the overall number of ions found in the w^{th} mass-to-charge range, A_w. Altogether,

$$\mathtt{N} = \begin{bmatrix} N_{11} & \dots & N_{1W} \\ \vdots & \ddots & \vdots \\ N_{M1} & \dots & N_{MW} \end{bmatrix} = \begin{bmatrix} \mathtt{N}_{1-}^t \\ \vdots \\ \mathtt{N}_{M-}^t \end{bmatrix} = \begin{bmatrix} \mathtt{N}_{|1} & \dots & \mathtt{N}_{|W} \end{bmatrix}, \tag{5.3}$$

where $\mathtt{N}_{m-}^t$ denotes the transposition of $\mathtt{N}_{m-}$. Denote by $\mathtt{N}_{\text{sums}}^{\text{row}}$ the vector of sums of the row entries of $\mathtt{N}$ and by $\mathtt{N}_{\text{sums}}^{\text{col}}$ the vector of sums of the column entries of $\mathtt{N}$,

$$\text{and} \quad \mathtt{N}_{\text{sums}}^{\text{row}} = \sum_{w=1}^W \mathtt{N}_{|w} = \begin{bmatrix} N_{1\cdot} \\ \vdots \\ N_{m\cdot} \end{bmatrix}, \quad \mathtt{N}_{\text{sums}}^{\text{col}} = \sum_{m=1}^M \mathtt{N}_{m-} = \begin{bmatrix} N_{\cdot 1} \\ \vdots \\ N_{\cdot W} \end{bmatrix}.$$

[3] So that the exchange rate is believed to be equal to one. This Gaussian choice is motivated by the ease of drawing from the conjugate distribution. That choice, of course, gives more probability to the region right to the mode than left to the mode. An alternative specification results from parametrizing C as $\text{cotan}(\alpha)$ and putting a uniform prior on α. This leads to the Cauchy distribution.

Let $D_{mw} = \sum_{\mathtt{mz}\in\mathtt{MZ}_m} p_{\mathtt{mz}}\mathtt{G}(A^w - \mathtt{mz})$. We can collect all D_{mw} in a matrix $\mathtt{D}$ calculated at the onset of the algorithm. Denote the columns and rows of $\mathtt{D}$ similarly to the notation used for $\mathtt{N}$ in Eq. (5.3),

$$\mathtt{D} = \Big[\Big(\sum_{\mathtt{mz}\in\mathtt{MZ}_m} p_{\mathtt{mz}}\mathtt{G}(A_w - \mathtt{mz})\Big)_{mw}\Big] = \begin{bmatrix} D_{11} & \dots & D_{1W} \\ \vdots & \ddots & \vdots \\ D_{M1} & \dots & D_{MW} \end{bmatrix} = \begin{bmatrix} \mathtt{D}^t_{1-} \\ \vdots \\ \mathtt{D}^t_{M-} \end{bmatrix} = \begin{bmatrix} \mathtt{D}_{|1} & \dots & \mathtt{D}_{|W} \end{bmatrix}.$$

$$\text{and} \quad \mathtt{D}^{\text{row}}_{\text{sums}} = \sum_{w=1}^{W} \mathtt{D}_{|w} = \begin{bmatrix} D_{1\cdot} \\ \vdots \\ D_{m\cdot} \end{bmatrix}, \quad \mathtt{D}^{\text{col}}_{\text{sums}} = \sum_{m=1}^{M} \mathtt{D}_{m-} = \begin{bmatrix} D_{\cdot 1} \\ \vdots \\ D_{\cdot W} \end{bmatrix}.$$

As mentioned before, the mass precision distribution $\mathtt{G}$ is either a truncated Gaussian or Lorentzian. In particular, most evaluations of $\mathtt{G}(A^w - \mathtt{mz})$ are zero, reducing the number of integral evaluations from $W \times M$ to a much smaller number of all generated isotopologues. Thus, matrix $\mathtt{D}$ will be sparse. What is more, the calculations necessary to establish $\mathtt{D}$ naturally parallelize. Evaluating $\mathtt{G}$ could be avoided altogether, if we assumed, similarly to what was done in the `MassTodon` project, that we cannot tell apart mass-to-charge ratios in small mass ranges; however, we will not follow this approach here, as we want to present a more general scheme.

The Posterior Distributions

The joint density[4] p of a list $(C, \Lambda, I, \mathtt{N}^{\text{col}}_{\text{sums}}, \xi, \nu^{-2})$ given hyperparameters $\Xi = (\mathtt{D}, a, b)$ can be written as

$$p\big(C, \Lambda, I, \mathtt{N}^{\text{col}}_{\text{sums}}, \xi, \nu^{-2}|\Xi\big) \propto \overbrace{f(C)g_{0,1}(\xi)\gamma_{1,1}(\nu^{-2})\prod_{m=1}^{M}\gamma_{a_m,b_m}(\Lambda_m)}^{\text{prior distribution}} \tag{5.4}$$

$$\times \prod_{w=1}^{W} \mathtt{poiss}\Big(N_{\cdot w}\Big|\Lambda^t\mathtt{D}_{|w}\Big) \tag{5.5}$$

$$\times \prod_{w=1}^{W} \exp\Big(-\frac{(I_w - \xi - CN_{\cdot w})^2}{2\nu^2}\Big). \tag{5.6}$$

We are ultimately interested in generating samples from $p(C, \Lambda, \xi, \nu^{-2}|I, \Xi)$, which is proportional to the function described above but integrated out over all possible values of the unobservable $\mathtt{N}^{\text{col}}_{\text{sums}}$,

$$p(C, \Lambda, \xi, \nu^{-2}, |I, \Xi) \propto \sum_{\mathtt{n}^{\text{col}}_{\text{sums}}} p\big(C, \Lambda, I, \mathtt{n}^{\text{col}}_{\text{sums}}, \xi, \nu^{-2}|\Xi\big).$$

[4] We will not distinguish between continuous and discrete random variables in the naming convention. Thus, a probability distribution function will be referred to as *density*.

The problem is too difficult to solve using paper and pencil methods, given that it involves complicated summations. We thus apply Markov Chain Monte Carlo methods, MCMC. Given that we have to integrate out one set of variables, we have to resolve to special *data augmentation* techniques.

Bayesian Calculations

To solve the problem in a fully Bayesian way we consider *data augmentation* (Tanner and Wong, 1987; Van Dyk and Meng, 2001). The general idea behind data augmentation is the following: denote by X a random variable that describes the outcomes, and by Y a random variable that describes the parameters. Also, assume that there exists a dummy variable Z, such that it is easier to draw from $Y|X,Z$ then it is from $Y|X$ alone, and such that it is possible (and easy) to draw $Z|X,Y$ is possible. Also, the distribution of Z must be consistent with that of $Y|X$, in the sense that the marginal distribution of $Y|X,Z$ must be equal to $Y|X$, $\mathbb{P}(Y|X) = \int \mathbb{P}(Y|X,Z)d\,\mathbb{P}(Z)$. If that is the case, then one can consider a *Gibbs-like* algorithm that alternates between conditional distributions $Y|X,Z$ and $Z|X,Y$ (Geman and Geman, 1984). To be more specific, one produces a Markov Chain $(Y^{[n]},Z^{[n]})$, such that $(Y^{[0]},Z^{[0]})$ can be drawn from any distribution, and $Y^{[n]}$ is drawn from given $(X,Z^{[n-1]})$, $Y^{[n]}|X,Z^{[n-1]}$, and $Z^{[n]}$ is drawn given $(X,Y^{[n]})$, $Z^{[n]}|X,Y^{[n]}$. The first coordinate of the chain, $(Y^{[n]})_{n=0}^{\infty}$, can then be shown to converge in distribution to $Y|X$ with the standard MCMC theory (Geyer, 1992; Gilks *et al.*, 1995). The *data augmentation* approach then boils down to drawing at random the quantities we miss, albeit with care, so as not to change the distribution we ultimately want to draw from. In the current setting, we slightly departure from the scheme presented above. However, we will now show that finding the appropriate set of *augmenting variables* is easy and natural.

In order to perform the Gibbs algorithm, we need to know how to draw C given that all other variables are fixed, and Λ given that all other variables are fixed, and so on, for all the remaining variables ξ, ν^{-2}. Observe, that if we knew how many ions of each substance was there in every mass range, then all the necessary calculations would be easy. Otherwise said, we want to augment the problem by considering matrix $\mathtt{N}$. Figure 5.3 presents the dependence structure of the augmented problem in form of a Bayesian net.

Drawing matrix $\mathtt{N}$ corresponds to disaggregating the middle part of the overall joint density p, i.e. that given by Eq. (5.5). Specifically,

$$\mathtt{poiss}\left(N_{\cdot w}\middle|\Lambda^t\mathtt{D}_{|w}\right) = \sum_{N_{1w}+\cdots+N_{Mw}=N_{\cdot w}} \prod_{m=1}^{M} \mathtt{poiss}\left(N_{mw}\middle|\Lambda_m D_{mw}\right).$$

So, after the deaggregation we substitute each $\mathtt{poiss}\left(N_{\cdot w}\middle|\Lambda^t\mathtt{D}_{|w}\right)$ by the corresponding

product $\prod_{m=1}^{M} \texttt{poiss}\big(N_{mw}\big|\Lambda_m\delta_{mw}\big)$, so that the overall density amounts to

$$\text{prior}(\Lambda, C, \xi, \nu^{-2}) \prod_{w=1}^{W}\prod_{m=1}^{M} \texttt{poiss}\big(N_{mw}\big|\Lambda_m D_{mw}\big) \prod_{w=1}^{w} \exp\Big(-\frac{(I_w-\xi-CN_{\cdot w})^2}{2\nu^2}\Big). \quad (5.7)$$

Above, we keep $N_{\cdot w}$ to ease the notation. Similarly, we will retain $N_{m\cdot}$. Given $\texttt{N}$, these values are entirely deterministic and they don't add up to the complexity of the problem. Otherwise said, the distribution of $\texttt{N}^{\text{col}}_{\text{sums}}$ is entirely replaced by that of $\texttt{N}$.

Generating Λ from conditional distribution

Note that the distribution of Λ_m depends only upon the values of $\texttt{N}_{m-}$ and their corresponding intensities $\texttt{D}_{m-}$: if we collect all the terms containing Λ_m in Eq. (5.7), i.e. consider its *Markov blanket* (Hasman, 1991), we end up with expression proportional to

$$\Lambda_m^{a_m-1}e^{-b_m\Lambda_m}\prod_{w=1}^{W} e^{-\Lambda_m D_{mw}}\frac{(\Lambda_m D_{mw})^{N_{mw}}}{N_{mw}!} \propto \Lambda_m^{a_m+N_{m\cdot}-1}e^{-(b_m+D_{m\cdot})\Lambda_m},$$

so that the posterior distribution of Λ_m is gamma with parameters $a'_m = a_m + N_{m\cdot}$ and $b'_m = b_m + D_{m\cdot}$. Also, different Λ_m are conditionally independent. Therefore, we now know how to draw vector Λ.

Generating $\texttt{N}$ from conditional distribution

Next, let us consider how to draw matrix $\texttt{N}$. For simplicity, we will draw this matrix in a two-step procedure, amounting to

1. drawing the sums of columns $\texttt{N}^{\text{col}}_{sums} = (N_{\cdot 1}, \ldots, N_{\cdot W})$
2. drawing the columns entries $\texttt{N}_{|w}$ given the appropriate sum $N_{\cdot w}$ drawn in step 1.

First, we will describe how to perform the second step. Observe, that the conditional distribution of $\texttt{N}$ depends only upon $\texttt{N}^{\text{col}}_{sums}$ (numbers of ions in each mass-to-charge range), $\texttt{D}$, and Λ. When we pull out part of formula Eq.(5.7) that is proportional to the entries of $\texttt{N}$, we end up with

$$\prod_{w=1}^{W} \mathbb{1}_{\{N_{1w}+\cdots+N_{Mw}=N_{\cdot w}\}}(\texttt{N}_{|w}) \prod_{m=1}^{M} \texttt{poiss}\big(N_{mw}\big|\Lambda_m D_{mw}\big).$$

Above, we added the characteristic functions of events $N_{1w}+\cdots+N_{Mw} = N_{\cdot w}$ to underline that the sums of columns are fixed to certain values. In particular, note that the columns of matrix $\texttt{N}$ are conditionally independent, as the above expression in product form. Each vector $\texttt{N}_{|w}$ can be seen to originate from a different multinomial distribution, as shown below.

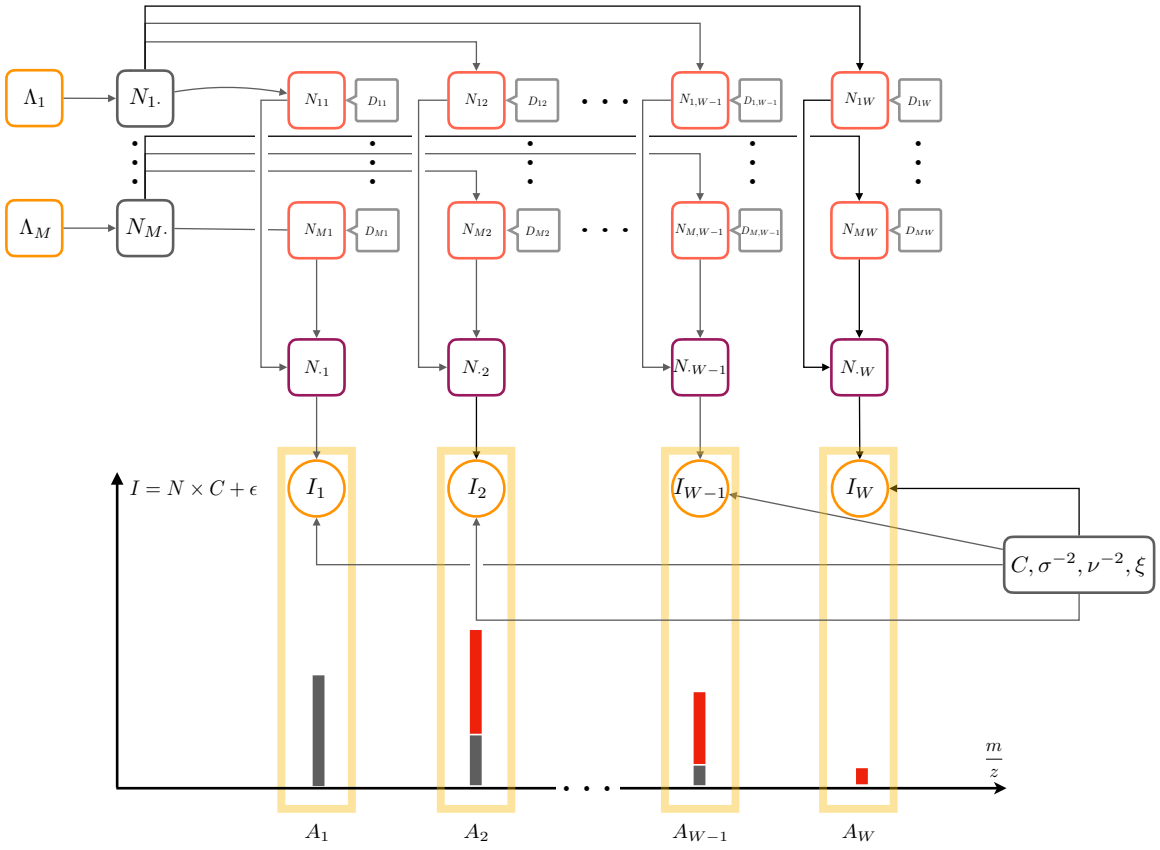


Figure 5.3: Bayesian net representation of the *data augmented* deconvolution problem attacked by MassOn. In top-left we put the expected numbers of ions inside the instrument, denoted by Λ_1 to Λ_M. Each Λ_m corresponds to a different source of ions. Ions in group m can be found in different mass ranges A_w (bottom), where intensity I_w has been observed. The number of ions the m ions in group w is denoted by N_{mw}. This number depends on the probability of finding ions of species m in that group, denoted by p_{wm}. These probabilities can be calculated as presented in Chapter 3 and are conditionally independent of all other variables given the values of the *matrix of ion counts* $\mathtt{N} = [N_{mw}]$. Sums of columns $\mathtt{N}^{\text{col}}_{\text{sums}} = (N_{\cdot 1}, \ldots, N_{\cdot W})$ and sums of rows $\mathtt{N}^{\text{row}}_{\text{sums}} = (N_{1\cdot}, \ldots, N_{M\cdot})$ are deterministic functions of $\mathtt{N}$ and are shown only for convenience, i.e. they do not introduce any randomness to the problem. Numbers $N_{\cdot w}$ are not directly observed; we observe the intensity they induce, which is assumed to follow equation $I_w = N_{\cdot w} \times C + \epsilon_w$.

Lemma. *Suppose that* $X_k \sim \text{Poisson}(\lambda_k)$ *are a sequence of* K *independent random variables. Then,* $(X_1, \ldots, X_K) | \sum_{k=1}^K X_k$ *is multinomial,*

$$\text{Multi}\Big(\sum_{k=1}^K X_k; \frac{\lambda_1}{\sum_{k=1}^K \lambda_k}, \ldots, \frac{\lambda_K}{\sum_{k=1}^K \lambda_k}\Big).$$

An easy proof is given by Kingman (1993). In our case, variables in column w are Poisson distributed with intensities equal to $\Lambda_m D_{mw}$, so that we can draw each column from a multinomial distribution with marginal probabilities proportional to $(\Lambda_m D_{mw})_{m=1}^M$. The above vector can be concisely written using the Hadamard's pointwise product operator $\odot$. Given vectors $v = (v_1, \ldots, v_k)$ and $w = (w_1, \ldots, w_k)$, by definition $v \odot w = (v_1 w_1, \ldots, v_k w_k)$. Then,

$$\mathtt{N} = \begin{bmatrix} \mathtt{N}_{|1} & \ldots & \mathtt{N}_{|W} \end{bmatrix} \sim \otimes_{w=1}^W \text{Multi}\Big(N_{\cdot w}; \frac{\Lambda \odot \mathtt{D}_{|w}}{\Lambda^t \, \mathtt{D}_{|w}}\Big).$$

It is possible that some $D_{mw} = 0$. In that case, we reduce the dimension of the multinomial distribution from M to that of the number of nonzero entries of the vector $\mathtt{D}_w$, ne-

glecting the zero probabilities (as surely these coordinates amount to zero counts). Finally, we do not deconvolve peaks outside of the support of any of the intensity measures μ_m and treat these peaks as not explainable. Therefore, it cannot happen that vector $\mathtt{D}_{|w}$ consists only of zeros. However, even that is easy to cope with, as then $\mathtt{N}_{|w} = \mathbf{0}$ with probability 1.

We now pass to the problem of drawing new values of counts of ions in each mass range, or $\mathtt{N}^{\text{col}}_{sums}$. In that stage, the values of the matrix $\mathtt{N}$ are not yet drawn (in fact, we are in the middle of drawing them), so that the *Markov blanket* of $\mathtt{N}^{\text{col}}_{sums}$ does actually include all other variables. The joint conditional density of these variables is proportional to

$$\prod_{w=1}^{W} \left\{ \mathtt{poiss}\left(N_{\cdot w} \middle| \Lambda^t \mathtt{D}_{|w}\right) \exp\left(-\frac{(I_w - \xi - CN_{\cdot w})^2}{2\nu^2}\right) \right\}.$$

It follows that different that ion counts in different mass-to-charge ranges are independent. The conditional distribution for each given w is certainly not any known distribution, given that its `pdf` is proportional to

$$\mathbb{P}(n|a,b,c^2) \propto \frac{a^n}{n!} \exp\left(-\frac{(n-b)^2}{2c^2}\right) = A(n), \tag{5.8}$$

where $n \in \mathbb{N} \cup \{0\}$, $a, b, c^2 > 0$. We have to invent a method of drawing random variables $\mathbb{P}(n|a,b,c)$. To this end, we could have used the Metropolis-Hastings algorithm, resulting in an overall *Metropolis-within-Gibbs*[5] setting. However, the straightforward form of $\mathbb{P}(n|a,b,c)$ suggests using a conceptually simpler rejection algorithm. What we need, is a function $B(n)$ that dominates $A(n)$ that is proportional to a `pdf` we know how to draw random samples from. In particular, consider Stirling's lower bound approximation of the factorial (Nemes, 2010),

$$n! \geq \sqrt{2\pi} n^{n+0.5} e^{-n},$$

valid for all positive integers. Using it, we can estimate $A(n)$ from above in the following way:

$$A(n) \leq \frac{a^n}{\sqrt{2\pi} n^{n+0.5} e^{-n}} \exp\left(-\frac{(n-b)^2}{2c^2}\right) = (2\pi)^{-0.5} \exp\left(-\frac{(n-b)^2}{2c^2} + n\log(a) - (n+0.5)\log(n) + n\right) \leq$$
$$(2\pi)^{-0.5} \exp\left(-\frac{(n-b)^2}{2c^2} + n\log(a) - (n+0.5)\log(2) + n\right) = \sqrt{2}\exp\left(\frac{m_B^2}{2c^2}\right)(2\pi)^{-0.5}\exp\left(-\frac{(n-m_B)^2}{2c^2}\right) = B(n),$$

where $m_B = b + c^2(\log a - \log 2 + 1)$ The above estimate works for $n \geq 2$. Otherwise, we set $B(0) = A(0)$ and $B(1) = A(1)$. To draw from a `pdf` proportional to $B(n)$, we consider the scheme presented in Figure 5.4. We note, that any distribution proportional to $B(n)$ on $\{2, 3, \dots\}$ can be naturally extended to a distribution that is continuous w.r.t. Lebesgue measure by considering an infinite mixture of uniform random variables defined over the natural grid $[n, n+1)$. That measure is dominated by a Gaussian density with mean m_B

[5] In that approach, instead of drawing directly from the conditional distribution, one performs a step of the Metropolis-Hastings algorithm.

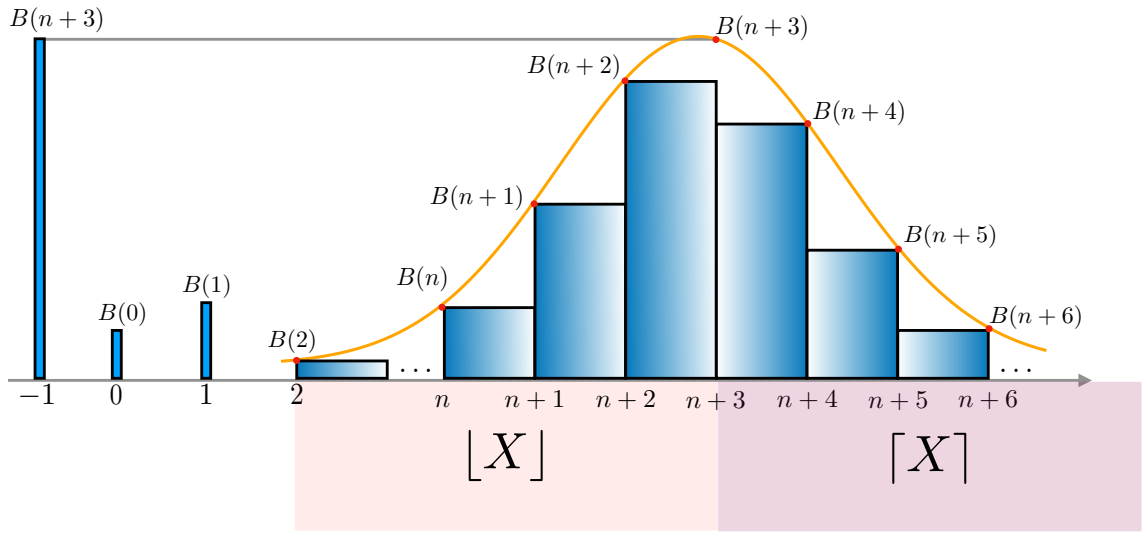


Figure 5.4: The rejection algorithm for drawing from density proportional to $B(n)$. The idea is to apply another rejection from a mix of a continuous distribution defined over $[2, \infty)$ and over a set of atoms $\{-1, 0, 1\}$. The -1 value is mapped to $n+3$, which is the global maximum of the $B(n)$ function on $\{2, 3, 4, \dots\}$. The natural candidate for the dominating distribution is the Gaussian with parameters m_B and c_B^2.

and variance c^2. To assure that the above density always dominates $B(n)$, we have to take out the rectangle corresponding to the maximum of B. We can then map all rectangles right to the mode to the numbers *next in line*. For example, in Figure 5.4 any draw X left from $n+3$ corresponds ultimately to integer $\lfloor X \rfloor$, and any draw X right from $n+3$ to $\lceil X \rceil$.

Recall that by $\Phi(x|m, \sigma^2)$ we denote the cumulative distribution function of the normal distribution with mean m and variance σ^2, and that its density is denoted by g_{m,c^2}. Also, denote by $\mathcal{N}(m_B, c^2)|_{[2,\infty}$ the normal distribution truncated to half-line $[2, \infty)$. The actual computation can be performed with the code snippet described in Algorithm 5. Observe that the inner while loop is called the less, the bigger is the chunk of probability we managed to get out in form of the atomic component of the distribution. This motivates a further improvement of the idea, which is similar in nature to one presented in the `IsoSpec` algorithm, which amounts to calculating a bigger set of configurations directly from A and use the above estimate for drawing from the tails of the distributions, that we would approximate from above with Gaussian tails. This can be achieved, as clearly the distribution $\mathbb{P}(n|a, b, c^2)$ in question is at most bimodal, being proportional to a Poisson distribution with intensity a and a normal distribution discretized over an equally spaced grid.

Ultimately, to restate the result in terms of other parameters of the model, the conditional distribution of the sums of columns of $\mathtt{N}$ can be seen to follow

$$\mathtt{N}^{\text{col}}_{\text{sums}} = (N_{\cdot 1}, \dots, N_{\cdot W}) \sim \otimes_{w=1}^{W} \mathbb{P}\left(\circ \left| \Lambda^t \mathtt{D}_{|w}, \frac{I_w - \xi}{C}, \frac{\nu^2}{C^2} \right. \right),$$

from which we can draw using Algorithm 5.4 W times independently.

Generating C from conditional distribution

Let us now focus on the function proportional to the conditional density of intensity-ion count exchange rate C,

$$\frac{\mathbb{1}_{[0,\infty)}}{[1-\Phi(0|1,1)]\sqrt{2\pi}}\exp\Big(-\frac{(C-1)^2}{2}\Big)\prod_{w=1}^{W}\exp\Big(-\frac{(I_w-\xi-CN_{\cdot w})^2}{2\nu^2}\Big).$$

The above distribution is proportional to another truncated normal distribution, with density

$$\frac{\mathbb{1}_{[0,\infty)}(c)}{[1-\Phi(0|m_C,\sigma_C^2)]\sqrt{2\pi}}\exp\Big(-\frac{(c-m_C)^2}{2\sigma_C^2}\Big),$$

with parameters

$$m_C=\frac{\nu^2+\sum_{w=1}^{W}N_{\cdot w}(I_w-\xi)}{\nu^2+\sum_{w=1}^{W}N_{\cdot w}^2}\quad\text{and}\quad\sigma_C^2=\frac{\nu^2}{\nu^2+\sum_{w=1}^{W}N_{\cdot w}^2}.$$

Both parameters can be easily retrieved by collecting terms of a binomial.

Generating ξ from conditional distribution

The function of ξ proportional to its conditional density is the following

$$\exp\Big(-\frac{\xi^2}{2}\Big)\prod_{w=1}^{W}\exp\Big(-\frac{(\xi-(I_w-CN_{\cdot w}))^2}{2\nu^2}\Big).$$

Again, by collecting terms of a binomial, we see that the distribution of ξ given all other parameters is Gaussian with mean m_ξ and variance σ_ξ^2 given by

$$m_\xi=\frac{\sum_{w=1}^{W}(I_w=CN_{\cdot w})}{\nu^2+W}\quad\text{and}\quad\sigma_\xi^2=\frac{\nu^2}{\nu^2+W}.$$

Generating ν^{-2} from conditional distribution

The function of ν^{-2} proportional to its conditional density is the following

$$\exp\big(-\nu^{-1}\big)\prod_{w=1}^{W}\sqrt{\nu^{-2}}\exp\Big(-\frac{(I_w-\xi-CN_{\cdot w})^2}{2\nu^2}\Big)=$$
$$(\nu^{-2})^{\frac{W}{2}+1-1}\exp\left(\nu^{-1}\Big[1+\sum_{w=1}^{W}(I_w-\xi-CN_{\cdot w})\Big]\right),$$

so that we see that we can generate ν^{-2} conditional on other variables from a gamma distribution with parameters

$$a_\xi=\frac{W}{2}+1\quad\text{and}\quad b_\xi=1+\sum_{w=1}^{W}(I_w-\xi-CN_{\cdot w}).$$

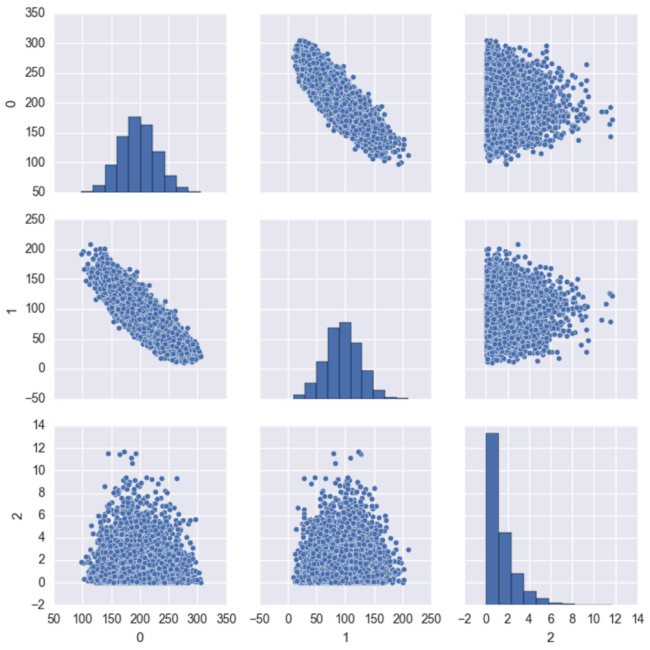


Figure 5.5: Results of Bayesian deconvolution. The plots shows the results of an attempt to estimate the content of three different molecular species based on four observed mass ranges. The original quantities were equal to 200, 100, and 0. Note that the computed approximations to the *a posteriori* distributions do concentrate around these values.

That was the last variable in the presented scheme. Algorithm 6 formalizes entire proposed scheme in a *pseudo code*.

Validation

A prototype of the approach has been implemented in Python. The prototype includes only the stage performing the deconvolution of the signal and currently does not perform inference of the parameters of the *intensity-ion count* dependence. We have tested the principles of the deconvolution on a toy example consisting of three convoluted spectra, m_1, m_2, m_3. The measurements were assumed to be gathered at four different mass-to-charge ratio bins. We assumed, that m_1 appears in reach bin with probabilities equal to $20\%, 70\%, 0\%$, and 10% respectively; m_2 with $40\%, 60\%, 0\%$, and 0%; and, finally, m_3 with $0\%, 30\%, 70\%$, and 0%respectively. The initial number of ions were fixed at 200, 100, and 0. Given the above setting, we have generated *in silico*. Fig. 5.5 presents the results of the deconvolution. Note how different substances compete to explain the results. More test will be carried out soon.

Similar Approaches & Extensions

It is not the first time that Gibbs algorithm was used for deconvolution, albeit with an entirely different context and independently of this line of research. Koronacki *et al.* (2005)

have used the same approach to deconvolve the signal obtained with positron emission tomography. This line of research investigated and implemented the deconvolution of the Poisson signal and did not extend this approach towards applications in mass spectrometry data signal processing.

It remains to note, that the same approach can be trivially generalized to continuous positive signals. In particular, consider non-negative least squares problem (Lawson and Hanson, 1995),

$$\text{argmin}_b \Big\{ ||y - Xb||_2 \ : \ \forall_{i=1}^{I} b_i \geq 0 \Big\},$$

where X is a data matrix and, y is the response vector, and b is the vector of non-negative coefficients. In mathematical terms, the problem reduces to finding the coefficients of a projection unto a cone $\sum_{i=1}^{I} \mathbb{R}_+ X^i$ spanned by the columns of X, so that it minimizes the Euclidean distance to y (Davis, 1954).

Consider now a special subcase of this problem, where all entries of the matrix X are non-negative too, $X \geq 0$. The problem can be cast into a fully Bayesian setting analogously to the problem of Poisson deconvolution. In particular, parameters b play the role of Λ, matrix X replaces the matrix $\mathtt{D}$, and the observation y replaces counts $(N^w)_{w=1}^{W}$. It is enough to replace all the Poisson distributions by the Gamma distribution and the multinomial distribution by the Dirichlet distribution. The only significant modification is required why considering the distribution of y, because the gamma distribution does not have a natural conjugate distribution. In that case, we can choose any distributions on the gamma parameters that make sense and perform all the calculations using the *Metropolis-within-Gibbs* update.

Also, note that the presented scheme can be generalized to model the detector in more details. In particular, we could model regions outside the dynamic range of the detector by introducing nonlinear functions. In particular, detector overfill could be modeled by using some sigma-shaped function.

Algorithm 5 Generating ion numbers in different mass ranges from $\mathbb{P}(\circ|a, b, c)$.

Compute $p_{-1} := B(\lceil b \rceil)$, $p_0 := B(0)$, $p_1 := B(1)$, and $p_2 := 1 - \Phi(2)$.
Generate $I \sim p_I$.
if $I > 1$ **then**
  repeat
    Generate $X \sim \mathcal{N}(m_B, c^2)|_{[2,\infty)}$
    Generate $U \sim \mathcal{U}(0, 1)$
    $I := \lfloor X \rfloor$
    if $X > b$ **then**
      $I := I + 1$
    end if
  until $A(I)p_2 \geq U\sqrt{2}\exp\left(\frac{m_B^2}{2c^2}\right)g_{m,c^2}(X)$
end if
if $I = -1$ **then**
  $I := \lceil b \rceil$
end if
return I

Algorithm 6 Data-augmented Gibbs generator of the average quantities of the convoluted molecular species Λ and the coefficients of the *intensity-ion count* dependence: C, ξ, ν^{-2}.

INPUT:

Data: $(I^w)_{w=1}^W$

Hyperparameters: $a, b, \mathtt{D}$

Initial values: $\Lambda, C, \xi, \nu^{-2}$

n = o

while n < N **do**

for $w \in \{1, \ldots, W\}$ **do**

Generate $N_{\cdot w} \sim \mathbb{P}\left(\circ \left| \Lambda^t \mathtt{D}_{|w}, \frac{I_w - \xi}{C}, \frac{\nu^2}{C^2} \right. \right)$,

end for

for $w \in \{1, \ldots, W\}$ **do**

Generate $\mathtt{N}_{|w} \sim \text{Multi}(N_{\cdot w}; \frac{\Lambda \odot \mathtt{D}_{|w}}{\Lambda^t \mathtt{D}_{|w}})$

end for

for $m \in \{1, \ldots, M\}$ **do**

Generate $\Lambda_m \sim \Gamma(a_m + \sum_{w=1}^W \mathtt{N}_{mw}, b_m + \sum_{w=1}^W D_{mw})$

end for

Generate $C \sim \mathcal{N}\left(\frac{\nu^2 + \sum_{w=1}^W N_{\cdot w}(I_w - \xi)}{\nu^2 + \sum_{w=1}^W N_{\cdot w}^2}, \frac{\nu^2}{\nu^2 + \sum_{w=1}^W N_{\cdot w}^2} \right) \Big|_{[0,\infty)}$

Generate $\xi \sim \mathcal{N}\left(\frac{\sum_{w=1}^W (I_w = CN_{\cdot w})}{\nu^2 + W}, \sigma_\xi^2 = \frac{\nu^2}{\nu^2 + W} \right)$

Generate $\nu^{-2} \sim \text{Gamma}(\frac{W}{2} - 1, 1 + \sum_{w=1^W} (I_w - \xi - CN_{\cdot w}))$

yield: Λ, C, ξ, ν^2

n++

end while

6

Conclusions and Future Research

"Hej młody Junaku
Smutek zwalcz i strach.
Przecież na tym piachu już za kilka lat
Przebiegnie, być może,
Jasna, długa, prosta,
Szeroka jak morze, Trasa Łazienkowska.
I z brzegiem zepnie drugi brzeg,
Na którym twój ojciec legł."

— Stanisław Bareja lub Stanisław Tym. Wszystko jedno – wszak wszystkie Staśki to fajne chłopaki.

In this dissertation several related topics in mass spectrometry have been approached. In particular, we have presented how to perform calculations involving fine isotopic structure with the `IsoSpec` algorithm. These results have been then applied to find products of reactions that can be found in ETD spectra, with the `MassTodon` workflow. What is more, while introducing the `MassTodon` workflow, we have also tumbled upon the problem of spectral deconvolution. The problem has been studied in detail in the `MassOn` project that, similarly to the *non-negative-regression-like* approach of `MassTodon`, is heavily relying on the `IsoSpec` algorithm. Finally, to study the ETD reactions in detail, we have shown how to use the `ETDetective` tool to estimate the reaction rates based on the results provided by the `MassTodon`.

As was surely noticed by the reader, the presented line of research was concerned more on fundamentals of the functioning of the mass spectrometry instruments, rather than on pure applications. In particular, it is true that the main applications of mass spectrometry prominently involve the use of Collision Induced Dissociation to study proteins, rather than ETD. What is more, most of the existing technology simply does not provide the necessary resolution to use the concept of the fine isotopic structure to the full. Finally, surely most of the *mass spec* community is typically not concerned with problems such as estimating the number of the observed ions. As pointed out by Alexey Nesvizhskii during the Bioinformatics for Protein Identification workshop during the 64^{th} Conference on Mass Spectrometry and Applied Topics in San Antonio,

Mass spectrometry is generally devoid of proper statistical reasonings.

So, is there any reason to actually continue this line of research? There are more and more high resolution mass spectrometers available on the market. These instruments will surely rely on data base searches and spectrum matches. Therefore, there will be an apparent need for tools such as `IsoSpec`. Even today, attempts are made by the developers of the popular `OpenMS` software to include fine structure calculators in their suite.

Bibliography

Andersen, M. S., Dahl, J. and Vandenberghe, L. (2013). Cvxopt: A python package for convex optimization, version 1.1. 6. *Available at cvxopt.org*, **54**.

Arnold, R. J., Jayasankar, N., Aggarwal, D., Tang, H. and Radivojac, P. (2006). A machine learning approach to predicting peptide fragmentation spectra. *Pac. Symp. Biocomput.*, pp. 219–230.

Ayaz-Guner, S., Zhang, J., Li, L., Walker, J. W. and Ge, Y. (2009). In vivo phosphorylation site mapping in mouse cardiac troponin i by high resolution top-down electron capture dissociation mass spectrometry: Ser22/23 are the only sites basally phosphorylated. *Biochemistry*, **48** (34), 8161–8170.

Beynon, J. H. (1960). Mass spectrometry and its applications to organic chemistry.

Blum, M., Floyd, R. W., Pratt, V., Rivest, R. L. and Tarjan, R. E. (1973). Time bounds for selection. *J. Comput. Syst. Sci.*, **7** (4), 448–461.

Böcker, S., Letzel, M. C., Lipták, Z. and Pervukhin, A. (2009). SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics*, **25** (2), 218–224.

Brand, W. A., Coplen, T. B., Vogl, J., Rosner, M. and Prohaska, T. (2014). Assessment of international reference materials for isotope-ratio analysis (IUPAC technical report). *Pure Appl. Chem.*, **86**, 425–467.

Breuker, K., Oh, H., Horn, D. M., Cerda, B. A. and McLafferty, F. W. (2002). Detailed unfolding and folding of gaseous ubiquitin ions characterized by electron capture dissociation. *J. Am. Chem. Soc.*, **124** (22), 6407–6420.

—, —, Lin, C., Carpenter, B. K. and McLafferty, F. W. (2004). Nonergodic and conformational control of the electron capture dissociation of protein cations. *Proc. Natl. Acad. Sci. U. S. A.*, **101** (39), 14011–14016.

Cai, W., Guner, H., Gregorich, Z. R., Chen, A. J., Ayaz-Guner, S., Peng, Y., Valeja, S. G., Liu, X. and Ge, Y. (2016). Mash suite pro: A comprehensive software tool for top-down proteomics. *Mol. Cell. Proteomics*, **15** (2), 703–714.

Chung, T. W. and Tureček, F. (2010). Backbone and side-chain specific dissociations of z ions from non-tryptic peptides. *J. Am. Soc. Mass Spectrom.*, **21** (8), 1279–1295.

Ciach, M. A., Łącki, M. K., Miasojedow, B., Lermyte, F., Valkenborg, D., Sobott, F. and Gambin, A. (2017). Estimation of rates of reactions triggered by electron transfer in top-down mass spectrometry. In *International Symposium on Bioinformatics Research and Applications*, Springer, pp. 96–107.

Cormen, T. (2009). *Leiserson C. Rivest R., Stein C. Introduction to Algorithms.-3rd*. MIT Press.

Cournoyer, J. J., Pittman, J. L., Ivleva, V. B., Fallows, E., Waskell, L., Costello, C. E. and O'Connor, P. B. (2005). Deamidation: Differentiation of aspartyl from isoaspartyl products in peptides by electron capture dissociation. *Protein Sci.*, **14** (2), 452–463.

Davis, C. (1954). Theory of positive linear dependence. *American Journal of Mathematics*, **76** (4), 733–746.

De Hoffmann, E. and Stroobant, V. (2007). *Mass spectrometry: principles and applications*. John Wiley & Sons.

Degroeve, S., Martens, L. and Jurisica, I. (2013). MS2PIP: A tool for MS/MS peak intensity prediction. *Bioinformatics*, **29** (24), 3199–3203.

Dittwald, P., Claesen, J., Burzykowski, T., Valkenborg, D. and Gambin, A. (2013). BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. *Anal. Chem.*, **85** (4), 1991–1994.

—, Valkenborg, D., Claesen, J., Rockwood, A. L. and Gambin, A. (2015). On the Fine Isotopic Distribution and Limits to Resolution in Mass Spectrometry. *J. Am. Soc. Mass Spectrom.*, **26** (10), 1732–1745.

Edmonds, J. and Karp, R. M. (1972). Theoretical improvements in algorithmic efficiency for network flow problems. *J. Assoc. Comput. Mach.*, **19** (2), 248–264.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Elias, J. E., Gibbons, F. D., King, O. D., Roth, F. P. and Gygi, S. P. (2004). Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature Biotechnology*, **22** (2), 214–219.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*. New York, London, Sydney: John Wiley & Sons, Inc.

Fenn, J., Mann, M., Meng, C., Wong, S. and Whitehouse, C. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science*, **246** (4926), 64–71.

Finucan, H. M. (1964). The Mode of a Multinomial Distribution. *Biometrika*, **51** (3/4), 513–517.

Fornelli, L., Damoc, E., Thomas, P. M., Kelleher, N. L., Aizikov, K., Denisov, E., Makarov, A. and Tsybin, Y. O. (2012). Analysis of intact monoclonal antibody igg1 by electron transfer dissociation orbitrap ftms. *Mol. Cell. Proteomics*, **11** (12), 1758–1767.

G. Marshall, A., T. Blakney, G., Chen, T., K. Kaiser, N., M. McKenna, A., P. Rodgers, R., M. Ruddy, B. and Xian, F. (2013). Mass Resolution and Mass Accuracy: How Much Is Enough? *Mass Spectrometry*, **2** (Spec Iss), S0009.

Gambin, A. and Kluge, B. (2010). Modeling proteolysis from mass spectrometry proteomic data. *Fundam. Informaticae*, **103** (1-4), 89–104.

Garcia, B. A., Shabanowitz, J. and Hunt, D. F. (2007). Characterization of histones and their post-translational modifications by mass spectrometry. *Curr. Opin. Chem. Biol.*, **11** (1), 66–73.

Ge, Y., Rybakova, I. N., Xu, Q. and Moss, R. L. (2009). Top-down high-resolution mass spectrometry of cardiac myosin binding protein c revealed that truncation alters protein phosphorylation state. *Proc. Natl. Acad. Sci. U. S. A.*, **106** (31), 12658–12663.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6), 721–741.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical science*, pp. 473–483.

Giannopoulos, A. A. and Milman, V. D. (2000). Concentration property on probability spaces. *Advances in Mathematics*, **156** (1), 77–106.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. CRC press.

Gillespie, D. T. (1977a). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, **81** (25), 2340–2361.

— (1977b). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, **81** (25), 2340–2361.

Göth, M., Lermyte, F., Schmitt, X. J., Warnke, S., von Helden, G., Sobott, F. and Pagel, K. (2016). Gas-phase microsolvation of ubiquitin: investigation of crown ether complexation sites using ion mobility-mass spectrometry. *Analyst*, **141** (19), 5502–5510.

Guner, H., Close, P. L., Cai, W., Zhang, H., Peng, Y., Gregorich, Z. R. and Ge, Y. (2014). Mash suite: a user-friendly and versatile software interface for high-resolution mass spectrometry data interpretation and visualization. *J. Am. Soc. Mass Spectrom.*, **25** (3), 464–470.

Hagberg, A. A., Schult, D. A. and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Pasadena, CA USA, pp. 11–15.

Håkansson, K., Cooper, H. J., Emmett, M. R., Costello, C. E., Marshall, A. G. and Nilsson, C. L. (2001). Electron capture dissociation and infrared multiphoton dissociation ms/ms of an n-glycosylated tryptic peptide to yield complementary sequence information. *Anal. Chem.*, **73** (18), 4530–4536.

Hasman, A. (1991). Probabilistic reasoning in intelligent systems: Networks of plausible inference. *International Journal of Bio-Medical Computing*, **28** (3), 221–225.

Hendrickson, C. L., Quinn, J. P., Kaiser, N. K., Smith, D. F., Blakney, G. T., Chen, T., Marshall, A. G., Weisbrod, C. R. and Beu, S. C. (2015). 21 Tesla Fourier Transform Ion Cyclotron Resonance Mass Spectrometer: A National Resource for Ultrahigh Resolution Mass Analysis. *J. Am. Soc. Mass Spectrom.*, **26** (9), 1626–1632.

Horn, D. M., Zubarev, R. A. and McLafferty, F. W. (2000). Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.*, **11** (4), 320–332.

Housecroft, C. E. and Constable, E. C. (2010). *Chemistry: An introduction to organic, inorganic and physical chemistry*. Pearson education.

Ipsen, A. (2014). Efficient calculation of exact fine structure isotope patterns via the multidimensional fourier transform. *Anal. Chem.*, **86** (11), 5316–5322.

— (2015). Derivation from First Principles of the Statistical Distribution of the Mass Peak Intensities of MS Data. *Anal. Chem.*, **87** (3), 1726–1734.

— and Ebbels, T. M. D. (2012). Prospects for a Statistical Theory of LC/TOFMS Data. *J. Am. Soc. Mass Spectrom.*, **23** (5), 779–791.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer Publishing Company, Incorporated, 1st edn.

Jackson, S. N., Dutta, S. and Woods, A. S. (2009). The use of ecd/etd to identify the site of electrostatic interaction in noncovalent complexes. *J. Am. Soc. Mass Spectrom.*, **20** (2), 176–179.

Jaitly, N., Mayampurath, A., Littlefield, K., Adkins, J. N., Anderson, G. A. and Smith, R. D. (2009). Decon2ls: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinf.*, **10** (1), 87.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning*, vol. 112. Springer.

Kallenberg, O. (1997). *Foundations of Modern Probability*. New York, Berlin, Heidelberg: Springer-Verlag.

Kaur, P. and O'Connor, P. B. (2004). Use of statistical methods for estimation of total number of charges in a mass spectrometry experiment. *Anal. Chem.*, **76** (10), 2756–2762.

Kienitz, H. (1961). Mass Spectrometry and its Applications to Organic Chemistry. *Angew. Chem.*, **73** (17-18), 634.

Kim, M.-S. and Pandey, A. (2012). Electron transfer dissociation mass spectrometry in proteomics. *Proteomics*, **12** (4-5), 530–542.

Kingman, J. F. C. (1993). *Poisson processes*. Wiley Online Library.

Koppenaal, D. W., Barinaga, C. J., Denton, M. B., Sperline, R. P., Hieftje, G. M., Schilling, G. D., Andrade, F. J., Barnes, J. H. and Iv, I. (2005). Ms detectors.

Koronacki, J., Lasota, S. and Niemiro, W. (2005). Positron emission tomography by markov chain monte carlo with auxiliary variables. *Pattern recognition*, **38** (2), 241–250.

Lawson, C. L. and Hanson, R. J. (1995). *Solving least squares problems*. SIAM.

Lermyte, F., Konijnenberg, A., Williams, J. P., Brown, J. M., Valkenborg, D. and Sobott, F. (2014). ETD allows for native surface mapping of a 150 kda noncovalent complex on a commercial q-twims-tof instrument. *J. Am. Soc. Mass Spectrom.*, **25** (3), 343–350.

—, Łącki, M. K., Valkenborg, D., Baggerman, G., Gambin, A. and Sobott, F. (2015a). Understanding reaction pathways in top-down ETD by dissecting isotope distributions: A mammoth task. *International Journal of Mass Spectrometry*.

—, Łącki, M. K., Valkenborg, D., Gambin, A. and Sobott, F. (2017). Conformational space and stability of etd charge reduction products of ubiquitin. *J. Am. Soc. Mass Spectrom.*, **28** (1), 69–76.

— and Sobott, F. (2015). Electron transfer dissociation provides higher-order structural information of native and partially unfolded protein complexes. *Proteomics*, **15** (16), 2813–2822.

—, Verschueren, T., Brown, J. M., Williams, J. P., Valkenborg, D. and Sobott, F. (2015b). Characterization of top-down ETD in a travelling-wave ion guide. *Methods*, **89**, 22–29.

—, Williams, J. P., Brown, J. M., Martin, E. M. and Sobott, F. (2015c). Extensive charge reduction and dissociation of intact protein complexes following electron transfer on a quadrupole-ion mobility-time-of-flight MS. *J. Am. Soc. Mass Spectrom.*, **26** (7), 1068–1076.

Li, L., Kresh, J. A., Karabacak, N. M., Cobb, J. S., Agar, J. N. and Hong, P. (2008). A hierarchical algorithm for calculating the isotopic fine structures of molecules. *J. Am. Soc. Mass Spectrom.*, **19** (12), 1867–1874.

—, Murat Karabacak, N., Cobb, J. S., Wang, Q., Hong, P. and Agar, J. N. (2010a). Memory-efficient calculation of the isotopic mass states of a molecule. *Rapid Commun. Mass Spectrom.*, **24** (18), 2689–2696.

Li, W., Song, C., Bailey, D. J., Tseng, G. C., Coon, J. J. and Wysocki, V. H. (2011). Statistical analysis of electron transfer dissociation pairwise fragmentation patterns. *Anal. Chem.*, **83** (24), 9540–9545.

Li, X., Lin, C. and O'Connor, P. B. (2010b). Glutamine deamidation: differentiation of glutamic acid and γ-glutamic acid in peptides by electron capture dissociation. *Anal. Chem.*, **82** (9), 3606–3615.

Łącki, M. K., Lermyte, F., Miasojedow, B., Olszański, M., Startek, M., Sobott, F., Valkenborg, D. and Gambin, A. (2017a). Assigning peaks and modeling etd in top-down mass spectrometry. *arXiv preprint arXiv:1708.00234*.

—, Startek, M., Valkenborg, D. and Gambin, A. (2017b). Isospec: Hyperfast fine structure calculator. *Anal. Chem.*, **89** (6), 3272–3277.

Loos, M., Gerber, C., Corona, F., Hollender, J. and Singer, H. (2015). Accelerated Isotope Fine Structure Calculation Using Pruned Transition Trees. *Anal. Chem.*, **87** (11), 5738–5744.

Lucy, L. B. (1974). An iterative technique for the rectification of observed distributions. *The astronomical journal*, **79**, 745.

Marty, M. T., Baldwin, A. J., Marklund, E. G., Hochberg, G. K., Benesch, J. L. and Robinson, C. V. (2015). Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. *Analytical chemistry*, **87** (8), 4370–4376.

Mayampurath, A. M., Jaitly, N., Purvine, S. O., Monroe, M. E., Auberry, K. J., Adkins, J. N. and Smith, R. D. (2008). Deconmsn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. *Bioinformatics*, **24** (7), 1021–1023.

McLuckey, S. A. and Stephenson, J. L. (1999). Ion/ion chemistry of high-mass multiply charged ions. *Mass Spectrom. Rev.*, **17** (6), 369–407.

McNaught, A. D. and Wilkinson, A. (1997). *IUPAC Gold Book*. Oxford: Blackwell Scientific Publications.

Michalski, A., Damoc, E., Lange, O., Denisov, E., Nolting, D., Muller, M., Viner, R., Schwartz, J., Remes, P., Belford, M., Dunyach, J.-J., Cox, J., Horning, S., Mann, M. and Makarov, a. (2012). Ultra High Resolution Linear Ion Trap Orbitrap Mass Spectrometer (Orbitrap Elite) Facilitates Top Down LC MS/MS and Versatile Peptide Fragmentation Modes. *Molecular & Cellular Proteomics*, **11** (3).

Mitchell Wells, J. and McLuckey, S. A. (2005). Collision-Induced dissociation (CID) of peptides and proteins. In *Methods in Enzymology*, pp. 148–185.

Morrison, L. J. and Brodbelt, J. S. (2016). Charge site assignment in native proteins by ultraviolet photodissociation (UVPD) mass spectrometry. *Analyst*, **141** (1), 166–176.

Nagao, T., Yukihira, D., Fujimura, Y., Saito, K., Takahashi, K., Miura, D. and Wariishi, H. (2014). Power of isotopic fine structure for unambiguous determination of metabolite elemental compositions: In silico evaluation and metabolomic application. *Anal. Chim. Acta*, **813**, 70–76.

Nemes, G. (2010). On the coefficients of the asymptotic expansion of n! *Journal of Integer Sequences*, **13** (2), 3.

Nikolaev, E. N., Jertz, R., Grigoryev, A. and Baykut, G. (2012). Fine structure in isotopic peak distributions measured using a dynamically harmonized fourier transform ion cyclotron resonance cell at 7 T. *Anal. Chem.*, **84** (5), 2275–2283.

Nocedal, J. (1980). Updating quasi-Newton matrices with limited storage. *Mathematics of computation*, **35** (151), 773–782.

Oh, H., Breuker, K., Sze, S. K., Ge, Y., Carpenter, B. K. and McLafferty, F. W. (2002). Secondary and tertiary structures of gaseous protein ions characterized by electron capture dissociation mass spectrometry and photofragment spectroscopy. *Proc. Natl. Acad. Sci. U. S. A.*, **99** (25), 15863–15868.

Okabe, A., Boots, B., Sugihara, K. and Chiu, S. N. (2000). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Series in Probability and Statistics, John Wiley and Sons, Inc., 2nd edn.

O'Connor, P. B., Lin, C., Cournoyer, J. J., Pittman, J. L., Belyayev, M. and Budnik, B. A. (2006). Long-lived electron capture dissociation product ions experience radical migration via hydrogen abstraction. *J. Am. Soc. Mass Spectrom.*, **17** (4), 576–585.

Posener, D. (1974). Precision in measuring resonance spectra. *Journal of Magnetic Resonance (1969)*, **14** (2), 121–128.

Richardson, W. H. (1972). Bayesian-based iterative method of image restoration. *JOSA*, **62** (1), 55–59.

Rockwood, A. L. (1995). Relationship of Fourier transforms to isotope distribution calculations. *Rapid Commun. Mass Spectrom.*, **9** (1), 103–105.

—, Van Orden, S. L. and Smith, R. D. (1996). Ultrahigh Resolution Isotope Distribution Calculations. *Rapid Commun. Mass Spectrom.*, **10** (November 1995), 54–59.

Roepstorff, P. and Fohlman, J. (1984). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom.*, **11** (11), 601.

Schnier, P. D., Gross, D. S. and Williams, E. R. (1995). On the maximum charge state and proton transfer reactivity of peptide and protein ions formed by electrospray ionization. *Journal of the American Society for Mass Spectrometry*, **6** (11), 1086–1097.

Schwudke, D., Schuhmann, K., Herzog, R., Bornstein, S. R. and Shevchenko, A. (2011). Shotgun lipidomics on high resolution mass spectrometers. *Cold Spring Harbor Perspect. Biol.*, **3** (9), 1–13.

Senko, M. W., Beu, S. C. and McLafferty, F. W. (1995). Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.*, **6** (4), 229–233.

Simons, J. (2010). Mechanisms for S-S and $N - C_\alpha$ bond cleavage in peptide ECD and ETD mass spectrometry. *Chem. Phys. Lett.*, **484** (4-6), 81–95.

Skinner, O. S., Breuker, K. and McLafferty, F. W. (2013). Charge site mass spectra: conformation-sensitive components of the electron capture dissociation spectrum of a protein. *J. Am. Soc. Mass Spectrom.*, **24** (6), 807–810.

—, McLafferty, F. W. and Breuker, K. (2012). How ubiquitin unfolds after transfer into the gas phase. *J. Am. Soc. Mass Spectrom.*, **23** (6), 1011–1014.

Slawski, M., Hussong, R., Tholey, A., Jakoby, T., Gregorius, B., Hildebrandt, A. and Hein, M. (2012). Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching. *BMC Bioinf.*, **13** (1), 291.

Snider, R. K. (2007). NIH Public Access. *J. Am. Soc. Mass Spectrom.*, **18** (8), 1511–1515.

Sohn, C. H., Chung, C. K., Yin, S., Ramachandran, P., Loo, J. A. and Beauchamp, J. L. (2009). Probing the mechanism of electron capture and electron transfer dissociation using tags with variable electron affinity. *J. Am. Chem. Soc.*, **131** (15), 5444–5459.

—, Yin, S., Peng, I., Loo, J. A. and Beauchamp, J. L. (2015). Investigation of the mechanism of electron capture and electron transfer dissociation of peptides with a covalently attached free radical hydrogen atom scavenger. *Int. J. Mass Spectrom.*, **390**, 49–55.

Startek, M. (2016). An asymptotically optimal, online algorithm for weighted random sampling with replacement. *arXiv preprint arXiv:1611.00532*.

Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. and Hunt, D. F. (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.*, **101** (26), 9528–9533.

Talagrand, M. (1996). A new look at independence. *Ann. Probab.*, **24** (1), 1–34.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, **82** (398), 528–540.

Thomson, J. J. (1913). Bakerian lecture: Rays of positive electricity. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, **89** (607), 1–20.

Tsybin, Y. O., Fornelli, L., Stoermer, C., Luebeck, M., Parra, J., Nallet, S., Wurm, F. M. and Hartmer, R. (2011). Structural analysis of intact monoclonal antibodies by electron transfer dissociation mass spectrometry. *Anal. Chem.*, **83** (23), 8919–8927.

—, He, H., Emmett, M. R., Hendrickson, C. L. and Marshall, A. G. (2007). Ion activation in electron capture dissociation to distinguish between n-terminal and c-terminal product ions. *Anal. Chem.*, **79** (20), 7596–7602.

Tureček, F. and Julian, R. R. (2013). Peptide radicals and cation radicals in the gas phase. *Chem. Rev.*, **113** (8), 6691–6733.

Tureček, F. (2003). N cα bond dissociation energies and kinetics in amide and peptide radicals. is the dissociation a non-ergodic process? *J. Am. Chem. Soc.*, **125** (19), 5954–5963.

Tureček, F. and Syrstad, E. A. (2003). Mechanism and energetics of intramolecular hydrogen transfer in amide and peptide radicals and cation-radicals. *J. Am. Chem. Soc.*, **125** (11), 3353–3369.

Valkenborg, D., Mertens, I., Lemière, F., Witters, E. and Burzykowski, T. (2012). The isotopic distribution conundrum. *Mass Spectrom. Rev.*, **31** (1), 96–109.

Van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, **10** (1), 1–50.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Wysocki, V. H., Tsaprailis, G., Smith, L. L. and Breci, L. A. (2000). Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.*, **35** (12), 1399–1406.

Xie, Y., Zhang, J., Yin, S. and Loo, J. A. (2006). Top-down esi-ecd-ft-icr mass spectrometry localizes noncovalent protein-ligand binding sites. *J. Am. Soc. Mass Spectrom.*, **128** (45), 14432–14433.

Yergey, J. a. (1983). A general approach to calculating isotopic distributions for mass spectrometry. *International Journal of Mass Spectrometry and Ion Physics*, **52** (2-3), 337–349.

Yin, S. and Loo, J. A. (2010). Elucidating the site of protein-atp binding by top-down mass spectrometry. *J. Am. Soc. Mass Spectrom.*, **21** (6), 899–907.

Zhang, H., Cui, W. and Gross, M. L. (2013). Native electrospray ionization and electron-capture dissociation for comparison of protein structure in solution and the gas phase. *Int. J. Mass Spectrom.*, **354**, 288–291.

—, —, Wen, J., Blankenship, R. E. and Gross, M. L. (2011). Native electrospray and electron-capture dissociation fticr mass spectrometry for top-down studies of protein assemblies. *Anal. Chem.*, **83** (14), 5598–5606.

Zhang, Y., Cui, W., Wecksler, A. T., Zhang, H., Molina, P., Deperalta, G. and Gross, M. L. (2016). Native ms and ecd characterization of a fab–antigen complex may facilitate crystallization for x-ray diffraction. *J. Am. Soc. Mass Spectrom.*, **27** (7), 1139–1142.

Zhang, Z. (2004). Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.*, **76** (14), 3908–3922.

— (2005). Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal. Chem.*, 77 (19), 6364–6373.

— (2010). Prediction of Electron-Transfer/Capture dissociation spectra of peptides. *Anal. Chem.*, **82** (5), 1990–2005.

—, Browne, S. J. and Vachet, R. W. (2014). Exploring salt bridge structures of gas-phase protein ions using multiple stages of electron transfer and collision induced dissociation. *J. Am. Soc. Mass Spectrom.*, **25** (4), 604–613.

Zhurov, K. O., Fornelli, L., Wodrich, M. D., Laskay, Ü. A. and Tsybin, Y. O. (2013). Principles of electron capture and transfer dissociation mass spectrometry applied to peptide and protein structure analysis. *Chem. Soc. Rev.*, **42** (12), 5014–5030.

Zubarev, R. A., Kelleher, N. L. and McLafferty, F. W. (1998). Electron capture dissociation of multiply charged protein cations. a nonergodic process. *J. Am. Chem. Soc.*, **120** (13), 3265–3266.