MASH Native: A Unified Solution for Native Top-Down Proteomics Data Processing

Eli J. Larson[1], Melissa R. Pergande[2], Michelle E. Moss[2], Kalina J. Rossler[2], R. Kent Wenger[2,3],

Boris Krichel[2], Harini Josyer[2], Jake A. Melby[1], David S. Roberts[1], Kyndalanne Pike[1], Zhuoxin

Shi[2], Hsin-Ju Chan[1], Bridget Knight[1], Holden T. Rogers[1], Kyle A. Brown[1], Irene M. Ong[4,5,6],

Kyowon Jeong[7], Michael Marty[8], Sean J. McIlwain[4,5*], Ying Ge[1,2,3*]

[1]Department of Chemistry, University of Wisconsin-Madison, Madison, WI 53705, USA.

[2]Department of Cell and Regenerative Biology, University of Wisconsin-Madison, Madison, WI 53705, USA.

[3]Human Proteomics Program, School of Medicine and Public Health, University of Wisconsin-Madison, Madison, WI 53705, USA.

[4]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53705, USA.

[5]University of Wisconsin Carbone Cancer Center, University of Wisconsin-Madison, Madison, WI 53705, USA.

[6]Department of Obstetrics and Gynecology, University of Wisconsin-Madison, Madison, WI 53705, USA.

[7]Department of Applied Bioinformatics, University of Tübingen, Tübingen, Germany 72704

[8]Department of Chemistry and Biochemistry, University of Arizona, Tucson, AZ, USA 85719

*Correspondence should be addressed to S.J.M. (sean.mcilwain@wisc.edu) or Y.G. (ying.ge@wisc.edu).

1   **<u>Abstract</u>**

2   *Motivation*: Native top-down proteomics (nTDP) integrates native mass spectrometry (nMS) with

3   top-down proteomics (TDP) to provide comprehensive analysis of protein complexes together with

4   proteoform identification and characterization. Despite significant advances in nMS and TDP

5   software developments, a unified and user-friendly software package for analysis of nTDP data

6   remains lacking.

7   *Results*: We have developed MASH Native to provide a unified solution for nTDP to process

8   complex datasets with database searching capabilities in a user-friendly interface. MASH Native

9   supports various data formats and incorporates multiple options for deconvolution, database

10   searching, and spectral summing to provide a "one-stop shop" for characterizing both native

11   protein complexes and proteoforms.

12   *Availability and implementation*: The MASH Native app, video tutorials, written tutorials and

13   additional   documentation   are   freely   available   for   download   at

14   https://labs.wisc.edu/gelab/MASH_Explorer/MASHSoftware.php. All data files shown in user

15   tutorials are included with the MASH Native software in the download .zip file.

16

17   *Contact*: sean.mcilwain@wisc.edu or ying.ge@wisc.edu

18   *Supplementary information*: Supplementary data have been submitted as separate documents for

19   review.

20

21

1 **<u>Introduction</u>**

2   Native mass spectrometry (nMS) analyzes intact proteins and protein complexes under

3   non-denaturing conditions to preserve their tertiary structure and non-covalent interactions in the

4   gas phase, which has emerged as a powerful structural biology tool to define protein structure-

5   function relationships (Loo, 1997; Sharon and Robinson, 2007; Leney and Heck, 2017; Keener *et*

6   *al.*, 2021; Karch *et al.*, 2022). Native top-down proteomics (nTDP) integrates nMS with top-down

7   proteomics (TDP) (Catherman *et al.*, 2014; Toby *et al.*, 2016; Chen *et al.*, 2018; Melby *et al.*,

8   2021), which enables structural characterization of protein complexes together with proteoform

9   sequencing to locate non-covalent ligand binding sites, posttranslational modifications (PTMs),

10   and mutations (Li *et al.*, 2018; Zhou *et al.*, 2020; Karch *et al.*, 2022; Jooß *et al.*, 2022). nTDP first

11   measures intact proteins and protein complexes under non-denaturing conditions (MS1) then

12   directly fragments proteins and protein complexes in the gas phase (MS2) to obtain primary

13   sequence information from a single dissociation event (Li *et al.*, 2018). Alternatively, nTDP may

14   be implemented in the "complex-down" mode using two separate dissociation events: 1)

15   dissociation of intact protein complexes (MS1) into protein subunits (MS2') by low-energy

16   collision-induced dissociation (CID) or surface induced dissociation (SID), and 2) fragmentation

17   of subunits (MS3) by tandem mass spectrometry techniques such as high-energy CID, electron

18   capture dissociation (ECD), electron transfer dissociation (ETD) or ultraviolet photodissociation

19   (UVPD) to provide primary sequence coverage and localize modifications (Skinner *et al.*, 2018b;

20   Stiving *et al.*, 2019; Jooß *et al.*, 2022).

21   Currently one of the major challenges in nTDP is the analysis of complex nTDP datasets

22   which include both isotopically resolved and isotopically unresolved MS1 and MS2' spectra as

23   well as the complicated MS2 and MS3 data, and difficulties in database searching. Although

multiple software packages have been developed for nMS of known proteins and complexes (Marty *et al.*, 2015; Cleary *et al.*, 2016, 2018; Reid *et al.*, 2018), the lack of any MS2/MS3 fragmentation assignment and database searching prevent the identification of unknown proteins. Meanwhile, significant efforts have been allocated towards the development of software packages for denatured TDP with capability in analyzing complicated MS2/MS3 datasets with database search algorithms to identify unknown proteins (Sun *et al.*, 2016; Kou *et al.*, 2016; Fellers *et al.*, 2015; Cai *et al.*, 2016; Wu *et al.*, 2020), but these denatured TDP software packages lack the capability to analyze the isotopically unresolved MS1/MS2' that are characteristic of nMS data. Hence, there is a critical need for a universal software package to address this major challenge in nTDP that can process MS1, MS2, MS2' and MS3 datasets with database search capabilities.

Herein, we introduce MASH Native (https://labs.wisc.edu/gelab/MASH_Explorer/MASHNativeSoftware.php), a unified solution for nTDP which can process isotopically unresolved MS1 and MS2' data together with isotopically resolved MS1, MS2, and MS3 deconvolution and database searching (**Figure 1**). MASH Native supports various nTDP applications in both targeted mode to characterize known proteins and discovery mode to identify unknown native proteins. It supports various MS file types with different vendor formats and integrates multiple deconvolution/search algorithms into one package. We detail the functions and features of MASH Native and provide examples of processing nTDP data to showcase its capabilities as a "one-stop shop" for nTDP.

<u>Results</u>

The MASH Native user interface is a multithreaded Windows desktop application written under a .NET framework environment in Visual Studio using the C# programming language (Wu

et al., 2020). MASH Native provides universal MS file support through ProteoWizard's file conversion engine, MSConvert (Chambers *et al.*, 2012), and directly imports both vendor-specific MS file types (Thermo *.RAW, Bruker *.d/*.baf/*.ascii) and general file formats (*.mgf, *.mzML, *.mzXML). It is recommended that users perform MASH Native data processing on a computer with at least 4 GB of memory ensure optimal function of all included algorithms and workflows, however, some deconvolution algorithms may require additional memory for multi-scan, isotopically resolved deconvolution events. The latest version of MASH Native is freely-available for download on the MASH website (https://labs.wisc.edu/gelab/MASH_Explorer/MASHNativeSoftware.php) along with licensing information, and written and video user tutorials (also included in the "Supporting Documents for Users" section of the Supplementary Information). All data files used to generate these tutorials are freely available for download on MassIVE as a complete submission (MSV000091693, doi:10.25345/C5NP1WV0N).

MASH Native software can deconvolute both isotopically resolved and isotopically unresolved data at the MS1, MS2, and MS3 level and enables database searching of nTDP results (**Figure 1, Table S1**). It can process nTDP, nMS, and complex-down proteomics data using multiple deconvolution and database search algorithms with flexible data output options (**Figure S1**). It also maintains the functions and capabilities previously developed for denaturing TDP so users can process both nTDP and TDP in the same software. To address challenges with low signal-to-noise (S/N) ratios of intact and fragment mass spectra, MASH Native includes a variety of spectral summing algorithms that may be applied prior to data processing workflows (**Figure S2 and S3**). To deconvolute isotopically unresolved MS1 spectra, MASH Native includes UniDec (Marty *et al.*, 2015), a powerful deconvolution algorithm, to characterize both isotopically

1   unresolved and isotopically resolved nMS data (**Figure S4**). Isotopically resolved spectral

2   deconvolution can also be performed in MASH Native (**Figure S5**), including TopFD (Kou *et al.*,

3   2016), MsDeconv (Liu *et al.*, 2010), eTHRASH (Horn *et al.*, 2000), and pParseTD (Yuan *et al.*,

4   2012). Users may also import previously deconvoluted results from external deconvolution

5   algorithms, such as FLASHDeconv (Jeong *et al.*, 2020), ProMEX (Park *et al.*, 2017) or Maximum

6   Entropy (Ferrige *et al.*, 1991). Deconvolution results of separate deconvolution workflows can be

7   combined into a single output table, allowing users to view MS1, MS2, and MS3 results

8   simultaneously and combine multiple deconvolution types to improve protein sequence coverage

9   (Mcilwain *et al.*, 2020). Results of deconvolution may be searched against a user-selected

10   *.FASTA file or user-defined protein sequence with TopPIC (Kou *et al.*, 2016), MS-Align+ (Liu

11   *et al.*, 2012), or pTop (Sun *et al.*, 2016) to identify proteoforms in a complex mixture. Search

12   results are reported as both gene-level and proteoform-level identifications. Identified proteoforms

13   are scored and ranked, with scoring techniques varying for each algorithm (Kou *et al.*, 2016; Liu

14   *et al.*, 2012; Basharat *et al.*, 2020; Sun *et al.*, 2016). Search results generated through MASH

15   Native or from additional search tools such as MSPathFinderT (Park *et al.*, 2017), may then be

16   imported in MASH Native to view identifications, generate fragment ion maps, view fragment

17   ions, and validate for all identified proteins and proteoforms.

18   To facilitate high-throughput data analysis, user-defined MASH Native processing

19   workflows can be designed, saved, and queued to allow batch processing of data files using two

20   different approaches: Discovery and Targeted Mode. Discovery Mode facilitates identification of

21   unknown proteins though database searching, a critical processing feature absent from current

22   nMS or native top-down software tools. This mode combines MS1 processing with isotopically

23   resolved MS2 or MS3 deconvolution and database searching in a single workflow for nTDP

datasets (**Figure S6**). To demonstrate MASH Native Discovery Mode for data processing, we accessed and reanalyzed data files from a previously published nTDP dataset of endogenous protein complex previously published by Kelleher and co-workers (MassIVE dataset # MSV000080328) (Skinner *et al.*, 2018a). The workflow to identify and characterize subunits of this complex is shown in figure S6A. Deconvolution of both the MS1 and MS2' spectra by UniDec finds the intact complex mass and released subunit masses. Subsequent isotopically resolved MS3 deconvolution by eTHRASH and database searching with TopPIC identified the two subunits and localized modifications sites on each subunit. This underlines that MASH Native is capable of analyzing complex nTDP data in the Discovery Mode. To identify novel complexes using a complex-down approach, users must begin at the MS3 level by database searching. Next, identified subunits are matched to associated MS2' spectra with intact subunit masses to protein complex interactors. Finally, users must match the detected MS1 mass by testing different stoichiometries of each detected subunit to determine complex stoichiometry and composition. Automation of this process will eliminate the need for manual testing of novel complexes in future MASH Native releases.

Targeted Mode allows users to comprehensively analyze native top-down or complex-down data for a known protein/protein complex, confirm results generated in Discovery Mode, or potentially find new possible complex associations with database searching. At the MS1 and MS2' level, MASH Native enables isotopically unresolved and isotopically resolved native deconvolution through UniDec (Marty *et al.*, 2015). Deconvolution and searching of MS2 or MS3 data in Targeted Mode may be performed using all high-resolution deconvolution algorithms and database search options (*vide supra*). We have used MASH Native to process a native top-down MS dataset of the bovine glutamate dehydrogenase (GDH) hexamer previously published by Loo

1   and co-workers (Li *et al.*, 2018) to demonstrated the utility of this targeted workflow (**Figure S7**).

2   MASH Native allowed isotopically unresolved MS1 deconvolution and isotopically resolved MS2

3   deconvolution along with sequence mapping and data visualization in a single software package

4   (**Figure S7**). Recently, our group has demonstrated the utility of MASH for targeted analysis in a

5   complex-down workflow for a native cysteine-linked antibody-drug conjugate (ADC) (**Figure S8**)

6   (Larson *et al.*, 2021). The presence of intrachain disulfide bonds limits the fragmentation

7   efficiency of the ADC and reduces sequence coverage by terminal fragment assignment. MASH

8   Native incorporates searching and assignment of internal fragment ions, increasing sequence

9   coverage and revealing sequence coverage of regions bounded by disulfide bonds (**Figure S9**) to

10 provide additional higher-order structural information for proteins and complexes (Lantz *et al.*,

11 2021, 2022).

12

13 **<u>Conclusion</u>**

14   MASH Native provides a unified software solution for the analysis of a variety of complex

15 nTDP data for the first time. As a freely available and universal processing tool, MASH Native is

16 a "one-stop shop" for nTDP data processing that can handle a variety of complex nTDP datasets

17 including isotopically unresolved and isotopically MS1, MS2', MS2, and MS3 in both Discovery

18 and Targeted Modes with database search algorithms as well as data visualization and validation

19 in a user-friendly interface. It can process raw data from various vendor formats and integrates

20 multiple deconvolution/search algorithms into one package. MASH Native has been well-

21 recognized since its release on April 7, 2022 (Liu *et al.*, 2022), and downloaded more than 1,400

22 times by users all around the world (66 % from North America, 22 % from Europe, 7 % from Asia,

23 4% from Oceania, 0.6% from South America, and 0.4% from Africa) (**Figure S10**). As the nTDP

community gains momentum to grow rapidly, MASH Native will play an increasingly important role to streamline nTDP data processing and accelerate the use of nTDP in structural biology and biomedical applications.

References

Basharat,A.R. *et al.* (2020) EnvCNN: A Convolutional Neural Network Model for Evaluating Isotopic Envelopes in Top-Down Mass-Spectral Deconvolution. *Anal. Chem.*, **92**, 7778–7785.

Cai,W. *et al.* (2016) MASH suite pro: A comprehensive software tool for top-down proteomics. *Mol. Cell. Proteomics*, **15**, 703–714.

Catherman,A.D. *et al.* (2014) Top Down proteomics: Facts and perspectives. *Biochem. Biophys. Res. Commun.*, **445**, 683–693.

Chambers,M.C. *et al.* (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.*, **30**, 918–920.

Chen,B. *et al.* (2018) Top-Down Proteomics: Ready for Prime Time? *Anal. Chem.*, **90**, 110–127.

Cleary,S.P. *et al.* (2018) Extracting Charge and Mass Information from Highly Congested Mass Spectra Using Fourier-Domain Harmonics. *J. Am. Soc. Mass Spectrom.*, 31–39.

Cleary,S.P. *et al.* (2016) Fourier Analysis Method for Analyzing Highly Congested Mass Spectra of Ion Populations with Repeated Subunits Sean. *Anal. Chem*, **88**, 6205–6213.

Fellers,R.T. *et al.* (2015) ProSight Lite: Graphical software to analyze top-down mass spectrometry data. *Proteomics*, **15**, 1235–1238.

Ferrige,A.G. *et al.* (1991) Maximum entropy deconvolution in electrospray mass spectrometry. *Rapid Commun. Mass Spectrom.*, **5**, 374–377.

Horn,D.M. *et al.* (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.*, **11**, 320–332.

1   Jeong,K. *et al.* (2020) FLASHDeconv: Ultrafast, High-Quality Feature Deconvolution for Top-

2   Down Proteomics. *Cell Syst.*, **10**, 213-218.e6.

3   Jooß,K. *et al.* (2022) Native Mass Spectrometry at the Convergence of Structural Biology and

4   Compositional Proteomics. *Acc. Chem. Res.*

5   Karch,K.R. *et al.* (2022) Native Mass Spectrometry: Recent Progress and Remaining Challenges.

6   *Annu. Rev. Biophys.*, **51**, 157–179.

7   Keener,J.E. *et al.* (2021) Native Mass Spectrometry of Membrane Proteins. *Anal. Chem.*, **93**,

8   583–597.

9   Kou,Q. *et al.* (2016) TopPIC: A software tool for top-down mass spectrometry-based proteoform

10   identification and characterization. *Bioinformatics*, **32**, 3495–3497.

11   Lantz,C. *et al.* (2021) ClipsMS: An Algorithm for Analyzing Internal Fragments Resulting from

12   Top-Down Mass Spectrometry. *J. Proteome Res.*, **20**, 1928–1935.

13   Lantz,C. *et al.* (2022) Native Top-Down Mass Spectrometry with Collisionally Activated

14   Dissociation Yields Higher-Order Structure Information for Protein Complexes.

15   Larson,E.J. *et al.* (2021) High-Throughput Multi-attribute Analysis of Antibody-Drug

16   Conjugates Enabled by Trapped Ion Mobility Spectrometry and Top-Down Mass

17   Spectrometry. *Anal. Chem.*, **93**, 10013–10021.

18   Leney,A.C. and Heck,A.J.R. (2017) Native Mass Spectrometry: What is in the Name? *J. Am.*

19   *Soc. Mass Spectrom.*, **28**, 5–13.

20   Li,H. *et al.* (2018) An integrated native mass spectrometry and topdown proteomics method that

21   connects sequence to structure and function of macromolecularcomplexes. *Nat. Chem.*, **10**,

139–148.

Liu,R. *et al.* (2022) Native top-down mass spectrometry for higher-order structural

characterization of proteins and complexes. *Mass Spectrom. Rev.*, 1–51.

Liu,X. *et al.* (2010) Deconvolution and database search of complex tandem mass spectra of

intact proteins: A combinatorial approach. *Mol. Cell. Proteomics*, **9**, 2772–2782.

Liu,X. *et al.* (2012) Protein identification using top-down. *Mol. Cell. Proteomics*, **11**, 1–13.

Loo,J.A. (1997) Studying noncovalent protein complexes by electrospray ionization mass

spectrometry. *Mass Spectrom. Rev.*, **16**, 1–23.

Marty,M.T. *et al.* (2015) Bayesian Deconvolution of Mass and Ion Mobility Spectra: From

Binary Interactions to Polydisperse Ensembles. *Anal. Chem*, **87**, 4370–4376.

Mcilwain,S.J. *et al.* (2020) Enhancing Top-Down Proteomics Data Analysis by Combining

Deconvolution Results through a Machine Learning Strategy. *J. Am. Soc. Mass Spectrom*,

31.

Melby,J.A. *et al.* (2021) Novel Strategies to Address the Challenges in Top-Down Proteomics. *J.

Am. Soc. Mass Spectrom.*, **32**, 1278–1294.

Park,J. *et al.* (2017) Informed-Proteomics: Open-source software package for top-down

proteomics. *Nat. Methods*, **14**, 909–914.

Reid,D.J. *et al.* (2018) MetaUniDec: High-Throughput Deconvolution of Native Mass Spectra

MS Data Set MetaUniDec Deconvolution Integration & Extraction. *J. Am. Soc. Mass

Spectrom*, **30**, 118–127.

Sharon,M. and Robinson,C. V (2007) The role of mass spectrometry in structure elucidation of

1   dynamic protein complexes. *Annu. Rev. Biochem.*, **76**, 167–193.

2   Skinner,O.S. *et al.* (2018a) Multiplexed mass spectrometry of individual ions improves

3   measurement of proteoforms and their complexes. *Nat. Chem. Biol.*, **14**, 36–41.

4   Skinner,O.S. *et al.* (2018b) Top-down characterization of endogenous protein complexes with

5   native proteomics. *Nat. Chem. Biol.*, **14**, 36–41.

6   Stiving,A.Q. *et al.* (2019) Surface-Induced Dissociation: An Effective Method for

7   Characterization of Protein Quaternary Structure. *Anal. Chem.*, **91**, 190–209.

8   Sun,R.X. *et al.* (2016) pTop 1.0: A High-Accuracy and High-Efficiency Search Engine for Intact

9   Protein Identification. *Anal. Chem.*, **88**, 3082–3090.

10   Toby,T.K. *et al.* (2016) Progress in Top-Down Proteomics and the Analysis of Proteoforms.

11   *Annu. Rev. Anal. Chem.*, **9**, 499–519.

12   Wu,Z. *et al.* (2020) MASH Explorer: A Universal Software Environment for Top-Down

13   Proteomics. *J. Proteome Res.*, **19**, 3867–3876.

14   Yuan,Z.F. *et al.* (2012) pParse: A method for accurate determination of monoisotopic peaks in

15   high-resolution mass spectra. *Proteomics*, **12**, 226–235.

16   Zhou,M. *et al.* (2020) Higher-order structural characterisation of native proteins and complexes

17   by top-down mass spectrometry. *Chem. Sci.*, **11**, 12918–12936.
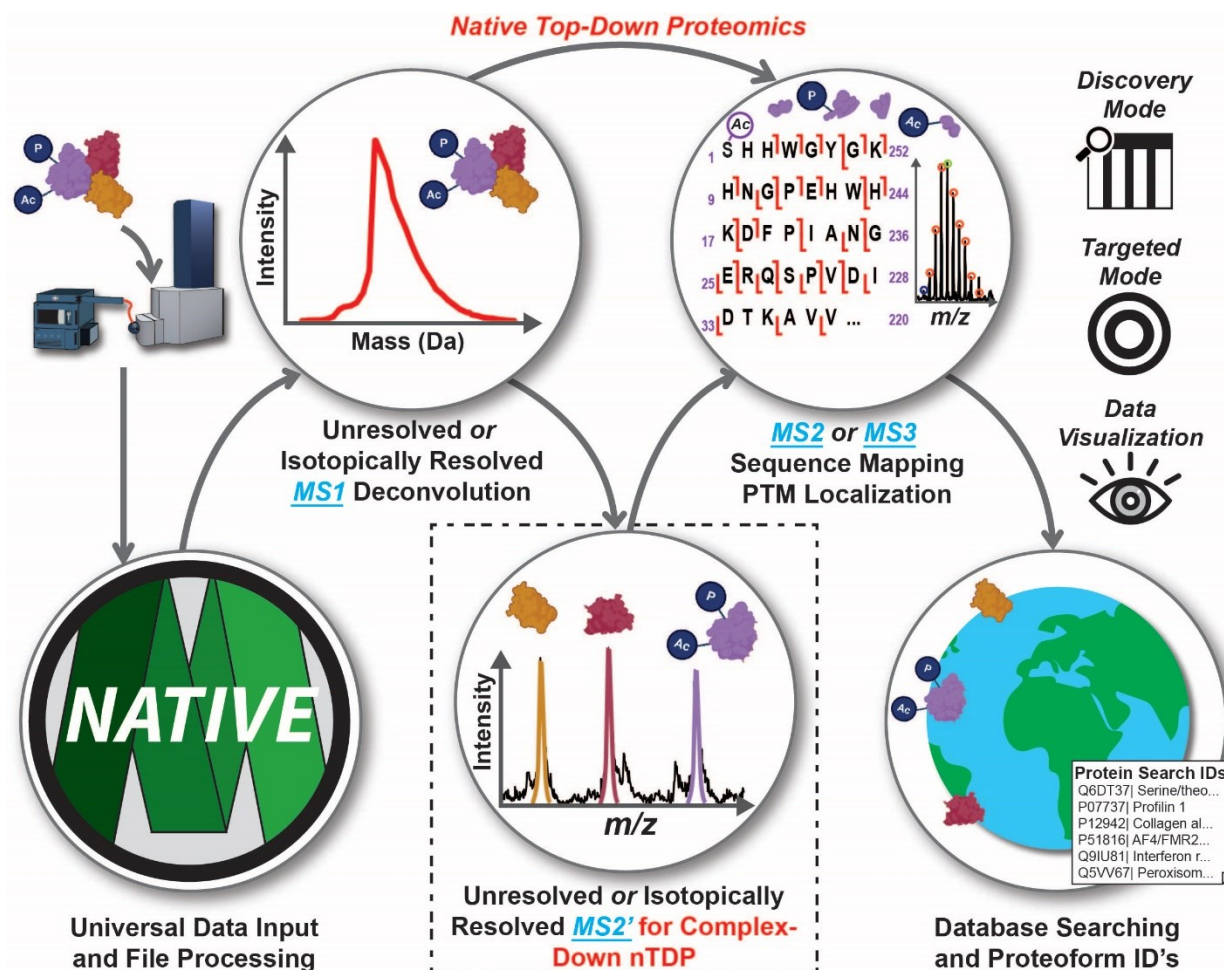
18

Figure 1. MASH Native provides a universal and comprehensive data processing software for a variety of nTDP analyses. MASH Native is capable of deconvoluting isotopically unresolved protein/protein complex (MS1) and released protein subunits (MS2') spectra, deconvoluting isotopically resolved MS1, MS2', MS2, and MS3 spectra, and performing database searches to identify unknown proteins. MASH Native can process nTDP data in both Discovery Mode and Targeted Mode approaches. It supports various MS file types and integrates multiple deconvolution/search algorithms into one package. MASH Native is a user-friendly software package capable of providing a "one-stop shop" for nTDP data processing.