Supplementary Methods

Novel PILOT_PTM Algorithm

The framework for PILOT_PTM (Fig. 1a) begins with a preprocessing algorithm which filters the raw spectrum to extract all globally and locally significant peaks based on their intensity (Fig. 1b). The preprocessor is capable of handling inputs from multiple fragmentation methods including Collision Induced Dissociation (CID), Electron Transfer Dissociation (ETD), and Electron Capture Dissociation (ECD) and will label candidate **b**-ion (CID) or **c**-ion peaks (ETD/ECD), the appropriate complementary **y**-ion (CID) or $\mathbf{z}^{\bullet}$-ion (ETD/ECD), and any supporting peaks (isotopes, neutral offsets, etc.) that may exist (Fig. 1c). The ILP model will derive a rank-ordered list of activated peaks for the template amino acid sequence based on one or more sets of candidate ion peaks (Fig. 1d). A complete list of modified sequences that satisfy the appropriate mass conservation constraints for each candidate ion peak set is then constructed. The postprocessor uses a cross-correlation function to mathematically verify the overlap between the experimental MS/MS and the theoretical spectrum created by a candidate modified sequence (Fig. 1e). Each sequence is assigned a cross-correlation score and placed in a rank-ordered list. The modified sequence that best explains the experimental data will have the highest cross-correlation score. Each portion of the algorithm (Fig. 1a) is discussed in further detail below.

Input

Input to the PILOT_PTM algorithm includes (a) the raw tandem mass spectrum (MS/MS), (b) the fragmentation method used, (c) a template amino acid sequence, and (d) a universal modification list. If necessary, the user can provide fragment and parent mass tolerances to override the values generally used by PILOT_PTM. Input from a MS/MS consists of the modified peptide parent mass, its charge state, and a mass ordered list of peak mass-to-charge (*m/z*) ratios and intensities. The MS/MS data can be generated from either CID, ETD, or ECD fragmentation patterns. To retain the precision required to analyze isotopic peaks, both modification masses and amino acid masses are monoisotopic. Modification entries in the Delta Mass database were converted from an integer value to the corresponding monoisotopic mass value [1]. If necessary, additional modifications may be added the universal list by locating the appropriate amino acid/terminus and supplying the modification type and monoisotopic mass.

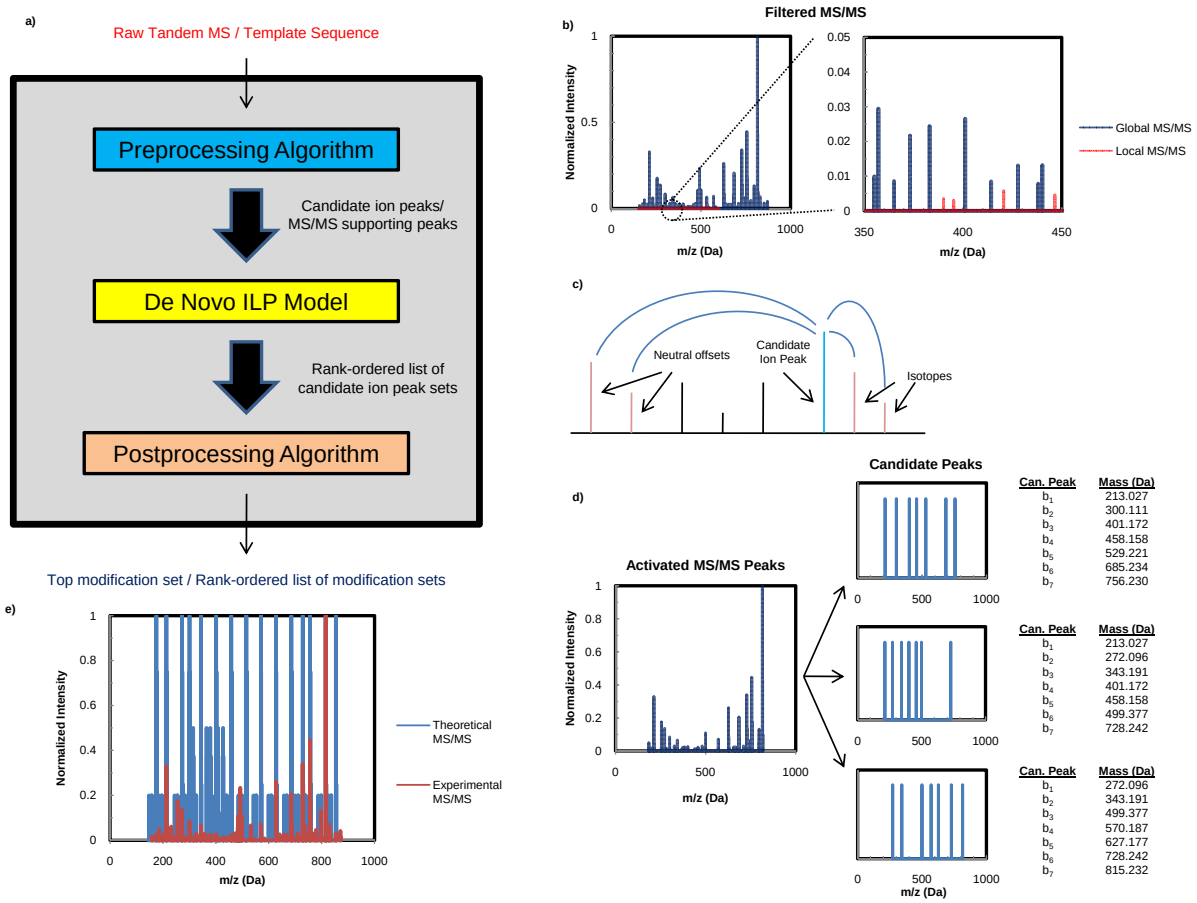

Figure 1: Description of the novel PILOT_PTM method. bf (a) Overall framework for PILOT_PTM. **(b)** Identification of globally and locally significant peaks. The highest-intensity filtered peaks are labeled as globally significant. Any other filtered peak is labeled as locally significant if the peak intensity is greater than all other peaks within a 5.0 Da mass window. **(c)** A set of singly charged support peaks (red) for a candidate ion peak (blue). **(d)** Output from an optimal solution of PILOT_PTM. The dark blue peaks on the left represent all MS/MS peaks that are activated for the optimal solution. The sets of blue peaks on the right represent all candidate ion peak sets that will activate the optimal combination of MS/MS peaks. **(e)** An example of cross-correlation showing the overlap of the model prediction (red) with that from a theoretical fragmentation (blue).

De Novo ILP Model

In this section, we discuss the integer linear optimization (ILP) model formulation for the identification of PTMs, with detailed descriptions of the input, parameters, sets, binary variables, constraints, and the objective function.

Indices & Parameters

The indices used throughout the ILP model are:

$$\begin{aligned} d &: \text{Modification index} \\ i &: \text{MS/MS peak index} \\ j &: \text{Candidate ion peak index} \\ k &: \text{Template amino acid sequence index} \end{aligned}$$

The following monoisotopic mass parameters have preset values.

$$\begin{aligned} m_P^{unmod} &= \text{Mass of the unmodified parent peptide} \\ m_k^{AA} &= \text{Mass of the amino acid located at template position } k \\ m_d &= \text{Mass of modification } d \\ m_k^{Min} &= \text{Minimum modification mass for template position } k \\ m_k^{Max} &= \text{Maximum modification mass for template position } k \\ m_{BC_1} &= \text{Mass of N-terminal template boundary condition} \\ m_{H^+} &= \text{Mass of a proton} \\ m_H &= \text{Mass of a hydrogen atom} \\ m_O &= \text{Mass of an oxygen atom} \\ m_N &= \text{Mass of a nitrogen atom} \end{aligned}$$

The parameters generated from the MS/MS are:

$$\begin{aligned} m_P &= \text{Mass of the modified parent peptide} \\ m_j &= \text{Mass of candidate ion peak } j \\ m_{BC_2} &= \text{Mass of C-terminal template boundary condition} \\ I_i &= \text{Intensity of MS/MS peak } i \end{aligned}$$

Note that the C-terminal boundary condition must be generated from the MS/MS because it is calculated based on the parent mass. The tolerance parameters used are:

$$\begin{aligned} tol_{fragment} &= \text{Fragment ion tolerance for MS/MS peaks} \\ tol_{parent} &= \text{Tolerance for error in parent mass} \end{aligned}$$

The parent ion tolerance, tol_{parent}, is generally set to 1.0 Da and the fragment ion tolerance, $tol_{fragment}$, is generally set to 0.5 Da. This represents the fact that PILOT_PTM is instrument independent and thus capable of handling spectra regardless of the accuracy of the instrument. The fragment tolerance is reduced by the preprocessor if either (1) high-quality ECD spectra is used, (2) the values are overridden by the user, or (3) the mass errors for isotopic offsets are small, indicating the use of a more accurate instrument. When the average isotopic offset in the tandem mass spectrum is less than 0.1 Da, then the fragment tolerance is set to 0.1 Da. The parent tolerance is altered by the preprocessor only if the value is overridden by the user.

Sets

Given a template amino acid sequence of length K, each amino acid is assigned an index, k, corresponding to the position in the template sequence. Without loss of generality, the N-terminal amino acid will correspond to k = 1 and the C-terminal amino acid to $k = K$. The set Mod_k is defined in Equation 1:

$$Mod_k = \{d : d \text{ is a known modification for amino acid position } k\} \quad \forall k \tag{1}$$

Mod_k will vary to represent the alternative modifications (defined from the universal list) that can be present on distinct amino acids (i.e., acetylation and methylation on K, phosphorylation on T, etc.). We note that Mod_1 and Mod_K will also include N-terminal and C-terminal modifications, respectively. Since we allow the terminal positions to have *both* an amino acid modification and a terminal modification (i.e., N-terminal acetylation), the sets Mod_1 and Mod_K consist of all possible enumerations of at most one amino acid modification and at most one terminal modification.

During the preprocessing stage, a list of peaks is generated that represent possible **b**-ions (for CID) or **c**-ions (for ETD/ECD). The peak masses will correspond to singly-charged ions and the choice of ion type is arbitrary as **y**-ions or $\mathbf{z}^\bullet$-ions could easily be used in the formulation of the problem. Though the set of masses will correspond to a single ion type, no loss of spectral information will occur since the *theoretical* **b**-ion or **c**-ion can be computed using a complementary **y**-ion or $\mathbf{z}^\bullet$-ion that is present and would thus validate the assignment. Each element j is called a *candidate ion peak* and the set $Candidate$ (Eqn. 2) is determined by the raw MS/MS and template sequence and is further detailed in the Preprocessing section.

$$Candidate = \{j : j \text{ is a possible } \mathbf{b}\text{-ion (CID) or } \mathbf{c}\text{-ion (ETD/ECD)}\} \tag{2}$$

For each candidate ion peak, we construct the set of supporting MS/MS peaks, $Support_j$ (Eqn. 3). The set $Support_j$ is intended to detail as much information about the candidate ion peak as possible and

is dependent on the fragmentation method used. For ETD/ECD spectra, $Support_j$ consists of the **c**-ion itself, the $\mathbf{c}^{+2}$-ion, $\mathbf{z}^{\bullet}$-ion, $\mathbf{z}^{\bullet+2}$-ion, **b**-ion, **y**-ion, and their corresponding isotopes. For CID spectra, the appropriate ions are the **b**-ion, $\mathbf{b}^{+2}$-ion, **y**-ion, $\mathbf{y}^{+2}$-ion, their isotopic peaks, and their neutral offsets (i.e. $-H_2O$, $-NH_3$, etc.). The **y**-ion or $\mathbf{z}^{\bullet}$-ion series can simply be calculated from the modified parent mass by the formula **c**-ion + $\mathbf{z}^{\bullet}$-ion = $m_P + m_H + 2 \cdot m_{H^+}$ for ETD/ECD spectra and **y**-ion + **b**-ion = $m_P + 2 \cdot m_{H^+}$ for CID spectra. When scanning for supporting offset, complementary, and multiply charged peaks, we use the $tol_{fragment}$ tolerance to determine whether the mass gap between the candidate ion peak and the supporting peak is acceptable. The scan tolerance is reduced to $0.4 \cdot tol_{fragment}$ for isotopic peaks to prevent the incorrect assignment of MS/MS peaks as isotopes. We then define $Mult_i$ as the set of all candidate ion peaks j that use MS/MS peak i as supporting information (Eqn. 4).

$$Support_j = \{i : i \text{ is a supporting MS/MS peak for candidate ion peak } j\} \ \ \forall j \quad (3)$$

$$Mult_i = \{j : i \in Support_j\} \ \ \forall i \quad (4)$$

The next set, CS_k, consists of all candidate ion peaks j that are valid peaks for the template amino acid sequence at position k. It is important to comment on the number of candidate ion peaks j at template position k. Enumeration of all possible j from the set $Candidate$ will result in $|Candidate|$ peaks for each k. This number overexaggerates the number of candidate ion peaks j that are needed for a given template position k. Each (j,k) combination will ultimately correspond to a binary variable in the ILP model, and hence it is desirable to create the smallest set CS_k while retaining all spectral information.

Bounding the Template Positions. Given the universal list of modifications, the theoretical bounds on the masses of the candidate ion peaks j used to construct the modified sequence can be easily calculated. We begin by specifying the N-terminal (BC_1) and C-terminal (BC_2) boundary conditions given the fragmentation method. For the **b**-ion series obtained in CID spectra, $m_{BC_1} = m_{H^+}$ and $m_{BC_2} = m_P - m_O - 2\ cdot m_H + m_{H^+}$. For **c**-ions in ETD/ECD spectra, $m_{BC_1} = m_N + 3 \cdot m_H + m_{H^+}$ and $m_{BC_2} = m_P + m_N - m_O + m_H + m_{H^+}$. Starting from BC_1, the template amino acid mass m_k^{AA} along with the smallest possible modification mass m_k^{Min} and the largest possible modification mass m_k^{Max} are used to find the lower and upper bounds, respectively, on the possible mass region for a candidate ion peak j at the first template position (k = 1). This is done recursively on these upper and lower bounds to find rigorous bounds on the range of a candidate ion peak j at subsequent template positions k. The same can be done when starting from the opposite boundary condition, BC_2, to compute these bounds in the reverse direction. Valid bounds on the masses of the candidate ion peaks j for a given template position k are then the tightest lower and upper bounds from these two sets of bounds. General formulae for computing the valid

lower m_k^L and upper bounds m_k^U for the mass of a candidate ion peak j as a function of template position k are given in Equations 5 and 6, respectively. The tol_{parent} factor is added (Eqn. 6) to account for potential error in the parent mass. Though this widens the possible boundary region for all template positions k, the increase is generally small with respect to the unadjusted width.

$$m_k^L = max\left\{m_{BC_1} + \sum_{k' \leq k}(m_{k'}^{AA} + m_{k'}^{Min}),\ m_{BC_2} - tol_{parent} - \sum_{k' \geq k+1}(m_{k'}^{AA} + m_{k'}^{Max})\right\}\ \forall k \quad (5)$$

$$m_k^U = min\left\{m_{BC_1} + \sum_{k' \leq k}(m_{k'}^{AA} + m_{k'}^{Max}),\ m_{BC_2} + tol_{parent} - \sum_{k' \geq k+1}(m_{k'}^{AA} + m_{k'}^{Min})\right\}\ \forall k \quad (6)$$

Determining Valid Paths. Even though the above equations give the tightest possible bounds for a given k, it is still inefficient to include all candidate ion peaks j that satisfy $m_k^L \leq m_j \leq m_k^U$ in the set CS_k. A list of *valid* peaks for the modified sequence is derived to eliminate peaks that are infeasible. For a given template position k, we are only interested in candidate ion peaks j belonging to at least one possible modified sequence which starts at an boundary BC_1 and ends at boundary BC_2 [2]. Thus, for any candidate ion peak j with $m_k^L \leq m_j \leq m_k^U$, j can only exist in the modified sequence if the mass difference between the candidate ion peak j and the boundary conditions is equal to any combination of the appropriate template amino acid masses plus their corresponding modifications. In other words, if we cannot get to the boundary conditions of BC_1 and BC_2 from candidate ion peak j at template position k by using the weight of a combination of template amino acids (modified or unmodified), then this peak cannot exist at template position k.

The set CS_k can be efficiently constructed by enumerating all possible amino acid paths in the forward and backward direction. We first define the set of all possible unmodified "jumps" from candidate ion peak j at template position k to candidate ion peak j' at template position k' ($k' > k$). This set, defined as J^{unmod}, is given in Equation 7.

$$J^{unmod} = \{(j,k,j',k')\ :\ k' > k,\ |m_{j'} - m_j - \sum_{k''=k+1}^{k''=k'} m_{k''}^{AA}| < tol_{fragment}\} \quad (7)$$

To incorporate the boundary conditions in a jump, we define two dummy candidate ion peaks for BC_1 and BC_2 that exist at $k = 0$ and $k = K$, respectively. Note that these dummy peaks will be the only candidate ion peaks that can exist at the given values of k. We then construct the sets J_1^{mod} and J_2^{mod} to represent the set of all modified jumps from candidate peaks j to peak j' that are separated by one or two template positions, respectively. For any $d \in Mod_k$, we define $m_{d,k}^{AA,mod} = m_k^{AA} + m_d$ and represent J_1^{mod} and J_2^{mod} by

Equations 8 and 9, respectively.

$$J_1^{mod} = \{(j,k,j',k+1) : \exists d \in Mod_{k+1}\, s.t. |m_{j'} - m_j - m_{d,k+1}^{AA,mod}| < tol_{fragment}\} \quad (8)$$

$$J_2^{mod} = \left\{ (j,k,j',k+2) : \begin{array}{c} \exists d \in Mod_{k+1}, d' \in Mod_{k+2}\, s.t. \\ |m_{j'} - m_j - m_{d,k+1}^{AA,mod} - m_{d',k+2}^{AA,mod}| < tol_{fragment} \end{array} \right\} \quad (9)$$

To reduce erroneous PTM assignment, we do not allow a modified amino acid connection between candidate ion peaks that are at least three template positions apart without the existence of any additional candidate ion peaks in between. The set of all possible jumps J is then defined as the union of the sets J^{unmod}, J_1^{mod}, and J_2^{mod} (Eqn. 10).

$$J = J^{unmod} \cup J_1^{mod} \cup J_2^{mod} \quad (10)$$

We now define the set of all forward paths between the start peak BC_1 and a candidate ion peak j at template position k as $P_{j,k}^{BC_1}$ whose elements p_f consist of a consecutive, non-overlapping set of jumps beginning with boundary BC_1 and ending with candidate ion peak j at position k (Eqn. 11). The set of all reverse paths from boundary BC_2 is defined similarly as $P_{j,k}^{BC_2}$ with elements p_r given by Equation 12. Note that each jump in a forward or reverse element must be a member of J.

$$p_f = \{(BC_1,0,j^1,k^1),(j^1,k^1,j^2,k^2),...,(j^{n-2},k^{n-2},j^{n-1},k^{n-1}),(j^{n-1},k^{n-1},j,k)\} \quad (11)$$

$$p_r = \{(j,k,j^1,k^1),(j^1,k^1,j^2,k^2),...,(j^{n-2},k^{n-2},j^{n-1},k^{n-1}),(j^{n-1},k^{n-1},BC_2,K)\} \quad (12)$$

The set CS_k is then defined as the set of candidate ion peaks j within the bounds defined above for which there exists at least one forward and reverse path (Eqn. 13). Note that CS_k is not defined for $k = K$ since this position will correspond to the boundary BC_2.

$$CS_k = \{j : j \in [m_k^L, m_k^U], \exists p_f \in P_{j,k}^{BC_1}, \exists p_r \in P_{j,k}^{BC_2}\} \;\; \forall\, 1 \leq k < K \quad (13)$$

Once the CS_k are constructed, we formulate Pos_j, which will simply give a list of template positions k where each candidate ion peak j may be found (Eqn. 14).

$$Pos_j = \{k : j \in CS_k\} \;\; \forall j \quad (14)$$

Binary Variables

We use binary variables to model the logical use of a candidate ion peak j at a template position k ($p_{j,k}$) as well as the logical use of a MS/MS peak i as supporting information (y_i). These variables are defined as:

$$p_{j,k} = \begin{cases} 1, \text{ if candidate ion peak } j \text{ is used at template position } k \\ 0, \text{ otherwise} \end{cases} \tag{15}$$

$$y_i = \begin{cases} 1, \text{ if MS/MS peak } i \text{ is used as supporting information} \\ 0, \text{ otherwise} \end{cases} \tag{16}$$

Constraints

The constraints of the problem are chosen to ensure proper use of the logical binary variables. At most one candidate ion peak j is able to be assigned to a template position k (Eqn. 17).

$$\sum_{j \in CS_k} p_{j,k} \leq 1 \quad \forall k \tag{17}$$

Additionally, we allow for missing candidate ion peaks j associated with a template position k. We require that there can be no more than three consecutive missing candidate ion peaks using Equation 18:

$$\sum_{k'=k}^{k'=k+3} \sum_{j \in CS_{k'}} p_{j,k'} \geq 1 \quad \forall\, k < K-2 \tag{18}$$

We can also enforce the constraint that a candidate ion peak j can be used at most once in the construction of a modified sequence (Eqn. 19).

$$\sum_{k \in Pos_j} p_{j,k} \leq 1 \quad \forall j \;\; s.t. \;\; |Pos_j| > 1 \tag{19}$$

Constraints are also introduced to ensure that a MS/MS peak i is used properly as supporting information. A MS/MS peak i can only be activated if at least one of the corresponding candidate ion peaks j in the set $Mult_i$ is activated for any valid template position k in the set Pos_j (Eqn. 20).

$$\sum_{j \in Mult_i} \sum_{k \in Pos_j} p_{j,k} \geq y_i \quad \forall i \tag{20}$$

The next set of constraints is added to ensure that a candidate ion peak j is not activated if the corresponding MS/MS ion peak i is not activated (Eqn. 21).

$$\sum_{k \in Pos_j} p_{j,k} \leq y_i \quad \forall i, j \in Mult_i \tag{21}$$

Objective Function

The objective of the problem is to maximize the intensity of the MS/MS peaks i used to construct the modified sequence, which is represented in Equation 22:

$$\max_{p_{j,k},y_i} \sum_i y_i \cdot I_i \tag{22}$$

ILP Model Summary

The entire ILP model is summarized as follows:

$$\max_{p_{j,k},y_i} \sum_i y_i \cdot I_i$$

$$s.t.$$

$$\sum_{j \in CS_k} p_{j,k} \leq 1 \qquad \forall k$$

$$\sum_{k'=k}^{k'=k+3} \sum_{j \in CS_{k'}} p_{j,k'} \geq 1 \qquad \forall\, k < K-2$$

$$\sum_{k \in Pos_j} p_{j,k} \leq 1 \qquad \forall j \;\; s.t. \;\; |Pos_j| > 1$$

$$\sum_{j \in Mult_i} \sum_{k \in Pos_j} p_{j,k} \geq y_i \qquad \forall i$$

$$\sum_{k \in Pos_j} p_{j,k} \leq y_i \qquad \forall i, j \in Mult_i$$

$$p_{j,k}, y_i = \{0,1\} \qquad \forall i, (j,k)$$

This ILP model can be solved to global optimality using CPLEX [3] to obtain a set of MS/MS peaks that correspond to one or more modified sequences. Using integer cuts [4], a rank-ordered list of the top 10 sets of MS/MS peak variables will be generated.

Cutting Plane Constraints

When incorporating all of the previous constraints, it is still possible to obtain linear programming relaxations that consider a set of $p_{j,k}$ at adjacent template positions that do not correspond to the mass difference of a modified or unmodified amino acid. For each $p_{j,k}$, we determine $Inv^L_{j,k,k'}$ and $Inv^U_{j,k,k'}$, the set of candidate ion peaks j' at template position k' ($k' < k$ and $k' > k$, respectively) such that no jump

exists between j and j' (Eqns. 23 and 24, respectively).

$$Inv^{L}_{j,k,k'} = \{j' : j' \in CS_{k'}, (j',k',j,k) \notin J\} \ \forall\, 1 < k < K, j \in CS_k, k' < k \tag{23}$$

$$Inv^{U}_{j,k,k'} = \{j' : j' \in CS_{k'}, (j,k,j',k') \notin J\} \ \forall\, 1 \leq k < K-1, j \in CS_k, k < k' \tag{24}$$

We prevent the misassignment of any invalid peak combination at adjacent template positions with Equations 25 and 26:

$$p_{j,k} + \sum_{j' \in Inv^{L}_{j,k,k-1}} p_{j',k-1} \leq 1 \qquad \forall\, 1 < k < K, j \in CS_k \tag{25}$$

$$p_{j,k} + \sum_{j' \in Inv^{U}_{j,k,k+1}} p_{j',k+1} \leq 1 \qquad \forall\, 1 \leq k < K-1, j \in CS_k \tag{26}$$

For candidate ion peaks j and j' at template positions k and k', respectively where $|k'-k| > 1$, we would like to prevent an invalid combination only if a candidate peak is not activated at any template position k'' between k and k'. This is illustrated using Equations 27 and 28:

$$p_{j,k} - \sum_{k''=k'+1}^{k''=k-1} \sum_{j' \in CS_{k''}} p_{j',k''} + \sum_{j' \in Inv^{L}_{j,k,k'}} p_{j',k'} \leq 1 \qquad \forall\, 1 < k < K, j \in CS_k, k' < k-1 \tag{27}$$

$$p_{j,k} - \sum_{k''=k+1}^{k''=k'-1} \sum_{j' \in CS_{k''}} p_{j',k''} + \sum_{j' \in Inv^{U}_{j,k,k'}} p_{j',k'} \leq 1 \qquad \forall\, 1 \leq k < K-1, j \in CS_k, k-1 < k' \tag{28}$$

The next set of constraints will be added when the linear relaxation activates candidate ion peaks at adjacent template positions where the mass difference between them is less than the smallest modified amino acid or greater than the largest modified amino acid. Thus, for each candidate ion peak j at template position k, we establish $m^{J,L}_{j,k}$ and $m^{J,U}_{j,k}$ which are the maximum (Eqn. 29) and minimum (Eqn. 30) masses that can be reached from j, respectively.

$$m^{J,U}_{j,k} = m_j + m^{Max}_k + tol_{fragment} \qquad \forall\, 1 \leq k < K-1, j \in CS_k \tag{29}$$

$$m^{J,L}_{j,k} = m_j + m^{Min}_k - tol_{fragment} \qquad \forall\, 1 \leq k < K-1, j \in CS_k \tag{30}$$

All $j' \in CS_{k+1}$ for which $m_{j'} > m^{J,U}_{j,k}$ and $m_{j'} < m^{J,L}_{j,k}$ correspond to candidate ion peaks outside the minimum and maximum possible mass peak boundaries. The improper assignment of peak variables can be prevented by Equations 31 and 32.

$$\sum_{\substack{j' \in CS_k \\ m_{j'} \leq m_j}} p_{j',k} + \sum_{\substack{j' \in CS_{k+1} \\ m_{j'} > m^{J,U}_{j,k}}} p_{j',k} \leq 1 \ \ \forall k < K-1, j \in CS_k \tag{31}$$

$$\sum_{\substack{j' \in CS_k \\ m_{j'} \geq m_j}} p_{j',k} + \sum_{\substack{j' \in CS_{k+1} \\ m_{j'} < m^{J,L}_{j,k}}} p_{j',k} \leq 1 \ \ \forall k < K-1, j \in CS_k \tag{32}$$

The first summation term in each equation allows for the consideration of all $p_{j',k}$ that would also be incorrectly assigned for any $p_{j',k+1}$ variable in the second summation.

Incorporating all of these equations in the initial formulation of the problem results in a large number of constraints, many of which are not activated for the optimal solution. To circumvent this computational burden, we apply them dynamically as cuts. That is, for a given ILP relaxation, the violations of Equations 25, 26, 27, 28, 31, and 32 are checked and cuts are then added when needed.

Candidate modified sequences

Once a solution of the ILP model is found, several candidate modified sequences are generated that must satisfy both (1) conservation of mass between adjacent ion peaks and (2) overall conservation of mass. We initially determine all modifications that can exist at a template position k that satisfy conservation of mass to within the $tol_{fragment}$ factor. All combinations of modifications along the template sequence are enumerated to form a potential list of PTM sets each with a given total mass. If the total mass of a PTM set is within tol_{parent} of $m_P - m_P^{unmod}$, then the modified sequence is added to the candidate list. If no PTM sets satisfy this total mass constraint, the optimal solution is removed via an integer cut [4] without being added to the rank-ordered list of peak variable sets.

Integer Cuts

Since a MS/MS peak intensity I_i contributes the same amount to the objective function (Eqn. 22) regardless of how it is used as supporting information, multiple sets of candidate ion peaks $p_{j,k}$ may give rise to the same optimal solution (Fig. 1d). After an optimal combination of MS/MS peaks y_i are found, all sets of candidate ion peaks $p_{j,k}$ which give this solution may be found by using an integer cut constraint [4]. We define B_p as the set of all $p_{j,k} = 1$ and NB_p as the set of all $p_{j,k} = 0$. Each set of $p_{j,k}$ may be removed from consideration in the next pass of the ILP solver through Equation 33:

$$\sum_{p_{j,k} \in B_p} p_{j,k} - \sum_{p_{j,k} \in NB_p} p_{j,k} \leq |B_p| - 1 \tag{33}$$

Once all sets of $p_{j,k}$ are found, the optimal solution of y_i is then removed through an integer cut. We define B_y as the set of all $y_i = 1$ and NB_y as the set of all $y_i = 0$ and define the constraint as in Equation 34:

$$\sum_{y_i \in B_y} y_i - \sum_{y_i \in NB_y} y_i \leq |B_y| - 1 \tag{34}$$

$|B_p|$ is the cardinality of B_p and $|B_y|$ is the cardinality of B_y. We utilize integer cuts on the MS/MS variables (Eqn. 34) until a rank-ordered list of 10 sets of peaks is generated.

Preprocessing

Much information is contained in the MS/MS that can be extracted prior to the creation of the ILP framework. Classification of a spectral peak will allow us properly utilize the peak as supporting information for a candidate **b**-ion for CID spectra or **c**-ion for ETD/ECD spectra. To begin, we remove all peaks that are associated with the precursor ion. For CID spectra, this includes the precursor ion, its +1 and +2 isotopes, and any neutral losses (i.e., -H_2O, $-NH_3$, $-CO$) [5]. For ETD/ECD spectra, we must remove all peaks that correspond to distinct charge states of the precursor ion and their isotopes. Additionally, all peaks that correspond to a common neutral loss of a charge reduced form of the precursor ion [6] are removed.

The MS/MS is then filtered to remove any peak that is within 0.4 Da of another peak of higher intensity. For MS/MS where the precursor ion has charge state greater than 3 and the fragment tolerance at most 0.10, this filter tolerance is reduced to 0.25 Da. The filter tolerance is reduced to 0.10 Da for ECD spectra to maintain the extremely high accuracy of the spectral data. The filtered MS/MS is scanned to extract peaks with the highest intensity (Fig. 1b). The quantity of peaks extracted depends on the parent mass and is equal to 125 for peptides of less than 1,000 Da, 175 for peptides between 1,000 and 2,000 Da, and 250 for peptides over 2,000 Da. This globally significant list of peaks comprises the set *Global* and is indexed over i based on decreasing intensity. The preprocessor creates a list of candidate ion peaks, $Candidate$, based upon the experimental data by considering all peaks in *Global* and any locally significant peaks. A peak is considered to be locally significant if the peak intensity is greater than all other peaks within a mass window of 2.0 Da (Fig. 1b). This mass window is reduced to 0.5 Da for ECD spectra.

All peaks in $Candidate$ are then scanned for the existence of isotopes using a tolerance of $0.4 \cdot tol_{fragment}$. Any labeled peaks are then removed from $Candidate$. If any doubly or triply charged peaks are found based on isotopic offsets, the appropriate singly charged peak of the same intensity is constructed and inserted into $Candidate$ if it is not already in the set. All peaks in $Candidate$ are scanned for the existence of a complementary ion within the set. If two ion peaks sum to the mass of the parent peptide ($m_P + 2 \cdot m_{H^+}$ for CID and $m_P + m_H + 2 \cdot m_{H^+}$ for ETD/ECD), they are labeled as complements. If multiple peaks satisfy the mass criteria for complementarity, then the set with the smallest difference from the recorded precursor mass is selected. To account for parent mass error and fragment error, the scan tolerance for complementary ions is $tol_{fragment} + 0.5 \cdot tol_{parent}$. For CID spectra, all neutral offsets within a tolerance of $tol_{fragment}$ are then removed from $Candidate$ if the offset does not have a complementary peak. To ensure the presence of a **b**-ion or **c**-ion, a dummy complement peak is constructed for all remaining peaks in $Candidate$ if the complementary ion not already in the set.

The preprocessor then queries all candidate peaks and determines a full list of supporting peaks (Fig. 1c) that exist in the set *Global* assuming that the candidate peak is singly charged. For CID spectra, this will include +1 and +2 isotopic offsets, neutral losses, and doubly charged peaks. For ETD/ECD spectra, this will include isotopic offsets and doubly charged peaks. If the preprocessor was able to identify doubly charged peaks based on isotopic information, then a peak $i \in Global$ will only be labeled as a doubly charged supporting peak if it has at least one isotope. If necessary, the preprocessor can scan for triply charged peaks as well. Note that the peaks in *Candidate* are merely "reference" peaks. That is, they only represent the *existence* of a candidate ion peak at a given mass. The candidate ion peak itself will only be used for supporting information if it is globally significant and thus contained in the set *Global*.

Postprocessing

A postprocessing algorithm is employed to score the candidate modified peptide sequences that are derived from the peak sets in the ILP rank-ordered list. Each modified sequence is susceptible to experimental error associated with the fragment and precursor *m/z* measurements. To validate a candidate modified sequence, a scoring function must be used to resolve these errors using monoisotopic mass values for the residues and PTMs. A cross-correlation technique is used to measure the mathematical overlap between the theoretical ions produced from the candidate PTM set and the experimental spectrum. Though an idealized model would take into account peptide cleavage chemistry and residue location [7], PILOT_PTM is designed to be instrument independent. A generalized model based on the SEQUEST algorithm [8] is established that is similar to the model used in PILOT [9, 10] and PILOT_SEQUEL [11].

Using a normalized scale, all singly charged **b**-ions and **y**-ions (CID spectra) or **c**-ions and $\mathbf{z}^{\bullet}$-ions (ETD/ECD spectra) are assigned an intensity of 1. The intensity assignment for almost all supporting ions associated with CID fragmentation will be equal to those chosen for the PILOT and PILOT_SEQUEL models [9–11]. These models assigned higher intensity fractions to supporting peaks that were expected to be observed more frequently during fragmentation. Neutral losses were generally assigned an intensity of 1/5, with exceptions given to residues which are more susceptible to loss than others. Neutral loss of water from D, E, S, and T residues and neutral loss of ammonia from Q and N residues [7, 12] tends to be more common than other residues, so these fragments are assigned an intensity of 1/3. Neutral losses to form **a**-ions and **x**-ions are assigned an intensity of 1/5. The authors note that while PILOT_PTM searches for **x**-ions, they were rarely, if ever, found in the annotated spectra. However, their inclusion in the postprocessing section did not impact the results of the overall algorithm.

Similar to PILOT and PILOT_SEQUEL, doubly charged **b**-ions and **y**-ions were generally assigned an intensity of 0 and 1/2, respectively. However, all **b**-ions and **y**-ions that contained 2 or more basic residues were assigned an intensity of 1. The additional basic residues may be present as the result of missed tryptic cleavage and tend to increase the possibility of observing a doubly charged fragment [5]. Additionally, neutral losses from the doubly charged ions with multiple basic residues were assigned the same normalized intensities as with the singly charged ion offsets. Isotopic offsets were searched for all singly and doubly charged **b**-ions and **y**-ions and their neutral offsets. Any +1 isotopes were assigned an intensity equal to 3/4 of the intensity of the base peak and any +2 isotopes were assigned an intensity equal to 1/4 of the base peak.

For ETD and ECD spectra, all **c**-ions and $\mathbf{z}^{\bullet}$-ions are assigned an intensity of 1. All doubly charged **c**-ions and $\mathbf{z}^{\bullet}$-ions were assigned an intensity of 1/2 unless the fragment peak has 2 or more basic residues. In these instances, the doubly charged ions were assigned an intensity of 1. For all isotopic **c**-ion peaks and the $\mathbf{z}^{\bullet}$-ion peaks under 800 Da, a normalized intensity assignment method similar to CID is used. For the $\mathbf{z}^{\bullet}$-ion peaks over 800 Da, the isotopic distribution begins to shift to favor the +1 isotope. Thus, we assign a normalized intensity that is equal to the base peak intensity for the +1 isotopic peak and equal to half the base peak intensity for the +2 isotopic peaks.

Once all peak intensities are assigned, the postprocessor scans each set of candidate ion peaks j output from the ILP model. If the mass difference between two candidate ion peaks j and j' that are at least two template positions apart is equal to the sum of the intermediate unmodified residue masses (within $tol_{fragment}$) but the activated candidate ion peaks in between j and j' indicate a possible modification, then these intermediate candidate ion peak assignments are checked by looking for the presence of peaks in the MS/MS that indicate unmodified residues. If enough supporting information exists, then the intermediate candidate ion peaks are reassigned to that of the unmodified sequence and subsequentally rescored. A mathematical overlap between the theoretical and experimental spectrum is then calculated based on monoisotopic masses for each candidate modified peptide (Fig. 1e). Each candidate modified peptide is assigned a cross-correlation score and inserted into a rank-ordered list. The modified peptide thought to best explain the experimental data is given the highest cross-correlation score.

Output

The output from the PILOT_PTM algorithm will consist of a rank-ordered list of modification sets associated with the template amino acid sequence. The ranking will be based on the score of the cross-correlation

function as defined above in the Postprocessing section. The modification set that best explains the experimental data will correspond to the highest score from this cross-correlation.

Algorithm Parameters

The following section discusses the parameters used for each algorithm for each test set. For each data set, the template amino acid sequence was provided to PILOT_PTM along with the fragmentation method used.

Test Set A - Phosphopeptides

The fragment ion tolerance for all spectra was set to 0.5 Da and the parent mass tolerance was set to 1.0 Da for data sets A1 and A2 and 0.2 Da for data set A3.

Test Set B - Histone H3 1-50 N-Terminal Tail

The fragment tolerance and parent tolerance were both set to 0.01 Da. The parameters for Mascot are as follows: (1) MSDB with human taxonomy, (2) V8-E protease with no missed cleavages, (3) FTMS-ECD instrument, (4) variable modification list: N-Terminal Acetylation, C-Terminal Methylation, K Methylation, K Dimethylation, K Acetylation, K Trimethylation, R Methylation, R Dimethylation, S/T Phosphorylation.

Test Set C - Propionylated Histone Fragments

The fragment tolerance was set to 0.5 Da and the parent tolerance was set to 0.1 Da. The parameters for Mascot, InsPecT, X!Tandem, VEMS, and Mod^i are as follows: (1) NCBInr database with mouse taxonomy and (2) trypsin protease with up to three missed cleavages. Additional parameters include: (1) up to five modification sites for InsPecT and VEMS, (2) marker ions ignored by VEMS, (3) ESI-TRAP instrument for Mascot and InsPecT, LTQ for Mod^i. For Mascot, InsPecT, X!Tandem, and VEMS, the variable modification list was chosen based on the protocol outlined in the manuscript. The universal list of modifications was used for PILOT_PTM and Mod^i. All additional parameters for the compared algorithms were left at the default values.

Test Set D - Total chromatin fraction

The fragment tolerance was set to 0.5 Da and the parent tolerance was set to 0.1 Da. The parameters for InsPecT, X!Tandem, and VEMS are as follows: (1) NCBInr database with human taxonomy and (2) trypsin protease with up to three missed cleavages. Additional parameters include: (1) up to three modification sites for InsPecT (Restricted) and VEMS, one modification site for InsPecT (Unrestricted), (2) marker ions ignored by VEMS, (3) ESI-TRAP instrument for InsPecT, LTQ for Mod^i. For InsPecT (Restricted), X!Tandem, and VEMS, the variable modification list was chosen based on the protocol detailed in the manuscript. The universal list of modifications was used for PILOT_PTM and Mod^i. All additional parameters for the compared algorithms were left at the default values.

Test Set E - Additional Unmodified Peptides

The fragment ion tolerance was set to 0.5 Da for data set E1, 0.2 Da for data set E2, and 0.1 Da for data set E3. The parent mass tolerance was set to be equal to the fragment tolerance for each data set.

References

[1] Mitchelhill, K. Delta Mass: A Database of Protein Post Translational Modifications. http://www.abrf.org/index.cfm/dm.home

[2] Hubler, S. L., Jue, A., Keith, J., McAlister, G. C., Craciun, G. and Coon, J. J. (2008) Valence Parity Renders z$^\bullet$-Type Ions Chemically Distinct. *J. Am. Chem. Soc.* **130**, 6388-6394

[3] CPLEX (2008) *ILOG CPLEX C++ API 11.1 Referece Manual*

[4] Floudas, C. A. (1995) *Nonlinear and Mixed-Integer Optimization*, Oxford University Press, New York

[5] Kinter, M. and Sherman, N. E. (2000) *Protein Sequencing and Identification Using Tandem Mass Spectrometry*, John Wiley & Sons, Inc., New York

[6] Good, D. M., Wenger, C. D., McAlister, G. C., Bai, D. L., Hunt, D. F. and Coon, J. J. (2007) Post-Acquisition ETD Spectral Processing for Increased Peptide Identifications. *J. Am. Soc. Mass. Spectrom.* **20**, 1435–1440

[7] Zhang, Z. (2004) Prediction of Low-Energy Collision-Induced Dissociation Spectra of Peptides. *Anal. Chem.* **76**, 3908–3922

[8] Eng, J. K., McCormack, A. L. and Yates III, J. R. (1994) An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989

[9] DiMaggio Jr., P. A. and Floudas, C. A. (2007) A Mixed-Integer Optimization Framework for De Novo Peptide Identification. *AIChE J.* **53**, 160–173

[10] DiMaggio Jr., P. A. and Floudas, C. A. (2007) De Novo Peptide Identification via Tandem Mass Spectrometry and Integer Linear Optimization. *Anal. Chem.* **79**, 1433–1446

[11] DiMaggio Jr., P. A., Floudas, C. A., Lu, B. and Yates III, J. R. (2008) A Hybrid Method for Peptide Identification Using Integer Linear Optimization, Local Database Search, and Quadrupole Time-of-Flight or OrbiTrap Tandem Mass Spectrometry. *J. Proteome Res.* **7**, 1584–1593

[12] Tabb, D. L., Smith, L. L., Breci, L. A., Wysocki, V. H., Lin, D. and Yates III, J. R. (2003) Statistical Characterization of Ion Trap Tandem Mass Spectra from Doubly Charged Tryptic Peptides. *Anal. Chem.* **75**, 1155–1163