Improved Peptide Search for Identification of SUMO and Sequence-Based Modifiers, in MaxSBM

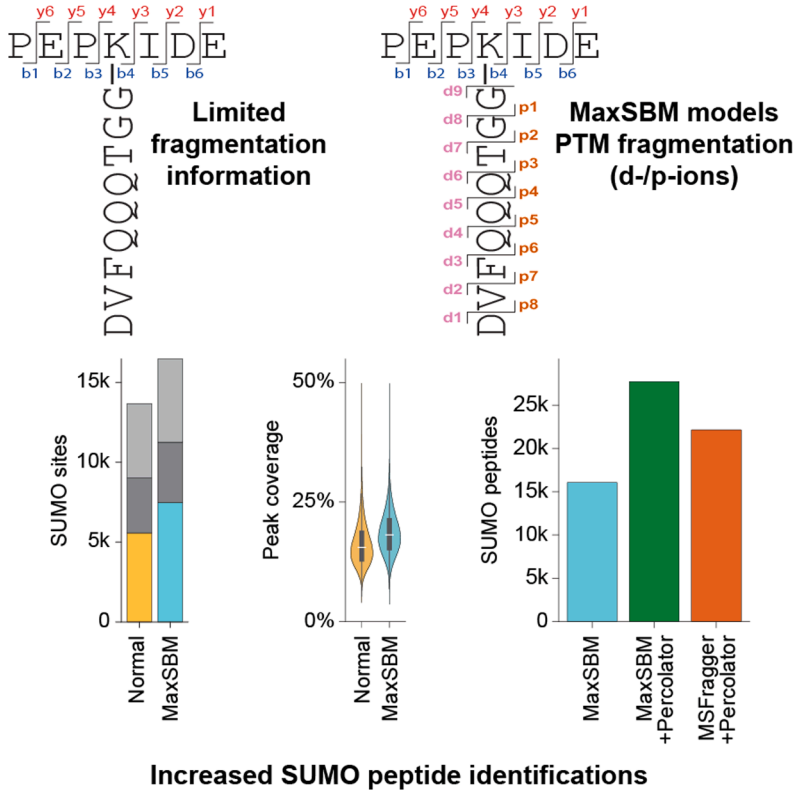
Authors

Caroline Lennartsson, Pelagia Kyriakidou, Michael Lund Nielsen, Jesper Velgaard Olsen, Jürgen Cox, and Ivo Alexander Hendriks

Correspondence

cox@biochem.mpg.de; ivo.hendriks@cpr.ku.dk

Graphical Abstract



In Brief

Here, we introduce MaxSBM, an optimized software module within the proteomics software MaxQuant for interpreting complex sequence-based modifiers, specializing in SUMOylation. Our approach incorporates PTM fragmentation into peptide scoring and annotation. By extending the theoretical spectral space to include fragments bearing partial modification peptide remnants, MaxSBM provides a more comprehensive spectral annotation that enhances peptide scoring, resulting in increased identification rates at a higher confidence. Beyond methodological refinement, we validated MaxSBM via reanalysis of published physiological SUMO datasets, ultimately unlocking new insights via mapping of previously obscured modification sites.

Highlights

- MaxQuant module for identification of sequence-based modifiers, including SUMO2/3.
- Andromeda search incorporates modifier-specific fragmentation patterns.
- Enhanced SUMO identification improves biological discovery from MS data.

2026, Mol Cell Proteomics *25*(*6*), 101589

https://doi.org/10.1016/j.mcpro.2026.101589

Improved Peptide Search for Identification of SUMO and Sequence-Based Modifiers, in MaxSBM

Caroline Lennartsson[1], Pelagia Kyriakidou[2], Michael Lund Nielsen[1,3], Jesper Velgaard Olsen[1,4], Jürgen Cox[2,*], and Ivo Alexander Hendriks[1,4,*]

Post-translational modifications (PTMs), such as Small Ubiquitin-like Modifier (SUMO)ylation and ubiquitination, regulate key cellular processes by covalently attaching to lysine residues. While mass spectrometry allows site-specific identification of PTMs, most existing search engines are optimized for small, non-fragmenting modifications and struggle to detect large, fragmenting protein-based modifiers. We refer to these as sequence-based modifiers (SBMs). To overcome this limitation, we developed an SBM-specific search strategy within MaxQuant that accounts for the fragmentation behavior of SBMs during peptide identification. Using publicly available datasets, we validated our approach for SUMO2/3. Our analysis identified distinct diagnostic features and characteristic mass shifts associated with SBM fragmentation, referred to in this study as d-ions (diagnostic ions) and p-ions (PTM ions). By leveraging these features, our method improved the identification of SUMOylated peptides from human cell lines by ~13%, SUMOylation sites in mouse embryonic cells by ~22%, and in mouse adipocytes by ~24%. Our search method improved spectral annotation of SBMs by up to 9% increase in the median Andromeda score. Taken together, we highlight the potential of our SBM search to enhance the discovery of protein-based modifications.

One of the major strengths of MS-based proteomics is the ability to identify post-translational modifications (PTMs). Identification of PTM sites is crucial for understanding biological mechanisms through direct identification of protein targets and regulatory dynamics. MS-based proteomics has become a powerful technique to study PTMs in a system-wide manner (1, 2). While proteomics has advanced in identifying PTMs, detecting and pinpointing unstable (labile) modifications remains difficult, as they can degrade or be lost during analysis. Such modifications are, for example, Small Ubiquitin-like Modifier (SUMO)ylation and ubiquitination. In this paper, we present a new methodology for improved site-specific identification of these modification types.

Sequence-based modifiers (SBMs) are very important in biology. SBMs refer to protein-based modifiers that covalently attach to other proteins to regulate their function, stability, and various other molecular mechanisms. The earliest discovered SBM is ubiquitin, best known for its role in targeting proteins for degradation (3). The family of ubiquitin-like modifiers has expanded over time, and includes SUMO1, SUMO2, SUMO3, NEDD8, ATG8, ATG12, URM1, UFM1, FAT10, and ISG15, each contributing to their unique functions (4). SUMOs are a prominent subgroup of these modifiers that play critical roles in a wide range of cellular processes, and are particularly known for modifying the majority of nuclear proteins (5, 6). SUMOs are covalently attached to lysine residues on target proteins *via* a conserved E1, E2, E3 enzymatic pathway (4, 7, 8). Due to their high sequence similarity, SUMO2 and SUMO3 are generally considered to have overlapping biological functions (9), while SUMO1 has different roles (10). Conjugation of SUMO2/3 contributes to many essential cellular mechanisms; nucleocytoplasmic exchange (11), DNA damage response (12–14), mRNA processing and metabolism (15), chromosome biology and chromatin organization (16–18), cell differentiation (19), gene transcription regulation (16, 17, 20–22), and far more. Consequently, the modification is also observed in many pathophysiological processes such as tumorigenesis (23), diabetes (24), heart failure (25, 26), neurological disorders (27), and liver diseases (28). Understanding SUMOylation by mapping of the subset of proteins that are SUMOylated (*i.e.,* the SUMOylome) is therefore crucial to elucidating key parts of cellular physiology.

The main characterization of SUMOylation sites in a systems-wide manner is done *via* mass spectrometry-based

From the [1]Novo Nordisk Foundation Center for Protein Research, Department of Cellular and Molecular Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; [2]Computational Systems Biochemistry, Max-Planck Institute of Biochemistry, Martinsried, Germany; [3]Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark; [4]Faculty of Health and Medical Sciences, Department of Cellular and Molecular Medicine, Copenhagen Center for Glycocalyx Research, University of Copenhagen, Copenhagen, Denmark

*For correspondence: Jürgen Cox, cox@biochem.mpg.de; Ivo Alexander Hendriks, ivo.hendriks@cpr.ku.dk.

https://doi.org/10.1016/j.mcpro.2026.101589

proteomics (5, 6). Endogenous SUMOylation identification is done with an extended MS workflow, involving antibody-based enrichment and enzyme serial digestion. This workflow generates a sequence remnant on the target peptides, which is difficult to identify due to its complex fragmentation pattern (5). MS computational workflows begin with centroiding and deisotoping of the raw spectral data. The cleaned spectra are searched against theoretical peptide spectra derived from in silico digestion of a protein database. Peptide-spectrum matches (PSMs) are made by comparing experimental spectra to the theoretical ones using search engines, such as Andromeda (29) (used in MaxQuant (30)) or MSFragger (31). The Andromeda search engine employs a probabilistic scoring algorithm to identify PSMs. PTM-containing peptides are identified, and the PTM itself is localized within the peptide by finding fragment ions bearing mass deltas that correspond to the mass of the modification. Andromeda evaluates a potential modification neutral loss from a modification by calculating the scores, both with and without the loss. The higher of the two scores is retained as the final score (29). To ensure statistical robustness, all scoring steps are mirrored identically against a reversed decoy database for false discovery rate (FDR) estimation (29).

There are many challenges related to the identification of SBMs, such as SUMOylation. They are difficult to study due to being biologically transient or low-abundant, and notably generate convoluted fragmentation patterns during MS analysis. Some progress has been made to improve their detection; for instance, MSFragger (31) can identify complex PTMs in two modes. Firstly, the tool PTM-Shepherd can search for diagnostic features of fragmenting PTMs (32). Secondly, MSFragger-Labile can increase labile PTM identification through consideration of PTM remainder masses on the peptide ions (33). These advancements have been shown to enhance identification of glycosylation and ADP-ribosylation (32–34). The software pLink (35) and OpenUaa (36) were developed for analysis of cross-linked peptides, and have previously been used for identification of SUMOylation because SUMO-modified peptides consist of two covalently linked sequences, a scenario similar to cross-linked peptides in MS/MS analysis. However, the currently available versions of this software no longer support SUMO-specific searches. Finally, the SUMmOn software was developed to identify SUMOylation (37). SUMmOn utilizes an initial database search with SEQUEST (38) and then looks for SUMO peak patterns. However, the software has only seen limited use in practice (39) and depends on software compilers that are not compatible with contemporary computer hardware. Taken together, the efforts made towards identification of SBMs and labile modifications highlight the overall interest and necessity for such approaches, and a need for optimized and up-to-date software still exists.

The major challenge to overcome is the analysis of convoluted MS/MS spectra resulting from the fragmentation of peptides bearing SBMs. Because the SBMs themselves fragment, the result is a much more complex fragmentation pattern that involves three peptide termini, rather than two. The SBM ion series integrates with the y and b ion series of the target peptide, hindering identification of the target peptide sequence. Here, we anticipate, evaluate, and incorporate series of fragment ions and remainder masses into the theoretical search space for PSM into a module in the MaxQuant search engine (30, 40). Since the strategy is to expand the theoretical search space, we must account for an increased probability of getting decoy matches. This problem is widely described in the database search field (41–43), especially in proteogenomics (44).

To address these challenges, we designed a dedicated MaxQuant module specifically for SBM identification. This module operates by incorporating PTM fragmentation into the theoretical search space during Andromeda scoring. This module annotated and enhanced the identification of SBMs during a dedicated data-dependent acquisition (DDA) database search. To optimize and evaluate our SBM module, we utilized several representative published data sets based on endogenous SUMOylation (5, 19, 45). With the MaxSBM module, we improved the identification of SUMOylated peptides in HEK cell lines by 13%. Additionally, we observed an increase in detected SUMOylation sites in mouse samples: 22% in embryonic cells and 24% in adipocytes. The method improved the spectra annotation of SBMs by higher PSM scoring and better overall intensity and peak coverage. We also compared our software with the existing tool for complex PTMs, MSFragger-Labile, as well as with a comparable pipeline using Percolator for significance estimation. Our results show that by using Percolator as post processor MaxSBM, we achieve a higher identification rate. However, these findings should be interpreted with caution, as the use of Percolator resulted in an increased number of “one-hit wonders.” Future studies should therefore focus on improving FDR estimation in this context, particularly through advanced or machine learning–based approaches. These methods could also be improved by expanding the input feature set that can better account for the increased complexity of these modifiers. Taken together, these results highlight the potential of our approach for SBM identification and localization in proteomics research, enabling deeper biological insights and improving our understanding of how these PTMs function in health and disease.

EXPERIMENTAL PROCEDURE

Experimental Design

Datasets—We used four published proteomics datasets to develop and evaluate the SBM module. For clarity, each dataset is assigned a shorthand name that will be used consistently throughout the manuscript. We analyzed three SUMOylation and one ubiquitination datasets: SUMO-HEK, SUMO-Adip, SUMO-MEC, and Ub-LysC.

The SUMO-HEK dataset (PXD008003) published by Hendriks *et al*. (5) was derived from HEK cells exposed to proteotoxic stress (heat shock). It was generated using a site-specific proteomics strategy that combines a dual-protease digestion method with peptide-level immunoprecipitation to map endogenous SUMO2/3 modification sites. Specifically, denatured lysates are first digested with Lys-C, peptides enriched *via* the SUMO2/3 targeting antibody, and then further digested with Asp-N before LC–MS/MS analysis (Fig. 1*A*).

The SUMO-Adip dataset (PXD024144), described by Zhao *et al*. (45), was generated to profile SUMOylation dynamics during mouse adipocyte differentiation. SUMO-modified peptides were enriched with a SUMO2/3-specific antibody and analyzed by LC–MS/MS to map stage-specific SUMO targets across differentiation.

The SUMO-MEC dataset, described by Theurillat *et al*. (19) (PXD017697), compares SUMOylomes between mouse embryonic stem cells and fibroblasts, highlighting developmental specificity.

The Ub-LysC dataset, reported by Akimov *et al*. (46) (PXD006201), was generated using a unique ubiquitination detection method based on Lys-C digestion, which leaves a longer covalently attached remnant with the sequence ESTLHLVLRLRGG (1431.83 Da).

Search Settings

MaxQuant—SUMO Datasets (HEK, Adip, MEC). All SUMO datasets were searched in MaxQuant (version 2.6.8.0) with default parameters unless specified. To accommodate SUMO remnants, the maximum peptide mass was set to 6000 Da, the second-peptide search was disabled, the maximum number of variable modifications (VMs) per peptide was set to three, SBM was enabled. Enzyme specificity used AspN, LysC/P, and GluN. The number of missed cleavages was set to 8. The fixed modification was set to Carbamidomethyl (C) and the VMs to Oxidation (M) and Acetyl (Protein N-term). SUMO was included as an additional variable modification using our variants: VM (variable modification) search (0 p-ions), standard SUMO (1 p-ion), SBM 3 (3 p-ions), and SBM 8 (all 8 p-ions), depending on each included modification setting, with d and p-ions selected from Table 1. The FDR filtering in MaxQuant was preset to default; that is the PSM FDR was set to 1% with an additional filtering step of 1% FDR at the protein level. Additionally, the search engine was set to only consider PSMs with a minimum score of 40 and delta score of 6 for modified peptides. The first search precursor tolerance was set to 20 ppm. The main search precursor tolerance was set to 4.5 ppm. The MS/MS (fragment) match tolerance was set to 20 ppm.

Optimization Runs (SUMO). Before the main analyses, we ran small optimization searches on a subset of SUMO-HEK raw files to tune the SBM search space efficiently. We tested different counts and selections of p-ions and evaluated alternative sets of diagnostic d-ions. These runs were used only to choose parameters for the main analyses. The optimization runs and the subsequent main HEK searches used the same UniProt (47) human proteome FASTA (downloaded December 2020, 97,795 entries including isoforms).

SUMO-HEK. HEK searches used the UniProt (47) human proteome (downloaded December 2020, 97,795 entries including isoforms). We performed four searches: VM, Standard SUMO, SBM 3, and SBM 8. We also performed a dedicated MaxQuant run to get the input for Percolator that was identical to SBM 3 except for the internal filtering parameters, which were set as peptideFdr, proteinFdr, and siteFdr set to 1 (100%), minScoreModifiedPeptides, and minDeltaScoreModifiedPeptides set to 0 to produce an unfiltered msms.txt for downstream rescoring in Percolator.

Small Ubiquitin-like Modifier Mouse Embryonic Cell. These searches used the UniProt (47) mouse proteome (downloaded March 2022, 63,654 entries including isoforms). For each of these datasets, we performed two searches: a Standard SUMO search and an SBM3 search.

Ubiquitin Group (Ub-LysC). Ubiquitin searches were performed in MaxQuant version 2.6.8.0. The maximum peptide mass was set to 6000 Da, and enzyme specificity to LysC/P with a maximum of 4 missed cleavages. The fixed modification was set to Carbamidomethyl (C) and the variable modifications to Oxidation (M) and Acetyl (Protein N-term). Ubiquitination was chosen as a variable modification using the remnant ESTLHLVLRLRGG (1431.83 Da), with d and p-ions selected from Supplemental Table S1. Ubiquitin searches used the UniProt (47) human proteome (downloaded December 2020, 97,795 entries including isoforms). The first search precursor tolerance was set to 20 ppm. The main search precursor tolerance was set to 4.5 ppm. The MS/MS (fragment) match tolerance was set to 20 ppm.

Percolator. To compare MaxQuant and FragPipe and MSFragger fairly, we rescored the MaxQuant PSMs with Percolator (48) (version 3.06) using the Crux toolkit (version 4.2) (49, 50). MaxQuant assigns each PSM a posterior error probability-based score (PEP score) to estimate the chance that a given PSM is incorrect (40). FragPipe uses Percolator, which trains a support vector machine on features derived from searches against target and decoy sequences (48). By running MaxQuant results through Percolator as well, both pipelines use the same PSM-level rescoring step for FDR estimation. This makes comparison possible because any differences will mainly reflect the search settings and engines, and not different PSM level rescoring methods. We note that the feature sets supplied to Percolator differ across setups. The Percolator standalone run based on the MaxQuant msms.txt uses a feature set that is not identical to the one passed to Percolator inside FragPipe. The FDR was set to 1% for all PSMs across all searches with Percolator.

As described in the MaxQuant SUMO-HEK section, we generated an SBM 3 export run to produce an unfiltered msms.txt for Percolator. In that export run, the MaxQuant FDR and some score thresholds were switched off to remove any filtering before rescoring.

Percolator is typically used with inputs that provide separate evidence for the target and the decoy for each spectrum, a best target score and a best decoy score. MaxQuant, in contrast, reports only the highest score overall after searching against both target and decoy sequences, so msms.txt contains one score per spectrum that can correspond to either a target or a decoy. Following community practice reported by others (51), we used msms.txt as Percolator input despite this difference. To do so, we reformatted the table to Percolator's expected column layout and names and assigned the appropriate target or decoy label to each PSM without constructing explicit target–decoy score pairs. The feature mapping used in this conversion is provided in Supplemental Table S2, and the conversion script is available at https://github.com/carolinelennartsson/SUMOylation. Percolator was executed with default settings except that the –only-psms flag was set to True to report on PSM level. Percolator uses three-fold cross-validation for support vector machine training (52). The dataset contained 606,759 targets and 256,486 decoys, with a size ratio of ~2.37.

FragPipe and MSFragger-Labile. We also processed the SUMO-HEK dataset with the MSFragger-Labile search (MSFragger version 4.4.1) (33) in the FragPipe MS analysis pipeline (version 24.0) to provide a comparison with MaxQuant. The same FASTA file was used: human proteome (downloaded December 2020 from UniProt (47), 97,795 entries including isoforms), decoy and contaminant hits were added in FragPipe. Statistical estimation of the MSFragger-Labile search output was calculated with Percolator (version 3.06.5) (48). The decoys and contaminants were added to the FASTA sequence database with Philosopher (53). Enzymes were set to LysC/P and a custom rule that cleaves N-terminal to D and E (to mirror the AspN/GluN behavior used in MaxQuant). The maximum number of missed cleavages was set to 2 for LysC/P and 6 for AspN/GluN,

which sums to 8 missed cleavages used in MaxQuant, and reasonably reflects the expected missed cleavages for each enzyme. This search was compared to setting 8 missed cleavages for both enzymes, which would allow MSFragger to generate at least as many sequences as MaxQuant. The fixed modification was set to Carbamidomethyl (C) and the variable modifications to Oxidation (M) and Acetyl (Protein N-term). The SUMO search was performed as in the MaxSBM searches, and SUMO was defined as a hybrid variable modification based on the MSFragger-Labile workflow. SUMO was set as a variable modification with mass 960.43 Da, on lysine residues (K). The diagnostic ions that were included in the search were d2 to d8 (Table 1). The peptide remainder fragment ions and the fragment remainder masses included were selected from p1 to p8 (Table 1), with either 0, 3, or 8 number of p-ions included. When three masses

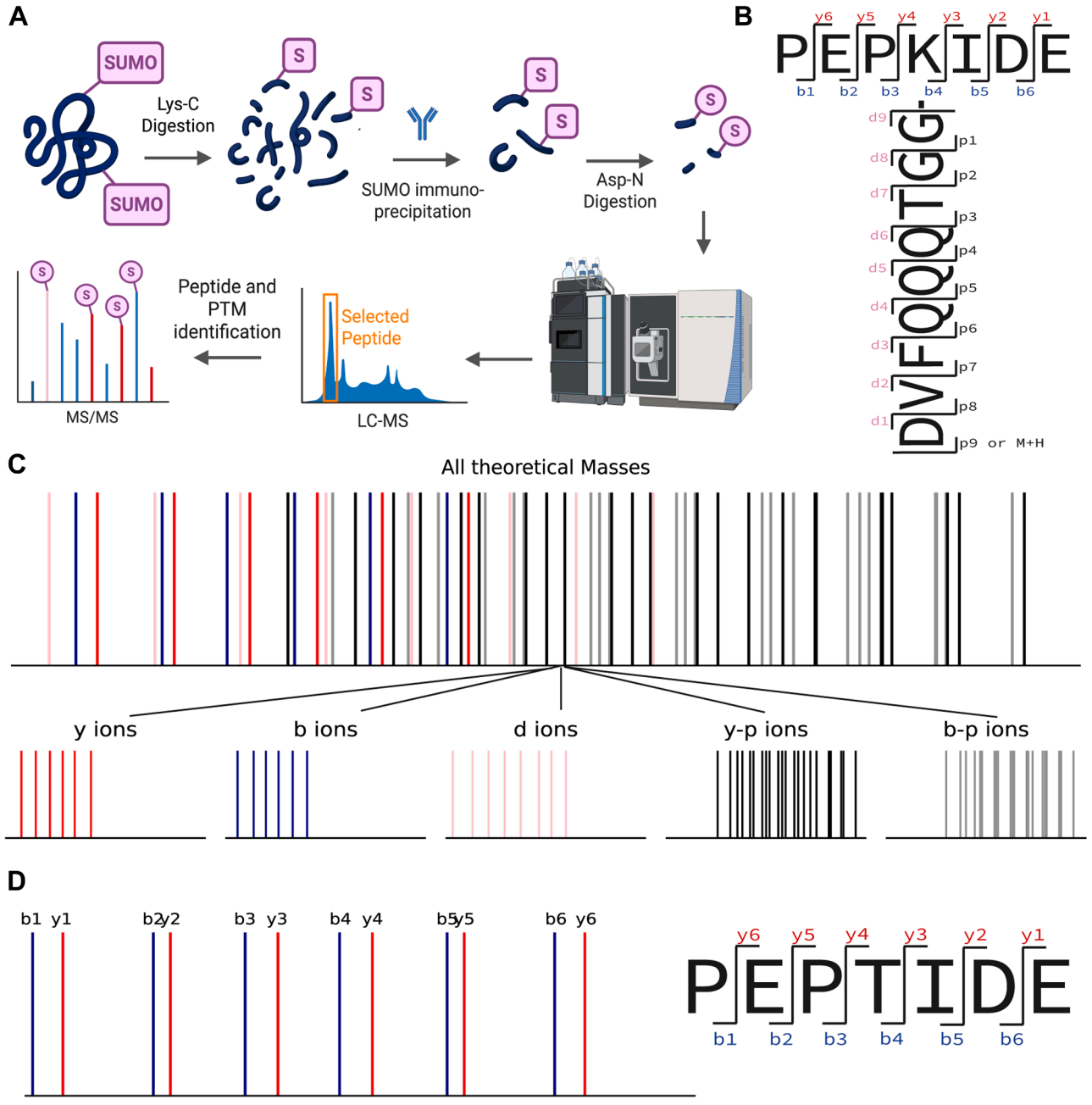


FIG. 1. **Workflow and fragmentation characteristics of SUMOylated peptides.** *A,* method workflow for MS-based identification of endogenously SUMOylated peptides (5). The samples are digested with AspN and LysC enzymes. The SUMOylated peptides are enriched with antibody-based enrichment. *B,* SUMOylated target peptide displaying the y-, b-, d-, and p-ion series. The y- and b-ions can potentially bear a SBM fragment (p-ion) and are annotated with both their corresponding y-/b- and p-ion labels. The diagnostic (d-ion) series, representing SBM fragments that are no longer attached to the target peptide, is shown in *pink*. *C,* theoretical spectrum showing all possible fragments for the SUMOylated peptide depicted in *panel B*. The theoretical spectra consist of y-, b-, d-, y-p, and b-p -ion series. *D,* theoretical spectrum of a standard (unmodified) peptide, shown for comparison to highlight the increased fragmentation complexity introduced by SUMOylation.

were included in the search, p2, p3, and p7 were selected. Both the peptide remainder masses and the fragment remainder masses were set to p2 (114.04 Da), p3 (215.09 Da), and p7 (746.33 Da). PSMs were validated with Percolator and filtered at 1% FDR for all searches, as per default. Proteins were also filtered at 1%. Both the precursor and fragment mass tolerance were set to 20 ppm. The MSFragger output file '*psm.tsv*' was used for the analysis.

Post Identification Processing and Data Analysis—All post identification processing was done in Python using pandas (54) and numpy (55). The results were visualized with matplotlib (56) and seaborn (57). The PSM data was extracted from the *'evidence.txt'*, *'msms.txt'*, and *'sites.txt'* MaxQuant output files. For analysis of the MaxSBM output, the modification table *'sites.txt'* is recommended. Reverse and contaminant hits were removed from the result when they are not explicitly visualized. Additionally, PSMs containing two or more SBMs were filtered out. To benchmark the number of detected gene names, the peptide to protein inference was done by finding all matching proteins for each peptide sequence from the protein database. A representative protein was chosen for each PSM through canonization. This was done by selecting the protein with the best annotation from the UniProt reference proteome (47). The supporting data for the FASTA files were downloaded from the UniProt web server. A protein was selected from the protein group by its annotation in the UniProt database in the order: review status (reviewed or unreviewed), annotation (0–5), number of Gene Ontology (GO) terms, alphabetical order of entry name, and finally entry name length (older entries have shorter names). During protein inference, the SUMO modification was only permitted at the C-terminus of a peptide if it was followed by an E or D in the protein sequence. The Python package Biotite was used to read and parse the FASTA files (58). Structural analysis of the SUMO sites was done with the AlphaFold predicted database of protein structures (59, 60). The AlphaFold database did not contain structures of the isoforms, which were therefore omitted in the analysis. Pairwise comparisons between the VMsearch and the other searches were performed using unpaired two-sample t-tests with unequal variance assumptions, implemented *via* the ttest_ind function from the scipy.stats Python library (61). This analysis was applied independently to intensity coverage, peak coverage, and the number of matched peaks. *p*-values from these tests were annotated on the corresponding violin plots in scientific notation alongside their significance levels. Significance thresholds were defined as follows: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***), and $p < 0.0001$ (****); values equal to or above 0.05 were considered not significant (ns). All scripts for visualization are available in the GitHub repository: https://github.com/carolinelennartsson/SUMOylation.

Statistical Rationale

No new experiments were conducted in this study. All analyses were based on previously published datasets and leveraged established statistical frameworks as described in earlier studies.

RESULT

Implementation of MaxSBM

We developed MaxSBM, a module integrated into MaxQuant, with the aim of enhancing the identification, annotation, and validation of SUMOylation sites. To this end, we leveraged previously published endogenous SUMOylation MS datasets from human HEK cells (SUMO-HEK) (5), SUMO-MEC (19), and mouse adipocytes (SUMO-Adip) (45). The protocol for identification of endogenous SUMOylation sites includes; serial enzymatic digestion, antibody-based enrichment, and MS-based peptide identification. The method is described in detail by Hendriks *et al*. (2018) (5) and summarized in Figure 1*A*. Digestion using both Lys-C and Asp-N of target proteins and the covalently attached SUMO2/3 protein, leaves a subset of peptides bearing the SUMO2/3-derived sequence remnant (DVFQQQTGG) on lysine residues (Fig. 1*B*). The theoretical combinations of single and double fragmentations of SUMOylated peptides are highly convoluted (Fig. 1*C*), and include the standard b- and y-ion series, as well as two additional ion series specific to SBM modification: the diagnostic fragment ion series (d-ions) and the modification loss series (referred to here as p-ions) (Table 1 and Fig. 1*B*). Compared to the theoretical spectra of an unmodified peptide (Fig. 1*D*), the y-p and b-p -ion peaks substantially increase the number of theoretical peaks (Fig. 1*C*). To consider this, we developed a new module MaxSBM, that builds on top of Andromeda by incorporating the SBM ion series into the search space (Fig. 2*A*).

MaxQuant Graphical User Interface Configuration

To enable the detection of SUMOylated peptides using MaxQuant, we configured the software to support sequence-based modifiers (SBMs) *via* the "MaxSBM" module. These

TABLE 1
Chemical compositions and monoisotopic neutral masses of Small Ubiquitin-Like Modifier; d-ions and corresponding p-ions, including doubly fragmented sequences FQ, FQQ, and QQ

Sequence	d-ion	d-ion Composition	d-ion neutral Mass (Da)	Corresponding p-ion
D	d1	H(5) C(4) N(1) O(3)	115.02	p8
DV	d2	H(14) C(9) N(2) O(4)	214.09	p7
DVF	d3	H(23) C(18) N(3) O(5)	361.16	p6
DVFQ	d4	H(31) C(23) N(5) O(7)	489.22	p5
DVFQQ	d5	H(39) C(28) N(7) O(9)	617.28	p4
DVFQQQ	d6	H(47) C(33) N(9) O(11)	745.33	p3
DVFQQQT	d7	H(54) C(37) N(10) O(13)	846.38	p2
DVFQQQTG	d8	H(57) C(39) N(11) O(14)	903.40	p1
DVFQQQTGG	d9	H(60) C(41) N(12) O(15)	960.43	-
FQQ	FQQ	H(25) C(19) N(5) O(5)	403.18	-
FQ	FQ	H(17) C(14) N(3) O(3)	275.12	-
QQ	QQ	H(16) C(10) N(4) O(4)	256.12	-

settings were implemented through the graphical user interface in the modification list, under the '*configuration'* tab (Fig. 2*B*). The MaxSBM module is activated by incorporating a VM in a DDA search with the modification type *'sequence based modifier'* (Fig. 2*B*). For SBMs, masses are calculated automatically when prompted a sequence input in the '*sequence'* box. After the search is finished, the identified p-ions are shown in the MaxQuant spectral viewer for the identified SUMOylated spectra (Supplemental Fig. S1, *A* and *B*). The identified p-ions are reported with annotation and mass in the *'msms.txt'* search output table. The MaxSBM output with the complete set of modification identifications is found in the modification output table *'sites.txt'*. We recommend analysis with the downstream proteomics analysis platform Perseus (62).

Optimization of the Search Space for Efficient Identification of SUMOylation

To ensure optimal performance of the new SBM module, we systematically optimized the theoretical search space by evaluating the impact of p-ion inclusions. All optimizations were done on the endogenous SUMOylation dataset (SUMO-HEK) (5). Although adding p-ions to the theoretical search space increases the possibility of identifying target peptides, the increased number of theoretical peaks also raises the chance of a decoy match (29). Therefore, careful tuning was necessary to strike a balance between sensitivity and specificity. We optimized on the HEK cell line dataset, as it's the most extensive identification of human SUMO sites (5). To fine-tune, we focused on two key aspects: identifying which p-ions yield the highest number of SBM identifications, and determining the optimal number of p-ions to include. Firstly, we found that while there was only a slight preference in the benefit of each, p2, p3, and p7 were optimal to include for getting the highest identification rates (Supplemental Fig. S2*A*). Secondly, we performed a series of searches wherein we searched a combination of 2, 3, 4, 5, 6, and 8 p-ions, preferentially selecting those that individually gave the highest identification rates. We found that with the HEK dataset, the optimum number of SUMO p-ions to include was three (Supplemental Fig. S2*B*). In addition to optimizing the p-ions, we also tested the effect of the diagnostic (d-ion) series on the identification rate. To illustrate the prevalence of the d-ions across the dataset, we summed the d-ion peaks from the SBM 3 search. We showed that the d-ion series are commonly observed across the data (Supplemental Fig. S3*A*). For optimization of the d-ions, we compared using none of the d-ions with using two sets of d-ions: d2-d9 and an optimized set reported by Hendriks *et al*. (5). This optimized set included d2-d9, along with masses corresponding to double-fragmented d-ions with the amino acid sequences 'FQQ', 'QQ', and 'FQ' (Table 1). Regardless of their prevalence, we showed that the number of d-ions did not influence the identification rate (Supplemental Fig. S3*B*). Nonetheless, including more d-ions increased spectral peak coverage (Supplemental Fig. S3*C*), the number of matched peaks (Supplemental Fig. S3*D*), as well as the spectral intensity coverage (Supplemental Fig. S3*E*). Thus, while d-ions can increase the quality of spectral annotation and served as an internal validation for the presence of SUMOylation, they otherwise do not positively or negatively affect identification rates.

Enhanced Identification Rates Using MaxSBM

We leveraged a previously published endogenous SUMOylation MS dataset from human HEK cells (SUMO-HEK) (5) to enhance identification rates using MaxSBM. With the SUMO-HEK datasets, we showed that including the SBM peaks in the search improves the annotation of SUMOylated peptides (Fig. 3). We evaluated the performance of the SBM module under different configurations of p-ion inclusion. We tested different p-ion configurations, comparing a VM search without p-ions, a standard MaxQuant search with one neutral loss, the optimal three-p-ion setup (SBM 3), and all eight p-ions (SBM 8). We observed an increase in total identified PSMs by ~7% (Fig. 4*A*). SBM 3 increased unique SUMOylated peptide IDs by 13% compared to the standard search and 20% compared to the VM search (Fig. 4*B*). Moreover, the data completeness increased from 44.1 to 45.3% when using SBM 3, indicative of more detection overall in 3/3 replicates. Intensity coverage (Fig. 4*C*), peak coverage (Fig. 4*D*), and the number of matched peaks, were all significantly increased with the addition of p-ions to the theoretical search space (Fig. 4*E*). When considering spectral annotation quality, we observed a notably increase for the SBM 3 search with 9% of the median Andromeda score (Fig. 4*F*). In general, increasing numbers of p-ions considered in the search resulted in larger numbers of matched peaks and correspondingly higher Andromeda scores of the SUMOylated peptides (Supplemental Fig. S4). Importantly, scores for reversed database hits also increased with larger numbers of p-ions in the search space (Supplemental Figs. S4*D* and S5*D*), stressing the need to carefully balance the number of included p-ions. Median PSM score remained identical for all non-SUMO target peptides (Supplemental Fig. S5*A*). The median score decreased by 13.6 (20.1%) for the non-SUMO reverse hits (Supplemental Fig. S5*B*). The median score increased by 6 (4.2%) for SUMO target hits (Supplemental Fig. S5*C*) and by ~1% for the SUMO reverse hits (Supplemental Fig. S5*D*). Overall, the majority of all SUMOylated peptides were identified by all searches, with most unique identifications resulting from the SBM 3 and SBM 8 searches (Fig. 4*G*). The number of protein groups identified was 2884 for SBM 3, with 339 unique protein groups for this search, mainly identified from the nucleus (Supplemental Fig. S6). The number of protein groups identified in SBM 8 was 2720, 2683 in Standard, and 2597 in the VM search (Supplemental Fig. S6).

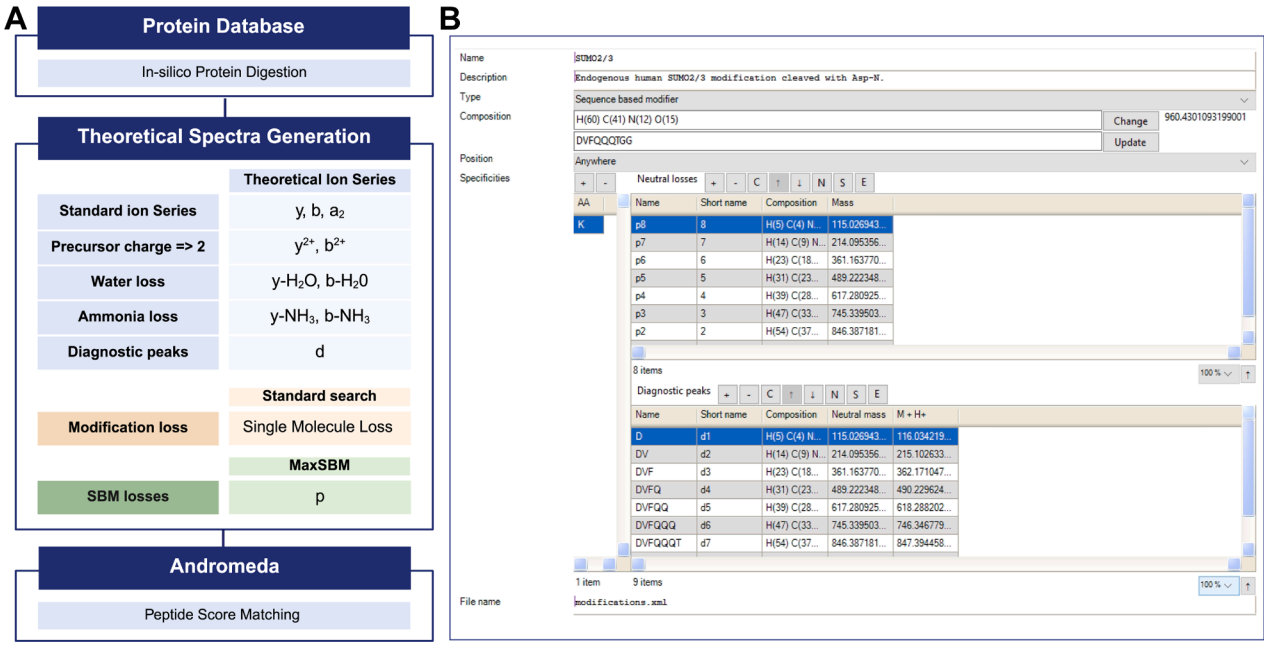


FIG. 2. **Implementation of sequence-based modifier searches in MaxQuant.** *A,* theoretical Peak List Generation for Peptide Score Matching in Andromeda (29). During peptide score matching, a protein database is in-silico digested into peptides. A theoretical peak list is generated for each peptide for the possible ion series. In the standard module, only a single modification loss is considered, whereas in the SBM module, all specified losses are utilized for identification. *B,* MaxQuant graphical user interface for defining SBM modifications, which allows users to define SBM modifications by inputting an amino acid sequence. The program then automates the process by calculating theoretical losses for SBM, where p-ions are displayed under "Neutral Losses" and the d-ions under "Diagnostic Peaks." SBM, sequence-based modifier.

MaxSBM-Exclusive SUMO Sites Conform to Known Biological Properties

To validate SUMO sites identified in the SUMO-HEK dataset exclusively through the SBM module, we investigated two known biological properties of SUMOylation sites: the tendency for SUMO to reside in disordered protein regions (6) and adherence to the canonical K x E modification motif (6, 63). To investigate SUMO-proximal structural properties, we extracted information from the AlphaFold structure database (60) and found that the predicted local distance difference test (predicted local distance difference test) is lower near SUMOylated lysine residues compared to unmodified lysine residues (Fig. 4*H*). As a low predicted local distance difference test correlates with a higher degree of disorder, this validates the tendency for SUMO to occur in disordered regions. There were no notable differences in SUMO-proximal structural context between the different searches, suggesting that the SBM searches identify SUMO sites in similarly disordered regions. Adherence to the K x E consensus motif was 26.9% for SBM 3, 27.8% for standard, and 28.5% for the VM search (Fig. 4*I*). These numbers were considerably higher than the randomly expected occurrence of 11%, and overall agree with previous observations of K x E adherence for similar SUMOylome depth (6, 63). For SUMO sites identified by both SBM 3 and standard search, K x E adherence was 27.9%, while peptides uniquely identified *via* SBM 3 or standard search were both ~22.5%, indicating that neither search has a bias towards non-KxE motifs (Fig. 4*J*). Taken together, our observations support the notion that SUMO sites exclusively identified using MaxSBM adhere to known SUMOylation sequence preferences, and thus are not more likely to be false positives. Moreover, when considering SUMO target proteins, we found that these overall adhered to the subcellular localization expected for SUMO, e.g., predominantly nuclear and chromatin. SBM-exclusive SUMO target proteins showed a similar distribution and a preference to modify nuclear proteins, further substantiating the validity of the SBM search (Supplemental Fig. S6).

Comparison with Existing Tools for the Identification of SBMs and Complex PTMs

We established that MaxSBM improved performance for the detection of SUMO within the MaxQuant search engine. To benchmark the performance of MaxQuant against other software that can similarly handle SBMs and fragmentation in the PTM itself, we utilized MSFragger-Labile (33) and compared performance on the SUMO-HEK dataset. The MaxQuant searches allowed up to 8 max missed cleavages across all enzymes. To compare with the enzyme settings in

MSFragger, the enzymes settings were tested with LysC/P set to 2 and AspN/GluN set to 6 missed cleavages (*i.e.,* 8 total and reflecting expected cleavage distribution), as well as with both enzymes set to 8 missed cleavages (*i.e.,* inclusive of at least all sequences also generated in MaxQuant). Both of the approaches performed comparably, with the search with 8 missed cleavages for both enzymes yielding slightly more (~2.4%) identifications (Supplemental Fig. S7, *A–C*).

Notably, MSFragger output is post-processed using Percolator, and MSFragger-Labile with Percolator considerably outperformed MaxQuant in the absence of Percolator. We therefore applied Percolator to post-process MaxQuant output generated using the SBM 3 search (Fig. 5). Strikingly, this increased the number of unique SUMOylated peptides from 16,053 to 27,728, representing a 72.7% increase (Fig. 5*A*), highlighting the substantial impact of semi-supervised rescoring on peptide identification. MSFragger identified 22,136 unique SUMOylated peptides, whereas MaxQuant with Percolator yielded 25.2% more identifications than the MSFragger-Labile pipeline. When applying a more stringent filter requiring at least two supporting PSMs, identification numbers were 11,862 for SBM 3, 17,405 for MaxQuant with Percolator, and 16,681 for MSFragger (Fig. 5*B*), indicating that the relative trends are maintained under stricter confidence criteria.

When considering all unique SUMOylated peptides, the majority of peptides (13,768) were detected by all three approaches, followed by subsets identified by one or both Percolator-enhanced workflows (Fig. 5*C*). Notably, 7607 peptides were identified exclusively by MaxQuant with Percolator, whereas 4969 were unique to MSFragger, and 3386 peptides were shared between the two Percolator-based approaches (Fig. 5*C*). The number of peptides supported by identified single PSM was highest for MaxQuant with Percolator and lowest in the absence of Percolator (Fig. 5*D*), consistent with increased sensitivity at lower evidence thresholds. Across all approaches, the mode and median number of PSMs per peptide was one (Fig. 5*E*). Analysis of Percolator score distributions showed the decoy population centered around −1, while the target PSMs displayed a bimodal distribution with maxima around −1 and 1 (Supplemental Fig. S8*B*), consistent with separation of low- and high-confidence identifications.

Together, these results demonstrate that Percolator-based rescoring substantially enhances identification sensitivity, while differences between search engines persist, resulting in partially overlapping but complementary sets of SUMOylated peptides.

Deriving Biological Insight

To further investigate the applicability of the MaxSBM module, we benchmarked the optimized search on two additional published endogenous SUMOylome datasets, with one based on mouse embryonic stem cells and fibroblasts (SUMO-MEC) (19), and the other based on differentiating mouse adipocytes (SUMO-Adip) (45). In the SUMO-MEC dataset, we found an increase of 21.7% unique SUMOylation sites when using the module, from 1308 (Standard) to 1592 (SBM 3) (Fig. 6*A*). Importantly, the MaxSBM search profiled 168 SUMO target protein groups that were not found using a standard search, and reassuringly the SBM-exclusive proteins displayed a prominent nuclear localization (Fig. 6*E*). We investigated this subset of exclusive SUMO target proteins using the STRING database (64), and found them to be highly interconnected and displaying enriched biological terms that are expected from SUMO target proteins (Supplemental Fig. S9), such as involvement in transcription and occurrence on phosphoproteins (Fig. 6*F*). In the SUMO-Adip dataset, the module resulted in a striking 24.1% gain of SUMOylation sites (Fig. 6*C*), from 3070 (Standard) to 3809 (SBM 3). In addition, we found 300 SUMO target proteins not found using the standard search (Fig. 6*G*). Similarly, these exclusive proteins were primarily localized to the nucleus (Fig. 6*G*). Proteins specifically identified using the module was found to be highly interconnected using the STRING database (Supplemental Fig. S10). Again, we found that these proteins were involved in DNA binding processes and transcriptional regulation (Fig. 6*H*). Altogether, by processing two physiological datasets using MaxSBM, we see substantial (21–24%) increases in the number of identified SUMO sites, along with overall higher spectral scores, and unveil hundreds of SUMO target proteins that align well with expected biological trends.

DISCUSSION

In this work, we demonstrate that inclusion of p-ions generated from SBMs (such as SUMO2/3) increases data completeness, peak coverage, and the intensity coverage, overall leading to a substantially better annotation and interpretation of existing data sets. To exemplify this, we used MaxSBM to re-process endogenous SUMO datasets, and can enhance peptide identification by 13% in a deep human cell line dataset (5), by expanding the theoretical peak list used for peptide score matching to include ions generated by fragmentation of the SUMO peptide remnant itself. We additionally show that we could increase the number of SUMOylation sites identified from mouse embryonic cells (19) by 21.7%, and from mouse adipocytes (45) by 24.1%. Importantly, the new sites derived *via* MaxSBM adhere to expected SUMO biological properties, including consensus motif adherence and structural preference. SUMO target proteins exclusively identified *via* MaxSBM reassuringly displayed predominant nuclear localization, as well as involvement in SUMO-related processes such as transcriptional regulation. Furthermore, the increased identification rate improves data completeness, that is detection of SUMOylation in more experimental replicates, which strengthens

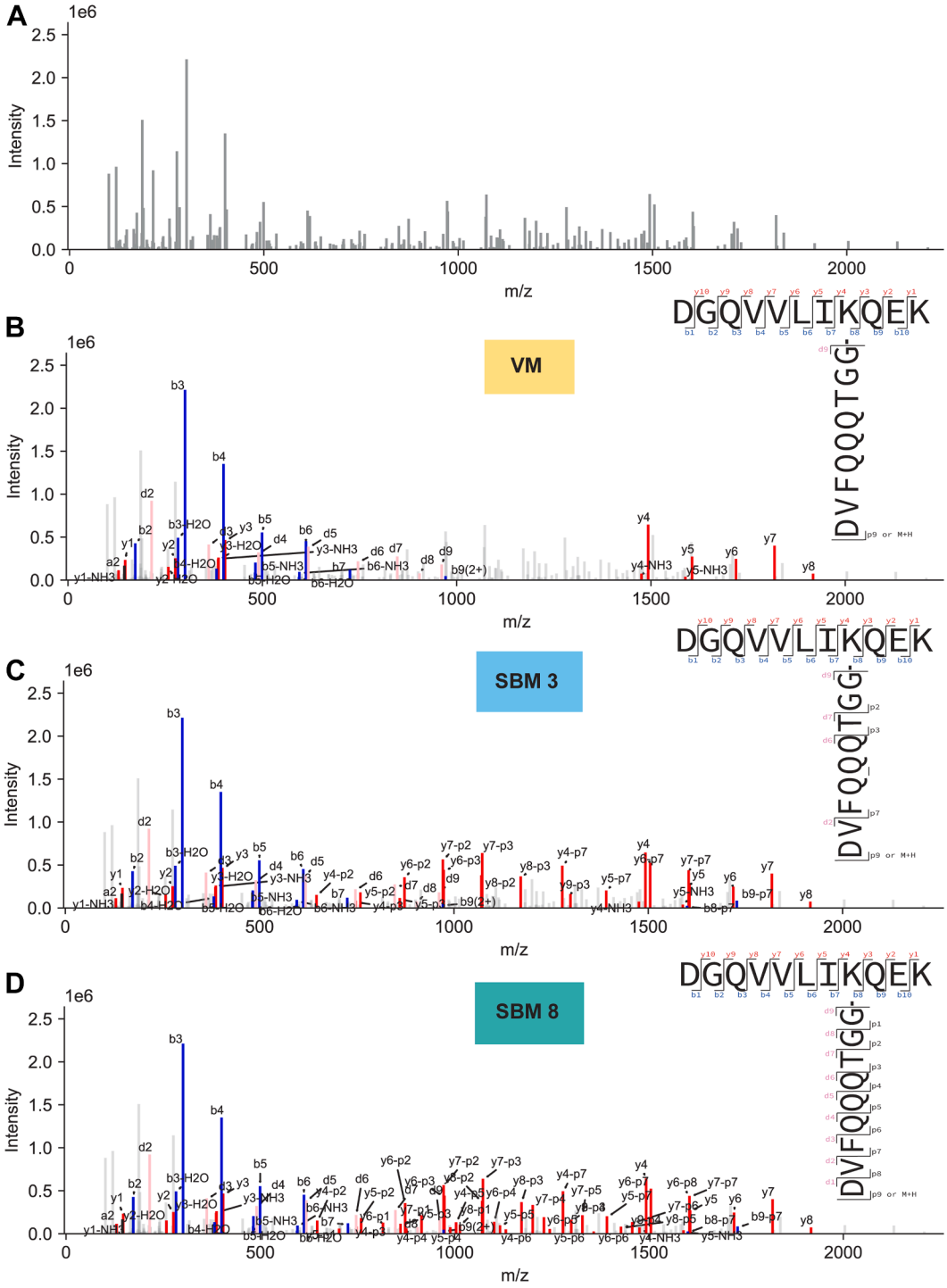


FIG. 3. **Spectral annotations of the SUMOylated peptide DGQVVLIK(sumo)QEK across different search strategies (SUMO-HEK dataset).** The top right SUMO peptide illustrations, coupled to each spectrum, shows the theoretical fragmentations that were included in each search, generating the theoretical peaks for PSM generation. *A,* all mass deconvoluted m/z peaks of the selected spectrum, deisotoped and centroided. *B,* highlighted annotations from the variable modification search space, where p-ions are not included, and SUMO is defined as a standard variable modification. *C,* annotations from the SBM 3 search space, showing annotations for p2, p3, and p7 ions. *D,* annotations from the SBM 8 search, displaying annotations for all included p-ions. PSM, peptide-spectrum match; SBM, sequence-based modifier.

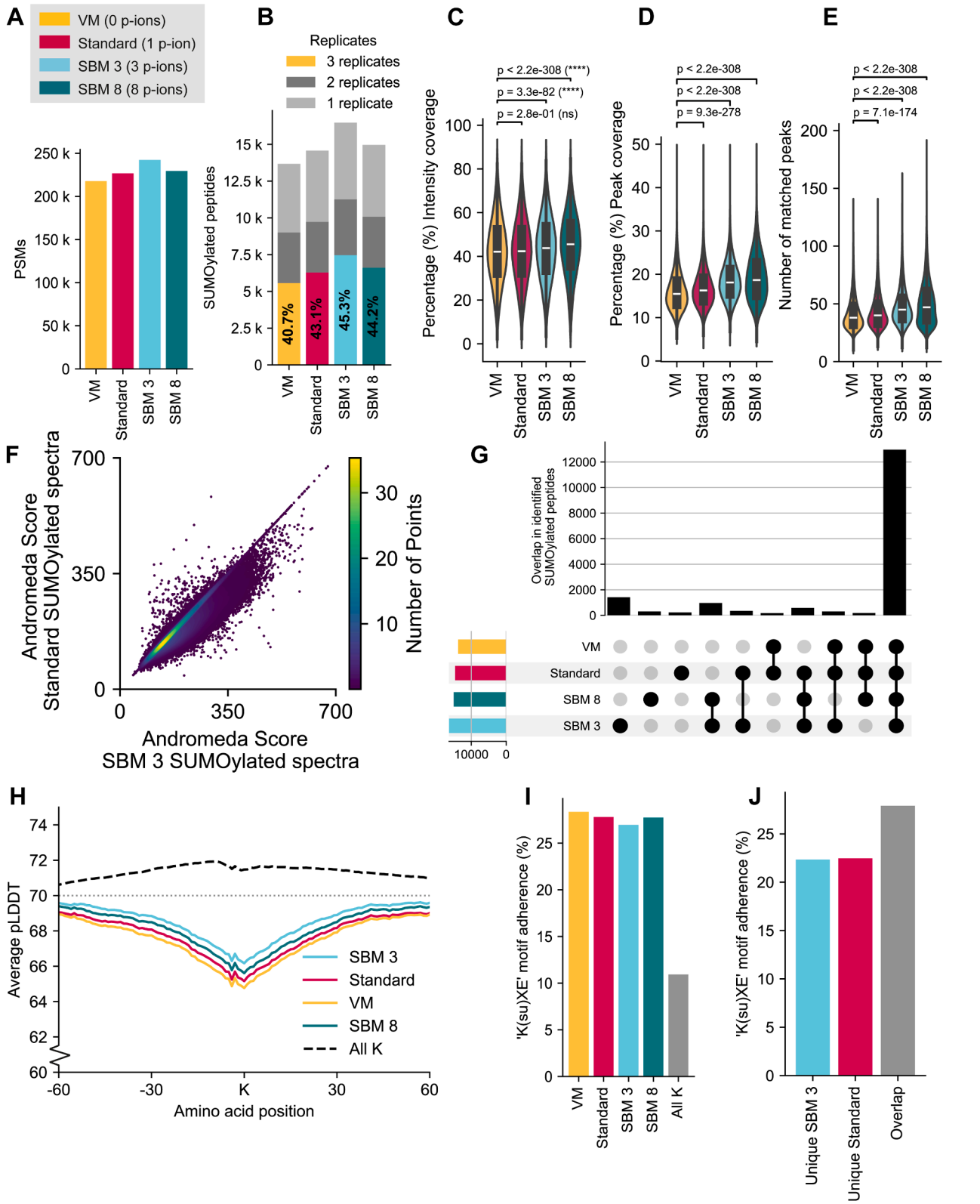


FIG. 4. **Enhanced SUMO identification using MaxSBM.** Comparison of search results from analysis of the SUMO-HEK dataset, with varying the number of p-ions in the searches: VM (0 p-ions), Standard (1 p-ion), SBM 3 (3 p-ions), or SBM 8 (8 p-ions). In all the figures, the reverse hits and contaminants have been removed. *Violin* plots illustrate the distribution of intensity coverage (%) across four datasets. Each *violin* includes an embedded boxplot showing the interquartile range, with the central line indicating the median. The number of data points (n)

downstream quantitative proteomics analysis. Overall, MaxSBM enables an increased SUMO peptide and target protein coverage, and can be applied to previously recorded datasets, and may be beneficial for studying this pivotal PTM.

Increasing the theoretical search space size is widely known in the field to decrease the identification rate (41, 42, 44). While adding theoretical peaks will allow for more matches, this comes with a concomitant increase in the probability of a decoy match. Therefore, during a DDA database search, the search space should be kept as concise and specific as possible, while still encompassing all reasonably expected peptides (41). Here, we find that the inclusion of the SBM p-ion series enhances both the sensitivity and specificity of SUMO peptide identification in database searches. However, the incorporation of the full set of eight p-ions resulted in fewer peptide identifications, likely due to highly increased odds of decoy matching. Our analysis indicated that the inclusion of three p-ions represented an optimal balance, which maximized identification performance without excessively inflating the search space. This is observed in both the highest identification rate in SBM 3 searches, as well as in the Andromeda score distributions. While the non-SUMO target hits remained the same, the SBM 3 module allowed for decreased scoring of non-SUMO reverse hits, enabling a better overall separation of the target-decoy distributions. In addition, the scoring increased for target SUMO hits in SBM 3, but interestingly not in SBM 8, showing that an additional search space increase was not beneficial. The observations are consistent with the well-established principle that expanding the number of theoretical fragment ions increases the probability of decoy matches, thereby reducing overall specificity. Consequently, while we here demonstrate the optimal p-ion assignment for the DVFQQQTGG peptide remnant, careful optimization would be essential for effective identification of other SBMs in DDA workflows.

We logically also considered processing ubiquitin data sets following Lys-C digest, that is the UbiSite method (46). However, after initial testing and benchmarking, we only saw very small increases (1%) in identifications using MaxSBM (Supplemental Fig. S11). We reason that this could be due to the overall large size (~1.4 kDa) of the ubiquitin Lys-C mass remnant, which contains multiple arginine residues, causing a very high (z = 4–6) peptide charge, ultimately causing highly convoluted fragmentation spectra that remain a technical challenge to overcome. Taking these observations together, we propose that PTMs most suitable for the MaxSBM workflow are those large enough to produce roughly 1 to 6 neutral losses or fragmentation sites, yet not so large that their fragmentation patterns overwhelm the spectra. In our endogenous analyses, SUMOylation met these criteria and aligned well with the workflow, whereas ubiquitin appeared too large for optimal performance. These findings suggest that additional PTM types within this intermediate size range may also be compatible; however, systematic evaluation will be necessary to clarify the broader applicability of the approach.

Applying Percolator to estimate peptide significance yielded a marked increase in identification numbers for the endogenous HEK SUMO dataset. Specifically, the MaxSBM module alone resulted in a 13% increase in peptide identifications relative to the standard MaxQuant search, with improvements reaching up to 70% when Percolator was applied as a post-processing step. This substantial gain is likely attributable to the inherent complexity of detecting SBM-modified peptides, which can limit score separation between target and decoy PSMs. Percolator's semi-supervised machine learning framework leverages additional PSM-derived features, such as charge state, precursor mass error, and fragment ion coverage, to improve target-decoy discrimination. This capability is particularly advantageous in settings where large search spaces and atypical fragmentation patterns challenge conventional scoring functions.

However, despite these benefits, several considerations warrant caution. Notably, both MaxQuant and MSFragger analyses incorporating Percolator identified substantial subsets of SUMOylated peptides that were not detected by the alternative search engine. These engine-specific

contributing to *violin* plots was 90,028 for VM, 97,699 for Standard, 112,409 for SBM 3, and 99,592 for SBM 8. *A,* the number of identified PSMs was 217,745 for VM, 226,671 for Standard, 242,272 for SBM 3, and 229,452 for SBM 8. *B*, unique SUMOylated peptides identified across 1, 2, or all 3 replicates. *C,* spectral intensity (abundance) coverage for the different searches. The median intensity coverage values for each dataset are as follows: VM (38%), Standard (40%), SBM 3 (45%), and SBM 8 (47%). *D,* spectral peak coverage for the different searches. The median peak coverage values for each dataset are as follows: VM (15%), Standard (16%), SBM 3 (18%), and SBM 8 (19%). *E,* average number of matched peaks for the different searches. The median number of matched peaks for each dataset is as follows: VM (38), Standard (40), SBM 3 (45), and SBM 8 (47). *F,* scatter plot comparison of the Andromeda score between the SBM 3 search and the standard search. The average Andromeda score for SBM 3 is 200.8, while for standard it's 183.9. *G,* UpSet plot visualizing unique and overlapping sets of peptides across the different searches. *H,* the average pLDDT scores from AlphaFold-predicted protein structures are shown for regions flanking SUMOylation sites. Lower pLDDT scores correspond to regions with lower structural confidence, typically indicating intrinsic disorder. For comparison, pLDDT scores for regions surrounding all unmodified lysine residues (All K) in SUMO target proteins are also shown. The enrichment of low pLDDT scores near SUMO-modified lysines adheres to the notion that SUMOylation preferentially occurs in disordered regions (6). *I,* K x E motif adherence for SUMOylated lysine residues across the different searches. The K x E motif adherence for unmodified lysine residues is shown for reference. *J,* K x E motif adherence across the overlapping SUMOylated peptides, as well as for search-specific peptides, comparing SBM 3 and standard. pLDDT, predicted local distance difference test; PSM, peptide-spectrum match; SBM, sequence-based modifier; VM, variable modification.

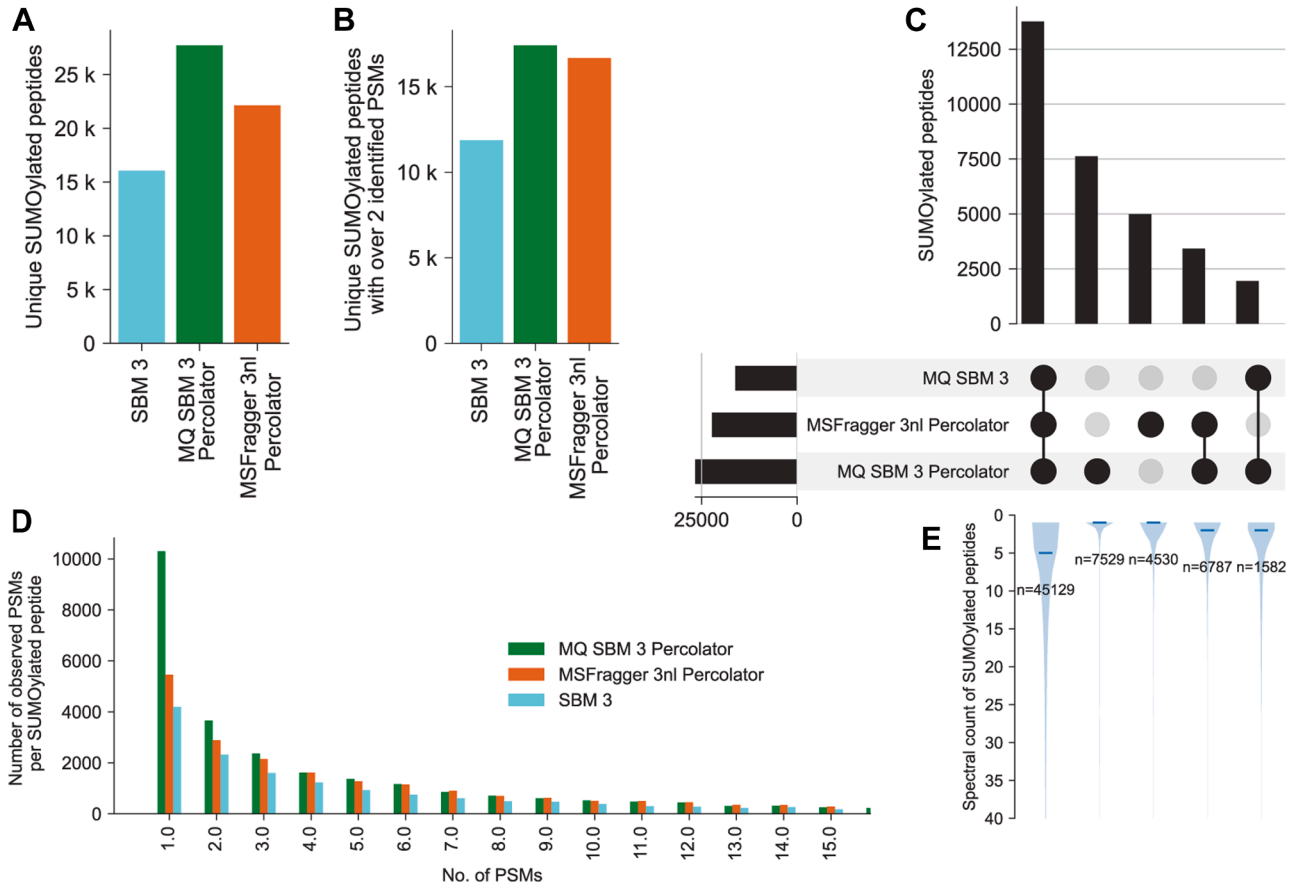


FIG. 5. **Benchmarking existing tools on the SUMO-HEK dataset.** *A,* the number of identified SUMOylated peptides across software. The MaxSBM 3 search with MaxQuant FDR filtering applied gave 16,348 SUMOylated peptides. Using MaxSBM with Percolator for significance estimation gave 27,762, and finally, using MSFragger-Labile with 3 p-ions gave 24,469. *B*, the number of identified SUMOylated peptides with at least two PSMs, compared across software. *C*, UpSet plot visualization of unique and overlapping sets of SUMO sites, comparing MaxSBM 3, MaxSBM 3 with Percolator, and MSFragger-Labile with Percolator, with 3 remainder fragment ions. *D,* comparing the number of observed PSMs per peptide for each search. *E,* distribution of PSMs per peptide for the unique and overlapping sets of identifications, shown in the upset plot. PSM, peptide-spectrum match; SBM, sequence-based modifier.

identifications were predominantly supported by a single PSM, with a median of one PSM per peptide, indicating that many of these additional identifications lie near the sensitivity limits of the respective workflows. While partial overlap between search engines is expected due to differences in scoring functions, search space definitions, and PTM handling, the enrichment of low-PSM identifications among unique peptide sets highlights the importance of carefully assessing confidence in these assignments.

Although Percolator is generally considered robust against overfitting due to its internal use of cross-validation and other regularization strategies (44, 52, 65), the absence of established well-defined ground truth in endogenous datasets complicates objective performance assessment. Moreover, recent studies by Freestone *et al*. have suggested that cross-validation procedures may be affected by the presence of multiple spectra originating from the same peptide species (66, 67), a factor that could be exacerbated in complex datasets with low signal. Consequently, the possibility that Percolator may inflate identification numbers through subtle overfitting effects cannot be entirely excluded, particularly for labile or atypically fragmenting modifications such as SBMs. In this context, the substantial increase in identifications underscores the need for careful validation.

Future work should further evaluate Percolator's FDR estimation specifically in the context of SBM identification. When comparing the search pipelines MaxSBM and MSFragger-Labile pipelines, MaxQuant yielded a higher number of identifications, albeit with a greater proportion supported by a single PSM. Together, these observations suggest that, particularly for complex PTM detection, different search strategies provide partially complementary views of the modified proteome, and that integrating results across workflows may be advantageous over reliance on a single approach. In addition, expanding and refining the feature space used for rescoring with SBM-specific attributes may further improve performance. For example, incorporating diagnostic fragment ions characteristic of SUMOylation could provide an orthogonal criterion for

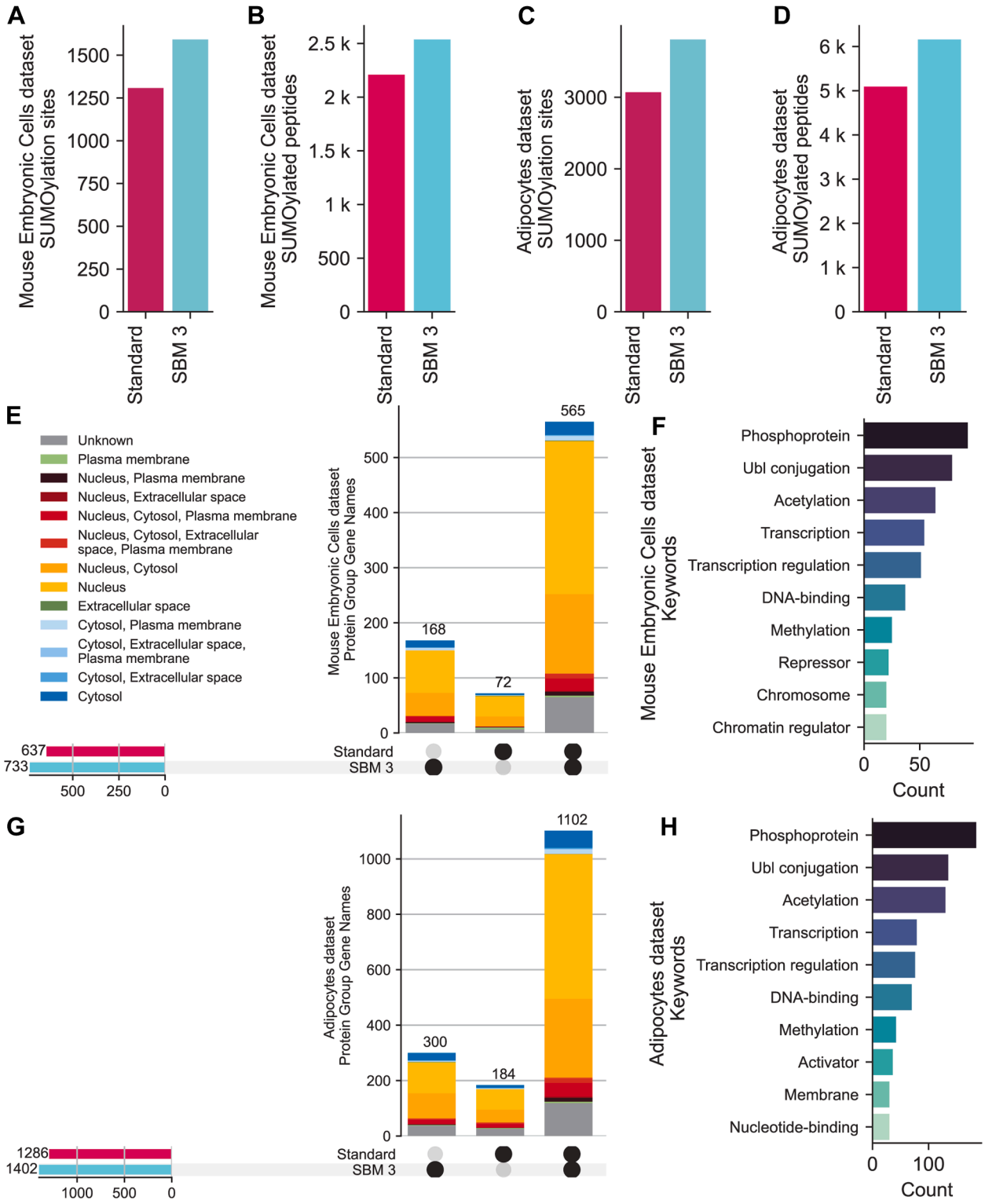


FIG. 6. **Biological insights enabled by enhanced SUMO identification.** *A,* number of identified SUMOylation sites in the SUMO-MEC dataset (19), derived from the 'sites.txt' MaxQuant output table. The optimized search SBM 3 is compared with a standard search. *B,* number of identified SUMOylated peptides in the SUMO-MEC dataset, derived from the 'msms.txt' MaxQuant output table. *C,* number of identified SUMOylation sites in the SUMO-Adip dataset (45). *D,* number of identified SUMOylated peptides in the SUMO-Adip dataset. *E,* the number of identified protein groups for the overlapping and unique sets in the SUMO-MEC dataset. The protein groups have been colored by their high-level cellular compartment (69). *F,* number of summed UniProt Keywords for the 160 unique proteins to SBM 3 in the SUMO-MEC dataset. *G,* the number of identified protein groups for the overlapping and unique sets in the SUMO-Adip dataset. *H,* number of summed UniProt Keywords for the 160 unique proteins to SBM 3 in the SUMO-Adip dataset. SBM, sequence-based modifier; SUMO-MEC, small ubiquitin-like modifier mouse embryonic cell.

distinguishing true from false matches. Finally, integrating emerging complementary approaches, such as deep learning-based spectrum prediction, may further enhance the discrimination between true and false identifications in modified peptide searches, although such strategies remain to be systematically evaluated for SBMs.

In summary, the MaxSBM module substantially enhances the identification of SUMOylated peptides, improving sensitivity and expanding proteome coverage in contemporary datasets. This improvement is achieved by explicitly modeling fragmentation within the modification itself through the incorporation of SBM-specific fragment ion series (p-ions) into the MaxQuant search framework, thereby increasing identification rates for these complex species. We further demonstrate that machine learning-based rescoring using Percolator can markedly boost identification numbers in this context, although its application to complex and labile PTMs warrants careful interpretation and further investigation. Notably, the underlying statistical framework implemented in MaxQuant remains robust, providing a stable foundation for peptide identification.

By enabling more confident and comprehensive identification of modified peptides while maintaining robust statistics, MaxSBM facilitates downstream biological interpretation, including the mapping of modification-specific pathways, identification of modification hotspots, and characterization of dynamic regulatory networks. As such, this module represents a valuable tool for researchers seeking to elucidate the complex proteomic landscapes underpinning cellular signaling and post-translational regulation. Ultimately, advancements in this field can provide a foundation for incorporating spectral prediction and DIA into the analysis of SBMs.

DATA AVAILABILITY

The module is integrated into MaxQuant search engine, which can be downloaded from www.maxquant.org [2025-02-07] [30], [40]. The software has been written in C# programming language using the Microsoft.NET 8 framework: https://dotnet.microsoft.com/. To run the application, NET 8 needs to be installed on the system. The software runs on Windows and Linux operating systems (68). Documentation on how to run the module can be found on the MaxQuant documentation page: https://cox-labs.github.io/coxdocs/MaxSBM.html [2026-01-13]. The Andromeda source code to calculate the theoretical SBM peaks can be found on GitHub: https://github.com/JurgenCox/mqtools7 [2025-06-18]. The scripts used for post-identification processing, diagnostic ion mining, protein inference and msms.txt to Percolator input conversion can be found on GitHub: https://github.com/carolinelennartsson/SUMOylation [2025-06-18]. The search output data, including processed tables (PSMs, peptides and proteins) have been deposited at https://doi.org/10.5281/zenodo.16994073.

Supplemental data—This article contains supplemental data (5, 33, 46, 50, 64, 69, 70).

Acknowledgments—We thank our colleagues at the Novo Nordisk Foundation Center for Protein Research and Max Planck Institute of Biochemistry for discussion and advice, in particular Nadezhda T. Doncheva and Jonas Damgaard Elsborg.

Author Contributions—C. L., P. K., and I. A. H. writing–review & editing; C. L. writing–original draft; C. L. and P. K. visualization; C. L. and J. C. software; C. L. methodology; C. L. investigation; C. L. formal analysis; C. L., M. L. N., and I. A. H. conceptualization; P. K., M. L. N., J. V. O., J. C., and I. A. H. validation; M. L. N. and I. A. H. project administration; M. L. N. and J. V. O. funding acquisition; J. V. O., J. C., and I. A. H. supervision; J. C. resources.

Funding and Additional Information—This work was funded in part by Independent Research Fund Denmark (0135-00096B and 8020-00220B), the European Union's Horizon 2020 research and innovation program (EPIC-XS-823839), the Danish Cancer Society (R146-A9159-16-S2), and the NNF CPR grant (ID: NNF24SA0098829).

Conflict of Interest—The authors declare no competing interests.

Abbreviations—The abbreviations used are: DDA, data-dependent acquisition; FDR, false discovery rate; PSM, peptide-spectrum matche; PTM, post-translational modification; SBM, sequence-based modifier; SUMO, small ubiquitin-like modifier; VM, variable modification.

Received September 4, 2025, and in revised form, April 6, 2026
Published, MCPRO Papers in Press, May 20, 2026, https://doi.org/10.1016/j.mcpro.2026.101589

REFERENCES

1. Olsen, J. V., and Mann, M. (2013) Status of large-scale analysis of post-translational modifications by mass spectrometry. *Mol. Cell. Proteomics* **12**, 3444–3452
2. Rigbolt, K. T. G., and Blagoev, B. (2012) Quantitative phosphoproteomics to characterize signaling networks. *Semin. Cell Dev. Biol.* **23**, 863–871
3. Ohtake, F. (2022) Branched ubiquitin code: from basic biology to targeted protein degradation. *J. Biochem.* **171**, 361–366
4. Cappadocia, L., and Lima, C. D. (2018) Ubiquitin-like protein conjugation: structures, chemistry, and mechanism. *Chem. Rev.* **118**, 889–918
5. Hendriks, I. A., Lyon, D., Su, D., Skotte, N. H., Daniel, J. A., Jensen, L. J., *et al*. (2018) Site-specific characterization of endogenous SUMOylation across species and organs. *Nat. Commun.* **9**, 2456
6. Hendriks, I. A., Lyon, D., Young, C., Jensen, L. J., Vertegaal, A. C. O., and Nielsen, M. L. (2017) Site-specific mapping of the human SUMO proteome reveals co-modification with phosphorylation. *Nat. Struct. Mol. Biol.* **24**, 325–336
7. Ulrich, H. D. (2009) The SUMO system: an overview. *Methods Mol. Biol.* **497**, 3–16
8. Johnson, E. S. (2004) Protein modification by SUMO. *Annu. Rev. Biochem.* **73**, 355–382

9. Wang, Y., and Dasso, M. (2009) SUMOylation and deSUMOylation at a glance. *J. Cell Sci.* **122**(Pt 23), 4249–4252
10. Vertegaal, A. C. O. (2022) Signalling mechanisms and cellular functions of SUMO. *Nat. Rev. Mol. Cell Biol.* **23**, 715–731
11. Ptak, C., and Wozniak, R. W. (2017) SUMO and nucleocytoplasmic transport. *Adv. Exp. Med. Biol.* **963**, 111–126
12. Borgermann, N., Ackermann, L., Schwertman, P., Hendriks, I. A., Thijssen, K., Liu, J. C., *et al*. (2019) SUMOylation promotes protective responses to DNA-protein crosslinks. *EMBO J.* **38**, e101496
13. Dantuma, N. P., and van Attikum, H. (2016) Spatiotemporal regulation of posttranslational modifications in the DNA damage response. *EMBO J.* **35**, 6–23
14. Dhingra, N., and Zhao, X. (2019) Intricate SUMO-based control of the homologous recombination machinery. *Genes Dev.* **33**, 1346–1354
15. Richard, P., Vethantham, V., and Manley, J. L. (2017) Roles of sumoylation in mRNA processing and metabolism. *Adv. Exp. Med. Biol.* **963**, 15–33
16. Talamillo, A., Barroso-Gomila, O., Giordano, I., Ajuria, L., Grillo, M., Mayor, U., *et al*. (2020) The role of SUMOylation during development. *Biochem. Soc. Trans.* **48**, 463–478
17. Keiten-Schmitz, J., Schunck, K., and Müller, S. (2019) SUMO chains rule on chromatin occupancy. *Front. Cell Dev. Biol.* **7**, 343
18. Wilson, N. R., and Hochstrasser, M. (2016) The regulation of chromatin by dynamic SUMO modifications. *Methods Mol. Biol.* **1475**, 23–38
19. Theurillat, I., Hendriks, I. A., Cossec, J.-C., Andrieux, A., Nielsen, M. L., and Dejean, A. (2020) Extensive SUMO modification of repressive chromatin factors distinguishes pluripotent from somatic cells. *Cell Rep.* **32**, 108146
20. Chymkowitch, P., Nguéa P, A., and Enserink, J. M. (2015) SUMO-regulated transcription: challenging the dogma. *Bioessays* **37**, 1095–1105
21. Rosonina, E., Akhter, A., Dou, Y., Babu, J., and Sri Theivakadadcham, V. S. (2017) Regulation of transcription factors by sumoylation. *Transcription* **8**(4), 220–231
22. Augustine, R. C., and Vierstra, R. D. (2018) SUMOylation: re-wiring the plant nucleus during stress and development. *Curr. Opin. Plant Biol.* **45**(Pt A), 143–154
23. Kessler, J. D., Kahle, K. T., Sun, T., Meerbrey, K. L., Schlabach, M. R., Schmitt, E. M., *et al*. (2012) A SUMOylation-dependent transcriptional subprogram is required for Myc-driven tumorigenesis. *Science* **335**, 348–353
24. Sireesh, D., Bhakkiyalakshmi, E., Ramkumar, K. M., Rathinakumar, S., Jennifer, P. S. A., Rajaguru, P., *et al*. (2014) Targeting SUMOylation cascade for diabetes management. *Curr. Drug Targets* **15**, 1094–1106
25. Kranias, E. G., and Hajjar, R. J. (2012) Modulation of cardiac contractility by the phospholamban/SERCA2a regulatome. *Circ. Res.* **110**, 1646–1660
26. Mendler, L., Braun, T., and Müller, S. (2016) The ubiquitin-like SUMO system and heart function: from development to disease: from development to disease. *Circ. Res.* **118**, 132–144
27. Chang, H.-M., and Yeh, E. T. H. (2020) SUMO: from bench to bedside. *Physiol. Rev.* **100**, 1599–1619
28. Zeng, M., Liu, W., Hu, Y., and Fu, N. (2020) Sumoylation in liver disease. *Clin. Chim. Acta* **510**, 347–353
29. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805
30. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372
31. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., and Nesvizhskii, A. I. (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520
32. Geiszler, D. J., Polasky, D. A., Yu, F., and Nesvizhskii, A. I. (2023) Detecting diagnostic features in MS/MS spectra of post-translationally modified peptides. *Nat. Commun.* **14**, 4132
33. Polasky, D. A., Geiszler, D. J., Yu, F., Li, K., Teo, G. C., and Nesvizhskii, A. I. (2023) MSFragger-labile: a flexible method to improve labile PTM analysis in proteomics. *Mol. Cell. Proteomics* **22**, 100538
34. Polasky, D. A., Yu, F., Teo, G. C., and Nesvizhskii, A. I. (2020) Fast and comprehensive N- and O-glycoproteomics analysis with MSFragger-Glyco. *Nat. Methods* **17**, 1125–1132
35. Yang, B., Wu, Y. J., Zhu, M., Fan, S. B., Lin, J., Zhang, K., *et al*. (2012) Identification of cross-linked peptides from complex samples. *Nat. Methods* **9**, 904–906
36. Liu, C., Wu, T., Shu, X., Li, S. T., Wang, D. R., Wang, N., *et al*. (2021) Identification of protein direct interactome with genetic code expansion and search engine OpenUaa. *Adv. Biol. (Weinh.)* **5**, e2000308
37. Pedrioli, P. G. A., Raught, B., Zhang, X. D., Rogers, R., Aitchison, J., Matunis, M., *et al*. (2006) Automated identification of SUMOylation sites using mass spectrometry and SUMmOn pattern recognition software. *Nat. Methods* **3**, 533–539
38. 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
39. Trulsson, F., and Vertegaal, A. C. O. (2022) Site-specific proteomic strategies to identify ubiquitin and SUMO modifications: challenges and opportunities. *Semin. Cell Dev. Biol.* **132**, 97–108
40. Tyanova, S., Temu, T., and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319
41. Bugyi, F., Szabó, D., Szabó, G., Révész, Á., Pape, V. F. S., Soltész-Katona, E., *et al*. (2021) Influence of post-translational modifications on protein identification in database searches. *ACS Omega* **6**, 7469–7477
42. Khatri, K., Klein, J. A., and Zaia, J. (2017) Use of an informed search space maximizes confidence of site-specific assignment of glycoprotein glycosylation. *Anal. Bioanal. Chem.* **409**, 607–618
43. Shanmugam, A. K., and Nesvizhskii, A. I. (2015) Effective leveraging of targeted search spaces for improving peptide identification in tandem mass spectrometry based proteomics. *J. Proteome Res.* **14**, 5169–5178
44. Verbruggen, S., Gessulat, S., Gabriels, R., Matsaroki, A., Van de Voorde, H., Kuster, B., *et al*. (2021) Spectral prediction features as a solution for the search space size problem in proteogenomics. *Mol. Cell. Proteomics* **20**, 100076
45. Zhao, X., Hendriks, I. A., Le Gras, S., Ye, T., Ramos-Alonso, L., Nguéa P, A., *et al*. (2022) Waves of sumoylation support transcription dynamics during adipocyte differentiation. *Nucleic Acids Res.* **50**, 1351–1369
46. Akimov, V., Barrio-Hernandez, I., Hansen, S. V. F., Hallenborg, P., Pedersen, A. K., Bekker-Jensen, D. B., *et al*. (2018) UbiSite approach for comprehensive mapping of lysine and N-terminal ubiquitination sites. *Nat. Struct. Mol. Biol.* **25**, 631–640
47. UniProt Consortium, T. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699
48. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925
49. Park, C. Y., Klammer, A. A., Käll, L., MacCoss, M. J., and Noble, W. S. (2008) Rapid and accurate peptide identification from tandem mass spectra. *J. Proteome Res.* **7**, 3022–3027
50. McIlwain, S., Tamura, K., Kertesz-Farkas, A., Grant, C. E., Diament, B., Frewen, B., *et al*. (2014) Crux: rapid open source protein tandem mass spectrometry analysis. *J. Proteome Res.* **13**, 4488–4491
51. Declercq, A., Bouwmeester, R., Hirschler, A., Carapito, C., Degroeve, S., Martens, L., *et al*. (2022) MS2Rescore: Data-driven rescoring dramatically boosts immunopeptide identification rates. *Mol. Cell. Proteomics* **21**, 100266
52. Granholm, V., Noble, W. S., and Käll, L. (2012) A cross-validation scheme for machine learning algorithms in shotgun proteomics. *BMC Bioinformatics* **13**, S3
53. da Veiga Leprevost, F., Haynes, S. E., Avtonomov, D. M., Chang, H. Y., Shanmugam, A. K., Mellacheruvu, D., *et al*. (2020) Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat. Methods* **17**, 869–870
54. McKinney, W. (2010) Data structures for statistical computing in python. *Proceedings of the Python in Science ConferenceProceedings of the Python in Science Conference* **445**,. Austin, Texas: 56–61
55. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., *et al*. (2020) Array programming with NumPy. *Nature* **585**, 357–362
56. Hunter, J. D. (2007) Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95

57. Waskom, M. (2021) Seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021
58. Kunzmann, P., Müller, T. D., Greil, M., Krumbach, J. H., Anter, J. M., Bauer, D., *et al*. (2023) Biotite: new tools for a versatile Python bioinformatics library. *BMC Bioinformatics* **24**, 236
59. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., *et al*. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589
60. Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., *et al*. (2024) AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375
61. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., *et al*. (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272
62. Tyanova, S., and Cox, J. (2018) Perseus: a bioinformatics platform for integrative analysis of proteomics data in cancer research. *Methods Mol. Biol.* **1711**, 133–148
63. Lamoliatte, F., McManus, F. P., Maarifi, G., Chelbi-Alix, M. K., and Thibault, P. (2017) Uncovering the SUMOylation and ubiquitylation crosstalk in human cells using sequential peptide immunopurification. *Nat. Commun.* **8**, 14109
64. Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., *et al*. (2023) The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646
65. The, M., MacCoss, M. J., Noble, W. S., and Käll, L. (2016) Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. *J. Am. Soc. Mass Spectrom.* **27**, 1719–1727
66. Freestone, J., Noble, W. S., and Keich, U. (2024) Reinvestigating the correctness of decoy-based false discovery rate control in proteomics tandem mass spectrometry. *J. Proteome Res.* **23**, 1907–1914
67. Freestone, J., Käll, L., Noble, W. S., and Keich, U. (2025) How to train a postprocessor for tandem mass spectrometry proteomics database search while maintaining control of the false discovery rate. *J. Proteome Res.* **24**, 2266–2279
68. Sinitcyn, P., Tiwary, S., Rudolph, J., Gutenbrunner, P., Wichmann, C., Yılmaz, Ş., *et al*. (2018) MaxQuant goes Linux. *Nat. Methods* **15**, 401
69. Binder, J. X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S. I., Schneider, R., *et al*. (2014) COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database (Oxford)* **2014**, bau012
70. Tyanova, S., Temu, T., Carlson, A., Sinitcyn, P., Mann, M., and Cox, J. (2015) Visualization of LC-MS/MS proteomics data in MaxQuant. *Proteomics* **15**, 1453–1456