Rapid and Accurate Global PTM Discovery (GPTM-D) Using Post-Acquisition Spectral Calibration and Defined Mass Windows

# Abstract

Posttranslational-modifications (PTMs) influence many aspects of protein function in biological processes, and correctly identifying the various protein modifications in biological samples is crucial for understanding proteins. Ways of identifying and localizing PTMs are limited, but emerging techniques in the field of mass spectrometry are becoming available. GPTM-D [Journal of Very Important Results, **1**, 1 (2016)] is a recently developed tool for global identification of PTMs using a single pass database search that is promising. Spectra file calibration prior to applying the tool, and algorithmic improvements in the peptide database search greatly improve the accuracy and efficiency of new PTM identification. We describe the calibration tool developed, and present numerical results that validate the proposed enhancement.

# Introduction

GPTM-d1 is a tool designed to discover new PTMs in samples acquired from tandem mass spectrometry. The GPTM-d workflow follows three steps: 1) An open mass database search2 that provides a spectral match to an unmodified peptide along with a mass difference. 2) A database augmentation step that adds plausible localized PTMs to a database based on the open mass search results. 3) A final, narrow mass, search with the augmented database that statistically confirms the presence of the modified PTMs added in the previous step.

Instrument noise, systemic drift and miscalibration limit the measured mass accuracy in acquired spectra. Increasing mass accuracy for both parent and fragment ions is crucial for peptide identification and PTM localization3. Higher mass accuracy provides more specificity, thereby decreasing the false discovery rate for the same number of correctly identified peptides and PTMs. Multiple calibration strategies that improve the mass accuracy of mass spectrometers have been devised. We extend the software lock mass concept4 to allow for more detailed peak matching, account for additional variables5, and apply the resulting calibration function to all of the acquired spectra.

Figure 1: The green components mark the proposed extensions to the GPTM-D workflow for identifying and localizing PTMs.

We add the additional calibration step to the GPTM-D workflow, see Figure 1. First, a database search is performed on the uncalibrated spectra file in order to provide the calibration mechanism with a (limited) list of peptide spectrum matches. As long as this initial search is able to identify at least a few peptides with high probability, the calibration procedure can be initiated. The calibrated spectra are then searched with a *Comb Search*. Once the comb search is complete, the original GPTM-d workflow is followed to completion. The calibration procedure and the comb search are described in the Algorithm and Experimental Results section.

## Previous work

One useful way of classifying spectra calibration techniques is to place them into three categories:

* **Pre-Calibration** is the standard manufacturer-recommended technique of using a calibration solution to calculate a set of *calibration constants* that are used in subsequent experiments.
* **Real-time Calibration** is an option in commercially available mass spectrometers that allows introducing a mass of an ion with a known value into an ion source together with the sample to be analyzed. A compound such as EEEEE can be present everywhere in the column, and is thus seen in every MS scan.
* **Post-Acquisition Calibration** is a purely computational method of shifting peaks in mass spectra to make them closer to their real value.

Another classification for calibration methods is:

* **Parametric Methods** that estimate some parameters of a calibration function, and then apply the function to every peak
* **Non-parametric Methods** use the peak match data directly without attempting to fit parameters to a function

By nature, Pre-Calibration must be parametric in nature, because the dataset to be calibrated is not available during the creation of the calibration function. Real-time calibration be at least partially non-parametric in order to implement the dependence on the scan number/retention time.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Parametric | Non-Parametric | Both |
| Pre-Calibration | Calibration solution |  |  |
| Real-time |  |  | Chemical Lock Mass |
| Post-Acquisition | Various Calibration Functions, e.g.5 | Software Lock Mass4 | mzCal |

### Parametric Calibration

The FT mass-analyzer Orbitrap records axial ion oscillations in the time domain, and uses a Fast Fourier Transform to extract the frequencies. The frequency and the mass-to-charge ratio are related by6:

where is the field curvature parameter. This relationship suggests a convenient form for the calibration function  
ibration equationslibration mixture. equation e amplitude of the electrostatic field and the trap geometry. s. ration procedu

where is the calibration parameter. Using the same parameter at every scan is a crude approximation because it ignores the dependency of the field curvature on the electrostatic voltage that can be different at every measurement. In practice, different values are used for different voltage ranges, and more involved calibration functions such as

have been proposed5. It is not clear what calibration functions are used internally by commercial mass spectrometers, but the effects of using the wrong parameters is evident.

### Non-Parametric Calibration

Nonparametric calibration functions have been used as well4, to model the dependence of the error on the retention time of a scan.

Figure 3: Systematic dependence of the error on the retention time on the Jurkat dataset

We propose to combine knowledge of the parametric dependence of the error on some variables known at scan time with a nonparametric dependence on the scan time.

# Experimental Procedures

The data analyzed comes from experiments described in 7 and in 8.

# Algorithm and Experimental Results

We propose an iterative calibration process that alternates between peak match extraction and the training and application of a calibration function, see Figure 2.

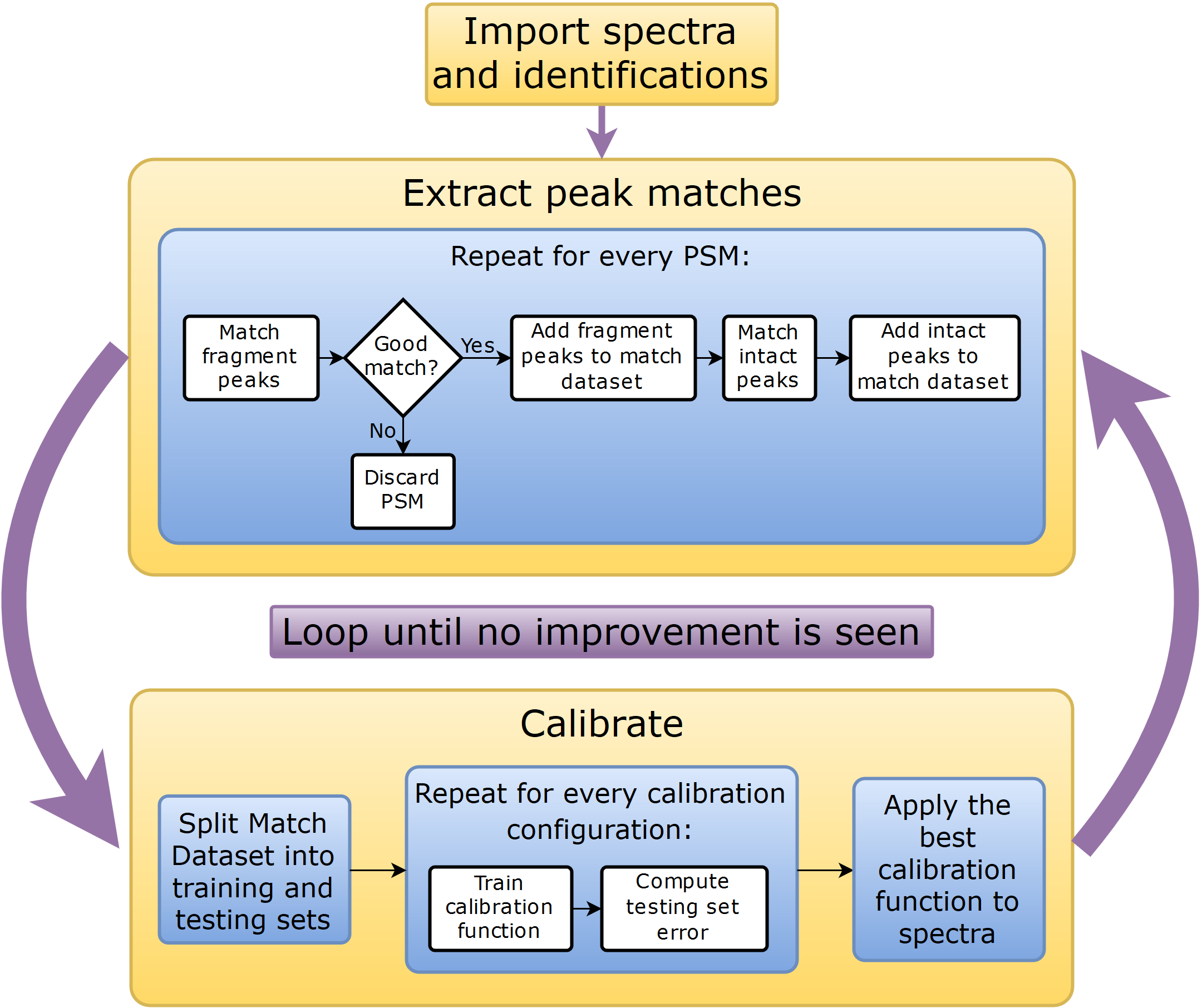
We start the section by providing general results that speak to the efficacy of the calibration, and in the following section we show the significant improvements in PTM discovery. We follow

Figure 2: Calibration process outline

## Theoretical-Experimental Peak Matching

Every peptide sequence identification corresponds to multiple peaks in the spectra. For every identification, the MS/MS scans should include peaks corresponding to the fragment ions of the peptide produced by the dissociation method employed in the mass spectrometer. The neighboring MS scans should have evidence of the un-fragmented peptide, over the elution profile of the peptide. Each of the matches correspond to peaks at different charge states, and different isotopic peaks. All of those have a true mz value, and most of them should have corresponding peaks in the acquired spectra.

For a concrete example, assume that an identification tells us that an MS/MS spectrum corresponds to peptide sequence HVVQSISTQQEKETIAK, identified with a precursor charge 3. Since the sequence contains 17 amino acids, the total number of b and y ions that should be present in the MS/MS spectrum is 32. Each of those ions can have either 1, 2, or 3 charges, so the number of monoisotopic peaks to look for is 96. Every ion-charge state match still corresponds to multiple peaks, since every peptide has an isotope distribution. The number of peaks in the isotope distribution can be large.

## Calibration

### Multiple Calibration Rounds

In order to both include more training points, and to exclude outliers that do not in reality correspond to any theoretical peak, after the initial data point search is finished, we use a simple constant-shift calibration to center our observations around zero. Then a new search of the data is done on the centered points. This helps with making the training set more symmetric with regards to outliers. Specifically, consider spectra that have all errors be of 0.01 m/z units, but we search within 0.02 m/z of zero. The number of outliers that underestimate the error are much greater than ones overestimating it, and therefore building a model based on this data would underestimate the error.

Another reason for doing the constant shifts is to calibrate the

We repeat the constant shift procedure until the number of observed matches between the theoretical and experimental peaks stops increasing.

### Calibration

Once the data is collected, we have training points that correspond to matches between theoretical and experimentally observed peaks.

Instead of pre-selecting the

### Possible Improvements

Different fractions of the same experiment are expected to have overlapping identifications.

Neighboring scans, look for peaks that are repeating.

## Comb Search

Using mass differences between the observed peptide mass and the theoretical mass is an efficient method for identifying post-translational modifications. A popular approach is to consider the MS/MS spectrum in isolation, and perform a database search that identifies

A search of the unimod database reveals that known modifications with mass difference within 200 daltons have values that are within [-0.1, 0.2] of every integer. PTM combinations also have this property. This allows us to ignore

## Calibration Quality

We first demonstrate the improvement in a standard protein database search.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Uncalibrated | mzCal | mzRefinery |
| 10ppm Precursor |  |  |  |
| 3ppm Precursor |  |  |  |
| 1ppm Precursor |  |  |  |

## Notch Search Time Improvement

## Peptide Shaker vs Morpheus

The calibration algorithm requires a list of identifications to work with, and these identifications usually come from a database or a de-novo search.

|  |  |  |
| --- | --- | --- |
|  | Initial Morpheus Search | Initial SearchGUI Search |
| Time for first search |  |  |
| FDR in Calibrated GPTMd Search |  |  |

## Mouse Data

Calibration successfully

### Sulfation and Phosphorylation differentiation

## Jurkat Data

## 11 Cell Lines Data

# Old Text

Physical measurements introduce noise, and instruments that take multiple measurements often have correlated noise between different samples. Knowledge of what some of the measurements should be, paired with an assumption of correlated error in measurement, enables us to make an intelligent guess of the error for the rest of the measurem­­­ents. This is indeed the scenario with mass spectrometry data, where the knowledge of the true mz values for some peaks comes from identified peptide sequences.

The numerical difference between a true, or *reference* value and an observed value is a sum of the *random error* and the *systematic error* of the measurement. The *random error* arises because of some inherent random variability, while the *error due to bias* is a **directed** error in the observed quantity caused by

The measurements’ *bias* (non-random or directed effects caused by a factor or factors unrelated to the independent variable) and error (random variability).

The numerical difference between a true, or *reference* value and an observed value always has a reason. This error can often be at least partially described by observable experimental conditions.

Note that the instrument *resolution* is another important measure of measurement quality, but it is unrelated to the error in an individual measurement.

The goal of the calibration process is to shift each peak in the MS and MS/MS spectra by an appropriate amount, to compensate for as much systemic error as possible. We observe that

a. Only spectra files calibration, no additional search done. Show MSE values for different calibration functions, pick one, and say why it’s better than others

b. A new normal Morpheus run comparison: 10ppm, vs 3ppm (4 runs)

i. Show that precursor mass errors are centered at zero and have smaller variance

ii. Show that there are more PTMs identified we are confident it

iii. Show that decoy PTMs are pushed down the list

Compare with 11 cell line paper, and with Jurkat/Mouse paper

c. Notches

i. Show that now are able to discern between PTMs with similar mass errors (ones that are only different because of the mass defect)

d. Show the improvement in search time due to the new notches option

The work does not look at MS/MS spectra in isolation, but attempts to reconcile the fragmentation patterns with the selected isolation m/z peak.

## Differences with Software Lock Mass

A recent paper suggests using known identifications to create a two-dimensional model of the error in the measurement. The two variables are the Retention Time and the m/z value of each peak. The model predicts the error in the measured m/z value based on these input variables.

The differences with the work presented here are as follows:

* We do not limit ourselves to two variables, but expand to use other useful information such as observed intensity, injection time and others.
* We separate the scan-wise variables from the individual peak variables (namely m/z value and intensity). This is an important consideration, since peaks that appear in the same scan have identical retention times, thus making the distribution of retention times discrete rather than the continuous m/z distribution of peaks.
* The calibration is done on both MS and MS/MS scans, as opposed to just MS scans.
* They calculate a single mass error value for each peptide, combining multiple peaks from multiple MS scans into a single datapoint. We consider each peak separately.
* They use a mass error value calculated by MaxQuant, we use the difference between the reference and observed peaks as the errors.
* We predict the error in *m/z* values, while they predict the mass errors
* They do not shift any peaks: Instead, they run a new database search with updated values for masses of MS isotope patterns. We shift the peaks: This is different, because some peaks if shifted can become a part of an isotope pattern, or fall out of one, or can create new isotope patterns. None of this can happen with their method.
* We publish our software both as a standalone tool and as a library, distributed along with its source code, in contrast to MaxQuant.

1. GPTM-D.

2. (a) Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P., A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat Biotechnol* **2015,** *33* (7), 743-9; (b) Na, S.; Bandeira, N.; Paek, E., Fast multi-blind modification search through tandem mass spectrometry. *Mol Cell Proteomics* **2012,** *11* (4), M111 010199.

3. Scherl, A.; Shaffer, S. A.; Taylor, G. K.; Hernandez, P.; Appel, R. D.; Binz, P. A.; Goodlett, D. R., On the benefits of acquiring peptide fragment ions at high measured mass accuracy. *J Am Soc Mass Spectrom* **2008,** *19* (6), 891-901.

4. Cox, J.; Michalski, A.; Mann, M., Software lock mass by two-dimensional minimization of peptide mass errors. *J Am Soc Mass Spectrom* **2011,** *22* (8), 1373-80.

5. Gorshkov, M. V.; Good, D. M.; Lyutvinskiy, Y.; Yang, H.; Zubarev, R. A., Calibration function for the Orbitrap FTMS accounting for the space charge effect. *J Am Soc Mass Spectrom* **2010,** *21* (11), 1846-51.

6. Makarov, A., Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal Chem* **2000,** *72* (6), 1156-62.

7. Shortreed, M. R.; Wenger, C. D.; Frey, B. L.; Sheynkman, G. M.; Scalf, M.; Keller, M. P.; Attie, A. D.; Smith, L. M., Global Identification of Protein Post-translational Modifications in a Single-Pass Database Search. *J Proteome Res* **2015,** *14* (11), 4714-20.

8. Cesnik, A. J.; Shortreed, M. R.; Sheynkman, G. M.; Frey, B. L.; Smith, L. M., Human Proteomic Variation Revealed by Combining RNA-Seq Proteogenomics and Global Post-Translational Modification (G-PTM) Search Strategy. *J Proteome Res* **2016,** *15* (3), 800-8.