rPTMDetermine: A Fully Automated Methodology for Endogenous Tyrosine Nitration Validation, Site-Localization, and Beyond

Naiping Dong [a], Daniel M. Spencer [a], Quan Quan [a, †], J. C. Yves Le Blanc [b], Jinwen Feng [a, ‡], Mengzhu Li [a], K. W. Michael Siu [a, c, d], Ivan K. Chu [a]*

[a] Department of Chemistry, The University of Hong Kong, Pokfulam, Hong Kong, China; [b] SCIEX, 71 Four Valley Drive, Concord, Ontario, L4K 4V8, Canada; [c] Department of Chemistry and Centre for Research in Mass Spectrometry, York University, Canada; [d] Department of Chemistry and Biochemistry, University of Windsor, Canada.

* Corresponding author: Ivan K. Chu, E-mail: ivankchu@hku.hk
† Present address: BeiGene (Beijing) Co., Ltd., Beijing, China
‡ Present address: Human Phenome Institute, Fudan University, Shanghai, China

ABSTRACT

We present herein rPTMDetermine, an adaptive and fully automated methodology for validation of the identification of rarely occurring post-translational modifications (PTMs), using a semi-supervised approach with a linear discriminant analysis (LDA) algorithm. With this strategy, verification is enhanced through similarity scoring of tandem mass spectrometry (MS/MS) comparisons between modified peptides and their unmodified analogues. We applied rPTMDetermine to (1) perform fully automated validation steps for modified peptides identified from an *in silico* database and (2) retrieve potential yet-to-be-identified modified peptides from raw data (that had been missed through conventional database searches). In part (1), 99 of 125 3-nitrotyrosyl–containing (nitrated) peptides obtained from a ProteinPilot search were validated and localized. Twenty nitrated peptides were falsely assigned because of incorrect monoisotopic peak assignments, leading to erroneous identification of deamidation and nitration. Five additional nitrated peptides were, however, validated after performing non-monoisotopic peak correction. In part (2), an additional 236 unique nitrated peptides were retrieved and localized, containing 113 previously unreported nitration sites; 25 endogenous nitrated peptides with novel sites were selected and verified by comparison with synthetic analogues. In summary, we identified and confidently validated 296 unique nitrated peptides—collectively representing the largest number of endogenously identified 3-nitrotyrosyl–containing peptides from the cerebral cortex proteome of a *Macaca fascicularis* model of stroke. Furthermore, we harnessed the rPTMDetermine strategy to complement conventional database searching and enhance the confidence of assigning rarely occurring PTMs, while recovering many missed peptides. In a final demonstration, we successfully extended the application of rPTMDetermine to peptides featuring tryptophan oxidation.

INTRODUCTION

Post-translational modification (PTM) is the alteration of a protein following mRNA translation; most PTMs arise from covalent additions to amino acid side chains. Some PTMs modulate cellular function, others result from environmental stress or disease. Because a protein can have multiple possible sites with differing functional significance for a given type of PTM, comprehensive sequence coverage and site mapping are desirable in proteome analyses. Advancements in mass spectrometry (MS) are rendering comprehensive and high-throughput proteomics analyses a reality. Combining proteolytic sample digestion, liquid chromatography (LC) separation, and high-throughput data-dependent tandem mass spectrometry (MS/MS) acquisition enables rapid peptide—and, thus, protein—identification using protein sequence database search engines (*e.g.*, Mascot[1], SEQUEST[2], ProteinPilot[3]), which can be employed for PTM analysis.

Confident identification of rarely occurring PTMs [of specific residue(s) in protein(s) in the proteome] is challenging because of the diverse range and low abundance of modifications. The former requires costly computationally intensive *in silico* database identification; the latter results in poor analytical detection, due to the low stoichiometric concentration. Comprehensive coverage and mapping of all possible PTM sites is essential to thoroughly characterize the biological function of any given protein modification. PTM-bearing peptides are reported to be responsible for 20–50% of false-positive identifications,[4] likely resulting in missed and misleading biological discoveries.

Confident PTM identifications are often claimed if they pass false discovery rate (FDR) control; this approach may, however, inaccurately estimate the rate of false-positives in rarely occurring PTM identifications.[5–7] The target-and-decoy strategy, which gauges the extent of false-positives by searching the target database and a decoy database of non-existent proteins (usually derived from the reversed target sequences), is routinely applied in database searching.[8] Accordingly, matches to decoy peptides signify incorrect identifications. The FDR estimates the confidence in the search results in terms of the ratio of the number of decoy-to-target peptide spectrum matches (PSMs) with scores above a threshold. False identifications can escape FDR control because of, for example, the low mass-accuracy of the instrumentation or restrictive search parameters.[9] Applying global FDR to confidently estimate PTM identifications assumes that the FDR for this subset is identical to that for the complete set of peptides. The PSM scores depend, however, on the PTMs and the compositions; thus, the assumption of uniformity may not be valid. Subgroup FDR control has been proposed to estimate the error rate in PTM identifications,[5,10] but this approach has its detractors because the smaller sample sizes broaden confidence intervals in the FDR.[8,11] An extra challenge arises from PTM localization in peptides with multiple potential sites; many search engines do not offer a means for PTM localization, so specialized software tools are required to serve this purpose.[12]

Manual validation has often been used for PTM identifications; unfortunately, manual validation is labor intensive, low throughput, and heavily dependent on the researchers' expertise. To circumvent the limitations of FDR control and the labor-intensive steps in manual validation, some previously developed tools have employed semi-supervised machine learning, which separates unknown correct identifications from incorrect (decoy) identifications, and evaluates PSMs from database searches.[13,14] Most such tools had been designed for validation of the whole set of search results; thus, the heterogeneity in peptide identifications remained, making the tools unamenable to identifying rarely occurring PTMs. The validation problem aside, database search engines leave most mass spectra unassigned (or with low-confidence assignments), including high-quality spectra that may derive from peptides bearing an untargeted PTM.[15,16] Clustering algorithms have been applied to group mass spectra, including unassigned spectra, into a cohort originating from a single peptide, thereby providing an avenue for identification.[17,18] This approach is ineffective for rare PTMs because the modified peptide is unlikely to be identified from multiple spectra.

An alternative method for retrieving PTM identifications uses an unrestricted search without specifying the allowed PTMs.[19] Although this approach can comprehensively identify peptide PTMs, the confidence of the identifications is not fully evaluated, limiting its utility.[20] Confidence is enhanced by comparison of the mass spectra of modified peptides with their unmodified counterparts; this approach does, however, rely on coexistence of these peptides.[21,22]

Herein, we present "rPTMDetermine," an adaptive and automated tool that adopts a semi-supervised approach with linear discriminant analysis (LDA)—a commonly used and effective machine learning algorithm—to differentiate modified peptide identifications from decoy identifications; it also uses similarity scoring between product ion spectra of unmodified peptides and those of their modified analogues, where PTMs do not significantly perturb fragmentation patterns, as an orthogonal criterion. Finally, PTM localization is performed through comparison of the LDA scores of all positional isoforms. By incorporating these two orthogonal approaches, we have employed rPTMDetermine successfully to search large numbers of unassigned mass spectra for the retrieval of missed PTM identifications. As a demonstration, we applied rPTMDetermine for the validation and site-localization of 3-nitrotyrosyl–containing peptides from database identification; we also used it for the retrieval of potential and yet-to-be-identified 3-nitrotyrosyl–containing peptides (given that the database search engines leave almost 60% of high-quality mass spectra unassigned or with only low-confidence assignments). Identifying ortho protein tyrosine nitration (PTN, **Scheme S1**) is important because this modification is a known biomarker of neurodegenerative disorders and age-related diseases.[23] Furthermore, because nitration is both regioselective and highly selective for specific tyrosyl residues,[24] detection of further endogenous nitration sites may aid in determining the factors that promote nitration. Previously, such identification has been challenging because of the low stoichiometry of PTN modification (estimated occurrence: <0.001%). We also demonstrate the extensibility of this strategy toward a second type of PTM: tryptophan oxidation.

EXPERIMENTAL SECTION

LC-MS/MS Data. The results described herein involve the re-analysis and re-interpretation of MS/MS spectra (with experimental details) that we have previously reported[25] and are available in the ProteomeXchange repository (http://www.proteomexchange.org/, PXD003173).[26] The data from four runs (I08, I17, I18, I19), which captured nitrated peptides during a four-dimensional LC separation with online MS/MS, are employed herein.

Data Extraction. The MS/MS data in the four runs were searched by means of ProteinPilot 4.5 against the *Macaca fascicularis* proteome database (January 2014; 56,572 entries; http://www.ncbi.nlm.nih.gov/). Previously, 40 identified nitrated peptides were confirmed by means of manual validation through comparison with the chromatographic and MS/MS properties of custom-synthesized nitrated peptides.[25] These peptides and their mass spectra are used herein as benchmarks to train the new validation procedure.

Mass Spectral Data Preprocessing. Isotope-encoded reporter ions for the iTRAQ 8plex (*i.e.*, *m*/*z* 113.1, 114.1, 115.1, 116.1, 117.1, 118.1, 119.1, and 121.1) were first removed from the mass spectra to avoid incorrect annotations to these fragments, which might have influenced subsequent de-noising and led to overfitting. A heuristic approach was then implemented to remove noise; it divided the spectra into 100 Da–wide segments, each retaining a maximum of eight peaks according to annotations derived from the theoretical ions of the assigned peptides (**Supporting Note I**).

Target and Decoy Set Construction. Target nitrated peptide identifications were collected by searching against the *M. fascicularis* proteome (target) database using ProteinPilot with a 1% global FDR control. Because ProteinPilot did not output decoy nitrated peptide identifications, the decoy peptide set was constructed *in silico* by reversing and theoretically digesting all protein sequences in the target protein database. All tyrosine residues in the decoy peptides were nitrated *in silico*, including all permutations of

nitration for peptides with multiple tyrosine residues (up to a maximum of three PTN modifications), thereby permitting construction of a decoy nitrated peptide database. Decoy nitrated peptide identifications were obtained by searching the mass spectra assigned to target nitrated peptides against the nitrated decoy database with a peptide tolerance of *m*/*z* 0.01. Details are available in **Supporting Note II**.

Validation Model Construction. For each target and decoy identification, 16 features (variables, see **Table S1**), which have been demonstrated as effective in evaluating PSMs,[13,27,28] were determined to assess the quality of each PSM (**Table S2**). Fisher scores, a measure of the difference in distribution of a variable between two or more data sets, were used to determine the features capable of discriminating between the target and decoy data sets.[29] The 16 features (variables) were calculated for each PTN identification; 13 features with Fisher scores much greater than zero (**Table S2**, **Figure S1**) were capable of discriminating between the target and decoy identifications. The full dataset (the combination of target and decoy data sets) was partitioned repeatedly into training and testing sets, and trained (using the training data set) and evaluated (using the testing data set) by an LDA model using the selected features. Each identification (including decoy) was assigned an LDA score, alongside a posterior probability (*p*) of being correct based on the target and decoy score distributions (**Supporting Note II**).[30] Those identifications with $p \geq 0.99$, theoretically guaranteeing an FDR below 1%,[31] were accepted (**Figure S2**).

Tyrosine Nitration Localization. To determine the specific site of nitration in peptides with multiple tyrosine sites, the peptide isoform candidates were generated by permuting nitration over all possible sites. Each candidate was validated against the given tandem mass spectrum, using the rPTMDetermine procedure, and assigned a site probability (**Supporting Note III**). If the site probability for an isoform exceeded the specified threshold, at 0.99, then the nitration was considered localized to the tyrosine site specified by the isoform. Otherwise, the site of modification was considered ambiguous.

Spectral Similarity. For each nitrated peptide, the non-nitrated analogue was validated using the LDA validation model with 11 of the abovementioned 13 features, because the other two (*TotalIntMod* and *MatchScoreMod*) are specific to modified peptides. The normalized dot product was used to calculate similarity scores, which quantified the overlap between the mass spectra of nitrated peptides and those of their validated non-nitrated analogues in terms of peak annotations and intensities. The formulae for score calculations are provided in **Supporting Note IV**.

Software Availability. rPTMDetermine was written in Python; it is freely available, along with instructions, at https://github.com/ikcgroup/rPTMValidation.

RESULTS AND DISCUSSION

Our study presented herein was motivated by the need for a fully automated methodology for the identification of rarely occurring post-translationally modified peptides in terms of (1) their validation from protein sequence database searches using MS/MS data and (2) the recovery of potential yet-to-be-identified, but unassigned, product ion mass spectra, missed by database search engines.

Validation of Post-Translationally Modified Peptide Identifications

rPTMDetermine (**Figure 1**) is a fully automated validation procedure that employs two sequential steps: (a) semi-supervised approach through LDA to validate and localize rarely occurring modified peptide identifications and (b) verification of these identifications through comparison of the mass spectra assigned to modified peptides and their unmodified counterparts.

*Step (a): LDA Validation (**Figure 1a**)*. This step involves an adaptive and fully automated validation tool developed for downstream analyses of target modified peptides. Briefly, target modified PSMs (mPSMs) are extracted from the database search results. An *in silico* modified decoy peptide database is constructed by reversing the target protein sequence database, performing theoretical digestion, then

permuting the modification over applicable target residues. Sixteen features (the most frequently acquired variables from low-energy CID spectra) measuring the match quality are used to characterize each mPSM (see **Experimental Section**). After classification, each PSM is measured and represented by a single LDA score (**Supporting Note II**, **Figure S2**). For each PSM having multiple residues susceptible to modification, all other potential alternative modified target sites are permuted to generate isoforms for each of the modified peptides (see **Figure 1c**, **Supporting Note III** for detailed implementations).

*Step (b): Orthogonal Assessment of LDA-Validated mPSMs (**Figure 1b**)*. Many rare PTM peptides have very low stoichiometry; therefore, in large-scale shotgun proteomics analyses, each modified peptide identification is typically complemented by a relatively high-abundance unmodified analogue. Each validated post-translationally modified peptide can, therefore, be further confirmed using additional orthogonal assessment—based on the similarity between the mass spectra of modified peptides and their unmodified analogues. Spectral similarity scores are computed through quantitative comparison of the spectra for the validated modified and unmodified PSMs (**Supporting Note IV**). This orthogonal measure could be widely applicable for modified peptides that are accompanied by unmodified analogues with similar peptide fragmentation patterns.

In this paper, we demonstrate the applicability of the proposed methodology with particular emphasis on 3-nitrotyrosyl–containing peptides (their validation, site-localization, and retrieval); we also demonstrate the extensibility of the established approach using oxidized tryptophan-containing peptides.

Verification of Identified 3-Nitrotyrosyl–Containing Peptides from ProteinPilot

As a demonstration, we re-evaluated the PTN peptides in a *M. fascicularis* cerebral cortex 3-nitrotyrosyl–containing proteome dataset—one of the largest collections of 3-nitrotyrosyl–containing peptides that has been characterized and validated for a non-human primate ischemic stroke model. Of the 125 unique nitrated peptides (233 PSMs) identified by ProteinPilot at global FDR 1%, 117 peptides [94%, 223 (96%) PSMs] were accompanied by non-nitrated analogues with the same charge-state, rendering applicable the validation approach using non-modified peptides proposed herein. The similarity score threshold was set at 0.42, derived by comparing the mass spectra of synthetic PTN benchmarks with those of non-nitrated analogues (**Supporting Note II**, **Figure S3**). After validation and localization by rPTMDetermine, 110 unique nitrated peptides (211 PSMs) were accepted (**Figure 2a**). **Figure 2b** displays MS/MS spectra of the nitrated (upper panel) and non-nitrated (lower panel) analogues of the undecapeptide $EAVCEVALDYK^{3+}$, revealing similar fragmentation patterns and the expected 45-Da shift arising from the nitration of residue Y (an LDA score of 8.12 and a similarity score of 0.80). It is a common observation that the non-nitrated peptides provide substantially better overall precursor and product ion signals, as expected from their higher stoichiometric abundances (see the extracted ion chromatograms in the inset to **Figure 2b**). In addition, the retention times of nitrated peptides are longer than those of their non-nitrated analogues; for example, in **Figure 2b**, the retention times are 71.5 and 57.1 min, respectively. Of all the pairs examined, the differences in retention times ranged from approximately 1 to 17 min, with those of most peptide pairs falling between 7 and 9 min (**Figure S4**). **Figure S5** provides additional examples of identification. **Figure S6** presents an example of an ambiguous localization.

Nitrated Peptides without Unmodified Analogues

The observation of eight unique nitrated peptide identifications featuring high LDA scores, but without their unmodified analogues, was unexpected (**Figure 2a**, upper-left quadrant with similarity score = 0), due to the higher stoichiometric concentration of non-nitrated analogues. Thus, we further characterized, and manually inspected, these product ion spectra, revealing that five of them featured false deamidation (**Figure 2a**, red triangles). The mis-assignments were due to mistakes in determining the monoisotopic m/z ratios for multiple charged precursor ions (deamidation and oxidation results in a net

mass increase of only 0.98 Da). The inset to **Figure 3** compares the experimental and theoretical isotopic distributions of an incorrect deamidation identification; the same MS/MS spectrum was annotated using the theoretical fragment ions of the deamidated and original (non-deamidated) versions of the peptide. Incorporating a deamidation correction algorithm into rPTMDetermine (as described in **Supporting Note V**) enhanced the fragment ion coverage and, as a result, improved the LDA and similarity scores, thereby resulting in successful recovery of those peptides with mis-assigned deamidation. Of the 20 deamidated nitrated peptides identified by ProteinPilot, we determined that 15 (of 20) had been erroneously assigned as deamidated (**Table S3** provides the LDA and similarity scores before and after applying the deamidation correction). Inspection of the remaining five unique nitrated peptides that failed validation revealed that they either had low-confidence non-nitrated analogues or displayed complex isotopic distribution patterns (**Supporting Note VI**).

In summary, the application of conservative validation and deamidation correction criteria resulted in a final list of 104 unambiguous and unique identifications, including 99 unique nitration sites, that were validated by rPTMDetermine (**Table S4**). This list contains all 40 of the high-confidence identifications that were previously validated with synthetic standards.[25]

Retrieval of Unassigned Product Ion Spectra of Modified Peptides

rPTMDetermine was further applied to retrieve and characterize additional 3-nitrotyrosyl–containing peptides from unassigned product ion spectra (a total of 954,619 or 79% of the total number of experimental mass spectra acquired) (**Supporting Note VII**). As demonstrated above, nitrated peptides were accompanied by non-nitrated analogues having the same charge-state. To retrieve unassigned spectra with tyrosine nitration, sets of high-confidence unmodified tyrosine-containing peptides were nitrated *in silico* by permuting nitration over all tyrosine residues to generate a theoretical peak list. The spectra potentially corresponding to 3-nitrotyrosyl–containing peptides were then extracted and filtered based on the list of theoretical precursor *m*/*z* values of the modified peptides.

The shortlisted mPSMs were then subjected to rPTMDetermine validation, as described above. mPSMs that met the established LDA and similarity score criteria were considered as new nitrated peptide identifications. From an initial set of 489 mPSMs, we identified an additional 236 nitrated peptides with localized sites (**Figures 4a** and **4b**). Of these mPSMs, 44 (35 unique nitrated peptides) were identified by ProteinPilot but failed FDR control.

Interestingly, a portion of the recovered unique nitrated peptide sequences (44 peptides, 19%) overlapped with the nitrated peptides that had been originally identified by ProteinPilot, but were missed/eluted at different retention times/sub-fractions; 16 were among the 40 that had been synthetically benchmarked (**Figure 4c**, **Table S4**). An example of such a retrieved nitration identification is $VGGY^{NO2}ILGEFGNLIAGDPR^{3+}$ (where nitration is indicated by "NO2"), which had been identified previously from product ion spectra. The newly assigned mass spectrum was of high quality (**Figures 5a** and **5b**). Complementary support for the proposed nitrated peptide includes an extended retention time (**Figure 5c**) and a fragmentation pattern similar to that of the non-nitrated analogue (**Figure 6a**, lower panel). The strongest evidence comes from comparison with the synthetic nitrated peptide (**Figure 5d**)—the resemblance is striking. It is evident that high-quality yet-to-be-identified product ion spectra were, indeed, frequently unassigned and missed by the database search engine.

The causes for failure in assignment are not completely clear; some are likely due to inaccurate charge and precursor *m*/*z* determinations, limited database search parameters (*e.g.*, cleavage rules and PTMs), unknown modifications, and low scores (*e.g.*, FDR control).[15,18,32,33] ProteinPilot employs a two-mode search to relax restrictions on proteolytic digestion and PTMs, and a complex statistical scoring scheme based on such factors as sequence tags, modifications (represented by feature probability which quantify modification occurring in nature), proteolytic digestion, and precursor mass error.[3] Our results

suggest that incorrect charge determination and low feature probabilities[3] for rare biological modifications might be responsible for most of the missed identifications. Although increasing the tyrosine nitration feature probability resulted in recovery of the majority of the retrieved nitrated peptides (>75%), it incurred a specificity penalty in the FDR-filtered results from ProteinPilot, leading to a considerable number of ambiguous identifications (**Supporting Note VIII**). Thus, such probability adjustment should be used with extreme caution; instead, we recommend that rPTMDetermine be used to recover the missed identifications.

The ability of the LDA similarity-based methodology to retrieve and validate a large population of nitrated peptides (236 unique endogenous 3-nitrotyrosyl–containing peptides) is particularly encouraging. It is intriguing that 214 are newly identified unique nitrated peptides; Quan *et al.* previously identified 125 endogenous nitrotyrosine-modified peptides using the same dataset from the cerebral cortex proteome of an *M. fascicularis* model of stroke.[25] In combination with the validated results from the ProteinPilot search, we confidently identified and validated 318 unique nitrated peptides using the rPTMDetermine methodology—representing, to the best of our knowledge, the largest number of endogenously identified nitrated peptides. A substantial portion of the retrieved nitrated peptides corresponded to 131 novel PTN sites (56%) and were mapped to a combined set of 98 nitrated proteins, which were missed through conventional database searching (**Table S4**). We selected 25 *novel* endogenous 3-nitrotyrosine–containing peptides from proteins known to be related to ischemic stroke for synthesis (**Table S5**) and analysis through MS/MS experiments under similar conditions (**Supporting Note XI**). The structures of these retrieved peptides were confirmed by comparing their MS/MS spectra with those of synthetic counterparts (**Figure S7**), further demonstrating the confidence of rPTMDetermine retrieval and the importance of recovering missed identifications for comprehensive biological discoveries.

Validation of Tryptophan Oxidized Peptide Identifications

We further established the applicability of the aforementioned methodology to study oxidized tryptophan-containing peptides as an example of the extensibility of rPTMDetermine. Oxidation is important in a range of pathological conditions, including neurodegenerative diseases.[34] Confident identification of oxidized peptides remains challenging, however, because oxidation may also target other residues (commonly phenylalanine, tyrosine, proline, and methionine);[16,35] these residues were considered as possible alternative oxidation sites during localization and the site probability criterion, nonetheless representing significant localization in all permutations. As an example, we explored this modification at tryptophan (**Scheme S2**) to demonstrate the extensibility of our approach. Oxidized tryptophan-containing peptide identifications were collected from ProteinPilot search results (at 1% global FDR); of these, 88 of 102 unique peptides were accompanied by their non-oxidized analogues. These identifications were validated, including deamidation correction, and localized using rPTMDetermine (**Supporting Note XII**; **Figures S8** and **S9**). We successfully validated 71 unique oxidized peptides (with LDA and similarity thresholds of 3.38 and 0.42, respectively; **Figure 6a**). Site localization lowered this number to 67 peptides—a consequence of product ion spectra lacking site-specific fragments to pinpoint the site of oxidation to tryptophan (**Figure 6c**). Applying the retrieval methodology for peptide nitration, we recovered a total of 126 oxidized tryptophan–containing peptides (**Figure 6b**). In total, we confidently identified and localized 164 unique oxidized tryptophan–containing peptides (**Figures 6d** and **6e**; **Table S6**). This example provides powerful, albeit anecdotal, evidence that the rPTMDetermine methodology is applicable to modifications other than nitration. Key details for consideration in any application are the anticipated shifts in the values of m/z in the MS/MS spectra between the modified and unmodified peptides and the site localization criteria. This present study provides a heuristic solution for the examination and site-determination of rare PTMs—a significant innovation given the biological relevance of many rare PTMs.

CONCLUSIONS

We have developed a fully automated procedure for validation of rarely occurring PTM peptide identifications, acquired by searching tandem mass spectra against a protein sequence database. We have demonstrated the utility of this multifaceted integrated strategy involving a semi-supervised machine learning approach with an LDA algorithm, an orthogonal similarity scoring assessment with comparisons of the mass spectra assigned to modified peptides with those assigned to their unmodified counterparts, and site-localization for the validation and retrieval of rarely occurring endogenous 3-nitrotyrosyl– and oxidized tryptophanyl–containing peptides from the cerebral cortex proteome of a *M. fascicularis* model of stroke. The power of such orthogonal criterion to filter false identifications and retrieve modified peptide identifications were hallmarked by the recovery of a number of potential yet-to-be-identified modified peptides from experimental raw data, but missed through conventional database searching. By preparing synthetic nitrated peptide analogues, we also confirmed 25 of the 131 newly retrieved novel endogenous 3-nitrotyrosyl–containing peptides with potential biological relevance in the post-stroke chronic phase. This study strongly suggests that the fully automated rPTMDetermine methodology enables confident identification and recovery of peptides with rarely occurring PTMs in untargeted proteomic analysis, enhancing the confidence of future biological discoveries.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website.

Supporting Notes, Tables, and Figures (pdf)
Table S4. List of validated and retrieved nitrated peptides (MS Excel)
Table S5. List of synthesized nitrated peptides (MS Excel)
Table S6. List of validated and retrieved tryptophan oxidation containing peptides (MS Excel)

AUTHOR INFORMATION

Corresponding Author

* Ivan K. Chu, E-mail: ivankchu@hku.hk.

Author Contributions

I.K.C., K.W.M.S., and N.P.D. conceived the project; I.K.C., and N.P.D. designed the experiments; Y.C.Y.L., Q.Q., and M.Z.L. performed the experiments; N.P.D., D.M.S., and J.W.F. performed the bioinformatics analyses. N.P.D., D.M.S., and I.K.C. wrote the manuscript.

ACKNOWLEDGMENT

We thank the Hong Kong Research Grants Council (Project Nos. HKU 17318616, 17305117, and 17304919) and the University of Hong Kong for financial support. We thank Dr. Leong Ting Lui for helpful discussions.

REFERENCES

(1) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-Based Protein Identification by Searching Sequence Databases Using Mass Spectrometry Data. *Electrophoresis* **1999**, *20* (18), 3551–3567. https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.

(2) Eng, J. K.; McCormack, A. L.; Yates, J. R. I. An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976–989. https://doi.org/10.1016/1044-0305(94)80016-2.

(3) Shilov, I. V.; Seymour, S. L.; Patel, A. A.; Loboda, A.; Tang, W. H.; Keating, S. P.; Hunter, C. L.; Nuwaysir, L. M.; Schaeffer, D. A. The Paragon Algorithm, a Next Generation Search Engine That Uses Sequence Temperature Values and Feature Probabilities to Identify Peptides from Tandem Mass Spectra. *Mol. Cell. Proteomics* **2007**, *6* (9), 1638–1655. https://doi.org/10.1074/mcp.T600050-MCP200.

(4) Bogdanow, B.; Zauber, H.; Selbach, M. Systematic Errors in Peptide and Protein Identification and Quantification by Modified Peptides. *Mol. Cell. Proteomics* **2016**, *15* (8), 2791–2801. https://doi.org/10.1074/mcp.M115.055103.

(5) Fu, Y.; Qian, X. Transferred Subgroup False Discovery Rate for Rare Post-Translational Modifications Detected by Mass Spectrometry. *Mol. Cell. Proteomics* **2014**, *13* (5), 1359–1368. https://doi.org/10.1074/mcp.O113.030189.

(6) Hart-Smith, G.; Yagoub, D.; Tay, A. P.; Pickford, R.; Wilkins, M. R. Large Scale Mass Spectrometry-Based Identifications of Enzyme-Mediated Protein Methylation Are Subject to High False Discovery Rates. *Mol. Cell. Proteomics* **2016**, *15* (3), 989–1006. https://doi.org/10.1074/mcp.M115.055384.

(7) Horlacher, O.; Lisacek, F.; Müller, M. Mining Large Scale Tandem Mass Spectrometry Data for Protein Modifications Using Spectral Libraries. *J. Proteome Res.* **2016**, *15* (3), 721–731. https://doi.org/10.1021/acs.jproteome.5b00877.

(8) Elias, J. E.; Gygi, S. P. Target-Decoy Search Strategy for Increased Confidence in Large-Scale Protein Identifications by Mass Spectrometry. *Nat. Methods* **2007**, *4* (3), 207–214. https://doi.org/10.1038/nmeth1019.

(9) Nesvizhskii, A. I. A Survey of Computational Methods and Error Rate Estimation Procedures for Peptide and Protein Identification in Shotgun Proteomics. *J. Proteomics* **2010**, *73* (11), 2092–2123. https://doi.org/10.1016/j.jprot.2010.08.009.

(10) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry-Based Proteomics. *Nat. Methods* **2017**, *14* (5), 513–520. https://doi.org/10.1038/nmeth.4256.

(11) Granholm, V.; Käll, L. Quality Assessments of Peptide-Spectrum Matches in Shotgun Proteomics. *Proteomics* **2011**, *11* (6), 1086–1093. https://doi.org/10.1002/pmic.201000432.

(12) Olsen, J. V; Mann, M. Status of Large-Scale Analysis of Post-Translational Modifications by Mass Spectrometry. *Mol. Cell. Proteomics* **2013**, *12* (12), 3444–3452. https://doi.org/10.1074/mcp.O113.034181.

(13) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets. *Nat. Methods* **2007**, *4* (11), 923–925. https://doi.org/10.1038/nmeth1113.

(14) Choi, H.; Nesvizhskii, A. I. False Discovery Rates and Related Statistical Concepts in Mass Spectrometry-Based Proteomics. *J. Proteome Res.* **2008**, *7* (1), 47–50.

(15) Pathan, M.; Samuel, M.; Keerthikumar, S.; Mathivanan, S. Unassigned MS/MS Spectra: Who Am I? In *Current Protocols in Protein Science*; Keerthikumar, S., Mathivanan, S., Eds.; Humana Press: New York, 2017; Vol. 1549, pp 67–74. https://doi.org/10.1007/978-1-4939-6740-7_6.

(16) Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P. A Mass-Tolerant Database Search Identifies a Large Proportion of Unassigned Spectra in Shotgun

Proteomics as Modified Peptides. *Nat. Biotechnol.* **2015**, *33* (7), 743–749. https://doi.org/10.1038/nbt.3267.

(17) Frank, A. M. A Ranking-Based Scoring Function for Peptide - Spectrum Matches. *J. Proteome Res.* **2009**, *8* (5), 2241–2252. https://doi.org/10.1021/pr800678b.

(18) Griss, J.; Perez-Riverol, Y.; Lewis, S.; Tabb, D. L.; Dianes, J. A.; Del-Toro, N.; Rurik, M.; Walzer, M.; Kohlbacher, O.; Hermjakob, H.; Wang, R.; Vizcano, J. A. Recognizing Millions of Consistently Unidentified Spectra across Hundreds of Shotgun Proteomics Datasets. *Nat. Methods* **2016**, *13* (8), 651–656. https://doi.org/10.1038/nmeth.3902.

(19) Na, S.; Paek, E. Prediction of Novel Modifications by Unrestrictive Search of Tandem Mass Spectra. *J. Proteome Res.* **2009**, *8* (10), 4418–4427. https://doi.org/10.1021/pr9001146.

(20) Ahrné, E.; Müller, M.; Lisacek, F. Unrestricted Identification of Modified Proteins Using MS/MS. *Proteomics* **2010**, *10* (4), 671–686. https://doi.org/10.1002/pmic.200900502.

(21) Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A. ModifiComb, a New Proteomic Tool for Mapping Substoichiometric Post-Translational Modifications, Finding Novel Types of Modifications, and Fingerprinting Complex Protein Mixtures. *Mol. Cell. Proteomics* **2006**, *5* (5), 935–948. https://doi.org/10.1074/mcp.t500034-mcp200.

(22) Xiu, L.-Y.; Ye, D.; Jia, W.; Qian, X.-H.; He, S.-M.; Sun, R.-X.; Fu, Y. DeltAMT: A Statistical Algorithm for Fast Detection of Protein Modifications From LC-MS/MS Data. *Mol. Cell. Proteomics* **2011**, *10* (5), M110.000455. https://doi.org/10.1074/mcp.m110.000455.

(23) Radi, R. Nitric Oxide, Oxidants, and Protein Tyrosine Nitration. *Proc. Natl. Acad. Sci. USA* **2004**, *101* (12), 4003–4008. https://doi.org/10.1073/pnas.0307446101.

(24) Lai, C. K.; Tang, W. K.; Siu, C. K.; Chu, I. K. Evidence for the Prerequisite Formation of Phenoxyl Radicals in Radical-Mediated Peptide Tyrosine Nitration In Vacuo. *Chem. Eur. J.* **2020**, *26* (1), 331–335. https://doi.org/10.1002/chem.201904484.

(25) Quan, Q.; Szeto, S. S. W.; Law, H. C. H.; Zhang, Z.; Wang, Y.; Chu, I. K. Fully Automated Multidimensional Reversed-Phase Liquid Chromatography with Tandem Anion/Cation Exchange Columns for Simultaneous Global Endogenous Tyrosine Nitration Detection, Integral Membrane Protein Characterization, and Quantitative Proteomics Mapping. *Anal. Chem.* **2015**, *87* (19), 10015–10024. https://doi.org/10.1021/acs.analchem.5b02619.

(26) Law, H. C. H.; Szeto, S. S. W.; Quan, Q.; Zhao, Y.; Zhang, Z.; Krakovska, O.; Lui, L. T.; Zheng, C.; Lee, S. M. Y.; Siu, K. W. M.; Wang, Y.; Chu, I. K. Characterization of the Molecular Mechanisms Underlying the Chronic Phase of Stroke in a Cynomolgus Monkey Model of Induced Cerebral Ischemia. *J. Proteome Res.* **2017**, *16* (3), 1150–1166. https://doi.org/10.1021/acs.jproteome.6b00651.

(27) Ulintz, P. J.; Zhu, J.; Qin, Z. S.; Andrews, P. C. Improved Classification of Mass Spectrometry Database Search Results Using Newer Machine Learning Approaches. *Mol. Cell. Proteomics* **2005**, *5* (3), 497–509. https://doi.org/10.1074/mcp.m500233-mcp200.

(28) Edwards, N.; Wu, X.; Tseng, C. W. An Unsupervised, Model-Free, Machine-Learning Combiner for Peptide Identifications from Tandem Mass Spectra. *Clin. Proteomics* **2009**, *5* (1), 23–36. https://doi.org/10.1007/s12014-009-9024-5.

(29) Bishop, C. M. Linear Models for Classification. In *Pattern Recognition and Machine Learning*; Springer-Verlag: New York, 2006; pp 137–173.

(30) Efron, B.; Tibshirani, R.; Storey, J. D.; Tusher, V. Empirical Bayes Analysis of a Microarray Experiment. *J. Am. Stat. Assoc.* **2001**, *96* (456), 1151–1160. https://doi.org/10.1198/016214501753382129.

(31) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. *J. Proteome Res.* **2008**, *7* (1), 40–44. https://doi.org/10.1021/pr700739d.

(32) Ning, K.; Fermin, D.; Nesvizhskii, A. I. Computational Analysis of Unassigned High-Quality MS/MS Spectra in Proteomic Data Sets. *Proteomics* **2010**, *10* (14), 2712–2718. https://doi.org/10.1002/pmic.200900473.

(33) Wetie, A. G. N.; Shipp, D. A.; Darie, C. C. Bottlenecks in Proteomics. In *Advancements of Mass Spectrometry in Biomedical Research*; Woods, A. G., Darie, C. C., Eds.; Springer: Heidelberg, 2014; pp 581–593.
(34) Grune, T.; Catalgol, B.; Jung, T. Protein Oxidation in Some Age-Related Diseases. In *Protein Oxidation and Aging*; John Wiley & Sons: New Jersey, 2012; pp 417–478. https://doi.org/10.1002/9781118493038.
(35) Spickett, C. M.; Pitt, A. R. Protein Oxidation: Role in Signalling and Detection by Mass Spectrometry. *Amino Acids* **2012**, *42* (1), 5–21. https://doi.org/10.1007/s00726-010-0585-4.

FIGURE LEGENDS

Figure 1. Workflow for PTM peptide validation and localization. Normal modified PSMs are generated by searching experimental mass spectra against the target protein sequence database. Decoy peptides are generated *in silico* from reversed target protein database sequences and decoy modified PSMs are collected by searching the mass spectra assigned to normal PTM peptides against the constructed decoy database. **(a)** LDA classifies the normal and decoy sets using discriminative PSM features (**Figure S1**), producing LDA scores. For PSMs with a probability of being correct (from the distribution of LDA scores) of ≥ 0.99, the corresponding non-modified peptide analogues are extracted from the database search results and validated using LDA. **(b)** If the non-modified PSM meets the same probability requirement, a similarity score compares the spectra of the modified and unmodified peptide pair, where fragment ions including the site of modification will demonstrate a shift in m/z (Δ Da) according to the modification mass. If a peptide has no alternative modification sites and meets the LDA and benchmark similarity score requirements, then the identification is confirmed. If modified peptide has alternative possible modification sites, then all modified isoforms are generated by permutation. **(c)** The LDA score is calculated for each isoform and used to infer the probability, p_{site}, of the given site bearing the modification. If an isoform has site probability exceeding a threshold probability, p_t, the modified peptide identification is confirmed and localized; otherwise, it is considered to be ambiguous.

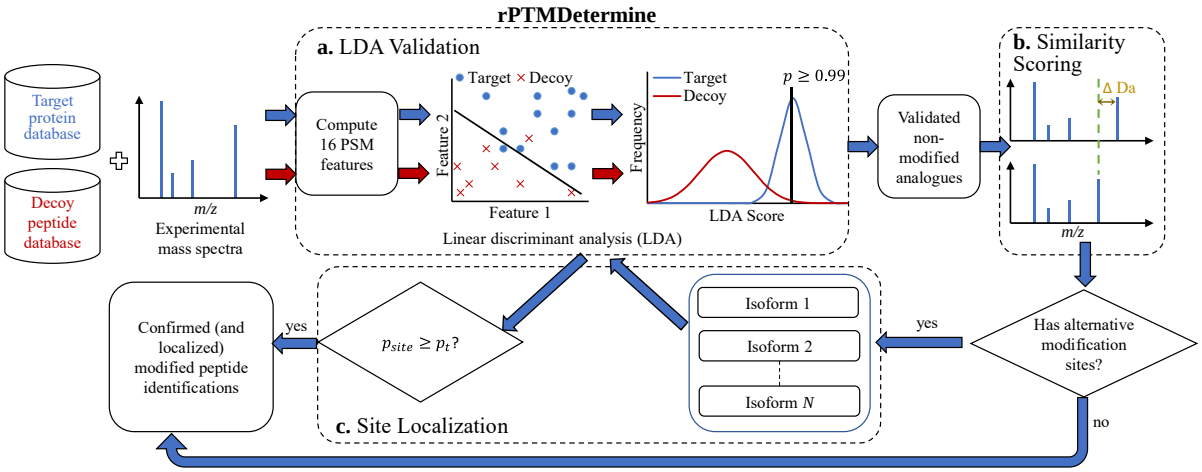
Figure 2. Validation results for PTN. **(a)** Scatter plot of unique nitrated peptide LDA and similarity scores for the *M. fascicularis* cerebral cortex data set, with each peptide represented by the highest-scoring PSM. Blue circles, peptides (Benchmark) previously confirmed by synthesis; red triangles, identifications to deamidated peptides. Nitrated peptide identifications without non-nitrated counterparts are assigned similarity scores of 0. Thresholds are indicated in the plots by dashed lines. **(b)** Demonstration of highly confident nitrated peptide identification: [IT8]EAVC*EVALDY^{NO2}K^{3+}, where "IT8" indicates iTRAQ 8-plex, "NO2" indicates nitration, and * indicates carbamidomethylation. CID mass spectrum of the nitrated peptide (upper panel) and its non-nitrated counterpart (lower panel). Single letter annotations indicate immonium ions. Purple annotations indicate ions containing the modified residue. **Inset:** Extracted ion chromatograms (XICs) for the nitrated peptide and its non-nitrated counterpart from the same subfraction.

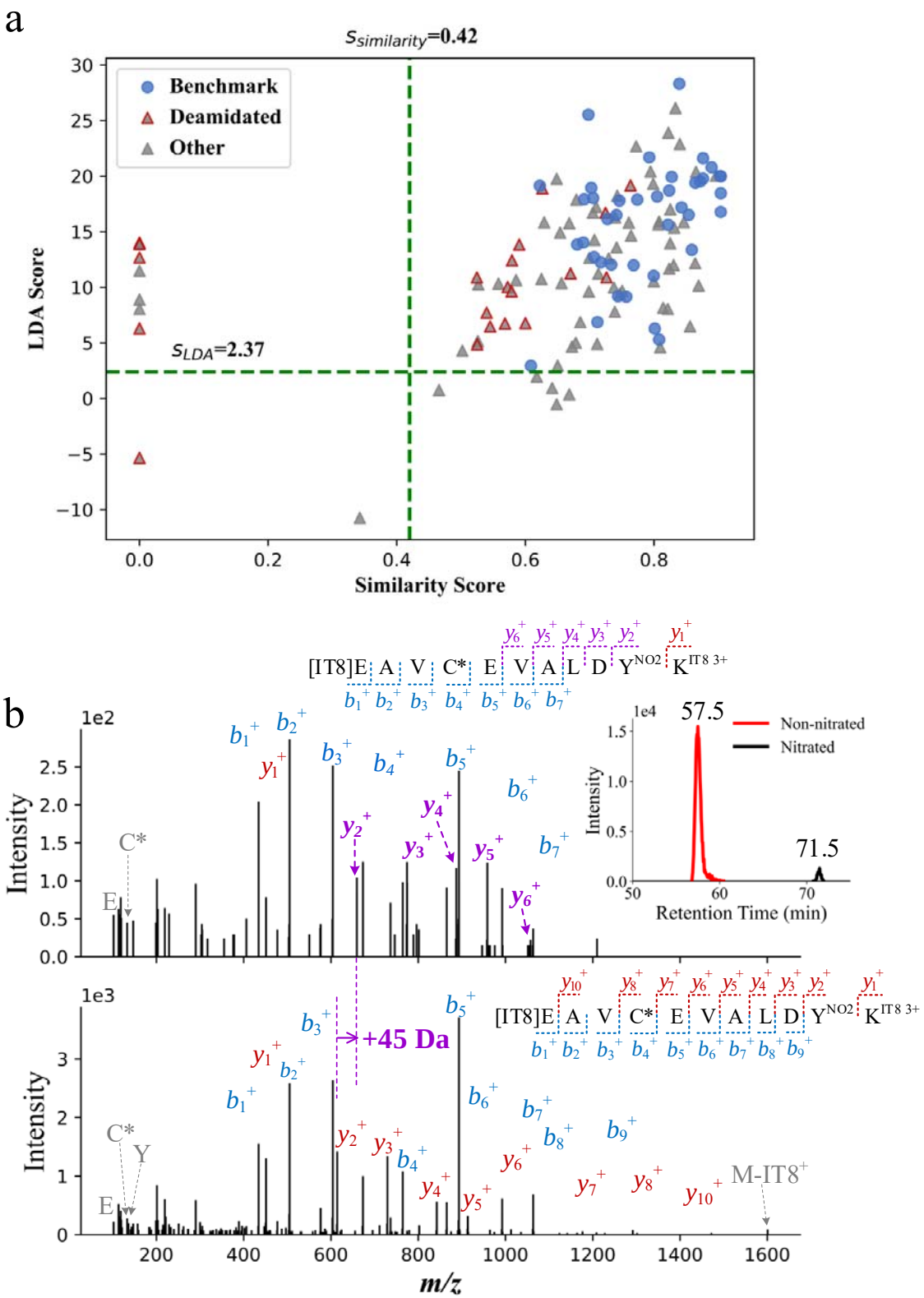
Figure 3. Example of an ambiguous deamidation assignment for the peptide [IT8]FLFPFFGSAY^{NO2}Q-^{N}GFASGNLER^{3+}, where "[IT8]," "-N," and "NO2" indicate iTRAQ 8-plex, deamidation, and nitration, respectively. The circled ions are exclusively assigned for the non-deamidated form of the peptide, notably including the only PTN-determining ions (purple). "F" represents the immonium ion of phenylalanine. The LDA score and similarity score before correction were 12.39 and 0.00, respectively; they were 22.65 and 0.83, respectively, following correction. Inset: Observed isotopic distribution in the MS spectrum from which this precursor (m/z 888.43) was selected. Overlaid are the theoretical isotopic distributions for the deamidated (upper panel) and non-deamidated (lower panel) forms of the assigned peptide, as predicted by the MS-Isotope toolkit of ProteinProspector (http://prospector.ucsf.edu/).

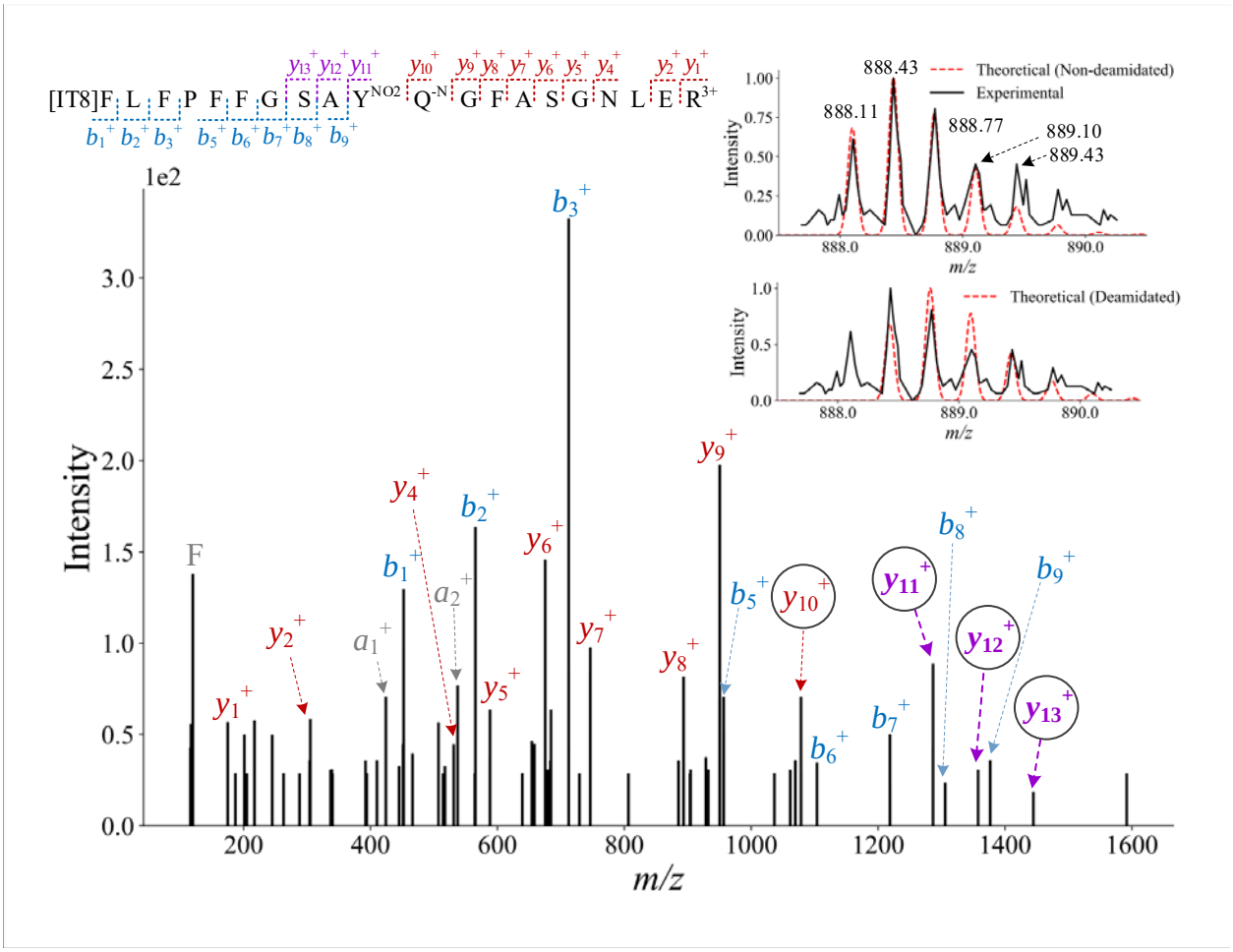
Figure 4. Results for retrieval of missed nitrated peptide identifications. (**a**) Distribution of LDA and similarity scores of unique nitrated peptide identifications not identified in ProteinPilot search. (**b**) Site probability of nitrated peptide identifications containing alternative tyrosine sites for nitration. Green dashed lines indicate score thresholds. **(c)** Venn diagram summarizing the distribution of retrieved nitrated peptide identifications, including the overlap with the ProteinPilot identifications validated by rPTMDetermine.

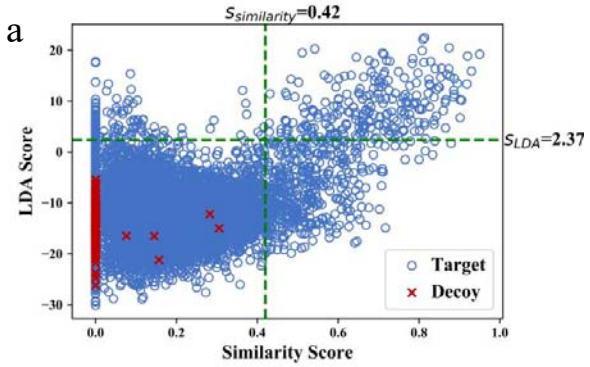
Figure 5. Retrieved nitrated peptide identification ($VGGY^{NO2}ILGEFGNLIAGDPR^{3+}$, where "IT8" indicates iTRAQ 8-plex and "NO2" indicates nitration). Golden annotations indicate those ions containing the modified residue. **(a)** CID mass spectrum of the nitrated peptide (upper panel) and its non-nitrated counterpart (lower panel). Single character annotations indicate immonium ions. The LDA score and similarity score for the nitrated peptide were 22.15 and 0.81, respectively. **(b)** Experimental and theoretical (MS-Isotope toolkit, ProteinProspector) isotopic distributions of the nitrated peptide. **(c)** XICs for the nitrated peptide and its non-nitrated counterpart. **(d)** Mass spectra obtained experimentally (upper panel) and for the synthesized nitrated peptide (lower panel).

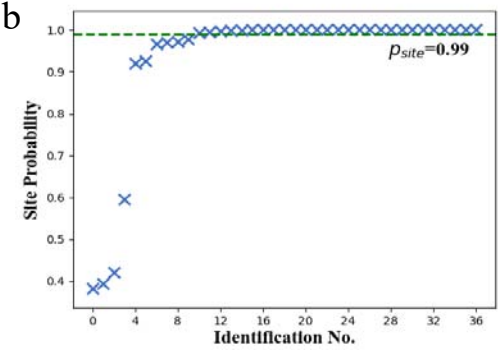
Figure 6. Results for validation and retrieval of missed oxidized tryptophan–containing peptide identifications. **(a)** Distribution of LDA and similarity scores for the unique tryptophan-oxidized peptides, with each represented by its highest scoring PSM. **(b)** LDA and similarity scores for the retrieval of missed tryptophan-oxidized peptide identifications during ProteinPilot search. **(c)** Site probabilities of modified peptides; score thresholds are represented by green dashed lines. **(d)** Venn diagram summarizing the unique tryptophan-oxidized peptides, including the peptides validated by rPTMDetermine (from ProteinPilot search results at 1% global FDR) and retrieved from the unassigned product ion spectra. **(e)** Summary of tryptophan oxidation validation and retrieval results; tryptophan-oxidized peptide identifications without non-oxidized counterparts have been assigned a similarity score of 0.

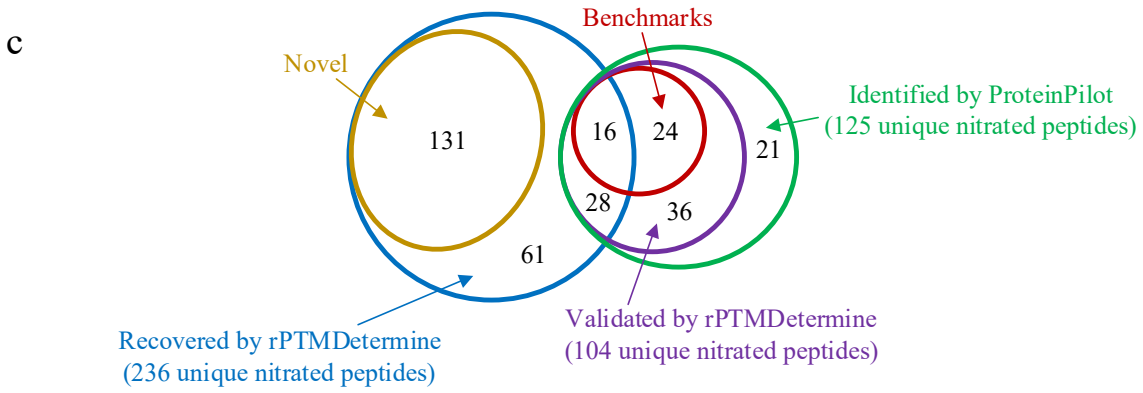

rPTMDetermine
a. LDA Validation
Target protein database
Decoy peptide database
m/z
Experimental mass spectra
Compute 16 PSM features
Target
Decoy
Feature 2
Feature 1
Target
Decoy
p ≥ 0.99
Frequency
LDA Score
Linear discriminant analysis (LDA)
Validated non-modified analogues
b. Similarity Scoring
Δ Da
m/z
Has alternative modification sites?
yes
Isoform 1
Isoform 2
Isoform N
no
c. Site Localization
p_site ≥ p_t?
yes
Confirmed (and localized) modified peptide identifications

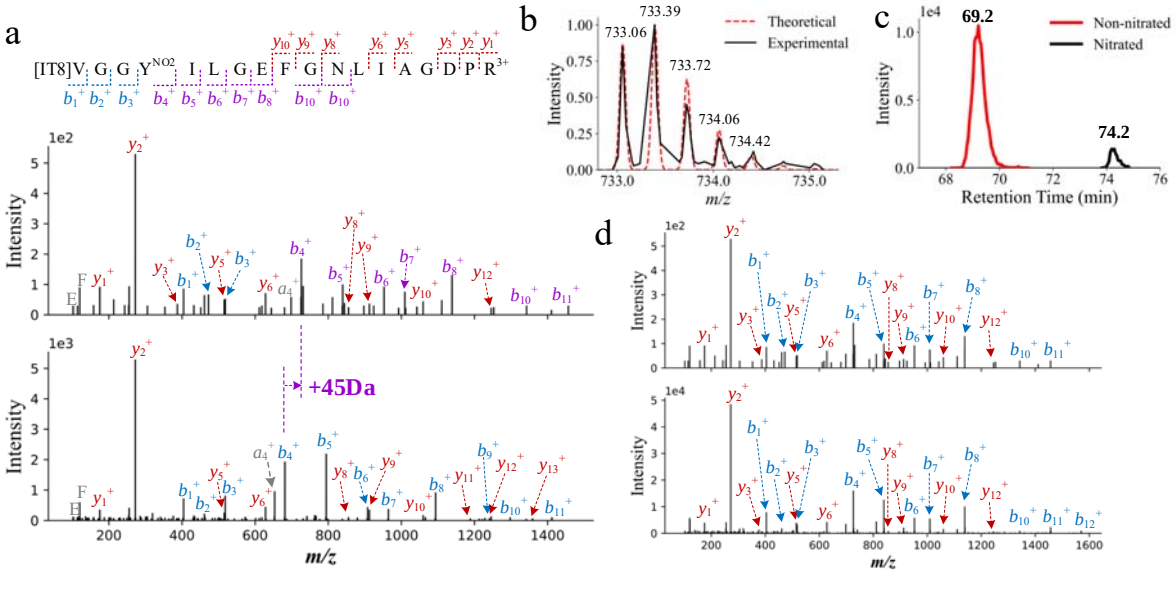

a
$s_{similarity}$=0.42
Benchmark
Deamidated
Other
LDA Score
s_{LDA}=2.37
Similarity Score
b
[IT8]E A V C* E V A L D Y^{NO2} K$^{IT8\ 3+}$
1e2
Intensity
C*
E
1e4
57.5
71.5
Non-nitrated
Nitrated
Intensity
Retention Time (min)
1e3
+45 Da
Y
M-IT8^{+}
m/z

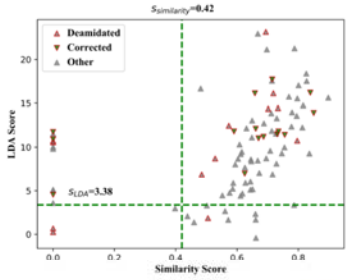

[IT8]F L F P F F G S A Y^NO2 Q^-N G F A S G N L E R^3+
1e2
Intensity
m/z
888.43
Theoretical (Non-deamidated)
Experimental
888.11
888.77
889.10
889.43
Theoretical (Deamidated)

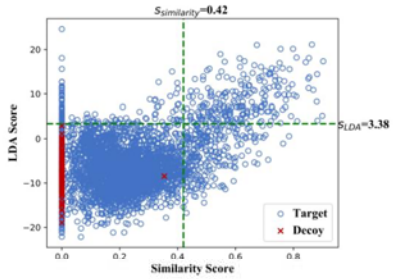

a
$S_{similarity}$=0.42
S_{LDA}=2.37
LDA Score
Similarity Score
Target
Decoy

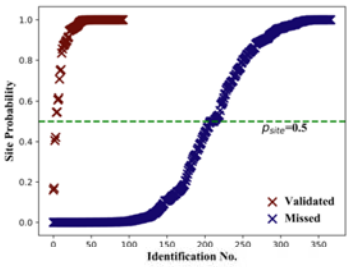

b
p_{site}=0.99
Site Probability
Identification No.


c
Benchmarks
Novel
131
16
24
21
28
36
61
Identified by ProteinPilot
(125 unique nitrated peptides)
Recovered by rPTMDetermine
(236 unique nitrated peptides)
Validated by rPTMDetermine
(104 unique nitrated peptides)


a
[IT8]V G G Y^{NO2} I L G E F G N L I A G D P R^{3+}
+45Da
Intensity
m/z
b
733.39
733.06
733.72
734.06
734.42
Theoretical
Experimental
c
69.2
74.2
Non-nitrated
Nitrated
Retention Time (min)
d


$s_{similarity}$=0.42
Deamidated
Corrected
Other
LDA Score
s_{LDA}=3.38
Similarity Score


$s_{similarity}$=0.42
LDA Score
s_{LDA}=3.38
Target
Decoy
Similarity Score


Site Probability
p_{site}=0.5
Validated
Missed
Identification No.