Title

Proteoform Analysis and Construction of Proteoform Families in Proteoform Suite

ISBN

9781071623244

Authors

Schaffer, Leah V
Shortreed, Michael R
Smith, Lloyd M

Publication Date

2022

DOI

10.1007/978-1-0716-2325-1_7

Peer reviewed

HHS Public Access

Author manuscript

Methods Mol Biol. Author manuscript; available in PMC 2023 January 01.

Published in final edited form as:

Methods Mol Biol. 2022 ; 2500: 67–81. doi:10.1007/978-1-0716-2325-1_7.

Proteoform Analysis and Construction of Proteoform Families in Proteoform Suite

Leah V Schaffer[1], **Michael R Shortreed**[1], **Lloyd M Smith**[2]

[1]Department of Chemistry, University of Wisconsin-Madison, Madison, WI, USA.

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

Abstract

Proteoform Suite is an interactive software program for the identification and quantification of intact proteoforms from mass spectrometry data. Proteoform Suite identifies proteoforms observed by intact-mass (MS1) analysis. In intact-mass analysis, unfragmented experimental proteoforms are compared to a database of known proteoform sequences and to one another, searching for mass differences corresponding to well-known post-translational modifications or amino acids. Intact-mass analysis enables proteoforms observed in the MS1 data without MS/MS (MS2) fragmentation to be identified. Proteoform Suite further facilitates the construction and visualization of proteoform families, which are the sets of proteoforms derived from individual genes. Bottom-up peptide identifications and top-down (MS2) proteoform identifications can be integrated into the Proteoform Suite analysis to increase the sensitivity and accuracy of the analysis. Proteoform Suite is open source and freely available at https://github.com/smith-chem-wisc/proteoform-suite.

Keywords

Proteoform; proteoform family; top-down proteomics; mass spectrometry; post-translational modification

1. Introduction

Much of the biological complexity in cells is due to variations at the protein level. Proteoforms are the different forms of a protein that arise due to biological processes in the cell, such as alternative splicing, amino acid variation and post-translational modifications (PTMs).[1] The set of proteoforms derived from the same gene is a proteoform family.[2] Proteoform identification is typically performed by top-down mass spectrometry, where intact proteins are analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS), which includes by definition fragmentation of the precursor.[3–5] However, due to sensitivity limitations, many proteoforms observed at the MS1-level are either not selected for MS2 fragmentation or do not have high enough quality MS2 fragmentation data for proteoform identification.

Correspondence: smith@chem.wisc.edu.

We developed the software program Proteoform Suite (https://github.com/smith-chem-wisc/proteoform-suite) to construct proteoform families and identify proteoforms based on analysis of MS1-level intact-mass data.[6] A list of unique experimental proteoforms from a sample is generated by aggregating results from top-down MS/MS search results (identified proteoforms) and/or MS1 deconvolution results (observed but unidentified proteoforms).[7] Aggregated experimental proteoform masses are compared to both a database of theoretical proteoform masses (experiment-theoretical comparisons) and to one another (experiment-experiment comparisons) by Proteoform Suite. Proteoform families are then constructed by grouping together experimental and theoretical proteoforms with mass differences corresponding to known PTMs and amino acids. In each proteoform family, the experiment-theoretical and experiment-experiment comparisons are used to identify additional experimental proteoforms by intact-mass. Proteoform families are visualized as a network of nodes (proteoform masses) and edges (mass differences corresponding to modifications or amino acid differences) using the software program Cytoscape.[8,9]

This chapter describes the workflow used to perform intact-mass analysis and construct proteoform families in Proteoform Suite. We first describe pre-processing steps performed on the raw mass spectrometry files to generate results for Proteoform Suite input. We then describe how to perform intact-mass analysis in Proteoform Suite and the visualization of proteoform families. Finally, we discuss remaining challenges in intact-mass analysis and caveats to be aware of when using Proteoform Suite.

Proteoform analysis in Proteoform Suite requires mass spectrometry data from an intact protein sample. This may include intact only data (MS1) or tandem MS/MS data, which contains both MS1 and MS2 spectra. See previous work for information on sample preparation, fractionation, and mass spectrometry settings.[10,11] To observe a greater number of proteoforms at the MS1-level, we recommend acquisition of both MS1-only and tandem MS/MS from the sample.

A release of Proteoform Suite can be downloaded from https://github.com/smith-chem-wisc/proteoform-suite. At least 8 GB of RAM is recommended, with more required for larger human databases. A 64-bit operating system should be utilized with .NET Core 3.1 installed. Release version 0.4.0 contains a user manual with detailed descriptions of the parameters and results on each page of Proteoform Suite.

2. Data Preprocessing

A typical Proteoform Suite analysis requires MS1 deconvolution results, top-down search results, and a protein database (Figure 1). The Proteoform Suite graphical user interface (GUI) opens on the Load Results page (Figure 2). On this page, results are loaded in for the analyses performed on subsequent pages (select Standard under Choose Analysis, labeled 1 in Figure 2). If necessary, this page can also be used to generate deconvolution and top-down results and to calibrate results. Sections 2.1 – 2.5 describe how to obtain results to input on the Load Results page using either external software programs or other analysis options on the Load Results page.

2.1 Deconvolution Results

Deconvolution results input into Proteoform Suite provide a list of observed experimental proteoform masses that are used for proteoform family construction and intact-mass identification of proteoforms (see Note 1). On the Load Results page under Standard analysis, set the drop-down menu (labeled 3 in Figure 2) to Deconvolution Results for Identification and add deconvolution results. There are three options to produce deconvolution results for Proteoform Suite:

1. Thermo Deconvolution 4.0 (see Thermo Fisher website for a quote and user guide). Run the Xtract algorithm for high resolution data, then open the result for each .raw file and export the results table. Save each file as an .xlsx in Microsoft Excel. The resulting .xlsx file is loaded under Deconvolution Results for Identification.
2. FLASHDeconv. FLASHDeconv is an ultrafast deconvolution algorithm developed by the OpenMS team for high resolution mass spectrometry data.[12] On the Load Results page of Proteoform Suite, select FLASHDeconv Deconvolution under Choose Analysis (labeled 1 in Figure 2). Input .mzML spectra files into the table. Set the desired parameters under Set Parameters (labeled 2 in Figure 2). Click the Deconvolute button (labeled 4 in Figure 2). For more advanced parameter options, a user can also externally run the command-line version of FLASHDeconv, available at https://www.openms.de/comp/FLASHDeconv/. The resulting .tsv file is loaded under Deconvolution Results for Identification.
3. Results from any external deconvolution software can also be input. Create a three-column tab-separated .tsv or .txt file: monoisotopic mass, intensity, retention time. The resulting .tsv or .txt file is loaded under Deconvolution Results for Identification.

For proteoform quantification, there is the option to label the Biological Replicate, Fraction, Technical Replicate, and Condition for each file. To change one of these labels for a single file, click the appropriate cell in the table. To change the label for more than one file or cell, select the cells, right click, enter a label, click Okay.

2.2 Top-Down Results

Top-down results input into Proteoform Suite provide a list of proteoforms identified by MS/MS analysis. These results are useful for improving intact-mass analysis by identifying additional proteoform families and proteoforms that might not be identified by intact-mass alone. On the Load Results page under Standard analysis, set the drop-down menu (labeled 3 in Figure 2) to Top-Down Hit Results and add top-down results. There are two options to produce top-down results for Proteoform Suite:

4. TDPortal. TDPortal is a high-throughput global proteome analysis software for top-down data[10] available through the National Resource for Translational and Developmental Proteomics. Request access to TDPortal at http://nrtdp.northwestern.edu/tdportal-request/. Under the Reports tab in TDViewer,

export the Hit Report. The resulting .xlsx file is loaded under Top-Down Hit Results.

5. MetaMorpheus. MetaMorpheus is an MS/MS search software program for both bottom-up and top-down high-resolution MS data. On the Load Results page of Proteoform Suite, select MetaMorpheus Top-Down Search under Choose Analysis (labeled 1 in Figure 2). Set the desired parameters under Set Parameters (labeled 2 in Figure 2). Set the drop-down menu (labeled 3 in Figure 2) to Spectra Files and add .raw or .mzML files. Set the drop-down menu1 to Protein Databases and add an .xml or .fasta database. Click the MetaMorpheus Top-Down Search button (labeled 4 in Figure 2). For more advanced parameter options or to search bottom-up data, externally run the GUI or command line versions of MetaMorpheus available at https://github.com/smith-chem-wisc/metamorpheus.The resulting AllPSMs.tsv file is loaded under Top-Down Hit Results.

2.3 Database

Download a protein database from UniProt (https://www.uniprot.org/proteomes/). Typically, only reviewed entries are employed but unreviewed entries can be included at the user's discretion. A database from MetaMorpheus generated from a bottom-up search and the global post-translational modification discovery strategy (G-PTM-D) can also be utilized, as previously described.[13–15] On the Load Results page under Standard analysis, set the drop-down menu (labeled 3 in Figure 2) to Protein Databases and add at least one database. There are two options for protein databases:

1. .xml or .xml.gz: contains annotated PTM information and subsequences
2. .fasta: optionally including protein isoforms

2.4 Bottom-Up Results

Bottom-up results input into Proteoform Suite provide a list of peptides identified by MS/MS analysis. To run a bottom-up search, download a release of MetaMorpheus and run a bottom-up search (https://github.com/smith-chem-wisc/metamorpheus). On the Load Results page under Standard analysis, set the drop-down menu (labeled 3 in Figure 2) to MetaMorpheus Bottom-Up Unique Peptides, and add the AllPeptides.psmtsv generated by the MetaMorpheus bottom-up search.

2.5 Mass and Retention Time Calibration

Mass and retention time calibration is an optional pre-processing step that is recommended when performing intact-mass analysis with deconvolution results.[7] Mass and retention time across runs can be calibrated for deconvolution and top-down results to improve mass accuracy and correct run-to-run variation, respectively. Calibrated files are output in the same file location as the input files; the calibrated results files can then be loaded on the Load Results page under Deconvolution Results for Identification and Top-Down Hit Results in the Standard analysis. For more details on how to perform calibration in Proteoform Suite, please see the user manual in Proteoform Suite release version 0.4.0.

3. Data Analysis

Proteoform Suite is an interactive software program for intact-mass analysis and proteoform family construction; each page is run in sequence, from left to right. To navigate to the next page, either click the right arrow in the top right of the GUI window or click the page tab name at the top of the GUI window. The user manual in release version 0.4.0 contains detailed information about each parameter and table column. The default parameters are set such that they will provide an acceptable starting point for most standard analyses.

1. Load Results. On this page, the user loads results files described in the Data Preprocessing section.
    1. Select Standard under Choose Analysis (labeled 1 in Figure 2).
    2. Under Load Data Using Drop Down Menu (labeled 3 in Figure 2), use the drop-down menu to select the result type to be loaded in (labeled 3 in Figure 2). Then use the Add button or drag-and-drop to add files of the selected result type to the table.
    3. Navigate to the Theoretical Proteoform Database page by clicking that page tab at the top of the GUI window or by clicking the right arrow at the top right of the GUI window.
2. Theoretical Proteoform Database. On this page, theoretical proteoforms are created using the file(s) input under Protein Databases on the Load Results page. The theoretical proteoform database includes theoretical proteoforms with combinations of annotated PTMs and subsequences. Theoretical proteoforms are used in the subsequent pages for identifying proteoforms by intact-mass analysis.
    1. Set the parameters as necessary. Parameters to note are the Average Mass checkbox (check if deconvolution results are reported as average masses instead of monoisotopic masses), the Carbamidomethylation checkbox (check if sample was carbamidomethylated), and the Max Modifications per Proteoform (increase or decrease to generate theoretical proteoforms with combinations of more or fewer annotated PTMs). Higher values for Max Modifications per Proteoform will include more theoretical proteoforms in the database for potential identification (see Experiment-Theoretical comparison) but may also result in an increased false discovery rate for intact-mass identifications (see Note 2).
    2. Click Run Page button at the top right of the GUI window.
    3. The main table will fill with a list of the theoretical proteoforms.
3. Top-Down. On this page, top-down results are read from the file(s) input under Top-Down Hit Results on the Load Results page. A top-down hit is a proteoform spectral match. Top-down hits are loaded, and then hits are aggregated into unique top-down proteoforms using both identification information and retention time. After aggregation, the theoretical database is supplemented with top-down

proteoform identifications not already present in the database (i.e., those with unannotated PTMs, unannotated subsequences, amino acid variations, or with a greater number of PTMs than the Max Modifications per Proteoform number on the Theoretical Proteoform Database page).

1. Set the parameters as necessary. Parameters to note are the Min. C-Score (the minimum proteoform characterization score[16] for TDPortal results) and the Ret. Time Tolerance (retention time used for merging top-down hits of the same proteoform identification). Higher values for the minimum C-score will exclude less characterized and potentially false top-down hits but may also exclude uncharacterized but true top-down identifications.
2. Click Run Page button at the top right of the GUI window.
3. The main table will fill with a list of the top-down proteoforms. Click a top-down proteoform to fill the bottom box with the sequence and PTM information. If bottom-up results were provided on the Load Results page, the second table will fill with bottom-up peptides that correspond to the selected top-down proteoform.

4. Raw Experimental Components. On this page, raw experimental components are read from the file(s) loaded under Deconvolution Results for Identification on the Load Results page. A raw experimental component is an individual deconvoluted proteoform observation at the MS1-level, as reported in the deconvolution result file(s). Deconvolution artifacts include missed monoisotopic mass errors and charge state harmonics are corrected using the mass tolerance set on the page.
    1. Set the parameters as necessary. Parameters to note are the cosine threshold between per-charge-intensities and fitted gaussian distribution and the cosine threshold between averagine and observed isotope pattern[12]. These parameters determine the thresholds for reported FLASHDeconv results to be included as raw experimental components. Higher values will exclude more raw experimental components that correspond to experimental artifacts but may also exclude raw experimental components that correspond to real observed experimental proteoforms.
    2. Click the Run Page button at the top right of the GUI window.
    3. The main table will fill with a list of raw experimental components. Click a raw experimental component to fill the bottom table with the observed charge state information.
5. Aggregated Proteoforms. On this page, experimental proteoforms are created by aggregating raw experimental components. These experimental proteoforms are analyzed and identified by intact-mass analysis and used to construct proteoform families. The Add Top-Down Proteoforms checkbox provides the option to supplement this list with top-down identified experimental proteoforms from the Top-Down page.

1. Set the parameters as necessary. Parameters to note are the Mass Tolerance and Ret. Time Tolerance, which are used to determine the mass and retention time tolerances used to merge raw experimental components into an experimental proteoform. Additionally, the Minimum Required Observations setting can be used to determine the minimum number of biological and/or technical replicates an experimental proteoform must have been observed in (determined by the raw experimental component files input and optionally labeled on the Load Results page). Higher values will exclude more raw experimental components that correspond to experimental artifacts but may also exclude raw experimental components that correspond to real observed experimental proteoforms.
2. Click the Run Page button at the top right of the GUI window.
3. The main table will fill with a list of aggregated experimental proteoforms. Click an experimental proteoform to fill the bottom table with raw experimental components aggregated into the selected proteoform. If the selected proteoform is a top-down proteoform, the bottom table will fill with the top-down hits aggregated into the selected proteoform.

6. Experiment-Theoretical Comparison. On this page, masses of experimental proteoforms created on the Aggregated Proteoforms page are compared to masses of theoretical proteoforms created on the Theoretical Proteoforms page, generating a list of experiment-theoretical pairs. Each experiment-theoretical pair has a mass difference (delta mass) between the experimental proteoform and the theoretical proteoform in the pair; pairs are generated for mass differences corresponding to a known single modification or set of modifications. Proteoform Suite creates a histogram of the mass differences for all generated experiment-theoretical pairs and a corresponding list of delta mass peaks observed in the histogram. Each experiment-theoretical pair with a delta mass in a delta mass peak accepted by the user is utilized to construct proteoform families (see Proteoform Families below).

    1. Set the parameters as necessary. There are several parameters of note on this page. The Peak Width Base determines the width of peaks when generating the delta mass histogram; a larger value will enable more experiment-theoretical pairs to be included in each peak but will also increase the false discovery rate of each peak. The checkbox Notch Search provides the option to perform a notch search.[17] If checked, a tolerance parameter will become visible, and experiment-theoretical pairs will only be generated if within tolerance from an accepted notch (corresponding to an exact match at 0 Da, a known PTM, a known PTM combination or an amino acid mass). The larger the tolerance, the more experiment-theoretical pairs that will be included at each notch, but also the false discovery rate will potentially be larger. The checkbox Add

Top-Down Theoreticals determines if theoretical proteoforms that were added to the database on the Top-Down page (not present in the original database) are included in the analysis.

2. Click the Run Page button at the top right of the GUI window.

3. The top right table will fill with all experiment-theoretical pairs. The bottom graph will display the delta mass histogram created from the experiment-theoretical pairs. The top left table will fill with the list of delta mass peaks from the histogram.

4. Browse the list of delta mass peaks (top left table). Accept peaks that have an acceptable false discovery rate and that correspond to common/likely modifications or amino acid differences (see Note 3). The Peak Threshold determines the minimum number of experiment-theoretical pairs that must be grouped in a delta mass peak in order for it to be accepted. This value can be changed to accept/unaccept peaks, or individual peaks can be accepted/unaccepted by checking/unchecking the accepted checkbox for the peak. Typically, only the delta mass peak closest to 0 Da (exact match experiment-theoretical pairs) is accepted. A higher number of accepted peaks results in more proteoform family connections and intact-mass identifications but also a larger false discovery rate.

7. Experiment-Experiment Comparisons. On this page, masses of experimental proteoforms created on the Aggregated Proteoforms page are compared to one another within a set retention time difference tolerance, generating a list of experiment-experiment pairs. Each experiment-experiment pair has a mass difference (delta mass) between the two experimental proteoforms in the pair. Proteoform Suite creates a histogram of the mass differences for all generated experiment-experiment pairs and a corresponding list of delta mass peaks observed in the histogram. Each experiment-experiment pair with a delta mass in a delta mass peak accepted by the user is utilized to construct proteoform families (see Proteoform Families below).

    1. Set the parameters as necessary. There are several parameters of note on this page. The Peak Width Base determines the width of peaks when generating the delta mass histogram; a larger value will enable more experiment-experiment pairs to be included in each peak but will also increase the false discovery rate of each peak. The checkbox Notch Search provides the option to perform a notch search[17]. If checked, a tolerance parameter will become visible, and experiment-experiment pairs will only be generated if within tolerance from an accepted notch (corresponding to a known PTM or a known PTM set). Larger tolerance values result in more experiment-experiment pairs included at each notch, but also increase the false discovery rate. The Max Retention Time Difference determines the maximum allowed difference in retention time between two experimental proteoforms to be eligible

to form an experiment-experiment pair. Larger values result in more experiment-experiment pairs, but also increase the false discovery rate.

2. Click the Run Page button at the top right of the GUI window.

3. The top right table will fill with all experiment-experiment pairs. The bottom graph will display the delta mass histogram created from the experiment-experiment pairs. The top left table will fill with the list of delta mass peaks from the histogram.

4. Browse the list of delta mass peaks (top left table). Accept peaks that have an acceptable false discovery rate and that correspond to common/likely modifications or amino acid differences (see Note 3). The Peak Threshold determines the minimum number of experiment-experiment pairs that must be grouped in a delta mass peak in order for it to be accepted. This value can be changed to accept/unaccept peaks, or individual peaks can be accepted/unaccepted by checking/unchecking the accepted checkbox for the peak. Typically, only the most abundant five to ten delta mass peaks corresponding to known and common PTMs and amino acid differences are accepted. A higher number of accepted peaks results in more proteoform family connections and intact-mass identifications but can also result in a larger false discovery rate.

8. Proteoform Families. On this page, accepted experiment-theoretical and experiment-experiment pairs are used to construct proteoform families. Experimental proteoforms are identified by intact-mass analysis; beginning with each theoretical proteoform in each family, connections between proteoforms are used to identify proteoforms first from experiment-theoretical pairs and then from subsequent experiment-experiment pairs. A false discovery rate is calculated for intact-mass identifications (see Note 4). A script for Cytoscape[8,9] can be exported to visualize proteoform families as a network of nodes (proteoforms) and edges (mass differences between proteoforms corresponding to modifications or amino acid differences).

    1. Set the parameters as necessary. There are several parameters to note on this page. If the checkbox Only Assign Common/Known Mods is checked, a modification will only be considered for each proteoform identification if annotated for that protein in the theoretical database or if a common modification (e.g., acetylation, methylation, phosphorylation). The Use ppm tolerance checkbox can be checked to require a proteoform identification to be within a certain ppm mass error tolerance.

    2. Click the Run Page button at the top right of the GUI window.

    3. The main table will fill with a list of proteoform families. Click a row in the table and the bottom table will fill with either a list

of experimental proteoforms, theoretical proteoforms, or proteoform relations depending on which column is selected.

4. To build proteoform families on this page, click the Browse button on the right and select a file path. Set the visualization parameters as desired (top right of the screen). Select either Build All Families in Cytoscape (writes a script to visualize all proteoform families) or Build Selected Families in Cytoscape (writes a script to visualize proteoform families selected in the main table).

9. Results

    1. Identified Proteoforms. Navigate to the Identified Proteoforms page to view all identified experimental proteoforms. This page displays two tables with intact-mass identifications (from deconvolution results and proteoform family construction) and top-down identifications (from input top-down results). If bottom-up results were provided on the Load Results page, click either an intact-mass or a top-down proteoform to fill the bottom table with bottom-up peptides that correspond to the selected proteoform identification.

    2. Results Summary. Navigate to the Results Summary page to view a summary of results from Proteoform Suite. Select a file path for the Results Folder and click the Save All button. A summary of the results text file, a method .xml file for subsequent runs, Cytoscape scripts, and various results tables will export. See the user manual in Proteoform Suite release 0.4.0 for a list and description of exported results.

    3. Visualize Proteoform Families. Scripts for Cytoscape are exported on either the Proteoform Families page or the Results Summary page. To visualize proteoform families in Cytoscape, install Cytoscape version 3.5.0 at https://cytoscape.org. In Cytoscape, select Execute Command File under Tools, then select the cytoscape_script_timestamp.tsv file output by Proteoform Suite.

4. Notes

1. Intact-mass analysis is highly dependent on the quality of the results of deconvolution. Missed monoisotopic mass errors, where the reported mass is one or more isotopic units from the monoisotopic mass, are a challenging issue for intact-mass analysis. Additionally, artifactual masses reported by deconvolution also increase the false discovery rate. The user should do their best to ensure high quality MS1 data and deconvolution results.

2. It is important for the user to consider the database size when performing an intact-mass analysis. There is a tradeoff between expanding the database to include more PTM combinations and amino acid variants and limiting increases to the FDR. Typically, our studies have used two or three PTM combinations per theoretical proteoform; this is because proteoforms with more modifications

are able to be identified via the experiment-experiment comparisons if a less modified proteoform is also identified from the same proteoform family. Proteoforms that only exist in ultra-modified forms are thus best identified by top-down MS/MS analysis.

3. When accepting and unaccepting peaks in the experiment-theoretical and experiment-experiment comparisons, we have found it is best to be conservative and accept the most abundant peaks closest to the delta mass of known, common, and expected post-translational modifications. If a large number of high-FDR and high mass error experiment-theoretical and experiment-experiment peaks are employed, the effect will be an increase in FDR and proteoform families may be contaminated with incorrect proteoform members. At this time, less common modifications are better identified either through top-down MS/MS analysis or by using bottom-up analysis to add the modification to the protein sequence, enabling it to be included in the theoretical proteoform database and thus identifiable through an exact match in the experiment-theoretical comparison.

4. In intact-mass analysis with Proteoform Suite, a remaining challenge is identifying previously unidentified, observed proteoforms by intact-mass alone without exceeding a reasonable false discovery rate. Proteoform Suite calculates a global false discovery rate by performing decoy experiment-theoretical and experiment-experiment comparisons to construct decoy proteoform families, described in detail previously.[6,7] This globally calculated false discovery rate is dependent on a several aspects of the analysis, including the size of the theoretical database, the deconvolution results, the top-down results, and the settings utilized when performing the experiment-theoretical and experiment-experiment comparisons.

Acknowledgements

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIH) under Award Number R35GM126914. L.V.S. was supported by the NIH Biotechnology Training Program, T32GM008349.

References

1. Smith LM, Kelleher NL, Consortium for Top Down Proteomics (2013) Proteoform: a single term describing protein complexity. Nat Methods 10 (3):186–187. doi:10.1038/nmeth.2369 [PubMed: 23443629]
2. Shortreed MR, Frey BL, Scalf M, Knoener RA, Cesnik AJ, Smith LM (2016) Elucidating Proteoform Families from Proteoform Intact-Mass and Lysine-Count Measurements. J Proteome Res 15 (4):1213–1221. doi:10.1021/acs.jproteome.5b01090 [PubMed: 26941048]
3. Cai W, Tucholski TM, Gregorich ZR, Ge Y (2016) Top-down Proteomics: Technology Advancements and Applications to Heart Diseases. Expert Rev Proteomics 13 (8):717–730. doi:10.1080/14789450.2016.1209414 [PubMed: 27448560]
4. Chen B, Brown KA, Lin Z, Ge Y (2018) Top-Down Proteomics: Ready for Prime Time? Anal Chem 90 (1):110–127. doi:10.1021/acs.analchem.7b04747 [PubMed: 29161012]
5. Toby TK, Fornelli L, Kelleher NL (2016) Progress in Top-Down Proteomics and the Analysis of Proteoforms. Annu Rev Anal Chem (Palo Alto Calif) 9 (1):499–519. doi:10.1146/annurev-anchem-071015-041550 [PubMed: 27306313]

6. Cesnik AJ, Shortreed MR, Schaffer LV, Knoener RA, Frey BL, Scalf M, Solntsev SK, Dai Y, Gasch AP, Smith LM (2018) Proteoform Suite: Software for Constructing, Quantifying, and Visualizing Proteoform Families. J Proteome Res 17 (1):568–578. doi:10.1021/acs.jproteome.7b00685 [PubMed: 29195273]
7. Schaffer LV, Shortreed MR, Cesnik AJ, Frey BL, Solntsev SK, Scalf M, Smith LM (2018) Expanding Proteoform Identifications in Top-Down Proteomic Analyses by Constructing Proteoform Families. Anal Chem 90 (2):1325–1333. doi:10.1021/acs.analchem.7b04221 [PubMed: 29227670]
8. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13 (11):2498–2504. doi:10.1101/gr.1239303 [PubMed: 14597658]
9. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27 (3):431–432. doi:10.1093/bioinformatics/btq675 [PubMed: 21149340]
10. Toby TK, Fornelli L, Kristina S, DeHart CJ, Fellers RT (2019) A comprehensive pipeline for translational top-down proteomics from a single blood draw. Nature Protocols 14:119–152. doi:10.1038/s41596-018-0085-7 [PubMed: 30518910]
11. Donnelly DP, Rawlins CM, DeHart CJ, Fornelli L, Schachner LF, Lin Z, Lippens JL, Aluri KC, Sarin R, Chen B, Lantz C, Jung W, Johnson KR, Koller A, Wolff JJ, Campuzano IDG, Auclair JR, Ivanov AR, Whitelegge JP, Pasa-Tolic L, Chamot-Rooke J, Danis PO, Smith LM, Tsybin YO, Loo JA, Ge Y, Kelleher NL, Agar JN (2019) Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. Nat Methods 16 (7):587–594. doi:10.1038/s41592-019-0457-0 [PubMed: 31249407]
12. Jeong K, Kim J, Gaikwad M, Hidayah SN, Heikaus L, Schluter H, Kohlbacher O (2020) FLASHDeconv: Ultrafast, High-Quality Feature Deconvolution for Top-Down Proteomics. Cell Syst 10 (2):213–218 e216. doi:10.1016/j.cels.2020.01.003 [PubMed: 32078799]
13. Dai Y, Shortreed MR, Scalf M, Frey BL, Cesnik AJ, Solntsev S, Schaffer LV, Smith LM (2017) Elucidating Escherichia coli Proteoform Families Using Intact-Mass Proteomics and a Global PTM Discovery Database. J Proteome Res 16 (11):4156–4165. doi:10.1021/acs.jproteome.7b00516 [PubMed: 28968100]
14. Dai Y, Buxton KE, Schaffer LV, Miller RM, Millikin RJ, Scalf M, Frey BL, Shortreed MR, Smith LM (2019) Constructing Human Proteoform Families Using Intact-Mass and Top-Down Proteomics with a Multi-Protease Global Post-Translational Modification Discovery Database. J Proteome Res 18 (10):3671–3680. doi:10.1021/acs.jproteome.9b00339 [PubMed: 31479276]
15. Li Q, Shortreed MR, Wenger CD, Frey BL, Schaffer LV, Scalf M, Smith LM (2017) Global Post-Translational Modification Discovery. J Proteome Res 16 (4):1383–1390. doi:10.1021/acs.jproteome.6b00034 [PubMed: 28248113]
16. LeDuc RD, Fellers RT, Early BP, Greer JB, Thomas PM, Kelleher NL (2014) The C-score: a Bayesian framework to sharply improve proteoform scoring in high-throughput top down proteomics. J Proteome Res 13 (7):3231–3240. doi:10.1021/pr401277r [PubMed: 24922115]
17. Solntsev SK, Shortreed MR, Frey BL, Smith LM (2018) Enhanced Global Post-translational Modification Discovery with MetaMorpheus. J Proteome Res 17 (5):1844–1851. doi:10.1021/acs.jproteome.7b00873 [PubMed: 29578715]

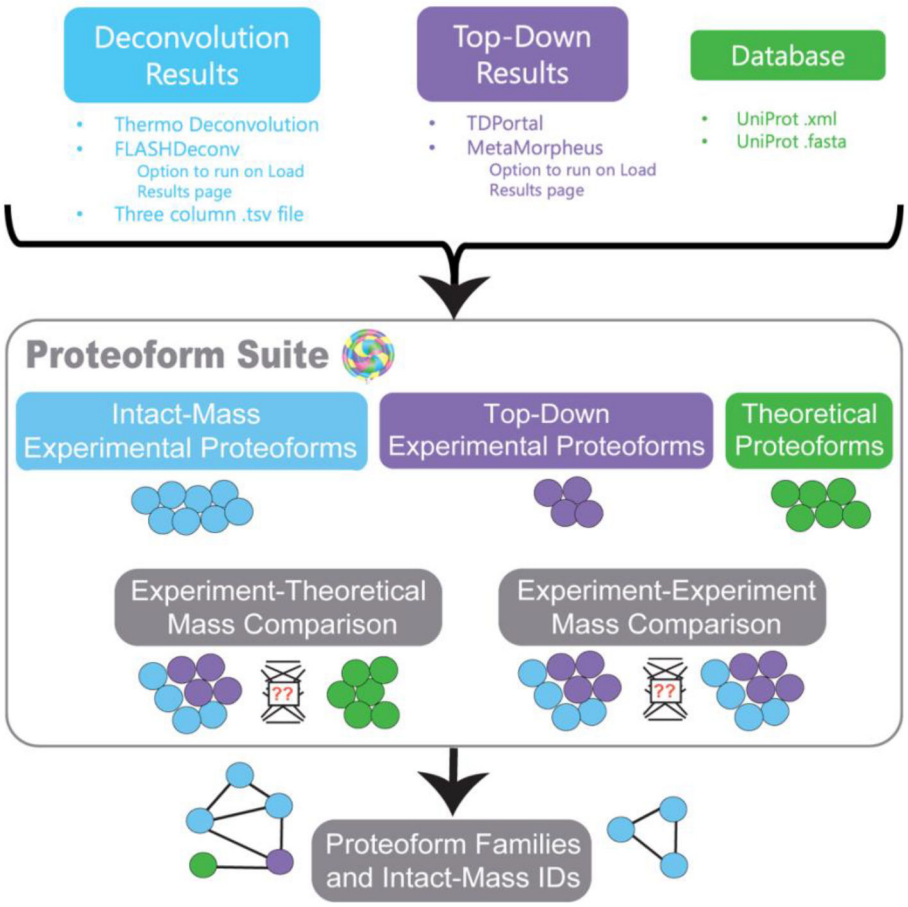

Figure 1.
Overview of inputs and outputs for Proteoform Suite analysis.

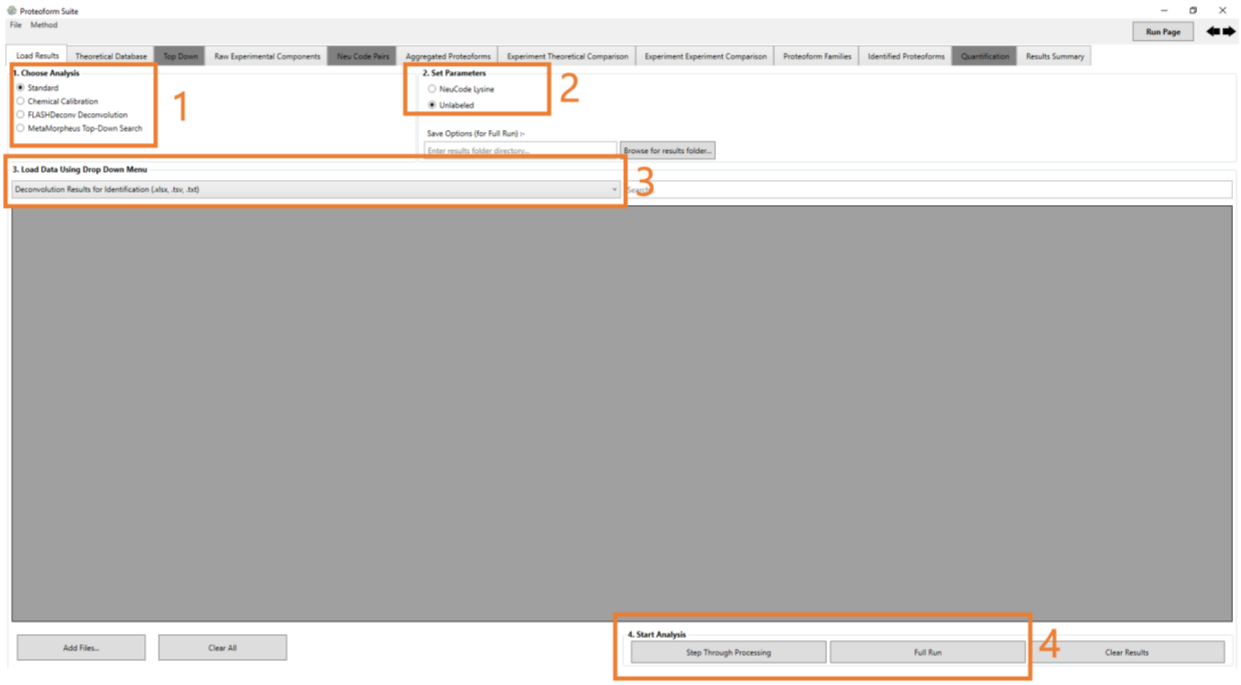

Figure 2.
Load Results page in Proteoform Suite.