# Derivation

For a peak match () two values are known:

The real value satisfies the relationship

Here is the electrostatic field constant and is the true frequency of the ion oscillation. The observed value is an output of the application of a calibration function by the mass spectrometer to the observed frequency. The two known possibilities are:

Other calibration functions with different dependency on the frequency might be used, but this is not known, since it is proprietary information of the mass spectrometer manufacturers.

If we assume that the observed and the real ion frequencies are the same, a single peak match is sufficient to generate a formula for computing the real peak m/z values from the observed ones. Indeed, assume we have an observed peak with certain m/z value:

The corresponding real m/z value is then computed (assuming the first form of the mass spectrometer calibration function):

This is the first clue for the multiplicative nature of the error in m/z observations.

A natural extension of this model to a dataset collecting multiple peak matches with different ratios of real to observed m/z values is by extracting a single value from all of the ratios. Multiplying any observed peak by would provide an estimate of the real peak value. A model that minimizes the mean squared error  
gives the estimate

The model above assumes that the errors of are… . A more careful analysis can be carried out if we instead assume that the errors of are normally distributed. We would like to minimize the mean squared error

To calculate the prediction, we would like to use

Without parametrizing the frequency, let’s go back to minimizing the difference between the predicted and the real frequency values:

The solution is

To calculate the prediction, we would like to use

Other forms of residual computation can be used, such as Theil-Sen estimator.

Assume there is a dependence of the error of the frequency on the intensity and some measures common to the spectrum:

Let us parametrize the error in the observed frequency instead of the observed mass to charge ratio.

The time domain is inherently unpredictable, so a random forest prediction procedure is used.

If a spline-based calibration is used.

## 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.8 | Software Lock Mass5 | 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 by9:

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 proposed8. 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 well5, 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 6 and in 7.

# 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.

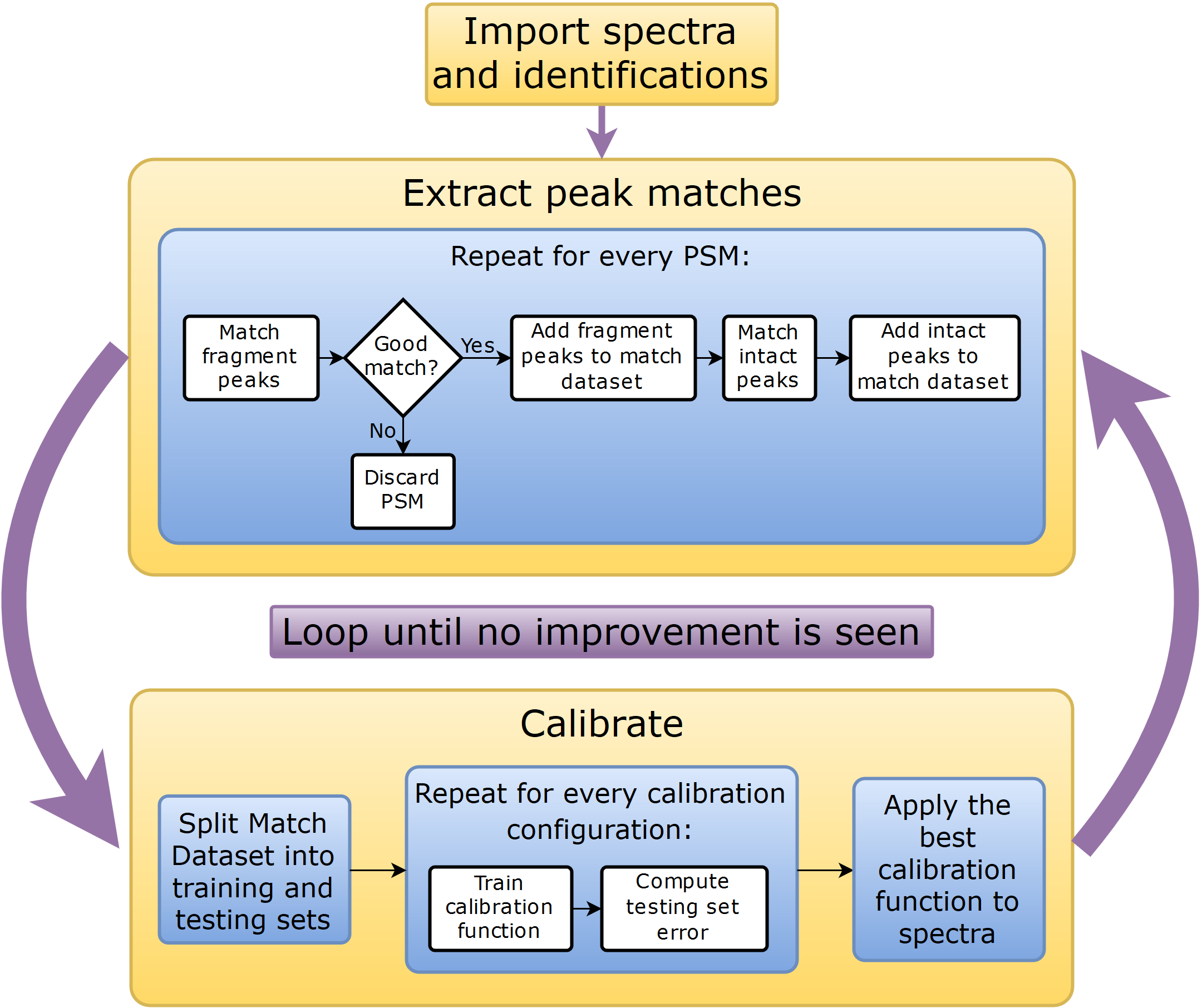
We start the section by providing general results that speak to the efficacy of the calibration, and in the following section we show the significant improvements in PTM discovery. We follow

Figure 4: 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.

# 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

## 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.

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.