Estimation of Rates of Reactions Triggered by Electron Transfer in Top-Down Mass Spectrometry

Micha Aleksander Ciach[1,*], Mateusz Krzysztof cki[1,*], Baej Miasojedow[1],
Frederik Lermyte[2,3], Dirk Valkenborg[3,4], Frank Sobott[2], and Anna Gambin[1]

[1] Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Poland,
[2] Biomolecular and Analytical Mass Spectrometry Group, Dept. of Chemistry, University of Antwerp, Belgium,
[3] Centre for Proteomics, University of Antwerp, Belgium,
[4] Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Hasselt, Belgium
m_ciach@student.uw.edu.pl, mateusz.lacki@biol.uw.edu.pl

Abstract. Electron transfer dissociation (ETD) is a versatile technique used in mass spectrometry for the high-throughput characterization of proteins. It consists of several concurrent reactions triggered by the transfer of an electron from its anion source to the sample cations. Transferring an electron causes peptide backbone cleavage while leaving labile post translational modifications intact. The obtained fragmentation spectra provide valuable information for sequence and structure analysis.
Here we propose a formal mathematical model of the ETD fragmentation process in form of a system of stochastic differential equations describing its joint dynamics. Parameters of the model correspond to the rates of occurring reactions. Their estimates for various experimental settings give insight into the dynamics of the ETD process.
We estimate the model parameters from the relative quantities of the fragmentation products in a given mass spectrum by solving a nonlinear optimization problem. The cost function penalizes for the differences between the analytically derived average number of reaction products and their experimental counterparts.
The presented method proves highly robust to noise *in silico*. Moreover, the model can explain a considerable amount of experimental results for a wide range of instrumentation settings. The implementation of the presented workflow, code-named ETDetective, is freely available under 2-clause BSD license.

Keywords: Mass Spectrometry, Electron Transfer Dissociation, Markov Jump Process, BFGS, ODEs.

1 Introduction

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 ($\mathrm{N-C_\alpha}$) bonds. The sample cations are positively charged during the electrospray ionization (ESI) step (Fenn et al., 1989), leading to the formation of $[\mathrm{M+nH}]^{\mathrm{n+}}$ 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 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, non-covalent 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.

$$\begin{aligned}
\textbf{PTR}\ [\mathrm{M}+\mathrm{nH}]^{n+}+\mathrm{A}^{\bullet-} &\longrightarrow [\mathrm{M}^{+}(\text{n-1})\,\mathrm{H}]^{(n-1)+}+\mathrm{AH}\\
\textbf{ETnoD}\ [\mathrm{M}+\mathrm{nH}]^{n+}+\mathrm{A}^{\bullet-} &\longrightarrow [\mathrm{M}^{+}\mathrm{nH}]^{(n-1)+\bullet}+\mathrm{A}\\
\textbf{ETD}\ [\mathrm{M}+\mathrm{nH}]^{n+}+\mathrm{A}^{\bullet-} &\longrightarrow [\mathrm{C}^{+}\mathrm{xH}]^{x+}+[\mathrm{Z}^{+}(\text{n-x})\mathrm{H}]^{(n-x-1)+\bullet}+\mathrm{A}
\end{aligned}$$

Fig. 1: Considered chemical reactions. M stands for a precursor or a fragment ion, C and Z stand for fragment ions.

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. 2. 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).

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 $\mathrm{N-C}_\alpha$ bond cleavage.

A specific type of $\mathrm{N-C}_\alpha$ 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.

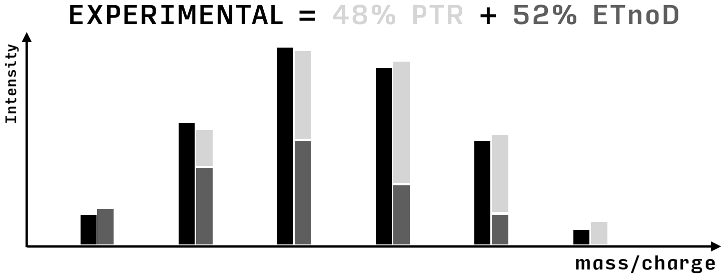

Fig. 2: The deconvolution of the observed isotopic envelopes performed by MassTodon. The observed signal (black) is represented as a combination of two theoretical isotopic patterns (gray).

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., 2017). 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 2).

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 3. ETDetective together with example data is available to download at `https://github.com/mciach/ETDetective` under the 2-clause BSD license.

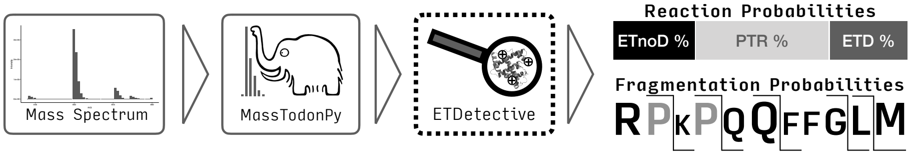

Fig. 3: The process of mass spectrum interpretation with MassTodon and ETDetective.

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. 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 paper. 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.

2 Formal model of the ETD reaction

2.1 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 5).

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:

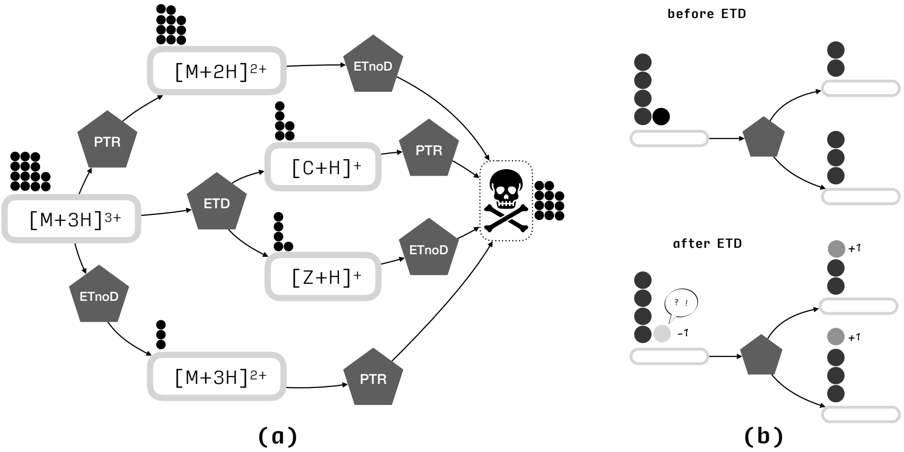

Fig. 4

Fig. 5: A model of the ETD reaction. (a) A fragment of the reaction graph for a triply charged precursor. The *molecular species* are depicted in pale grey and the *reactions* in dark grey. The 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 black circles. Only one ETD transition has been shown for clarity of the image. (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).

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 $q_u + g_u \leq q$. ■

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 reaction 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

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

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:

$$\lambda_{uu} = -Iq_u^2.$$

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 5. 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

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

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

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

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.

2.2 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 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,*

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

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*

$$\frac{\partial}{\partial t}\mathbb{E}X_u(t) = \sum_{w\colon w\to u} \lambda_{wu}\mathbb{E}X_w(t) + \lambda_{uu}\mathbb{E}X_u(t) \tag{1}$$

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

$$\begin{aligned}\frac{\partial}{\partial t}\mathrm{Cov}(X_u(t), X_v(t)) = &\sum_{w\colon w\to u} \lambda_{wu}\mathrm{Cov}(X_w(t), X_u(t)) \\ &+ \sum_{w\colon w\to v} \lambda_{wv}\mathrm{Cov}(X_w(t), X_v(t)) \\ &+ (\lambda_{uu} + \lambda_{vv})\mathrm{Cov}(X_u(t), X_v(t)) \\ &- \lambda_{uv}\mathbb{E}X_u - \lambda_{vu}\mathbb{E}X_u.\end{aligned} \tag{3}$$

Since we have defined λ_{uv} to be zero when $u \not\to v$, Equation (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 3. *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,*

$$\mathbb{E}X_u(t) = NP_R\frac{q_0^2}{q_0^2 - q_u^2}(\exp(-Iq_u^2 t) - \exp(-Iq_0^2 t)) \tag{4}$$

$$\mathrm{Var}X_u(t) = \mathbb{E}X_u(t) - (\mathbb{E}X_u(t))^2/N = N\frac{\mathbb{E}X_u(t)}{N}\Big(1 - \frac{\mathbb{E}X_u(t)}{N}\Big) \tag{5}$$

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,*

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

2.3 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 (1) is

$$\mathbb{E}X_u(t) = \sum_{i=1}^{n_u} A_i^u \exp(B_i^u t), \tag{6}$$

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 (see Lemma 5 in the appendix and following Corollaries). From Corollary 1 it follows that the coefficients for the precursor molecular species are $n_u = 1$, $A_1^r = X_r(0)$ and $B = -Iq_0^2$. The coefficients for the other molecules satisfy a recursive dependence,

$$
\begin{aligned}
n_u &= 1 + \sum_{w\colon w\to u} n_w, \\
\{(A_i^u, B_i^u) : i = 1, \dots, n_u - 1\} &= \bigcup_{j=1}^{p} \left\{ \left(A_k^{w_j} \frac{\lambda_{w_j}}{B_k^{w_j} - \lambda_{uu}}, B_{w_j}^k \right) : k = 1, \dots, n_{w_j} \right\}, \\
(A_{n_u}^u, B_{n_u}^u) &= \left(\sum_{w\colon w\to u} \sum_{i=1}^{n_w} A_i^w \frac{-\lambda_{wu}}{B_i^w - \lambda_{uu}}, \lambda_{uu} \right), \qquad (7)
\end{aligned}
$$

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, as formalized in Algorithm 1. 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, it results from the proof of Lemma 5 that 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.

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)
```

This leads to the following theorem.

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

3 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.

3.1 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, 1977), 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 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 6.

The fitting procedure turned out to be fairly robust toward a moderate noise and missing data, see Figure 6. 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).

3.2 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 investigated by computing the percentage of the experimental spectrum accounted for by the

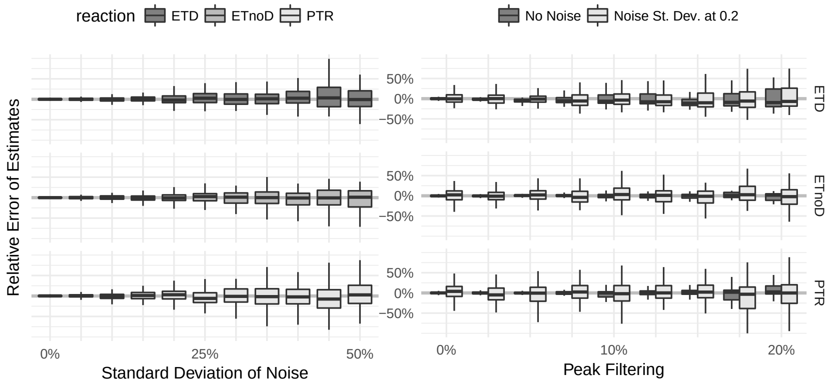

Fig. 6: Relative errors of the fitting procedure on in silico Substance P data. The known true values of parameters are respectively $P_{\text{ETD}} = 30\%, P_{\text{ETnoD}} = 25\%, P_{PTR} = 45\%$. Cleavage probabilities were assumed to be uniform (proline being the obvious exception). Each boxplot 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).

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 7: 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

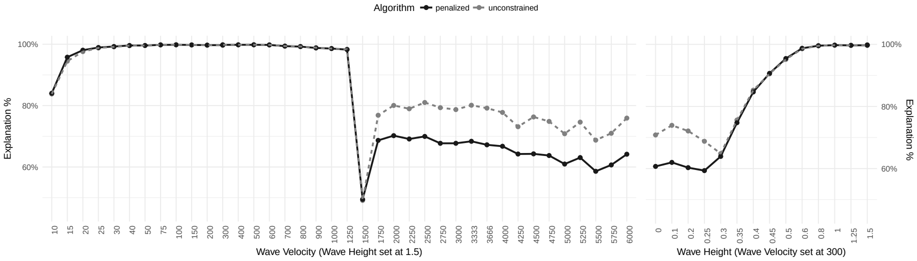

Fig. 7: Explanation Percentage (EP) for experimental Substance P spectra. Gray line: EP for model fit without decharging penalty ($\rho = 0$). Black line: EP for model fit with decharging penalty ($\rho = 1$). Left: EP for different values of Wave Velocity with Wave Height set to 1.5. Right: EP for different values of Wave Height with Wave Velocity set to 300.

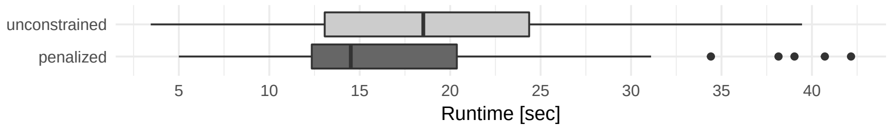

Fig. 8: The distribution of the runtime of ETDetective, both for the unconstrained ($\rho = 0$) and the penalized ($\rho = 1$) versions of the fitting procedure.

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.

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 9 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 8 show that the actual runtime of ETDetective is fairly limited on the considered Substance P results.

4 Discussion & Conclusions

In this article, 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 (Morrison 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.

Acknowledgements. This work was partially supported by the National Science Centre grants number 2013/09/B/ST6/01575, 2014/12/W/ST5/00592 and 2015/17/N/ST6/03565, the SBO grant *InSPECtor* (120025) of the Flemish agency for Innovation by Science and Technology (IWT). The authors thank the Research Foundation Flanders (FWO) for funding a Ph.D. fellowship (F.L.). The Synapt G2 mass spectrometer is funded by a grant from the Hercules Foundation Flanders.

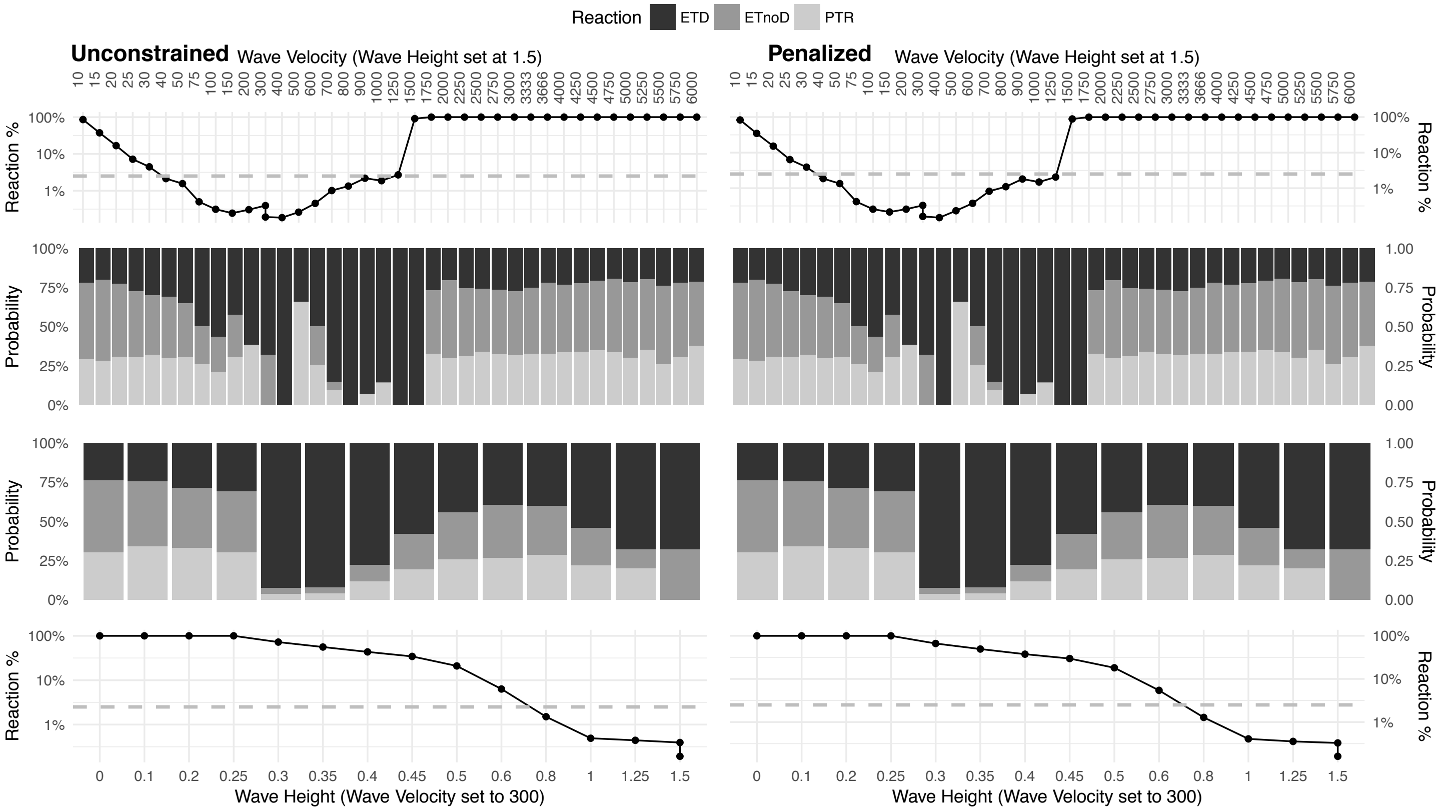

Fig. 9: Application of ETDetective to experimental data preprocessed by MassTodon. Left: Fitting with penalty for discharging ($\rho = 1$). Right: No penalty for discharging ($\rho = 0$). Top plots show results of model fit for different values of Wave Velocity with Wave Height set to 1.5; Bottom plots 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.

Bibliography

Randy J Arnold, Narmada Jayasankar, Divya Aggarwal, Haixu Tang, and Predrag Radivojac. A machine learning approach to predicting peptide fragmentation spectra. *Pac. Symp. Biocomput.*, pages 219–230, 2006.

Kathrin Breuker, Hanbin Oh, Cheng Lin, Barry K Carpenter, and Fred W McLafferty. Nonergodic and conformational control of the electron capture dissociation of protein cations. *Proc. Natl. Acad. Sci. U. S. A.*, 101(39):14011–14016, 2004.

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

Joshua E Elias, Francis D Gibbons, Oliver D King, Frederick P Roth, and Steven P Gygi. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature Biotechnology*, 22(2):214–219, jan 2004. doi: 10.1038/nbt930.

J. Fenn, M Mann, C. Meng, S. Wong, and C. Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, oct 1989. doi: 10.1126/science.2675315.

Anna Gambin and Bogusław Kluge. Modeling proteolysis from mass spectrometry proteomic data. *Fund. Inform.*, 103(1-4):89–104, 2010.

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

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

Min-Sik Kim and Akhilesh Pandey. Electron transfer dissociation mass spectrometry in proteomics. *Proteomics*, 12(4-5):530–542, 2012.

Frederik Lermyte and Frank Sobott. Electron transfer dissociation provides higher-order structural information of native and partially unfolded protein complexes. *Proteomics*, 15(16):2813–2822, jul 2015. doi: 10.1002/pmic.201400516.

Frederik Lermyte, Albert Konijnenberg, Jonathanp Williams, Jefferym Brown, Dirk Valkenborg, and Frank Sobott. 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, 2014.

Frederik Lermyte, Mateusz Krzysztof Łącki, Dirk Valkenborg, Geert Baggerman, Anna Gambin, and Frank Sobott. Understanding reaction pathways in top-down ETD by dissecting isotope distributions: A mammoth task. *Int. J. Mass Spectrom.*, 390:1–9, 2015a.

Frederik Lermyte, Tim Verschueren, Jeffery M Brown, Jonathan P Williams, Dirk Valkenborg, and Frank Sobott. Characterization of top-down ETD in a travelling-wave ion guide. *Methods*, 89:22–29, 2015b.

Frederik Lermyte, Jonathan P Williams, Jeffery M Brown, Esther M Martin, and Frank Sobott. 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, 2015c.

Frederik Lermyte, Mateusz Krzysztof Łącki, Dirk Valkenborg, Anna Gambin, and Frank Sobott. Conformational space and stability of ETD charge reduction products of ubiquitin. *J. Am. Soc. Mass Spectrom.*, 28(1):69–76, 2017.

Wenzhou Li, Chi Song, Derek J Bailey, George C Tseng, Joshua J Coon, and Vicki H Wysocki. Statistical analysis of electron transfer dissociation pairwise fragmentation patterns. *Anal. Chem.*, 83(24):9540–9545, 2011.

Mateusz K. Łącki, Frederik Lermyte, Błażej Miasojedow, Mikołaj Olszański, Michał Startek, Frank Sobott, Dirk Valkenborg, and Anna Gambin. Assigning peaks and modeling etd in top-down mass spectrometry. *arXiv preprint arXiv:1708.00234*, 2017.

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

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

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

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

Paul D Schnier, Deborah S Gross, and Evan R Williams. 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, 1995.

Jack Simons. 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, 2010.

Chang Ho Sohn, Cheol K Chung, Sheng Yin, Prasanna Ramachandran, Joseph A Loo, and J L Beauchamp. Probing the mechanism of electron capture and electron transfer dissociation using tags with variable electron affinity. *J. Am. Chem. Soc.*, 131(15):5444–5459, 2009.

Chang Ho Sohn, Sheng Yin, Ivory Peng, Joseph A Loo, and J L Beauchamp. 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, 2015.

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

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

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

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

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

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

Konstantin O Zhurov, Luca Fornelli, Matthew D Wodrich, Ünige A Laskay, and Yury O Tsybin. Principles of electron capture and transfer dissociation mass spectrometry applied to peptide and protein structure analysis. *Chem. Soc. Rev.*, 42(12):5014–5030, 2013.

Appendix

The following lemma will be used in proofs:

Lemma 4. *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 $-Iq_u^2 < -Iq_v^2$. ■

Proof of Theorem 1

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 $\exp(-Iq_0^2 t)$. ■

Proof of Proposition 1

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. ■

Proof of Theorem 2

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

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

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

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

After basic algebraic manipulations, we get

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

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

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

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

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

which proves Equation (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

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

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

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

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

$$\begin{aligned}\frac{\partial}{\partial t}\mathbb{E}X_u^2 = &\sum_{w\colon w\to u} 2\lambda_{wu}\mathbb{E}X_u(t)X_w(t) + \sum_{w\colon w\to u} \lambda_{wu}\mathbb{E}X_w(t) \\ &+ \sum_{w\colon u\to w} \lambda_{uw}\mathbb{E}X_u(t) - \sum_{w\colon u\to w} 2\lambda_{uw}\mathbb{E}X_u^2(t).\end{aligned}$$

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 (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.

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

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

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

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 (3). ■

Proof of Lemma 3

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

Derivation of general form of $\mathbb{E}X_u(t)$.

Lemma 5. *The general form of solution for system from Theorem 2 is*

$$\mathbb{E}X_u(t) = \sum_{i=1}^{n_u} A_i^u \exp(B_i^u t) \tag{8}$$

for some positive integer n_u and time-independent coefficients A_i^u, B_i^u.

Proof. Proceed by induction. For the root molecule, the Equation (8) follows from Corollary 1. Now, consider a non-precursor molecular species u, and assume that the Equation (8) is true for all molecular species v such that $v > u$. From Theorem 2, we have

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

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

$$\frac{\partial}{\partial t}\mathbb{E}X_u(t) = \sum_{w\colon w\to u} \sum_{i=1}^{n_w} \lambda_{wu} A_i^w \exp(B_i^w t) + \lambda_{uu}\mathbb{E}X_u(t). \tag{9}$$

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 9 is

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

By differentiating and substituting again into (9), 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 4). 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

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

■

From the proof of Lemma 5, we immediately get an important corollary:

Corollary 2. *Let $w_1, w_2, \dots, w_p$ be the parent molecules of a molecular species u. Then, the coefficients in Equation* (8) *satisfy a recursive dependence*

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

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

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

The next corollary follows straight from Corollary 1.

Corollary 3. *Let u be a molecular species, let r be the precursor molecular species, and let $N = X_r(0)$. Then, we have $n_r = 1$, $A_1^r = N$, and $B_1^r = -Iq_0^2$.*

Proof of Theorem 3

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 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 is $O(Lq_0^2)$, and the time complexity of the whole procedure is $O(L^2q_0^4)$.