Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2025, **16**, 18243

All publication charges for this article have been paid for by the Royal Society of Chemistry

Framework for *de novo* sequencing of peptide mixtures *via* network analysis and two-dimensional tandem mass spectrometry

MyPhuong T. Le,[a] Yu Zhu,[b] Eric T. Dziekonski,[a] Dylan T. Holden,[a] David F. Gleich[b] and R. Graham Cooks *[a]

Received 23rd May 2025
Accepted 30th August 2025

DOI: 10.1039/d5sc03762j

rsc.li/chemical-science

Two-dimensional tandem mass spectrometry (2D MS/MS) provides in-depth biopolymer structural information previously not directly accessible with traditional one-dimensional MS/MS workflows, and in significantly less time (<1 second per sample). In this study, we enhance 2D MS/MS data analysis for greater applicability in omics workflows and address challenges in sequencing peptides in mixtures. We designed a graph-theory-based framework to efficiently manage, visualize, and maximize the structural information extractable from 2D MS/MS spectra. Graph analysis algorithms, including a PageRank-based method, are shown to deconvolve MS/MS signals and group together product ions from the same presursor peptide, enabling the reconstruction of peptide fragmentation trees. From this, MS^n information can be extracted to improve sequencing accuracy relative to current MS/MS methods. We also introduce a computationally efficient *de novo* sequencing approach that leverages this structural information to reduce reliance on databases and sample separation, while also enabling the rapid sequencing of post-translationally modified peptides. Tests on simulated 2D MS/MS spectra, designed to mimic those from proteomic samples, achieved high precision in signal assignment. Proof-of-concept studies were conducted on real data from simple mixtures of short chain peptides, showing the potential applicability of combining network analysis with *de novo* sequencing to analyze unknown peptide mixtures. We anticipate that this technique will complement proteomics workflows and facilitate direct biopolymer structural analysis.

1 Introduction

Mass spectrometry (MS) lies at the heart of omics studies[1,2] and is fundamental to the sequencing of biomolecules[3–5] owing to the predictable nature of gas-phase fragmentation, especially under collision-induced dissociation (CID) conditions.[6–11] This allows for structural elucidation of biomolecules from their MS/MS spectra *via* sequencing methods such as database search and *de novo* sequencing.[12–21] A wide range of software has been developed for this purpose and tailored to specific sample types, applications, and experiments, allowing for the significant expansion of omics studies over the past several decades.[22–24] Challenges remain. Accuracy of database searches relies heavily on the quality and specificity of the database, not only in terms of sequence composition but also in the experimental conditions chosen. *De novo* sequencing methods, in concept, remove the need for databases and allow for the discovery of novel structures.[13,25–27] However, this generally requires high-resolution data and is sensitive to the presence of noise and incomplete fragmentation.[26,28,29] Structural modifications, or post-translational modifications (PTMs) in the context of proteomics, remain a major problem as they greatly expand the search space and require databases specific to each modification, the curation of which requires significant effort while the discovery of novel PTMs remains difficult.[30,31] Finally, complex mixtures containing multiple biopolymers as well as other components add another layer of challenge to biopolymer studies.[26,32–34]

Two-dimensional mass spectrometry encompasses techniques that correlate precursor ions with their corresponding product ions without isolating the former. This experiment was first developed using FT-ICR MS instruments.[35–37] Newer two-dimensional methods, Partial Covariance MS (2D-PC-MS) and Fragment Correlation MS (FC-MS), rely solely on signal processing and achieve similar results without significant instrumentation modifications.[38–41] These methods generate high-quality data, although they are generally time-consuming, requiring many minutes to hours of analysis. Recently, our lab developed a method of two-dimensional tandem mass spectrometry (2D MS/MS) based on a single quadrupole ion trap analyzer. The experiment has relatively low mass resolution but data acquisition times are only a few seconds,[42] providing in-

[a]*Department of Chemistry, Purdue University, West Lafayette, IN 47907, USA. E-mail: cooks@purdue.edu*

[b]*Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA*

depth structural information previously inaccessible with 1D MS/MS.[43] In these experiments, biopolymer samples are subjected to two stages of fragmentation, first before mass analysis and second during mass analysis so preserving precursor–product relationships without sacrificing analysis time or sample usage. Essentially the experiment provides – for a single population of ions – product ion spectra of all precursor ions. Stairstep patterns can be extracted from the resulting 2D MS/MS spectrum to reconstruct MS^2 pathways of the molecular ion, from which information regarding the interconnectivity of the monomers and any structural modifications is inferred. The ability to recover such information makes this method a promising route to biopolymer structural analysis, potentially addressing some of the challenges outlined above. Specifically, the preservation of precursor–product relationships between a molecular ion and its fragment allows efficient deconvolution of fragmentation signals associated with each component of the mixture. The precursor–product relationships among fragments generated from the same biopolymer can improve sequencing accuracy compared to traditional methods which give only one-dimensional MS/MS data.[44]

In this paper, we propose a framework that combines the two-stage fragmentation method of 2D MS/MS with a graph theory-based data analysis scheme to efficiently manage, visualize, and extract information. A network analysis algorithm was developed to deconvolute signals from 2D MS/MS spectra of biopolymer mixtures, and it was evaluated against synthetic data that mimic the complexity of typical proteomics data to test potential applicability to current proteomics workflows. We also introduce a *de novo* sequencing algorithm to support peptide structural elucidation using the fragmentation tree reconstructed from the 2D MS/MS signals. By taking advantage of well-defined neutral loss values, we demonstrate how PTM information can be recovered without the need for preexisting databases, all the while mitigating isobaric interferences and reducing the reliance on high-resolution data in comparison to traditional *de novo* sequencing approaches. Finally, in a proof-of-concept study of peptide mixtures, we apply the proposed network analysis and *de novo* sequencing algorithms to 2D MS/MS data. This framework should be applicable to 2D MS/MS-like data[38–41] generated *via* data independent or dependent acquisitions methods[45] and so enhance the accuracy of current biopolymer analysis studies.

2 Methods

2.1. 2D MS/MS analysis of biopolymer

2.1.1 Two-step fragmentation. Aqueous solutions of biopolymer (50 mM, 1% formic acid) are directly ionized by nano-electrospray ionization (nESI) and sprayed into a linear ion trap instrument (Thermo Electron Finnigan LTQ) modified for 2D MS/MS experiments.[71] Prior to their arrival in the ion trap, injected ions are fragmented simultaneously and indiscriminately by in-source fragmentation (IS-CID) or broadband excitation waveforms, gas-phase methods somewhat akin to the effect of non-specific enzymatic digestion to convert longer (bio) polymers into shorter chain versions.[72] The resulting fragments are analyzed by 2D MS/MS, during which each precursor is sequentially activated and fragmented to form product ions. The m/z values of each precursor are plotted on the x-axis, while the m/z values of their corresponding products are plotted on the y-axis. The resulting 2D MS/MS spectrum is therefore a combination of product ion profiles from each precursor ion generated from the fragmentation of the original biopolymer molecule (Fig. 1A). Following data processing, during which peak centroid identification and m/z calibration against standard were done to obtain precise unit resolution m/z values, stairstep patterns were reconstructed from the signals to recover MS^n information (Fig. S1A). Detailed description of the 2D MS/MS instrumentation, 2D MS/MS data processing, and data interpretation can be found in our earlier publications.[42,43]

For proof-of-concept experiments, an incompletely phosphorylated pentapeptide, Ac-IYGEF-NH_2, and an equimolar mixture of three enkephalin analogs, YGGFL, YAGFL, and YGGFM, were used to examine the capability of the method in analyzing heavily overlapping structures. Additionally, a single-blind study was carried out using a sample containing five peptides covering a mass range from m/z 300 to m/z 800. In addition, 2D MS/MS spectra of these five individual peptides were collected to establish a "ground truth" to test the performance of the graph partitioning algorithms. Final 2D MS/MS spectra were generated by combining spectra from various IS-CID energies and instrument settings to (1) cover a wide range of fragmenting precursors and (2) capture lower mass product ions of high mass precursors.

2.2. A graph-theory based data analysis framework

An example of 2D MS/MS data analysis using graph data structure is shown in Fig. 1B. Data from a 2D MS/MS spectrum is defined as a collection of multiple reaction monitoring (MRM) transitions,[73] or pairwise relationships between two m/z values (Fig. 1B, step 1). Consequently, a graphical representation can be used for efficient data management and visualization. Graphs are data structures with a set of nodes and a set of edges. Each edge represents a connection between two nodes. Accordingly, we model each m/z value as a node and add a directed edge, or an arrow, connecting one m/z value to another if there exists such a precursor-product pair in the 2D MS/MS spectrum. This construction forms a directed acyclic graph (Fig. 1B, step 2). The advantage of this construction is that paths in this graph represent potential fragmentation sequences. A path in graph theory is defined as a sequence of edges connecting a sequence of nodes within a graph, such that no node is repeated along the path. In a directed graph, these paths must follow the direction of each edge.

To organize the resulting complex network, the molecular ion(s) of each biopolymer(s) present in the sample must first be identified. These values are often associated with peaks that (1) are accompanied by common adducts (+22 Da and +38 Da for Na^+ and K^+ adducts relative to +1 in positive mode), (2) decrease in intensity following the application of IS-CID if singly charged, and (3) have a relatively more abundant product ion profile in comparison to other peaks.[74,75] If the sample contains

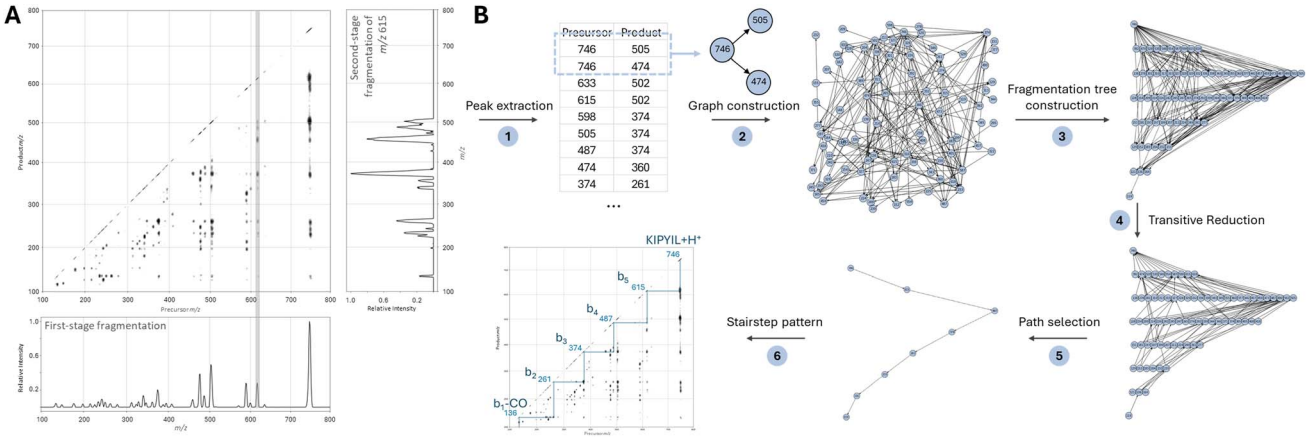

Fig. 1 (A) 2D Tandem mass spectrum of Neuromedin N collected using the two-step fragmentation experiment. Summing all signals onto the *x*-axis allows for the recovery of the full mass spectrum of fragments generated from KIPYIL during IS-CID. The product ion profile of each precursor ion, *e.g.* *m*/*z* 615, can be recovered by extracting signals in the *y*-dimension that coincide with the *x*-axis at the *m*/*z* value of the precursor. Signals from unfragmented precursors are often not observable for reasons associated with the instrument's scan function, but they should fall along the diagonal autocorrelation line where product ion *m*/*z* equals precursor ion *m*/*z*, as reconstructed in the spectrum. (B) Step-by-step data analysis of a 2D mass spectrum begins by locating the coordinates of all signals within the spectrum (step 1). For each pair of coordinates, nodes with *x* and *y* values are created and an edge is added from *x* to *y* (step 2). If the molecular ion is identified, the longest path from each node in the graph to the molecular ion node is calculated, and the length of this path is used to classify each node with respect to layers beneath the molecular ion (step 3). Further simplification can be achieved by applying a restricted transitive reduction to remove redundant edges (step 4). An MS^n path can be reconstructed by tracing a path from the molecular ion through the layers to any lower node of choice, or to the lowest node for the longest path (step 5). Signals with *x* and *y* coordinates corresponding to each pair of nodes in the selected path can then be used to map the stairstep pattern onto the original 2D spectrum (step 6).

a molecular species, a subgraph is extracted by selecting all nodes accessible from the node representing the *m*/*z* value of the identified molecular ion. This subgraph is subsequently plotted as a layered graph, with the molecular ion at the top and descendant nodes arranged in layers based on the length of the longest path between them and the molecular ion. The resulting graph resembles a fragmentation tree (Fig. 1B, step 3).

Further simplification can be done for visualization purposes *via* transitive reduction,[76] an algorithm that removes redundant edges (Fig. 1B, step 4). Given the tree-like structure of the graph, directed paths can be extracted starting from the molecular ion and extending to any descendant node (Fig. 1B, step 5). These paths represent MS^n pathways and, therefore, are equivalent to the stairstep pattern mentioned earlier (Fig. 1B, step 6).

2.3. Network analysis for deconvolution of peptide mixture fragmentation

2.3.1 Graph partitioning algorithm. If multiple biopolymers are determined to be present in the mixture, signal deconvolution is required before the fragmentation trees of individual species can be constructed (Fig. 1B, step 3). The 2D MS/MS data generated from biopolymer mixtures often produce fragments originating from different species that share the same *m*/*z*, resulting in an entangled network of fragmentation trees with "shared" nodes. To resolve this, we developed a graph partitioning algorithm based on Personalized PageRank (PPR)[77] to identify nodes originating from the same biopolymer. Specifically, for each species with identified molecular ions, the PPR score of each node in the graph relative to this ion is calculated, as detailed in the SI.

A two-threshold system was implemented, where high and low thresholds, T1 and T2, are calculated as percentiles of all PPR scores between all nodes and the molecular ion node. Nodes with a PPR score above T1 are classified as fragments of the molecular ion under consideration, while those with scores below T2 are excluded. For scores between T1 and T2, a node is assigned to the species only if its score for this biopolymer is higher than its scores for all other species in the mixture. This process is repeated for each molecular ion identified in the mixture, resulting in nodes that may originate from a single biopolymer, multiple biopolymers, or neither (*e.g.*, noise).

2.4. *De novo* sequencing of unknown peptides using 2D MS/MS data

Once the fragmentation tree of a peptide is constructed, a *de novo* sequencing method is used to recover the peptide's structure based on the following assumptions and principles:

(1) The molecular ion of the peptide is correctly identified.

(2) There is a finite list of possible peptide structures at each *m*/*z* and neutral loss value.

(3) Ions of higher mass must contain the amino acids present in the ions of lower mass within the same MS^n pathway.

(4) Structural information for each fragment can be inferred from the ion type of its precursor and product, as well as the nature of the associated neutral losses.

2.4.1 Library generation. Based on the second assumption above, an *in silico* library of all possible peptide ions and neutral loss structures was generated at unit mass level. Peptide

sequences with masses up to 1000 Da, formed from the 20 common amino acids, were calculated. The nominal mass of each sequence was calculated as b-type, y-type, a-type, and x-type ion; c- and z-ions were excluded as they are not commonly generated under conventional CID conditions.[10,58,78] Neutral losses of up to 300 Da consisting of up to two amino acid residues generated from the fragmentation of intact peptides as well as fragmentation of common product ions (b- and y-types) were calculated (Fig. S3). Post-translationally modified amino acids were included only in the neutral loss library. This latter is based on the weak assumption that the PTM moieties remain attached to the modified amino acid during fragmentation. Neutral losses of small molecules (*e.g.*, H_2O, CO_2, CO, NH_3, *etc.*)[78] were also considered depending on the presence of specific amino acids, based on known fragmentation trends.

2.4.2 Three-step sequencing. A three-step sequencing algorithm that takes an MS^n pathway as input and returns the best possible sequence(s) of molecular ion(s) that may fragment into ions with the corresponding series of *m*/*z* values. An example of this algorithm being applied to a pathway extracted from a fragmentation tree is shown in Fig. 2.

In the first step, the structures of ions at each *m*/*z* value are narrowed down in a bottom-up direction, *i.e.*, from the lowest to the highest mass in each pathway. For each pair of *m*/*z* values, $m_2 \rightarrow m_1$ and their corresponding neutral loss value $n_{12} = m_2 \rightarrow m_1$, a list of candidate structures for m_1, m_2, and n_{12} is retrieved from the library. If a modification P corresponding to a mass shift of ΔP is present in the candidate list of n_{12}, the candidate pool for m_2 is extended to include structures with mass $m_2 - \Delta P$. The third assumption is then applied to narrow down the candidate pool for m_2. This process is repeated until the molecular ion is reached. The first step is repeated across as many pathways as necessary to obtain the smallest possible number of potential structures of the molecular ion.

The second step is performed in a top-down manner. The narrowed-down structure(s) of the molecular ion is used to further refine and narrow down the structural candidates identified in the first step to arrive at the component *m*/*z* values and neutral losses of each pathway.

The third and final step is a bottom-up sequencing process, in which the ion types and neutral loss types are used to accurately determine the relative order of the amino acids in each fragment, ultimately reconstructing the sequence of the molecular ion.

2.4.3 Path selection. To efficiently sequence a molecular ion from its fragmentation tree, pathways are submitted to *de novo* sequencing algorithm based on their information content.

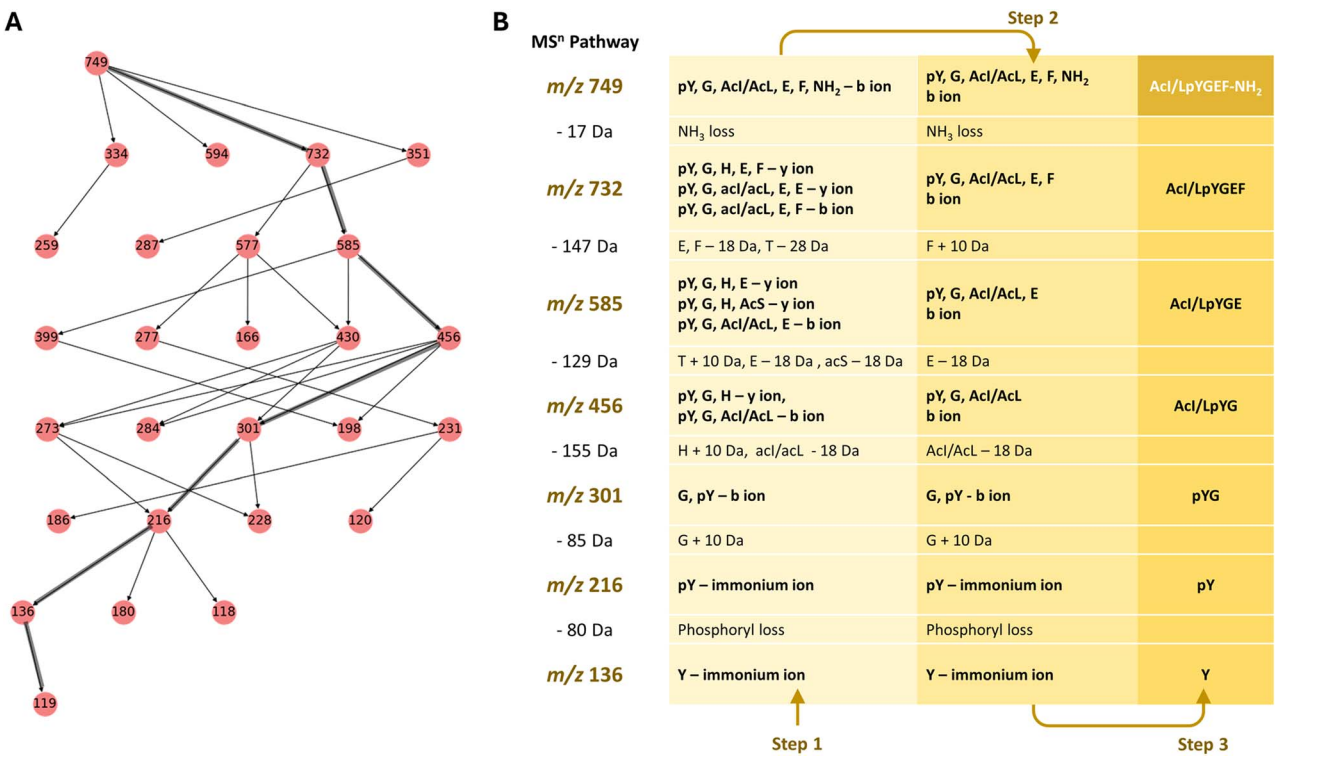

Fig. 2 Application of the *de novo* sequencing algorithm. (A) Stairstep sequence (or MS^n sequence) extracted from a fragmentation tree. (B) The difference between each node was calculated as neutral loss values and annotated as + x Da. Listed next to the node's *m*/*z* and NL values are the corresponding structure candidate pools resulting from the application of structural constraints imposed by the structure pools of the previous node and NL value in the direction indicated by the arrow for each step. Node *m*/*z* 119 is not included in the steps as it does not correspond to any structure in the library and is a known product ion formed by loss of NH_3 from Y (immonium) ions. Examples of how modifications are accounted for in the sequencing algorithm can be found in several $m_2 \rightarrow m_1$ pairs, such as *m*/*z* 456 → *m*/*z* 301. Here, the NL value of 155 Da includes the possibility of an acetylated I/L. As a result, candidates for m_2 in the first sequencing step include fragments of both *m*/*z* 456 and *m*/*z* 456 + 42, with 42 corresponding to the mass shift resulting from the acetyl group.

To identify the most informative pathways, *m*/*z* values providing little structural information, such as those resulting from the loss of small molecules or a neutral loss (NL) not in the library, are disregarded. Typically, these are NLs exceeding 300 Da in mass or those involving rearrangements. After the removal of such *m*/*z* values, the length of the remaining pathway reflects its relative information content. If multiple pathways have similar lengths, priority is given to those with greater interconnectivity among their nodes and stronger connections between each node and the molecular ion node. Interconnectivity can be evaluated by the number of edges within each pathway, while connectivity to the molecular ion can be assessed by the average shortest path from each node to the molecular ion node.

Additionally, pathways with lowest *m*/*z* values should also be prioritized, as fewer structures are possible at low *m*/*z* values (Fig. S4). These ions are often comprised of just one amino acid, which can more efficiently narrow down the possible structures in the subsequent steps. To mitigate interference from potential internal isobaric and misassignments, pathways lacking chemical senses, such as those containing multiple consecutive small losses, are also excluded.

3 Results and discussion

3.1. Analysis of 2D MS/MS data using graph structure

Due to the polymeric nature of the analyte being considered, a particular fragment can be generated during either of the two stages of fragmentation, leading to the presence of signal pairs in which the precursor ion of one event has an *m*/*z* value that is also the product ion *m*/*z* value of the other.[43] By connecting these pairs of signals, stairstep patterns can be drawn to reconstruct an MS^n pathway (Fig. S1), information previously inaccessible using traditional 1D MS and conventional 1D MS/MS methods. Data generated from 2D MS/MS experiments, is therefore rich in structural information, particularly the interconnectivity among substructures of a biopolymer. By arranging 2D MS/MS data into a graphical structure, this valuable information can be efficiently stored, accessed, and utilized in structural studies. This approach should also facilitate the automation of data analysis which would be challenging to perform manually as the complexity of the data increases. In addition, graphical organization makes MS/MS data suitable for diverse network analysis algorithms, many of which offer innovative approaches to addressing MS challenges, such as facilitating the comparison and correlation of MS/MS fragmentation patterns among structurally similar compounds.[46,47]

During graph construction (Fig. 1), multiple disconnected subgraphs may be generated. However, most nodes typically connect to form one large component due to the interconnectivity of the biopolymer's fragments. As a result, noise signals will seldom overlap with significant *m*/*z* values so enabling the removal of noise and background signals. Using graph format also allows simplification and visualization of 2D MS/MS in a more intuitive manner compared to stairstep patterns, facilitating manual data inspection and interpretation when necessary. If the molecular ion(s) can be identified, a layered graph can be constructed from all *m*/*z* values that can be "reached" from the molecular ion, structurally resembling a fragmentation tree. Paths within this graph, therefore, represent potential fragmentation sequences. Such structures are widely recognized by mass spectrometrists as supporting structural elucidation and serve as the basis for various MS software platforms.[13,48–50]

When interpreting the reconstructed fragmentation tree, it is important to note that no relationship between any two *m*/*z* values can be assumed without a direct connection between them. Accordingly, we simplify the fragmentation tree for visualization purposes without compromising the information presented within the graph by employing a restricted transitive reduction method, *i.e.*, an edge can only be removed if there exists a two-step path connecting the nodes. In addition, the absence of a connection between two *m*/*z* values does not negate their relationship. This situation often arises when data is collected using ion trap mass spectrometers, in which smaller fragments may not be detected due to inherent low mass cutoff values.[51,52]

3.2. Graph partitioning algorithm and its evaluation using synthetic peptide mixture fragmentation

3.2.1 Graph partitioning algorithm. Elucidating the structures of individual biopolymers in complex mixtures remains challenging, particularly in omics workflows where co-eluting species generate convoluted MS/MS spectra. This is especially relevant in data-independent (DIA) workflows, where multiple peptides are usually simultaneously co-fragmented in the MS/MS scan.[45] As a result, individual MS/MS profiles must be deconvoluted to be effectively analyzed using common sequencing methods.[53,54] In some current omics workflows, precursor–product relationships are reconstructed by correlating changes in signal intensity, assuming that fluctuations in precursor ion intensity are mirrored by those of its product ions. These correlations are established by analyzing relative ion intensity against LC elution time[55–57] or by leveraging the intrinsic signal intensity fluctuations during MS/MS measurements (2D-PC-MS and FCMS).[38–41] However, both methods require significant analysis time and substantial sample quantities.

As previously noted, isobaric ions, combined with data complexity, pose significant challenges to the application of 2D MS/MS techniques for peptide mixture analysis.[40] This issue is especially pronounced in unit-resolution data.[26,58,59] Along with shared fragments from structurally overlapping peptides, these ions introduce "shared nodes" which convolute the fragmentation trees. Simply extracting fragmentation trees from source nodes, therefore, can lead to misassigned descendant nodes. To address this, we developed a graph partitioning algorithm based on the assumption that an ion with a particular *m*/*z* value is more likely to be a product ion of a molecular ion if there are other fragmentation paths connecting that molecular ion to the product ion *m*/*z* value.

The PageRank algorithm,[60,61] developed to rank components of a network based on how well connected they are *via* the number and quality of their links to others, has been

successfully applied in various scientific disciplines due to its versatility, robustness to noise, and the intuitive nature of the results.[62–64] Personalized PageRank, specifically, allows one to introduce biases into specific nodes (*e.g.*, molecular ions) and to prioritize nodes with direct connections, which is important in the context of MS data and therefore is used in this application. Effective clustering methods must be capable of assigning the same m/z values to single or multiple peptides while also filtering out noise, thus necessitating a two-threshold system. Specifically, threshold T1 accounts for shared structures between different peptides, allowing some nodes to be grouped into multiple clusters simultaneously. Threshold T2 is used to identify and exclude noisy nodes loosely connected to some clusters through isobaric ion contributions. By carefully tuning T1 and T2, we can mitigate isobaric interference and strike a balance between identifying as many peptide fragments as possible while avoiding the inclusion of isobaric ions from other peptides or noise, or in other words, balance recall and precision.

3.2.2 Simulated 2D MS/MS proteomics data. To evaluate the performance and applicability of the proposed graph partitioning algorithm, we generated simulated 2D MS/MS datasets to mimic the complexity of MS data typically obtained during traditional proteomics experiments, specifically co-eluted peptides from a preceding LC separation step (SI, Section 3). Specifically, peptide sequences ranging from 8 to 15 amino acids in length were randomly generated from the 20 common amino acids. Three types of noise signals were added from: (1) coeluting peptides whose molecular ion is not identified, (2) Na^+ and K^+ adducts of peptides in the mixture, and (3) random precursor-product pairs mimicking chemical noise and contaminants. All of these components were subjected to *in silico* two-stage fragmentation, during which arbitrary rearrangements might occur, including internal switching of amino acid positions and losses of neutral fragments of arbitrary mass. The mass of each fragment was recorded at unit resolution while its isotopic distribution was discarded given that these relatively low intensity signals are often not recorded or are removed during signal processing. Signals from all fragments were combined and subsequently deconvoluted using the graph partitioning algorithm. We evaluated the performance using two metrics: precision, which calculates the percentage of correctly assigned nodes out of all nodes assigned to a peptide, and recall, which calculates the percentage of correctly assigned nodes out of all the nodes that actually belong to a peptide.

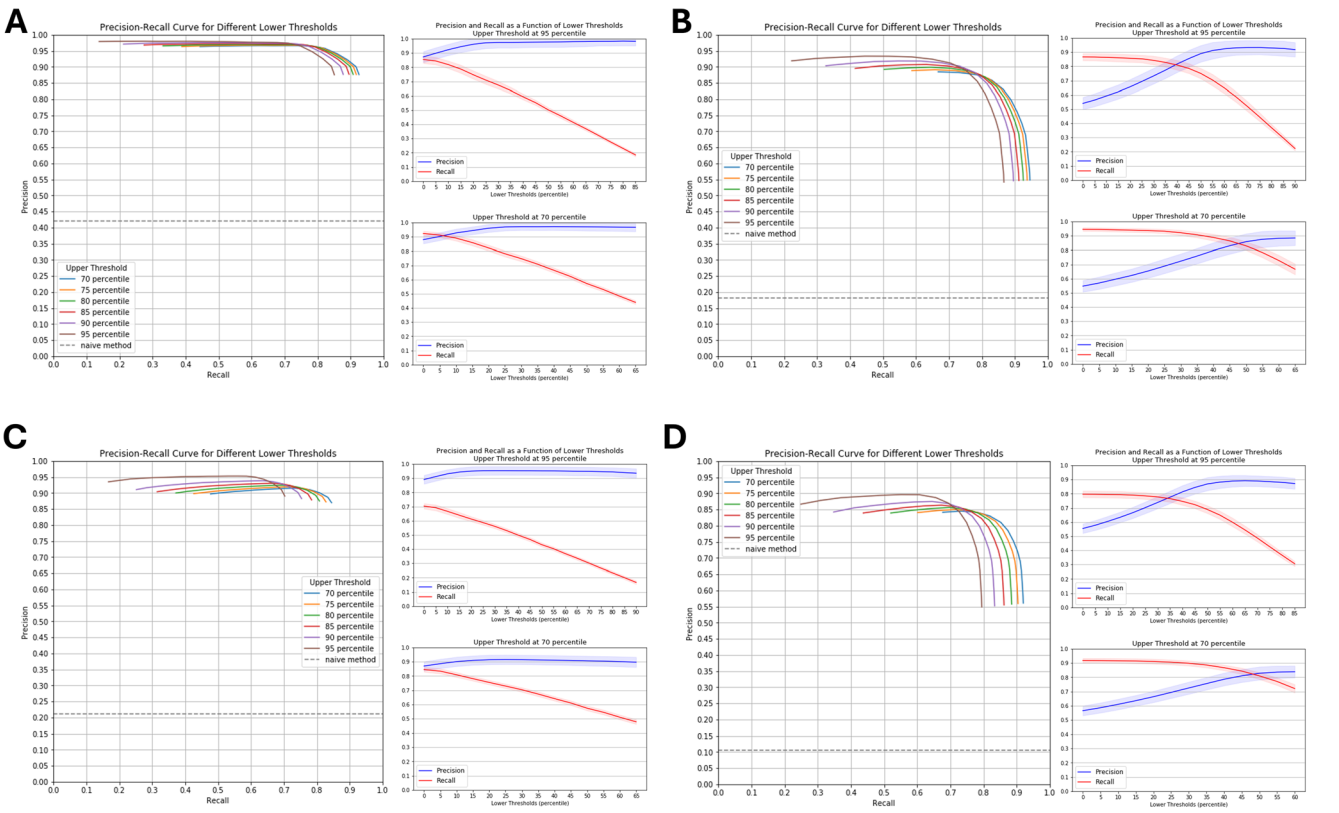

Fig. 3 Precision-recall (red) curves evaluate the graph partitioning algorithm using varying T1 and T2 thresholds. Each curve represents a constant T1 with varying T2 values in intervals of 5 percentile, with the lowest T2 being 0 percentile and highest being T1 – 5 percentile, from right to left of each curve. For the highest and lowest T1 values, precision and recall are plotted as a function of T2 value. Crossover points correspond to the T2 value that represent a compromise between precision and recall. Plots are generated for four types of samples, for each the curves are plotted using the average value over $N = 50$ individual samples. (A) Mixture of 5 peptides with all molecular ions identified and with low background noise, (B) mixture of 10 peptides with only 5 molecular ions identified and high background noise, (C) mixture of 15 peptides with all the molecular ion identified and low background noise, and (D) mixture of 20 peptides with only 10 molecular ions identified and high background noise.

3.2.3 Evaluation using simulated proteomics data. We tested the performance of the partitioning algorithm against several factors including the number of peptides in the mixture, the number of peptides with identified molecular ions, the relative noise levels, and different combinations of thresholds T1 and T2 (Fig. 3). Results were compared against a "naïve" method, which assigns nodes to peptide groups based solely on path existence, resulting in a recall of one.

As the complexity of the mixture increases, the precision of the naïve method drops significantly, consistent with the increased number of overlapping nodes as more species are introduced. In contrast, the precision of data partitioned using the graph partitioning algorithm remains markedly higher than that of the naïve method, even in cases where there are high levels of noise – conditions that can be detrimental to traditional 1D-MS sequencing methods.[28,65,66]

Generally, there is a clear trade-off between precision and recall as T1 and T2 vary. Increasing the T1 threshold improves precision but reduces recall. Similarly, raising T2 increases precision, but its effect plateaus before recall drops rapidly. Optimal threshold values depend on sample characteristics, including the number of peptides in the mixture, the ratio of peptides with identifiable molecular ions to unidentified signals, and the overall presence of noise and contaminants. For example, in high-noise samples, setting higher T2 values helps filter interference and increase precision without significantly sacrificing recall.

Threshold selection also depends on the goal of the experiment. For targeted studies using database matching, precision should be prioritized over recall due to the inherent specificity of 2D MS/MS data. Preliminary surveys therefore can be conducted to determine the minimum recall needed to accurately retrieve the peptide sequence, thereby maximizing precision. In contrast, for *de novo* sequencing of unknown/novel structures, balancing precision and recall is necessary as more information is required to improve the accuracy of structural assignment. Incorrect node assignments can often be eliminated based on the chemist's expertise, which can in turn be incorporated into the *de novo* sequencing program, allowing for lower precision to be tolerated (Fig. S17).

3.3. *De novo* peptide sequencing using 2D MS/MS data

Construction of the fragmentation tree provides an opportunity for *de novo* sequencing, which is advantageous over database searches for analysis of unknown or novel peptide structures. Many traditional *de novo* sequencing algorithms rely implicitly on graph theory, where fragmentation pathways are constructed by connecting m/z values (nodes) when their mass difference corresponds to an amino acid residue.[13,25,26,67,68] As these connections are presumed from the observed mass differences, the method is vulnerable to isobaric interference and thus heavily dependent on mass resolution. Handling PTMs, especially unknown PTMs, is especially challenging as these exponentially expand the database size and search space.[69,70] This often requires the user to define expected modifications to constrain the search space and reduce computation time. By leveraging "true" neutral loss values, traditional methods can be improved to enable the recovery of PTMs information without the need for preexisting databases. By placing constraints on the size of neutral loss values, one can freely incorporate PTM information without exponential expansion of the fragment library. As PTM information is defined by mass difference, unknown PTMs can be identified by adding theoretical mass differences and rerunning the sequencing algorithm, thus facilitating discovery of unknown PTMs.

Fig. 2 presents an example of automated sequencing of a peptide with three different modifications using unit-resolution 2D MS/MS data and following the three-step sequencing algorithm described earlier. In this case, only one pathway (bolded in Fig. 2A) was needed to successfully retrieve the original peptide sequence. Note that certain isobaric ions inherent to unit mass resolution, such as differentiating *I vs. L* or *K vs. Q*, persist in the absence of further distinguishing fragments. Another example, examining the step-by-step sequencing of YGGFL, which requires multiple MS^n pathways, can be found in the SI, Fig. S5–S7.

The proposed sequencing algorithm is especially useful for applications involving the exploration of novel peptide structures, as it is grounded in fundamental chemical principles of fragmentation. This approach enables the integration of chemical knowledge and insights throughout the sequencing process, allowing chemists to actively engage with and refine their understanding of the underlying chemistry. In contrast to "black box" methods, which provide results with little transparency into the decision-making process, this algorithm offers clarity and fosters a deeper connection to the chemistry driving the analysis.

3.4. Proof-of-concept experiments

We conducted a series of experiments to further explore the performance of graph partitioning on simple mixtures of short-chain peptides and the capabilities of the sequencing algorithm in elucidating peptide structures from MS^n information. The 2D MS/MS spectrum of each mixture was generated by combining multiple spectra collected using different IS-CID energies and scan parameters. This approach is crucial for graph partitioning, as recovering the direct relationship between lower mass ions and the molecular ion allows more accurate assignment of the low mass ions to the intact molecule.

3.4.1 Mixtures of structurally related peptides. One of the challenges in studying peptide mixtures is examining peptides with similar core structures but differing modifications, resulting in convoluted MS/MS profiles, which often contain a relatively high percentage of fragments generated from multiple species. To illustrate, we examined a sample derived from the phosphorylation of a pentapeptide, Ac-IYGEF-NH_2 (Fig. S8A). Two molecular ions were readily identified, each accompanied by loss of a neutral of 17 Da. This was particularly evident for the ion at m/z 749, accompanied by m/z 732. Both ions exhibited almost identical product ion profiles, except for a few fragments such as m/z 350. This suggests the presence of a terminal modification, causing the two species to share the

same b-ion series but differing in their y-ions, or *vice versa*. Additionally, the relatively low intensity of the *m*/*z* 669 and *m*/*z* 652 pair indicates that the phosphorylation process was incomplete. Signal extraction followed by the use of the graph partitioning algorithm was applied to the spectrum (Fig. S8B), resulting in three subsets of nodes, those that belong to either peptide and those that belong to both peptides (Fig. S8D). Fragmentation trees of each peptide were constructed (Fig. S4C), and the peptide of *m*/*z* 749 was sequenced using the sequencing algorithm as previously shown in Fig. 2. Presumably, without prior knowledge of the mixture, the structure of *m*/*z* 669 could be easily inferred from the structure of *m*/*z* 749 based on the presence of MS^n pathways in the former's fragmentation tree that are similar but 80 Da apart from those found in the latter. Alternatively, the structure of *m*/*z* 669 could be quickly elucidated using the sequencing algorithm as well (Fig. S9A).

To further examine the capability of the graph partitioning algorithm in separating heavily convoluted signals, we analyzed a mixture of three enkephalin analogs using IS-CID 2D MS/MS (Fig. 4). Given their highly similar sequences, with only one or two differing amino acids, considerable overlap among the fragmentation trees of each peptide was anticipated. Specifically, this includes shared "nodes" resulting from overlapping structures, and both fully and partially overlapping pathways. For example, YGGFM differs from YGGFL by an N-terminal residue, meaning they share b-ion fragments, while pathways involving y-ion fragments are offset by 18 Da. Similarly, YAGFL and YGGFL, would share b-ion pathways up to the b2 ion, and y-ion pathways from y3 ions onward. This example illustrates that the partitioning algorithm can effectively manage samples with a high degree of shared nodes among the component peptides. Finally, we applied the graph partitioning algorithm to recover fragmentation trees for each peptide (Fig. 4B). The results indicate that all expected shared nodes and pathways were accurately assigned to the fragmentation trees. Partially shared pathways were reconstructed from the trees and color-coded as shown in Fig. 4B and D.

3.4.2 Complex peptide mixtures. A mixture of three peptides with overlapping stairstep patterns, arising from both isobaric ions and shared partial structures, was re-processed (Fig. S10) using the PageRank-based method described above. Individual stairstep patterns recovered from known fragmentation pathways of each component peptide could be reconstructed using only their molecular mass as inputs. In addition, *m*/*z* values belonging exclusively to one peptide, two peptides, or all three peptides could be correctly identified (Fig. S9).

The IS-CID 2D-MS/MS spectrum of a five-peptide mixture was recorded (Fig. 5), and the graph partitioning algorithm was used to recover fragmentation trees for each of the five components

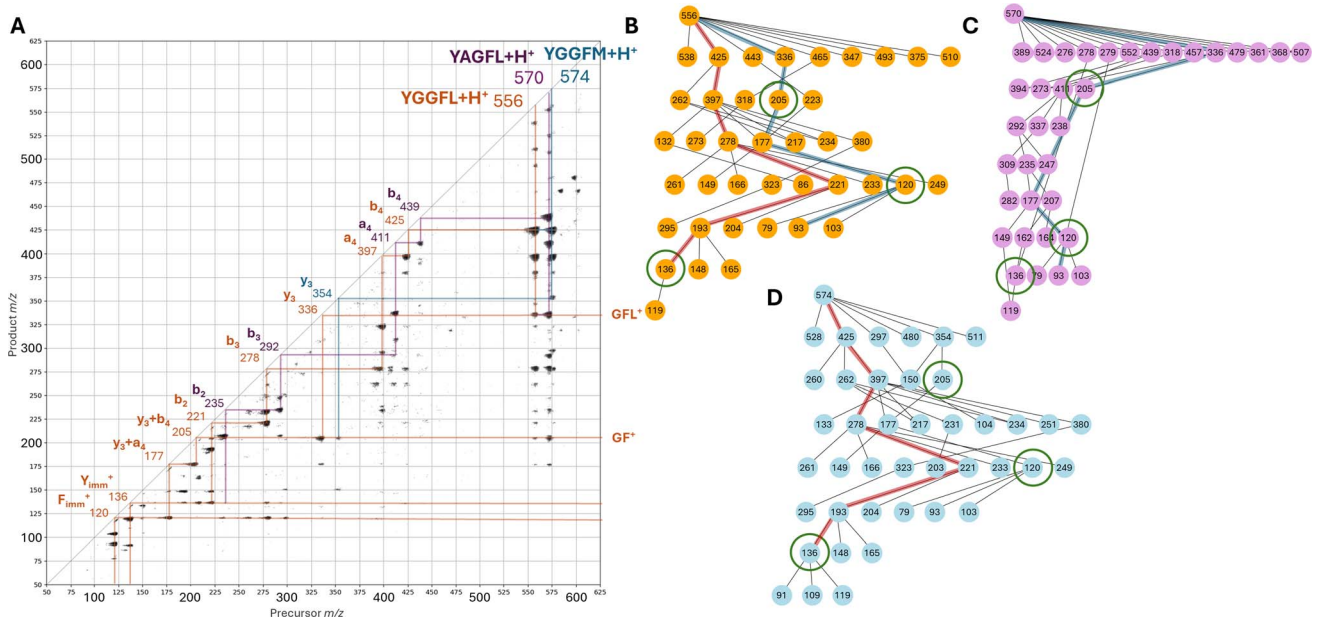

Fig. 4 (A) 2D MS/MS spectrum of a three-peptide mixture containing different enkephalin analogs. Graph partitioning algorithm was used to deconvolute the signals and reconstruct fragmentation trees for the individual peptides, from which pathways were extracted to map stairstep patterns. Due to overlapping structures, multiple fully and partially shared pathways are observed. Orange stairstep patterns represent the two main fragmentation pathways of YGGFL: (1) M + H+ →b4 → a4 → b3 → b2 → iminium ion and (2) M + H^+ → y_3 → y_3 + b_4 → y_3 + a_4 → immonium ion. Purple and blue colored patterns represent the partially diverged patterns from the orange pathways due to the mass difference and position of the differing amino acid of YAGFL and YGGFM in comparison to YGGFL. Fragmentation trees of each peptide were reconstructed from the results of applying graph partitioning algorithms to the 2D MS/MS of the three enkephalin analog mixture: (B) YGGFL with molecular ion of *m*/*z* 556, (C) YAGFL with molecular ion of *m*/*z* 570, and (D) YGGFM with molecular ion of *m*/*z* 574. Nodes representing *m*/*z* values of ions generated by all three peptides are correctly assigned to all three fragmentation trees, such as *m*/*z* 136, *m*/*z* 120, *m*/*z* 205, as circled. Pathway 1, bolded in red, shared by YGGFL and YGGFM, can be extracted from both fragmentation tree (but not in YAGFL's). Pathway 2, bolded in dotted line, can be found in both YGGFL's and YAGFL's and partially in YGGFM's.

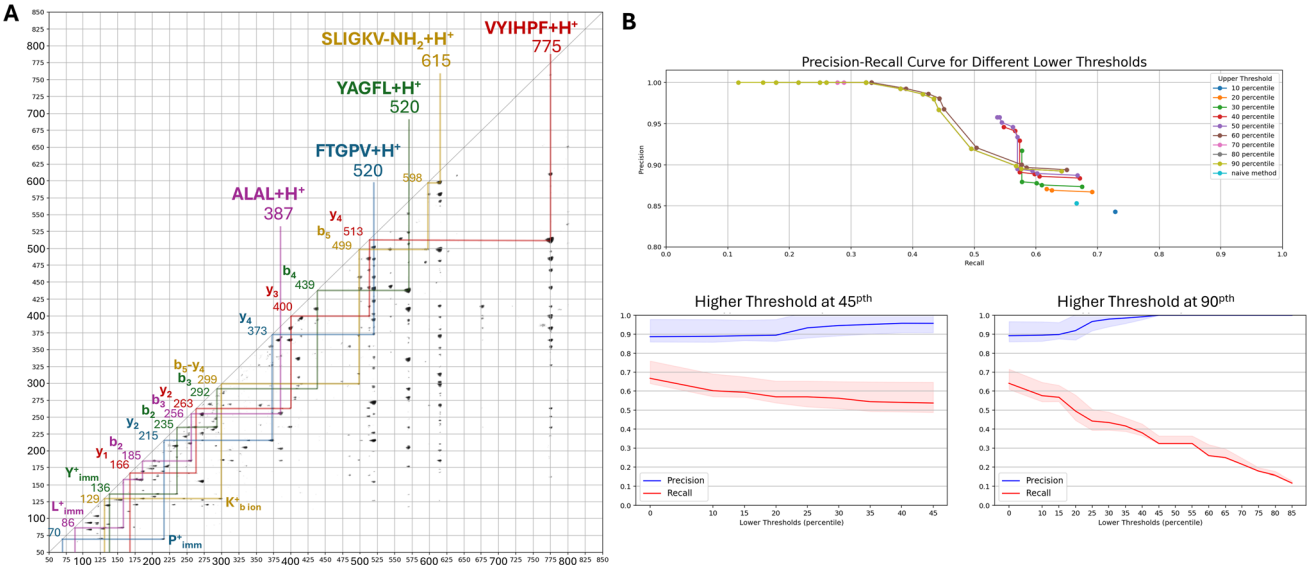

Fig. 5 (A) 2D MS/MS with annotation of a five-peptide mixture. Stairstep patterns were recovered from the graph partitioning algorithm, while structures and ion types were labelled with support from the results of *de novo* sequencing. (B) Precision-recall curve evaluates the performance of the graph clustering algorithm on a five-peptide mixture with varying T1 and T2 thresholds. Each curve represents a constant T1, with T2 values varying in intervals of 5 percentile. T2 ranges from 0 percentile (lowest) to T1 − 5 percentile (highest), progressing from right to left along each curve. Precision and recall are plotted as a function of T2 for T1 values of 90 percentile and 50 percentiles.

(Fig. S12). Using data from the 2D MS/MS spectrum of individual peptides as the 'ground truth' (Fig. S11), we plotted the precision-recall curve for the partitioning as a function of different thresholds (Fig. 5B) to evaluate the effect of varying T1 and T2 for a relatively simple mixture with low noise. The precision trend resembles those observed in the simulated data, specifically in the cases where the molecular ions of all peptides in the mixture are identified (Fig. 3A and C). The recall peaked at approximately 75% for the naïve method instead of 100%. This difference can be attributed to signals present in the individual spectra that are absent in the mixture spectrum, resulting in the loss of precursor-product relationship information among certain pairs of m/z values. Consequently, some of the m/z values, despite belonging to specific peptides, could not be linked to the molecular ion. The naive method achieved an 80% precision, which is relatively high compared to the more complicated mixtures of longer peptides observed in the simulation. This observation is consistent with 20–25% of the signals observed in the spectrum being products of multiple peptides at once (shared nodes).

The threshold pair of the 50th percentile and 40th percentile for the first and second threshold (Fig. 5B), respectively, was selected for partitioning and fragmentation tree construction, as it maximizes both precision and recall. Isobaric ion occurrences (which increase in frequency for unit-resolution data), however, cannot be distinguished in the absence of differentiating fragments, a limitation best illustrated in the case of SLIGKV-NH2 (Fig. S16). This challenge is inherent to CID fragmentation and low mass resolution and could be addressed by combining different fragmentation methods or utilizing data of higher resolution. Selected stairstep patterns for each peptide in the mixture are shown and color-coded. These patterns were identified using the fragmentation trees resulting from the graph partitioning algorithm (Fig. S12), while the ion type annotations were determined using the sequencing algorithm (Fig. S13–S17) and applied to the dataset in Fig. 5. Together with previously discussed data, these results demonstrate that the proposed data analysis framework enables the deconvolution of complex unit-resolution 2D MS/MS spectra of peptide mixtures and the reconstruction of their individual fragmentation trees. Supported by the *de novo* sequencing algorithm, we were able to recover the sequences of all five peptides in the mixture using the fragmentation trees recovered from the partitioning algorithm. This highlights the potential for further development of sequencing software capable of elucidating longer and more complex peptide sequences, with promising applications in challenging tasks such as identifying and characterizing unknown PTMs, higher-order structure determinations, as well as drug-binding studies.

4 Conclusions

We propose a method that leverages the capabilities of 2D MS/MS to preserve precursor-product ion relationships, along with a data analysis framework designed to efficiently exploit this information as an alternative approach to traditional MS-based biopolymer analysis. In doing so, we demonstrate how certain challenges associated with conventional 1D MS workflows can be circumvented. Specifically, we show how this method can be applied to structural analysis and *de novo* sequencing of peptide mixtures. Our sequencing approach still relies on the ion types most commonly observed under CID (b-, y-, and to a lesser extent a-ions). The difference lies in how ion relationships are established. Conventional *de novo* sequencing algorithms

typically infer connectivity from mass differences alone, whereas our framework uses the explicit precursor-product connectivity preserved in 2D MS/MS data. This enables us to confirm whether two candidate ions are truly linked by a fragmentation event, providing an added layer of certainty. In practice, this should reduce misassignments, mitigates noise interference, and improve sequencing accuracy—particularly valuable for unknown peptides that are not represented in databases.

While 2D MS/MS ion trap instrumentation is still being optimized to achieve the resolution and processing speed required for modern proteomics applications, this method holds significant promise for studying biopolymer mixtures in general, particularly for structural analysis tasks in synthesis, functionalization, and degradation studies. Specifically, the graph-theory-based data analysis framework described here is agnostic to mass analyzer – if equivalent precursor-product maps are generated, for example, with FT-ICR 2D MS, data-dependent acquisition (DDA) or data-independent acquisition (DIA) with narrow isolation windows, the same graph-partitioning and sequencing algorithms can be applied. Such an implementation would extend the benefits of high-resolution methods to the 2D MS/MS framework and potentially open more advanced applications of network analysis, such as screening for unknown PTMs in terms of both types and locations, or the automated identification of molecular ions through graph partitioning algorithms. Furthermore, the core of the *de novo* sequencing algorithm can be further enhanced by integrating established sequencing software and databases, which would strengthen the graph partitioning and structural analysis capabilities of the network analysis framework. This method therefore provides a novel perspective on handling 2D MS/MS data and highlights opportunities to refine and enhance MS-based techniques for broader applications in biopolymer structural studies.

Author contributions

MPL designed and carried out the experiments, analyzed the data, developed the *de novo* sequencing algorithms, and drafted the manuscript. YZ contributed to the research ideas, wrote the data simulation, and co-developed the graph partitioning algorithms. ETD and DTH contributed to 2D MS/MS instrument development, data acquisition and processing, and edited the manuscript. DFG and RGC provided guidance, critical insights, and led the investigation.

Conflicts of interest

There are no conflicts to declare.

Abbreviations

2D MS/MS	Two-Dimensional Tandem Mass Spectrometry
2D-PC-MS	Two-Dimensional Partial Covariance Mass Spectrometry
CID	Collision-Induced Dissociation
DDA	Data Dependent Acquisition
DIA	Data Independent Acquisition
FC-MS	Fragment Correlation Mass Spectrometry
FT-ICR	Fourier Transform Ion Cyclotron Resonance
IS-CID	In-Source Fragmentation
MRM	Multiple Reaction Monitoring
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
MS^n	Multi-Stage Mass Spectrometry
NL	Neutral Loss
PPR	Personalized PageRank
PTM	Post-Translational Modification

Data availability

The scripts for the data simulation section of this article are openly available at **https://github.com/MyPhuongLe/2DMSMSPeptideSimulation**.

The data supporting this article have been included as part of the SI. Supplementary Information is available. See DOI: **https://doi.org/10.1039/d5sc03762j**.

Acknowledgements

The work at Purdue was supported in part by a grant from the National Center for Advancing Translational Sciences (grant 5UG3TR004139). DFG and YZ acknowledge support from the Department of Energy (DE-SC0023162). The authors acknowledge Dr Lingqi Qiu for insights and assistance with the *de novo* sequencing validation experiment.

References

1 X. Dai and L. Shen, *Front. Med.*, DOI: **10.3389/fmed.2022.911861**.
2 V. G. Zaikin and R. S. Borisov, *J. Anal. Chem.*, 2021, **76**, 1567–1587.
3 K. Biemann, *Biomed. Environ. Mass Spectrom.*, 1988, **16**, 99–111.
4 V. H. Wysocki, K. A. Resing, Q. Zhang and G. Cheng, *Methods*, 2005, **35**, 211–222.
5 X. Wei and L. Li, *Int. J. Clin. Exp. Pathol.*, 2008, **2**, 132–148.
6 J. S. Cottrell, *J. Proteomics*, 2011, **74**, 1842–1851.
7 H. Chen, K. Tabei and M. M. Siegel, *J. Am. Soc. Mass Spectrom.*, 2001, **12**, 846–852.
8 F. Hannauer, R. Black, A. D. Ray, E. Stulz, G. J. Langley and S. W. Holman, *Rapid Commun. Mass Spectrom.*, 2023, **37**, e9596.
9 P. Roepstorff and J. Fohlman, *Biomed. Mass Spectrom.*, DOI: **10.1002/bms.1200111109**.
10 V. H. Wysocki, G. Tsaprailis, L. L. Smith and L. A. Breci, *J. Mass Spectrom.*, 2000, **35**, 1399–1406.
11 M. Girod, D. Arquier, A. Helms, K. Juetten, J. S. Brodbelt, J. Lemoine and L. MacAleese, *J. Am. Soc. Mass Spectrom.*, 2024, **35**, 1040–1054.

12 J. M. Mata, E. van der Nol and S. J. Pomplun, *J. Am. Chem. Soc.*, 2023, **145**, 19129–19139.
13 C. C. A. Ng, Y. Zhou and Z.-P. Yao, *Anal. Chim. Acta*, 2023, **1268**, 341330.
14 N. H. Tran, R. Qiao, L. Xin, X. Chen, C. Liu, X. Zhang, B. Shan, A. Ghodsi and M. Li, *Nat. Methods*, 2019, **16**, 63–66.
15 M. Yilmaz, W. E. Fondrie, W. Bittremieux, C. F. Melendez, R. Nelson, V. Ananth, S. Oh and W. S. Noble, *Nat. Commun.*, 2024, **15**, 6427.
16 K. Verheggen, H. Ræder, F. S. Berven, L. Martens, H. Barsnes and M. Vaudel, *Mass Spectrom. Rev.*, 2020, **39**, 292–306.
17 M. Grabarics, M. Lettow, C. Kirschbaum, K. Greis, C. Manz and K. Pagel, *Chem. Rev.*, 2022, **122**, 7840–7908.
18 Y. S. Ting, J. D. Egertson, S. H. Payne, S. Kim, B. MacLean, L. Käll, R. Aebersold, R. D. Smith, W. S. Noble and M. J. MacCoss, *Mol. Cell. Proteomics*, 2015, **14**, 2301–2307.
19 C.-Y. Yen, S. Houel, N. G. Ahn and W. M. Old, *Mol. Cell. Proteomics*, DOI: **10.1074/mcp.M111.007666**.
20 K. L. Schey, J. M. Luther and K. L. Rose, *Methods*, 2015, **87**, 75–82.
21 L. Fields, M. Ma, K. DeLaney, A. Phetsanthad and L. Li, *Proteomics*, 2024, **24**, 2300285.
22 R. B. Kitata, J.-C. Yang and Y.-J. Chen, *Mass Spectrom. Rev.*, 2023, **42**, 2324–2348.
23 W. Peng, C. D. Gutierrez Reyes, S. Gautam, A. Yu, B. G. Cho, M. Goli, K. Donohoo, S. Mondello, F. Kobeissy and Y. Mechref, *Mass Spectrom. Rev.*, 2023, **42**, 577–616.
24 B. C. Orsburn, *Proteomes*, 2021, **9**, 15.
25 B. Behsaz, H. Mohimani, A. Gurevich, A. Prjibelski, M. Fisher, F. Vargas, L. Smarr, P. C. Dorrestein, J. S. Mylne and P. A. Pevzner, *cels*, 2020, **10**, 99–108e5.
26 A. M. Frank, M. M. Savitski, M. L. Nielsen, R. A. Zubarev and P. A. Pevzner, *J. Proteome Res.*, 2007, **6**, 114–123.
27 D. V. Petrovskiy, K. S. Nikolsky, L. I. Kulikova, V. R. Rudnev, T. V. Butkova, K. A. Malsagova, A. T. Kopylov and A. L. Kaysheva, *Sci. Rep.*, 2024, **14**, 15000.
28 K. McDonnell, E. Howley and F. Abram, *Comput. Struct. Biotechnol. J.*, 2022, **20**, 1402–1412.
29 D. Schulte and J. Snijder, *J. Proteome Res.*, 2024, **23**, 3552–3559.
30 J. S. Brodbelt, *Curr. Opin. Chem. Biol.*, 2022, **70**, 102180.
31 S. Ramazi and J. Zahiri, *Database*, 2021, baab012.
32 A. Michalski, J. Cox and M. Mann, *J. Proteome Res.*, 2011, **10**, 1785–1793.
33 F. Kryuchkov, T. Verano-Braga, T. A. Hansen, R. R. Sprenger and F. Kjeldsen, *J. Proteome Res.*, 2013, **12**, 3362–3371.
34 R. Peckner, S. A. Myers, A. S. V. Jacome, J. D. Egertson, J. G. Abelin, M. J. MacCoss, S. A. Carr and J. D. Jaffe, *Nat. Methods*, 2018, **15**, 371–378.
35 C. W. I. Ross, S. Guan, P. B. Grosshans, T. L. Ricca and A. G. Marshall, *J. Am. Chem. Soc.*, 1993, **115**, 7854–7861.
36 M. A. van Agthoven, A. M. Lynch, T. E. Morgan, C. A. Wootton, Y. P. Y. Lam, L. Chiron, M. P. Barrow, M.-A. Delsuc and P. B. O'Connor, *Anal. Chem.*, 2018, **90**, 3496–3504.
37 M. A. van Agthoven, Y. P. Y. Lam, P. B. O'Connor, C. Rolando and M.-A. Delsuc, *Eur. Biophys. J.*, 2019, **48**, 213–229.
38 T. Driver, V. Averbukh, L. J. Frasiński, J. P. Marangos and M. Edelson-Averbukh, *Anal. Chem.*, 2021, **93**, 10779–10788.
39 T. Driver, B. Cooper, R. Ayers, R. Pipkorn, S. Patchkovskii, V. Averbukh, D. R. Klug, J. P. Marangos, L. J. Frasinski and M. Edelson-Averbukh, *Phys. Rev. X*, 2020, **10**, 041004.
40 Y. Li, G. Cavet, R. N. Zare and T. Driver, *Proc. Natl. Acad. Sci. U. S. A.*, 2024, **121**, e2409676121.
41 Y. Li, G. Cavet, R. N. Zare and T. Driver, *Anal. Chem.*, 2024, **96**, 15081–15084.
42 L. J. Szalwinski, D. T. Holden, N. M. Morato and R. G. Cooks, *Anal. Chem.*, 2020, **92**, 10016–10023.
43 M. T. Le, D. T. Holden, J. M. Manheim, E. T. Dziekonski, K. Iyer and R. Graham Cooks, *Angew. Chem.*, 2024, **136**, e202315904.
44 T. Lin and G. L. Glish, *Anal. Chem.*, 1998, **70**, 5162–5165.
45 L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner and R. Aebersold, *Mol. Cell. Proteomics*, 2012, **11**, 016717.
46 J. Y. Yang, L. M. Sanchez, C. M. Rath, X. Liu, P. D. Boudreau, N. Bruns, E. Glukhov, A. Wodtke, R. de Felicio, A. Fenner, W. R. Wong, R. G. Linington, L. Zhang, H. M. Debonsi, W. H. Gerwick and P. C. Dorrestein, *J. Nat. Prod.*, 2013, **76**, 1686–1699.
47 J. Watrous, P. Roach, T. Alexandrov, B. S. Heath, J. Y. Yang, R. D. Kersten, M. van der Voort, K. Pogliano, H. Gross, J. M. Raaijmakers, B. S. Moore, J. Laskin, N. Bandeira and P. C. Dorrestein, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, E1743–E1752.
48 C. Bartels, *Biomed. Environ. Mass Spectrom.*, 1990, **19**, 363–368.
49 F. Rasche, A. Svatoš, R. K. Maddula, C. Böttcher and S. Böcker, *Anal. Chem.*, 2011, **83**, 1243–1251.
50 S. Böcker and F. Rasche, *Bioinformatics*, 2008, **24**, i49–i55.
51 M.-Y. Zhang, N. Pace, E. H. Kerns, T. Kleintop, N. Kagan and T. Sakuma, *J. Mass Spectrom.*, 2005, **40**, 1017–1029.
52 R. E. March, *Mass Spectrom. Rev.*, 2009, **28**, 961–989.
53 J. D. Venable, M.-Q. Dong, J. Wohlschlegel, A. Dillin and J. R. Yates, *Nat. Methods*, 2004, **1**, 39–45.
54 G. Xu, J. Stupak, L. Yang, L. Hu, B. Guo and J. Li, *Rapid Commun. Mass Spectrom.*, 2018, **32**, 763–774.
55 B. C. Searle, L. K. Pino, J. D. Egertson, Y. S. Ting, R. T. Lawrence, B. X. MacLean, J. Villén and M. J. MacCoss, *Nat. Commun.*, 2018, **9**, 5128.
56 C.-C. Tsou, D. Avtonomov, B. Larsen, M. Tucholska, H. Choi, A.-C. Gingras and A. I. Nesvizhskii, *Nat. Methods*, 2015, **12**, 258–264.
57 Y. Y. Lu, J. Bilmes, R. A. Rodriguez-Mias, J. Villén and W. S. Noble, *Bioinformatics*, 2021, **37**, i434–i442.
58 K. F. Medzihradszky and R. J. Chalkley, *Mass Spectrom. Rev.*, 2015, **34**, 43–63.
59 B. Spengler, *J. Am. Soc. Mass Spectrom.*, 2004, **15**, 703–714.
60 L. Page, S. Brin, R. Motwani and T. Winograd, 1999.
61 D. F. Gleich, *SIAM Rev.*, 2015, **57**, 321–363.
62 J. González-Gomariz, G. Serrano, C. M. Tilve-Álvarez, F. J. Corrales, E. Guruceaga and V. Segura, *J. Proteome Res.*, 2020, **19**, 4795–4807.

63 J. D. Hogan, J. Wu, J. A. Klein, C. Lin, L. Carvalho and J. Zaia, *Mol. Cell. Proteomics*, 2021, **20**, 100093.
64 M. A. Pethe, A. B. Rubenstein and S. D. Khare, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 168–176.
65 S. L. Hubler, P. Kumar, S. Mehta, C. Easterly, J. E. Johnson, P. D. Jagtap and T. J. Griffin, *J. Proteome Res.*, 2020, **19**, 161–173.
66 A. J. Seneviratne, S. Peters, D. Clarke, M. Dausmann, M. Hecker, B. Tully, P. G. Hains and Q. Zhong, *Bioinformatics*, 2021, **37**, 4719–4726.
67 Y. Yan, S. Zhang and F.-X. Wu, *Proteome Sci.*, 2011, **9**, S17.
68 Z. Mao, R. Zhang, L. Xin and M. Li, *Nat. Mach. Intell.*, 2023, **5**, 1250–1260.
69 P. Minguez, I. Letunic, L. Parca and P. Bork, *Nucleic Acids Res.*, 2013, **41**, D306–D311.
70 M. R. Shortreed, C. D. Wenger, B. L. Frey, G. M. Sheynkman, M. Scalf, M. P. Keller, A. D. Attie and L. M. Smith, *J. Proteome Res.*, 2015, **14**, 4714–4720.
71 D. T. Snyder, L. J. Szalwinski, Z. St. John and R. G. Cooks, *Anal. Chem.*, 2019, **91**, 13752–13762.
72 T. N. J. Fouquet, R. B. Cody and L. Charles, *Mass Spectrom. Rev.*, 2025, 1–24.
73 R. W. Kondrat, G. A. McClusky and R. G. Cooks, *Anal. Chem.*, 1978, **50**, 2017–2021.
74 T. De Vijlder, D. Valkenborg, F. Lemière, E. P. Romijn, K. Laukens and F. Cuyckens, *Mass Spectrom. Rev.*, 2018, **37**, 607–629.
75 A. A. Ramos, H. Yang, L. E. Rosen and X. Yao, *Anal. Chem.*, 2006, **78**, 6391–6397.
76 K. Simon, in *Graph-Theoretic Concepts in Computer Science*, ed. M. Nagl, Springer, Berlin, Heidelberg, 1990, pp. 245–259.
77 T. H. Haveliwala, in *Proceedings of the 11th international conference on World Wide Web*, Association for Computing Machinery, New York, NY, USA, 2002, pp. 517–526.
78 B. Paizs and S. Suhai, *Mass Spectrom. Rev.*, 2005, **24**, 508–548.