Journal of proteome research

pubs.acs.org/jpr

Article

Conditional Fragment Ion Probabilities Improve Database Searching for Nonmonoisotopic Precursors

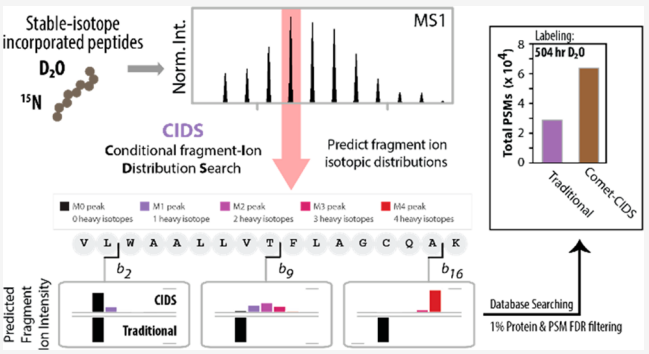
Jonathon J. O'Brien,*[,§] Meagan Gadzuk-Shea,[§] Phillip M. Seitzer, Ramin Rad, Fiona E. McAllister, and Devin K. Schweppe*

Cite This: *J. Proteome Res.* 2023, 22, 334−342

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Stochastic, intensity-based precursor isolation can result in isotopically enriched fragment ions. This problem is exacerbated for large peptides and stable isotope labeling experiments using deuterium or ^{15}N. For stable isotope labeling experiments, incomplete and ubiquitous labeling strategies result in the isolation of peptide ions composed of many distinct structural isomers. Unfortunately, existing proteomics search algorithms do not account for this variability in isotopic incorporation, and thus often yield poor peptide and protein identification rates. We sought to resolve this shortcoming by deriving the expected isotopic distributions of each fragment ion and incorporating them into the theoretical mass spectra used for peptide-spectrum-matching. We adapted the Comet search platform to integrate a modified spectral prediction algorithm we term Conditional fragment Ion Distribution Search (CIDS). Comet-CIDS uses a traditional database searching strategy, but for each candidate peptide we compute the isotopic distribution of each fragment to better match the observed m/z distributions. Evaluating previously generated D_2O and ^{15}N labeled data sets, we found that Comet-CIDS identified more confident peptide spectral matches and higher protein sequence coverage compared to traditional theoretical spectra generation, with the magnitude of improvement largely determined by the amount of labeling in the sample.



KEYWORDS: *peptide spectrum matching, protein turnover, isotopic envelope ^{15}N, D_2O, stable isotope labeling, database searching*

INTRODUCTION

Standard proteomics identification algorithms and modern machine learning models have been designed and optimized to identify monoisotopic peptide precursors.[1,2] When a peptide is isolated with an unknown number of stable isotope-incorporated amino acids, two problems occur. First, the isolated precursor mass may not match the masses created during in-silico digestion of the protein. This first challenge has been addressed previously through work to estimate monoisotopic peaks or the inclusion of multiple precursor windows during search.[3] Second, the MS2 fragmentation spectra present shifts in ion masses consistent with the number of heavy atoms contained in each fragment. Adjusting for these shifts is trivial when labeling occurs consistently and completely at a known amino acid residue. However, when the isolated precursor peak contains a population of structural isomers—for example, isotopomers with incomplete ^{13}C or deuterium labeling at a subset of amino acids—a distribution of fragment ions at known mass shifts will be present that were previously difficult to predict. As an example, in a D_2O pulse-chase experiment fragment ions from deuterated peptides often have isotopic distributions spread across amino acids and the location of the largest peak depends on both the amount of protein turnover that has occurred and the number of heavy atoms contained in the isolated precursor.[4] Consequently, protein turnover experiments based on ubiquitous labeling strategies, such as using heavy water (D_2O) or ^{15}N, inevitably result in vanishing numbers of successful peptide identifications as the amount of stable isotope incorporation increases. It is plausible that the challenges presented by isolating populations of structural isomers could be resolved by modifying the theoretical spectra to include isotopic envelope estimations for each fragment. While some research has been done to predict fragment ion isotopic distributions based on mass and other molecular attributes,[5] this work was not considered in the context of the peptide-spectrum-matching problem.

Here, we aimed to explore the probabilistic effects of mass isolation on isotopic distributions and determine whether

Special Issue: Software Tools and Resources 2023

Received: April 26, 2022
Published: November 22, 2022

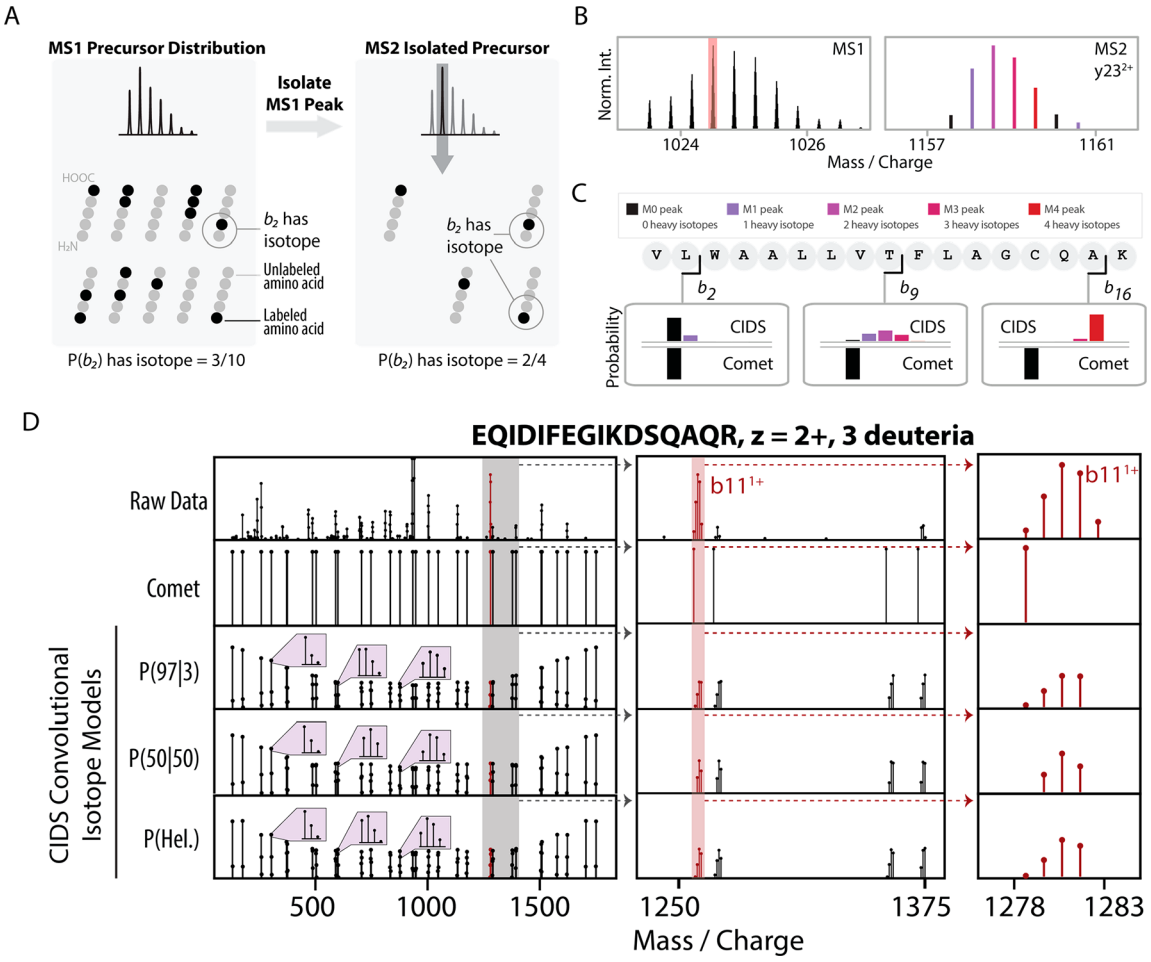


Figure 1. Theoretical background for Comet-CIDS. (A) Isolation of specific precursor peaks alters the isotopic distributions fragment ions. (B) Example precursor isotope distribution for peptide AAELGAELGAQAISHLEEVSDEGIAAMAAAR and fragment $y23^{2+}$ showing isotopic distributions after isolation and peptide fragmentation. (C) Convolutional isotopic model estimation of theoretical fragment peaks based on Comet-CIDS and Comet. The isotopic fragment peaks are colored based on estimated incorporation of stable isotopes. (D) Comet-CIDS predicted spectra for peptide EQIDIFEGIKDSQAQR using a range of labeling estimations across theoretical spectra. Each fragment is depicted with a capped line (line and point). Comet spectra have a single peak per fragment. Small insets are shown for several fragments in the full spectra (purple boxes). Insets show fragment isotope distribution estimates for the range of 1250−1375 m/z and the $b11^{1+}$ fragment alone (red). While the same total number of fragments are used, the number of peaks considered increases when using CIDS for each of the fragment isotopic distributions. For the convolutional isotope modeling of CIDS, the indicated amino acid isotopic distribution estimates were used (see Experimental Procedures). P(Hel.) refers to empirical distributions computed based on work from the Hellerstein group.[12]

knowing the entire state of isotopic labeling in a biological system a priori could enable prediction of fragment ion distributions. Because mass isolation for peptide fragmentation selects the most intense peak from a precursor distribution, generally with a single number of heavy atoms (Figure 1), the distribution of heavy atoms on each fragment ion will be altered by the laws of conditional probability. For a protein turnover experiment, this would require knowing a priori the amount of labeling in each pool of free-floating amino acids as well as the total amount of turnover that had occurred at the time of sample collection. This information is unlikely to be known prior to an experiment being performed. However, exploring a large number of plausible configurations suggests that fragment ion distributions are primarily determined by the number of heavy atoms in the precursor, with minimal impact from the underlying labeling rates. Accordingly, even crude approximations of the underlying state of amino acid labeling could lead to substantial gains in peptide spectral matching.

To test this, we introduce a framework for identifying populations of structural isomers and derive a new search modality for identifying peptides, which we term "Conditional fragment-Ion Distribution Search" (CIDS, pronounced "kids"). The primary aim of CIDS is to improve peptide identification by accounting for the deviations that occur when data-dependent acquisition selects non-monoisotopic precursor peaks. To this end, we derive the theoretical isotope-enriched distributions for *b*- and *y*-fragment ions and integrate them into the theoretical spectra used in the open-source search algorithm Comet.[6] We evaluate our new software (Comet-CIDS) on previously published stable isotope incorporation experiments based on both D_2O and ^{15}N labeling strategies. Comparisons against established search algorithms, Mascot and Comet highlight the benefits of incorporating fragment ion distributions during search and how CIDS models affect peptide-spectral matching, peptide scoring, and protein identifications.

https://doi.org/10.1021/acs.jproteome.2c00247
J. Proteome Res. 2023, 22, 334−342

EXPERIMENTAL PROCEDURES

Source Data and Data Analysis for CIDS Testing

Previously collected source data was obtained from Sadygov et al.[4] Raw data files were downloaded from ProteomeXchange via identifier PXD009493.[4] Briefly, protein samples were collected from male mice ($LDLR^{-/-}$) at 8 weeks of age. These mice were given a 20 μL bolus injection of D_2O followed by ad libitum access to 5% D_2O drinking water for different numbers of days (0−21 days, or 0−504 h). Murine liver proteins were fractionated by SDS-PAGE into 9 gel bands (only 8 available at the 0 h time point). From this experiment we selected RAW files from the first replicate of healthy mice at the beginning middle and end of the experiment (0, 168, and 504 h).

The second data set we analyzed came from a ^{15}N labeling experiment designed to explore protein turnover rates in murine eyes.[7] Raw files were downloaded from ProteomeXchange via identified PXD016212. Since the eye is believed to have a highly stable proteome, the authors collected only a single sample 12 weeks after replacing the food source with ^{15}N spirulina. Multiple RAW files were provided by the authors corresponding to the region of the eye collected and we tested our search algorithm on the single file corresponding to the lens-cortex.

Raw File Conversion and Database Search

File conversion to mzXML format and monoisotopic peak estimation was performed using Monocle.[6] Spectra were then searched using either Comet (stable release Comet 2018.01 rev. 5) or Comet-CIDS based on the same revision.[6] Mascot search results were derived from the original publication's peptide spectral match output, along with their FDR estimates. Comet-CIDS was run using various amino acid isotopic distributions, as discussed throughout the manuscript. For ease of use, running Comet-CIDS requires only small adjustments to the common Comet parameters file (Table S1 of the Supporting Information, SI). Comet search parameters were otherwise kept at defaults for the stable release using ion trap tolerances (Peptide Mass Tolerance = 20 ppm, Fragment Ion Tolerance = 1.0005, Fragment Bin Offset = 0.4, Theoretical Fragment Ions = 1, Allowed Missed Cleavages = 2, Max Variable Mods in peptide = 5), including variable modification of methionines (oxidation −15.9949146221) and static modification of cysteines (carboxyamidomethylation −57.02146374). Spectra were searched against a forward−reverse Uniprot mouse database (downloaded: 05/2017). For searches termed "high-resolution" all parameters were kept the same with the exception of: Fragment Ion Tolerance = 0.02, Fragment Bin Offset = 0.0, Theoretical Fragment Ions = 0. Comet's "mass_offset" parameter was used for all deuterium Comet-CIDS, and otherwise noted, searches (mass_offsets = 0 1.006262, 2.012524, 3.018785, 4.025047, 5.031309, 6.037571, 7.043832, 8.050094, 9.056356, 10.062618). Data for ^{15}N were searched with the following mass offsets: mass_offsets = 0.0000000, 0.9970349, 1.9940698, 2.9911047, 3.9881396, 4.9851745, 5.9822094, 6.9792443, 7.9762792, 8.9733141, 9.9703490, 10.9673839, 11.9644188, 12.9614537, 13.9584886, 14.9555235, 15.9525584, 16.9495933, 17.9466282, 18.9436631, 19.9406980. Note that, unless otherwise specified, we compare Comet-CIDS to Comet where Comet has no additional offset masses as this is currently standard practice in the field.

For all search results, PSMs were filtered to a 1% peptide and protein false-discovery rate using linear discriminant analysis and assembled by parsimony.[3,8] Mascot search results were taken from the original data set results files and filtered based on those reported FDRs. Comet CIDS is freely available https://github.com/SchweppeLab/CometCIDS. For later analyses, peptides and proteins were considered deuterated when after monoisotopic peak correction with Monocle,[3] the best scoring PSM had a delta mass greater than 1 Da.

Conditional Probability Calculations

The fragment ion peak masses become difficult to predict when isolating a precursor that contains a heavy atom at an unknown position. For example, if we isolate ions containing a single 13C isotope, that isotope could have been present in any of the amino acids within the peptide. The whole peptide must contain exactly one heavy amino acid, but the location of the heavy amino acid will vary across the population of ions. Consequently, each *b*- and *y*-fragment ion in the MS2 spectra will present a set of peaks proportional to the distribution of isotopes for each ion. We now show that these distributions can be derived exactly with a small number of assumptions.

Let H be a random variable representing the number of heavy isotopes contained in an isolated precursor peptide, and let h be an observed instance of this variable. For each candidate peptide, we know h, since only one possibility will be within the mass tolerance of the instrument. We let θ represent the percentage of peptides that, at the time of sampling, had been synthesized after label administration. Marginal isotopic distributions of *b*- and *y*-ions (the proportions we would expect to see if all peptides were fragmented) can be generated using the concepts described for calculating peptide isotope distributions. For a peptide of length m, let B_i and Y_j represent random variables for the number of heavy isotopes contained in randomly sampled b_i and y_j ions respectively ($i = 1, \cdots, m$ and $j = 1, \cdots, m$). Further, let P^{B_i}, P^{Y_j}, $Q_t^{B_i}$ and $Q_t^{Y_j}$ represent a shorthand for the marginal isotopic probability distributions of each b and y ion where, as before, P denotes a distribution prior to labeling and Q_t represents the isotopic distribution of fragments synthesized between $(0, t)$. These are the distributions of the respective populations as they would be seen in a cell, without restricting the total number of heavy isotopes found in the precursor. We present the derivations for the b ions (they are analogous to the y ions). Mass isolation results in the following conditional probability distribution:

$$p(B_i = b|H = h, \theta) = \frac{p(H = h|B_i = b, \theta)p(B_i = b|\theta)}{p(H = h|\theta)} \quad (1)$$

Both $p(B_i = b|\theta)$ and $p(H = h|\theta)$ can be calculated using convolutions of the underlying amino acid labeling probabilities. Note that in these equations "b" does not specify an ion type, rather b and h denote the number of heavy atoms in the b_i ion and the precursor, respectively. In the case of a protein turnover experiment these are not generally known. However, even rough approximations prove to be valuable, as we will demonstrate later.

Calculating $p(H = h \mid B_i = b,\theta)$ requires separating out old and new peptides. If L defines the old, $L = o$, versus new, $L = e$, peptides in the population, then partitioning the ions by L gives us

$$p(B_i|H, \theta) = p(B_i|H, L = o)p(L = o) + p(B_i|H, L = e)p(L = e) \quad (2)$$

https://doi.org/10.1021/acs.jproteome.2c00247
J. Proteome Res. 2023, 22, 334−342

where conditioning on L lets us drop θ, since the proportion of "new" peptides is either 0 or 1 within each stratum of L. Thus,

$$p(B_i|H, L = o)(1 - \theta) + p(B_i|H, L = e)\theta$$
$$\frac{p(H|B_i, L = o)p(B_i|L = o)}{p(H|L = o)}(1 - \theta) + \frac{p(H|B_i, L = e)p(B_i|L = e)}{p(H|L = e)}\theta \quad (3)$$

Each of the probabilities in eq 3 can be found using convolutions of natural and observed amino acid sequences.

$$p(H|L = o) = P$$
$$p(H|L = e) = Q_t$$
$$p(B_i|\theta, L = o) = P^{B_i}$$
$$p(B_i|\theta, L = e) = Q_t^{B_i}$$
$$p(H = h|B_i = b, L = o) = P^{Y_{(m-i)}}(h - b)$$
$$p(H = h|B_i = b, L = e) = Q_t^{Y_{(m-i)}}(h - b) \quad (4)$$

The last two expressions follow from the complementary nature of b and y ions and the observation that

$$p(H = h|B_i = b) = p(B_i + Y_{m-i} = h|B_i = b) = p(Y_{m-i} = h - b) \quad (5)$$

A shorthand analytic expression for the conditional distribution of a b ion (as will be observed in an MS2 spectra) is given by

$$p(B_i = b|H = h, \theta) = \frac{P^{Y_{(m-i)}}(h - b)P^{B_i}}{P}(1 - \theta) + \frac{Q_t^{Y_{(m-i)}}(h - b)Q_t^{B_i}}{Q_t}\theta \quad (6)$$

With these equations, we could generate alternative theoretical MS2 spectra. However, the above expression is still dependent on the amount of turnover than has occurred, θ. In order to fully determine the isotopic proportions we would either need to know the turnover of a given protein in advance or we need to allow for a simplifying assumption.

Model Simplifications

A simple solution to the problem of having an unknown amount of turnover in the population of protein molecules, is to treat θ as a uniform random variable and to integrate it out of our distribution. This results in a simple average of the pre- and postlabeling distributions.

$$p(B_i = b|H = h) = \int_0^1 \frac{P^{Y_{(m-i)}}(h - b)P^{B_i}}{P}(1 - \theta) + \frac{Q_t^{Y_{(m-i)}}(h - b)Q_t^{B_i}}{Q_t}\theta d\theta$$
$$= \frac{1}{2}\left(\frac{P^{Y_{(m-i)}}(h - b)P^{B_i}}{P} + \frac{Q_t^{Y_{(m-i)}}(h - b)Q_t^{B_i}}{Q_t}\right). \quad (7)$$

More complex prior distributions for θ might be more realistic, but as we will show, the specific amino acids distributions used work well even with very crude approximations. This model depends on taking many convolutions of the amino acid isotopic distributions. Accordingly, we will refer to the class of models generated from eq 7 as convolutional isotope models, where the specific model is defined in conjunction with a set of amino acid isotopic distributions. From this perspective, it is worth noting that a few further simplifying assumptions allow us to skip the convolutions altogether.

Suppose that every amino acid has an equal isotopic distribution, regardless of when translation occurred, and that each amino acid that can be labeled can only be labeled once. Under these assumptions, the unconditional probability of observing h heavy amino acids in the peptide, $p(H = h)$, is given by the probability mass function of the Binomial distribution, $\text{binomial}\left(h, m, \frac{1}{m}\right)$. Similarly, the unconditional probability that a b-ion contains b heavy labels, $p(B_i = b)$, is given by a $\text{binomial}\left(b, i, \frac{1}{m}\right)$, and

$$p(H = h|B_i = b) = P^{Y_{(m-i)}}(h - b) = \text{binomial}\left(h - b, m - i, \frac{1}{m}\right)$$

The conditional fragment ion distributions reduce to

$$p(B_i = b|H = h) = \frac{\binom{i}{b}\left(\frac{1}{m}\right)^b\left(1 - \frac{1}{m}\right)^{i-b}\binom{m-i}{h-b}\left(\frac{1}{m}\right)^{h-b}\left(1 - \frac{1}{m}\right)^{m-i-(h-b)}}{\binom{m}{h}\left(\frac{1}{m}\right)^h\left(1 - \frac{1}{m}\right)^{m-h}} = \frac{\binom{i}{b}\binom{m-i}{h-b}}{\binom{m}{h}} \quad (8)$$

which is the probability mass function of a hypergeometric distribution. When the assumptions seem remotely plausible, the hypergeometric distribution could therefore be used to skip the computationally intensive amino acid convolutions entirely, as eq 8 depends only on peptide length and the number of heavy atoms in the precursor.

Model Assumptions and Configurations

The convolutional isotope models were derived under the assumption that underlying amino acid isotopic distributions are known, and that mass isolation selects only precursor molecules containing a single fixed number of heavy atoms. Equation 8 further requires that every amino acid has the same probability of labeling and that only a single heavy atom will be found on each residue. None of these assumptions are likely to be true in practice.

It is plausible that metabolomics could be used to measure the isotopic envelopes of amino acids, but many laboratories would prefer to skip this step. Furthermore, mass isolation is not sufficiently precise to completely exclude precursors from adjacent isotopic masses. While the target analyte will surely dominate the signal, interference from adjacent isotopic peaks could still result in deviations to our derived results, especially when analyzing higher charge states. Despite these limitations, it is plausible that observed fragment ion structures will adhere closely enough to our theoretical results for them to serve as useful approximations.

From either of the equations for conditional fragment ion distributions, we are able to generate alternative theoretical spectra for use in a proteomics search algorithm. Placing these distributions at their corresponding masses results in a new theoretical MS2 spectrum (see Figure 1). To explore how variations in model assumptions impact the theoretical results we generated spectra for the peptides VLWAALLVTFLAGCAK (Figure S1) and EQIDIFEGIKDSQAQR (Figure S2)

https://doi.org/10.1021/acs.jproteome.2c00247
J. Proteome Res. 2023, 22, 334−342

using 11 different models for 0−5 heavy atoms (Figure S1). In addition to standard Comet spectra (peaks of intensity 1 at each monoisotopic m/z) we test a set of amino acid distributions that were derived using previously described logic for how 3% D_2O in blood would be synthesized into amino acids.[9] We refer to the latter set of amino acid distributions as P(Hellerstein) or P(Hel.). Additionally, we generated spectra from a range of possible amino acid isotopic distributions applied equally across all 20 amino acids. These are denoted with the notation P(M0|M1|M2|M3|M4) which implies a probability mass function on each isotopologue (with trailing zeroes left blank), and we test search results setting the mass probabilities as P(99|1), P(97|3), P(80|20), P(75|25), P(70|30), P(70|10|10|10), P(50|50), P(40|60), P(30|70), P(25|75), and P(33|33|33).

RESULTS AND DISCUSSION

Stochastic isolation of stable isotope containing peptide precursors during data-dependent experiments results in nonmonoisotopic fragmentation spectra. While the monoisotopic mass of the precursor can be estimated by searching for more precursor offsets or by incorporating monoisotopic peak estimation,[3,6,10] the effect on fragment isotopic distributions has not been widely explored for peptide spectral matching. Crucially, the isotopic distribution of fragment ions is substantially altered by process of mass isolation (Figure 1A), so that the expected isotopic envelopes for a given fragment are not the same as they are for the precursor (Figure 1B). Furthermore, the distribution of fragment ion isotopologues follows a predictable pattern where the smaller peaks are dominated by a single M0 peak while the largest are dominated by a single peak at the offset matching the number of heavy atoms in the precursor (Figure 1C). All the fragments between the smallest and largest ions inevitably display a transition between these two states. Consequently, variations in the assumptions about amino acid labeling appear to have a minimal impact on our theoretical MS2 spectra.

Visualizing theoretical MS 2s for the peptide EQIDIFEGIKDSQAQR (Figure 1D) while varying our assumptions about the underlying state of amino acid labeling, we see that the spectra are largely similar to all convolutional models providing a closer approximation to the real data than the peak structure generated by standard search engines. This relationship held true across a large range of potential amino acid distributions and variations to the number of heavy atoms isolated in the precursor (Figures S1 and S2). The M4 peak observed in the real data for the b_{11} ion shows another imperfection of our theory as the assumption of perfect mass isolation does not allow for the existence of this peak. Yet all of the Convolutional Isotope Models are far closer to reality than the theoretical spectrum built with only monoisotopic fragment ions. These results suggest that any approximation of the underlying amino acid labeling could serve as a useful tool in the context of peptide-spectrum-matching.

To test the utility of Comet-CIDS, we searched published data from a deuterium labeling protein turnover experiment.[4] Briefly, murine liver samples were collected from $LDLR^{-/-}$ mice—a model of nonalcoholic fatty liver disease—fed a normal diet at 0, 168, and 504 h after a bolus injection of D_2O and subsequent replacement of their drinking water with 5% D_2O. The authors calculated the deuterium body water enrichment of these samples to be 3%.[4] Proteins were separated by SDS-PAGE into individual bands and each band was processed for LC-MS/MS analysis on a Q-Exactive Plus (Thermo). Samples were originally searched with the Mascot search engine[11] with no modifications for the presence of deuterium, wide fragment ion tolerances (0.6 Da), and originally filtered to a PSM FDR of 5%. When researching the data, we used more stringent filtering to reduce the PSM and protein FDR to 1% for all search results and maintained the low-resolution search parameters used in the original d2ome analysis. This was done to enable comparison to the original search data and because of the slow speed of the CIDS convolutional isotope modeling (Figure S3).

All searches (Comet and Comet-CIDS) were preprocessed with Monocle's monoisotopic peak estimation[3] and multiple peak offsets to account for—and estimate—peptide deuteration (Figure 3A). For the convolutional isotope model, we used the P(97|3) (probability of 97% with 0 isotopes and 3% with at least one isotope) as an approximation to the underlying state of free-floating amino acids based on deuterium body-water enrichment. The full set of additional Comet-CIDS parameters is provided in Table S1. Note that these assumptions are very crude approximations. D_2O only labels nonessential amino acids through the biosynthesis of amino acids, many of which will incorporate more than one deuteria.[12] As a result, our assumptions both under and overestimate the amount of labeling. Despite this crude approximation, the Comet-CIDS strategy still greatly improved peptide identifications (Figure 2).

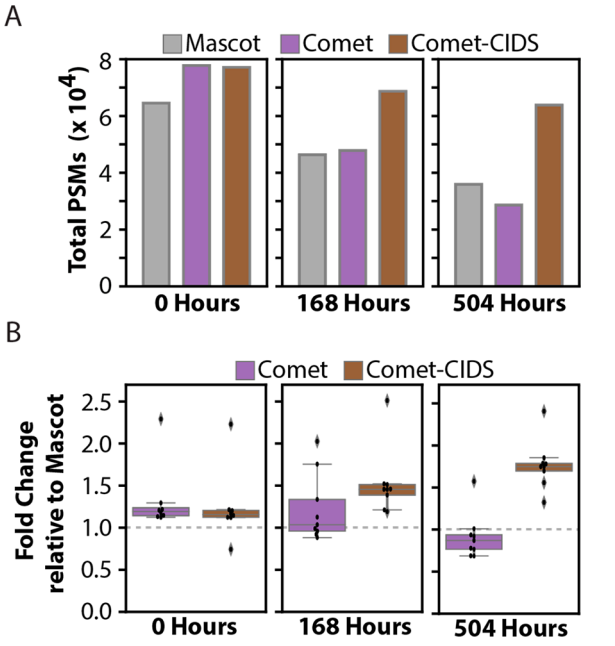


Figure 2. Peptide spectral match results. (A) PSMs results for Mascot, Comet, and Comet-CIDS at each of the three time points tested. (B) Relative improvement compared to published (Mascot) results. Longer time points resulted in increased improvement with Comet-CIDS.

Comet and Comet-CIDS performed approximately equally well at the 0-h time point with both sets of search results providing a moderate boost over the previously reported Mascot search results (Figures 2 and S4). Consistent scoring at the 0-h time point was expected, as no isotopic incorporation would have occurred at 0 h (Figure 2B, 3A). Importantly, reproducible PSM sensitivity at 0 h built confidence that the

https://doi.org/10.1021/acs.jproteome.2c00247
J. Proteome Res. 2023, 22, 334−342

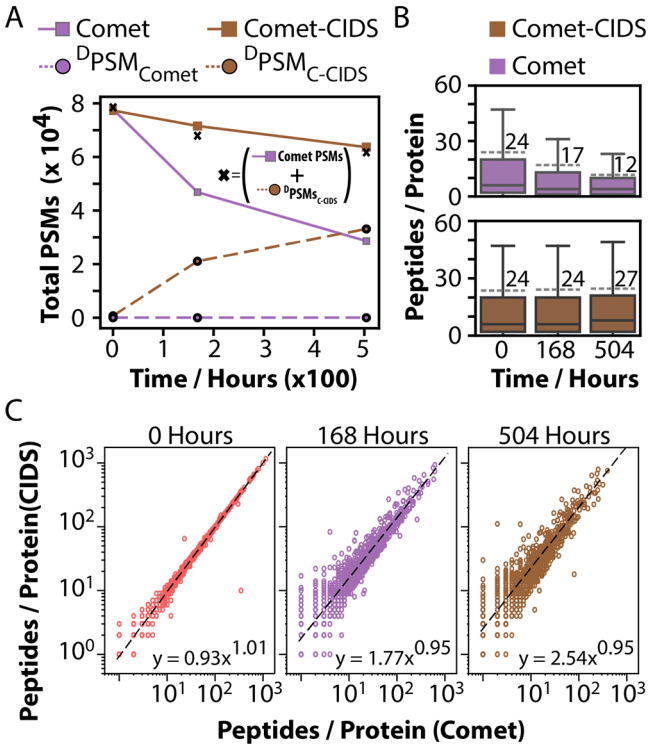


Figure 3. Time dependent effects on PSMs. (A) Increased labeling time results in more Comet-CIDS identifications compared to Comet. The majority of PSMs gained by Comet-CIDS are derived from deuterated precursors (dotted lines) and the sum of the PSMs identified in Comet plus the deuterated peptides from Comet-CIDS comes close to the total number of Comet-CIDS identifications (shown with an x). (B) Comparison of the total peptides observed for each protein identified using Comet versus using Comet-CIDS. Average peptides per protein are indicated for each by dotted lines. (C) The increased number of peptides-per-protein was driven by a consistent improvement in identifications across time points. Dotted line is a log−linear regression.

modeled fragment distributions were not artifactually affecting the PSM sensitivity. Building on this, at longer time points Comet-CIDS increased the total number of confidently identified PSMs compared to Mascot by 76% and 110% at the 168- and 504-h time points, respectively (Figure S4). By comparison, Comet identified more peptides at the 0- and 168-h time points (30% and 15%) but actually generated fewer PSMs than Mascot at 504-h (6%). Thus, Comet-CIDS' sensitivity improvement was consistently observed in every gel band, across all time points (Figure S4).

The improvement from Comet-CIDS increased the total number of peptides observed per protein (Figure 3B). Regression analysis of the total peptides observed per protein identified estimated a 1.8-fold increase at 168 h, and 2.5-fold increase at 504 h (Figure 3C). The time dependent gain in PSMs suggests that the increased sensitivity is not randomly occurring across the entire set of PSMs. Notably, the gains predominantly occur in a set of highly valuable peptides—the ones that contain deuteria (Figure 3A). At the 504-h time point, more than 50% of observed PSMs were derived from deuterated peptides when using Comet-CIDS (Figure S5) and these are precisely the peptides that enable estimation of protein turnover rates. These data suggest that Comet-CIDS is highly effective at identifying deuterated PSMs. Comet and Comet-CIDS had highly correlated PSM scoring (XCorr) for most peptides with a subset of the data demonstrating substantial improvements (Figure S6).

Improved PSM sensitivity also resulted in gains to the total number of protein identifications and our ability to detect relevant gene set enrichments (Figure 4, S7). We found that Comet-CIDS improved protein identifications at the 504 h time point by 11% − 1903 with Comet to 2119 with Comet-CIDS. This resulted in an identification rate in line with the 0 h time point with no stable-isotope incorporation (2491/2465). We observed that, as with the PSMs, these gains were dependent on labeling time with a greater improvement seen after 504 h compared to 168 h. A small number of peptides were identified by Comet but not by Comet-CIDS (Figure 4B). However, this effect was not dependent on labeling time and thereby not likely to result in the loss of deuterated protein identifications. In addition to the time dependent gains in protein identifications, we observed that for protein components of CORUM complexes, Comet-CIDS improved the peptide coverage of proteins in a similar time dependent manner (Figure S7). Moreover, the 381 proteins identified exclusively by Comet-CIDS are associated with proteostasis and protein processing (Figure S7B). In particular, across multiple annotation classes, we observed enrichment for terms associated with metabolism, translation, mRNA processing, and protein translocation. There is a clear link of these processes to protein turnover,[13] but it is additionally interesting that proteins within these general functional classes

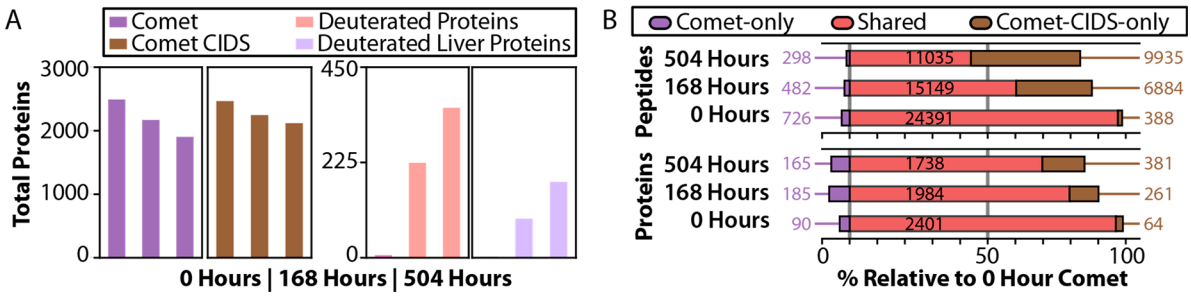


Figure 4. Comet-CIDS increase protein identifications across time points. (A) Comet-CIDS improved protein identifications at longer time points. Similar to PSM results, longer time points resulted in more deuterated proteins. Interestingly these include an increase in proteins annotated as liver proteins demonstrating improved coverage of tissue-specific proteins collected from murine liver tissue. (B) Comet-CIDS improved total peptide and protein identifications largely intersected with the original Comet protein identifications. Proteins or peptides identified by Comet but not Comet-CIDS (Lost − Purple), identified with both (Shared − Red), or identified with only Comet-CIDS (Gained − Brown) are noted. Bars represent the percentage of peptides or proteins identified relative to the 0 h Comet identifications.

https://doi.org/10.1021/acs.jproteome.2c00247
J. Proteome Res. 2023, 22, 334−342

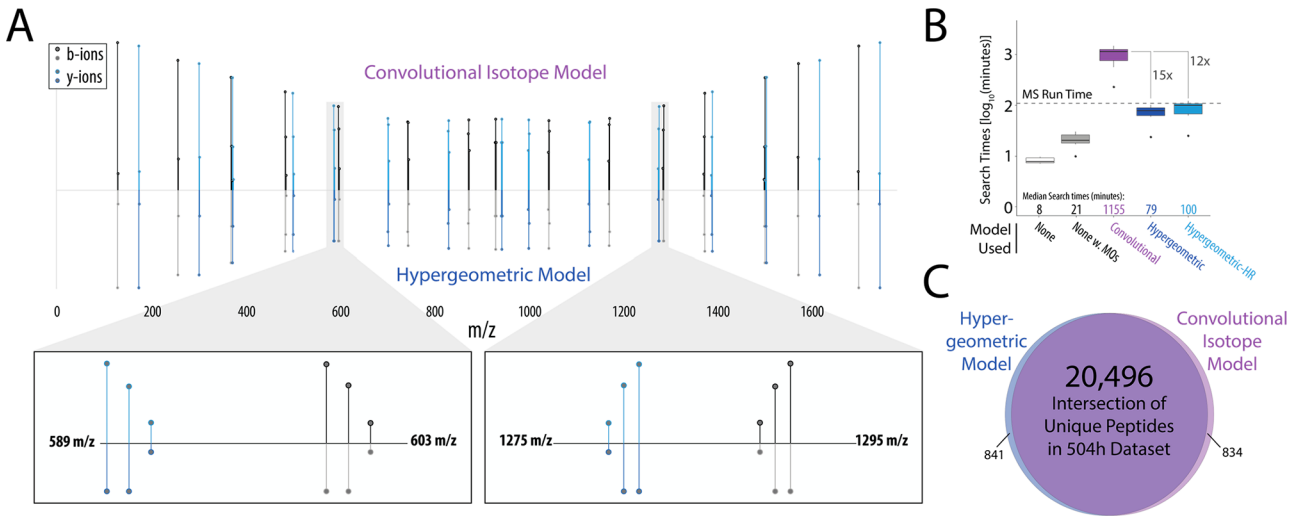


Figure 5. Comparison of convolutional isotope and hypergeometric models for CIDS. (A) Convolutional isotope and hypergeometric models produce similar fragment ion distributions. Insets show specific fragment ions from the larger spectra. Both b- (gray) and y-ions (blue) are highlighted in the insets below. For all spectra, fragments from the convolutional isotope model are on top and from the hypergeometric model are on the bottom of the reciprocal plots. (B) Search times for different CIDS models for the 504 h d2ome data set. "None" refers to no model being used (Comet), "None-MOs" refers to no model being used with 10 mass offset windows used in the Comet search. Hypergeomteric modeling reduces CIDS search times by 15× for the 504 h D2Ome data set. The speed increase of the hypergeometric modeling enables high-resolution searching as well. Total MS run time is noted with the dotted line and median search times are shown for each set of searches. (C) Comparison of unique peptides identified by the hypergeometric model or the original convolutional isotope model.

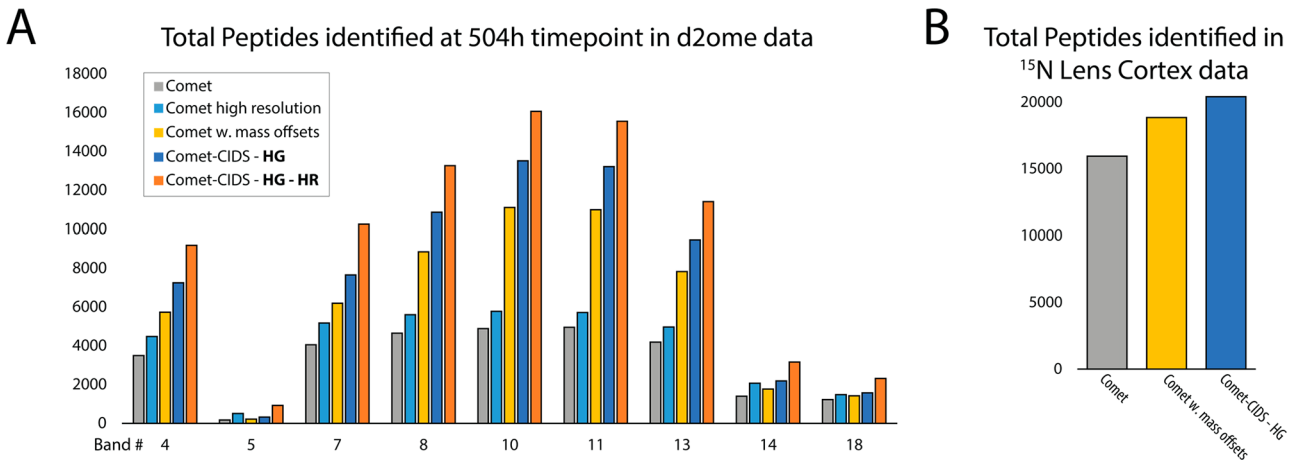


Figure 6. Hypergeometric models for CIDS increases peptide identification for isotopically labeled peptides. (A) Hypergeometric (HG) modeling is fast enough to enable high-resolution (HR) Comet searching which generates more peptide identifications than Comet (with or without mass offsets or high-resolution searching) or the low resolution hypergeometric CIDS searches. (B). CIDS can be applied to additional isotopically enriched data types such as ^{15}N labeled, murine lens cortex samples. The hypergeometric CIDS modeling improves detection of peptides by 28% compared to Comet alone and 8% compared to Comet with multiple mass offsets.

are likely deuterated to a degree that obfuscates their identification without consideration of stable-isotope incorporation during searching.

While the results from applying Comet-CIDS to the d2ome data were very encouraging, they came at a substantial cost in processing time. The convolutional isotope models are built in real time during database searches and, owing to the time necessary to model each fragment, require a substantial increase in total search time (Figure S3). However, all the results suggested that the number of heavy atoms in the precursor would be the most important variable in determining the structure of the isotopologues. Accordingly, we reran the analyses implementing the hypergeometric model (eq 8), which generates very similar MS2 spectra despite the relaxed assumptions that underlie the model (Figure 5A). Our implementation of the hypergeometric model relies on precalculated intensities and searching the data in this way reduced computing times by 15-fold (Figure 5B) with minimal differences in the number of unique peptides identified (Figure 5C). Compared to traditional Comet search, the hypergeometric model in Comet-CIDS was only 3.9-fold slower, whereas with the convolutional isotope model, Comet-CIDS was 56-fold slower. Thereby, even the most time and memory intensive searches (for example, multiple precursor offsets or high-resolution fragment binning) can be accomplished relatively quickly—that is, less than MS acquisition times. Note that we also included a comparison against Comet searched with all the offsets used in Comet-CIDS. This comparison was attempted to determine how much of the increase in computing time resulted from searching more

https://doi.org/10.1021/acs.jproteome.2c00247
J. Proteome Res. 2023, 22, 334−342

masses, however it also revealed that simply searching for all the likely precursor masses provides a substantial increase in PSMs (Figure 6A). It was not obvious that increasing the search space without adjusting the MS2 spectra would be beneficial, and we are unaware of any examples in the literature where other groups have employed such a strategy. Nonetheless, it does appear to be the best strategy other than using Comet-CIDS. Importantly, the speed and memory improvements of the hypergeometric modeling enabled searches using high-resolution Comet fragment binning resulting in a further increase in total PSMs identified for each of the d2ome gel bands (Figure 6A).

The peptide and protein identification improvements highlighted in this manuscript would be expected for any protein turnover technology that partially labels amino acids but up to this point we have only considered deuterium labeling. To demonstrate this, we analyzed ^{15}N labeled murine lens cortex epithelial proteins.[7] These samples were derived from mice that had been exclusively fed a ^{15}N diet after weaning. Since the eye is believed to have a highly stable proteome, the authors collected only one sample 12 weeks after replacing the food source with ^{15}N spirulina. ^{15}N labeling does not share the toxicity concerns that limit label administration in D_2O experiments. Therefore, labeling rates tend to be far higher, resulting in wider isotope distributions in MS1 scans.

In this experiment, there was also an a priori expectation that many proteins would not be turned over at all. Therefore, the authors chose to study only a single time point 12 weeks after label administration began. To account for the diffuse MS1 isotopes we increased the number of offset masses in addition to the monoisotopic precursor searched from 10 (in the D_2O data) to 20. This created the unfortunate situation where we might routinely search for more masses than would be possible (small peptides) and still potentially fail to search for enough mass offsets for larger peptides with high nitrogen content. Furthermore, the overall lack of turnover in the eye suggests that for most PSMs, Comet-CIDS would provide little benefit over Comet, at least on average. Nonetheless, we found that our Comet-CIDS pipeline improved identifications compared to Comet by 28%, with an 8% gain in PSMs relative to the default Comet search with the 20 mass offsets included (Figure 6B).

CONCLUSIONS

By modeling the isotopologue distributions of MS2 fragment ions, we found that it is possible to improve peptide spectral matching compared to conventional database search methods. Both the theoretical and observed benefits occur specifically when a nonmonoisotopic precursor, composed of many structural isomers, has been isolated for fragmentation. By integrating the CIDS approach with the Comet database search platform, we provide a novel means to integrate isotopically labeled peptide fragment ion distributions at database scale. Relative to standard search strategies, the benefits for protein turnover experiments can be substantial.

We explored the utility of the Comet-CIDS approach in the context of protein turnover experiments. Using previously acquired data, we demonstrated that Comet-CIDS can improve the detection of isotopically labeled peptides and proteins in pulse-chase protein turnover experiments. In addition to the improved coverage of the proteome, we anticipate that the increased PSMs per protein will have a highly beneficial impact on quantitative performance. Identifying PSMs that were lost due to peptide deuteration leads to larger sample sizes and could mitigate problems associated with missing data, which persist despite the best efforts to plug in missing values with imputations or by matching masses across runs.[14−16] Furthermore, since Comet-CIDS specifically improves the identification of labeled peptides, we anticipate that the quantitative benefits for protein turnover estimation will predominantly fall upon the proteins undergoing turnover during the course of the experiment.

Finally, we note that fragmentation biases are inherent to nearly all DDA experiments. Therefore, CIDS has the potential to provide benefits beyond stable isotope-based protein turnover experiments as CIDS better predicts the isotopic fragment distributions resulting from selection of non-monoisotopic precursor peaks. Traditional DDA experiments continue to rely on stochastic, intensity-dependent selection of $^{13}C/^{15}N$ isotopic peaks from a given precursor's isotopic distribution[17] and theoretical spectra in standard database searching strategies—and modern machine learning models—were designed for detecting monoisotopic peaks.[1,2] Yet, in cases where high intensity ^{13}C-containing precursor peaks are isolated, resulting in the presence of large ^{13}C fragment peaks, CIDS should be able to match peptides that would go unidentified with traditional searches. In addition, future applications could target structural isomers where the probability of labeling individual amino acids is completely unknown, as happens routinely in protein footprinting experiments.[18] Because the process of precursor mass isolation has a strong and highly predictable impact on the isotopic distribution of fragment ions, we believe that incorporating this structure into search algorithms could have many unexpected benefits, enabling the use of a large set of previously unusable observations.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.2c00247.

Table S1. Additional parameters for Comet-CIDS implementation. Figure S1. Comet-CIDS theoretical spectra for the peptide VLWAALLVTFLAGCAK. Individual fragment ions are shown with capped lines to highlight the fragment distributions. Insets to the right of each full spectra highlight the b91+ fragment ion distribution as a function of the number of incorporated deuterium. Each individual peak in the CIDS theoretical spectra has similar isotopic distribution profiles which can be built using CIDS on a per peptide/spectrum basis. Models with blank spectra indicate where the marginal probability of observing a particular number of deuteria in the precursor (Q_t in eq 7), was below the default setting of 0.0001 ('isotope_min_prob = 0.0001', Table S1). Figure S2. MS2 spectra of EQIDIFEGIKDSQAQR and the Comet-CIDS predicted spectrum as a function of labeling estimation for free-floating amino acid distribution. Figure S3 Search times for Comet, Comet with 10 mass offset windows, and Comet-CIDS (convolutional isotope model) with 10 mass offsets. Median search times and relative time differences are highlighted in the plot. Figure S4. (A) Boxplot of total PSMs for individual runs within each time point data set.

https://doi.org/10.1021/acs.jproteome.2c00247
J. Proteome Res. 2023, 22, 334−342

(B) Total PSMs per run for each individual band across the full data set of the three time points. Figure S5. Total PSMs identified at individual time points and those attributed to deuterated peptides. Deuterated peptides were annotated as those with a mass shift of a 1 or more deuterons after initial monoisotopic peak correction (see Experimental Procedures). Figure S6. XCorr comparison between PSMs from Comet and Comet-CIDS at 0 h (A) and 504 h (B). Left plots show matching scans for PSMs filtered to 1% protein FDR. Plots shown on the right correspond to matched scans from unfiltered data. (C) XCorr distributions for PSMs unique to Comet (top, purple) and Comet-CIDS (bottom, brown). Figure S7. Improved peptide and protein identifications are involved in proteostasis. (A) Within CORUM complexes, Comet-CIDS consistently improved protein coverage for constituent proteins. (B) Significant enrichment for the 381 proteins identified with Comet-CIDS, but not Comet, at the 504-h time point. BioPlanet: NCATS BioPlanet; GO-BP: GO Biological Process; GO−CC: GO Cellular Component; GO-MF: GO Molecular Function; WikiPathway. Enrichment calculated using Enrichr (PDF)

AUTHOR INFORMATION

Corresponding Authors

Devin K. Schweppe − *University of Washington, Seattle, Washington 98105, United States;* orcid.org/0000-0002-3241-6276; Email: dkschwep@uw.edu

Jonathon J. O'Brien − *Calico Laboratories, South San Francisco, California 94080, United States;* orcid.org/0000-0001-9660-4797; Email: obrienj@calicolabs.com

Authors

Meagan Gadzuk-Shea − *University of Washington, Seattle, Washington 98105, United States*

Phillip M. Seitzer − *Calico Laboratories, South San Francisco, California 94080, United States;* orcid.org/0000-0002-7379-8960

Ramin Rad − *Calico Laboratories, South San Francisco, California 94080, United States*

Fiona E. McAllister − *Calico Laboratories, South San Francisco, California 94080, United States;* orcid.org/0000-0002-7862-6711

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.2c00247

Author Contributions

[§]These authors contributed equally to this work.

Notes

The authors declare the following competing financial interest(s): J.J.O., P.S., R.R., and F.M. are employees of Calico Labs, LLC.

ACKNOWLEDGMENTS

We would like to thank the Schweppe Lab and Jimmy Eng of the UWPR for helpful advice and Adam Baker for help designing the figures.

REFERENCES

(1) Eng, J. K.; McCormack, A. L.; Yates, J. R. 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*, 976−989.

(2) Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; Reimer, U.; Ehrlich, H.-C.; Aiche, S.; Kuster, B.; Wilhelm, M.; et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **2019**, *16*, 509.

(3) Rad, R.; et al. Improved Monoisotopic Mass Estimation for Deeper Proteome Coverage. *J. Proteome Res.* **2021**, *20*, 591−598.

(4) Sadygov, R. G.; et al. d2ome, Software for in Vivo Protein Turnover Analysis Using Heavy Water Labeling and LC-MS, Reveals Alterations of Hepatic Proteome Dynamics in a Mouse Model of NAFLD. *J. Proteome Res.* **2018**, *17*, 3740−3748.

(5) Goldfarb, D.; Lafferty, M. J.; Herring, L. E.; Wang, W.; Major, M. B. Approximating Isotope Distributions of Biomolecule Fragments. *ACS Omega* **2018**, *3*, 11383−11391.

(6) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13*, 22−24.

(7) Liu, P. Long-lived metabolic enzymes in the crystalline lens identified by pulse-labeling of mice and mass spectrometry. *Elife* **2019**, *8*, No. e50170, DOI: 10.7554/eLife.50170.

(8) Schweppe, D. K.; et al. Full-Featured, Real-Time Database Searching Platform Enables Fast and Accurate Multiplexed Quantitative Proteomics. *J. Proteome Res.* **2020**, *19*, 2026−2034.

(9) Holmes, W. E.; Angel, T. E.; Li, K. W.; Hellerstein, M. K. Dynamic Proteomics: In Vivo Proteome-Wide Measurement of Protein Kinetics Using Metabolic Labeling. *Methods Enzymol* **2015**, *561*, 219−276.

(10) Senko, M. W.; Beu, S. C.; McLaffertycor, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 229−233.

(11) 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*, 3551−3567.

(12) Holmes, W. E.; Angel, T. E.; Li, K. W.; Hellerstein, M. K. Dynamic Proteomics: In Vivo Proteome-Wide Measurement of Protein Kinetics Using Metabolic Labeling. *Method Enzymol* **2015**, *561*, 219−276.

(13) Hipp, M. S.; Kasturi, P.; Hartl, F. U. The proteostasis network and its decline in ageing. *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 421−435.

(14) O'Brien, J. J.; et al. The Effects of Nonignorable Missing Data on Label-Free Mass Spectrometry Proteomics Experiments. *Ann. Appl. Stat* **2018**, *12*, 2075−2095.

(15) O'Connell, J. D.; Paulo, J. A.; O'Brien, J. J.; Gygi, S. P. Proteome-Wide Evaluation of Two Common Protein Quantification Methods. *J. Proteome Res.* **2018**, *17*, 1934−1942.

(16) Lim, M. Y.; Paulo, J. A.; Gygi, S. P. Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model. *J. Proteome Res.* **2019**, *18*, 4020−4026.

(17) Alevra, M.; et al. A mass spectrometry workflow for measuring protein turnover rates in vivo. *Nat. Protoc* **2019**, *14*, 3333−3365.

(18) Johnson, D. T.; Di Stefano, L. H.; Jones, L. M. Fast photochemical oxidation of proteins (FPOP): A powerful mass spectrometry-based structural proteomics tool. *J. Biol. Chem.* **2019**, *294*, 11969−11979.