MIND: A Double-Linear Model To Accurately Determine Monoisotopic Precursor Mass in High-Resolution Top-Down Proteomics

Peer-reviewed author version

Lermyte, Frederik; Dittwald, Piotr; CLAESEN, Jurgen; Baggerman, Geert; Sobott, Frank; O'Connor, Peter B.; Laukens, Kris; HOOYBERGHS, Jef; Gambin, Anna & VALKENBORG, Dirk (2019) MIND: A Double-Linear Model To Accurately Determine Monoisotopic Precursor Mass in High-Resolution Top-Down Proteomics. In: ANALYTICAL CHEMISTRY, 91(15), p. 10310-10319.

DOI: 10.1021/acs.analchem.9b02682
Handle: http://hdl.handle.net/1942/29122

MIND: A double-linear model to accurately determine monoisotopic precursor mass in high-resolution top-down proteomics

Frederik Lermyte[1,2,3,4]*, Piotr Dittwald[5]*, Jürgen Claesen[6], Geert Baggerman[2,7], Frank Sobott[1,8,9], Peter B. O'Connor[4], Kris Laukens[10,11], Jef Hooyberghs[7], Anna Gambin[5], Dirk Valkenborg[2,6,7]

[1]Biomolecular and Analytical Mass Spectrometry Group, Department of Chemistry, University of Antwerp, Antwerp, Belgium
[2]UA-VITO Center for Proteomics, University of Antwerp, Antwerp, Belgium
[3]School of Engineering, University of Warwick, Coventry, United Kingdom
[4]Department of Chemistry, University of Warwick, Coventry, United Kingdom
[5]Institute of Informatics, University of Warsaw, Warsaw, Poland
[6]Interuniversity Institute of Biostatistics and Statistical Bioinformatics, Hasselt University, Hasselt, Belgium
[7]Applied Bio and Molecular Systems, Flemish Institute for Technological Research (VITO), Mol, Belgium
[8]Astbury Centre for Structural Molecular Biology, University of Leeds, Leeds, United Kingdom
[9]School of Molecular and Cellular Biology, University of Leeds, Leeds, United Kingdom
[10]Adrem Data Lab, Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium
[11]Biomedical Informatics Network Antwerp (biomina), University of Antwerp, Antwerp, Belgium

*These authors contributed equally to this work

Corresponding author: Dirk Valkenborg (dirk.valkenborg@uhasselt.be)

Keywords: Mass spectrometry, Bioinformatics, Data processing, Top-down proteomics, Intact protein mass spectrometry, Post-translational modification, smart algorithms

Abstract: Top-down proteomics approaches are becoming ever more popular, due to the advantages offered by knowledge of the intact protein mass in correctly identifying the various proteoforms that potentially arise due to point mutation, alternative splicing, post-translational modifications, *etc.* Usually, the average mass is used in this context; however, it is known that this can fluctuate significantly due to both natural and technical causes. Ideally, one would prefer to use the monoisotopic precursor mass, but this falls below the detection limit for all but the smallest proteins. Methods which predict the monoisotopic mass based on the average mass, are potentially affected by imprecisions associated with the average mass. To address this issue, we have developed a framework based on simple, linear models, which allows prediction of the monoisotopic mass based on the exact mass of the most abundant (aggregated) isotope peak, which is a robust measure of mass, insensitive to the aforementioned natural and technical causes. This linear model was tested experimentally as well as *in silico*, and typically predicts monoisotopic masses with an accuracy of only a few parts per million. A confidence measure is associated to the predicted monoisotopic mass to handle the off-by-one-Da prediction error. Furthermore, we introduce a correction function to extract the 'true' (*i.e.* theoretically) most abundant isotope peak from a spectrum, even if the observed isotope distribution is distorted by noise or poor ion statistics. The method is online available as a R shiny app: https://valkenborg-lab.shinyapps.io/mind/

1. Introduction

To a large extent, the spectacular growth observed in the field of mass spectrometry (MS)-based proteomics over the last two decades can be viewed as a triumph of bioinformatics. In particular, the ability to correlate peptide tandem MS data to protein sequences in databases has been invaluable to making bottom-up proteomics a practical technology[1-3]. In contrast, top-down approaches dispense with the use of enzymatic digestion, and instead rely on the isolation and fragmentation of an intact protein ion in the gas phase[4-6]. This is conceptually much more straightforward, and more akin to the approach conventionally used for the identification of small molecules. However, more stringent requirements are imposed on the type of mass spectrometer used (including transmission of intact protein ions, extensive fragmentation, high sensitivity, *etc.*), explaining why the field of top-down proteomics has only recently emerged as more than an academic curiosity[7-9]. Besides the conceptual simplicity of top-down proteomics, the second major benefit is knowledge of the intact protein precursor mass. This allows the efficient detection and identification of protein variants, collectively known as proteoforms[10], resulting from *e.g.* alternative RNA splicing, post-translational modification, truncation, or site-specific mutation. To illustrate this point, a pilot project was recently undertaken, in which no fewer than 74 proteoforms of the human histone H4 were identified in a top-down workflow[6].

It is beneficial that the aforementioned determination of the precursor mass is performed with a high accuracy, as this reduces the ambiguity of identification, particularly for unknown proteoforms. For instance, lysine trimethylation and acetylation, two PTMs relevant in histone characterization, lead to very similar observed mass differences, just 0.036 Da apart. The same difference also exists between a lysine *versus* a glutamine residue[11], a substitution that can be caused by a single A-to-C nucleotide mutation. For a typical 11 kDa protein (the approximate size of the aforementioned histone H4), the resulting difference in precursor mass can only be discerned at a mass accuracy better than 3.2 parts per million (ppm). Also, in intact protein analysis, the concept of 'precursor mass' must be further specified, as the large number of atoms present leads to the occurrence of broad, complex isotope distributions[12,13]. As such, the monoisotopic and average mass, and the mass of the most abundant (aggregated) isotope peak, are three potentially different values even for a medium-sized protein, as illustrated in **Figure S1**. The average mass is experimentally the most easily accessible value (as it can be determined even if resolving power is insufficient to separate the isotope peaks) and therefore the most used in practice. However, we have recently shown that this value is rather sensitive to both natural and technical variations of the elemental isotope definitions[14]. Ideally, one would prefer to use the monoisotopic precursor mass, as this value does not change with fluctuations of elemental isotopic abundances. Unfortunately, as is fairly trivial to show, the probability of occurrence of the monoisotopic variant is very low even for proteins of intermediate size[12], and as such, the monoisotopic variant typically falls well below the limit of detection. In the same study where we demonstrated the low precision of average protein masses, we also showed that the mass of the most abundant (aggregated) isotope peak (henceforth referred to as 'most abundant mass' for convenience) is considerably more robust, and as such, potentially offers a good compromise between ease of detection and reduction of ambiguity[14]. As mentioned earlier, though, it would be convenient to know the monoisotopic mass for ease of identification and database searching.

In the current work, we present an algorithm which relates the most abundant to the monoisotopic mass of a protein. We have named this algorithm MIND (MonoIsotopic liNear preDictor). The method

was trained and tested *in silico* on unmodified human protein sequences found in the UniProt database (release 2014), in the 8 – 60 kDa mass range. The difference between monoisotopic mass as calculated from the molecular formula and predicted by MIND was typically well below one part per million (ppm) in these *in silico* experiments, comparing favorably to existing methods for estimating monoisotopic protein mass[15]. With the exception of ubiquitylation, *in silico* addition of the ten most commonly occurring post-translational modifications[16] (PTMs) to these sequences had only a negligible effect on the quality of the prediction.

The greater stability of the MIND-predicted monoisotopic mass, compared to average mass, was shown experimentally through the analysis of a set of four proteins. In these experiments, performed to simulate the single-scan precursor spectra typically available in top-down LC-MS^n experiments, the measured average mass consistently fluctuated in a significantly wider range scan-to-scan, and with a larger systematic deviation, compared to the predicted monoisotopic mass using MIND.

In contrast to previously developed averagine-scaling methods, which require that the observed isotope distributions match the theoretical ones reasonably well, our method requires only the confident selection of the most abundant isotope peak. In case of poor ion statistics (*i.e.* a low number of ions), this might not be the most intense isotopic signal observed. However, we show that while the average mass has a high random fluctuation, it operates in a limited range (*i.e.* this value is measured with a high accuracy, but a low precision), whereas poor ion statistics will lead to an error on the most abundant mass of a (very nearly) integer number of atomic mass units (*i.e.* a high-precision, low-accuracy measurement). By exploiting a correlation between the average and most abundant mass, we can therefore confidently identify the 'true' most abundant isotope peak with both high-precision and high-accuracy.

As this approach combines the benefits of measurement of the most abundant (sensitivity, robustness) and monoisotopic (confident identification, highly interoperable) mass values, we believe this method will be of considerable interest to the ever-growing field of top-down proteomics.

2. Materials and Methods

2.1. Description of the MIND algorithm

To develop and test this algorithm, all 95,616 human proteins (78,328 after removing redundant elemental compositions) with a mass between 8 – 60 kDa in the UniProt database (release 2014) were used. Of the 78,328 unique masses, 10,000 were randomly removed from the training data set and used as an independent validation set. While we expect the resulting model to work well for most mammalian proteins, the same workflow depicted in Figure 3 can easily be tailored to generate a MIND model for different classes of proteins (or other polymers) if necessary.

The BRAIN algorithm[12,17,18] was used to compute the isotope distribution of each unmodified protein. In this manner, we obtain the monoisotopic, most abundant, and average masses for each sequence (post-translational modifications are only considered at a later stage; see **Section 3.3**). As already suggested by Dittwald *et al.*[17], if we plot the monoisotopic (y-axis) *vs.* most abundant (x-axis) mass for each sequence, an approximately linear relation with intercept (α) and a slope (β) just below unity is observed, as can be intuitively expected (**Figure 1a**). This relation is described by **Equation 1**:

$$M_{mono} = \alpha + \beta \times M_{MostAb} + \varepsilon \qquad [1]$$

where M_{mono} is the monoisotopic mass, and M_{MostAb} is the most abundant mass. After fitting this linear model we obtain the following parameter estimates for $\alpha = 0.6074$ and $\beta = 0.9994$ (**Table S1**). The errors or residuals, depicted by ε, can be calculated for each protein *i* as the difference between the monoisotopic masses M_{mono} from the training data set and the predicted monoisotopic masses $\widehat{M_{mono}}$ using the model in **Equation 1**. The distribution of the residuals expressed in Da can be observed in the inset of **Figure 1a**. It turns out that for approximately 65% of all proteins, $|\varepsilon| \leq 0.5$ Da, and for 99% of proteins, $|\varepsilon| \leq 2$ Da. These deviations are too large for practical use and render our simple linear model not useful for the accurate prediction of the monoisotopic mass. However, we argue that there is a special structure in the residuals of this linear model and we illustrate this structure by some straightforward mathematical manipulations applied to the residuals. Consider following null addition highlighted in grey and consecutive rearrangements of terms leading to the relation in **Equation 2** for every protein *i*:

$$\varepsilon_i = M_{mono_i} - \widehat{M_{mono_i}} + M_{MostAb_i} - M_{MostAb_i}$$

$$\varepsilon_i = M_{mono_i} - \alpha - \beta \times M_{MostAb_i} + M_{MostAb_i} - M_{MostAb_i}$$

$$\varepsilon_i = (1 - \beta) \times M_{MostAb_i} - \alpha - M_{MostAb_i} + M_{mono_i}$$

$$\varepsilon_i = (1 - \beta) \times M_{MostAb_i} - \alpha - \Delta_i \qquad [2]$$

$$\text{with } \Delta_i = M_{MostAb_i} - M_{mono_i} \qquad [3]$$

This relation expresses that for every protein *i*, the residual ε_i is a linear function of the most abundant mass with a slope of (1- β) and an intercept that is composed out of two terms being α from model [1] and Δ_i from equation [3], i.e., the difference between the most abundant mass and the monoisotopic mass for that particular protein *i*. This equation allows us to explain the trends in **Figure 1b**, where the residuals are plotted against the most abundant masses M_{MostAb}. Here, a clear structure can be seen, in which the values of the residuals ε are found on a limited number of parallel trend lines in the residual plot[19]. The lines indeed have a slope of (1- β) and the data points on a line have nearly identical values for the intercept. The reason why these residuals are concentrated on parallel lines is that the Δ value can be categorized in discrete subsets. The discrete nature of the Δ values can be explained by two factors:

1) when disregarding sulphur, proteins with a similar atomic compositions have a similar isotope distribution[20], hence proteins within a certain mass range may have a very similar Δ value.
2) slight changes in the isotope distribution can induce a change in the denomination of the most abundant peak. This shift results in a disruptive change in the Δ values with a mass of approximately 1 Da, *i.e.*, the mass of an average neutron.

The existence of subsets in Δ values is illustrated by the vertical lines in **Figure S2a** that displays the residuals ε against the Δ values – only a small scatter on the x-axis can be observed when zooming into this plot**.** For example, the proteins that are composing the parallel trend lines in **Figure S2a** indicated by blue, black and red have nearly identical Δ value as indicated in **panel b** by the corresponding colors.

Moreover, the average Δ values for the blue, black and red lines are equal to 9.0226, 10.0252 and 11.0277 Da respectively. The variance associated to these mean values are very small and below 1.5e-6. Therefore, only a small error is made when replacing the protein's exact Δ_i value in **Equation 2** by its averaged value $\overline{\Delta_s}$ of the proteins belonging to subset *s*:

$$\varepsilon_i = (1 - \beta) \times M_{MostAb_i} - \alpha - \overline{\Delta_s} \qquad [4]$$

The model in **Equation 4** now defines a set of proteins *i* that have a constant value for $\overline{\Delta_s}$. This equation can now be used to further fine-tune the prediction of the monoisotopic mass from **Equation 1** by the simple addition of the result from **Equation 1** and the prediction from **Equation 4**. For example, in the case of myoglobin with a monoisotopic mass of 16940.965 Da and a most abundant isotope mass of 16950.992, a vertical (**Figure 1b**) and horizontal (**Figure S2b**) line will intersect three of these parallel vertical trend lines. As a consequence, for the given most abundant isotope mass there will be three corresponding $\overline{\Delta_s}$ values: $\overline{\Delta_9} = 9.0226$, $\overline{\Delta_{10}} = 10.0252$ and $\overline{\Delta_{11}} = 11.0277$ as depicted in **Figure S2a**. Do note that the index *s* indicates the nominal mass difference between the most abundant isotope mass and monoisotopic mass. Next, **Equation 4** yields three possible residual values for myoglobin: 1.000434 (blue arrow), -0.002109 (black arrow), and -1.004583 (red arrow) depicted in **Figure 1b**. In turn, the three predicted residual values can now be added to the predicted value from **Equation 1** yielding three estimates for the monoisotopic mass: (16940.899 + 1.000434) Da, (16940.899 + 0.002109) Da, and (16940.899 -1.004583) Da. Note that in previous example only one of the three predictions is correct. The procedure to assign a confidence score to the three cases will be explained later on, and typically we focus on the predictions for the most likely case.

In somewhat a similar approach to our method, Tsay *et al.*[15] have described a serendipitously discovered method for predicting $\overline{\Delta_s}$ as a function of M_{MostAb}. They have observed that their estimate for $\overline{\Delta_s}$ could be off by approximately +/- 1 Da. This off-by-one error can also be observed in the averaged values $\overline{\Delta_9}$, $\overline{\Delta_{10}}$ and $\overline{\Delta_{11}}$. In the case of myoglobin, the correct averaged value is $\overline{\Delta_{10}}$ thus leading to a difference between the other possible delta values of $\overline{\Delta_9} - \overline{\Delta_{10}} = -1.00254$ and $\overline{\Delta_{11}} - \overline{\Delta_{10}} = 1.00247$. It is trivial to see that the differences in the $\overline{\Delta_s}$values will shift the predicted values from **Equation 4** by exactly the same values. Hence, the predicted values differ by approximately 1 Da.

The model explained in the previous paragraphs suffers two major drawbacks:

- In order to estimate the residual value, we need information about the approximate $\overline{\Delta_s}$ value, which leads to an improper statistical model. Ideally, the residual values could be estimated based on the most abundant mass alone.
- The off-by-one Da error as a result of ambiguity in the actual Δ_i value cannot be avoided (see **Figure S2b**), however, a formal method that could provide confidence in the estimated residual is wanted.

In order to improve the prediction, it would be convenient to have information about which trend line or equivalently, which $\overline{\Delta_s}$value is truly associated with a given protein *i*, but as discussed previously we do not have this information and *a fortiori*, the inclusion of such information would result in an improper statistical model. We will demonstrate that, despite these issues, there exists an elegant way to avoid the incorporation of the Δ values such that we can arrive at a comprehensive yet compact model. To achieve this goal, for each protein *i*, we decompose its residual value ε into an integer part ε_{int} and a fractional part ε_{frac}, where

$$\varepsilon_{int} = [\varepsilon] \text{ and } \varepsilon_{frac} = \varepsilon - [\varepsilon]$$

with [.], the rounding operator. For now, we will focus our discussion on the estimation of the fractional part ε_{frac} only. Notice that the blue, black and red dots in **Figure 1b** are projected on shared parallel lines. This projection is shown in **Figure 1c**, where ε_{frac} (varying between -0.5 and +0.5 Da) is plotted as a function of M_{MostAb}, neglecting the integer part, ε_{int}. The result is a saw-tooth pattern. These saw-tooth parallel lines are easy to model by a piecewise linear model (**Table S2**), and it has the considerable advantage that ε_{frac} is only a function of the most abundant mass M_{MostAb}:

$$\varepsilon_{frac} = \alpha_p + \beta_p \times M_{MostAb} + \varepsilon_p \qquad [5]$$

with ε_p being a very small remaining error which is not easily corrected for.

Instead of considering the value of 1.00235 Da for adjacent isotopic peaks reported by Horn[21] presented in **Figure S2a** and used by Tsay[15], we approximate this value as 1 Da by pooling together the fractional parts ε_{frac}. Note that by this simplification, we will introduce only a small (mDa-range) error. Disregarding this very small error, a consistent mapping of the residuals ε to the fractional part ε_{frac} is achieved.

To obtain an estimate for the monoisotopic mass, we need to evaluate **Equation 1** and **Equation 5** given a most abundant peak mass and add the resulting estimates. However, in this estimation we have ignored the integer part of the residual, ε_{int}, that models the off-by-one-Da error. Indeed, in **Figure 4b**, it can be observed that more than 98% of the proteins yield ε values that round off to 0, -1, or +1 Da, with only 1.5% of ε values rounding off to ±2, and 0.2% rounding off to higher values. The question now is whether we can predict ε_{int} as a function of M_{MostAb}. For this purpose, we apply a moving window with a bandwidth of 500 Da and steps of 10 Da across **Figure 1c.** Every step of the moving window evaluates the number of residuals that did not fold back to the zero range and the number of residuals where ε_{int} equals +1 or -1. The resulting plot is displayed in **Figure 1d** and takes the form of a dampened periodic signal in function of the most abundant peak mass. Because of our choice to use the rounding operator instead of the floor or ceil function to determine the fractional part, the most probable value of ε_{int} is nearly always zero as indicated by the black line in **Figure 1d**. Averaging the probabilities over the 8 to 60kDa mass range results in the inset histogram in **Figure 1d**, we see that ε_{int} = 0 Da in 66% of the proteins in the training set. Therefore, the model illustrated in **Figure 1a** and **Figure 1c** in combination with ε_{int} of zero is sufficient to allow the prediction of the monoisotopic mass with highest probability. However, a few trends can be observed in **Figure 1d** and its close-up presented in Figure 3. First, notice that the probabilities do not have to sum to one, as ε_{int} values outside of the [-1, +1] range are not included in this analysis. Second, in general the probabilities for the zero values (black line) decrease for higher masses, whilst the probabilities for the -1/1 values (blue/red line) increase with higher masses. Third, the probabilities can differ locally in function of the most abundant mass. Because a user might want a more accurate estimate about these probabilities than the average probability, we provide the harmonic signal as a look-up table as **Supplementary information** to estimate the probabilities for ε_{int} in relation to the most abundant mass. For example, when looking at a mass of 16,000 Da, the probability for -1, 0 and 1 is equal to 0.3005, 0.6102 and 0.0874, respectively. Notice that the zero value has the largest probability, but the -1 value for ε_{int} has also a considerable probability, whilst the +1 value has low probability. A user of the MIND method

can use this confidence score to allow better decision-making about the correct monoisotopic mass in a downstream analysis.

Once the model is constructed for a given type of analyte (in this case, mammalian proteins), this calculation only involves the evaluation of two linear models (*i.e.* those shown in **Figures 1a** and **1c** for which the coefficients are provided in the supplementary **Table S2**), making the method fast and computationally inexpensive. *In silico* validation of the method, as well as a comparison to the method developed by Tsay (which has already been shown[15] to outperform those developed by Senko[22] and Zubarev[23]), was performed using the 10000 sequences that were removed from the training set used to calibrate the MIND method (see **Section 3.3**).

2.2. Mass spectrometry

For the proof-of-concept experiments, spectra were acquired on a Thermo LTQ Orbitrap Velos, operated at a resolving power of 100,000 at 400 *m/z* and 1,000,000 charges were accumulated in the LTQ for analysis in the Orbitrap. Immediately prior to infusion of the protein, external calibration was performed *via* an automatic routine, using a standard calibration mix containing n-butylamine, caffeine, MRFA, and Ultramark 1621 (Pierce LTQ Velos ESI Positive Ion Calibration Solution, Thermo catalog #88323). Bovine insulin (Sigma catalog # I5500; monoisotopic mass 5729.60 Da, average mass 5733.58 Da), equine cytochrome c (Sigma catalog # C2506; monoisotopic mass 12352.23 Da, average mass 12360.21 Da), and equine *apo*-myoglobin (Sigma catalog #M0630; monoisotopic mass 16940.97 Da, average mass 16951.50 Da) were acquired from Sigma (St. Louis, MO, USA) and infused at a concentration of 1 µM in 49:50:1 H_2O/MeCN/HCOOH, without further purification, using nano-ESI with an Advion Triversa Nanomate inlet system. The monoclonal antibody adalimumab (Thermo), used for the treatment of arthritis, was also used to test the MIND method. The antibody was reduced using TCEP, followed by LC-MS analysis using a Dionex Ultimate 3000 RSLC system (Thermo Fisher Scientific, Waltham, MA, USA) coupled to a maXis II ETD QTOF (Bruker Daltonik, Bremen, Germany). The isotope distributions observed in 61 spectra for the 24+ charge state of the light chain (monoisotopic mass 23397.61 Da, average mass 23412.32 Da) of the antibody was used as input for the MIND algorithm.

3. Results and discussion

3.1. Proof-of-concept experiments on LTQ-Orbitrap and maXis II mass spectrometers

In order to evaluate both the real-world performance and scan-to-scan stability of the monoisotopic mass predicted using MIND, 50 spectra of intact insulin, and 200 spectra of cytochrome c, and *apo*-myoglobin were acquired using a Thermo LTQ Orbitrap Velos and processed for each spectrum independently. Additionally, 61 spectra of adalimumab were acquired on a Bruker maXis II instrument. For each single scan, the monoisotopic mass was calculated using the MIND method. Because no spectral averaging was performed (in order to simulate real-world LC-MS conditions), the most intense isotope signal did not always match the theoretically most abundant one due to poor ion statistics. A remedial measure was found to accommodate poor ion statistics, and will be discussed in **Section 3.2**. We also calculated the average mass for each scan by a weighted sum of the well-resolved isotope masses and intensities. Results are summarized as histograms showing count (y-axis) *versus* mass deviation expressed in ppm (x-axes) from the ground truth values and the computed average and predicted monoisotopic masses (**Figure 4)**. The scales of the axes are modified for each molecule to maximize the information in the figures.

For the Orbitrap spectra of insulin, cytochrome c, and myoglobin, the observed average protein mass fluctuated in a broad range, over 20 ppm wide (over 100 ppm wide for cytochrome c, as spectral quality in this case was rather poor). For the maXis II spectra of adalimumab, scan-to-scan variation of average mass was less, but still fluctuated in a range about 10 ppm wide. Not only did $M_{Average}$ consistently show significant scan-to-scan variability, but systematic biases between -5 and -150 ppm (the extreme value again occurred for the cytochrome c spectra) were also observed. In all four spectra, it is clear that the average protein mass could not be confidently determined with an accuracy comparable to the instrument specifications (low-ppm mass accuracy for both instruments).

Applying MIND in all four cases, both the accuracy and precision (*i.e.* scan-to-scan variability) were significantly improved compared to the use of the average mass. The mean mass error over the experimental spectra was reduced to less than 5 ppm in all cases, while the variability was consistent to within 2 ppm, except for cytochrome c that fell within a range of 10 ppm. For adalimumab, these results were more than sufficient to confidently conclude that this was the unmodified sequence, with no glycosylation, oxidation, or (significant) deamidation. It is worth noting that although the quality of the cytochrome c spectra was rather poor, we could still predict the monoisotopic mass with acceptable accuracy and precision. The reason is that the MIND prediction relies on the most abundant peak mass that was still of reasonable quality and acceptable signal-to-noise ratio.

3.2. Further refinement: correction for poor ion statistics

The workflow outlined in **Section 2.1** applies to 'perfect' experimental data for which the true most abundant mass in known; however, for the processing of real-world experimental data, the possibility of data imperfections must be taken into account. Using modern high-performance mass spectrometers such as Fourier Transform Ion Cyclotron Resonance (FTICR), Orbitrap instruments or time-of-flight instruments, masses for the (aggregated) isotope peaks can easily be measured with accuracies on the order of only a few ppm or better. However, for low-abundance analytes a problem with poor ion statistics can occur. As a result, it is quite possible that the observed relative isotopic abundances differ significantly from those calculated based on the elemental composition such that a wrong peak is nominated as most abundant peak. This problem is exacerbated in large ions, which exhibit a broad isotope distribution, in which several peaks have theoretical intensities only a few percent below that of the most abundant mass. Indeed, in top-down proteomics, it can be shown that the probability of the experimentally observed most abundant isotope peak not matching the theoretically predicted one, is sufficiently large such that it should not be neglected (*vide infra*). This probability is further increased by the introduction of a small amount of noise, which can also cause a minor distortion of the relative intensities of isotope peaks.

Neither poor ion statistics, nor the presence of a degree of electronic noise will significantly influence the measured *m/z* values for the individual isotope peaks, and thus their masses are still measured with an error of at most a few ppm. As isotope peaks are by definition spaced approximately 1 Da apart, erroneously selecting a peak adjacent to the theoretically most abundant one introduces an error of $10^6/M_{MostAb}$ ppm. In the precursor range between 10 – 100 kDa, typical for top- and middle-down proteomics, this would therefore introduce an error of 10 – 100 ppm, and is thus a much more significant source of error than the mass accuracy of the instrument in a 'naive' MIND implementation, in which the peak for which the highest intensity is observed, is assumed to be the theoretically most abundant one.

By contrast, the observed scan-to-scan variability of the average mass will be much greater than that of the masses measured for the individual isotope peaks, but (particularly for large molecules) will not display any sudden 'jumps' by plus or minus 1 Da (*i.e.* this value will consistently be relatively accurate, if less precise than that of the most abundant isotope peak). It is easy to see that $M_{Average}$ is relatively insensitive to limited random fluctuations in the abundance of different isotope peaks, as these tend to largely cancel each other out. Thus, we want to combine the robust, but imprecise measurement of the average mass with the precise, but less accurate most abundant isotope mass to reliably identify the 'true' most abundant isotope peak, even for a low number of ions and relatively poor ion statistics. The procedure in **Figure 5** illustrates the simple strategy that combines the computed average mass with the observed most abundant mass to in order to denominate the theoretical most abundant mass. Note that the average mass is always consistently higher than the theoretical most abundant mass (assuming naturally occurring isotope abundances); therefore, the difference between both values should lie in a well specified mass range between 0.1 and 1.1 Da. As the candidates for M_{MostAb} are by definition spaced 1 Da apart, this mass range is in most cases sufficient to uniquely identify the theoretical most abundant peak from a single spectrum. In other words, if the difference between the average and most abundant mass is outside the range of 0.1 and 1.2 Da a correction of the most abundant peak is applied as explained in the next paragraph. This correction function was used in the MIND analysis of the protein spectra discussed in **Section 3.1**. on, e.g., the myoglobine data.

A histogram of the number of spectra in which each signal between 16949 and 16954 Da occurred as the base peak in our myoglobin data set is shown in **Figure 5**. It is clear from this figure that, although the signal at approximately 16950.9 (*i.e.* the theoretically most abundant aggregated isotope peak) occurs as the experimentally most intense peak in a majority of the spectra, the experimentally most abundant peak is located 1 or 2 Da away from the theoretical value in nearly 50% of cases. As expected, however, the measured average mass from the single spectra is relatively constant between scans, and only fluctuates in a range of approximately 20 ppm (0.36 Da) wide. As a result, a histogram of the values of ($M_{Average} - M_{MostAb}$) shows clear clusters, around -1.5, -0.5, +0.5, +1.5, and +2.5 Da. As these correspond to selection of $M_{MostAb}+2$, $M_{MostAb}+1$, M_{MostAb}, $M_{MostAb}-1$, and $M_{MostAb}-2$, it is trivial to correct for a discrepancy between the experimentally and theoretically most abundant peak due to poor ion statistics, by selecting the next or previous peaks. The result of this correction is shown in **Figure 5**, were essentially the same value for M_{MostAb} is consistently generated from all 200 spectra. Latter mass values are used as an input for our predictive model.

To evaluate whether the method for accurate most abundant peak selection is robust against very poor ion statistics, a simple simulation study was conducted based on the assumptions of a multinomial model[24]. For a particular ion number (x-axis in bottom-right panel of **Figure 5**) 5000 isotope distribution were simulated based on the theoretical distribution of myoglobin. Our robust selection method is then applied. If the difference of the average mass minus the most abundant mass is situated in the bin between 0.1 and 1.1 Da, we leave the selection of the most abundant mass unchanged. If is lies left or right of this bin, we apply a correction by selecting the previous or next peak, respectively. Since this is a simulation study, we know the exact most abundant mass and can compute the accuracy of our selection method. An accuracy of 100% means perfect selection of the most abundant peak, 0% means complete failure to select the most abundant peak mass. In **Figure 5**, we report the accuracy based on the 5000 simulations for varying ion statistics. It can be seen that the naive approach, *i.e.*, just selecting the most intense peak, even with relatively good ion statistics (10000 ions) fails to correctly select the theoretical most abundant peak mass in many cases. Our correction

method operates perfectly until we have fewer than 1000 ions in the trap. The reason for this breakdown is that the uncertainty on the average mass becomes larger than 1 Da, so that the discrimination of the mass bins becomes unclear.

3.3. In silico validation and effect of post-translational modifications

Encouraged by the good performance of MIND under real-world conditions, we decided to perform a more in-depth test using a data set of 10,000 human protein sequences not used to calibrate the prediction model (see **Materials and Methods**). We also compared the performance of MIND to that of the method developed by Tsay *et al.*[15] (**Figure S3** and **Table S1**). For these 10,000 sequences, we find that MIND predicts the monoisotopic value with an accuracy better than 0.5 ppm in 66.5% of cases. The probability for the 0 Dalton error was always the largest for the 10,000 sequences, therefore, 31.9% of cases, an error of ±1 Da is made. This error corresponds to the percentages observed when constructing the model, as can be seen from the histogram in **Figure 1d** (see Materials and Methods). It should be noted, however, that these percentages represent averaged values for the entire mass range of the model. Upon inspecting the look-up table in **Figure 1d** and Figure 3, one can conclude that the frequencies related to the off-by-one-Da error changes in function of the most abundant mass. However, it is clear that the 0 Da bin represents the majority case and the -1 and +1 bins describe the tail of the distribution. Nevertheless, for larger molecules the difference in these frequencies becomes smaller. Therefore, the lookup table can be used to provide more accurate confidence measures for the off-by-one Da error if desired for the three cases A smoothed version of the look-up table can be found in the **Supplementary Information**. For the 66.5% of cases where the ppm error is around zero, we find the median error is 0.008 ppm, with 95% of values in the interval between -0.125 and +0.118 ppm (interval width 0.243 ppm). Meanwhile, for the Tsay method, the mass error is close to 0 ppm in only 50.2% of cases, and an error of -2 Da is far more common than when using MIND (4.6% *versus* 0.6% of cases). For those values close to zero, 95% of values occur in a range 0.275 ppm wide, similar (albeit slightly wider) to MIND. In contrast to MIND, however, we find that there is a small, but systematic bias in the results using the Tsay method, as the median error is +0.100 ppm rather than +0.008 ppm (**Supporting Figure S3**).

As mentioned before, the main advantage of knowledge of the intact protein mass is the ease with which different proteoforms can be detected. One of the most common ways in which different proteoforms can arise, is the occurrence of post-translational modifications (PTMs). It is therefore worthwhile to consider to which extent the monoisotopic masses predicted by the MIND algorithm are affected by the presence of one or more PTMs, as the model was developed considering only unmodified protein sequences. We therefore investigated the effect of the ten most commonly occurring post-translational modifications[16], by introducing a single instance of these modifications in each of the sequences in our validation sets, and predicting the monoisotopic masses with MIND as well as the Tsay method. As before, we then again determined the proportion of sequences for which the prediction was 'close to' the actual monoisotopic mass (*i.e.* not off by *ca.* 1 or 2 Da), the median accuracy within this 'near 0 Da' cluster, and the interval width containing 95% of the sequences in this cluster. Do note that we are using the MIND method that was trained on unmodified protein sequences. Nevertheless, for 9 of the 10 PTMs, the effect on each of these PTMs on the performance characteristics was negligible for both the MIND and Tsay methods. The exception to this is ubiquitination, which resulted in a somewhat decreased size and width of the 'near 0 Da' cluster, and an increased median deviation (**Supporting Table S1**). As this effect occurs using both methods, we

conclude that the ease of use and the benefits of MIND over existing methods for predicting the monoisotopic mass are consistent and independent of PTM state.

4. Conclusion

As the use of top-down mass spectrometry methods for the identification and quantification of known and unknown proteoforms becomes more widespread, the demand for methods to accurately and reliably determine precursor mass will only increase. Due to the low probability of occurrence of the monoisotopic variant of an intact protein, as well as significant fluctuation of the average mass due to natural and technical causes, it is unlikely that this solution will be provided by improvements in mass spectrometry technology alone. In this work, we have described a simple yet powerful computational method that relates the observed most abundant isotopic variant mass to the true monoisotopic mass, showing low-ppm precision and excellent mass accuracy compared to the average mass. Undoubtedly, this method will prove to be of value to the ever-growing top-down proteomics community and is made available as a shiny app in the R software framework.

Acknowledgements

The authors thank the FWO and the Polish Academy of Sciences for facilitating this collaboration via the project "Computational methods for high-resolution mass spectrometry data and massive parallel sequencing" (project number VS.028.19N). F.L. is grateful to the Engineering and Physical Sciences Research Council (EP/ N033191/1) for postdoctoral funding. P.B.O. acknowledges the Horizon 2020 grant: EU_FT-ICR_MS Network, project ID 731077. P.D. was supported by START fellowship by the Foundation for Polish Science. A.G. acknowledges Polish National Science Center grant 2018/29/B/ST6/00681. We thank the reviewers for their valuable comments.

Supporting Information

- Figure S1: Simulated isotope distribution of equine apo-myoglobin.
- Figure S2: Residuals of model 1 in function of Δ_i.
- Figure S3: Comparison fo MIND with the mathd of Tsay *et al.* (2013).
- Table S1: Result on the *in silico* analysis of the capabilities of MIND to account for post-translational modification.
- Table S2: Coefficient of the MIND model.
- Look-up table to assign confidence score whether the result fits into +0 cases (majority) or ±1 cases (tail).

Figures

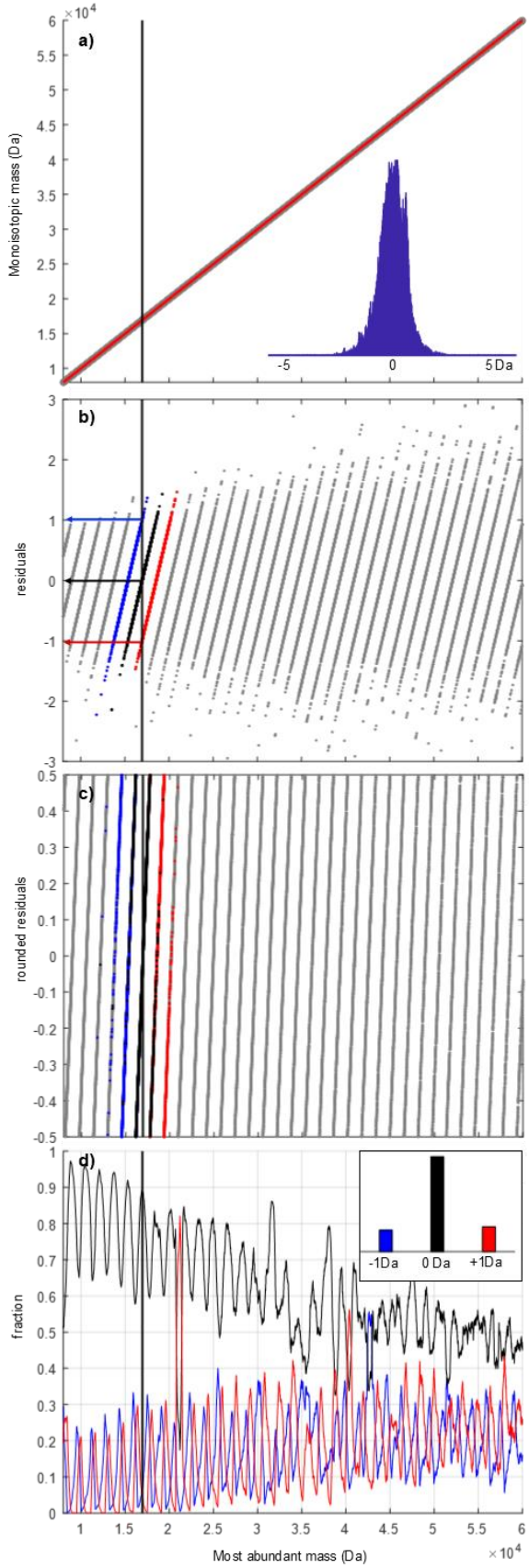

Figure 1: (a) approximately linear correlation observed between monoisotopic and most abundant mass of human proteins in the training data set (*ca.* 68000 human protein sequences). The histogram

(inset) shows the deviation from this simple linear model across the entire set. (b) Plot of the residuals (ε) *versus* M_{MostAb}, revealing a structure in this deviation, with the fractional part of ε (ε_{frac}) shown in Panel (c). Panel (d) shows periodicity in the nearest integer value to ε, allowing prediction of the integer part of ε (ε_{int}). The inset in Panel (D) shows the frequency of ε_{int} values of -1, 0, and +1 Da, across all 68000 sequences in the training set. The vertical line across all four panels is located at 17 kDa, illustrating the case of human *apo*-myoglobin.

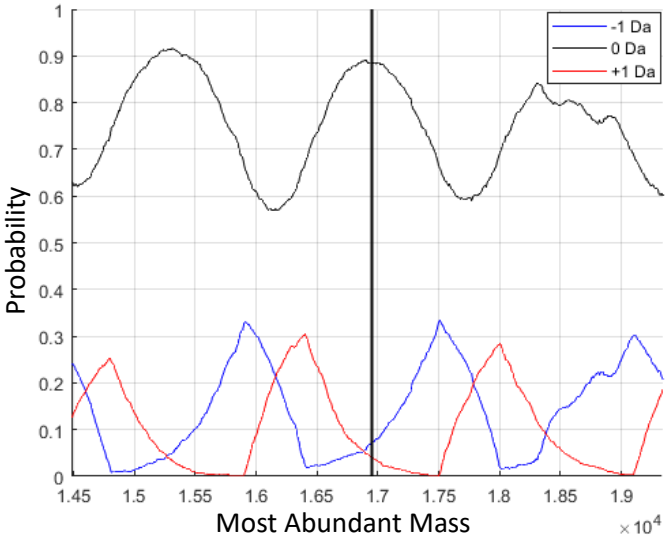

Figure 2: Close-up of Panel d) from Figure 1. In this plot it can be clearly seen that the confidence score related to predicted monoisotopic mass will fit in the majority case (i.e., 0 Da) or tail cases (-1, 1 Da) changes in function of the most abundant mass.

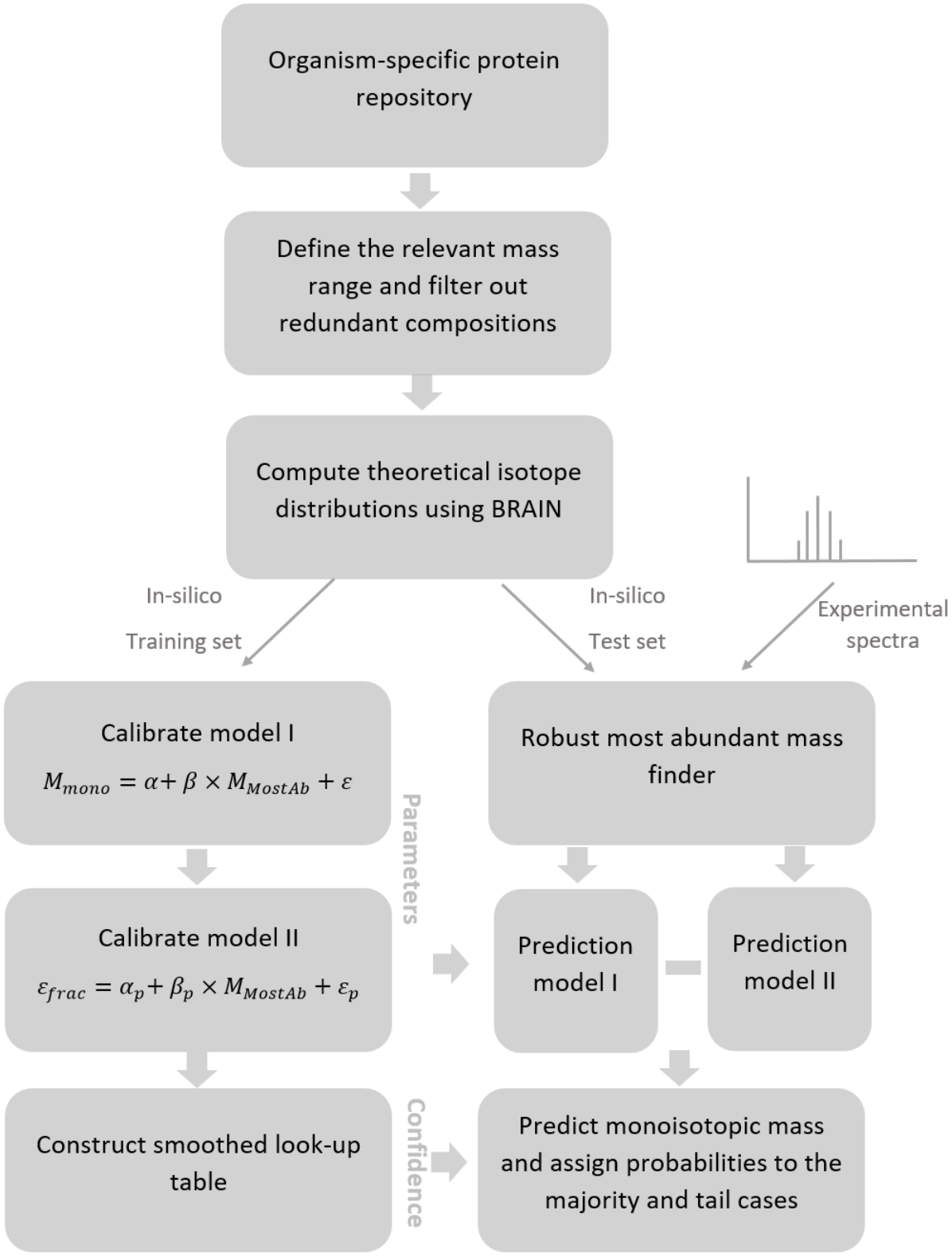

Figure 3: Flow chart illustrating the main steps in the MIND methodology and *in silico* validation.

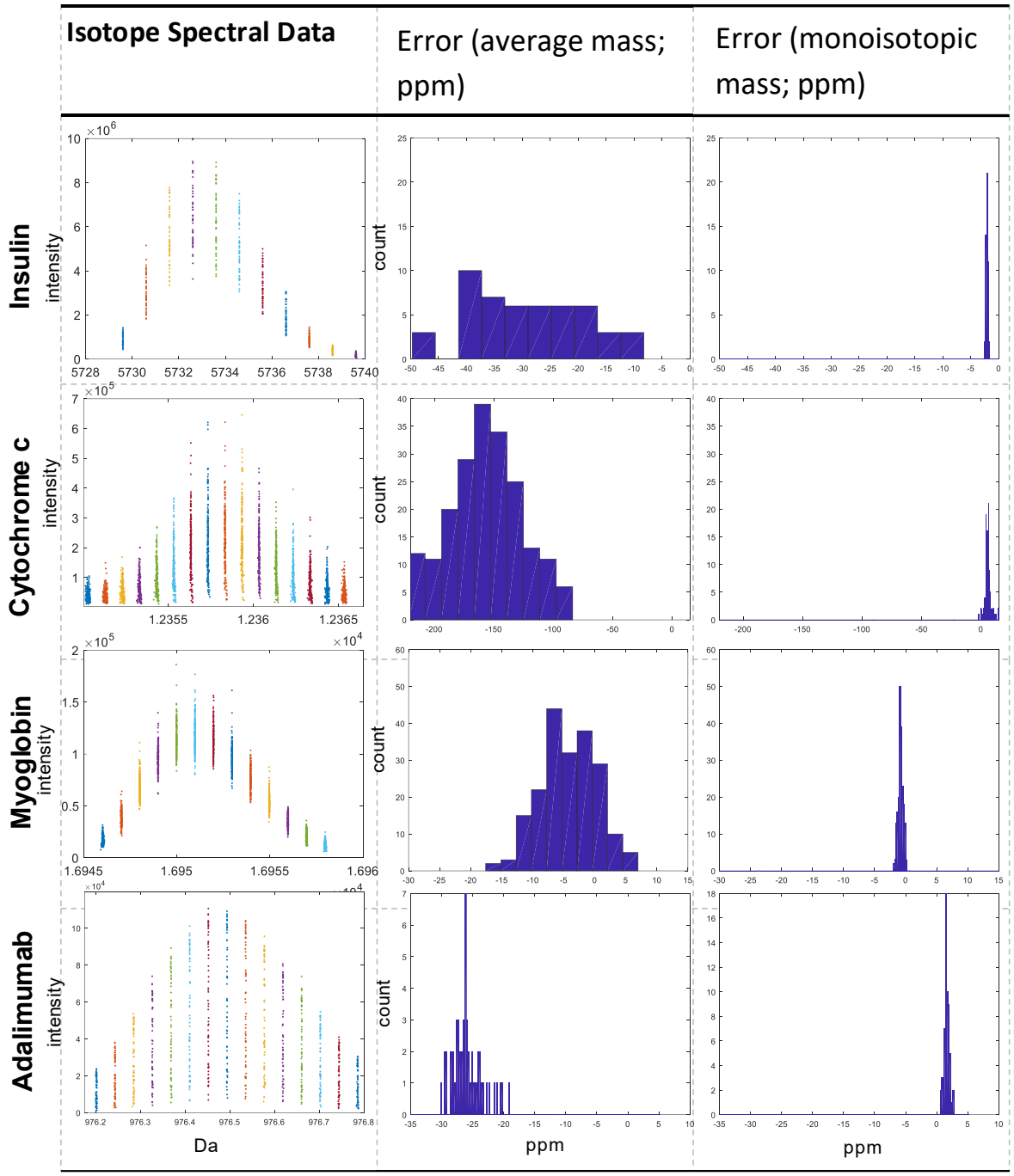

Figure 4: Summary of the results from spectra of bovine insulin, equine cytochrome c, equine apo-myoglobin (all acquired using an LTQ Orbitrap Velos instrument), and light chain adalimumab (on a maXis II ETD). The first column shows the individual isotopically resolved signals that were detected (50 individual spectra for the top three rows, and 61 for adalimumab); the second column shows histograms of the deviation of the experimental *vs.* actual $M_{Average}$, while the third column shows the deviation of M_{mono}, calculated using the MIND algorithm. The probabilities for the -1, 0, and 1 Da error are Insulin = [0.2899, **0.5208**, 0.1892], Cytochrome c = [0.0854, **0.9128**, <0.001], Myoglobin = [0.0732, **0.8844**, 0.0405], Adalimumab = [0.1265, **0.8102**, 0.0617]. Notice that Insulin is outside the mass range for which the model was trained (8-60 kDa), but still gives good results when extrapolating.

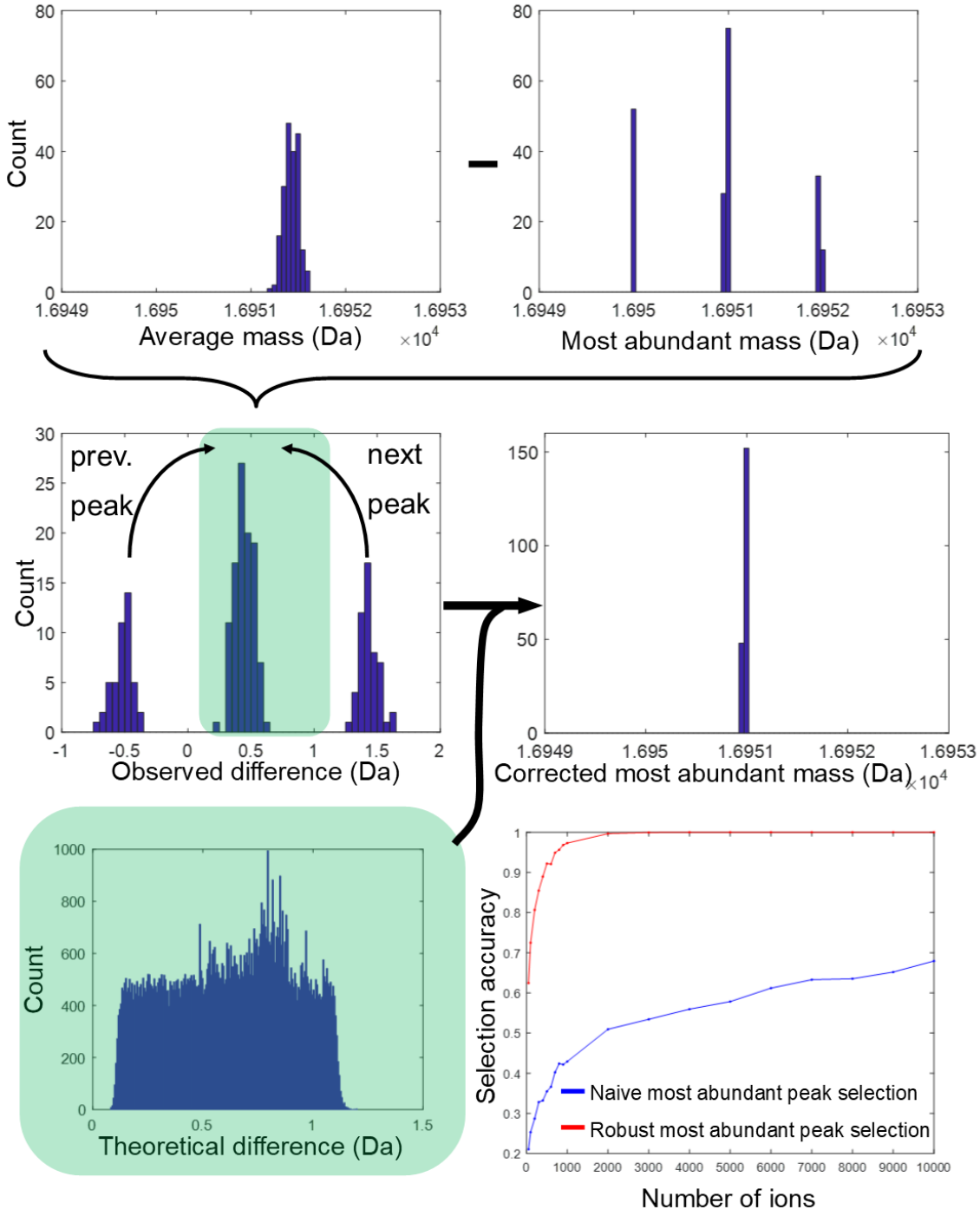

Figure 5: Workflow to find the 'true' most abundant mass in case of poor ion statistics. Subtracting the 'true' most abundant mass from the average mass, always (for the *ca.* 78000 protein sequences in our training and validation data sets) yields a value between 0.1 and 1.2 Da. Therefore, in the vast majority of cases, when this value is greater than 1 Da, the signal to the left of the 'true' most abundant isotope peak was inadvertently selected, and the next peak in the series needs to be used as input for the MIND algorithm. Similarly, a value below 0 indicates that the selected peak is too heavy, and the previous peak in the isotope distribution needs to be chosen. In terms of concrete implementation, this is equivalent to simply taking the floor function of [$M_{Average} - M_{MostAb}$ ('naively' based on the most intense observed signal)], and adding the resulting integer (-1, 0, or 1) to the index of the most intense peak.

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) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551-3567.
(3) Yates, J. R., 3rd. Pivotal role of computers and software in mass spectrometry - SEQUEST and 20 years of tandem MS database searching. *J Am Soc Mass Spectrom* **2015**, *26*, 1804-1813.
(4) Kelleher, N. L.; Lin, H. Y.; Valaskovic, G. A.; Aaserud, D. J.; Fridriksson, E. K.; McLafferty, F. W. Top Down versus Bottom Up Protein Characterization by Tandem High-Resolution Mass Spectrometry. *J Am Chem Soc* **1999**, *121*, 806-812.
(5) Loo, J. A.; Edmonds, C. G.; Smith, R. D. Primary sequence information from intact proteins by electrospray ionization tandem mass spectrometry. *Science* **1990**, *248*, 201-204.
(6) Dang, X.; Scotcher, J.; Wu, S.; Chu, R. K.; Tolic, N.; Ntai, I.; Thomas, P. M.; Fellers, R. T.; Early, B. P.; Zheng, Y.; Durbin, K. R.; Leduc, R. D.; Wolff, J. J.; Thompson, C. J.; Pan, J.; Han, J.; Shaw, J. B.; Salisbury, J. P.; Easterling, M.; Borchers, C. H., et al. The first pilot project of the consortium for top-down proteomics: a status report. *Proteomics* **2014**, *14*, 1130-1140.
(7) Doerr, A. Top-down mass spectrometry. *Nat Methods* **2008**, *5*, 1.
(8) Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L. Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **2011**, *480*, 254-258.
(9) Lermyte, F.; Tsybin, Y. O.; O'Connor, P. B.; Loo, J. A. Top or Middle? Up or Down? Toward a Standard Lexicon for Protein Top-Down and Allied Mass Spectrometry Approaches. *J Am Soc Mass Spectrom* **2019**, *30*, 1149-1157.
(10) Smith, L. M.; Kelleher, N. L.; Proteomics, C. T. D. Proteoform: a single term describing protein complexity. *Nat Methods* **2013**, *10*, 186-187.
(11) Korangy, F.; Julin, D. A. Enzymatic effects of a lysine-to-glutamine mutation in the ATP-binding consensus sequence in the RecD subunit of the RecBCD enzyme from Escherichia coli. *J Biol Chem* **1992**, *267*, 1733-1740.
(12) Claesen, J.; Dittwald, P.; Burzykowski, T.; Valkenborg, D. An efficient method to calculate the aggregated isotopic distribution and exact center-masses. *J Am Soc Mass Spectrom* **2012**, *23*, 753-763.
(13) Valkenborg, D.; Mertens, I.; Lemiere, F.; Witters, E.; Burzykowski, T. The isotopic distribution conundrum. *Mass Spectrom Rev* **2012**, *31*, 96-109.
(14) Claesen, J.; Lermyte, F.; Sobott, F.; Burzykowski, T.; Valkenborg, D. Differences in the Elemental Isotope Definition May Lead to Errors in Modern Mass-Spectrometry-Based Proteomics. *Anal Chem* **2015**, *87*, 10747-10754.
(15) Chen, Y. F.; Chang, C. A.; Lin, Y. H.; Tsay, Y. G. Determination of accurate protein monoisotopic mass with the most abundant mass measurable using high-resolution mass spectrometry. *Anal Biochem* **2013**, *440*, 108-113.
(16) Khoury, G. A.; Baliban, R. C.; Floudas, C. A. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci Rep* **2011**, *1*.
(17) Dittwald, P.; Claesen, J.; Burzykowski, T.; Valkenborg, D.; Gambin, A. BRAIN: a universal tool for high-throughput calculations of the isotopic distribution for mass spectrometry. *Anal Chem* **2013**, *85*, 1991-1994.
(18) Dittwald, P.; Valkenborg, D. BRAIN 2.0: time and memory complexity improvements in the algorithm for calculating the isotope distribution. *J Am Soc Mass Spectrom* **2014**, *25*, 588-594.
(19) Searle, S. R. Parallel lines in residual plots. *Am Stat* **1988**, *42*, 211-211.

(20) Valkenborg, D.; Assam, P.; Thomas, G.; Krols, L.; Kas, K.; Burzykowski, T. Using a Poisson approximation to predict the isotopic distribution of sulphur-containing peptides in a peptide-centric proteomic approach. *Rapid Commun Mass Spectrom* **2007**, *21*, 3387-3391.
(21) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J Am Soc Mass Spectrom* **2000**, *11*, 320-332.
(22) Senko, M. W.; Beu, S. C.; McLafferty, 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.
(23) Zubarev, R. A.; Demirev, P. A. Isotope depletion of large biomolecules: Implications for molecular mass measurements. *J Am Soc Mass Spectrom* **1998**, *9*, 149-156.
(24) Kaur, P.; O'Connor, P. B. Use of statistical methods for estimation of total number of charges in a mass spectrometry experiment. *Anal Chem* **2004**, *76*, 2756-2762.