# The Hunt Lab Guide to *De Novo* Peptide Sequence Analysis by Tandem Mass Spectrometry


## Authors

Lissa C. Anderson, Dina L. Bai, Greg T. Blakney, David S. Butcher, Larry Reser, and Jeffrey Shabanowitz

## Correspondence

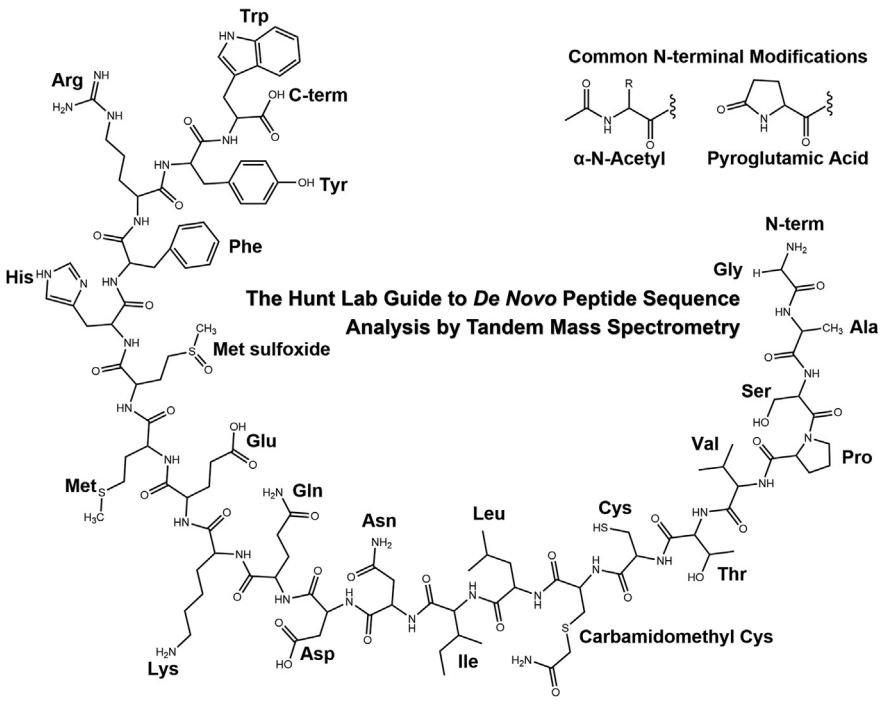
anderson@magnet.fsu.edu



## In Brief

This perspective celebrates the role Donald Hunt played in training researchers to manually interpret tandem mass spectra of peptides. We have adapted his teaching materials into *de novo* sequencing tutorials and have made available software tools to aid in calculations of fragment ion mass-to-charge ratios. We hope that the combination of these educational tools will continue to benefit students and researchers by empowering them to interpret data on their own.


## Graphical Abstract




## Highlights

- Donald Hunt taught the world to *de novo* sequence peptides by MS/MS.
- His teaching materials were digitized and adapted into freely available tutorials.
- Free software tools aid manual interpretation MS/MS spectra.

2024, Mol Cell Proteomics *23*(*12*), 100875

https://doi.org/10.1016/j.mcpro.2024.100875

# The Hunt Lab Guide to *De Novo* Peptide Sequence Analysis by Tandem Mass Spectrometry

Lissa C. Anderson[1,2,*], Dina L. Bai[3], Greg T. Blakney[1], David S. Butcher[1], Larry Reser[3], and Jeffrey Shabanowitz[3]

**Donald Hunt has made seminal contributions to the fields of proteomics, immunology, epigenetics, and glycobiology. The foundation of every important work to come out of the Hunt Laboratory is *de novo* peptide sequencing. For decades, he taught hundreds of students, postdocs, engineers, and scientists to directly interpret mass spectral data. To honor his legacy and ensure that the art of *de novo* sequencing is not lost, we have adapted his teaching materials into "The Hunt Lab Guide to *De Novo* Peptide Sequence Analysis by Tandem Mass Spectrometry". In addition to the *de novo* sequencing tutorials, we present two freely available software tools that facilitate manual interpretation of mass spectra and validation of search results. The first, "Hunt Lab Peptide Fragment Calculator", calculates precursor and fragment mass-to-charge ratios for any peptide. The second program, "Predator Protein Fragment Calculator", was inspired in part by the fragment calculator developed in the Hunt Lab. Its capabilities are enhanced to facilitate interpretation of mass spectral data derived from intact proteins. We hope that the combination of these educational tools will continue to benefit students and researchers by empowering them to interpret data on their own.**

## THE "BLUE BOOK"

On our first day in the Hunt lab, new graduate students were each given a blue, one-inch, three-ring binder. We were expected to read and master its contents by the end of the summer. When we were not being trained by senior students, we sat in our cubicles and diligently worked through the binder. In addition to chemical and laboratory safety information, relevant review articles, and groundbreaking articles from the lab, most of the binder's contents were dedicated to the *de novo* sequencing "short courses". One was dedicated to collision activated dissociation (CAD), and the other to electron transfer dissociation (ETD). There were also more than 100 unannotated spectra for practicing our newly learned *de novo* sequencing skills.

The CAD short course was originally developed in the late '80s, when Don was consulting for Finnigan MAT (now Thermo Fisher Scientific Inc). He, Jeffrey Shabanowitz, and several talented students and postdocs had laid the foundation for ultrasensitive *de novo* protein sequence analysis by tandem mass spectrometry (MS/MS) (1–5). Finnigan had introduced the first commercial triple quadrupole mass spectrometer, the triple-stage quadrupole (TSQ). Edman degradation was still the standard method for sequencing proteins, and magnetic sector instruments were favored, so Don (and Finnigan) needed to educate the scientific community of the merits of the new instruments and methods (6). From 1989-95, Finnigan sponsored the CAD short courses at the University of Virginia up to four times a year. In that time, Don personally trained over 400 scientists, from the US and abroad, to *de novo* sequence peptides (7).

The weaknesses of CAD for analysis of phosphopeptides, O-GlcNAcylated peptides, larger, more highly charged peptides, and intact proteins spurred Don on to pursue methods for improved sequence coverage. After the development and commercialization of ETD (8), he created an analogous ETD short course and, over the next decade, traveled extensively to disseminate it at conferences and workshops. In 2015, the course was adapted into a web-based tutorial (9). Unfortunately, the tutorial has been subject to the entropy of the internet and succumbed to link death.

Donald Hunt taught the world to sequence peptides by mass spectrometry, so the impact of the lessons contained in the "blue book" is incalculable. Their application propelled proteomics into a new era of discovery that has transformed our understanding of protein biology. The best way to honor his contributions is to ensure they retain their utility as educational tools. Here, we pick up that baton, and present "The Hunt Lab Guide to *De Novo* Peptide Sequence Analysis" by CAD and ETD.

From the [1]National High Magnetic Field Laboratory, and [2]Department of Chemistry and Biochemistry, Florida State University, Tallahassee, Florida, USA; [3]Department of Chemistry, University of Virginia, Charlottesville, Virginia, USA
*For correspondence: Lissa C. Anderson, anderson@magnet.fsu.edu.





https://doi.org/10.1016/j.mcpro.2024.100875

TUTORIALS

The original CAD short course used spectra generated by the TSQ. The peaks were 3 mass units wide, but the mass differences could be reliably calculated using the peak centroids. At some point in the '90s, it was updated to include spectra collected using a 3D ion trap, the Finnigan LCQ. Edman degradation was still widely in play as evidenced by Don's designation of N-terminal acetylation and pyroglutamate as, "Common N-terminal Blocking Groups" instead of N-terminal modifications. He recognized that most scientists might only have access to less expensive, low-resolution instruments or data. Consequently, in addition to the spectrum recorded on the unmodified peptide, he included spectra of the methyl ester and acetylated derivatives. This was a common practice that enabled one to determine the number of carboxylic acids (Asp, Glu, C terminus) the peptide possessed, and to distinguish Gln (128.06 Da) residues from Lys (128.10 Da) residues. Since the TSQ and the LCQ used low-energy collisional activation, formation of d-, w-, and other high-energy fragments was not discussed, and the isomers Leu/Ile could not be differentiated (annotated as Lxx or X) (10).

When Don taught the courses, he used an overhead projector and transparencies. The figures (which were Don Hunt "originals") and spectra in our blue books were photocopies of these transparencies (Fig. 1). When the ETD course was adapted into a web-based tutorial the figures were digitized, but the CAD course was overlooked. The raw spectral data had been archived to magnetic tape long before. To revive the tutorials here, some spectra were traced in CorelDRAW X8 vector graphics software (https://www.coreldraw.com). Other figures were replicated in ChemDraw (https://revvitysignals.com/products/research/chemdraw). Exact monoisotopic masses and chemical formulae of fragments were incorporated, but the text is largely unchanged. Our guiding principle was to err on the side of Don.

The tutorials are provided in PDF format and included as supplemental materials accompanying this article. We encourage readers to print the unannotated spectra at the beginning of each tutorial and label them as they follow along. A basic four-function calculator and a pen or pencil is all that is required. Fully annotated spectra are included at the end of the tutorials along with five unlabeled "practice problems". These are presented in order of difficulty and include examples of peptides containing labile posttranslational modifications (PTMs) (phosphorylation – pS, pT, or pY; O-GlcNAcylation – gS or gT). The answer key is provided on the last page of each tutorial.

SOFTWARE TOOLS

*The Hunt Lab Peptide Fragment Calculator*

In 2004, with ETD coming to life, few existing programs supported interpretation of ETD MS/MS spectra, so a suite of customized software was developed in the Hunt Lab to aid students in validation of search results and *de novo* sequencing. The Hunt Lab Peptide Fragment Calculator software (http://doi.org/10.17605/OSF.IO/XRG3Q) calculates theoretical precursor and fragment masses for any peptide sequence. The code is written in Java. Fragment masses include ions for CAD (b-, y- and a-ions) and ETD (c-, $z^{\bullet}$-, y-, and $a^{\bullet}$-ions). The program returns the monoisotopic or average m/z of precursor and fragment ions over a user-defined range of charge states and ion types in a vertically displayed format. This output can be printed or copied to other programs such as Microsoft Excel.

Customizable static and variable PTMs can be added directly to multiple residues of the sequence entered in the Sequence Window (Fig. 2). For example, carbamidomethylation can be applied to all Cys residues by selecting it in the Static Modification Window. Variable modifications can be

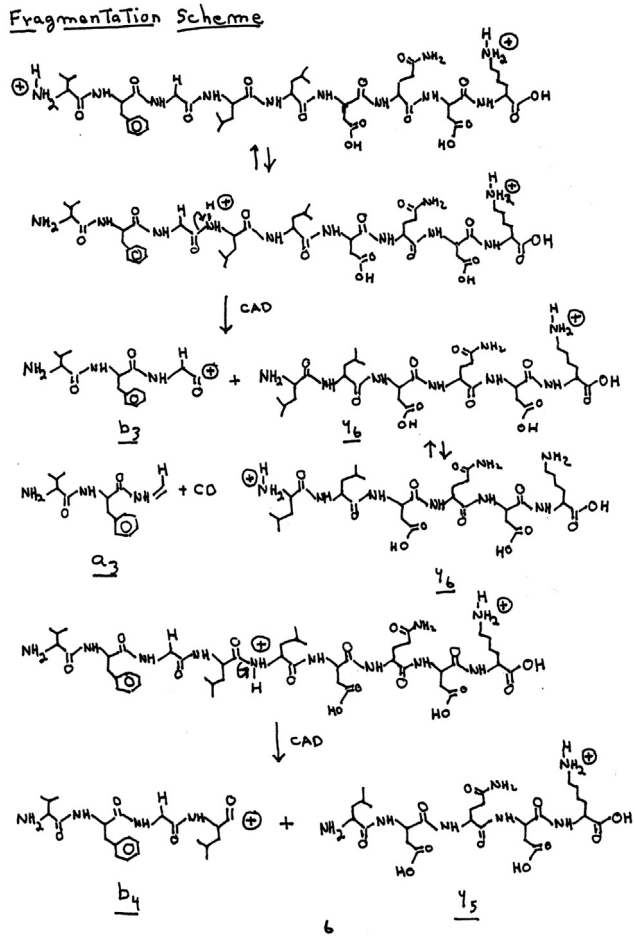



FIG. 1. **A Donald Hunt original (c. 1990)**. On collision with helium atoms in the ion trap, the $(M+2H)^{++}$ ions fragment to produce ions of type b and y. Ions of type b can also lose carbon monoxide and form ions of type a. The fragmentation shown in the top of the figure results from protonation of the third amide bond and generates the complementary pair, $b_3^+$ and $y_6^+$. Fragmentation products resulting from protonation of the fourth amide bond, $b_4^+$ and $y_5^+$, are shown at the *bottom*.

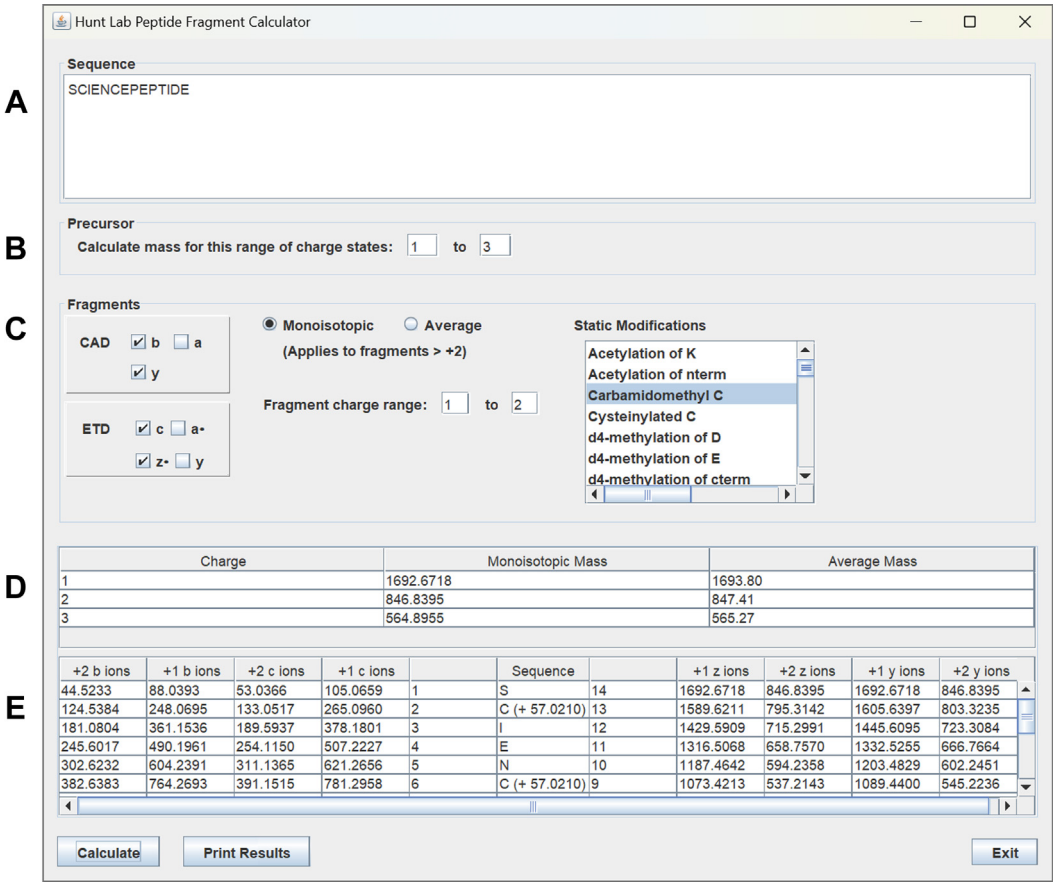



FIG. 2. **Hunt Lab Peptide Fragment Calculator graphical user interface**. A sequence can be typed or manually copied/pasted in the sequence window (*A*). Users define the range of charge states for intact peptide m/z calculation in the precursor parameters window (*B*). Dissociation or specific fragment ion types, and range of charge states are defined in the fragment parameters window (*C*). Users can also select monoisotopic or average masses (only applies to fragments of charge state 2+ or greater). One or more (Ctrl-Click) static modifications can be applied to the sequence in the static modifications window. Variable PTMs are specified by highlighting the modified residue in the sequence window; a list of available PTMs will "pop-up". The program output is divided into two sections. The *upper pane* (*D*) displays the monoisotopic and average precursor masses in the selected charge states. The *lower pane* (*E*) displays the m/z values for all requested ion types. PTM, posttranslational modification.

added to individual residues by highlighting them in the Sequence Window. Information about PTMs is stored in a separate text file, "FragmentMods.txt". This includes the name of each modification, its monoisotopic and average delta mass (Da), and the residue being modified in a comma-separated format. To make new modifications available, new lines can be added to the file using any text editor.

The Hunt Lab Peptide Fragment Calculator is legacy software that is comparable to several free web-based tools developed before and since 2004 including MS-Product in Protein Prospector (http://prospector.ucsf.edu/), NIST Mass and Fragment Calculator Software (11), and MS/MS Fragmentation Calculator from the University of Washington Proteomics Resource (https://proteomicsresource.washington.edu/). Users may find it useful to install the Hunt Lab Peptide Fragment Calculator program on computers that are air gapped or offline for security reasons.

### *Predator Protein Fragment Calculator*

Interpretation of MS/MS spectra becomes more difficult as precursor mass increases because fragment ion signal is distributed among more channels (*i.e.* charge states, isotopologues, and fragment ion types). For fragment ion assignment, major issues arise when the monoisotopic peak is not detected or not distinguishable from the noise in the spectrum. This renders fragment calculators that report only monoisotopic or average mass (or m/z) ineffective. While a simple internet search will reveal that there are several free web-based isotope pattern calculators, few were designed for analysis of peptides or intact proteins. Common limitations include (1) a requirement that the program input be a

chemical formula instead of an amino acid sequence, (2) returning only neutral or singly charged isotopologue masses, (3) simulation of only intact (no fragment) isotope distributions, and (4) simulation of only one sequence (or chemical formula) and charge state at a time. To address these issues, the Predator Protein Fragment Calculator (Fig. 3) (http://doi.org/10.17605/OSF.IO/QV874) was developed at the National High Magnetic Field Laboratory. This software is based on the Yergey algorithm (12) and the IsoPro 3.1 MS/MS software (13). It breaks each precursor or fragment into its elemental composition, based on amino acid sequence plus any chemical modifications, ion type, and charge state. The neutral mass of all corresponding isotopologue masses and their abundances above a reporting threshold are then calculated. The abundance weighted m/z average and abundance of all isotopologues of the desired

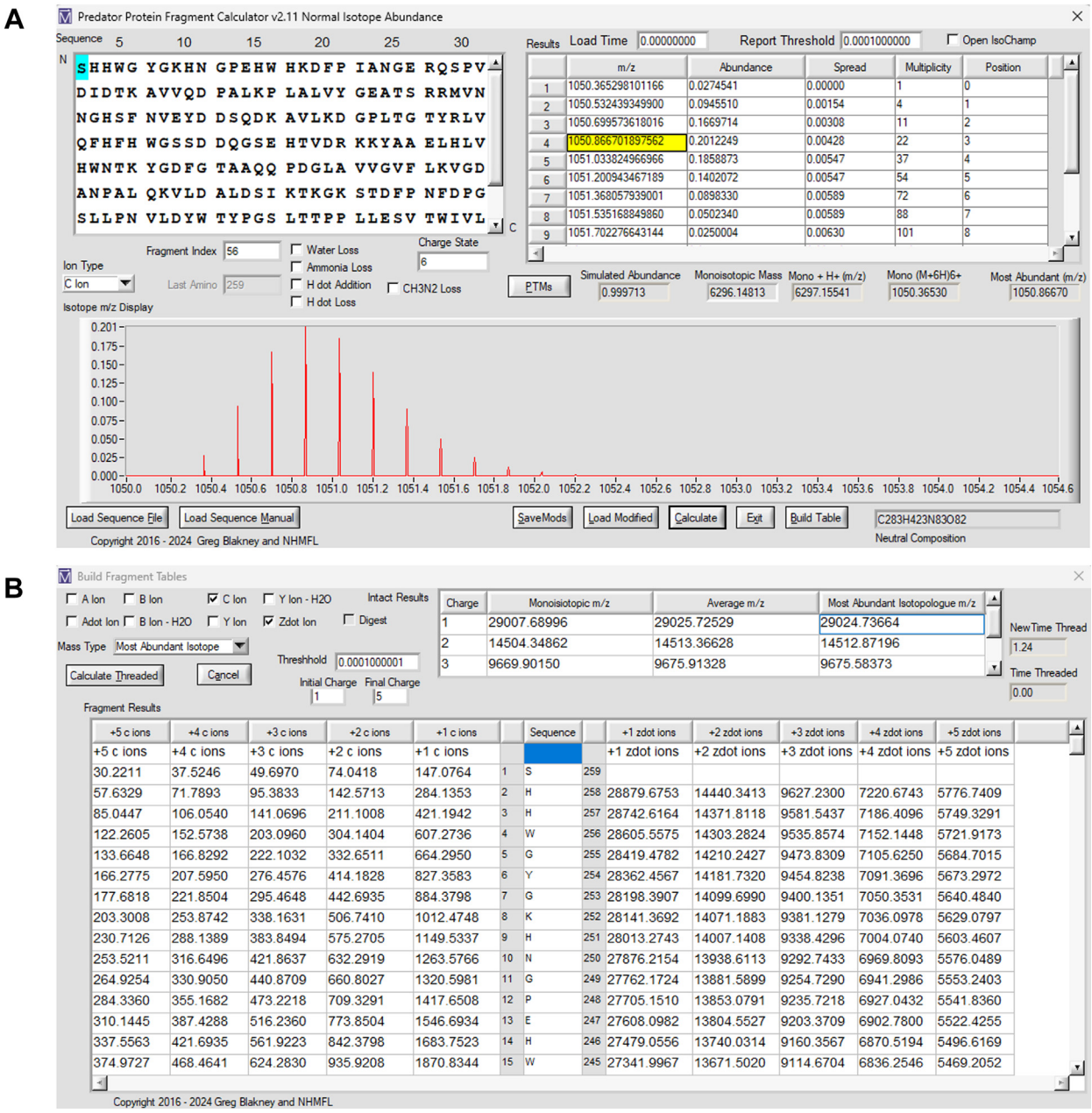



FIG. 3. **Predator Protein Fragment Calculator graphical user interface**. Users input a protein sequence, PTMs, ion type, index, and charge state in the main panel (*A*). The simulated isotope distribution is returned in tabular format and plotted. Selecting "Build Table" from the main panel opens the Build Fragment Tables panel (*B*). Here, users specify desired fragment ion types and range of charge states, and the m/z of the most abundant isotopologues are calculated for the entire sequence and shown in the Fragment Results table. PTM, posttranslational modification.

fragment ion are plotted and displayed in a table for comparison with raw data.

*Calculation Termination*—In any isotope simulation, one must determine when the simulation is "finished" or, in the case of large biomolecules, "good enough". Here, this is done by setting a lower abundance threshold, labeled as the "Report Threshold", for any calculation path to continue. The scale of this value assumes a value of 1 to be the maximum value of the abundance of all isotopologues when summed (labeled as "Combined Abundance"). Isotopologues with abundances below the report threshold are not returned. The minimum value allowed for the reporting threshold is $1 \times 10^{-10}$. Users may choose a less stringent, higher reporting threshold to speed calculations, but the recovered simulated combined abundance may be lower.

*Main Panel*—The user can load a sequence file (FASTA format) or manually enter the protein sequence. Chemical formulae of PTMs can be added to the sequence by double-clicking on the desired residue in the sequence box, or by clicking the "PTMs" button and specifying the amino acid index. PTM indices and chemical compositions can be saved to be used again later. Users specify the charge state, amino acid index, and select the desired ion type from a drop-down menu that includes several fragment ions, intact, or protein digest options. Once the neutral isotopologue masses and abundances are determined, the charge state is applied and the m/z and abundance of all isotopologues above the report threshold are displayed in a table and plotted. The chemical composition of the neutral is displayed in the bottom right-hand corner. Above the isotope m/z display, some common neutral losses (*e.g.* water, ammonia, H atom exchange, and so on.) can be selected and applied. The most abundant isotopologue is highlighted in the table and displayed above the isotope m/z display on the right side along with the monoisotopic m/z, monoisotopic mass (neutral), and the monoisotopic mass of the $[M + H]^+$.

*Build Table*—The sequence and PTM input from the main panel is also the basis for the "Build Fragment Tables" panel, the GUI design of which was modeled after the Hunt Lab Peptide Fragment Calculator. Given the information in this panel, the user is provided quick access to the most helpful single piece of information required to annotate an isotopically resolved intact protein MS/MS spectrum in one table—the most abundant peak for each fragment in each charge state. The user defines fragment ion types and charge states in the upper left-hand portion of the panel. The m/z of the most abundant isotopologues for the given fragments and charge states are displayed in the "Fragment Results" table. N- and C-terminal fragments are displayed on the left and right side of the sequence, respectively. C-terminal ion types are listed in reverse order such that complementary pairs are represented in consecutive rows. The "Intact Results" table in the upper righthand side is provided so that users can quickly verify the intact m/z across the selected range of charge states.

*Limitations*—To decrease computation time, element isotope distributions were precalculated, sorted, and converted to C libraries as lookup tables. The element count limits are as follows: carbon ≤25,000; hydrogen ≤35,000; nitrogen ≤5000; oxygen ≤6000; and sulfur ≤600. Phosphorus is unlimited because it has only one stable isotope. These limits enable simulations for proteins up to approximately 200 kDa. We also assume that the modest change in hydrogen (proton) count related to the charge state of the protein will not change the most abundant isotopologue or shape of the distribution appreciably. All calculations are performed at zero charge and the m/z of the isotopologues are adjusted by adding the appropriate number of protons without recalculation of isotopologue abundance.

## CONCLUDING REMARKS

The advent of the Orbitrap and higher resolution time-of-flight instruments along with advances in computer and software technology have made sequencing peptides and proteins easier than it was in the '90s. In fact, it's fair to say that *de novo* sequencing is a dying art. However, any Hunt Lab alum will attest that their *de novo* sequencing skills helped them discover novel peptides or proteoforms that would have otherwise been misidentified or gone unidentified. In practice, proteins with undetermined sequences, those that contain PTMs, or those that result from splice variants and genes containing single nucleotide polymorphisms require, at least to some degree, manual validation of the MS/MS data. *De novo* sequencing remains the only available approach to analyze proteins from organisms with unknown genomes, novel splice variants, and antibodies (14). In the years to come, it will be interesting to see how artificial intelligence and machine learning improve the accuracy of automated protein sequencing. But it's been 30 years since SEQUEST (15), and the algorithms still routinely misidentify peptides. Any peptide or protein containing an unexpected sequence or PTMs cannot be identified correctly via database or spectral library search. Misidentifications plague our datasets, and analytes of interest are frequently missed. In fact, it is estimated that only one out of every four spectra collected during an MS/MS experiment can be mapped to a protein sequence (16, 17). Our advice to readers—take the time to ensure the accuracy of data interpretation. Question and validate important search results. An understanding of *de novo* sequencing might allow you to find something you would have missed. That same understanding may help us to build better algorithms for the future as the field expands into instruments that acquire more data faster than ever before. Good luck to everyone, and thanks for everything, Don.

## DATA AVAILABILITY

All software referenced in this article is publicly available from Open Science Framework. Hunt Lab Peptide Fragment Calculator is available at http://doi.org/10.17605/OSF.IO/

XRG3Q. Predator Protein Fragment Calculator is available at http://doi.org/10.17605/OSF.IO/QV874. No other software or datasets were used in this publication. These tools are available for reuse under the terms of the Creative Commons Attribution Noncommercial-Noderivatives 4.0 International Public License (https://creativecommons.org/licenses/by-nc-nd/4.0/). Hunt Lab Peptide Fragment Calculator is Copyright © 2024 Dina L. Bai (dlb6z@virginia.edu). Predator Protein Fragment Calculator is Copyright © 2016 to 2024 Greg Blakney (blakney@magnet.fsu.edu) and the National High Magnetic Field Laboratory.

Hunt Lab: https://osf.io/xrg3q/?view_only=6467ab93a05142e59429a0aa95aafedb

Predator: https://osf.io/qv874/?view_only=9c24d402d26146a598b9a2c7d1a4ab3f

## SUPPLEMENTAL MATERIALS

This article contains supplemental data (10, 18).

*Acknowledgments*—The authors thank Lydia Babcock-Adams, and Joseph Frye, Kristina Håkansson, and Christopher Hendrickson for helpful discussions.

*Funding and additional information*—This work was supported by the National Institutes of Health grant R01 GM037537 (D. L. B, L. R. and J. S.), by the National Science Foundation Divisions of Materials Research and Chemistry grants DMR-1644779 & DMR-2128556 and the State of Florida (L. C. A., G. T. B., and D. S. B.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

*Author contributions*—L. C. A., D. L. B., D. S. B., and L. R., visualization; L. C. A., D. L. B., G. T. B., L. R., and J. S. writing–review and editing; L. C. A., D. L. B., and G. T. B., writing–original draft; L. C. A. and J. S. supervision; L. C. A. and J. S. project administration; L. C. A., conceptualization; D. L. B. and G. T. B. software; D. S. B. validation; D. S. B. data curation.

*Conflict of interest*—The authors declare no competing interests.

*Abbreviations*—The abbreviations used are: CAD, collision activated dissociation; ETD, electron transfer dissociation; MS/MS, tandem mass spectrometry; PTM, posttranslational modification; TSQ, triple-stage quadrupole.

Received August 12, 2024, and in revised form, October 23, 2024
Published, MCPRO Papers in Press, November 7, 2024, https://doi.org/10.1016/j.mcpro.2024.100875

## REFERENCES

1. Hunt, D. F., Bone, W. M., Shabanowitz, J., Rhodes, J., and Ballard, J. M. (1981) Sequence analysis of oligopeptides by secondary ion/collision activated dissociation mass spectrometry. *Anal. Chem.* **53**, 1704–1706
2. Fitton, J. E., Hunt, D. F., Marasco, J., Shabanowitz, J., Winston, S., and Dell, A. (1984) The amino acid sequence of delta haemolysin purified from a canine isolate of S. aureus. *FEBS Lett.* **169**, 25–29
3. Hunt, D. F., Buko, A. M., Ballard, J. M., Shabanowitz, J., and Giordani, A. B. (1981) Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer. *Biomed. Mass Spectrom.* **8**, 397–408
4. Hunt, D. F., Yates 3rd, J. R., Shabanowitz, J., Winston, S., and Hauer, C. R. (1986) Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 6233–6237
5. Hunt, D. F., Shabanowitz, J., Yates, J. R., 3rd, Zhu, N. Z., Russell, D. H., and Castro, M. E. (1987) Tandem quadrupole Fourier-transform mass spectrometry of oligopeptides and small proteins. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 620–623
6. Hunt, D. F. (2002) Personal commentary on proteomics. *J. Proteome Res.* **1**, 15–19
7. Yates, J. R., Stafford, G. C., Shabanowitz, J. Foreword: Donald F. Hunt. Inter. J. Mass Spectr., 259(1-3), vii-xi
8. Syka, J. E., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9528–9533
9. Hunt, D. F., Shabanowitz, J., and Bai, D. L. (2015) Peptide sequence analysis by electron transfer dissociation mass spectrometry: a web-based tutorial. *J. Am. Soc. Mass Spectrom.* **26**, 1256–1258
10. Medzihradszky, K. F., and Chalkley, R. J. (2015) Lessons in *de novo* peptide sequencing by tandem mass spectrometry. *Mass Spectrom. Rev.* **34**, 43–63
11. Kilpatrick, E. L., Liao, W., Camara, J. E., Turko, I. V., and Bunk, D. M. (2012) Expression and characterization of 15N-labeled human C-reactive protein in Escherichia coli and Pichia pastoris for use in isotope-dilution mass spectrometry. *Protein Expr. Purif.* **85**, 94–99
12. Yergey, J. A. (1983) A general approach to calculating isotopic distributions for mass spectrometry. *Int. J. Mass Spectrom. Ion Phys.* **52**, 337–349
13. [software] Senko, M. W. (1998) *IsoPro Version 3.1*. M.W. Senko, Sunnyvale, VA
14. Vyatkina, K., Wu, S., Dekker, L. J., VanDuijn, M. M., Liu, X., Tolic, N., *et al*. (2015) De novo sequencing of peptides from top-down tandem mass spectra. *J. Proteome Res.* **14**, 4450–4462
15. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989
16. Skinner, O. S., and Kelleher, N. L. (2015) Illuminating the dark matter of shotgun proteomics. *Nat. Biotechnol.* **33**, 717–718
17. Griss, J., Perez-Riverol, Y., Lewis, S., Tabb, D. L., Dianes, J. A., Del-Toro, N., *et al*. (2016) Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat. Methods* **13**, 651–656
18. Camacho, C., Boratyn, G. M., Joukov, V., Vera Alvarez, R., and Madden, T. L. (2023) ElasticBLAST: accelerating sequence search via cloud computing. *BMC Bioinformatics* **24**, 117