Theory

We report in the Section the details of the four ingredients upon which the method is based: the derivation of the potential function (i.e. the probability score) for a parent sequence $P$ from the phenomenological distribution of peaks in the learning datase LSET; the costruction of a one-dimensional lattice model whose states encode all possible sequences of the given parent mass; the description of the energy function of such model, that includes both the potential function and the constraints that rule out meaningless states (i.e. states that do not correspond to a peptide sequence), and finally, the discussion of how the best sequence and probabilities profiles can be related to the thermodynamical averages of the model, and how the latter can be efficiently calculated.

---

*To whom correspondence should be addressed
†Universidad de Zaragoza
‡Current address: Institute for Scientific Interchange (ISI), Torino, Italy
A. Construction of the energy potentials from a phenomenological distribution of the peaks

This section describes the three steps that lead to the construction of the potential function: the characterization of the phenomenological distributions of peaks proceeding from different fragment species (i.e. $y$, $b$, $y^{++}$, etc.); the factorization of the likelihood of a parent sequence given a spectrum, in terms of independent fragmentation probabilities at the fragmentation sites, and the connection between the above steps to write the latter probabilities in term of the phenomenological distributions.

1. Characterization of the phenomenological distribution of the peaks

Here we describe the methods used to select the spectra of the learning dataset, to define which peaks match to a product ion, to filter out isotopic peaks, and also how we bin the mass-intensity space to characterize the phenomenological distribution, and how we define the species that are relevant for the interpretation. The final outcome will be a list of ion-products types (i.e. $y$, $b$, $y^{++}$, $y-H_2O$, etc), in order of their relevance for spectrum interpretation, and of the corresponding distributions, in the (mass,intensity)-plane, of the peaks associated to them, as obtained from the learning dataset LSET. To draw such distributions, both the fragment masses and the intensities are normalized to their biggest values in the spectrum. The resulting distributions are later used to score fragment ions according to the intensity of the experimental peak that can be associated to them.

1a. Definition of the Learning and Extended Learning Database. We collected a database of spectra interpreted with SEQUEST from the freely-available internet resource PeptideAtlas (http://www.peptideatlas.org/) by the Seattle Proteome Center:

- PAe000032 LCQ-DECA ion mass spectrometer (Thermo-Finnigan) and a micro-electrospray source (Brechbuehler);
• PAe000035\(^1\) LCQ-DECA (Thermo-Finnigan);

• PAe000142\(^2\) 2D HPLC coupled with the classical LCQ-ESI ion trap (Thermo-Finnigan);

• PAe000219\(^3\) \(\mu\)LC-ESI-MS/MS using LCQ-DECA (Thermo-Finnigan);

• PAe000244\(^3\) \(\mu\)LC-ESI-MS/MS using LCQ-DECA (Thermo-Finnigan);

We selected the most reliable according to the SEQUEST\(^4\) Xcorr and \(\Delta C_n\) scores

Table S1: Cut-off values for XCorr and \(\Delta C_n\) used to select the spectra in the learning database

<table>
<thead>
<tr>
<th>q</th>
<th>XCorr cut-off</th>
<th>(\Delta C_n) cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.5</td>
<td>0.3</td>
</tr>
<tr>
<td>2</td>
<td>3.0</td>
<td>0.3</td>
</tr>
<tr>
<td>3</td>
<td>3.5</td>
<td>0.3</td>
</tr>
</tbody>
</table>

We filtered out those spectra whose total monoisotopic mass, derived from the theoretical sequence, is different from the mass specified in the .dta file by more than 0.5 Da. This also excludes PTMs from the learning dataset. We selected at most 10 spectra associated to the same peptide, to avoid bias from overexpressed peptides. We end up with 138 spectra of charge 1, 4590 of charge 2, and 754 of charge 3, that represent our “learning database” (LSET).

The “extended learning database” (ELSET), that we will use for tests with greater statistics, is obtained with the same procedure, but without imposing the limit of at most 10 spectra for each peptide: this yields 158, 7839, and 1390 spectra of charge 1, 2, 3 respectively.

We recognized isotopic “trains” of peaks, and just kept the monoisotopic one of each train.

1b. Identification of the peaks. Normalization of the spectra  Thanks to the provided “true” parent peptide, it is possible to associate the peaks in each spectrum to the \(\{a,b,c\}\) and \(\{x,y,z\}\) series, and the corresponding analogues under neutral losses Table S2, produced at each fragmentation site. An experimental peak \(\pi_\alpha\) is identified as representing the corresponding theoretical one of the known precursor sequence if

\[
d(\rho_\alpha, \rho^T_i) \leq \epsilon = \min(\gamma_1, \gamma_2 \rho_\alpha)
\] (S1)
where \( \rho_\alpha, \rho_\alpha^T \) are the m/z ratio of the experimental and theoretical peak, and the parameters are \( \gamma_1 = 2.0 \) and \( \gamma_2 = 0.0006 \).

The above expression, that will be used also to define matching peaks in spectra interpretation, allows for an error in the peak position which increases linearly with the mass, up to the value \( \gamma_1 \). With this procedure, a peak in the spectrum matching a theoretical fragment from the precursor sequence is tagged with the labels \( s = (f, q, l) \) describing the theoretical fragment’s series \( f \), charge \( q \), and neutral-loss state \( l \). This identification is the basis to derive (independently for parent peptides of different charge) a probability distribution for each kind of ion.

Table S2: Types of neutral losses considered during the identification of experimental peaks from the known precursor spectrum in the learning dataset. The loss-prone amino-acids and the corresponding mass variations are also reported.

<table>
<thead>
<tr>
<th>( i )</th>
<th>Type</th>
<th>affected aa</th>
<th>( \Delta m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>water loss (-w)</td>
<td>S, T</td>
<td>-18.01</td>
</tr>
<tr>
<td>2</td>
<td>ammonia loss (-n)</td>
<td>Q,K,R</td>
<td>-17.03</td>
</tr>
<tr>
<td>3</td>
<td>water gain (+w)</td>
<td>H</td>
<td>+18.01</td>
</tr>
<tr>
<td>5</td>
<td>urea loss (-u)</td>
<td>R</td>
<td>-97.98</td>
</tr>
</tbody>
</table>

In order to have an homogenous description for peptide of different length, and inspired by the observations in Elias et al.\(^5\), we normalize the mass/charge coordinate each spectrum both on the precursor mass and the intensity on that of the most intense peak, reducing the distributions of all the spectra to the common region \([0 : 1] \otimes [0 : 1]\).

1c. Definition of the phenomenological distributions; identification of the relevant families

To characterize the probability distributions of the peaks of different ions emerging from the learning dataset, we bin such common region according to the Bayesian Information Criterion\(^6\) (BIC) to obtain the simplest histogram structure that still contains the important features of the original distribution.

We start by considering the different models \( A_r \) where the m/z axis is discretized in a number of parts which is an integer divisor of 12000: \( r = 1, 2, \ldots, 6000, 12000 \), while the intensity axis is discretized in 12000 parts. For each of these models, we consider each “column” of given m/z
coordinate, and look for the optimal definition of the bins, inside the columns, that yields the lower value of the quantity

\[ B(A_r) = -2 \ln \mathcal{L}(A_r) + k(A_r) \ln N_{tot} \]  

(S2)

where \( k(A_r) \) is the total number of bins, \( N_{tot} \) the total number of sample data, and \( \mathcal{L}(A_r) = \prod_{\alpha \in A_r} p(\pi \in \alpha)^{n_\alpha} \) is the likelihood, with \( p(\pi \in \alpha) \) the probability that peak \( \pi \) falls in bin \( \alpha \), with population \( n_\alpha \). In this expression, we assume that all the peaks are independent events; moreover, we assume that, within a cell, there is a uniform probability to find a peak in any position. To minimize \( B(A_r) \), we use a Simulated Annealing procedure that, at a given m/z coordinate, tries the following Monte Carlo movements: move a separation edge up or downward; join a bin with the one above or below it; split a bin in two, one on top of the other. For each model \( A_r \), such procedure produces an irregular binning (in the intensity axis), with bigger cells in regions of lower peak density. We finally choose the model \( A_r^* \) yielding the lower value of \( B(A_r) \): for singly charged precursors, the optimal value corresponds to \( r = 25 \) divisions of the m/z axis; for doubly-charged ones, \( r = 50 \), and for triply-charged, \( r = 40 \).

Once defined the binning grids for each charge state of the precursor ion, we characterize the distributions of the different product ions, for the six series together with possible neutral losses. Peaks that are not associate to any fragmentation are considered as “noise”, and contribute to the noise distribution. In this way, we can define the probability that a peak is caused by a product ion of some kind, or by noise, according to the population fraction in the corresponding bin. Analogously, we derive a probability distribution of missing peaks associated to product ions of different kind as a function of the m/z coordinate.

Not all types of product ions have the same importance in the interpretation process. Instead of focusing a-priori on \( y \) and \( b \) ions, a common choice due to their prominent abundance, we choose to determine the set of significant ions in an objective way, basing on the reduction in Shannon Entropy associated to the separation of each family.\(^5\) Namely, starting from the full distribution \( \Sigma \) of all the peaks, we identify, for each type of product ion \( s \), the set \( \Sigma_s \) of peaks associated to such
ions, and the complementary one, $\Sigma \setminus \Sigma_s$. Then we calculate the entropy reduction:

$$\Delta H(s_i) = p(\Sigma)H(\Sigma) - p(\Sigma_s)H(\Sigma_s) - p(\Sigma \setminus \Sigma_s)H(\Sigma \setminus \Sigma_s)$$  \hspace{1cm} (S3)$$

where $H(\Sigma)$ is the Shannon Entropy, $^7$ $p(X)$ is the fraction of peaks that belong to $X$. We remove from $\Sigma$ all the peaks associated to the ion $s^+$ yielding the biggest reduction, and repeat the procedure on $\Sigma \setminus \Sigma_s$. Since the entropy loss can be interpreted as information gain upon the identification of each product type, this procedure identifies which ions carry more information and are more important: remarkably, the resulting 2 most significant ions agree with the common choice used for spectra identification, even if there are differences at lower positions between differently charged precursor ions, Table S3.

Table S3: List of top-ranking product-ion types, according to the Shannon Entropy reduction criterion, along with their occurrences $N$ in the database. “-w”: water loss; “-n”: ammonia loss.

<table>
<thead>
<tr>
<th>seq</th>
<th>$Q = 1$</th>
<th>$Q = 2$</th>
<th>$Q = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fragment</td>
<td>N</td>
<td>fragment</td>
</tr>
<tr>
<td>1</td>
<td>y</td>
<td>1173</td>
<td>y</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>1003</td>
<td>b</td>
</tr>
<tr>
<td>3</td>
<td>$b - w$</td>
<td>546</td>
<td>$y^{++}$</td>
</tr>
<tr>
<td>4</td>
<td>a</td>
<td>515</td>
<td>$(y - n)^{++}$</td>
</tr>
</tbody>
</table>

2. Spectrum Interpretation as a Bayesian Inference problem

The distribution probabilities of the peaks associated to different kind of ions derived above will serve as a basis to associate a probability to a precursor peptide, but before doing so, we need to state in a precise way the statistical rules by which a spectrum is produced. We start by considering the spectrum $\Sigma$, the precursor sequence $P$ and the noise $R^\Sigma$ as three random variables, extracted from different distributions. Their joint probability $p(P, R^\Sigma, \Sigma)$ can be written as:

$$p(P, R^\Sigma, \Sigma) = p(\Sigma | P, R^\Sigma) p(P, R^\Sigma)$$  \hspace{1cm} (S4)$$

$$= p(P | \Sigma, R^\Sigma) p(\Sigma, R^\Sigma)$$  \hspace{1cm} (S5)$$
introducing the conditional probabilities \( p(\Sigma|P,R^\Sigma) \) of observing the spectrum \( \Sigma \) given the parent sequence \( P \) and the source \( R^\Sigma \), and \( p(P|\Sigma,R^\Sigma) \) of the parent sequence \( P \) given the spectrum and the noise; the latter is the quantity that the true precursor ion is expected to maximize. Assuming that the noise is independent from the precursor ion, and it is the same for all the spectra, we write
\[
p(P,R^\Sigma) = p(P)p(R), \quad p(R^\Sigma,\Sigma) \approx p(R)p(\Sigma).
\]
We observe that in any de-novo framework the actual sequence \( P \) cannot be resolved by the experiment, because of the mass degeneracies as between I and L: the actual observables are the fragmentation sites \( \mathcal{F}(P) = \{v_k, k = 1, \ldots, L-1\} \) (with \( L \) the length in residues of \( P \)) and \( \mathcal{T} = \{\mathcal{T}_k, k = 1, \ldots, L-1\} \), where \( \mathcal{T}_k = \{s_{i}, i = 1, \ldots, N^s(v)\} \) is the set of all the product ions that can be produced by fragmentations at the peptide bond located at \( v \), with \( N^s(v) \) the number of expected types of ions at peptide bond \( v \). Notice that \( N^s(v) \) and \( \mathcal{T}_v \) depend in general on the fragmentation site \( v \), since they are related to the charge and neutral losses that the residue terminating at \( v \) can bear.

On the other hand, once given \( P \), the sets \( \mathcal{F}(P) \) and \( \mathcal{T}(P) \) are completely determined. Therefore we can write, for any random variable \( Y \):
\[
p(P|Y) \equiv p(P,\mathcal{T}(P)|Y) = p(P|\mathcal{T}(P),Y)p(\mathcal{T}(P)|Y)
\]
\[\text{(S6)}\]
This is true in particular for
\[
p(P|R,\Sigma) = p(P|\mathcal{T}(P))p(\mathcal{T}(P)|R,\Sigma).
\]
\[\text{(S7)}\]
where we have used the fact that \( p(P|\mathcal{T}(P),\Sigma,R) = p(P|\mathcal{T}(P)), \) since no extra information about \( P \), with respect to the one contained in \( \mathcal{T}(P) \), is carried by \( \Sigma \) or \( R \).

Substituting in Eq. (S5) finally yields:
\[
p(\mathcal{T}(P)|R,\Sigma) = p(\Sigma|P,R)\frac{p(\mathcal{T}(P))}{p(\Sigma)}
\]
\[\text{(S8)}\]
expressing Bayes theorem. Interpreting the spectrum will be equivalent to finding the peptide
sequence $P'$ that maximize probability Eq. (S8):

$$P' = \arg \max_p P(\mathcal{T}(P)|R,\Sigma) = \arg \max_p [P(\mathcal{T}(P))],$$  \hspace{1cm}  \text{(S9)}$$

where we neglect the denominator, which is the same in all the maximization process. Such $P'$ is the best prediction we can give of the true precursor ion $P^*$. Notice that Eq. (S8) and Eq. (S9) are the correct expressions of Eqs. (1,2) in the main text, when we consider the degeneracy $I \leftrightarrow L$: we omitted this detail in the main text for the sake of simplicity since neither the method nor the results are affected in any way by the substitution.

We propose for $p(\mathcal{T}(P))$ a simple expression, based on the assumption of equal probability of generating any theoretical product ion, and complete statistical independence of such generation:

$$p(\mathcal{T}(P)) = \prod_{\nu \in \mathcal{F}(P)} \prod_{s \in \nu} p_{\text{ion}}$$  \hspace{1cm}  \text{(S10)}$$

where $p_{\text{ion}}$ is a constant for all fragmentation sites $\nu$ and species $s$.

To estimate the probability $p(\Sigma|P,R)$ in terms of the phenomenological distributions we have derived from the database, we introduce $\Theta_\nu$, a subset of $\mathcal{T}_\nu$, as the ions that have been actually generated by fragmentation at $\nu$, and $\Theta = \{\Theta_k, k = 1, \ldots, L-1\}$, the corresponding collection of ions generated at any fragmentation position of $P$. Let $\Theta^M$ be the maximal subset of fragments in $\Theta$ that are matched by an experimental peak, and $\Sigma^M$ the corresponding subset of the spectrum $\Sigma$ matching $\Theta^M$. We resort to a number of assumptions:

1. an experimental peak that matches a product-ion $s$ of the proposed precursor peptide $P$, is indeed the “image” of $s$ alone. The possibility that multiple fragments of the same parent $P$ match the same peak are neglected, as well as the possibility that both a product ion and the noise, or just the latter, may contribute to the observed intensity of a peak.

2. If a product-ion $s$ of $P$ matches more than one experimental peak in $\Sigma$, then the true image of $s$ will be the peak $\pi$ associated to the highest probability.
3. All experimental peaks that do not correspond to any possible product-ion of $P$ are due to noise.

4. Peaks due to noise are independent random events, so that

$$p(R : \Sigma \setminus \Sigma^M | R) = \prod_{\pi \in \Sigma \setminus \Sigma^M} p(R : \pi | R)$$  
(S11)

where $p(R : \pi | R)$ is the probability that, given the noise source $R$, peak $\pi$ was produced by noise.

5. For a given $P$, probabilities of association between theoretical and experimental peaks at different fragmentation sites $\nu_k \in \mathcal{F}(P)$ of $P$ are independent and can be factorized.\(^8\) Also, within a given choice of $\Theta_{\nu_k} \subseteq \mathcal{T}_{\nu_k}(P)$ produced in $\nu_k$, the probabilities are independent and can be factorized on $s_i(\nu)$.

With the above assumptions, $p(\Sigma|P,R)$ becomes:

$$p(\Sigma|P,R) = \sum_{\Theta \subseteq \mathcal{F}(P)} p(R : \Sigma \setminus \Sigma^M | R)p(\Theta^M : \Sigma^M | P)p(\Theta \setminus \Theta^M | P)$$  
(S12)

which represents the sum, over the sets of fragment ion actually generated, of the product of three independent terms, related to the probability that the noise has produced all the unmatched peaks, the probability that the fragment ions $\Theta^M$ have produced the matched peaks with the experimental intensities and positions, and finally the probability that the set of fragments $\Theta \setminus \Theta^M$ has produced no peak in the spectrum. Multiplying and dividing by $p(R : \Sigma^M | R)$, together with assumption 4 above, allows us to collect out the probability $p(R : \Sigma | R)$ that the whole spectrum is due to noise. Moreover, assumption 5 above allows us to factorize the second and third term on the fragmentation sites of $P$, finally yielding:

$$p(\Sigma|P,R) = p(R : \Sigma) \prod_{\nu \in \mathcal{F}(P)} \sum_{\Theta_{\nu} \subseteq \mathcal{T}_{\nu}(P)} \prod_{s_i \in \Theta_{\nu}} \left( \delta I(s_i^\nu) w^M(s_i^\nu, I(s_i^\nu)) + (1 - \delta I(s_i^\nu)) w^M(s_i^\nu) \right)$$  
(S13)
where $I(s^Y_i)$ is the experimental peak matching $s^Y_i$ (if it exists, otherwise $I(s^Y_i) \equiv 0$; $\delta_{I(s^Y_i)} = 1$ if there is an experimental peak sufficiently close to the position of the theoretical fragment $s^Y_i$, and $\delta_{I(s^Y_i)} = 0$ otherwise;

$$w^M(s^Y_i, \mathcal{F}(s^Y_i)) = \frac{p^{\text{loc}}(s^Y_i : I(s^Y_i)|P)}{p(R : I(s^Y_i)|R)} \tag{S14}$$

$$w^M(s^Y_i) = p^{\text{loc}}(s^Y_i : \emptyset|P) \tag{S15}$$

$$p(R : \Sigma) = \prod_{\pi \in \Sigma} p(R : \pi|R) \tag{S16}$$

where $p^{\text{loc}}(*)$ indicates that the probability is normalized in a small neighborhood of the theoretical position $\rho(s^Y_i)$, according to the fact that, once determined the precursor peptide $P$, the fragmentation sites $\nu \in \mathcal{F}(P)$ and the expected ions $s^Y_i$ are completely fixed.

Introducing Eq. (S13) and Eq. (S10) in Eq. (S8), and dropping the term $p(R : \Sigma)$ that does not depend on $P$, and introducing the dynamical variable $\xi^{s_i}_\nu$ that indicates if there fragmentation of the peptide in $\nu$ produced the ion $s_i$ or not:

$$\xi^{s_i}_\nu = \begin{cases} 
1 & \text{if } s_i \in \Theta_{\nu} \\
0 & \text{if } s_i \in \mathcal{F}_\nu \setminus \Theta_{\nu}
\end{cases} \tag{S17}$$

we write equation Eq. (S8) as:

$$p(\mathcal{F}(P)|\Sigma,R) \propto \prod_{\nu \in \mathcal{F}(P)} \prod_{i=1}^{N^\nu} \left[ \sum_{\xi^{s_i}_\nu=0,1} p^{\text{ion}} \left( \delta_{\xi^{s_i}_\nu,0} + \delta_{\xi^{s_i}_\nu,1} \left( \delta_{I(s^Y_i)} w^M(s^Y_i,I(s^Y_i)) + (1 - \delta_{I(s^Y_i)}) w^M(s^Y_i) \right) \right) \right] \tag{S18}$$

where $N^\nu$ is the number of expected types of ions at peptide bond $\nu$. Using the fact that the “deltas” above just take the values 0,1, it is easy to convince oneself that the above expression can
be recast as in Eq.(5) in the main text:

\[ p(\mathcal{T}(P)|\Sigma, R) \propto \prod_{v \in \mathcal{T}(P)} \left[ \prod_{i=1}^{N_v(v)} e^{-H_v(s_i^v, \xi_i^v)} \right] , \]  

(S19)

where we have introduced:

\[ H_v(s_i^v, \xi_i^v) = \mu - \delta_{\xi_i^v, 1} \left( \delta_{I(s_i^v)} \log(w^M(s_i^v, I(s_i^v))) + (1 - \delta_{I(s_i^v)}) \log(w^{\overline{M}}(s_i^v)) \right) , \]

(S20)

defining the chemical potential as \( \mu = -\log p_{\text{ion}} \). The role of the latter is to penalize the system against matching the experimental spectrum with too many fragment ions, thus compensating a bias towards sequences with many light residues. Its actual value cannot be learnt from experimental distributions, and we fix it \textit{a posteriori}, optimizing the agreement between the predicted peptide and the “true” parent, as predicted by SEQUEST, for the spectra in the learning database. Eq. (S20) coincides with Eq. 6 in the main text, with the identification:

\[ h^M_v(s_i, I(s_i)) = -\log(w^M(s_i^v, I(s_i^v))) = -\log \frac{p_v(s_i^v : I(s_i^v)|P)}{p(R : I(s_i)|R)} \]  

(S21)

\[ h^{\overline{M}}_v(s_i) = -\log(w^{\overline{M}}(s_i^v)) = -\log p_v(s_i^v : \emptyset|P) \]  

(S22)

Notice that in the derivation of Eq. (S19) and in its final expression, the \( v \)’s represent the fragmentation sites of the parent peptide \( P \) (i.e., their peptide bonds), and are not yet identified with the discrete sites in the mass array. Such identification will be made in Eq. (S32), when presenting the energy function of the lattice model.

3. Characterization of \( p(\mathcal{T}(P)|\Sigma, R) \) in terms of the phenomenological distributions.

To complete the description of the potentials \( H_v(s_i^v, \xi_i^v) \) introduced in the previous section, we have to relate them with the phenomenological distribution of the peaks derived in Footnote ‡. The quantities in Eq. (S16) can be calculated from the distributions of the learning dataset with the
Calling $B(\pi)$ the bin containing peak $\pi$, here $x(\pi)$ represents a rectangular subregion of $B(\pi)$ of sides $(2\varepsilon, h(B(\pi)))$, with $h(B(\pi))$ the height of $B(\pi)$, i.e. its length in the intensity axis. $\mathcal{I}_\varepsilon(\rho)$ represents the region $[\rho - \varepsilon, \rho + \varepsilon] \otimes [0, 1]$, including peaks of all intensities in an interval around $\rho$. $N(R)$ is the total number of noise peaks, $N(s,y)$ represents the numbers of peaks matched to product ions of type $s$ inside the region $y$, calculated assuming uniform probabilities within bins (for instance, $N(s,x(\pi)) = N(s,B(\pi))A(x(\pi))/A(B(\pi))$, with $A$ the area). $N^T(t, \mathcal{I}_\varepsilon(\rho(t)))$ represents the total number of expected fragments of type $t$ in the region $\mathcal{I}_\varepsilon(\rho(t))$: $N^T(t, \mathcal{I}_\varepsilon(\rho(t))) = N(\bar{t}, \mathcal{I}_\varepsilon(\rho(t))) + N(t, \mathcal{I}_\varepsilon(\rho(t)))$. Finally, $N(\bar{t}, \mathcal{I}_\varepsilon(\rho(t)))$ represents the number of missing peaks that should be associated to fragment ions of type $t$ in the region $\mathcal{I}_\varepsilon(\rho(t))$.

**B. Definition of the physical model**

In this section we describe how the MS/MS interpretation problem can be mapped to a 1-dimensional, on-lattice physical model, discussing on one hand the choice of the site variables that describe the state of the model, and the resulting configuration space, and on the other the energy function that determines the equilibrium probability population of each state.

**1. Mapping the sequence space on a 1D lattice model**

**1a. Definition of the 1-D lattice and of the representation of the residues.** We define a discrete one-dimensional lattice of $M + 1$ sites, numbered from 0 to $M$, where $M$ is the discretized mass
of the parent peptide (more precisely, it is the sum of the masses of the residues that compose it, neglecting the extra groups H and OH at the N- and C- terms, and neglecting the extra proton of the precursor MH\(^+\)). \(M\) is estimated from that provided in the .dta file as

\[
M = \left[ \frac{1}{\eta} \left( M^{\text{dta}} - m(\text{N-term}) - m(\text{C-term}) - m(H) \right) \right] \tag{S26}
\]

where \([x]\) indicates the closest integer to \(x\), and the lattice spacing \(\eta\) is derived from the learning dataset LSET as:

\[
\eta = \left\langle \frac{1}{M} \left( M^{\text{dta}} - m(\text{N-term}) - m(\text{C-term}) - m(H) \right) \right\rangle_{\text{LSET}} = 1.0011312210 \tag{S27}
\]

\(\eta\) represents the average, over the “learning dataset” LSET, of the ratio between the parent-peptide’s mass provided in the “dta” file, \(M^{\text{dta}}\), stripped of the N- and C- terminal masses as well of the extra proton, and the total discrete mass, \(M = \sum_i m(a_i)\), calculated from the sequence of the parent peptide, that comes with the database.

Table S4 lists all the residues with their discretized mass. For the moment we ignore post-translational modification, however, the algorithm is still applicable when PTMs are present, by simply augmenting Table S4 with the corresponding modified residues masses and corresponding properties.

1b. Dynamical Variables and constraints. Any sequence of total mass \(M\) will be identified, in the lattice, by the terminal points of each residue, that will coincide (thanks to the mass discretization) with different lattice sites, situated at a distance corresponding to the residue’s mass. We map the possible amino-acid sequences to the model configurations by introducing, for each site, a variable \(r \in [0, r_{\text{max}}]\), where \(r_{\text{max}}\) is the biggest residue mass (if post-translational modifications are included, it will correspond to the heaviest species, be it a wild-type or a modified residue). We implement the following rule as a constraint in the configuration space: the only values allowed at \(v\) are \(r_v = r_{v-1} + 1\) or \(r_v = 0\), the latter just holding when \(r_{v-1} = m(a) - 1\), for some amino-acid
Table S4: List of residues with the corresponding discrete mass $m_a$ as a multiple of $\eta$, along with the residue natural frequency $f(a)$ (in percentage). The capacity of each residue to accept a charge $q$ or to be modified by a neutral group (water loss ($l_1$); ammonia loss ($l_2$); water gain ($l_3$); urea loss ($l_4$) is reported in the following columns (0 for “not allowed”; 1 for “allowed”). Globally $m_a$, $q$ and $l_i$ represent the characteristic numbers $\omega(a)$, characterizing residue $a \in \mathcal{A}$: from their knowledge, it is possible to identify the residue uniquely, apart from the degeneracy Ile-Leu.

<table>
<thead>
<tr>
<th>$a$</th>
<th>$m_a$</th>
<th>$f(a)$</th>
<th>$q$</th>
<th>$l_1$</th>
<th>$l_2$</th>
<th>$l_3$</th>
<th>$l_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>57</td>
<td>7.49</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>71</td>
<td>5.22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S</td>
<td>87</td>
<td>4.53</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P</td>
<td>97</td>
<td>5.22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>V</td>
<td>99</td>
<td>1.82</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>101</td>
<td>6.26</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>103</td>
<td>4.11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>I</td>
<td>113</td>
<td>7.10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L</td>
<td>113</td>
<td>2.23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N</td>
<td>114</td>
<td>5.45</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>115</td>
<td>9.06</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Q</td>
<td>128</td>
<td>5.82</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>K</td>
<td>128</td>
<td>3.25</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>129</td>
<td>2.27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>131</td>
<td>3.91</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>137</td>
<td>5.12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>147</td>
<td>7.34</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>R</td>
<td>156</td>
<td>6.48</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Y</td>
<td>163</td>
<td>5.96</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>W</td>
<td>186</td>
<td>1.32</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
a from the list of chemical species allowed as component of the parent peptide.

Given a sequence $S$ of the correct total mass, the system configuration that describes it will be characterized by having $r_\nu = 0$ at the sites $\nu$ that correspond to the possible fragmentation sites of the sequence $S$. We do not consider different fragmentation sites for different species $a,b,c,x,y,z$, etc. that can be produced during CID. Instead, we consider as fragmentation sites the “end” of each residue of $S$ (i.e. the sites $\nu$ corresponding to the sum of the mass of all residues of each N-terminal subsequence of $S$). All the relevant fragment ions originating from each of those fragmentation site are considered, and the corresponding peaks are looked for in the experimental spectrum.

The fragment ions that can be generated by a fragmentation site $\nu$, and the position $m/z$ of the corresponding peaks, depend not only on the mass, but also on the presence of residues, N-terminal or C-terminal to the cut, that can get charged, or undergo neutral losses. Again, just the total number of charges and neutral losses N- or C-terminal to the fragmentation site, and not the details fo the sequence, are needed to determine the position of the peaks. For this reason, we introduce at each site $\nu$ two variables $q^N_\nu$, $q^C_\nu$, representing the maximal “intrinsic” charge that can be found in the subsequence N-terminal or C-terminal to $\nu$, respectively, according to the parent peptide’s charge and the number of basic residues in such subsequences. Namely,

$$q^X_\nu = \min(Q - 1, n_X)$$  \hspace{1cm} (S28)

($X = N, C$), with $Q$ the total charge of the parent peptide as detected by the first mass spectrometer and $n_N$, $n_C$ the number of basic residues K, R, H, contained in the N- or C- fragment, respectively. In practice, $q^X_\nu \in [0, Q - 1]$ counts the number of basic residues $X$-terminal to $\nu$, so that its maximal value is $Q - 1$ (since a unit charge is always present from the extra proton attached to the parent peptide). In the fragmentation process, $q^X_\nu$ imposes an upper limit on the number of charges species that can be generated: for instance, if $q^N_\nu = 0$ and $q^C_\nu = 2$ at some $\nu$ for a triply charged spectrum, then the N-fragments generated at $\nu$ can just have charge $+1$, while the C-fragments can have charge in the range $[1, 3]$, so that the algorithm will look for a match for the all the above fragment
ions in the experimental spectrum.

The constraints on the charge variables are a little more complicated than those for the mass variable \( r_\nu \), but again they can be written in terms of the variables at sites \((\nu - 1)\) and \( \nu \) in the following way:

- If \( r_\nu \neq 0 \), that is, if \( \nu \) does not correspond to a peptide bond (a fragmentation sites), or if \( r_\nu = 0 \) but the residue \( a \) ending in \( \nu \) is not H,K or R, then \( q^X_\nu = q^X_{\nu - 1} \) for both \( X = N,C \).

- If \( r_\nu = 0 \) and the residue \( a \), N-terminal to \( \nu \), is H,K or R, then

\[
q^N_\nu = \begin{cases} 
q^N_{\nu - 1} + 1 & \text{if } q^N_{\nu - 1} < Q - 1 \\
q^N_{\nu - 1} & \text{if } q^N_{\nu - 1} = Q - 1 
\end{cases} \tag{S29}
\]

\[
q^C_\nu = \begin{cases} 
q^C_{\nu - 1} - 1 & \text{if } q^C_{\nu - 1} \leq Q - 1 \\
q^C_{\nu - 1} & \text{if } q^C_{\nu - 1} = Q - 1 
\end{cases} \tag{S30}
\]

Notice the difference in the behavior of \( q^N_\nu \) and \( q^C_\nu \): the latter “forks” when \( q^C_{\nu - 1} \leq Q - 1 \), since the number of basic residues C-terminal to the current position is unknown, due to the fact that the number of basic residues of the parent sequence may exceed \( Q - 1 \) (but obviously, just \( Q - 1 \) of them can be charged). The above rule guarantees that we can set \( q^N_0 = 0 \) and \( q^C_M = 0 \), independently from the total number of basic residues.

The possible fragments generated in a dissociation at a certain position \( \nu \) also depend on the number of residues that can present neutral losses (such as loss of water, ammonia, phosphate groups, etc.: the most common cases are reported in Table S4). This results in a different mass of the corresponding peak. To carry this information, we introduce two vectors of integers \( l^X_\nu = \{l^X_{\nu,1}, \ldots, l^X_{\nu,L}\} \), \((X = \{N,C\})\), at each site \( \nu \), specifying the maximal number of neutral losses of each kind \( l^X_{\nu,\alpha} \) that can be seen in X-terminal fragments. Here \( L \) is the number of neutral losses type considered: we choose \( L = 2 \), (corresponding to \( H_2O \) and \( NH_3 \) loss, see S.I.) The constraints between the values of these vectors at neighboring sites are similar to those for the charge, with the
difference that the maximal number of neutral losses of each kind to consider, $L_\alpha (\alpha = 1, \ldots , \mathcal{L})$ is not known a priori (while the total charge of the precursor peptide is measured by the first Mass Spectrometer, and hence an input of the interpretation algorithm). We fix $L_\alpha = 1$ according to the findings reported in Table S3.

Again, the idea is that such constraints do not limit the total number of residues that can lose neutral species, but provide a way to rule out fragment ions at $\nu$ corresponding to neutral losses that are incompatible with the sequence composition. For instance, fixing a maximal total number $L_\alpha$ for neutral losses of type $\alpha$ means that in every case we will look, at most, for fragments with $L_\alpha$ losses of that kind, even if the state of the system (i.e., the composition of the sequence) would allow for more. However, if at a certain $\nu$ we have $l_{\nu,\alpha}^N < L_\alpha$, we will just generate N-terminal fragment ions at $\nu$ with at most $l_{\nu,\alpha}^N$ losses of type $\alpha$, since that means that the composition of the sequence N-terminal to $\nu$ does not allow for more.

In detail, the constraint are the following:

- If $r_\nu \neq 0$ or if $r_\nu = 0$ but the residue $a$ ending in $\nu$ does not allow a neutral loss of type $\alpha$, then $l_{\nu,\alpha}^X = l_{\nu-1,\alpha}^X$ for both $X = N, C$ and for all $\alpha$.

- If $r_\nu = 0$ and the residue $a$, N-terminal to $\nu$, allows a neutral loss of type $\alpha$, then

$$l_{\nu,\alpha}^N = \begin{cases} l_{\nu-1,\alpha}^N + 1 & \text{if } l_{\nu-1,\alpha}^N < L_\alpha \\ l_{\nu-1,\alpha}^N & \text{if } l_{\nu-1,\alpha}^N = L_\alpha \end{cases}$$  \hspace{1cm} (S31)

$$l_{\nu,\alpha}^C = \begin{cases} l_{\nu-1,\alpha}^C - 1 & \text{if } l_{\nu-1,\alpha}^C \leq L_\alpha \\ l_{\nu-1,\alpha}^C & \text{if } l_{\nu-1,\alpha}^C = L_\alpha \end{cases}$$  \hspace{1cm} (S32)

Another important information to carry at each site is related to characteristics of enzymatic digestion. In particular, we specify ourselves to trypsin rules, which is the most commonly used enzyme for digestion: if a different enzyme is used, the rules should be changed accordingly. Trypsin cleaves the protein at the carboxyl side of the residues Lysine (K) and Arginine (R), but the cleavage is inhibited if the following residue is either a Proline (P) or another K or R. For
simplicity we start considering as cleavage inhibitor only the Proline. Therefore, we distinguish between three types of residue: the cleaving residues (K and R), the residue that prevent a previous cleavage (P), and the rest of residues. We introduce a binary variable, $\pi_\nu$, that specifies the nature of the previous residue and force the following residue to respect the tryptic rules: a value $\pi_\nu = 1$ means that the previously added residue was a cleaving one, so that a P is expected at the next $\nu$ where $r_\nu = 0$, unless it is the end of the chain. Without this binary flag, sequences containing K and R at arbitrary positions would be considered in the configuration space, enlarging the allowed state space. Actually, this constraint can be made softer than the rest, since it is known that sometimes trypsin fails resulting in one (or at most, a few) missing cleavage sites, signalled by some isolated K or R.

The following rules implement the correct constraints on $\pi_\nu$:

- if $r_\nu \neq 0$, $\pi_\nu = \pi_{\nu-1}$;
- if $r_\nu = 0$ and $a$ is the species of the residue ending at $\nu$, then
  - if $\pi_{\nu-1} = 0$ and $a$ is not K or R, then $\pi_\nu = 0$,
  - if $\pi_{\nu-1} = 0$ and $a$ is K or R, then $\pi_\nu = 1$,
  - if $\pi_{\nu-1} = 1$ and $a$ is P, then $\pi_\nu = 0$.

Every other combination is forbidden.

The above defined variables provide all the ingredients to calculate, at every $\nu$ of the lattice, the positions in the spectra of all possible fragment ions that could be generated at $\nu$ by the current system configuration. However, it is not known if all the possible fragments at $\nu$ will be effectively generated in the fragmentation, and which subsets will be detected: this would require an accurate knowledge of the physics of the CID process, to find out the correct distribution probabilities and correlations in the fragmentation pattern at each site. To circumvent this problem, we introduce a binary fragmentation variable $\xi_\nu = 0, 1$ that accounts for the event of ion formation, taking the value 1 if the corresponding ion species $s_i$ is produced in the CID, and 0 otherwise. The
constraints on these variables just depend on the state variables at ν, and are non zero if and only if \( r_ν = 0 \) and the ion species considered are compatible with the local charge and neutral losses states (i.e. \( \xi^{v \leftrightarrow \leftrightarrow} \) can be 1 only if \( q^C_ν \geq 1 \), analogously \( \xi^{b-NH_3}_v \) can take value 1 if \( l^N_{ν,NH_3} \geq 1 \)). Notice that at \( ν = 0, M \) fragmentation is considered as impossible, so that \( T(σ_0) = T(σ_M) = 0 \) and \( ξ_0, ξ_M \) are all zero. We collect all the state-variables, characterized above, in a global one, \( σ_ν ≡ \{ r_ν, q^N_ν, q^C_ν, l^N_ν, l^C_ν, π_ν, ξ_ν \} \), that will contain all the relevant information needed at site ν to yield the correct fragmentation.

1c. **Boundaries Conditions.** At the N-terminal (ν = 0) and at the C-terminal (ν = M), the variables \( σ_0 \) and \( σ_M \) are subjected to special constraints. At the N-term, the first residue starts at ν = 0 with \( r_0 = 0 \); there is no possible N-terminal charge or neutral loss, so that \( q^N_0 = 0, l^N_0 = 0 \), while C-terminal ions can express with the maximum of the charge, \( q^C_0 = Q - 1 \), and, as there is no strict a-priori constraint on the number of neutral losses of each kind \( α \) (apart from the maximal value \( L_α \) that we decide to consider), all possible values between 0 and the maximum are allowed: \( l^C_0, α \in [0, L_α] \) for every \( α \) in \([1, L]\). Moreover, all residue types are admitted so that \( π_0 = 0, 1 \);

At the C-term, we have \( r_M = 0 \) and \( q^C_M = 0 \) and \( l^C_M = 0 \), since there is no C-terminal fragment, and associated charge or neutral loss. For the N-terminal ions we have \( q^N_M = Q - 1 \) and \( l^N_M, α \in [0, L_α] \) for every \( α \) in \([1, L]\). Moreover, we strictly impose the trypsin rule on the terminal residue, asking that it is a “cleaving residue” (K or R, for trypsin).

2. **The Energy Function**

The definition of the state variables, in the previous section, was meant to carry, at each site, all the information on the relevant factors on which the fragmentation pattern and the association between fragments ions and peaks depend. This allows us to introduce an energy function (Eq. (4) in the main text), to “score” model configurations \( Φ = \{ σ_ν, ν = 0, . . . , M \} \), that just depends on local
(on-site and next-neighbours) interactions:

\[ H(\phi, \Sigma) = \sum_{\nu=1}^{M-1} H_{\nu-1,\nu}(\sigma_{\nu-1}, \sigma_{\nu}), \quad \text{(S33)} \]

where

\[ H_{\nu-1,\nu}(\sigma_{\nu-1}, \sigma_{\nu}) = H_{\nu}^1(\sigma_{\nu}; \Sigma) + H_{\nu-1,\nu}^2(\sigma_{\nu-1}, \sigma_{\nu}). \quad \text{(S34)} \]

The equilibrium probability of any model configuration will be given by the Boltzmann probability:

\[ p(\phi|\Sigma, T) = Z^{-1} e^{-\beta H(\phi, \Sigma)} \quad \text{(S35)} \]

where \( \beta = 1/T \). The partition function \( Z = \sum_{\phi} e^{-\beta H(\phi, \Sigma)} \) involves a sum over all the states of the model, and thus over all the sequences with total mass and charge as the parent peptide.

In the expression of the energy above, the second part \( H^2 \) represents all the constraints that rule out meaningless model states (as those corresponding to distances, between fragmentation sites, that do not match any residue’s mass), or enforce the correct enzymatic rules. \( H_{\nu-1,\nu}^2(\sigma_{\nu-1}, \sigma_{\nu}) \) can take just two possible values: zero if the constraint is satisfied, and infinity if it is not:

\[ H_{\nu-1,\nu}^2 = \begin{cases} 
0 & \text{if } \Delta_{\nu-1,\nu}^0 = 1, \text{ or if } \exists a \in \mathcal{A} \text{ such that } \Delta_{\nu-1,\nu}^a = 1 \\
\infty & \text{otherwise} 
\end{cases} \quad \text{(S36)} \]

where we have introduced the binary function \( \Delta_{\nu-1,\nu}^0 = 0, 1 \) as a shortcut for the constraints:

\[ \Delta_{\nu-1,\nu}^0 = \begin{cases} 
1 & \text{if } r_{\nu} = r_{\nu-1} + 1, \ q_{\nu}^X = q_{\nu-1}^X, \ l_{\nu}^X = l_{\nu-1}^X, \ \pi_{\nu} = \pi_{\nu-1} \\
0 & \text{otherwise} 
\end{cases} \quad \text{(S37)} \]

\[ \Delta_{\nu-1,\nu}^a = \begin{cases} 
1 & \text{if } r_{\nu} = 0, \ r_{\nu-1} = m_a - 1, \ q_{\nu}^X = q_{\nu-1}^X + \delta_q^X(a), \ l_{\nu}^X = l_{\nu-1}^X + \delta_l^X(a), \ \pi_{\nu} = \pi_{\nu-1} + \delta_{\pi}^X(a) \\
0 & \text{otherwise} 
\end{cases} \quad \text{(S38)} \]
for both \( X = \{ N, C \} \), where \( \delta_q^X(a) \), \( \delta_l^X(a) \) and \( \delta(\sigma) \) take the values \(-1, 0, 1\) and represent the changes associated to the constraints explained above and the values reported in Table S4.

Notice that, for the sake of simplicity, at present we treat all the constraints in the same way, even if those corresponding to enzymatic rules should not be so strict as the others: we leave the appropriate refinements, implementing a finite energy cost instead of an infinite one, for future versions of the model.

The single-site term \( H^1 \) of Eq. (S33) represents the energy of a fragmentation at \( \nu \) (requiring \( r_\nu = 0 \)), and depends only on the experimental spectrum \( \Sigma \) and fragmentation pattern at site \( \nu \). According to the fact that there cannot be fragmentation at 0 and M, we set \( \mathcal{F}(\sigma_0) = \mathcal{F}(\sigma_M) = \emptyset \), whence \( H^1_0 = H^1_M = 0 \) for all possible values of \( \sigma_0 \) and \( \sigma_M \).

\( H^1 \) considers which and how many ions of the different species, together with their neutral losses, are generated and match some peaks. This is a completely local term, acting like an “external field” rewarding or penalizing fragmentations at \( \nu \). Due to its local nature, it rules out the possibility of describing non-local events, such as correlations between ion types produced at fragmentations at neighboring residues, correlations between intensities of fragments from different sites, or the possibility that a peak is actually produced by two fragmentations, at different sites, that happen to produce ions with the same \( m/z \) ratio.

Comparing Eq. (S35) with Eq. (S19), we see that we can complete the mapping between the interpretation problem and the physical model behavior by: first, identifying the fragmentation sites in the latter equation with the discrete lattice sites of the physical model; second, identifying

\[
H^1_\nu(\sigma_\nu, \Sigma) \equiv \sum_{i=1}^{N^r(\sigma_\nu)} H^r_\nu(s^\nu_i, \xi^\nu_i),
\]

(with \( N^r(\sigma_\nu) \) the number of ion types compatible with the state variable at \( \sigma_\nu \)), and third, establishing the equivalence:

\[
\sum_{\{\xi^\nu_{i} = 0.1\}} p(\phi|\Sigma, T) \equiv p(\mathcal{F}(P)|\Sigma, R).
\]

Notice that the latter equation reflects the fact that the relation between model states and sequences
is not one-to-one: a sequence is described by several configurations of the system, with the same fragmentation sites but corresponding to the different choices of the variables $\xi_{s,i}^{\mu}$, that is, to the different fragmentation patterns at each site.

C. Characterization of the Equilibrium State of the Model, and of Thermodynamics Variables

From equation Eq. (S35) it is clear that the calculation of the partition function,

$$Z = \sum_{\{\sigma_v\}_{v=0..M}} e^{-\beta H}$$  \hspace{1cm} (S41)

is the key step to produce properly defined probabilities. Moreover, it is strictly related to the free-energy $F = -T \ln Z$, which is minimal when the system reaches equilibrium. Other relevant thermodynamic variables (that can be derived from $Z$ by derivatives on $T$, but are more conveniently written as ensemble averages) are the average energy $U = \langle H \rangle$ and entropy $S = \beta (\langle H \rangle - F)$, that will be related to the average “energy score” of the solution (and, at $T = 0$, to the energy of the best sequence $P'$), and to the population of alternative configurations of the model (and hence, alternative sequences), other than the best one, at a given $T$. Another fundamental quantity is that defined in Eq. (10) in the main text,

$$p_v(a) = \langle \Delta^a_{v-1,v} \rangle,$$  \hspace{1cm} (S42)

which represents the probability of the residue species $a$ to end in $v$. Notice that for any $v$, it holds that $\bar{p}_v + \sum_{a} p_v(a) = 1$, where we have defined $\bar{p}_v = \langle \Delta^0_{v-1,v} \rangle = \sum_{r>0} \langle \delta_{r,v,r} \rangle$ the probability that $v$ does not coincide with a fragmentation point.
In all the above expressions, the average of any quantity $X$, $\langle X \rangle$, is expressed as:

$$\langle X \rangle = \frac{1}{Z} \sum_{\{\sigma_v\}} X e^{-\beta H}. \quad \text{(S43)}$$

From the above discussion, it is clear that both the partition function and any other relevant observable involve the sum over all the configurations of the system. The latter can be performed efficiently, in an iterative way, thanks to the simple form of the energy function, that involves just next-neighbors interactions.

First, let us notice that that the state-variables $\xi_{v, \nu}$ depend only on the local state at the cleavage site $v$ so one can integrate them out and use an effective energy in the calculations. The resulting system is described by the reduced variable $\sigma_v^\dagger = (r_v, q_v^X, l_v^X, \pi_v)$. Actually, Eq. (S36) shows that $H^2_{v-1,v}(\sigma_{v-1}, \sigma_v) = H^2_{v-1,v}(\sigma_{v-1}^\dagger, \sigma_v^\dagger)$ is already a function of the reduced variables only, so that the integration is related to the $H^1$.

Indeed, starting from Eq. (S41), and the definitions Eqs. (S33) and (S34), one can write:

$$Z = \sum_{\{\sigma_v^\dagger\}} e^{-\beta H^\dagger} \quad \text{(S44)}$$

where we have introduced:

$$e^{-\beta H^\dagger} \equiv Z^\dagger = \sum_{\{\xi_{v, \nu} = 0, 1\}} e^{-\beta H} = e^{-\beta \sum_{v=1}^M H^\dagger_{v-1,v}} \prod_{v=1}^M \left( \prod_{i=1}^{N_v(\sigma_v^\dagger)} \left( 1 + e^{-\beta H^\dagger_{v, i}} \right) e^{-\beta \mu} \right). \quad \text{(S45)}$$

From the latter equation one can see that the integration of the $\xi$ variables in the expression of the energy Eq. (S34), is accounted for by simply substituting $H^1_v(\sigma_v, \Sigma)$ with

$$H^1_{v}(\sigma_v^\dagger, \Sigma) = \sum_{i=1}^{N_v(\sigma_v^\dagger)} \left( \mu - \frac{1}{\beta} \ln(1 + e^{-\beta H^\dagger_{v, i}}) \right) \quad \text{(S46)}$$

with $H^\dagger_{v, i} = (\delta_{I(s_i)} h^M_{v}(s_i, I(s_i)) + (1 - \delta_{I(s_i)}) h^{-M}_{v}(s_i))$ (see Eqs. (S20) to (S22)). Notice that the above
equation can be considered as the energy associated to a peptide sequence, after taking the marginals over all fragmentation patterns at a given \( \nu \), produced by the different combinations of the \( \xi \) variables.

On the other hand, Eq. (S43) can be written in term of the reduced variable \( \sigma^\dagger_\nu \) as:

\[
\langle X \rangle = \frac{1}{Z} \sum_{\{\sigma^\dagger_\nu\}_{\nu=0\ldots M}} \langle X \rangle \xi e^{-\beta H^\dagger}, \quad (S47)
\]

where

\[
\langle X \rangle_\xi = \frac{1}{Z^\dagger} \sum_{\{\xi_\nu\}_{\nu=0\ldots M}} X e^{-\beta H}, \quad (S48)
\]

is the average over just the \( \xi \) variables. Thus, it is generally possible to recast the averages in Eq. (S43) in term of the reduced variables \( \sigma^\dagger_\nu \) by the formal substitution \( \langle X \rangle = \langle X^{(\xi)} \rangle_{\sigma^\dagger} \), with \( X^{(\xi)} = \langle X \rangle_\xi \) provided by the above equation. In the end, we are left with the simple recipe that integration on the \( \xi \) variables leads to the substitution \( \sigma_\nu \to \sigma^\dagger_\nu, e^{-\beta H} \to e^{-\beta H^\dagger} \) and \( \langle X \rangle = \langle X^{(\xi)} \rangle_{\sigma^\dagger} \) in Eq. (S41), Eq. (S43). In the following, we will show how the sum and averages over the reduced variables \( \sigma^\dagger_\nu \) can be performed. For the sake of simplicity in the notation, we will drop the \( \dagger \) in all formulas, assuming that the integration over the \( \xi \) variables has already been performed, yielding the proper changes in the arguments.

**Calculation of the Partition Function.** The integration over the \( \xi \) variables does not change the simple, next-neighbor structure of the energy function, so that it is possible to resort to an iterative procedure for the evaluation of the partition function and equilibrium averages. We write \( Z \) as:

\[
Z = \sum_{\{\sigma_\nu\}_{\nu=1}} \prod_{\nu=1}^{M} e^{-\beta H_{\nu-1,\nu}(\sigma_{\nu-1},\sigma_\nu)} \equiv Z_{\mu | \mu=M} \quad (S49)
\]
where we have introduced the reduced partition function $Z_\mu$, restricted to the subsystem characterized by the sites $\nu$ in $[0, \mu]$:

$$Z_\mu = \sum_{\nu=1}^{\mu} e^{-\beta H} = \sum_{\nu=1}^{\mu} \prod_{v=1}^{\mu} e^{-\beta H_{\nu-1, \nu}(\sigma_{\nu-1, \sigma_{\nu}})} \tag{S50}$$

$$= \sum_{\sigma_\mu} W_\mu(\sigma_\mu) \tag{S51}$$

with $W_\mu(\sigma_\mu)$ defined as:

$$W_\mu(\sigma_\mu) = \sum_{\nu=1}^{\mu-1} \prod_{v=1}^{\mu} e^{-\beta H_{\nu-1, \nu}(\sigma_{\nu-1, \sigma_{\nu}})} \tag{S52}$$

We can actually write the vector $W_\mu(\sigma_\mu)$ recursively, as a function of $W_{\mu-1}(\sigma_{\mu-1})$:

$$W_\mu(\sigma_\mu) = \sum_{\sigma_{\mu-1}} W_{\mu-1}(\sigma_{\mu-1}) e^{-\beta H_{\mu-1, \mu}(\sigma_{\mu-1, \sigma_{\mu}})} \tag{S53}$$

In practice, it is convenient to introduce the normalized $\zeta_\mu(\sigma_\mu) = W_\mu(\sigma_\mu)/Z_\mu$, that can be efficiently calculated as:

$$\zeta_\mu(\sigma_\mu) = \frac{1}{\phi_{\mu-1, \mu}} \sum_{\sigma_{\mu-1}} \zeta_{\mu-1}(\sigma_{\mu-1}) e^{-\beta H_{\mu-1, \mu}(\sigma_{\mu-1, \sigma_{\mu}})} \tag{S54}$$

where we have defined $\phi_{\mu-1, \mu} = Z_\mu/Z_{\mu-1}$. The latter can be calculated at each step by imposing that $\sum_{\sigma_\mu} \zeta_\mu(\sigma_\mu) = 1$.

**Calculation of the Average Energy.** After integration of the $\xi$ variables, we can write the average energy as:

$$U = \frac{1}{Z} \sum_{\nu=1}^{M} \left( \sum_{\alpha=1}^{M} H^{(\xi)}_{\alpha-1, \alpha}(\sigma_{\alpha-1, \sigma_{\alpha}}) \right) \prod_{v=1}^{\nu} e^{-\beta H_{\nu-1, \nu}(\sigma_{\nu-1, \sigma_{\nu}})} \tag{S55}$$
We notice that we can introduce the average energy \( \langle E^{(\xi)}_{\mu} \rangle_{(0,...,\mu)} \) for the subsystem \( v = 0 \ldots \mu \) as:

\[
\langle E^{(\xi)}_{\mu} \rangle_{(0,...,\mu)} = \sum_{\sigma_{\mu}} \varepsilon_{\mu}(\sigma_{\mu})
\]

(S56)

where:

\[
\varepsilon_{\mu}(\sigma_{\mu}) = \frac{1}{Z_{\mu}} \sum_{\{\sigma_{v}\}} \left( \sum_{\alpha=1}^{\mu} H^{(\xi)}_{\alpha,\alpha-1,\alpha} \right) \prod_{v=1}^{\mu} e^{-\beta H_{v-1,v}(\sigma_{v-1},\sigma_{v})}
\]

(S57)

As for the partition function above, this quantity can be calculated resorting to the recursive expression:

\[
\varepsilon_{\mu}(\sigma_{\mu}) = \frac{Z_{\mu-1}}{Z_{\mu}} \sum_{\sigma_{\mu-1}} \left[ \varepsilon_{\mu-1}(\sigma_{\mu-1}) + H^{(\xi)}_{\mu-1,\mu}(\sigma_{\mu-1},\sigma_{\mu}) \zeta_{\mu-1}(\sigma_{\mu-1}) \right] e^{-\beta H_{\mu-1,\mu}}
\]

(S58)

with \( \zeta_{\mu}(\sigma_{\mu}) \) defined in Eq. (S54) The final value of \( U \), can then be calculated from the value of \( \varepsilon_{M}(\sigma_{M}) \) as:

\[
U = \sum_{\sigma_{M}} \varepsilon_{M}(\sigma_{M})
\]

(S59)

The entropy is calculated as \( S = \beta U + \log Z \).

**Calculation of the Fragmentation Probability.** To identify the parent peptide sequence we need to calculate, at every site \( \mu \), the probability of a residue \( a \) to end in \( \mu \) (see Eq. (S42)). To this end, we introduce the analogue of \( \zeta_{\mu}(\sigma_{\mu}) \), calculated iterating down from \( M \), as:

\[
\tilde{\zeta}_{\mu}(\sigma_{\mu}) = \frac{1}{Z_{\mu}} \sum_{\{\sigma_{v}\}} \prod_{v=\mu+1}^{M} e^{-\beta H_{v-1,v}(\sigma_{v-1},\sigma_{v})}
\]

\[
= \frac{Z_{\mu+1}}{Z_{\mu}} \sum_{\sigma_{\mu+1}} e^{-\beta H_{\mu+1,\mu+1}(\sigma_{\mu},\sigma_{\mu+1})} \tilde{\zeta}_{\mu+1}(\sigma_{\mu+1})
\]

(S60)
that, if we define \( \tilde{Z}_\mu = \frac{Z}{Z_\mu} \) we can write

\[
p_\mu(a) = \langle \Delta a_{\nu-1,\nu} \rangle
\]

\[
= \frac{Z_{\mu-1}}{Z_\mu} \sum_{\sigma_\mu} \sum_{\sigma_{\mu-1}} \Delta a_{\nu-1,\nu} \xi_{\mu-1}(\sigma_{\mu-1}) \xi_\mu(\sigma_\mu) e^{-\beta H_{\mu-1,\nu}(\sigma_{\mu-1},\sigma_\mu)}
\]

with \( \Delta a_{\nu-1,\nu} \) defined in Eq. (S38).

**Results**

**Treatment of the Parent Mass errors**

As described in the main text, the discretization of the mass in a mass array \( \nu = 0, \ldots, M \) involves some truncation of the residues masses, that, according to the parent sequence and length, can produce a cumulative effect significant enough to cause a unit shift in the estimate of the parent mass.

Figure S1 shows the distribution of the mismatch

\[
\Delta m = \left( M^{\text{dta}} - m(\text{N-term}) - m(\text{C-term}) - m(H) \right) - M \eta
\]

between the mass extracted from the “.dta” input file and that corresponding to the theoretical sequence, for the spectra in the test database. The latter mass is calculated as \( M \eta \), where \( \eta \) is obtained from Eq. (S26). Our protocol of estimation of the discretized parent mass, Eq. (S26), yields a wrong result in 47 out of 280 spectra of the test dataset.

To deal with this problem, we repeat the evaluation of the average energy \( U \) assuming \( M \) and \( M + 1 \) as the parent mass, where \( M \) is the discrete mass estimated according to Eq. (S27). We select the parent mass according to the value that yields the lowest value of the average energy, according to the idea that the correct parent mass should yield more matches to the experimental peaks, and hence a lower energy. We have seen that this choice improves the figures for the parent mass error,
Figure S1: Distribution of the parent mass mismatch ($\Delta m$) for the test dataset. The plot illustrates the difference between the precursor mass reported in the .dta file, stripped of the mass of N- and C-term and of the extra proton, and the theoretical mass obtained multiplying the discrete mass, corresponding to the parent sequence, by $\eta$. In 47 cases this distribution exceeds the $[-0.5,0.5]$ range, where an exact estimation of the precursor mass can be done.

even if it is not enough to eliminate it completely: see Table 1 in the main text.

**De-novo results:**

1. **Probability profiles at different temperatures.**

Figure S2 reports the probability profile for the same spectrum of Fig. 2 in the main text, at different temperatures. Notice that the $T = 0.1$ there is essentially one possible state, corresponding to the most likely sequence.

At $T = 3$, just a few fragmentation sites maintain a clear relevance, while probabilities of alternative fragmentations are different from zero at most sites. This tendency increases upon increasing the temperature, until no interesting information can be extracted any more from the equilibrium state.
Figure S2: Example of probability profiles obtained for the same spectrum at increasing temperatures. At each site \( \nu \), the height of the bar represents \( p_\nu(s) \), and the label on it expresses the species \( s \) to which it refers. Bars exceeding 1 represent the true parent sequence, as predicted by SEQUEST.

Figure S3: Temperature dependence of phi-values \( \Phi \) (continuous line), \( \Phi' \) (points) and entropy (dashed line) for two sample spectra, with “true” sequence (a) ALAEHGIVFGEPK (same spectrum as in Figure S2). (b) DLEHPIEVPGVK.
2. T-dependence of F-values and thermodynamics quantities

Figure S3 reports the behaviour of two sample spectra, characterized by a high quality and low quality quality \textit{T-novoMS} prediction, respectively. As expected, \( \Phi \) and \( \Phi' \) values are always very close at low temperatures, when the profile F-value \( \Phi \) is completely dominated by the low temperature configuration, accounted for also by \( \Phi' \). For the first spectrum we see that at low \( T \), both the profile indicator \( \Phi \) and the sequence one \( \Phi' \) have a high value, indicating a good quality of the prediction. Notice that \( \Phi \) is a continuous function of \( T \), while \( \Phi' \) is discontinuous, due to the fact that the most likely sequence changes abruptly at increasing temperatures, and so do \( TP \) and \( PP \). On the contrary, the profile F-value entails a sum over fragmentation probabilities at the correct sites, corresponding to the “true” sequence, which are, in general, decreasing function of \( T \), as the probability distribution of \( p_{\nu}(s_i) \) spreads on an increasing number of sites \( \nu \), and lowers as a consequence of normalization as the temperature increases.

Notice that \( \Phi' \) may also increase upon increasing the temperature: this uncommon event is associated to wrongly predicted sequences, where the probability of fragmentation at the correct site can benefit from the initial increase of temperatures, and it is related to the fact that raising the temperature can increase the coverage of true fragmentations predicted. This can be attributed to an imperfect energy landscape that sometimes induces a ground state which is quite different from the true solution, and traps the system into an energy pit: increasing the temperature, and hence fluctuations, helps the system access configurations more similar to the true solution.

Notice that neither \( \Phi \) nor \( \Phi' \) are known in the real-life case when the “true” parent peptide of a spectrum is unknown, and is precisely what one is seeking for. For this reason, it is necessary to find a thermodynamics observable, like the entropy, that correlates with the F-values and may indirectly inform about the quality of the interpretation.

3. Correlation between the quality of prediction and symbol entropy

Table S5 complements Fig. 3 in the main text, reporting several indicators of the relationship between sequence entropy and prediction F-value. We see that only at very low entropies \( S^s < 1 \)
Table S5: Analysis of the relations between sequence entropy and quality parameter F-value Φ. The data refer to all 7839 spectra with charge 2 in the extended learning dataset; 941 are of good quality (Φ > Φ₀ = 0.8). True positives TP=n(S < S₀, Φ > Φ₀), false positives FP=n(S < S₀, Φ < Φ₀), precision Pr=TP/(TP+FP) negative predicted value NP= TN/(TN+FN) (with TN=n(S > S₀, F < F₀) and FN= n(S > S₀, F > F₀) ), recall Rc = TP/(TP+FN), specificity Sp = TN/(TN+FP), F-value Fv=2PrRc/(Pr + Rc). Here n(●) means “number”; probabilities p(●) are expressed as a percentage. Notice that Fv suggests selecting S₀ = 4 as a cutoff to distinguish good from bad predictions; however, the precision and recall values are limited to 0.44 and 0.75, pointing at a still poor “resolution” of good and bad predictions. On the other hand, NP(8)=1: all but one of the 3435 spectra with S > 8 are of poor quality Φ < 0.8

<table>
<thead>
<tr>
<th>S₀</th>
<th>nₛ&lt;ₛ₀</th>
<th>TP</th>
<th>FP</th>
<th>Pr</th>
<th>NP</th>
<th>Rc</th>
<th>Sp</th>
<th>Fv</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>131</td>
<td>98</td>
<td>33</td>
<td>75</td>
<td>89</td>
<td>10</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>478</td>
<td>329</td>
<td>149</td>
<td>69</td>
<td>92</td>
<td>35</td>
<td>2</td>
<td>46</td>
</tr>
<tr>
<td>3</td>
<td>985</td>
<td>524</td>
<td>461</td>
<td>53</td>
<td>94</td>
<td>56</td>
<td>7</td>
<td>54</td>
</tr>
<tr>
<td>4</td>
<td>1609</td>
<td>703</td>
<td>906</td>
<td>44</td>
<td>96</td>
<td>75</td>
<td>13</td>
<td>55</td>
</tr>
<tr>
<td>5</td>
<td>2337</td>
<td>838</td>
<td>1499</td>
<td>36</td>
<td>98</td>
<td>89</td>
<td>22</td>
<td>51</td>
</tr>
<tr>
<td>6</td>
<td>3042</td>
<td>899</td>
<td>2143</td>
<td>30</td>
<td>99</td>
<td>96</td>
<td>31</td>
<td>45</td>
</tr>
<tr>
<td>7</td>
<td>3752</td>
<td>928</td>
<td>2824</td>
<td>25</td>
<td>100</td>
<td>99</td>
<td>41</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>4404</td>
<td>940</td>
<td>3464</td>
<td>21</td>
<td>100</td>
<td>100</td>
<td>51</td>
<td>35</td>
</tr>
<tr>
<td>9</td>
<td>4939</td>
<td>941</td>
<td>3998</td>
<td>19</td>
<td>100</td>
<td>100</td>
<td>58</td>
<td>32</td>
</tr>
</tbody>
</table>

we are able to single out good predictions with a 75% reliability, but the coverage is low (just 98 of 941 good sequences present such a low entropy). On the other hand, predictions with S > 5 are of poor quality 89% of the times.

**Database-search results**

Table S6 for TSET1 shows the dependence of the number of true and false positives, together with a few other derived quantities like precision, recall and F-value, as a function of the z-score threshold z₀ used to define correct and wrong predictions. Table S6 reveals that a cutoff z₀ = 6.27 yields a FDR<1%, with a coverage of 45%: there is less than 1% probability that a sequence characterized by z > z₀ is not the correct one, and 45% of all spectra in TSET1 are predicted with this level of reliability.

Figure S4 shows the signals obtained with different definitions of the False Discovery Rate, as a function of coverage or of z-score threshold. The agreement of our definition of FDR as
Table S6: Reliability of the z-score $z$ (see text) as a blind indicator of the quality of the database search, for the 280 spectra in the test database TSET1, at $T = 2$, with noise pre-filtering. Here a true positive $TP = n(z > z_0, r=1)$ is a correct prediction (the best ranking sequence in the database coincides with the “true” peptide) with $z > z_0$; a false positive $FP = n(z > z_0, r>1)$ has $z > z_0$ but the best ranking is not the true parent sequence. Analogously, precision $Pr = p(r=1|z > z_0)$, recall $Rc = p(z > z_0|r=1)$, F-value $Fv = 2PrRc/(Pr + Rc)$, coverage $C = n(z > z_0)/n($spectra$)$. Here $n(\bullet)$ means “number”; probabilities $p(\bullet)$ are expressed as a percentage. Since the False Discovery Rate $FDR = 1 - Pr$, we see that FDR is less than 1% at least for $z > 6.27$, a condition satisfied by the 45% of all the spectra (coverage, last column). The best balance between precision and recall is obtained for $z_0 = 4.63$, when the F-value is maximal (94.19%).

<table>
<thead>
<tr>
<th>$z_0$</th>
<th>$n_{z&gt;z_0}$</th>
<th>$TP$</th>
<th>$FP$</th>
<th>$Pr$</th>
<th>$Rc$</th>
<th>$Fv$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.50</td>
<td>255</td>
<td>231</td>
<td>24</td>
<td>90.59</td>
<td>97</td>
<td>94.09</td>
<td>91</td>
</tr>
<tr>
<td>4.63</td>
<td>246</td>
<td>227</td>
<td>19</td>
<td>92.28</td>
<td>96</td>
<td>94.19</td>
<td>87</td>
</tr>
<tr>
<td>4.70</td>
<td>243</td>
<td>224</td>
<td>19</td>
<td>92.18</td>
<td>94</td>
<td>93.53</td>
<td>86</td>
</tr>
<tr>
<td>5.10</td>
<td>221</td>
<td>210</td>
<td>11</td>
<td>95.02</td>
<td>88</td>
<td>91.90</td>
<td>78</td>
</tr>
<tr>
<td>5.50</td>
<td>193</td>
<td>188</td>
<td>5</td>
<td>97.41</td>
<td>79</td>
<td>87.65</td>
<td>68</td>
</tr>
<tr>
<td>5.90</td>
<td>158</td>
<td>154</td>
<td>4</td>
<td>97.47</td>
<td>65</td>
<td>78.17</td>
<td>56</td>
</tr>
<tr>
<td>6.27</td>
<td>127</td>
<td>126</td>
<td>1</td>
<td>99.21</td>
<td>53</td>
<td>69.42</td>
<td>45</td>
</tr>
<tr>
<td>6.30</td>
<td>124</td>
<td>124</td>
<td>0</td>
<td>100.00</td>
<td>52</td>
<td>68.89</td>
<td>44</td>
</tr>
</tbody>
</table>

$FDR = (do + 2db)/(db + tb + to)$, adopted from Navarro and Vázquez $^9$, with the independent one $FDR' = FP/PP$ (where $PP$ is the number of spectra $\Sigma$ with $z_T(\Sigma) > z_0$, and $FP$ is the number of spectra whose true precursor sequence does not coincide with the best in the database, yet $z_T > z_0$) suggests that it is a better definition than the standard one, $FDR_{st} = ND/NT$, where ND and NT are the number of spectra in the decoy and target database with a $z$-score above the $z$-score threshold $z_0$.

On the other hand, the comparison with MASCOT results can be done just in terms of the standard definition of FDR. For this reason, for our database-search results, we calculate the values of the FDR with both definitions, and compare the coverage at the same level of $FDR_{st}$ obtained with Mascot. Figure S5 shows the relationship between FDR and $z_0$ or the coverage for TSET1 and TSET2, and also for ELSET, for comparison. A first remarkable thing is that the results obtained for $FDR$ and $FDR'$ in TSET1 are quite close, despite their very different nature (the former is based on a comparison between targets and decoys, the latter relies on the knowledge of the “true parents”, for spectra in TSET1). The definition used for the $FDR^9$ appears indeed better than
Figure S4: Comparison of the performance of the definition of False Discovery Rate adopted here (FDR; see Methods) with the standard one, \( FDR_{st} = \frac{ND}{NT} \), against the signal of \( FDR' \), which is calculated from the knowledge of the true sequences, for spectra in TSET1. The black horizontal line signals the threshold FDR=0.01. Notice that the definition we adopt is closer to the \( FDR' \), and quite different from the standard definition.

alternative ones. The very good performance, in terms of high coverage at low FDR, on ELSET is expected, since the latter is an extension of the learning dataset, so that its signals could be considered as a reference for comparisons. However, it is not trivial that TSET1 behaves similarly to ELSET for \( FDR < 0.05 \), and they both approach the \( FDR = 0.01 \) line for \( z_0 \approx 6.3 \) (top panel), even if the method obtains a lower coverage for TSET1 (bottom panel). Such good performance is not repeated with TSET2, where low values of FDR are only obtained for high z-threshold, that correspond to a small fraction of all the spectra, yielding a small coverage, of around 2% of all spectra, when \( FDR = 0.01 \). However, the comparison with MASCOT in the main text reveals that the behavior of the latter on TSET2 is not very different from that of \textit{T-novoMS}. So, the very different performance on TSET1 and TSET2 in Figure S5 can be mostly attributed to a different quality of the two datasets.
Figure S5: FDR as a function of the z-score threshold \( z_0 \) and of coverage, for the ELSET, TSET1, and TSET2 datasets. The score was calculated with the *de novo* probability profiles at \( T = 2 \), with pre-filtering of the spectra (see Methods). Coverage is defined as the fraction, of all the spectra in the dataset, satisfying \( z > z_0 \). The dashed line corresponds to FDR= 1%.

**Notes and References**


(8) Notice that this assumption can be made only for the fragmentation sites $v_k \in \mathcal{F}(P)$ of $P$, and not for any site: the sequence $P$ determines which sites are fragmentation sites, and the probability to observe a fragment of type $s$ at $v$ and $s'$ at, e.g., $v + 1$ are not independent. However, we can still assume that the generations of fragment ions at the different fragmentation sites of $P$ are independent processes.