A Fast Neural Network for Isotopic Charge State Assignment

John G. Pavek,[1] Nicholas E. Bollis,[2] Josiah Grimes,[1] Michael R. Shortreed[2], Lloyd M. Smith,[2] and Michael T. Marty[1,*]

[1]Department of Chemistry and Biochemistry, University of Arizona, Tucson, AZ 85721, USA

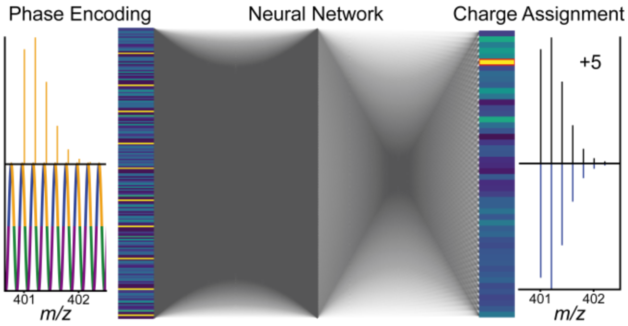
[2]Department of Chemistry, University of Wisconsin-Madison, Madison, WI 53706, USA

*mtmarty@arizona.edu

Abstract

Electrospray ionization (ESI) mass spectrometry is an essential technique for chemical analysis in a range of fields. In ESI, analytes can produce multiple charge states, which must be correctly assigned for identification. Existing approaches to charge state assignment can suffer from limited accuracy and/or poor speed. Here, we developed a fast deep learning neural network to perform isotopic cluster charge assignment. The performance of our algorithm, IsoDec, was demonstrated on top-down proteomics spectra collected on diverse instruments. On these highly complex individual spectra, we found that IsoDec produces similar sequence coverage to existing software tools but with improved accuracy. Importantly, this performance enhancement stems directly from the neural network charge assignment approach, not simply improved scoring and filtering of isotopic clusters. Finally, when applied to large top-down proteomics data sets, we discovered that IsoDec produces proteoform-spectrum matches with a better combination of coverage and accuracy. Overall, IsoDec provides a compelling demonstration of the potential of lightweight neural networks in mass spectrometry data analysis for diverse applications.

For Table of Contents Use Only:



https://doi.org/10.26434/chemrxiv-2025-f2p26 **ORCID:** https://orcid.org/0000-0001-8115-1772 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

Introduction

Electrospray ionization (ESI) mass spectrometry (MS) is an essential technique in chemical analysis, enabling a diverse range of environmental, industrial, and biological applications. One benefit of ESI is that it imparts multiple charges to analytes, which enables larger species to be analyzed on instruments with limited *m/z* ranges. Even on instruments with higher *m/z* ranges, having multiple charge states aids in fragmenting molecules for different applications.[1,2]

However, the presence of different charge states complicates data analysis because the charge must be assigned to determine the mass. For low resolution data, with resolved charge states but unresolved isotope distributions, multiple approaches have been developed to determine the charge from the spacing between adjacent charge states.[3–14] For unresolved data lacking resolved charge state distributions, charge detection-mass spectrometry can be used to directly measure the charge on individual ions.[15,16] For high resolution data with resolved isotope distributions, the spacing between adjacent isotopes is used for charge determination.

Early work from Senko *et al.* used Fourier transforms and Patterson charge maps to assign the charge of isotope clusters.[17] Horn *et al.* incorporated this approach into a sequential algorithm that automatically assigned peaks and then removed assigned clusters from the spectrum.[18] This algorithm, THRASH, has been adapted in many commercial and non-commercial offspring.[19,20] For example, Hardklör used a similar overall approach but developed a simpler algorithm, called QuickCharge, based on inverse *m/z* differences to speed up charge assignment.[21]

Another approach to charge assignment is with pattern matching. For example, ProMex scans theoretical isotopic envelopes for each possible charge across a spectrum to identify and assign isotopic envelopes in *m/z* space.[22,23] Other tools use a targeted approach and generate a set of theoretical isotope distributions for predicted fragments, which are then compared against the data.[24,25] In contrast, rather than assigning a charge to isotope clusters, FLASHDeconv implements log-transformation and pattern scanning to identify groups of isotope peaks with the same mass but different charges.[26]

One common problem with tools for isotopic charge state assignment has been low accuracy. Tabb *et al.* found that different charge assignment algorithms gave significant heterogeneity in top-down proteomics searches.[27] Examining manually annotated data with different algorithms and settings, McIlwain *et al.* found that there were significant tradeoffs between precision and recall, and some peaks were never assigned correctly.[28] These studies demonstrate the need for new approaches for charge assignment in complex mass spectrometry data.

To address these challenges, we developed a deep learning neural network model for charge state assignment. Machine learning (ML) and artificial intelligence (AI) approaches have been used in MS data analysis for a range of applications.[29,30] For example, AI/ML has been used in imaging MS segmentation[30–32] and in peptide fragmentation prediction.[33] It has also been used for isotope cluster grouping[34,35] and isotope envelope scoring.[36,37] However, we are not aware of any approaches using AI/ML for charge state assignment.

Using publicly available top-down proteomics (TDP) data sets, we assembled a training set of over 6 million assigned isotope clusters and trained a simple neural network to assign the charge of these clusters. We have incorporated our charge state assignment neural network into a broader workflow to analyze top-down proteomics data. Here, we demonstrate that our neural network approach matches the coverage of existing tools but produces fewer unassigned fragments and greater speed per feature identified.

Experimental Section

Training and Test Data

To enable our deep learning approach, we downloaded TDP data files from MassIVE and the Proteomics Data Exchange. Full details on the training data used are provided in the Supplemental Methods (SM1). Additional data files were downloaded to test the full workflow and were not included in either the training or testing of the neural network. For these workflow tests, we first selected three averaged MS2 spectra from top-down proteomics experiments: ultraviolet photodissociation (UVPD)-fragmented carbonic anhydrase II on an Orbitrap Fusion Lumos, electron-transfer dissociation (ETD)-fragmented carbonic anhydrase II on a 21 T FT-ICR, and electron-capture dissociation (ECD)-fragmented apomyoglobin on an Agilent 6560c Q-ToF.[38–40] These spectra were selected to provide a rich dataset of top-down fragments that could be compared between instruments and fragmentation methods. Beyond these isolated spectra, we also selected a set of six replicate *E. coli* top-down LC-MS/MS injections to assess speed, reproducibility, and performance of IsoDec in the context of a top-down proteomics database search with data with multiple mixed MS1 and MS2 scans.[41]

Computational Overview

There were four key steps in the development of IsoDec. First, we processed the training data into collections of centroids that represented isotopic clusters of known charge. Second, we transformed each of these small, isolated clusters of data using a mass defect approach into an encoded matrix. Third, we trained the neural network on the encoded training matrices. Fourth, we incorporated the neural network algorithm into a larger workflow to pick peaks, assign charges, and export results.

Processing Training Data into Discrete Isotope Clusters

To train the neural network, we broke the data down into a simple classification problem, where each isotope cluster would be classified as a discrete charge from 1 to 49. Although not implemented yet, the classification of charge 0 has been reserved for pure noise, yielding a classification vector of length 50. It is possible to extend this limit to higher charge states, but we found that it was hard to get examples of charges above 30. There were several charge states above 40 that had very few examples in our training data, with no representatives for z=46 (see Figure S1). It is possible to synthesize higher charge states for training, and preliminary data indicates this is highly effective at improving model performance for charge 30 to 50 (as described below). However, the initial model presented here is primarily limited to below 30.

To assign charge states prior to training, we used a brute force assignment strategy, as described in the Supplemental Methods (SM2). After charge states had been assigned, 90% of the clusters were kept as training data, and 10% were randomly separated into testing data. The result was a large set of narrow, centroided isotopic clusters surrounding charge states of known charge. The charge distribution was heavily skewed toward lower charge states, especially 1, as shown in Figure S1. In the final data collection, we had over 6.1 million assigned isotope clusters in the training data and over 680,000 clusters in the test data.

Mass Defect Encoding

After creating collections of assigned isotope clusters, we next encoded the data using mass defect encoding. The premise of our strategy is that an isotope cluster should have a uniform *m/z* spacing. For peaks that have a uniform spacing, they should have the same mass defect relative to a repeating unit of *1.0033/z*. Here, 1.0033 is used rather than 1 because it is the average isotope spacing, corresponding to roughly the mass difference between ^{13}C and ^{12}C. Another way to describe this effect is that the phase of each cluster is consistent with respect to a wave of frequency *1.0033/z* (Figure 1A and D). To capture this spacing, we summed the intensity from peaks with similar mass defect/phase values.

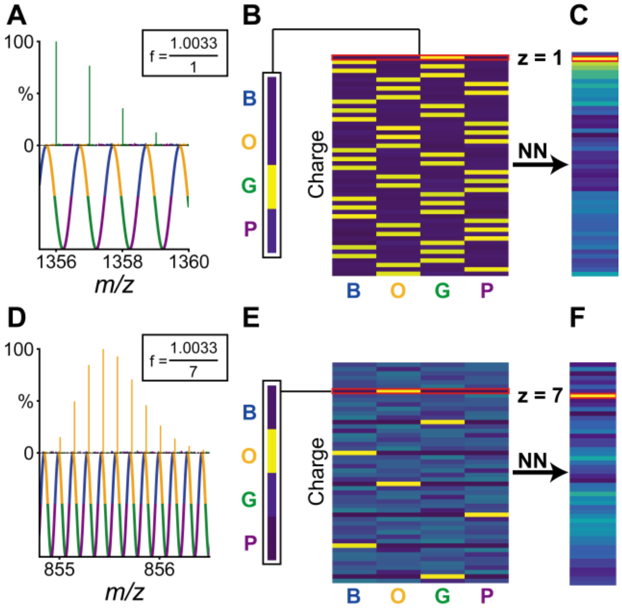


Figure 1. Graphical representation of the mass defect encoding scheme. (**A**) A wave with a frequency of 1.0033/1, separated into four phase bins by color (*blue*, *orange*, *green*, and *purple*), is shown under mock data containing a +1 isotope distribution. Note that each of the isotope peaks fall in the *green* bin, showing that they are in phase with the plotted wave. (**B**) The z=1 row of the full encoding vector, as well as its position in the full encoded matrix (*yellow* = most intense, *purple* = least intense). (**C**) The output charge prediction vector, with a maximum at 1. The same process for a +7 isotope distribution is shown in **D**–**F**. Notice that there are hotspots in the encoded matrix corresponding to +7, but also +14, +21, and subsequent harmonic multiples of 7.

To calculate the phase for each peak at each *z*, we first transformed the experimental *m/z* by multiplying by *z/1.0033*. In this space, each isotope peak is exactly 1 apart when *z* matches the true charge state. To calculate the phase, we simply extract the decimal part of the rescaled *m/z* by taking the value modulo 1 to yield a phase between 0 and 1. Because *m/z* is continuously sampled, we need to bin these phase values. The bin was calculated by multiplying the phase by the total number of bins and taking the floor of the value.

We tested several binning resolutions. Surprisingly, very coarse binning worked just as well as finer binning. After training, using 8 bins yielded similar accuracy to 16 bins. Using 4 bins was slightly worse than 8 bins, but the difference was very small, less than 0.1% in accuracy. Bin sizes of 2 or 3 worked but had more noticeable losses in accuracy. Thus, we chose to train the model at both a 4- and 8-bin size resolution, where the resolution of 4 bins is slightly faster, and 8 bins is slightly more accurate.

After calculating the bin number for each peak at each possible charge state, we added the intensity into a *50×N* matrix (Figure 1B and E), where *N* is the total number of bins. Each row represents a different charge state from 1–50, used in the calculations above. Thus, the phase must be calculated for each *m/z* data point and each possible charge state, but the relatively simple math operations make this process fast. After summing all points into the matrix, there are strong intensities in the matrix for rows where the charge state is equal to the correct charge state or a higher harmonic.

Note, it is not possible to reverse the operation of phase encoding. Information on both the absolute *m/z* and each peak intensity is lost in the process. Thus, this encoding is likely not useful for other applications like segmenting which peaks belong to a cluster. However, we found that this phase/mass defect encoding was both fast and informative for charge state assignments, yielding superior accuracy to other encoding approaches tested.

Training the Neural Network

After converting all the centroid clusters into encoded matrices, the next step was to train the neural network using PyTorch.[42] First, data was loaded and grouped into batches. Although slower, we found that a batch size of 32 for training data was best. Training was performed on the University of Arizona Ocelote High Performance Computer (HPC) system using a single Nvidia P100 graphics processing unit (GPU). Training took under an hour on the HPC, but it could be done overnight on a laptop with a GPU.

The neural network was set up as a simple two-layer model. First, the *50×N* encoding matrix (Figure 1B and E) was flattened into the input vector (Figure 2A). This was then passed through a linear neural network (Figure 2, Operation 1) to a hidden layer vector of the same length, *50×N*. A rectified linear unit (ReLU) operation was then performed on the hidden layer, which simply sets the values below 0 to 0 (Figure 2D). Finally, a second linear network (Figure 2, Operation 2) was used with an output dimension of 50. The location of the highest value in the output vector (Figure 2G) was returned as the charge state.

The advantage of this simple model is its speed and simplicity. By using only linear transformations, it was easy to translate the model into C with simple matrix multiplication operations. The 8-bin model had 180,450 parameters, and the 4-bin model had 50,250 parameters. More complex linear models with more depth and larger hidden layers were tested, but these did not significantly improve accuracy. Other activation functions besides ReLU were also tested, but ReLU was slightly more accurate and faster.

The model was trained for 10 epochs (iterations) over the full set of training data. The loss function was set to cross entropy loss, and the SGD optimizer was used with an initial learning rate of 0.1 and a scheduler that stepped down by a factor of 0.95 each step. After each epoch, the test data was evaluated for accuracy.

To simulate crowded spectra, we developed an approach to synthetically mix two known clusters. First, a primary and secondary cluster were chosen at random from the full data set. The *m/z* of the secondary distribution was shifted to be within the same encoding window as the primary cluster. The intensity of the secondary distribution was normalized randomly to between 5% and 20% of the primary distribution, so that a clear charge state assignment was still present. After testing different amounts of mixed data, we trained against data with 40% synthetic double mixes. Additional alternative training strategies were tested that did not improve the model performance, and they are briefly described in the Supplemental Methods (SM3).

The final model had an accuracy of 99.2% for both the 8-bin and 4-bin versions. After manually examining the mis-assigned data, most of these 0.8% were clusters that were mis-assigned in the original brute force assignment, often with mixed up 1 vs. 2 charge data. Using stricter criteria on the initial charge assignment could help improve this (a strategy we plan on testing in the future), but we suspect that removing lower-quality data that is correctly assigned might make the model less robust.

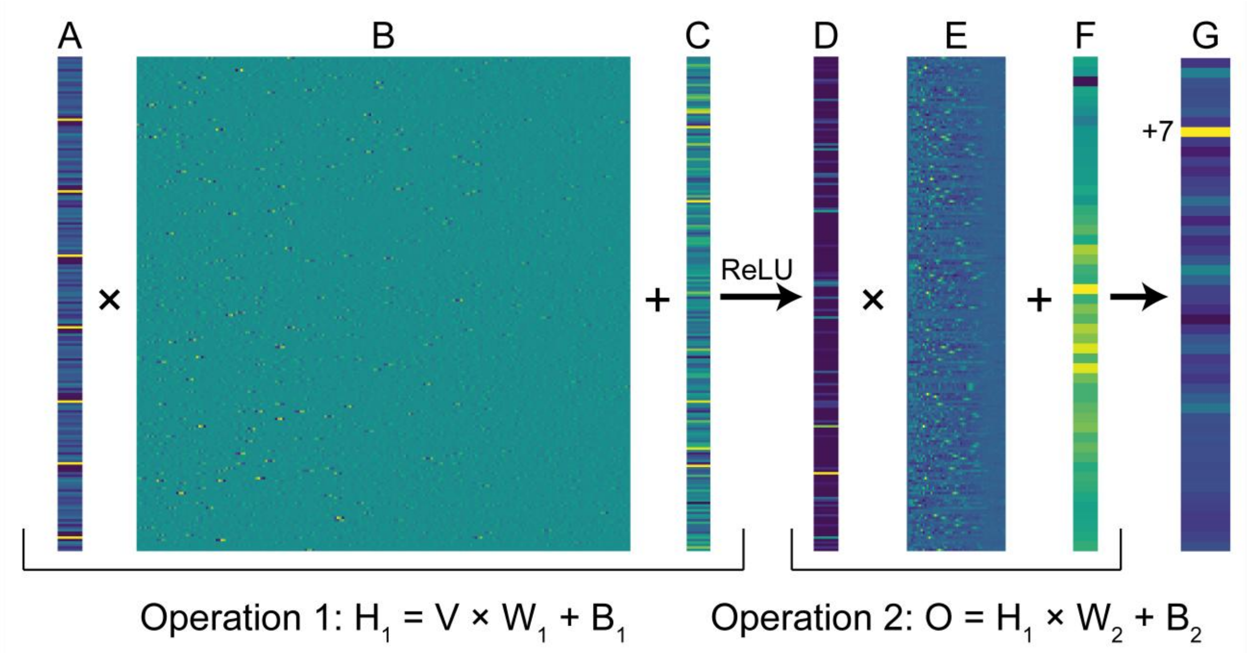

A
B
C
D
E
F
G
×
+
ReLU
×
+
+7
Operation 1: $H_1 = V \times W_1 + B_1$
Operation 2: $O = H_1 \times W_2 + B_2$

Figure 2. Schematic of the neural network operations. The matrices and vectors are visualized from the actual model parameters and an example input: **A**, input vector; **B**, layer 1 weights; **C**, layer 1 bias; **D**, hidden layer 1 after ReLU; **E**, layer 2 weights (transposed for visualization purposes); **F**, layer 2 bias; **G**, output charge prediction vector, the maximum position of which is the assigned charge of +7.

Developing a Workflow for Spectrum Assignment

To integrate this charge assignment neural network into a workflow that can assign an entire spectrum, we use an approach similar to THRASH and related algorithms.[18,20,43,44] First, we centroided the data. Then, we selected clusters from the data. Isolated regions were selected around each cluster, similar to the training data. These isotope clusters were then encoded, and the charge state was predicted by the neural network algorithm. We allowed two possible charge states to be returned from the neural network in cases where the second highest position in the prediction vector is within 95% of the maximum position. A full description of parameters is provided in Table S1.

After predicting the charge state for a cluster (Figure 3, step 1), the next step is to check that the predicted charge matches the data. First, a theoretical isotope distribution for the predicted charge state is generated using a Fourier transform-based algorithm[45] and an averagine model.[43] Note, the averagine model could present problems for non-protein analytes, as discussed below. The predicted *m/z* peaks are then matched against the experimental data. The cosine similarity score is calculated for the match. If the cosine similar score is above 70% and meets other criteria, the peak is considered a match. However, this cutoff can be adjusted by the user.

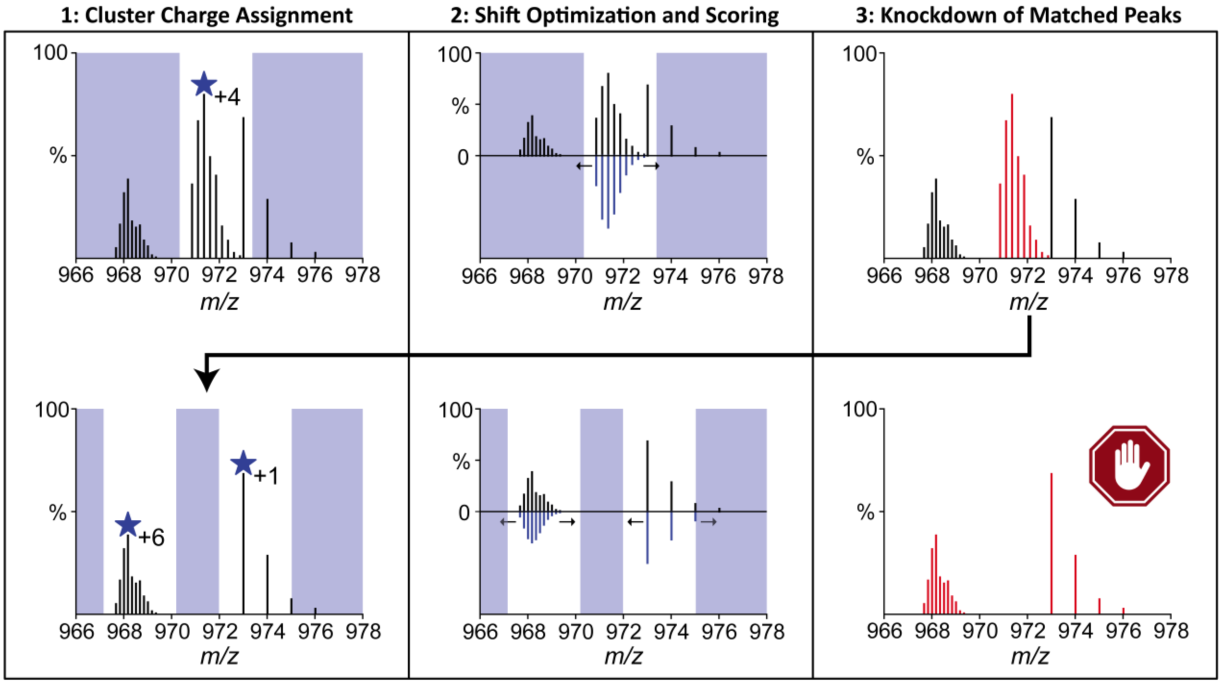


Figure 3. Schematic outline of the full IsoDec workflow. (**Step 1**) Prospective isotope clusters are selected, encoded, and the charge is predicted. (**Step 2**) A theoretical isotope distribution is generated and aligned to the data. The optimal shift of the theoretical distribution relative to the experimental data is determined by cosine similarity score. (**Step 3**) Peaks matched to the theoretical distribution at the optimal shift from step 2 are removed from the spectrum. The process, steps 1 through 3, is then repeated on the spectrum with matched peaks removed. This loop is repeated until the maximum number of knockdown rounds has been reached, no peaks were accepted, or insufficient peaks remain.

To fix any mis-monoisotopic alignment issues, the spectrum is shifted, usually by a range of up to ±3 isotopes, and the score for each of those shifts is calculated (Figure 3, step 2). The shift with the best score is used. If multiple shifts meet the cosine similarity score threshold, then these shifts are also returned as possible monoisotopic masses (see Supplemental Discussion SD1).

For each correct peak, the key information about the scan, such as monoisotopic mass, charge, and isotope distribution are stored in a data structure. After performing the first cycle of charge state assignments, the matched peaks are removed from the spectrum, and the process is repeated. After 5–10 rounds of assignments and removals, the algorithm moves on to the next scan. After completing all scans in the data, the matched clusters can be output in a variety of formats, such as MsAlign.[46,47] Note, unlike other machine learning algorithms for feature identification,[34,48,49] IsoDec does not require the LC dimension. Currently, each scan is processed independently.

Implementation and Incorporation into Existing Tools

All code and model parameters for IsoDec are shared open-source with a modified BSD 3-clause license at https://github.com/michaelmarty/UniDec. IsoDec was initially developed and trained in Python using PyTorch,[42] but a shared library in C was also developed to speed up execution and simplify distribution. A GUI for deconvolution has been included in the UniDec software suite (Figure S2), and IsoDec has also been incorporated into MetaMorpheus for proteomics searches (https://github.com/smith-chem-wisc/MetaMorpheus). Additional information on the computational implementation is provided in the Supporting Methods (SM4).

Assessing Performance

To characterize IsoDec, we first evaluated sequence coverage and fragments matched for the 3 TDP averaged spectra from different instrument types. These metrics were calculated using an in-house Python script, with results benchmarked against ProSight Lite[50] to ensure accuracy. Additional evaluation was performed by calculating precision-recall curves. Detailed methods for both types of evaluation are provided in the Supplemental Methods (SM5, SM6, and SM7).

Results from full LC-MS/MS files were searched using a new version of MetaMorpheus in which importing MsAlign files had been enabled, as described in the Supplemental Methods (SM8). For each of the six TDP files, MS2 MsAlign files were generated using IsoDec (both 4- and 8-bin models), FLASHDeconv,[26] and TopFD.[20] MsAlign files from each tool were searched separately but using the same search parameters. Comparison of IsoDec to the native deconvolution implemented in MetaMorpheus was also performed (Figure S3, Supplemental Discussion SD2). Details on how proteoform-spectrum matches (PrSMs) were calculated and compared are provided in the Supplemental Methods (SM9). Additional Supplemental Methods detail assessing deconvolution quality (SM10), performance on synthetic spectra (SM11), and statistical methods (SM12).

Results and Discussion

Evaluating Performance on Isolated Spectra

We initially benchmarked the performance of IsoDec on a set of complex intact protein fragmentation spectra. These averaged spectra were selected because of their complexity and because they covered three different common mass analyzers (Orbitrap, FT-ICR, and ToF) and fragmentation types (UVPD, ETD, and ECD). Different instrument vendors, mass analyzers, and fragmentation strategies produce data with different characteristics, and it was important to ensure that model performance was consistent across different platforms.

First, we determined the sequence coverage and percentage of fragments explained for each of the three selected spectra. To compare the performance of IsoDec against existing tools, we benchmarked the results of IsoDec against those of TopFD and FLASHDeconv.[20,28] IsoDec and TopFD produced the highest sequence coverage on all three spectra (Figure 4A). The two tools performed similarly, yielding between 63–77% sequence coverage across the three spectra. FLASHDeconv achieved from 41–55% coverage.

In contrast, IsoDec and FLASHDeconv had the highest percentage of fragments matched. These two tools produced between 40–60% fragments matched, while TopFD yielded between 25–40% (Figure 4B). Interestingly, the 8-bin and the 4-bin model of IsoDec performed very similarly on these metrics. Overall, IsoDec exhibited a balanced combination of sequence coverage and fragments matched, demonstrating its ability to accurately assign isotopic envelopes and do so without loss of sensitivity. Importantly, IsoDec achieves this balance across the different fragmentation types and mass analyzers represented in these spectra.

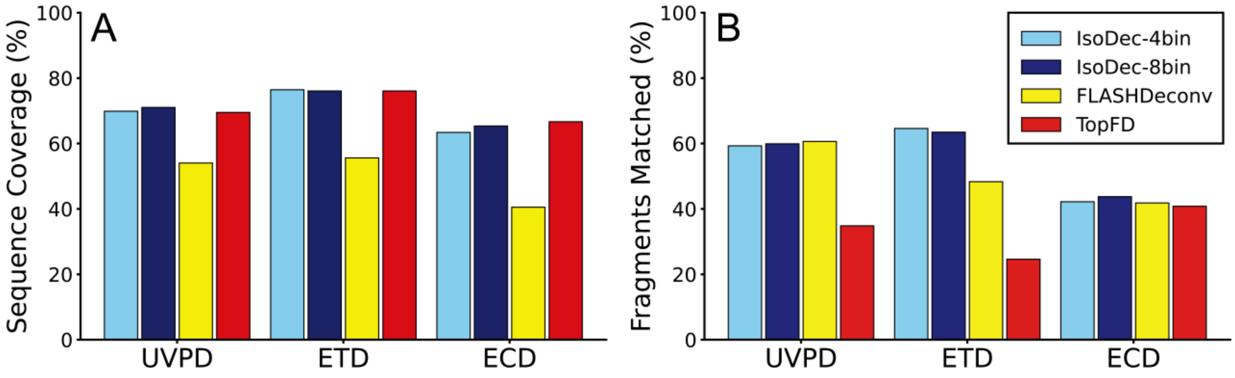


Figure 4. Summary of single spectrum analysis results. Sequence coverage (**A**) and percentage of fragments matched (**B**) produced by each software tool on the three selected spectra. No error bars are shown because each analysis was performed on a single averaged spectrum of each type.

One important question is whether this improvement in accuracy came from the neural network charge prediction or from the filtering that was performed after assignment. To test this, we implemented THRASH-like charge assignment in Python that replaced only the neural network piece of the IsoDec workflow and repeated the same evaluation discussed above. THRASH uses a combination of Fourier and Patterson charge maps to assign charges to isotope clusters, and it is the basis for TopFD.[17] Although the fragments matched was consistently slightly lower, the sequence coverage and percentage of fragments matched were largely similar when replacing the neural network with Fourier/Patterson charge assignment (Figure S4).

However, sequence coverage and percentage of fragments matched are imperfect metrics because not all true isotope clusters are from terminal fragments of the protein.[51] There can be true isotope clusters from contaminants or non-terminal fragments that are real but do not match a theoretical terminal fragment. To address this limitation, we performed manual annotation of the selected spectra so that a precision-recall curve, as presented by McIlwain *et al.*, could be constructed to more thoroughly investigate performance.[28]

A clear difference in performance between neural network models and THRASH charge assignment can be observed in the precision-recall curve (Figure 5D). IsoDec with the 8- and 4-bin models had optimized F_1 values of 0.69 and 0.67, respectively. Under optimal parameters the 8-bin model produced precision and recall values of 0.77 and 0.63, and the 4-bin model yielded 0.74 and 0.60. Here, it is clearer that the 8-bin model outperforms the 4-bin model. Both outperform IsoDec with THRASH-like charge prediction, which had an optimal F_1 value of 0.64, with a precision of 0.72 and a recall of 0.58. These results show that there is an advantage to using the neural net for charge assignment over existing strategies.

It is encouraging that the curve for IsoDec-THRASH passes very close to the point for TopFD, which also uses a THRASH-based approach. This agreement indicates that our implementation of THRASH performs similarly to existing implementations in TopFD. FLASHDeconv is more stringent than TopFD, as shown by the higher precision but lower recall.

Interestingly, there is a clear recall ceiling for IsoDec, likely because the charge for these clusters is never correctly assigned (see Figure 5C), no matter how loose the filtering parameters are. In the case of the cluster in Figure 5C, the distribution to the left is selected first, and multiple shared isotope peaks are knocked down. The knockdown of these shared isotope peaks results in incorrect charge assignment for the distribution highlighted in Figure 5C. Similar recall ceilings were previously observed for THRASH-based algorithms.[28] To achieve higher recall, new strategies to handle cases where the charge is incorrectly assigned are required. Additional details on limitations and future directions are provided below.

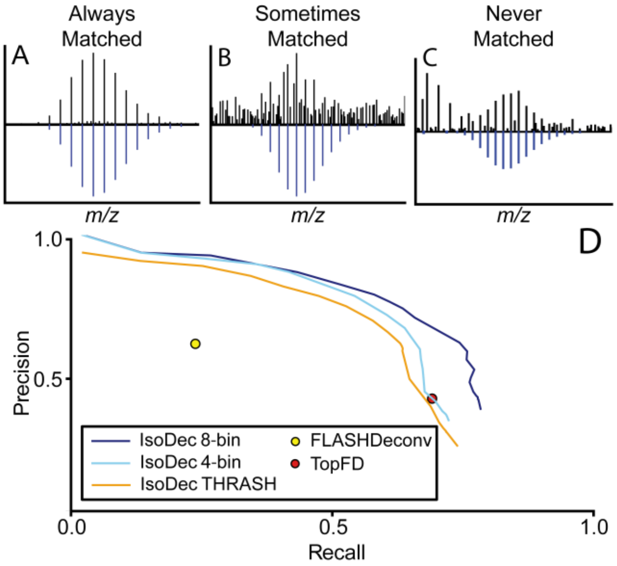


Figure 5. Precision (the percentage of identified clusters that were manually annotated) versus Recall (the percentage of manually annotated clusters that were identified) curve for IsoDec with different charge assignment methods as well as single points for FLASHDeconv and TopFD with default parameters. Above the curve are three isotope clusters that are vertically aligned approximately with where they are first matched on the precision-recall curve. (**A**) is matched at a precision and recall of 1 and 0.027. (**B**) is first matched at precision and recall of 0.895 and 0.45. (**C**) has its charge assigned incorrectly by IsoDec, so it is never matched.

Evaluating Performance on Large Top-Down Data Sets

Next, we tested how IsoDec performed on large TDP LC-MS/MS data sets consisting of thousands of individual scans and millions of features. We selected a set of six files from replicate injections of intact *E. coli* lysate[41] and benchmarked the performance of IsoDec against FLASHDeconv and TopFD. In addition to sensitivity and accuracy, speed also becomes a concern when working with large files, so we first compared the total time to deconvolve the selected files. Here, IsoDec performed similarly to FLASHDeconv and significantly faster than TopFD (Figure 6A). We also compared the time per cluster to account for differences in depth of coverage (Figure 6B). Although FLASHDeconv was slightly faster overall, IsoDec has a faster rate per-cluster.

After comparing speed, we next compared deconvolution reproducibility. The six selected files are replicate injections under the same chromatographic conditions, so the same species should be found at approximately the same retention time in each file. Features were constructed by aggregating clusters with similar masses from nearby scans, as described in the Supplemental Methods (SM10). Each possible pairwise combination of the six files was compared from the same deconvolution algorithm, and we calculated the percentage of features that agreed between the files. Here, IsoDec slightly outperformed FLASHDeconv and TopFD, matching an average of over 75% of features (Figure S5 and Supplemental Discussion SD3). However, all algorithms performed similarly with respect to feature intensity, with over 95% of total feature intensity matched between replicate injections. These data suggest all algorithms perform well on abundant features, and the differences lie in the less abundant proteoforms.

We then investigated how IsoDec performs in the context of a proteomics database search. We found that each tool produced a similar number of proteoform-spectrum matches (PrSMs; Figure S6A). A PrSM refers to a match of the combination of an experimental precursor mass and fragmentation spectrum to their synthetic or catalogued counterparts from a database or spectral library. The number of unambiguous PrSMs was statistically similar between tools, but FLD and TopFD produced higher numbers of ambiguous PrSMs.

We next compared the scans that yielded unambiguous PrSMs from each deconvolution tool. In total, 2160 scans across the six files yielded a PrSM with every algorithm (Figure S6C). From these unambiguous PrSMs, we compared the proteoforms identified using a strict definition of matching that required the same charge in the same scan on the same proteoform, defined as having the same sequence with the same set of post-translational modifications (PTMs) on the same specific residues.[52] Of the 2160 scans that produced PrSMs in all three searches, 1754 (81%) produced the same proteoform identification (Figure 5C), showing remarkable agreement. If we allowed less strict matching criteria where the PTMs could vary, 2015 (93%) common PrSMs were found (Figure S6D). However, as previously observed,[27] there is still significant variability from each deconvolution tool because only about half of the unique scans that yielded PrSMs in a particular search were shared between all three (Figure S6C).

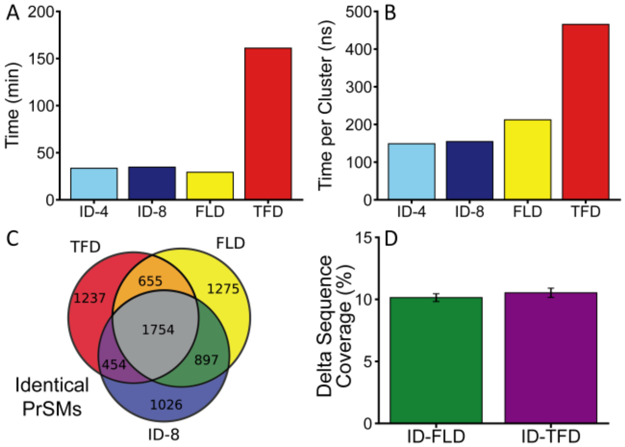


Figure 6. Summary of results of top-down file analysis. (**A**) The time required to deconvolve the set of six files and (**B**) the time per assigned cluster across all six files for each algorithm. No error bars are included in **A** of **B** because these metrics report the summed time across all six data files. (**C**) Overlap of unambiguous PrSMs between the three different deconvolution methods. (**D**) Average difference in sequence coverage on 1754 identical PrSMs between IsoDec-8bit and FLASHDeconv (*green*) or TopFD (*purple*), error bars represent 95% confidence intervals (N=1754). TFD, TopFD; FLD, FLASHDeconv; ID-4, IsoDec-4bit; ID-8, IsoDec-8bit.

To further characterize the performance of the three algorithms in a search context, we dug deeper into the cases where all three tools (IsoDec 8-bin, FLASHDeconv, and TopFD) produced the same proteoform identification. The sequence coverage and percentage of fragments matched were computed for each matched PrSM (see SM9), and the difference in these two parameters between IsoDec 8-bin and both FLASHDeconv and TopFD were calculated. IsoDec produced approximately 10% greater sequence coverage than FLASHDeconv and TopFD (Figure 6D), but FLASHDeconv had a higher percentage of fragments matched (Figure S6B). Manually inspecting three representative MS2 spectra revealed that this decrease in fragments matched with IsoDec is likely due to the detection of more low-abundance clusters that are real but do not correspond to a terminal fragment that would be matched in a traditional search. Overall, as observed with the individual spectra, TopFD tends to assign more clusters with lower accuracy, FLASHDeconv tends to assign fewer clusters with higher accuracy, and IsoDec tends to fall in the middle, balancing depth of coverage with accuracy of assignment.

Limitations and Future Directions

IsoDec and its neural network approach to charge assignment show promise in addressing some of the challenges with high-resolution deconvolution. However, this initial version has several limitations. The first limitation is the inability to reliably assign high (>30) charge states (Figure S7A), which stems from the underrepresentation of high charge states in the training data (Figure S1). In preliminary testing, balancing the charge states present in the training data with addition of synthetic high-charge clusters dramatically improved accuracy on high charge states (Figure S7B). Further testing and optimization are needed to examine how this new training strategy impacts overall model performance, but this initial result is encouraging.

The workflow also struggles with overlapping isotope distributions, particularly in cases where there is harmonic overlap or distributions of similar intensity (Figure S8A). In some cases (Figure S8A), the charge state is correctly assigned to a minor distribution, but it is not accepted as a match because another distribution has a higher apex intensity. Other problems occur when shared peaks are knocked down on earlier versions of the algorithm, and what is left behind cannot be assigned (Figures 5C and S8B). Future work is exploring ways to limit the focus on the encoding to improve charge assignment when multiple peaks are present and ways to improve peak matching, fitting, and removal after assignment.

Finally, we have largely focused on protein analysis in our initial development of IsoDec. We expect that other types of molecules, including DNA, RNA, and polymers, should perform similarly with the neural network charge assignment. However, adjustments may be needed in isotope distribution prediction to score peaks properly after charge assignment. Future work will implement alternative isotope distribution models beyond the averagine used here.[53]

Conclusion

Here, we describe a neural network for isotopic cluster charge assignment using a unique mass defect encoding method. The performance of the new workflow, IsoDec, was compared on protein fragmentation spectra from different instruments and on large top-down proteomics LC-MS/MS files. Overall, IsoDec produced similar sensitivity to existing tools but with improved accuracy and speed. TopFD tended to have a higher number of clusters assigned, but a smaller percentage of those clusters matched protein fragments. In contrast, FLASHDeconv tended to have a higher percentage of clusters matched but at the cost of assigning fewer total clusters. Importantly, the improvement in performance with IsoDec is due to the neural network strategy for charge assignment, not simply stricter filtering of assigned clusters. Overall, IsoDec presents a novel strategy for encoding raw mass spectrometry data for AI applications and reveals the potential for lightweight and specialized AI to help automate analysis of challenging data.

Acknowledgements

The authors thank the community for public sharing of data that was used as both training and testing data, especially from the labs of Liangliang Sun, Julia Chamot-Rooke, Oliver Kohlbacher, Jan Enghild, Lissa Anderson, Jennifer Broadbelt, and Neil Kelleher. This work was funded by the National Science Foundation (CHE-2403561 to M.T.M.) and the National Institutes of Health (R01GM147653 to M.R.S and T32GM135066 to N.E.B). This work used High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and Research, Innovation, and Impact (RII) and maintained by the UArizona Research Technologies department.

References

(1) Chanthamontri, C.; Liu, J.; McLuckey, S. A. Charge State Dependent Fragmentation of Gaseous α-Synuclein Cations via Ion Trap and Beam-Type Collisional Activation. *Int. J. Mass Spectrom.* **2009**, *283* (1), 9–16. https://doi.org/10.1016/j.ijms.2008.12.007.

(2) Reid, G. E.; Wu, J.; Chrisman, P. A.; Wells, J. M.; McLuckey, S. A. Charge-State-Dependent Sequence Analysis of Protonated Ubiquitin Ions via Ion Trap Tandem Mass Spectrometry. *Anal. Chem.* **2001**, *73* (14), 3274–3281. https://doi.org/10.1021/ac0101095.

(3) Ferrige, A. G.; Seddon, M. J.; Green, B. N.; Jarvis, S. A.; Skilling, J.; Staunton, J. Disentangling Electrospray Spectra with Maximum Entropy. *Rapid Commun. Mass Spectrom.* **1992**, *6* (11), 707–711. https://doi.org/10.1002/rcm.1290061115.

(4) Ferrige, A. G.; Seddon, M. J.; Jarvis, S.; Skilling, J.; Welch, J. The Application of Maxent to Electrospray Mass Spectrometry. In *Maximum Entropy and Bayesian Methods: Seattle, 1991*; Smith, C. R., Erickson, G. J., Neudorfer, P. O., Eds.; Springer Netherlands: Dordrecht, 1992; pp 327–335. https://doi.org/10.1007/978-94-017-2219-3_24.

(5) Ferrige, A. G.; Seddon, M. J.; Jarvis, S.; Skilling, J.; Aplin, R. Maximum Entropy Deconvolution in Electrospray Mass Spectrometry. *Rapid Commun. Mass Spectrom.* **1991**, *5* (8), 374–377. https://doi.org/10.1002/rcm.1290050810.

(6) Mann, M.; Meng, C. K.; Fenn, J. B. Interpreting Mass Spectra of Multiply Charged Ions. *Anal. Chem.* **1989**, *61* (15), 1702–1708.

(7) Zhang, Z.; Marshall, A. G. A Universal Algorithm for Fast and Automated Charge State Deconvolution of Electrospray Mass-to-Charge Ratio Spectra. *J. Am. Soc. Mass Spectrom.* **1998**, *9* (3), 225–233. https://doi.org/10.1016/S1044-0305(97)00284-5.

(8) Morgner, N.; Robinson, C. V. Massign: An Assignment Strategy for Maximizing Information from the Mass Spectra of Heterogeneous Protein Assemblies. *Anal. Chem.* **2012**, *84* (6), 2939–2948. https://doi.org/10.1021/ac300056a.

(9) van Breukelen, B.; Barendregt, A.; Heck, A. J. R.; van den Heuvel, R. H. H. Resolving Stoichiometries and Oligomeric States of Glutamate Synthase Protein Complexes with Curve Fitting and Simulation of Electrospray Mass Spectra. *Rapid Commun. Mass Spectrom.* **2006**, *20* (16), 2490–2496. https://doi.org/10.1002/rcm.2620.

(10) Lu, J.; Trnka, M. J.; Roh, S.-H.; Robinson, P. J. J.; Shiau, C.; Fujimori, D. G.; Chiu, W.; Burlingame, A. L.; Guan, S. Improved Peak Detection and Deconvolution of Native Electrospray Mass Spectra from Large Protein Complexes. *J. Am. Soc. Mass Spectrom.* **2015**, *26* (12), 2141–2151. https://doi.org/10.1007/s13361-015-1235-6.

(11) Marty, M. T.; Baldwin, A. J.; Marklund, E. G.; Hochberg, G. K. A.; Benesch, J. L. P.; Robinson, C. V. Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to

Polydisperse Ensembles. *Anal. Chem.* **2015**, *87* (8), 4370–4376. https://doi.org/10.1021/acs.analchem.5b00140.

(12) Tseng, Y.-H.; Uetrecht, C.; Yang, S.-C.; Barendregt, A.; Heck, A. J. R.; Peng, W.-P. Game-Theory-Based Search Engine to Automate the Mass Assignment in Complex Native Electrospray Mass Spectra. *Anal. Chem.* **2013**, *85* (23), 11275–11283. https://doi.org/10.1021/ac401940e.

(13) Cleary, S. P.; Thompson, A. M.; Prell, J. S. Fourier Analysis Method for Analyzing Highly Congested Mass Spectra of Ion Populations with Repeated Subunits. *Anal. Chem.* **2016**, *88* (12), 6205–6213. https://doi.org/10.1021/acs.analchem.6b01088.

(14) Peris-Díaz, M. D.; Guran, R.; Zitka, O.; Adam, V.; Krężel, A. Mass Spectrometry-Based Structural Analysis of Cysteine-Rich Metal-Binding Sites in Proteins with MetaOdysseus R Software. *J. Proteome Res.* **2021**, *20* (1), 776–785. https://doi.org/10.1021/acs.jproteome.0c00651.

(15) Keifer, D. Z.; Shinholt, D. L.; Jarrold, M. F. Charge Detection Mass Spectrometry with Almost Perfect Charge Accuracy. *Anal. Chem.* **2015**, *87* (20), 10330–10337. https://doi.org/10.1021/acs.analchem.5b02324.

(16) Su, P.; McGee, J. P.; Hollas, M. A. R.; Fellers, R. T.; Durbin, K. R.; Greer, J. B.; Early, B. P.; Yip, P. F.; Zabrouskov, V.; Srzentić, K.; Senko, M. W.; Compton, P. D.; Kelleher, N. L.; Kafader, J. O. Standardized Workflow for Multiplexed Charge Detection Mass Spectrometry on Orbitrap Analyzers. *Nat. Protoc.* **2025**, 1–24. https://doi.org/10.1038/s41596-024-01091-y.

(17) Senko, M. W.; Beu, S. C.; McLafferty, F. W. Automated Assignment of Charge States from Resolved Isotopic Peaks for Multiply Charged Ions. *J. Am. Soc. Mass Spectrom.* **1995**, *6* (1), 52–56. https://doi.org/10.1016/1044-0305(94)00091-D.

(18) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. Automated Reduction and Interpretation of High Resolution Electrospray Mass Spectra of Large Molecules. *J Am Soc Mass Spectrom* **2000**, *11* (4), 320–332. https://doi.org/10.1016/s1044-0305(99)00157-9.

(19) Zabrouskov, V.; Senko, M. W.; Du, Y.; Leduc, R. D.; Kelleher, N. L. New and Automated MSn Approaches for Top-down Identification of Modified Proteins. *J. Am. Soc. Mass Spectrom.* **2005**, *16* (12), 2027–2038. https://doi.org/10.1016/j.jasms.2005.08.004.

(20) Basharat, A. R.; Zang, Y.; Sun, L.; Liu, X. TopFD: A Proteoform Feature Detection Tool for Top–Down Proteomics. *Anal. Chem.* **2023**, *95* (21), 8189–8196. https://doi.org/10.1021/acs.analchem.2c05244.

(21) Hoopmann, M. R.; Finney, G. L.; MacCoss, M. J. High-Speed Data Reduction, Feature Detection, and MS/MS Spectrum Quality Assessment of Shotgun Proteomics Data Sets Using High-Resolution Mass Spectrometry. *Anal Chem* **2007**, *79* (15), 5620–5632. https://doi.org/10.1021/ac0700833.

(22) Liu, X.; Inbar, Y.; Dorrestein, P. C.; Wynne, C.; Edwards, N.; Souda, P.; Whitelegge, J. P.; Bafna, V.; Pevzner, P. A. Deconvolution and Database Search of Complex Tandem Mass Spectra of Intact Proteins: A Combinatorial Approach. *Mol Cell Proteomics* **2010**, *9* (12), 2772–2782. https://doi.org/10.1074/mcp.M110.002766.

(23) Park, J.; Piehowski, P. D.; Wilkins, C.; Zhou, M.; Mendoza, J.; Fujimoto, G. M.; Gibbons, B. C.; Shaw, J. B.; Shen, Y.; Shukla, A. K.; Moore, R. J.; Liu, T.; Petyuk, V. A.; Tolić, N.; Paša-Tolić, L.; Smith, R. D.; Payne, S. H.; Kim, S. Informed-Proteomics: Open-Source Software Package for Top-down Proteomics. *Nat. Methods* **2017**, *14* (9), 909–914. https://doi.org/10.1038/nmeth.4388.

(24) Palasser, M.; Breuker, K. FAST MS: Software for the Automated Analysis of Top-Down Mass Spectra of Polymeric Molecules Including RNA, DNA, and Proteins. *J. Am. Soc. Mass Spectrom.* **2025**, *36* (2), 247–257. https://doi.org/10.1021/jasms.4c00236.

(25) Robey, M. T.; Durbin, K. R. Improving Top-Down Sequence Coverage with Targeted Fragment Matching. *J. Am. Soc. Mass Spectrom.* **2024**, *35* (12), 3296–3300. https://doi.org/10.1021/jasms.4c00161.

(26) Jeong, K.; Kim, J.; Gaikwad, M.; Hidayah, S. N.; Heikaus, L.; Schlüter, H.; Kohlbacher, O. FLASHDeconv: Ultrafast, High-Quality Feature Deconvolution for Top-Down Proteomics. *Cell Syst.* **2020**, *10* (2), 213-218.e6. https://doi.org/10.1016/j.cels.2020.01.003.

(27) Tabb, D. L.; Jeong, K.; Druart, K.; Gant, M. S.; Brown, K. A.; Nicora, C.; Zhou, M.; Couvillion, S.; Nakayasu, E.; Williams, J. E.; Peterson, H. K.; McGuire, M. K.; McGuire, M. A.; Metz, T. O.; Chamot-Rooke, J. Comparing Top-Down Proteoform Identification: Deconvolution, PrSM Overlap, and PTM Detection. *J. Proteome Res.* **2023**, *22* (7), 2199–2217. https://doi.org/10.1021/acs.jproteome.2c00673.

(28) McIlwain, S. J.; Wu, Z.; Wetzel, M.; Belongia, D.; Jin, Y.; Wenger, K.; Ong, I. M.; Ge, Y. Enhancing Top-Down Proteomics Data Analysis by Combining Deconvolution Results through a Machine Learning Strategy. *J. Am. Soc. Mass Spectrom.* **2020**, *31* (5), 1104–1113. https://doi.org/10.1021/jasms.0c00035.

(29) Meyer, J. G. Deep Learning Neural Network Tools for Proteomics. *Cell Rep Methods* **2021**, *1* (2), 100003. https://doi.org/10.1016/j.crmeth.2021.100003.

(30) Beck, A. G.; Muhoberac, M.; Randolph, C. E.; Beveridge, C. H.; Wijewardhane, P. R.; Kenttämaa, H. I.; Chopra, G. Recent Developments in Machine Learning for Mass Spectrometry. *ACS Meas. Sci. Au* **2024**, *4* (3), 233–246. https://doi.org/10.1021/acsmeasuresciau.3c00060.

(31) Abdelmoula, W. M.; Lopez, B. G.-C.; Randall, E. C.; Kapur, T.; Sarkaria, J. N.; White, F. M.; Agar, J. N.; Wells, W. M.; Agar, N. Y. R. Peak Learning of Mass Spectrometry Imaging Data Using Artificial Neural Networks. *Nat. Commun.* **2021**, *12* (1), 5544. https://doi.org/10.1038/s41467-021-25744-8.

(32) Abdelmoula, W. M.; Stopka, S. A.; Randall, E. C.; Regan, M.; Agar, J. N.; Sarkaria, J. N.; Wells, W. M.; Kapur, T.; Agar, N. Y. R. massNet: Integrated Processing and Classification of Spatially Resolved Mass Spectrometry Data Using Deep Learning for Rapid Tumor Delineation. *Bioinformatics* **2022**, *38* (7), 2015–2021. https://doi.org/10.1093/bioinformatics/btac032.

(33) Tran, N. H.; Qiao, R.; Xin, L.; Chen, X.; Liu, C.; Zhang, X.; Shan, B.; Ghodsi, A.; Li, M. Deep Learning Enables de Novo Peptide Sequencing from Data-Independent-Acquisition Mass Spectrometry. *Nat Methods* **2019**, *16* (1), 63–66. https://doi.org/10.1038/s41592-018-0260-3.

(34) Zohora, F. T.; Rahman, M. Z.; Tran, N. H.; Xin, L.; Shan, B.; Li, M. DeepIso: A Deep Learning Model for Peptide Feature Detection from LC-MS Map. *Sci Rep* **2019**, *9* (1), 17168. https://doi.org/10.1038/s41598-019-52954-4.

(35) Boiko, D. A.; Kozlov, K. S.; Burykina, J. V.; Ilyushenkova, V. V.; Ananikov, V. P. Fully Automated Unconstrained Analysis of High-Resolution Mass Spectrometry Data with Machine Learning. *J Am Chem Soc* **2022**, *144* (32), 14590–14606. https://doi.org/10.1021/jacs.2c03631.

(36) Basharat, A. R.; Ning, X.; Liu, X. EnvCNN: A Convolutional Neural Network Model for Evaluating Isotopic Envelopes in Top-Down Mass-Spectral Deconvolution. *Anal. Chem.* **2020**, *92* (11), 7778–7785. https://doi.org/10.1021/acs.analchem.0c00903.

(37) Zhong, J.; Song, X.; Wang, S. FREE: Enhanced Feature Representation for Isotopic Envelope Evaluation in Top-Down Mass Spectra Deconvolution. *Anal. Chem.* **2024**, *96* (31), 12602–12615. https://doi.org/10.1021/acs.analchem.4c00152.

(38) Mikawy, N. N.; Ramírez, C. R.; DeFiglia, S. A.; Szot, C. W.; Le, J.; Lantz, C.; Wei, B.; Zenaidee, M. A.; Blakney, G. T.; Nesvizhskii, A. I.; Loo, J. A.; Ruotolo, B. T.; Shabanowitz, J.; Anderson, L. C.; Håkansson, K. Are Internal Fragments Observable in Electron Based Top-

Down Mass Spectrometry? *Mol. Cell. Proteomics MCP* **2024**, *23* (9), 100814. https://doi.org/10.1016/j.mcpro.2024.100814.

(39) Weisbrod, C. R.; Kaiser, N. K.; Syka, J. E. P.; Early, L.; Mullen, C.; Dunyach, J.-J.; English, A. M.; Anderson, L. C.; Blakney, G. T.; Shabanowitz, J.; Hendrickson, C. L.; Marshall, A. G.; Hunt, D. F. Front-End Electron Transfer Dissociation Coupled to a 21 Tesla FT-ICR Mass Spectrometer for Intact Protein Sequence Analysis. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (9), 1787–1795. https://doi.org/10.1007/s13361-017-1702-3.

(40) Dunham, S. D.; Wei, B.; Lantz, C.; Loo, J. A.; Brodbelt, J. S. Impact of Internal Fragments on Top-Down Analysis of Intact Proteins by 193 Nm UVPD. *J. Proteome Res.* **2023**, *22* (1), 170–181. https://doi.org/10.1021/acs.jproteome.2c00583.

(41) Dupré, M.; Duchateau, M.; Malosse, C.; Borges-Lima, D.; Calvaresi, V.; Podglajen, I.; Clermont, D.; Rey, M.; Chamot-Rooke, J. Optimization of a Top-Down Proteomics Platform for Closely Related Pathogenic Bacterial Discrimination. *J. Proteome Res.* **2021**, *20* (1), 202–211. https://doi.org/10.1021/acs.jproteome.0c00351.

(42) Imambi, S.; Prakash, K. B.; Kanagachidambaresan, G. R. PyTorch. In *Programming with TensorFlow: Solution for Edge Computing Applications*; Prakash, K. B., Kanagachidambaresan, G. R., Eds.; Springer International Publishing: Cham, 2021; pp 87–104. https://doi.org/10.1007/978-3-030-57077-4_10.

(43) Senko, M. W.; Beu, S. C.; McLaffertycor, F. W. Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions. *J Am Soc Mass Spectrom* **1995**, *6* (4), 229–233. https://doi.org/10.1016/1044-0305(95)00017-8.

(44) Kaur, P.; O'Connor, P. B. Algorithms for Automatic Interpretation of High Resolution Mass Spectra. *J. Am. Soc. Mass Spectrom.* **2006**, *17* (3), 459–468. https://doi.org/10.1016/j.jasms.2005.11.024.

(45) Rockwood, A. L.; Palmblad, M. Isotopic Distributions. In *Mass Spectrometry Data Analysis in Proteomics*; Matthiesen, R., Ed.; Humana Press: Totowa, NJ, 2013; pp 65–99. https://doi.org/10.1007/978-1-62703-392-3_3.

(46) Liu, X.; Sirotkin, Y.; Shen, Y.; Anderson, G.; Tsai, Y. S.; Ting, Y. S.; Goodlett, D. R.; Smith, R. D.; Bafna, V.; Pevzner, P. A. Protein Identification Using Top-Down Spectra. *Mol. Cell. Proteomics* **2012**, *11* (6), M111.008524. https://doi.org/10.1074/mcp.M111.008524.

(47) Kou, Q.; Xun, L.; Liu, X. TopPIC: A Software Tool for Top-down Mass Spectrometry-Based Proteoform Identification and Characterization. *Bioinformatics* **2016**, *32* (22), 3495–3497. https://doi.org/10.1093/bioinformatics/btw398.

(48) Dai, Y.; Yang, Y.; Wu, E.; Shen, C.; Qiao, L. Deep Learning Powers Protein Identification from Precursor MS Information. *J. Proteome Res.* **2024**, *23* (9), 3837–3846. https://doi.org/10.1021/acs.jproteome.4c00118.

(49) Zohora, F. T.; Rahman, M. Z.; Tran, N. H.; Xin, L.; Shan, B.; Li, M. Deep Neural Network for Detecting Arbitrary Precision Peptide Features through Attention Based Segmentation. *Sci Rep* **2021**, *11* (1), 18249. https://doi.org/10.1038/s41598-021-97669-7.

(50) Fellers, R. T.; Greer, J. B.; Early, B. P.; Yu, X.; LeDuc, R. D.; Kelleher, N. L.; Thomas, P. M. ProSight Lite: Graphical Software to Analyze Top-down Mass Spectrometry Data. *Proteomics* **2015**, *15* (7), 1235–1238. https://doi.org/10.1002/pmic.201400313.

(51) Lyon, Y. A.; Riggs, D.; Fornelli, L.; Compton, P. D.; Julian, R. R. The Ups and Downs of Repeated Cleavage and Internal Fragment Production in Top-Down Proteomics. *J. Am. Soc. Mass Spectrom.* **2018**, *29* (1), 150–157. https://doi.org/10.1007/s13361-017-1823-8.

(52) Smith, L. M.; Kelleher, N. L. Proteoform: A Single Term Describing Protein Complexity. *Nat. Methods* **2013**, *10* (3), 186–187. https://doi.org/10.1038/nmeth.2369.

(53) Agten, A.; Prostko, P.; Geubbelmans, M.; Liu, Y.; De Vijlder, T.; Valkenborg, D. A Compositional Model to Predict the Aggregated Isotope Distribution for Average DNA and RNA Oligonucleotides. *Metabolites* **2021**, *11* (6), 400. https://doi.org/10.3390/metabo11060400.