# Optimizing Protein Characterization using Machine Learning-Guided Mass Spectrometry

Alexander Pelletier

Thesis submitted to the University of Ottawa in partial Fulfillment of the requirements for the degree of Masters in Biochemistry with a specialization in Bioinformatics

Department of Biochemistry, Microbiology and Immunology

Faculty of Medicine

University of Ottawa

# Abstract

Mass spectrometry-based proteomics excels at high-throughput identification of proteins expressed in complex biological samples. However, the technology struggles to identify low abundance proteins due to large amounts of redundant data acquired for high abundance proteins with little collected for low abundance proteins. To improve the identification sensitivity of these proteins, I designed a machine learning classifier that assesses protein identification confidence on-the-fly, during mass spectrometry analysis. Proteins deemed confidently identified are excluded from further analysis, saving mass spectrometry resources for lower abundance proteins. Simulating data from a HEK293 cell lysate mass spectrometry analysis, our algorithm uses 16.2% - 66.2% fewer mass spectrometry resources with a 2.6% - 39.5% drop in protein identifications. When applied to live mass spectrometry experiments, these saved resources will likely improve the overall protein identification sensitivity of the experiment, particularly for lower abundance proteins, and will therefore provide a better understanding of the cell's biology.

# Acknowledgements

First and foremost, I would like to acknowledge and thank my supervisors, Dr. Mathieu Lavallée-Adam and Dr. Daniel Figeys. Dr. Lavallée-Adam inspired me to pursue a career in proteomics-related bioinformatics and has encouraged me every step along the way. His insight and technical knowledge was instrumental to the planning and execution of my project. I admire and appreciate his dedication to his students and his research. I am grateful for my discussions with Dr. Figeys, concerning technical aspects related to this project, his insightful career advice, and for reminding myself to always keep in mind the big picture biological impact of my project.

Secondly, I would like to acknowledge the lab members from both my supervisors' labs. I would like to thank Zhibin Ning from the Figeys's lab for his contribution of mass spectrometry data used in my project as well as his discussions on aspects of my project. I would like to thank Nora Wong, a former undergraduate researcher in Dr. Lavallée-Adam's lab, as her summer project formed the foundation on which my project began. I would like to thank Yun-En Chung, a current undergraduate researcher in Dr. Lavallée-Adam's lab for his discussions and contributions to my project as he moves forward with this research. I would like to thank Krystal Walker from the Figeys's lab for allowing me to shadow her proteomics sample preparation procedure so I can gain a deeper understanding and appreciation for the techniques necessary for the big picture of my project. As well, I would like to thank and acknowledge all the other members of these laboratories for their helpful feedback on my project and presentations, continual encouragement, and for helping me with day-to-day technical difficulties.

Thirdly, I would like to acknowledge the members of the community I have been a part of during my time at the University of Ottawa. I am grateful for the partial admissions scholarship from the University of Ottawa and for my scholarship through the NSERC Create in

Technologies for Microbiome Science and Engineering (TECHNOMISE) program. I am thankful for all the members of the TECHNOMISE program whose research and discussions excite me and have helped my professional development as a researcher. I would like to thank members of the Ottawa Institute of Systems Biology and the Biochemistry, Microbiology and Immunology department at the University of Ottawa Faculty of Medicine. I would also like to thank the student organization Let's Talk Science for fostering the confidence and passion when discussing science to both technical and lay audiences.

Finally, I would like to acknowledge that a majority of the research concerning this thesis was conducted on land which is the traditional and unceded territory of the Algonquin Nation. As well, a portion of the writing of my thesis was conducted on land which is the traditional and unceded territory of the Tongva, Chumash, and Kumeyaay Native American people of Southern California.

# Table of Contents

# List of Abbreviations

| | |
|---|---|
| DDA | Data Dependent Acquisition |
| FDR | False Discovery Rate |
| XCorr | Cross-Correlation |
| numDB | Number of peptides identified for a protein in a database |
| MS | Mass Spectrometry |
| MS1 | Mass Spectrometry scan of precursor ions |
| MS2 | Mass Spectrometry scan of fragment ions |
| API | Application Programming Interface |
| ppm | Parts per million |
| PSM | Peptide Spectrum Match |
| m/z | Mass-to-charge ratio |
| LC | Liquid Chromatography |
| MS/MS | Tandem Mass Spectrometry |
| LC-MS/MS | Liquid Chromatography coupled to Tandem Mass Spectrometry |

# List of Figures

# List of Tables

# 1. Introduction

## 1.1 Mass Spectrometry-based Proteomics

Mass spectrometry (MS)-based proteomics has emerged as a state-of-the-art technology for identifying a large number of proteins in complex biological samples. This technology is useful for various clinical and basic research applications, including the ability to characterize the proteome of an organism or cell type,[1–4] elucidating protein-protein interactions,[5–7] and for highlighting protein differential expression between a disease versus non-disease state.[8,9] MS also enables the identification of post-translational modifications[10,11] and the absolute quantification of proteins when used with the appropriate standards.[12] The basic principle behind mass spectrometry is that it measures the mass-to-charge ratio (*m/z*) of analytes. These *m/z* values are then used to derive knowledge about the identity of these analytes.

High-throughput identification of proteins in a biological sample is best achieved by an approach named bottom-up proteomics.[13] In bottom-up proteomics, proteins are extracted from cells of interest and are digested with a restriction enzyme to create smaller peptides. These smaller peptides are more easily amenable to MS analysis, which has poor sensitivity with larger molecules such as proteins.[14] Trypsin is typically used as a restriction enzyme due to the enzyme's high specificity and for the size of peptides it generates, based on the frequency of arginines and lysines in proteomes (the amino acids after which it cleaves), which is ideal for MS analysis.[14,15] After digestion, a large pool of peptides is created, many of which may differ in amino acid sequence, but not in masses. Multiple different peptides analyzed at the same time cannot be differentiated, therefore samples will typically be separated based on some biochemical property before MS. Most approaches will separate peptides based on

hydrophobicity using liquid chromatography (LC) prior to MS analysis.[16,17] Therefore, peptides with similar masses but different levels of hydrophobicity will not be analyzed at the same time in the mass spectrometer.

## 1.2 Tandem Mass Spectrometry

Since two peptides from two different proteins with different amino acid sequence may have the same mass, simply measuring their *m/z* ratio is not sufficient to unravel their identity. Therefore, mass spectrometry analysis is typically performed in tandem (MS/MS).[18] In MS/MS, as peptides elute off the liquid chromatography column, the mass spectrometer measures their *m/z* values, generating an MS spectrum of precursor ions (referred to as MS1 spectrum). The time at which a given peptide elutes off the chromatography column is referred to as the peptide's retention time. From the MS1 spectrum, a subset of the observed ions, corresponding to the eluted peptides, is selected for fragmentation based on their *m/z* values. The *m/z* values of the resulting peptide fragments are then measured to generate a tandem MS spectrum of the fragment ions (referred to as MS2 spectrum). It is from these MS2 spectra that the peptide sequences and therefore proteins are computationally identified. In other words, the MS1 spectra show the *m/z* value of the peptides entering the mass spectrometer and the MS2 spectra provide the mass of multiple fragments of a given peptide.

## 1.3 Peptide and Protein Identification

After MS data acquisition, the resulting MS2 spectra are searched against a protein sequence database to determine which peptides and therefore proteins were present in the original sample. Many database search tools exist for this purpose. Some of the most widely used include SEQUEST[19], X!Tandem[20], and Mascot[21]. Although these tools differ in their implementation, a database search tool typically computationally compares the experimental

mass spectra to a database of theoretical mass spectra to statistically determine the proteomic composition of the sample.

The theoretical mass spectrum database is generated from a protein sequence database specified by the user. These protein sequence databases are typically specific to the biological samples being analysed with the assumption that most of the proteins expected to be found in the sample are contained within the database. A set of theoretical peptides are generated by the in silico digestion of each protein sequences found in the protein sequence database. This in silico digestion breaks up proteins into peptides based on the cleavage sites of the restriction enzyme used during sample preparation. To account for the possibility of peptides that have not been fully digested during sample preparation, theoretical peptides with miscleavage sites can also be included in the peptide database. As well, peptides with certain amino acids with post-translational modifications can be included in the peptide database. The number of miscleavages and post-translational modifications is specified by the user and are necessary to identify these peptides in a sample at the cost of greater computational time. For all of these peptides, theoretical peptide mass fragments generated upon MS fragmentation are calculated to generate a theoretical MS2 spectrum.

Each MS2 spectrum acquired by the mass spectrometer is compared against these theoretical MS2 spectra and a confidence score is calculated indicating the confidence that this MS2 spectrum matches a given peptide. These peptide-spectrum pairings are called a peptide-spectrum match (PSM). PSMs with low confidence scores are likely matched by chance, while PSMs with high confidence scores are more likely to correspond to correct matches. Typically, only PSMs above a certain confidence score threshold are considered to be significant. This threshold can be chosen based on a desired false discovery rate (FDR) of PSMs. FDR is

estimated by comparing these confidence scores against the confidence scores acquired by searching a set of theoretical MS2 spectra generated from peptides that are not expected to be present in the sample, known as decoy peptides. One way these decoy peptides are generated is by reversing the amino acid sequences of each peptide in the peptide database, preserving the amino acid composition and co-occurrence of each pair of adjacent amino acids.[22,23] The FDR is therefore estimated as the proportion of PSMs that match a decoy peptide over all PSMs above a given confidence threshold. Users can specify their desired FDR (typically set to 1%).[24,25] This translates to filtering the set of PSMs such that no greater than 1% of these PSMs obtained their corresponding confidence score solely by chance. However, acquiring a set of PSMs alone is usually not sufficient to estimate which peptides and proteins are present in a sample.

After the mass spectra are searched with a database search tool, the results are processed using algorithms that statistically assess the probability that a set of peptides and therefore proteins are confidently identified. Some of the acquired spectra can have a significant amount of noise, some due to contaminating analytes. Since the database search tool attempts to match every spectrum (including these noisy spectra) to a peptide, statistical assessment for the confidence of a peptide identification is necessary to control for spectral quality and estimate false positive peptide identifications.[26] To assess the confidence of a protein identification, further statistical analysis is necessary to account for peptides which correspond to multiple proteins in the protein sequence database and proteins identified by a single peptide.[27] The approaches that assess protein identification confidence require completed mass spectrometry datasets to make such statistical inferences. Assessing protein identification confidence first requires a database search to be performed on the mass spectrometry data followed by the statistical assessment of peptide and protein identification confidence. These statistical

assessments require sufficient mass spectrometry data to be acquired and therefore current tools such as ProteinProphet[27] and Percolator[28–30] are not suitable for assessing protein identification confidence on-the-fly.

## 1.4 Sensitivity Issue of Mass Spectrometry-based Proteomics

While mass spectrometry-based proteomics can identify thousands of proteins in a single experiment, the technology still struggles at identifying proteins of low abundance without specific targeting or enrichment.[16,17] These low abundance proteins are vital for obtaining a comprehensive understanding of the biology of the cell, identifying the interactions between low abundance proteins and their interactors, characterizing low abundance post-translational modifications, and increasing the discovery power for protein differential expression in a disease state. For example, when assessing protein differential expression between a disease versus a non-diseased state, it is difficult to differentiate between a protein that is absent from a condition versus a protein that was unsuccessfully detected by the mass spectrometer due to its low abundance. Identifying more of these low abundance proteins would increase our ability to perform such analyses.

Such low abundance proteins often remain unidentified due to the MS/MS data acquisition approach. Indeed, due to the large number of potential ions the instrument can select for fragmentation and the limited amount of time peptides are available for MS analysis during liquid chromatography elution, systematically acquiring an MS2 spectrum for each peptide species in a sample is virtually impossible for a single experiment.[31] Although technological improvements have accelerated MS analysis to allow the acquisition of MS2 spectra for more peptides, the field is unlikely to achieve a comprehensive characterization of all possible peptides in extremely complex samples, such as microbiome and human plasma samples, anytime soon.

Typical MS analysis favours the acquisition of data from the most abundant proteins. Therefore, a common strategy for selecting peptides to fragment is to choose those that have the highest intensity detected by the mass spectrometer. Those peptides are more likely to generate MS2 spectra that have a greater signal-to-noise ratio and therefore are easier to identify compared to lower intensity peptides.[32] This method is referred to as Top *N* data dependent acquisition (DDA), where the mass spectrometer collects MS2 spectra for the *N* most intense peptides in an MS1 spectrum. Clearly, acquiring MS data in this manner consistently analyzes peptides of high abundance while obtaining little to no data for peptides of lower abundance, thereby limiting the protein identification sensitivity for the experiment.

Low abundance proteins are often not identified in an MS experiment due to this repeated analysis of peptides of high abundance. This issue is illustrated in Figure 1. In this toy example, Protein 2 is of relatively low abundance and remains unidentified. Peptide B' is a peptide, which would uniquely identify this protein, differentiating this protein from other proteins of higher abundance. However, this peptide is also of very low abundance. Since mass spectrometers acquire the MS2 spectra of peptides based on the *N* most intense precursor ions, this peptide is less likely to be identified in this manner. Therefore, proteins identified with high confidence are predominantly from proteins with high abundance, and proteins of low abundance often remain unidentified.[31,32]

## 1.5 Experimental Mitigation Strategies to Favour Low Abundance Protein Identification

### 1.5.1 Multiple MS Technical Replicates

Approaches have been proposed to mitigate the lack of identification sensitivity for low abundance proteins. For instance, the identification of additional proteins can be achieved by

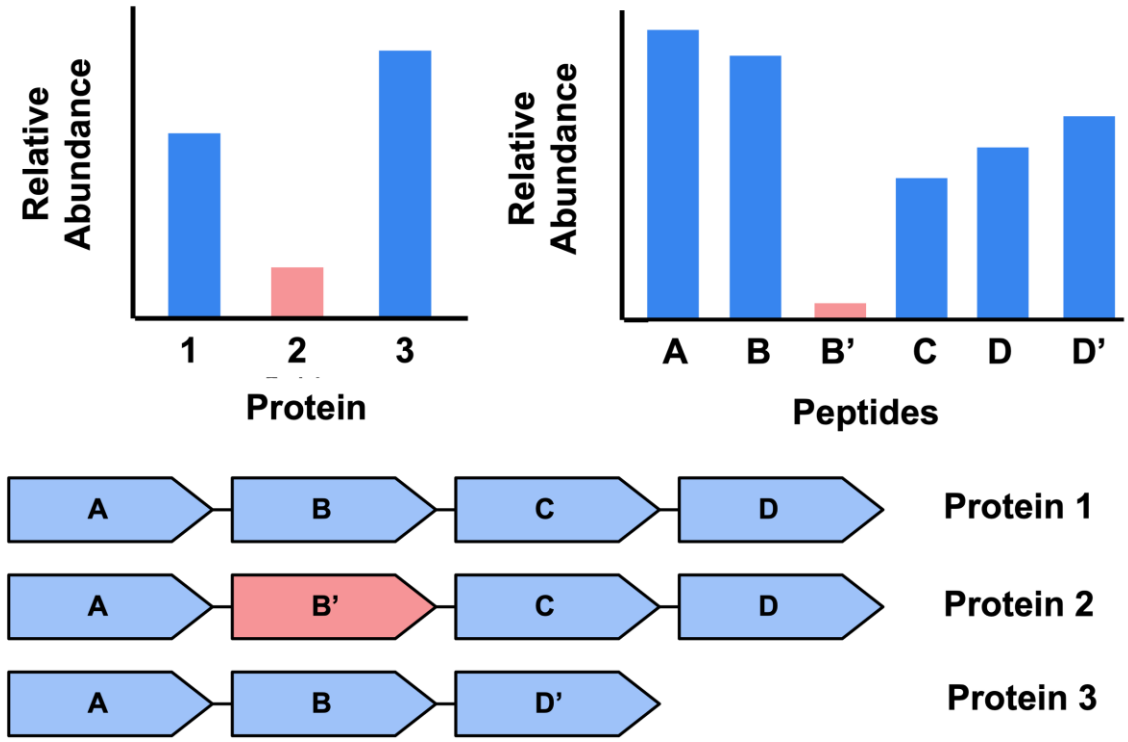


***Figure 1. Low Abundance Protein Identification Problem***

Three highly conserved proteins are shown in this toy example. Protein relative abundance is shown in the top left with blue and red bars indicating confident and not confident protein identifications, respectively. The tryptic peptides of these proteins are shown with their relative abundance in the top right with blue and red bars indicating confident and not confident peptide identifications, respectively. Peptide B' is different in sequence from Peptide B and uniquely identifies Protein 2, however it is not identified due to its low relative abundance. Hence, Protein 2 cannot be unambiguously identified without specific peptide evidence.

replicate analyses using multiple technical replicates of an MS experiment.[31] Fluctuations in the chromatographic separation, relative peptide abundance, and mass spectrometer sensitivity typically result in the mass spectrometer acquiring MS2 spectra from proteins that were not detected in previous replicates. While this strategy may identify a greater number of overall proteins, increases in protein identification sensitivity are generally minimal due to the repeated analysis of highly abundant peptides, which are usually confidently identified in most replicate MS analyses.[33] This technique also requires large amounts of sample and MS resources.

### 1.5.2 Liquid Chromatographic Separation

Additional proteins can also be identified by increasing the length of an MS experiment, resulting in an increase in the liquid chromatography separation resolution.[34] This increased resolution decreases the number of peptides that co-elute, thereby decreasing the number of lower abundance peptides that will be present in the mass spectrometer at the same time as higher abundance peptides. Although this approach can still result in the co-elution of peptides of vastly differing abundances, a fewer number of co-eluting peptides will yield more opportunities for lower abundance peptides to be analyzed. This typically results in an increase in the number of proteins identified overall, at the cost of utilizing far greater MS resources.[34]

Multiple chromatographic separation strategies can be used to reduce the complexity of a proteomics sample, including reverse-phase high-performance liquid chromatography, size-exclusion chromatography, and strong cation exchange chromatography to name a few.[35] Further separation of these co-eluting peptides can be achieved by coupling orthogonal separation strategies during liquid chromatographic separation.[35] For example, coupling strong cation exchange[36,37] or coupling ion mobility separation[38] both to reverse-phase high-performance liquid chromatography have been shown to increase protein identification sensitivity.

Low abundance proteins can be enriched using techniques such as affinity purification coupled to MS.[39] This technique utilizes antibodies to purify a protein of interest from a protein sample, either by direct binding or binding an epitope tag. By reducing the complexity of the sample and enriching for specific proteins, this method increases the likelihood that lower abundance proteins are identified by the mass spectrometer. This technique requires additional sample preparation, a priori knowledge of a protein of interest, and is difficult to perform in a high-throughput fashion.

### 1.5.3 Conventional Dynamic Exclusion

When a mass spectrometer acquires data in a Top-*N* Data Dependent fashion, high abundance proteins are often analyzed in excess of what is needed for a confident protein identification. As a result of the liquid chromatography separation, a given peptide species is only present during a given elution time window during the experiment. While liquid chromatography reduces the complexity of samples such that the mass spectrometer can acquire MS2 spectra for multiple proteins by separating peptides in time during the MS analysis, peptides from proteins with a very high abundance can elute over a long period of time.[40] Taking the mass spectrometer's Top *N* DDA strategy along with the long elution time periods of high abundance peptides, the instrument will redundantly analyze the same highly abundant peptides, while gathering little to no data related to less abundant peptides.

To resolve this issue, a common strategy used in the field is to employ a dynamic exclusion approach. Dynamic exclusion prevents the fragmentation of ions of a given *m/z* value (referred to as a precursor ion) for a short period of time after an MS2 spectrum was acquired for that same *m/z* value.[41] In other words, after an MS2 spectrum was acquired for a given peptide, the *m/z* value of that corresponding peptide ion will not be considered for further data acquisition

for a fixed period of time. The mass spectrometer therefore dynamically maintains a list over time of *m/z* values that are excluded from fragmentation, thus giving the opportunity for ions of peptides of lower intensity to be analyzed. Nevertheless, because abundant proteins generate numerous peptides and these abundant peptides can elute over periods of times much longer than their exclusion periods (usually up to 3 minutes[42]), MS approaches acquire more MS2 spectra than necessary for their confident identification. To illustrate this, I investigated a previously acquired MS dataset of a HEK 293 cell lysate run on a 120-minute gradient using dynamic exclusion. Assuming five MS2 spectra are necessary to identify a protein, a very conservative bar that would yield very confident identifications, over 60% of the acquired MS2 spectra associated with a confident protein identification are acquired in excess (Figure 2).

### 1.5.4 Bespoke Exclusion List

Several techniques take advantage of dynamic exclusion lists to achieve greater protein identification sensitivity. Among these, we count the Bespoke exclusion list approach which excludes peptides from fragmentation based on a pre-defined list of high-abundance and expected analytes that are not informative for a typical proteomics experiment.[40] Such an exclusion list often includes protein contaminants, such as keratins and proteolytic enzymes used in sample preparation, as well as small molecules, such as organic solvents and detergents used in cleaning lab equipment.[40] While this approach improves MS data acquisition by reducing the analytical time devoted to contaminant molecules that represent experimental artefacts, high abundance proteins that are biologically relevant remain likely to prevent the fragmentation of peptides with low abundance levels.

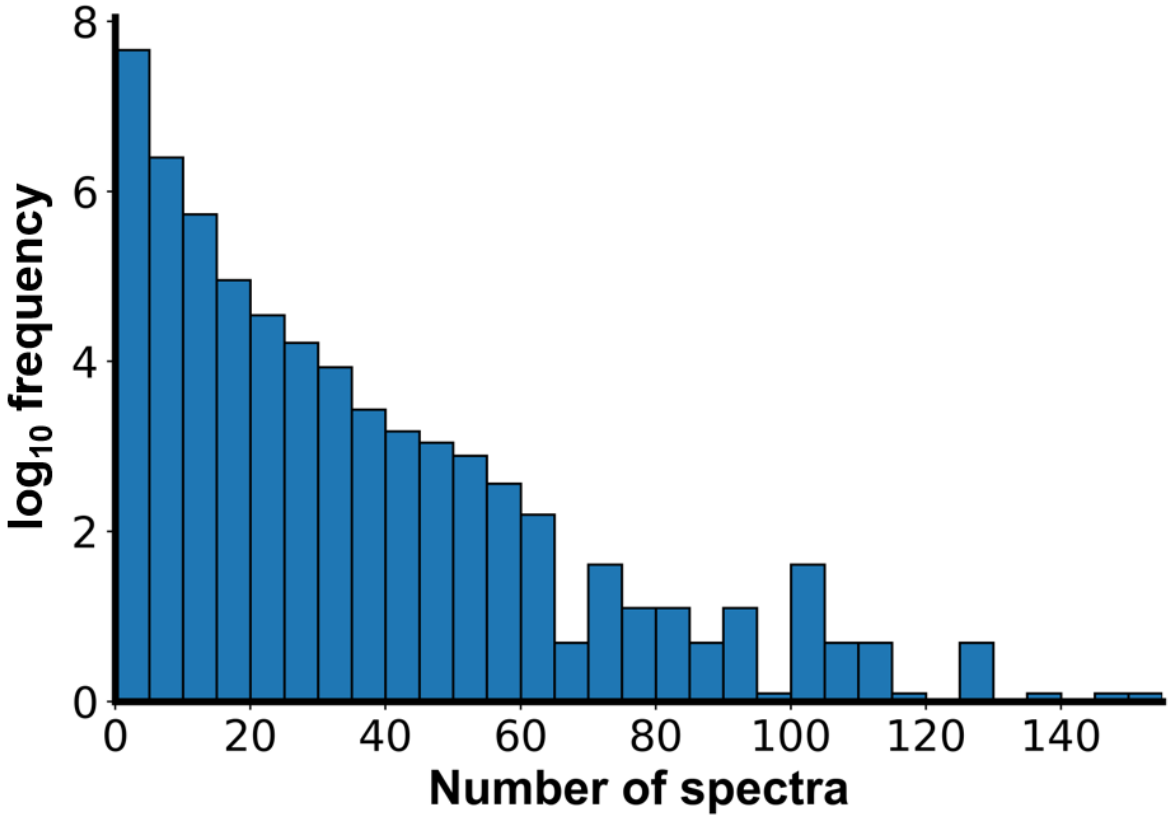


***Figure 2. Number of MS2 Spectra Acquired per Confident Protein Identification***

Number of MS2 spectra acquired per confident protein identification. Confident proteins were identified at a 1% FDR. In a 120-minute MS experiment using HEK293 cell lysate, over 60% of the acquired MS2 spectra associated with a confident protein identification are acquired in excess of 5 MS2 spectra per confident protein identification.

### 1.5.5 Excluding *m/z* Values Observed in Previous Replicated Experiments

Kreimer et al. also addressed low abundance protein identification issues with their SmartMS algorithm.[33] Their strategy uses indexed exclusion lists to prevent the selection of *m/z* values likely corresponding to peptides that have been confidently identified in multiple experiments, with the goal of increasing the number of confidently identified proteins. After each technical replicate MS analysis from aliquoted samples, confidently identified peptides are added to an exclusion list of the following replicate's MS analysis.

Although Kreimer et al. demonstrated that their approach was successful in identifying more proteins when compared to conventional Top *N* DDA, their algorithm relies on repeated analysis of a sample and thus requires a greater overall amount of sample and consumes more MS resources to iteratively exclude these highly abundant peptides.

### 1.5.6 Pseudo Real-time Exclusion of Peptides from Previously Identified Proteins

Further improvements on protein identification utilizing a dynamic exclusion list was proposed by McQueen et al.[43] Briefly, this method periodically updates an exclusion list based on proteins it has identified so far in the experiment. The approach uses spiked-in peptides with known retention times to signify multiple chromatographic "tripping points" to update their dynamic exclusion list. These tripping points help to ensure peptides are excluded at the correct time during the experiment. After each tripping point, the acquired data is used to assess protein identification confidence. Confident protein identifications must have acquired a minimum number of mass spectra with confidence scores above a specified threshold. The *m/z* values of all identified and expected peptides from the sequence of confidently identified proteins are added to the exclusion list at the following tripping point.

Though this approach reduces the redundant analysis of peptides from proteins identified with high confidence, the periodic update of the exclusion list does not fully take advantage of the information available in real-time, as peptides are only added to the exclusion list at the following tripping point, which can occur several minutes after a protein is confidently identified. Additionally, since their approach uses a series of thresholds for assessing a confident protein identification, optimal values for these thresholds can be difficult to determine. Choosing too lenient thresholds will result in the premature exclusion of peptides. As well, mass spectra acquired with borderline confidence values below this threshold cannot contribute to determining a protein's identification confidence.

## 1.6 Real-time Mass Spectrometry Analysis

Recently, some technologies have emerged involving real-time analysis of MS data to guide the behaviour of the mass spectrometer. For instance, the Real Time Search-$MS^3$ algorithm aims to improve the relative quantification of peptides from multiplexed proteomics samples, where multiple MS samples are analyzed simulatneously.[44,45] MS3 spectra are required for relative quantification of peptides, but can be acquired even for peptides that have not been identified with high confidence. This approach performs an on-the-fly database search for the acquired MS2 spectra to assess a peptide's identification confidence, only acquiring MS3 spectra for peptides that have been identified with high confidence. This prevents the mass spectrometer from unnecessarily acquiring MS3 spectra for peptides with poor identification confidence and in turn allows for the confident quantification of a greater number of peptides.

Another real-time proteomics analysis method is the Advanced Peak Determination method, which aims to increase the identification sensitivity of peptides and proteins by analyzing MS1 spectra as they are acquired to increase the number of suitable *m/z* candidates for

fragmentation by the instrument.[46] Mass spectrometers typically construct a list of candidate precursors suitable for fragmentation by annotating the peaks generated from the isotopic distribution and charge state of peptides in an MS1 spectrum. The key innovation for their Advanced Peak Determination method over the standard method is its improved ability to annotate peaks from multiple peptides that co-elute and occupy similar *m/z* regions. In the standard method, peaks corresponding to high abundance peptides are considered suitable for fragmentation while peaks from lower abundance peptides are often left unannotated due to interference from these high abundance peptides. This approach annotates the peaks from lower abundance peptides by iteratively annotating peaks from the MS1 spectra in order of abundance. This improves the number of suitable peptides for fragmentation by the mass spectrometer and in turn yields a greater number of peptide and protein identifications.

Finally, the MaxQuant.Live algorithm increases the number of peptides targeted for MS quantification.[47] The algorithm utilizes a targeted list of peptides the user wishes to quantify and recognizes these peptides from the MS1 spectra using the peptide's *m/z* value, time, and intensity. Peptides recognized by this algorithm can then be selected for standard fragmentation or targeted for a peptide specific proteomics analysis. One such analysis is the acquisition of a breakdown curve for a peptide, acquiring multiple mass spectra for a single peptide using different peptide fragmentation collision energies to determine the optimal collision energy for future proteomics analysis.

However, the laboratories developing these real-time approaches have made special agreements with the instrument manufacturers to gain full control of the instrument. This is made possible by the instrument's complete Application Programming Interface (API), which is not publicly available. As a result, other groups developing real-time analysis strategies without API

access have relied on simulations to test their methods. One such approach is Novor,[48] a real-time de novo peptide sequencing algorithm that uses simulations of previously completed mass spectrometry experiments to assess its performances.

These real-time MS methods show the wide range of applications possible by processing MS data on-the-fly to control the instrument's behaviour. However, none of these approaches aim to identify proteins during MS experiments in real-time with the goal of increasing the identification sensitivity of low abundance proteins.

## 1.7 Protein Classification using Machine Learning

### 1.7.1 Machine Learning

Machine learning algorithms are a set of artificial intelligence methods that learn from data or experience without being explicitly programmed to. Machine learning algorithms generally fall into one of two categories: unsupervised learning and supervised learning. The goal of unsupervised machine learning is typically to identify structure and patterns among data points in the input data. Examples of unsupervised machine learning include clustering, outlier detection, and dimensionality reduction. In contrast, supervised learning learns patterns in the data based on a training dataset with labeled examples to make predictions for unlabeled data. Examples of supervised learning include regression, where the algorithm predicts the value of a dependent variable, and classification, where the algorithm predicts the class of an observation. These algorithms learn relationships between data features to achieve the desired output. The question of whether a protein is confidently identified can be framed as a problem that can be solved by a supervised learning approach. More specifically, one could frame this question as a binary classification problem in which the two labels are "confidently identified" and "not confidently identified". The supervised learning algorithm would therefore learn from labeled

mass spectrometry data how the different features of the data can help to predict in unlabeled data whether a protein is confidently identified or not.

Some examples of supervised learning methods include artificial neural network, decision tree, support vector machine, and logistic regression classifier. An artificial neural network uses an approach composed of layers of computational nodes, mimicking the biological neural network of the brain. This approach can learn complex, non-linear relationships in training data, but requires greater computational resources and a large number of training examples to effectively train the model. As well, this approach often suffers from a lack of interpretability due to the large number of nodes and layers.[49,50] A decision tree on the other hand classifies its examples by learning a flowchart-like structure for its data features. This approach is highly interpretable as the rules learned for decision trees are based on real values on the data.[51,52] However, this approach is prone to overfitting to the training data such that learned rules might not generalize well to novel input data.[53,54] A support vector machine classifies its data by learning a decision boundary that best separates different classes in the training data. Since this approach aims to reduce the risk of misclassification on the data, it is typically less prone to overfitting.[55] However, in binary classification, a the classical support vector machine approach either classifies a datapoint as one class or the other and can hardly differentiate borderline classifications made with low confidence.[56] Finally, a logistic regression classifier uses a sigmoid function to determine the relationship between data features from its training data and estimates a confidence score or probability that a new data point falls into a given category. Although a logistic regression classifier may be more prone to overfitting than a support vector machine, the approach is advantageous in that borderline classifications can be identified and only predictions above a certain confidence or probability threshold can be classified as one class

or the other. Both the support vector machine and logistic regression classifier excel in their interpretability and its speed of calculation.[57,58]

### 1.7.1 Applications in Proteomics

Several machine learning algorithms have been employed with the goal of classifying correct and incorrect peptide and protein identifications based on their MS data features. Among these, DTASelect[59,60] and Percolator[61] use database search result features to discriminate between correct and incorrect PSMs, utilizing linear discriminant analysis and support vector machine algorithms, respectively. PeptideProphet[26] and ProteinProphet[27] both utilize expectation-maximization algorithms to discriminate between correct and incorrect peptide and protein identifications using spectral counts, respectively. Another approach by Elias et al.[62] uses a decision tree algorithm to discriminate between a correct and incorrect protein identification using spectral intensity patterns. Compared to simply applying a confidence score threshold to the database search results, all of these software packages are capable of controlling the number of false positive identifications. As are with most software used to process database search results to assess the confidence of peptide and protein identifications, these machine learning algorithms are utilized after the MS data acquisition is completed.

Taken together with the recent capability to analyse MS data as it is acquired to guide the instrument's behaviour in real-time and considering that a large percentage of MS2 are acquired in excess for a confident protein identification, we believe it is possible to use these resources to guide MS data acquisition to increase protein identification sensitivity.

# 2. Hypothesis and Objectives

## 2.2 Hypothesis

I hypothesize that I can develop a machine learning algorithm that increases the identification sensitivity of low abundance proteins in MS experiments, thereby providing a more comprehensive proteomic characterization and a better understanding of the biology of complex biological samples. This is achieved by assessing the identification confidence of a protein in real-time using machine learning and excluding from further analysis any peptides related to confidently identified proteins. Our algorithm is tested in silico by simulating a real-time MS analysis on a previously completed MS experiment and processing mass spectra in the order they were acquired to determine whether the precursor *m/z* value of the mass spectra should be excluded from analysis or its ion be fragmented to obtain an MS2 spectrum to be used for protein identification. We expect the exclusion of peptides from confidently identified proteins will favour the identification of lower abundance proteins by freeing up MS resources to analyze mass spectra from less abundant proteins when applied, to guide MS data acquisition in real time.

To achieve our goals presented for this project, the following aims were devised:

### 2.2.1 Aim 1

Design a supervised machine learning algorithm to confidently identify proteins in real-time during an MS analysis.

### 2.2.2 Aim 2

Build and evaluate a real-time exclusion list of *m/z* values from the proteins deemed confidently identified by the machine learning algorithm.

# 3. Methods

The goal of my approach is to free up MS resources to enable the detection of low abundance peptides and therefore their parent proteins, illustrated in Figure 3. Briefly, to improve the protein identification sensitivity of the mass spectrometer, I have developed a machine learning algorithm that uses MS data features to estimate the confidence of protein identifications in real-time and exclude the unnecessary data acquisition of mass spectra from peptides that belong to already confidently identified proteins. The algorithm's pipeline is depicted in Figure 4. This algorithm is tested in silico, reading through previously acquired MS data files to mimic data acquisition, processing the MS1 and MS2 spectra to make decisions whether to analyze a mass spectrum or exclude it from further analysis.

## 3.1 Datasets Analyzed

For software development and testing, MS data was generated by Zhibin Ning from the Daniel Figeys's laboratory. The samples used for our analyses are from the Human Embryonic Kidney 293 (HEK293) cell line, a widely studied cell line in proteomics that allows us to easily verify and benchmark our approach. These HEK293 cell lysate were also processed by Zhibin Ning and were analyzed for a 120-minute and a 240-minute liquid chromatography gradient on the Q-Exactive mass spectrometer (Thermo Electron, Waltham, MA). The 120-minute experiment was used to assess our algorithm's performance with an in silico simulation of real-time MS data acquisition. The 240-minute experiment was used to train the machine learning classifier evaluating the confidence of protein identifications, while the 120-minute experiment was used to test the classifier. We chose these samples for their respective roles with the assumption that a confident protein identification from the 240-minute experiment will share similar properties to a confident protein identification for the 120-minute experiment, while illustrating the ability of

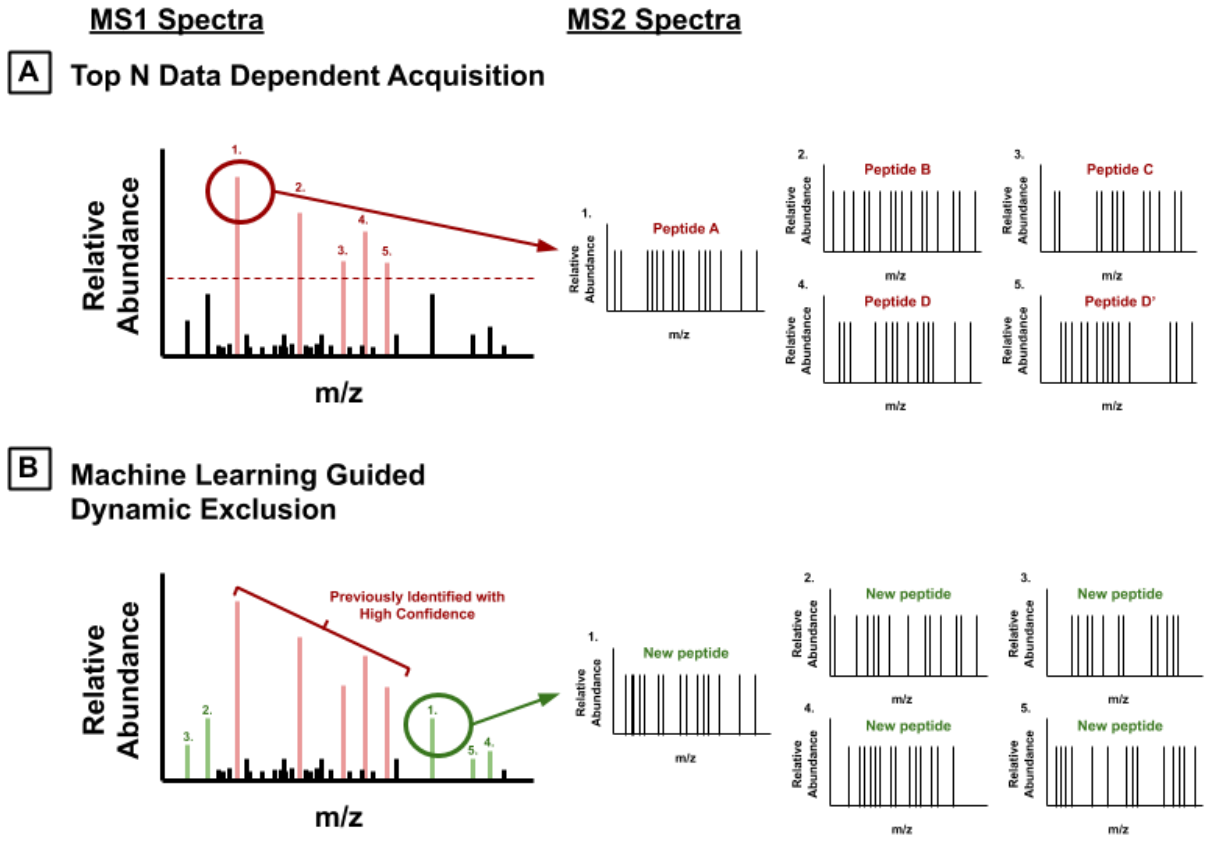


***Figure 3. Top N Data Dependent Acquisition vs Machine Learning Guided Exclusion***

Conventional Top *N* Data Dependent Acquisition acquires MS2 spectra on the top *N* peptides with the highest intensity, illustrated with *N*=5 (A). The top 5 peptides shown in red are fragmented and their MS2 spectra are shown to the right. Our algorithm excludes peptides from proteins that have been already been identified with high confidence and fragments the less abundant peptides shown in green, with their MS2 spectra shown to the right. All spectra are toy examples.

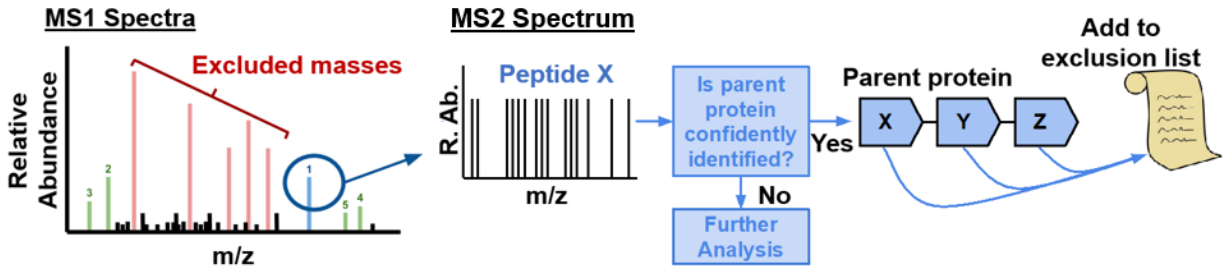


***Figure 4. Machine Learning Guided Exclusion Pipeline***

Our algorithm obtains an MS1 spectrum and determines which peptides to fragment based on the on-the-fly exclusion list. The resulting MS2 spectra are processed and the confidence of the identification of their corresponding proteins is determined. Peptides from confidently identified proteins are added to the exclusion list, while the remaining peptides are subject to further investigation.

our approach to be generalizable to different datasets when trained on another. Although these datasets are from the same cell line and were processed using the same mass spectrometer, the samples were analyzed using different experimental protocols in their data acquisition and identified a different number of proteins (120-minute: 2702 proteins; 240-minute: 3239 proteins).

## 3.2 Protein Extraction for HEK 293 Cells

Briefly, HEK293 cells were grown to 80% confluence in 15cm dishes and harvested in RIPA buffer (25 mM Tris-HCl, pH 7.6), 150 mM NaCl, 1% NP-40, 1% sodium deoxycholate, 0.5% Sodium Dodecyl Sulfate (SDS, Cell Signalling Technology), after two washes with phosphate-buffered saline (PBS). Cells were sonicated for 1 min with 20 seconds pulse using 30% power on Sonic Dismembrator 500 w/ Branson 1020 Sonicator (Fisher Scientific) to increase protein recovery. Proteins were precipitated to remove detergent by adding 5X volume of cold acetone overnight followed by two washes with cold acetone as well. The protein pellet was then reconstituted in 8 M urea (Sigma-Aldrich) in 50 mM ammonium bicarbonate (ABC, Sigma-Aldrich) and quantified by the DC protein assay kit (BioRad). Reduction and alkylation were done by adding dithiothreitol (DTT, Sigma-Aldrich) to a final concentration of 10 mM at 56°C for 30min followed by 20 mM iodoacetamide (IAA, Sigma-Aldrich) at room temperature. The solution was then diluted 5 times by 50mM ABC. Digestion was performed by adding trypsin (Worthington Biochemical Corp., Lakewood, NJ) at a protein-enzyme ratio of 50:1, 37 °C overnight, with continuous head-to-end rotating. Digested peptides were then desalted on Sep-Pak C18 SPE column (Waters, US), aliquoted and dried down by SpeedVac (ThermoFisher Scientific, San Jose, CA). Dried peptides were reconstituted in 0.5% (v/v) FA, 2% ACN in water.

## 3.3 LC-MS/MS Analysis

Eksigent 2D+ nanoLC system (Dublin, CA) was hooked up with a Q-Exactive mass spectrometer (Thermo Electron, Waltham, MA), equipped with a nano-electrospray interface operated in positive ion mode. The solvent system consists of buffer A of 0.1% FA in water, and buffer B of 0.1%FA in 80% acetonitrile. Dried down protein digests were acidified with 0.5% (v/v) formic acid and loaded on 200 μm I.D. × 20 mm fused silica pre-column packed in-house with 5μm ReproSil-Pur C18 beads (100 Å; Dr. Maisch GmbH, Ammerbuch, Germany) at a flow rate of 4uL/min for 10min, and analyzed on 75 μm I.D. × 150 mm fused silica analytical column packed in-house with 1.9μm ReproSil-Pur C18 beads (100 Å; Dr. Maisch GmbH, Ammerbuch, Germany) at a flow rate of 200 nL/min for either 120min or 240min. Gradient elution was set from 5 to 35% buffer B (80% ACN). The spray voltage was set to 2.0 kV and the temperature of the heated capillary was 300 °C. The instrument method consisted of one full MS scan from 300 to 1800 *m/z* followed by the acquisition of data-dependent MS/MS scan of the 12 most intense ions, a dynamic exclusion repeat count of 1 in 30s, and an exclusion duration of 30s. The full mass was scanned in Orbitrap analyzer with R = 70,000 (defined at *m/z* 400) for MS1 and 17,500 for MS2. To improve the mass accuracy, all the measurements in the Orbitrap mass analyzer were performed with a real-time internal calibration by the lock mass of background ion 445.120025. The charge state rejection function was enabled, and charge states with unknown and single charge state were excluded for subsequent MS/MS analysis. All data were recorded with Xcalibur software (ThermoFisher Scientific, San Jose, CA).

## 3.4 Mass Spectrometry File Processing

During MS analysis, the mass spectrometer regularly acquires MS1 spectra and selects for fragmentation the top *N* ions with the highest intensity that are not on the mass

spectrometer's exclusion list. This results in the generation of MS2 spectra for these peptides that are later matched to a peptide sequence. These MS1 and MS2 spectra are gathered by the mass spectrometer in a proprietary, vendor specific RAW file, which must be converted to file formats usable by the database search tool. Thermo RAW files from previously completed MS experiments were converted to the mzML open file format using MSConvert from the ProteoWizard package.[63]

The protein sequence database search algorithm Comet[64] was used to search the resulting mzML files against the UniProt Swiss-Prot human protein sequence database (downloaded on 2017-01-11; number of protein entries: 20130).[65] These database searches were performed with a precursor ion mass tolerance of 20 ppm, considering only fully tryptic peptides with a maximum of 2 miscleavages allowed. Carbamido-methylation of cysteine was considered as a fixed modification. To maximize the usability of our software package and its flexibility in handling results from different database search tools, the Comet results were converted to the HUPO-Proteomics Standard Initiative mzIdentML[38] standard file format using tpp2mzid from the Trans-Proteomics Pipeline.[66] Comet reports a cross-correlation score (XCorr) for each peptide, indicating the quality of a match between an MS2 spectrum and the peptide sequence with a higher XCorr score indicating a more confident match.[64] At the time the work in this thesis was performed, the manufacturer's software necessary to fully control the mass spectrometer was not available. As well, the real-time database search capabilities of Comet were not yet released. In this thesis, MS2 spectra were searched with Comet ahead of the simulation of the real-time analysis.

## 3.5 Aim 1: On-the-fly Protein Confidence Assessment

The key innovation of this project is the construction and real-time usage of an on-the-fly dynamic exclusion list using a machine learning algorithm identifying proteins as MS data are acquired. As the mass spectrometer acquires an MS2 spectrum, a supervised machine learning approach assesses the likelihood that the corresponding protein is confidently identified and excludes from MS analysis its corresponding peptides. Specifically, the confidence that a protein *p* is identified is assessed using a logistic regression classifier, using the following features: 1) the cardinality of the set *S* of MS2 spectra matched to peptides belonging to *p* (also called spectral count[67]), 2) the highest XCorr of the PSMs from *S*, 3) the mean, median, and variance of the XCorr of the PSMs from *S*, and the pairwise products of these features. From these features, the classifier computes a confidence score that *p* is reliably identified. Proteins which achieved an identification confidence score above a defined confidence score threshold are deemed confident protein identifications. We use a logistic regression classifier trained on the above protein features obtained from the completed 240-minute MS analysis of a HEK293 cell lysate from which positive examples of confident protein identifications and negative examples of proteins that have not been identified with high confidence were extracted. PeptideProphet and ProteinProphet were used to assess the confidence of a protein identification obtained using Comet.[26,27] These approaches provide a false discovery rate (FDR) estimate of the identifications made by Comet.

Proteins identified with less than 1% FDR and spectral count less than or equal to 30 were used as positive examples of highly confident protein identifications. Proteins identified with less than a 25% protein probability score from Protein Prophet (corresponding to an FDR > 2.4%) were used as negative examples of protein identifications. The number of training

examples were 2834 and 524 for the positive and negative training sets, respectively. The positive training set for a confident protein identification was chosen using a 1% FDR threshold, which is considered to be standard in the field for a significant protein identification. The spectral count threshold of 30 was chosen to prevent our classifier from overfitting confident protein identifications as only those with very high spectral counts. Similarly, the negative training set for a poor protein identification were chosen as proteins identified with greater than 2.4% FDR to ensure that the set is enriched for very poor-quality identifications and few border-line confident protein identifications close to 1% FDR.

The protein identification confidence score output by the logistic regression classifier follows a scale of 0.0 to 1.0, with 1.0 corresponding to a confident protein identification and 0.0 corresponding to a poor protein identification. A protein obtaining a classifier score above a given threshold is deemed confidently identified. Confidently identified proteins in real-time can then be used to build and maintain a dynamic exclusion list of *m/z* values.

## 3.6 Aim 2: Using On-the-fly Confidence Assessment of Proteins to Build a Real-time Exclusion List

### 3.6.1 Peptide Retention Time Prediction and Correction

Because multiple distinct peptides in a complex sample, such as a cell lysate have a very similar and even equal *m/z* values, the exclusion of a given *m/z* value may result in the accidental exclusion of the ions of peptides that were not meant to be excluded from MS2 spectrum acquisition. However, much less peptides have both similar *m/z* values and retention times due to their difference in amino acid composition and therefore hydrophobicity. To minimize such unintentional exclusions, we restrict fragmentation exclusion of a given *m/z* value for a time

duration centered around its peak retention time. This is similar to the amount of time a given *m/z* value is excluded in a conventional dynamic exclusion approach. This duration will be referred to as its *m/z* exclusion time window.

The peak retention time of each peptide was predicted using RTCalc from the Trans-Proteomics Pipeline[66] and a *m/z* exclusion time window is built around this prediction as described above. However, peak peptide retention time can vary due to a number of factors, including variations in liquid chromatography gradient,[68] different sample preparation techniques,[69,70] and relative abundance of the peptide in a sample,[40] and is therefore difficult to predict. Therefore, during the analysis, when a peptide is matched to an MS2 spectrum with a XCorr score > 2.5, its *m/z* exclusion time window is adjusted such that the time this peptide was observed by the mass spectrometer corresponds to the beginning of its *m/z* exclusion time window. This XCorr threshold was chosen, as the proportion of false positive peptide identifications is < 0.12% at this threshold.[22] Additionally, the time difference between a peptide's observed and predicted retention time is also used to calibrate the predicted retention time of all peptides. To increase the accuracy of all *m/z* exclusion time window, the peptide retention time predictions from RTCalc are calibrated in real-time during the experiment. For each peptide with a XCorr > 2.5, we calculate the difference between the peptide's predicted retention time peak and the time at which this peptide is observed. This time difference is then used to calculate an offset using the *M* most recent differences. An offset is calculated using a harmonic series, which weighs the differences in retention times in favour of the more recent values:

$$offset = \frac{\sum_{m=1}^{M} \frac{1}{m} d_m}{\sum_{m=1}^{M} \frac{1}{m}},$$

Where $d_m$ is the time difference of the m$^{th}$ peptide from *M*. The offset is then added to the RTCalc predicted retention time to more accurately reflect when a peptide might be observed in an experiment. The XCorr threshold of 2.5 was chosen to prevent low quality and false positive peptide identifications from negatively affecting the retention time calibration, while providing enough data points to accurately calibrate.

This retention time offset calculation is referred to as the Weighted Mean offset calculation since it weights the differences based on a harmonic series. This approach is benchmarked against the Mean offset calculation and the Median offset calculation, which take the mean and median of the *M* most recent differences, respectively.

### 3.6.2 On-the-fly Exclusion List Construction

The mass spectrometer uses an exclusion list, consisting of *m/z* values it will not select for fragmentation and therefore excluding the acquisition of certain MS2 spectra.. Ions with a *m/z* value within a given mass tolerance *t* measured in parts per million (ppm) of the excluded *m/z* value on the exclusion list will not be selected for fragmentation. This mass tolerance accounts for inaccuracies in *m/z* value measurements by the mass spectrometer. Mass spectrometers with a Top *N* data dependent acquisition strategy will select for fragmentation the *N* most abundant ions that are not on the dynamic exclusion list for fragmentation. Our algorithm uses this dynamic exclusion list to guide MS data acquisition.

The dynamic exclusion list is constructed as the algorithm processes MS data and assesses the confidence of protein identifications. Peptides analyzed by the mass spectrometer that are matched with an MS2 spectrum with XCorr above 2.5 and peptides from proteins identified with high confidence will see their *m/z* values added to the dynamic exclusion list by our software. The latter peptides are excluded based on a peptide sequence database. This

database is generated before the MS run using the UniProt Swiss-Prot sequence database (downloaded on 2017-01-11; number of entries: 20130)[65] and digested in silico using the chainsaw tool from the ProteoWizard suite[63] with Trypsin/P as the digestion enzyme, which cleaves after arginines and lysines, even when followed by a proline. Peptides generated are fully tryptic with at most one miscleavages and have minimum and maximum length of 6 and 200 amino acids, respectively. The masses are then calculated using the amino acid constitution of the resulting peptides. Hence, once a protein is deemed confidently identified, the masses of all peptides associated with it in the peptide sequence database are added to the exclusion list for the period defined by its exclusion time window (described in previous section).

Our algorithm mimics the instrument's dynamic exclusion list, generating an exclusion list of peptide masses at a given experimental time corresponding to confidently identified proteins (Aim 1), which the mass spectrometer would be able to use to guide its data acquisition. In this way, the algorithm is capable of dynamically excluding peptides during MS data acquisition (Aim 2), freeing up resources to analyze mass spectra from peptides that will lead to the identification of more proteins in an experiment.

### 3.6.3 Simulated Real-time Dynamic Exclusion in Silico

To assess the performance of our algorithm, we simulate MS data acquisition in silico using a previously acquired dataset. Our algorithm processes the spectra from the mzML file in the order that they were previously acquired by the instrument. When our algorithm reads a spectrum, it then makes a real-time decision based on the dynamic exclusion list whether to perform a database search on that spectra or ignores it to simulate its exclusion. The database search of the MS2 spectra for peptide identification is not performed in real-time. Instead, the MS datasets are searched using Comet (see search parameters above) prior to processing the

datasets with the simulation. Our machine learning package processes this file by associating the Comet peptide identifications to the corresponding MS2 spectra. Our tool updates the dynamic exclusion list based on the peptide identifications.

To assess the number of proteins identified, our software package outputs a partial database search result file of the MS2 spectra that were not excluded from MS/MS analysis. The resulting file is processed using PeptideProphet and ProteinProphet to assess the identification confidence of peptide and protein identifications, respectively.[26,27] Putative protein identifications were filtered at a 1% FDR. At the same time, the software package keeps track of the MS2 spectra that were excluded from the analysis and not used in protein identification downstream. These are considered saved resources that could be redistributed to acquire MS2 spectra from proteins that remain uncharacterized.

## 3.7 Algorithm Performance and Benchmarking

### 3.7.1 Heuristic Approach to Protein Identification Confidence

Our logistic regression classifier assessing protein identification confidence for dynamic exclusion building was benchmarked against a heuristic approach based on a method proposed by McQueen et al.[43] The heuristic exclusion strategy matches the implementation of our machine learning-guided exclusion in all aspects of the algorithm except for its protein identification confidence assessment. In place of the logistic regression classifier we use for assessing protein identification confidence, a heuristic strategy using both a XCorr threshold and a PSM count threshold is employed. A protein *p* is considered confidently identified when the number of spectra associated with *p,* with confidence scores above a given XCorr threshold, is greater than the PSM count threshold. In this way, we assess the performance of our logistic regression

classifier in improving overall protein identification compared to using this heuristics-based approach.

Determining the optimal thresholds for this heuristic approach can be challenging. As well, any scores below these thresholds do not contribute to determining a protein's identification confidence. Our logistic regression classifier approach addresses these issues, requiring only the selection of a single threshold for protein identification confidence scores that uses training examples of confidently identified proteins to estimate protein identification confidence. In this way, our approach is more flexible in that it can include scores which would have otherwise been discarded with the heuristic approach.

### 3.7.2 Using Saved Resources: A Simulation

The MS dataset used for simulation was acquired using a Top 12 DDA strategy, meaning that (at maximum) the top 12 most intense ions from a given MS1 mass spectrum are fragmented to generate an MS2 spectrum. Since our approach excludes certain *m/z* values from MS2 spectrum acquisition, some of these 12 ions would not be fragmented and therefore allow the mass spectrometer to fragment different ions than those 12. However, despite software tools now freely available to access MS data in real-time during an experiment, current software supplied by mass spectrometer vendors do not readily allow for our software to update in real-time the instruments' exclusion list. Hence, to estimate further improvements in protein identification sensitivity our algorithm could generate, we simulate a limited Top *N-X* DDA experiment, artificially limiting the number of MS2 spectra obtained after each MS1 spectrum by assuming that *X* of those are not acquired after each MS1 spectrum. Therefore, if our algorithm excludes from fragmentation one or more of the top *N-X* ions, our software can simulate the acquisition of additional spectra from the *X* remaining spectra, for which we have previously collected

information. These additional MS2 spectra makes it possible to estimate how our approach has the capability to improve identification sensitivity of lower abundance proteins.

# 4. Results

The goal of identifying proteins with fewer mass spectra to free up analytical resources is achieved using a machine learning algorithm to confidently identify proteins in real-time during an experiment to guide MS data acquisition. My approach processes MS data as it is being acquired to make decisions in-real-time to prevent the mass spectra acquisition of peptides from proteins previously identified with high confidence. Our algorithm includes two key components: Aim 1) training a logistic regression classifier for estimating protein identification confidence and Aim 2) utilizing a dynamic exclusion list to prevent the analysis of peptides from proteins identified with high confidence. Our algorithm is tested and benchmarked on a previously completed MS experiment *in silico*.

## 4.1 Logistic Regression Classifier Training

We utilize a logistic regression classifier to estimate the likelihood that a protein is confidently identified based on database search results and spectral count data. The training procedure of the classifier assigns weights to each feature representing how useful for determining whether a protein is confidently identified or not. These feature weights are learned from protein identifications obtained in 240-minute LC-MS/MS analysis of a HEK 293. The weights learned for each feature used in the logistic regression classifier are reported in Table 1. Since most features are gathered from the peptide match quality score (XCorr), these features are expected to be correlated. For our classifier, the most discriminating features were the standard deviation, median, and mean of the obtained XCorr scores.

We then assessed our classifier on a 120-minute LC-MS/MS analysis of a HEK 293 to evaluate that a classifier assessing protein identification confidence is generalizable between similar MS experiments despite the difference in LC gradient duration as confident protein

***Table 1. Logistic Regression Classifier Feature Weights[1]***

| **Feature** | **Weight** |
|---|---|
| Standard Deviation of XCorr Scores | -1.28211 |
| Median of XCorr Scores | -0.54573 |
| Mean of XCorr Scores | -0.4574 |
| Highest XCorr Score | 0.221874 |
| Cardinality | -0.01056 |
| Mean of XCorr Scores X Median of XCorr Scores | 0.22983 |
| Highest XCorr Score X Standard Deviation of XCorr Scores | 0.182569 |
| Highest XCorr Score X Median of XCorr Scores | 0.166113 |
| Highest XCorr Score X Mean of XCorr Scores | 0.15695 |
| Mean of XCorr Scores X Standard Deviation of XCorr Scores | 0.151736 |
| Median of XCorr Scores X Standard Deviation of XCorr Scores | 0.141475 |
| Cardinality X Standard Deviation of XCorr Scores | -0.00338 |
| Cardinality X Mean of XCorr Scores | -0.00168 |
| Cardinality X Median of XCorr Scores | -0.0015 |
| Cardinality X Highest XCorr Score | -0.00084 |

[1] Intercept: 1.23629

identifications in both cases should share similar MS features. Proteins that obtain a probability of confident identification as given by the classifier, which is above a defined probability threshold are classified as confidently identified. By varying this threshold, one can yield a greater or fewer number of proteins deemed as confidently identified.

Given a probability threshold, the classifier will predict a protein as confidently identified or not. The performance of the classifier is tested against a 120-minute LC-MS/MS experiment, comparing the classifier's prediction versus the test dataset consisting of proteins identified at a 1% FDR using PeptideProphet and ProteinProphet.[26,27] Proteins confidently identified by our classifier as well as in the test dataset are considered true positives. True negative classifications are proteins not confidently identified by our classifier as well as in the test dataset. On the other hand, false positive classifications are those deemed confidently identified by our classifier but not in the test dataset. False negative classifications are those deemed not confidently identified by our classifier but are confidently identified in the test dataset. The true positive rate is defined as the number of true positives over the sum of true positives and false negatives. The false positive rate is defined as the number of false positives over the sum of false positives and true negatives. A receiver operating characteristics curve showing the true positive and false positive rates of the classifier at a varying probability threshold when applied to the 120-minute LC-MS/MS experiment is shown in Figure 5. From this curve it was calculated that the area under the curve is 0.812.

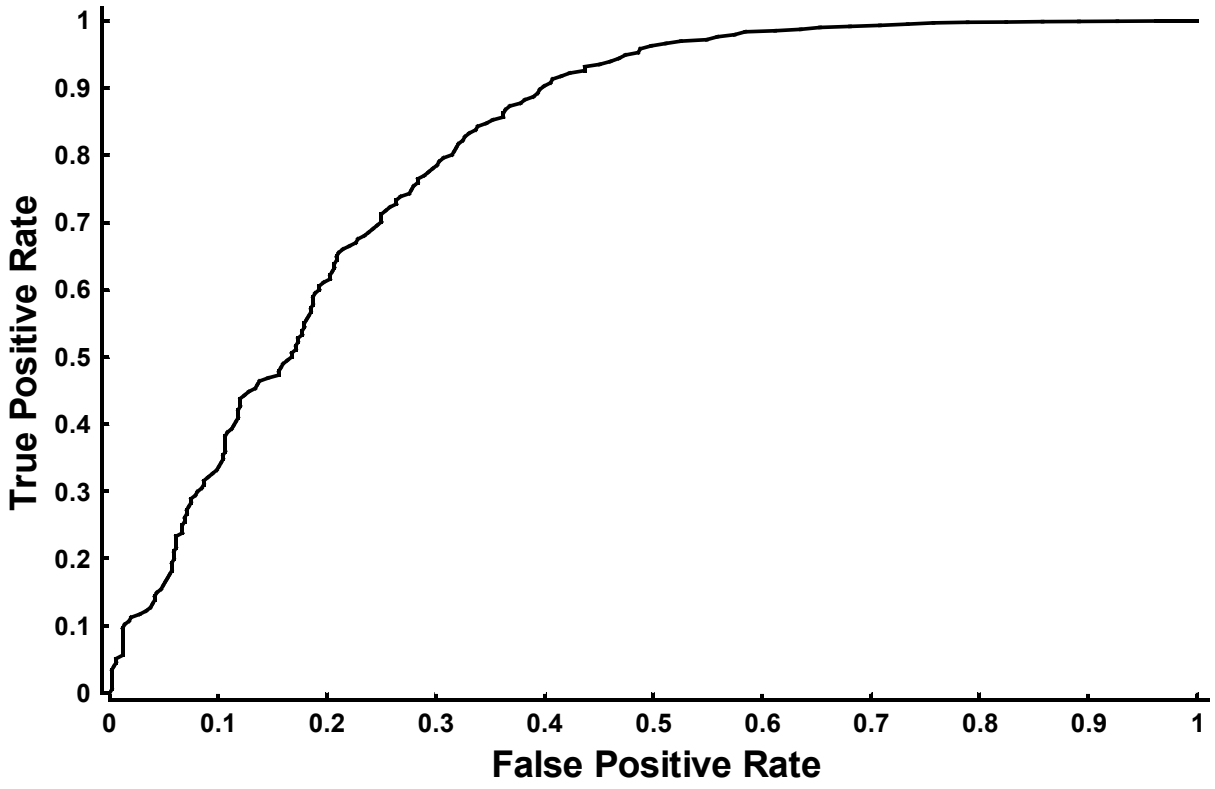


***Figure 5. Receiver Operating Characteristic Curve for Protein Identification Logistic Regression Classifier***

Receiver operating characteristic performed for logistic regression classifier assessing protein identification confidence, trained on the 240-minute LC-MS/MS experiment and tested on the 120-minute LC-MS/MS dataset. Classifier performance is compared to proteins confidently identified at a 1% FDR using ProteinProphet. True Positive and False Positive Rates are reported at a varying probability threshold.

## 4.2 On-the-fly Peptide Retention Time Calibration

Peptides from proteins deemed confidently identified see their masses added to the dynamic exclusion list for a given mass exclusion time window, such that the ions with the corresponding *m/z* values are not redundantly fragmented. RTCalc from the Trans Proteomics Pipeline[66] was used to determine the expected retention time of peptides added to the exclusion list. However, accurate peptide retention time prediction is challenging as differences in experimental design can result in vastly different retention times between experiments. Indeed, peptide retention time predictions for our 120-minute LC-MS/MS experiment were poor. Retention time prediction error was calculated as the difference between a peptide's predicted retention time and the experimental time when this peptide was detected by the mass spectrometer. In the experiment, only 0.57% of the peptide-spectrum matches were detected within 2 minutes of their predicted retention time (Figure 6). Some retention time errors were greater than 75 minutes from the predicted time. Although peptide retention time prediction performs poorly in this sense, since these predictions are based on the biochemical properties of these peptides, we expect these peptides to elute through the chromatography column in a similar order as their predicted retention times. Figure 6 shows that the difference between a peptide's predicted retention time and experimental time at which it was observed vary predictably throughout the experiment. Our software package therefore performs a predicted retention time correction to increase the accuracy of the predicted retention times and therefore improves the usability of our on-the-fly exclusion list. An offset is computed to adjust retention time prediction using the *M* of most recent differences between a peptide's predicted and observed retention times of peptide-spectrum matches with XCorr > 2.5 (see methods).

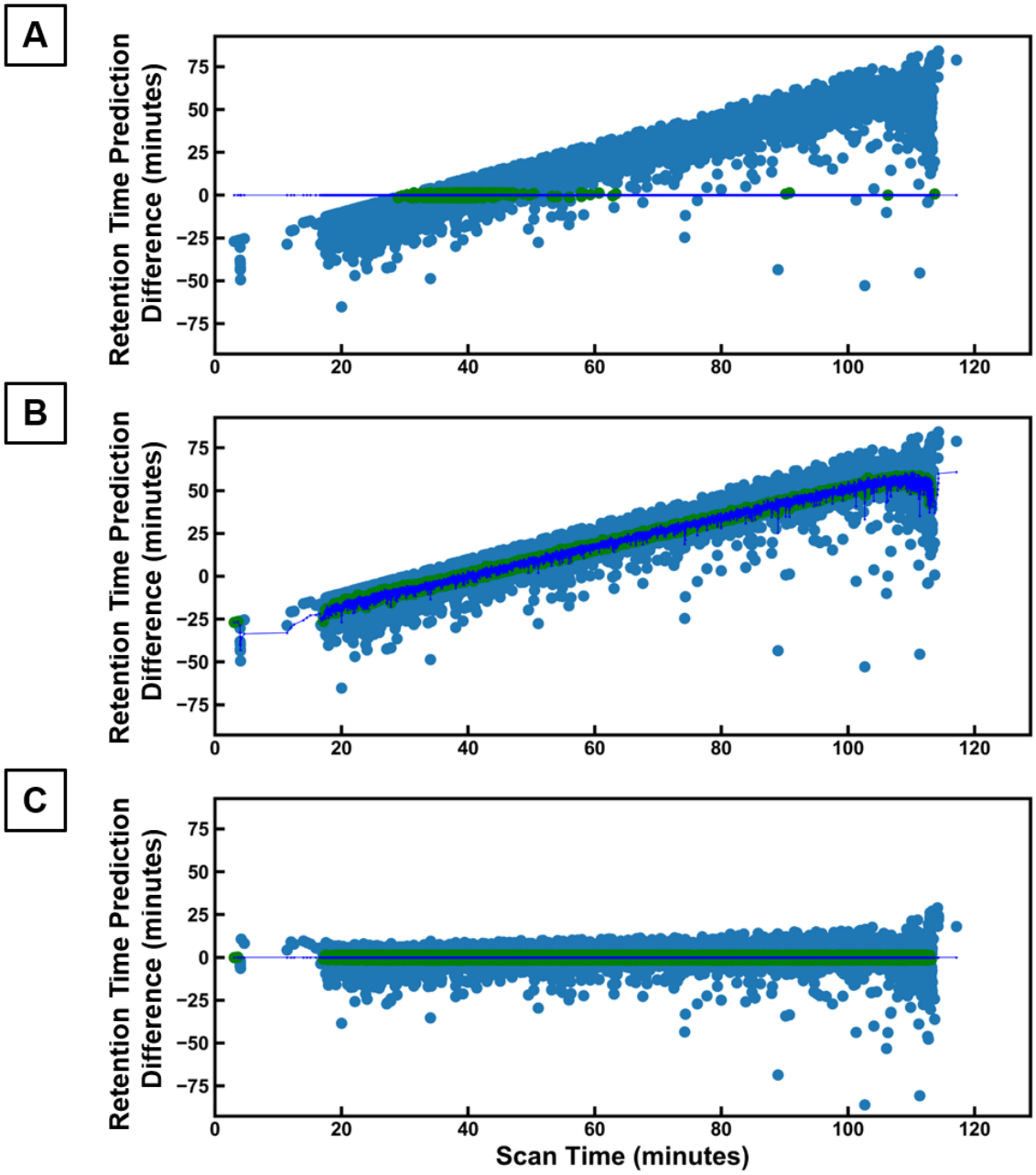


***Figure 6. Predicted Peptide Retention Time Before and After Retention Time Correction***

Predicted peptide retention time was corrected using experimental data from a 120-minute LC-MS/MS experiment. Retention time prediction error is shown for each peptide-spectrum matches at its experimental time in the top panel. Peptides falling within 2.0 minutes of predicted retention time are shown in green. Using the retention time prediction errors, a time offset was

calculated at each time point, using a weighted mean of at most the 100 most recent retention time prediction errors shown in blue in the middle panel. This value is added to the predicted peptide retention time at each time point and shown in the bottom panel.

The performance of this correction was assessed by comparing the fraction of peptides that actual retention times fall within a given mass exclusion time window *w* (in minutes) of its predicted retention time. We assessed the performance of our calibration method against the arithmetic mean and median for the *M* most recent differences in retention times for peptide-spectrum matches with XCorr > 2.5. For each method, *M* and *w* were varied. In most cases, the weighted mean method outperformed the arithmetic mean and median methods for offset calculation with regards to the number of peptides with a corrected retention time now falling within *w* of the predicted retention time being higher (Table 2). Of note, incorporating a greater number of *M* error values in the offset calculation reduces the number of peptides with retention times falling within *w* of the predicted retention time. As expected, when increasing *w*, a greater percentage of peptides fall within its predicted retention time. However, increasing *w* too much runs the risk of incorrectly excluding peptides with similar masses and retention times when applied to our machine learning guided dynamic exclusion. Using the weighted mean method and *M* as 100 most recent retention time prediction errors for retention time offset calculation yielded the highest percentage of observed peptides falling within two minutes of its predicted retention time (bold in Table 2, Figure 6 bottom). This increased the percentage of peptides within 2 minutes of its predicted retention time from 0.57% to 48.2%.

For peptide retention time calibration to be achieved on-the-fly, the offset values must be calculated only from MS2 spectra acquired by the mass spectrometer. Since our algorithm excludes from analysis a number of these spectra, we assess the performance of the peptide retention time calibration using only the MS2 spectra that were not excluded using our algorithm. The 120-minute LC-MS/MS analysis of HEK 293 cell lysate was simulated with machine learning-guided exclusion with on-the-fly retention time correction.

***Table 2. Retention Time Correction Benchmarking***

| Mass Exclusion Time Window (*w*) | Number of most recent predicted retention time errors* used for retention time correction (*M*) | Fraction of peptides within *w* minutes of predicted retention time without retention time correction | Fraction of peptides within *w* minutes of predicted retention time using Mean correction | Fraction of peptides within *w* minutes of predicted retention time using Median correction | Fraction of peptides within *w* minutes of predicted retention time using Weighted Mean correction |
|---|---|---|---|---|---|
| 0.25 | 100 | 0.00098687 | 0.05595263 | 0.06946978 | 0.06662879 |
| 0.5 | 100 | 0.00164479 | 0.11136697 | 0.12404677 | 0.13256975 |
| 0.75 | 100 | 0.0023027 | 0.16636263 | 0.17955082 | 0.19528096 |
| 1 | 100 | 0.00305033 | 0.21914531 | 0.23442687 | 0.25867998 |
| 1.5 | 100 | 0.00433625 | 0.32240797 | 0.34136786 | 0.37294776 |
| 2 | 100 | 0.00568199 | 0.41825413 | 0.43604773 | **0.48216155** |
| 0.25 | 500 | 0.00098687 | 0.05681988 | 0.05909268 | 0.06597087 |
| 0.5 | 500 | 0.00164479 | 0.11304166 | 0.11510512 | 0.12811388 |
| 0.75 | 500 | 0.0023027 | 0.17060917 | 0.17452675 | 0.19076527 |
| 1 | 500 | 0.00305033 | 0.2248273 | 0.2311672 | 0.25269894 |
| 1.5 | 500 | 0.00433625 | 0.33191782 | 0.33849696 | 0.36672747 |
| 2 | 500 | 0.00568199 | 0.42818266 | 0.43544962 | 0.47019947 |
| 0.25 | 1000 | 0.00098687 | 0.05676007 | 0.0598104 | 0.06594097 |
| 0.5 | 1000 | 0.00164479 | 0.11411825 | 0.11597237 | 0.1270971 |
| 0.75 | 1000 | 0.0023027 | 0.17028021 | 0.17273244 | 0.19082508 |
| 1 | 1000 | 0.00305033 | 0.22521607 | 0.2265618 | 0.25114387 |
| 1.5 | 1000 | 0.00433625 | 0.32964503 | 0.33284488 | 0.36669757 |
| 2 | 1000 | 0.00568199 | 0.42689674 | 0.42483328 | 0.46726876 |
| 0.25 | 5000 | 0.00098687 | 0.01916923 | 0.02093364 | 0.0618738 |
| 0.5 | 5000 | 0.00164478 | 0.0377104 | 0.04120936 | 0.1249738 |
| 0.75 | 5000 | 0.00230270 | 0.0578665 | 0.06220281 | 0.1886719 |
| 1.0 | 5000 | 0.00305033 | 0.0784413 | 0.08540925 | 0.2483925 |
| 1.5 | 5000 | 0.00433625 | 0.1246747 | 0.13313795 | 0.3598791 |
| 2.0 | 5000 | 0.00568198 | 0.1747958 | 0.18514309 | 0.4629624 |
| 0.25 | 5000 | 0.00098687 | 0.0191692 | 0.02093364 | 0.0618738 |
| 0.5 | 5000 | 0.00164478 | 0.0377104 | 0.04120936 | 0.1249738 |
| 0.75 | 5000 | 0.00230270 | 0.0578665 | 0.06220281 | 0.1886719 |
| 1.0 | 5000 | 0.00305033 | 0.0784413 | 0.08540925 | 0.2483925 |
| 1.5 | 5000 | 0.00433625 | 0.1246747 | 0.13313795 | 0.3598791 |

***** Predicted retention time errors are calculated as the difference between a peptide's predicted retention time and the time the peptide was identified by the mass spectrometer.

Masses of peptides from proteins deemed confidently identified were excluded, with a precursor mass tolerance of 3 ppm, for 1.5 minutes around their offset-corrected predicted peptide retention time. With this approach, 0.0127% of peptide-spectrum matches fell within 1.5 minutes of their predicted retention time before retention time correction, compared to 28.5% with calibration (Figure 7). Although a mass exclusion window of 2.0 minutes achieved the greatest fraction of peptides within their predicted retention time window, results using a mass exclusion window of 1.5 minutes was used to prevent too large of a mass exclusion window from negatively affecting the overall performance of the algorithm.

## 4.3 Assessing the Machine Learning Algorithm Performance

The goal of our project is to be able to identify a similar number of proteins using fewer MS resources. To assess the performance of our algorithm, we simulate MS data acquisition on a previously completed 120-minute LC-MS/MS analysis of a HEK293 cell lysate acquired using a conventional dynamic exclusion strategy. Our software processes the spectra in the order they were originally acquired. For every MS2 spectra, our machine learning-guided exclusion algorithm makes decisions whether to analyze or exclude the acquired spectrum. Therefore, for every simulated MS experiment, we calculate the number of MS2 spectra included for analysis in the simulation as well as the number of confidently identified proteins using these MS2 spectra. These results are compared against a simulated data acquisition in which no MS2 spectra are excluded from analysis. Algorithm performances are assessed using the fraction of resources used for protein identification and the protein identification fold change, the number of proteins identified using fewer resources compared to using all available spectra for protein identification. The algorithm's performance is therefore benchmarked against that of a conventional dynamic exclusion strategy.

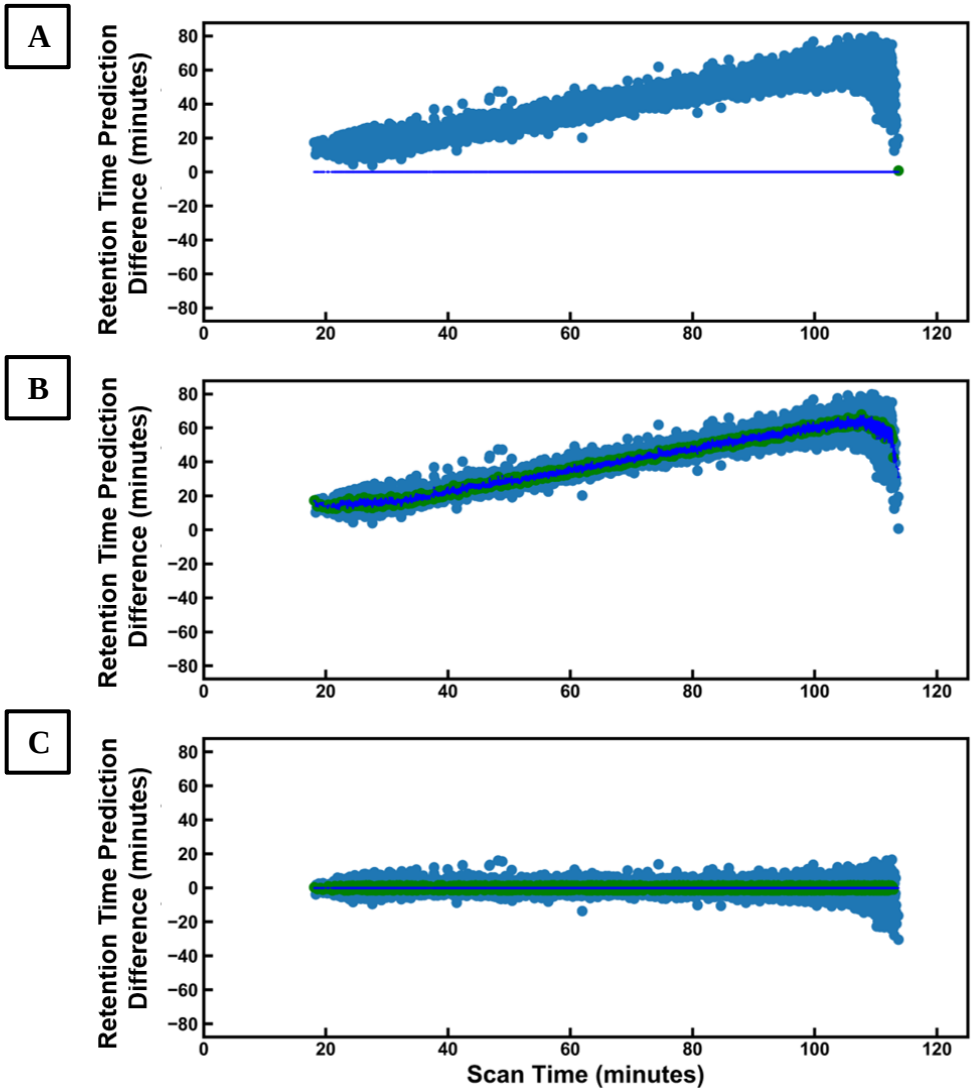


***Figure 7. On-the-fly Predicted Peptide Retention Time Correction During Simulated Real-time Dynamic Exclusion***

Predicted peptide retention time was corrected using experimental data from a 120-minute LC-MS/MS experiment. Retention time prediction error is shown for each peptide-spectrum matches at its experimental time in the top panel. Peptides falling within 1.5 minutes of predicted retention time are shown in green. Using the retention time prediction errors, a time offset was calculated at each time point, using a weighted mean of at most the 100 most recent retention

time prediction errors shown in blue in the middle panel. This value is added to the predicted peptide retention time at each time point and shown in the bottom panel.

This experimental simulation was repeated using a variety of algorithm settings, including mass exclusion time window ranging from 0.5, 0.75, 1.5, and 2.0 minutes, a ppm mass tolerance of 3, 5, and 10 ppm, and a protein identification probability threshold of 0.1, 0.3, 0.5, 0.7, and 0.9. Our logistic regression classifier was benchmarked against a heuristic exclusion approach for assessing protein identification based on identifying *t* PSMs with a XCorr threshold *x*, where *t* and *x* were set to 2, 3 and 4 and 2.5, 3, and 3.5, respectively (see methods). The heuristic simulations were performed with the same ppm mass tolerance and mass exclusion time window described above.

Each machine learning guided exclusion and each heuristic guided exclusion simulation is matched by a set of ten simulations with the same number of MS2 spectra that were excluded at random. Figure 8 shows results from the simulations of the 120-minute LC-MS/MS experiment with the above algorithms. Algorithm performances for all experimental strategies are assessed using the fraction of MS2 spectra used (i.e. not excluded) for protein identification and the fold-change of the number of proteins identified using only non-excluded resources over the total number of protein identified at a 1% FDR using ProteinProphet[27] using all available spectra. Our machine learning guided exclusion algorithm resulted in the identification of 60.5 to 97.4% of the originally identified proteins with all MS2 spectra, while using between 33.8 to 83.8 % of the available MS resources (Figure 8). The machine learning-guided exclusion simulation with the highest protein identification fold-change identified 2.6% fewer proteins, while using 16.2% fewer MS2 spectra.

In an actual MS experiment, these saved resources would be available to acquire MS2 spectra corresponding to lower abundance proteins. When compared to randomly excluding the same number of spectra, our algorithm consistently identifies more proteins (Figure 8).

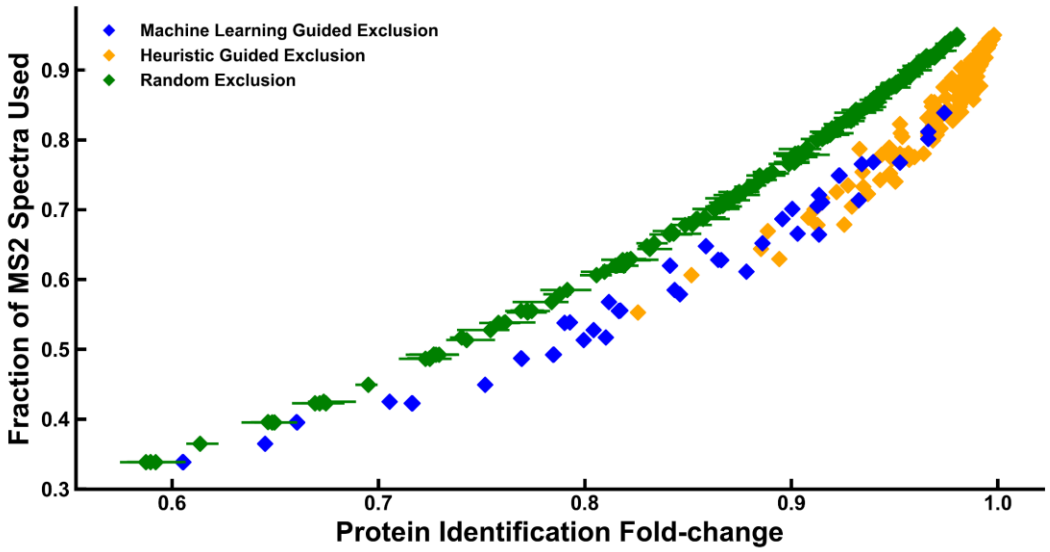


***Figure 8. Fraction of MS2 Spectra Used and Protein Identification Sensitivity in Real-time Dynamic Exclusion for Different Exclusion Strategies***

Fraction of MS2 spectra used and protein identification fold-change for the Machine Learning-guided (blue), heuristic-guided (orange) and random (green) exclusion algorithms. The average protein identification sensitivity for randomly excluded simulations are shown with their minimum and maximum range represented by the error bars.

Our machine learning exclusion strategy can save far greater MS resources at the cost of identifying fewer proteins in some cases, when compared to the heuristic exclusion strategy. Nevertheless, the heuristic approach is the one that identified the most proteins.

Next, we assess the correctness of the machine learning algorithm's peptide exclusion. To assess the accuracy of our algorithm's peptide exclusion, we determine the number of correct and incorrect peptide exclusions. A correct exclusion is defined as an MS2 spectrum that is excluded because its corresponding peptide has indeed been added to the exclusion list. An incorrect exclusion on the other hand corresponds to an MS2 spectrum for which the corresponding peptide was not in the exclusion list, but its precursor ion mass was within the ppm mass tolerance of another excluded peptide and they shared similar retention time predictions. During the simulated MS experiment, for every MS2 spectra excluded from analysis, we search if its corresponding peptide sequence is on the software's exclusion list. It is deemed a correct exclusion if its peptide sequence is found. Figure 9 depicts the number of correct and incorrect exclusions of spectra. Figure 9 shows that for our simulated experiments, a smaller ppm mass tolerance results in a similar number of correctly excluded spectra with significantly fewer incorrectly excluded spectra than a large ppm mass tolerance. For example, the experiment with the greatest proportion of correct exclusions (3-ppm experiment correctly excluded 9824 spectra and incorrectly excluded 4192 spectra) correctly excluded a greater number of spectra with half as many incorrect exclusions when compared to the experiment with the smallest proportion of correct exclusions (10-ppm experiment correctly excluded 9039 spectra and incorrectly excluded 8337 spectra). This result is expected since increasing the ppm mass tolerance for the masses on the exclusion list allows the exclusion from fragmentation of more ions with a similar mass to those on the exclusion list. Figure 9 also illustrates that larger

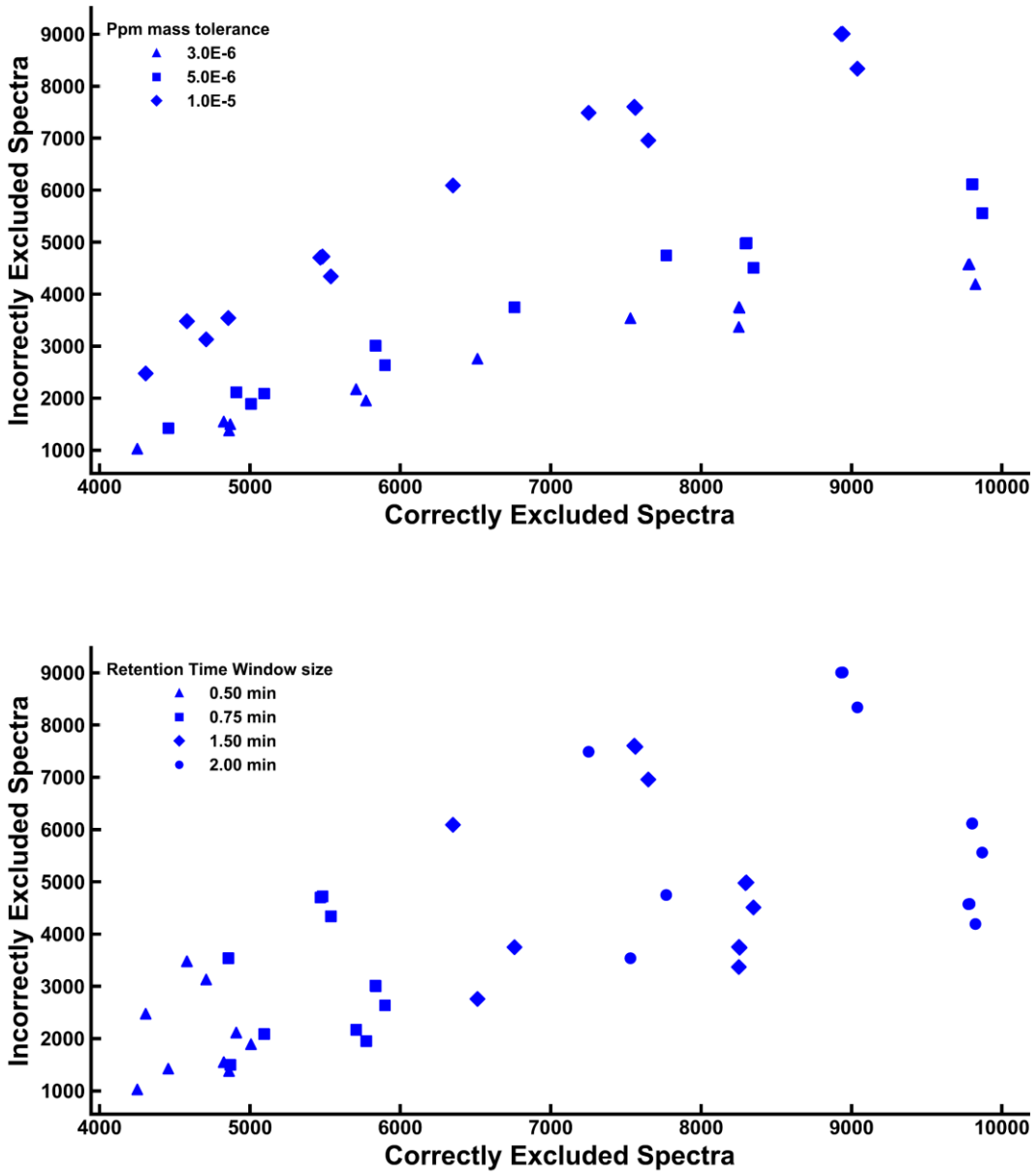


***Figure 9. Peptide Dynamic Exclusion Quality Assessment***

Number of incorrectly excluded MS2 spectra and correctly excluded MS2 spectra with our machine learning-guided exclusion executed under different algorithms parameters (ppm mass tolerance, mass exclusion time window size, and probability threshold). Differing ppm mass

tolerances are labeled in the top panel and differing mass exclusion time window sizes are labeled in the bottom panel.

mass exclusion time windows result in a greater number of overall spectra excluded. Mass exclusion time window size does not appear to have a strong influence on the number of incorrect versus correct exclusions made by our algorithm. This is a little bit more surprising since, increasing the mass exclusion time window size of the masses in the exclusion list increases the duration around a peptide's predicted retention time for which a peptide will be excluded, resulting in more spectra being excluded overall. When comparing the greatest number of overall spectra excluded for each mass exclusion time window, the largest mass exclusion time window of 2 minutes (excluding 17947 total spectra) excluded over twice as many spectra as the smallest mass exclusion time window of 0.5 minutes (excluding 8060 total spectra).

**4.4 Runtime Analysis**

For our algorithm to be utilized to its full potential during an MS experiment, it must be able to assess protein identification confidence and update the instrument's exclusion list within a reasonable timespan, i.e. at least as fast as MS2 spectra are acquired. For every simulated MS experiment, our software tracks the computational time for the algorithm to assess protein identification confidence and construct an exclusion list as well as tracks the total number of peptides excluded from analysis. Figure 10 shows our algorithm running time for simulations under the different mass tolerance and mass exclusion time window parameters listed above. As expected, simulations with more lenient parameters result in a larger number of peptides added to the exclusion list, thereby increasing the time required for our simulation. Depending on the number of peptides added to the exclusion list, our algorithm takes between 1,867 to 6,134 seconds of total computational time to assess protein identification confidence, decide whether an MS2 spectrum should be excluded or processed, and update the dynamic exclusion list. While

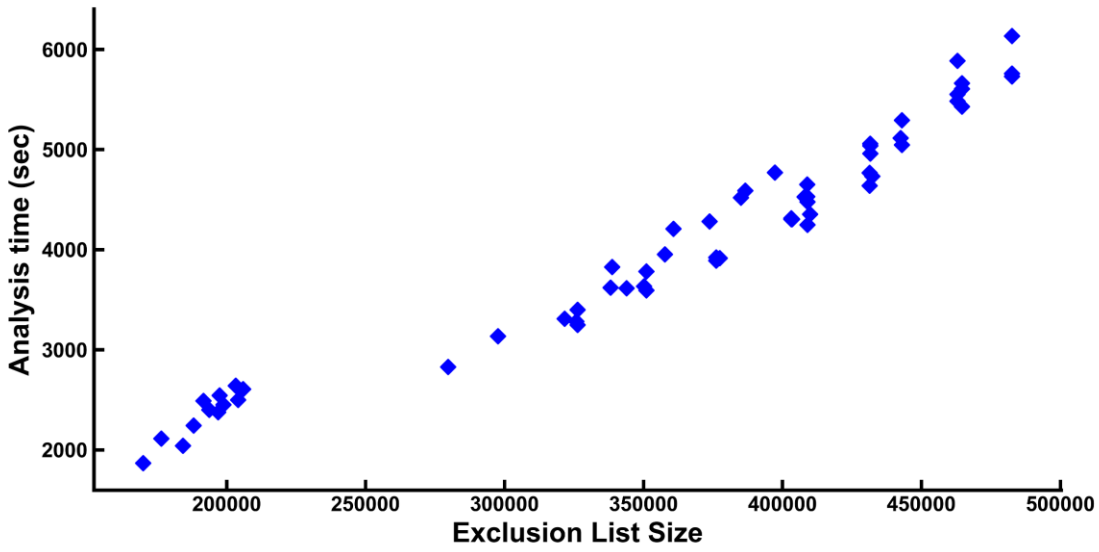


***Figure 10. Algorithmic Running Time for Machine Learning Guided Exclusion***

The software package running time as a function of the total number of peptides added to the exclusion list during a simulated experiment is assessed for multiple simulated experiments using varying algorithms parameters (ppm mass tolerance, mass exclusion time window size, and probability threshold).

the compiled running time does not include the time necessary for on-the-fly database search of MS2 spectra nor the latency involved with communicating an updated exclusion list to the mass spectrometer, all of the simulations are well below the 7,200 second duration of the analyzed MS dataset.

To verify that our algorithm does not derive a vast majority of its saved resources from a handful of highly abundant proteins and does not bias its exclusion towards these proteins, our software tracks the number of MS2 spectra excluded from analysis per confident protein identification. We therefore investigated the abundance of the proteins from which MS2 spectra are excluded with our algorithm. If most of the excluded spectra originated from a very small number of highly abundant proteins, this would indicate a similar effect could be achieved merely by arbitrarily excluding a handful of highly abundant proteins. Figure 11 shows the number of MS2 spectra that are excluded from analysis from each protein in a given experiment, with proteins sorted in increasing abundance, based on spectral counts. While these results are shown for a single experiment this trend is reproduced between simulated experiments with machine learning guided exclusion under different parameters (data not shown). This illustrates that our exclusion strategy does not derive a vast majority of its saved resources from a handful of highly abundant proteins but appears to exclude peptides from a wide array of proteins detected in the experiment.

## 4.5 Simulated Limited Data Dependent Acquisition

To assess our machine learning algorithm's potential in identifying a greater number of proteins than traditional data acquisition strategies currently being used, we built a framework that allows to estimate potential gains from MS2 spectra exclusion in in silico MS simulations. In our 120-minute LC-MS/MS experiment, the mass spectrometer acquisition was using a Top 12

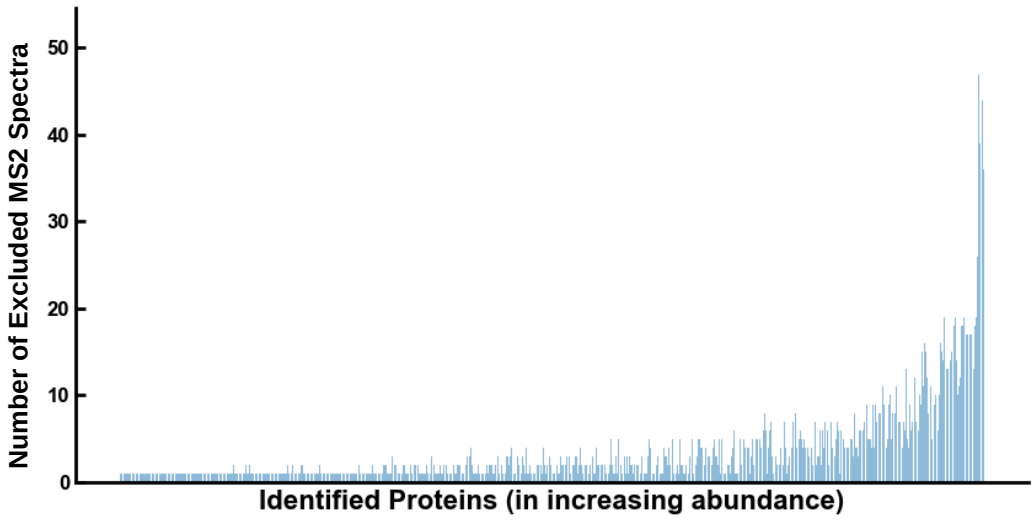


***Figure 11. Frequency of Saved Resources from Identified Proteins from Simulated Real-time Dynamic Exclusion***

The number of MS2 spectra excluded per proteins using the machine learning guided dynamic exclusion algorithm on the 120-minute LC-MS/MS dataset. Peptides from proteins achieving a protein identification confidence above 90% were excluded for 1.5 minutes around its peak retention time. MS2 spectra with a precursor mass within 3 ppm of a mass on the exclusion list were excluded. Identified proteins on the x-axis are sorted in order of lowest to highest abundance based on their spectral counts in the original experiment. There were 2551 proteins identified in total at a 1% FDR.

DDA approach. This means that (at maximum) 12 MS2 spectra were previously acquired for a MS1 spectrum. Further improvements in protein identification sensitivity can therefore be estimated artificially simulating that a limited Top *N-X* DDA approach was used to acquire MS data, therefore allowing the machine learning algorithm to explore the remaining *X* MS2 spectra if some of the top *N-X* are excluded.

Figure 12 and 13 represents results using the same experiment with a Top 6 and a Top 5 DDA respectively. These values were chosen since 6 represents half of the original Top 12 DDA approach and 5 was chosen, as it is an acquisition strategy often used in the scientific literature.[71] In the limited Top 6 DDA simulation, the machine learning-guided exclusion algorithm identified between 66.7 to 101.1% of the total number of proteins identified in the original experiment using only the top 6 most abundant MS2 spectra all of this using between 40.5 to 89.8% of the available MS resources (Figure 12). When compared to randomly excluding the same number of spectra, the machine learning algorithm consistently identified more proteins. The machine learning-guided exclusion identifies a similar number of proteins compared to the heuristic approach, while also capable of saving more MS resources in some cases. Although the number of additional proteins identified using our machine learning approach is minor, the saved MS resources could allow for future increases in the number of proteins identified.

For the limited Top 5 DDA simulation, the machine learning guided exclusion algorithm identified between 69.2 to 102.3% of the number proteins identified using only the top 5 most abundant MS2 spectra (Figure 13). Our algorithm achieves this performance while using between 44.6 to 92.2% of the available MS2 spectra. Our algorithm's performance in the Top 5 DDA simulations remains comparable to the performance in the Top 6 DDA simulations. The Top 5 DDA simulation identifies slightly more proteins in the best cases. However, nine of these

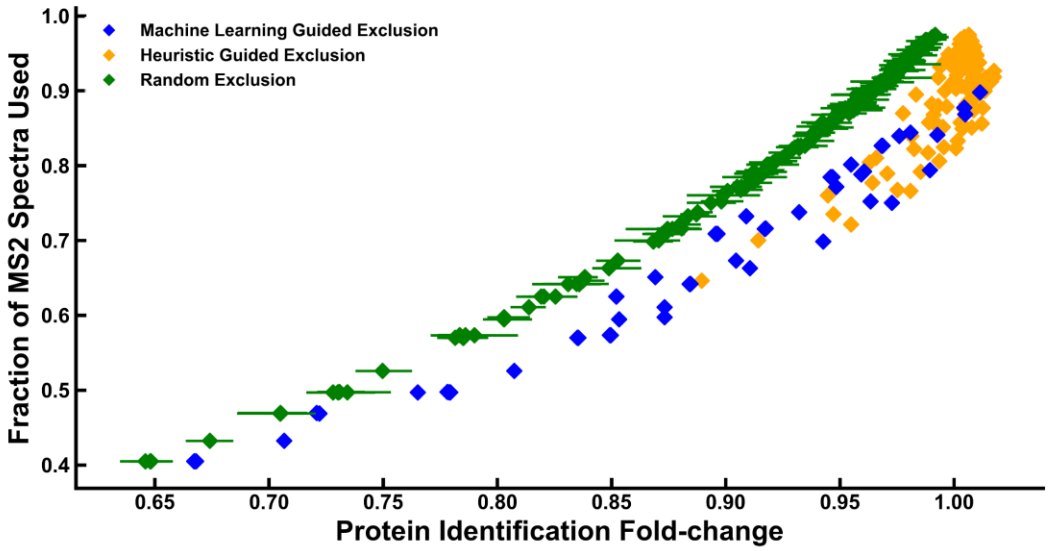


***Figure 12. Fraction of MS2 Spectra Used and Protein Identification Fold-change Simulating Limited Top 6 DDA Real-time Dynamic Exclusion for Different Exclusion Strategies***

Fraction of MS2 spectra used and protein identification fold-change for the Machine Learning-guided (blue), heuristic-guided (orange) and random (green) exclusion algorithms. The average protein identification sensitivity for randomly excluded simulations are shown with their minimum and maximum range represented by the error bars.

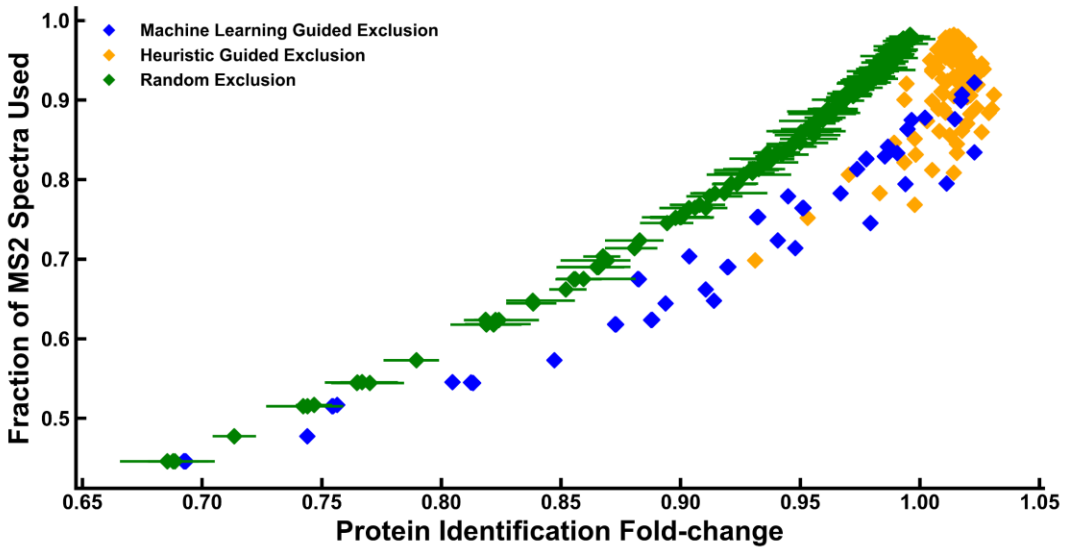


***Figure 13. Fraction of MS2 Spectra Used and Protein Identification Fold-change Simulating Limited Top 5 DDA Real-time Dynamic Exclusion for Different Exclusion Strategies***

Fraction of MS2 spectra used and protein identification fold-change for the Machine Learning-guided (blue), heuristic-guided (orange) and random (green) exclusion algorithms. The average protein identification sensitivity for randomly excluded simulations are shown with their minimum and maximum range represented by the error bars.

cases overlap with simulations using the same number of MS resources by randomly excluding spectra. These occurred for the simulations which used a ppm mass tolerance of 10ppm, probability threshold of 0.1, 0.3, and 0.5, with mass exclusion time window of 0.75, 1.5, and 2.0 minutes. These experiments use a large ppm mass tolerance and a low probability threshold which would increase the likelihood a spectrum is excluded by chance and not due to its presence on the exclusion list. As well, since there are fewer MS2 spectra acquired between each MS1 spectrum, there is a higher chance of a spectrum excluded at random will result in acquiring an additional MS2 spectrum which will contribute to a confident protein identification.

With this limited acquisition simulation approach, our algorithm can guide MS data acquisition to exclude an MS2 spectrum from analysis, and instead acquiring an MS2 spectrum not acquired by the baseline comparison. In the best case, these excluded MS2 spectra are from proteins already identified with high confidence and the newly acquired MS2 spectrum can contribute to the identification of uncharacterized proteins. When comparing the artificially limited Top 6 and Top 5 DDA to utilizing all Top 12 DDA MS2 spectra, we illustrate our algorithm's ability to repurpose its MS resources to identify a greater proportion of proteins (Figure 14). When comparing our algorithm on a simulated Top 6 to Top 12 DDA, the Top 6 experiments typically use a greater fraction of MS2 spectra relative to the maximum number of MS2 spectra available.

Our algorithm and in silico simulated data acquisition demonstrates the promise of utilizing a machine learning algorithm to guide MS data acquisition. By analyzing MS data features as they are acquired and utilizing these features to assess protein identification confidence in real-time, one can prevent the redundant analysis of peptides from proteins identified with high confidence. The MS2 spectra saved by this approach will free up resources

for the mass spectrometer to acquire data from peptides of proteins of lower abundance, thus increasing the chance of identifying them.

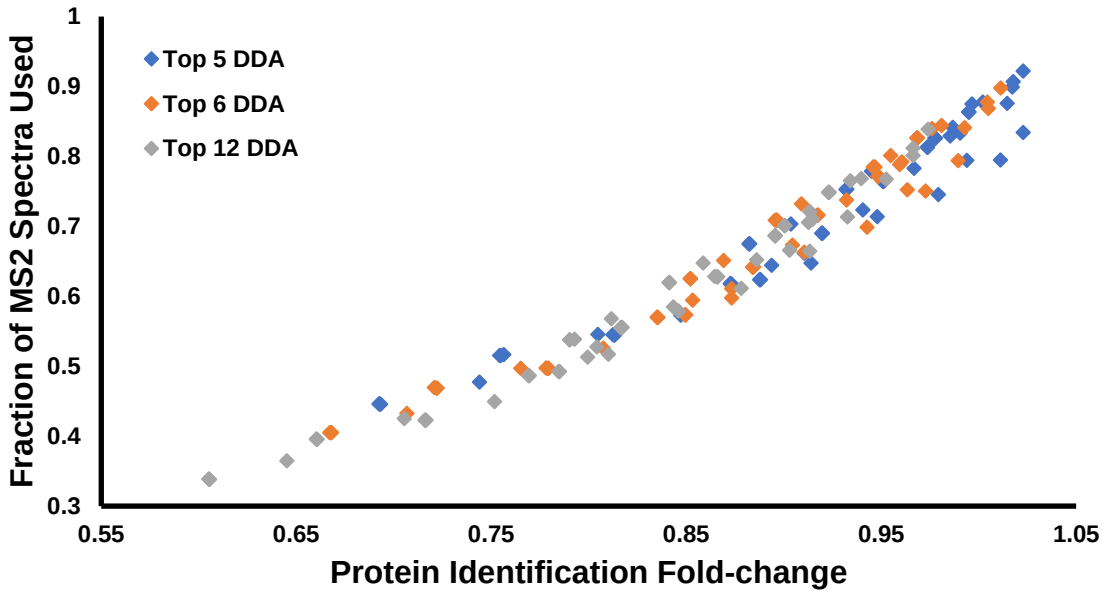


***Figure 14. Fraction of MS2 Spectra Used and Protein Identification Fold-change for Real-time Dynamic Exclusion for Differing Top N DDA Simulations***

A subset of the MS2 spectra from the original MS experiment were used for protein identification. The maximum number of MS2 spectra available between each MS1 spectra was artificially limited to 5 and 6 MS2 spectra for datapoints in blue and orange respectively. Datapoints in grey have the maximum number of MS2 spectra available between each MS1 spectra.

# 5. Discussion

I was able to design a machine learning algorithm capable of guiding MS data acquisition with the goal of identifying a similar number of proteins using fewer MS resources than conventional MS approaches. This is achieved by analyzing MS data as they are acquired by the instrument and using these data to assess protein identification confidence on-the-fly. Since the analysis of peptides from proteins identified with high confidence will not further contribute to a protein identification, we prevent further analysis of these peptides by adding their masses to a dynamic exclusion list. Peptide masses excluded in this manner will not be analyzed for a period of time around its predicted retention time. In this way, we prioritize the analysis of peptides from proteins that have not yet been confidently identified.

Although our algorithm succeeds in decreasing the amount of resources utilized during an MS experiment (e.g. using 83.3% of the original MS2 spectra, yielding a protein identification sensitivity of 97.4%), its ability to utilize these saved resources in an actual experiment remains to be determined. Several factors must be considered to ensure its optimal usage in a laboratory setting. These include 1) runtime parameters for our algorithm, 2) appropriate datasets for classifier training, 3) generalizability of our simulated MS experiments, 4) technical barriers to the deployment of this technology, and 5) algorithmic design improvements.

## 5.1 Runtime Parameters and Settings

The performance of our algorithm is dependent on a number of parameters that will affect the behaviour of our machine learning-guided MS data acquisition. These parameters include ppm mass tolerance, mass exclusion time window, probability threshold, and the number of miscleavages used for the construction of the exclusion database. Understanding the behaviour

the machine learning algorithm will take under a given set of parameters will determine its overall performance on guiding MS data acquisition.

### 5.1.1 Ppm Mass Tolerance

Firstly, the ppm mass tolerance plays a role in the decision to exclude a mass from analysis during an MS experiment. If a given mass *m* is within the ppm mass tolerance of another peptide's mass on the dynamic exclusion list, *m* will be excluded from analysis. Increasing the ppm mass tolerance is useful to account for variations in mass due to isotopic variation and instrument measurement errors. However, increasing the ppm mass tolerance too much will result in the unintentional exclusion of a peptide with a similar mass to one found on the exclusion list. On the other hand, a ppm mass tolerance that is too small may result in a fewer number of overall excluded peptides. This behaviour is illustrated in Figure 9 of the results section, as we increase the ppm mass tolerance, a greater number of overall peptides are excluded, however a greater proportion of these exclusions are due to the incorrect exclusions of a distinct peptide than one found on the exclusion list.

### 5.1.2 Mass Exclusion Time Window

Secondly, a mass on the dynamic exclusion list will be excluded for a period of time around the predicted retention time of its corresponding peptide. We refer to this period of time as the mass exclusion time window. Restricting this time window to a smaller time frame will reduce the chance the algorithm excludes from the MS analysis a distinct and novel peptide with a similar mass to that of the one on the exclusion list. Making the time window too large will result in a larger fraction of these exclusions. A time window that is too small however will result in a smaller fraction of these exclusions at the cost of fewer overall exclusions. This increases the odds of the redundant analyses of a peptide on the exclusion list. This behaviour is

illustrated in Figure 9 of the results section, as we increase the mass exclusion time window, a greater number of overall peptides are excluded from analysis.

### 5.1.3 Protein Identification Confidence Probability Threshold

Next, the algorithm uses a probability threshold for the classification of a protein as being confidently identified or not confidently identified. Our logistic regression classifier assesses the probability that a protein is confidently identified based on its MS data features acquired so far. Proteins with an identification probability above this threshold are deemed confident protein identifications. A probability threshold above 0.9 yielded a better algorithmic performance in that a good proportion of MS resources could be saved without sacrificing a large number of protein identifications. A stringent probability threshold is beneficial to decrease the likelihood that our classifier deems a protein to be confidently identified that is, in fact, not considered not confidently identified. Such proteins would be excluded from analysis for the remainder of the duration of the experiment and result in an unsuccessful identification. Although a lower probability threshold results in a greater number of MS resources being freed up for future analyses, this comes at the cost of having a greater number of unsuccessful protein identifications that would otherwise have been identified. Simulated experiments with a lower probability threshold in general yield a smaller fraction of MS2 spectra used with a smaller protein identification fold change than simulated experiments with a greater probability threshold. However, it is important to note that in these simulated experiments, the MS2 spectra excluded for analysis are not used to gather additional MS2 spectra that would otherwise be acquired in an actual MS experiment. In this way, it is difficult to estimate the true effect of the probability threshold without first applying this method online with a mass spectrometer.

### 5.1.4 Number of Miscleavages

Finally, the number of miscleavages when considering the peptide exclusion database will influence the number of peptides excluded from analysis. When a protein is deemed confidently identified, the set of masses of the peptides associated with the protein are added to the dynamic exclusion list. These peptides are determined by the in silico digestion of a protein database described in methods section. The number of allowed miscleavages affects the composition of this exclusion database. For our simulated experiments, we used 1 miscleavages.

Using 0 miscleavages yields a smaller set of peptides associated with a given protein while increasing this number will increase the number of peptides associated with a single protein. This will cause our algorithm not to exclude spectra from any peptides with miscleavages in the sample. Increasing this value will result in the number of excluded peptides overall to increase, as more peptides are added to the exclusion list. However, this will result in longer algorithmic running time due to a greater number of peptides added to the exclusion list per protein excluded from analysis and the increased time required to maintain and search this larger exclusion list.

## 5.2 Training Set Choice for Classifier

Utilizing an appropriate dataset for training our machine learning classifier is vital for the accurate classification of a protein as confidently identified or not. The training data must consist of positive examples of confidently identified proteins and negative examples of proteins which have not been identified with high confidence. Necessarily, these training examples must closely resemble positive and negative examples in the dataset to be analyzed to be useful. We trained our classifier using a HEK293 240-minute experiment with proteins identified with a 1% FDR with less than or equal to 30 spectral counts as the positive training examples and proteins

identified with greater than 2.4% FDR as the negative training examples. Proteins identified between 1 and 2.4% FDR were not included to make the negative dataset and positive dataset more distinct and not consider borderline protein identifications. However, in removing these borderline protein identifications, may cause our classifier to misclassify these borderline cases. If borderline cases are deemed confidently identified in real-time and are added to the exclusion list, peptides from those proteins are no longer analyzed. As a result, these proteins may never be confidently identified. Further iterations of this algorithm can address this issue by prioritizing MS analysis of peptides specific to these borderline protein identifications.

The dataset used to train our logistic regression classifier for protein identification confidence can be modified to yield performance increases. It is hypothesized that the features necessary for a confident protein identification in the HEK 293 240-minute experiment are similar to those of the 120-minute experiment since these experiments come from the same cell line and analyzed on the same instrument. An alternative for this training dataset would be to include merged datasets of multiple HEK 293 120-minute experiments as these may more accurately represent positive and negative examples of the dataset to be tested. Training examples can be gathered and merged from online MS dataset repositories such as PRIDE[72] to train our classifier to be more generalizable to other MS experiments on different cell types and experimental conditions.

We also considered the need to re-train our classifier periodically throughout the experiment. The initial feature weights for assessing protein identification confidence may not properly identify proteins with high confidence if our training data was not be generalizable to current experiment. As a result, our classifier might be assessing protein identification confidence poorly. This can be resolved by specifying time points to retrain our logistic

regression model based on the data accumulated so far. The MS data up to that time point could be passed through the protein identification pipeline to generate a new positive and negative training dataset with similar criterion for the original training dataset. This new training dataset could be used to train a new logistic regression classifier for estimating protein identification confidence until the next timepoint. Since this step is computationally intensive, it should be implemented in parallel with respect to the machine learning-guided exclusion pipeline.

## 5.3 Generalizability of Simulated Results

Our algorithm's performance and the ability of our algorithm to identify additional proteins is estimated using a previously acquired MS experiment simulated in silico. The saved resources from our machine learning-guided exclusion can only demonstrate the potential for these resources to be re-utilized in an MS experiment to identify additional proteins. The extent to which our algorithm improves on the protein identification sensitivity in an MS experiment was assessed by artificially limiting the number of MS2 spectra between each MS1 spectra available for protein identification. This allows more MS resources to become available when preventing the analysis of redundant peptides. However, the additionally available spectra from the simulated dataset are quickly depleted and our simulation runs out of potential MS2 spectra resources to analyze when other spectra are excluded. The overall increase in protein identification is hence restricted and it becomes challenging to accurately estimate an increase in protein identification rate by a simulated MS experiment.

In short, to further assess the potential for our algorithm in identifying more proteins, we would require assessing its performance using a mass spectrometer to guide data acquisition in real-time.

## 5.4 Technical Barriers

Of the next steps for the further development of the method presented in this thesis is to apply this algorithm with a mass spectrometer to guide its data acquisition in real time. To realize this, one must implement on-the-fly protein sequence database search capabilities, acquire the proper API to communicate with the mass spectrometer, and consider run-time limitations of the pipeline.

For the algorithm to guide MS data acquisition on-the-fly, MS data must be analyzed as they are acquired in order for our algorithm to use the resulting data features. Currently, our data originates from a previously complete MS experiment and the protein sequence database search is performed offline, prior to the simulation of the MS experiment. When implemented in the context of a live MS experiment, mass spectra must be continuously collected from the mass spectrometer, converted to a usable file format, searched against a peptide database, and the necessary data features for our algorithm to assess protein identification confidence must be extracted. This can be achieved using a real-time database search platform such as Orbiter.[45]

For our algorithm to properly guide MS data acquisition, it must communicate to the mass spectrometer and update its dynamic exclusion list to prevent the analysis of these peptides. Both the acquisition of MS data for protein identification and the communication of a dynamic exclusion list is made possible by the API to communicate with the instrument. Some vendors, such as ThermoFisher, can allow their instruments to be controlled in this way. These APIs are vendor and instrument specific. For our software to be compatible with the mass spectrometer's API, specific programming languages must be considered. For example, the ThermoFisher instrument API for the Q-Exactive mass spectrometer is limited to the C# programming language. As well, any 3$^{rd}$ party software used for database search and statistical assessment

must be able to analyze data as it is acquired. For example, a version of the Comet[64] database search tool is capable of single spectrum searching against a protein database. In this way, any MS resources our algorithm saves can be used in real-time to acquire mass spectra from proteins of lower abundance and potentially improve protein identification sensitivity.

Currently, our software package takes around 2-6000 seconds of total computational time depending on its hyperparameters. This computational time includes the time it takes for assessing protein identification confidence and maintaining a dynamic exclusion list. This however, does not include the time it will take for acquiring mass spectra from the mass spectrometer, performing a database search of these mass spectra, the communication of an updated dynamic exclusion list, nor any network latency from communication between the software and mass spectrometer. All these factors must be taken into account to more accurately estimate the necessary computational time and address any technical limitations of our approach.

For our algorithm to perform at its full potential, it must perform all analysis before the next set of mass spectra are acquired. For example, with a Q-Exactive mass spectrometer, MS1 spectra can be acquired every second, during which 10 to 20 MS2 spectra can be generated. The software package therefore needs to make decisions in a similar timespan. If these cannot be achieved within the available time, compromises will have to be made in order for the software to run in real-time. For example, occasionally skipping the analysis of a set of spectra to give the program time to complete the analysis of the previous data.

A significant fraction of the computational time for the program is used to check if a peptide is on the program's exclusion list. The greatest algorithmic running time improvements can be achieved by improving the exclusion list's data structure efficiency, implementing parallelization or multi-threading of the code. Runtime particularly suffers when a very large

number of peptides are added to the exclusion list. A query peptide mass must be searched against the exclusion list to determine if the mass should be excluded from MS2 spectra acquisition. Our implementation utilizes a modified binary search algorithm to query if a mass is on the exclusion list. The runtime can be improved by breaking up the large exclusion list into many smaller exclusion lists binned based on their mass and querying multiple MS2 spectra at one time.

Another running time consideration to take into account is the speed of proteomics database search algorithms used for on-the-fly processing of MS data. Our plans moving forward with the algorithm development is to use the Orbiter real-time search platform[45] to search mass spectra as they are acquired by the mass spectrometer. In this thesis, our database search results were run prior to simulated MS data acquisition so the run-time requirements for these database searches to be performed on-the-fly have not been assessed. While we believe this tool will be able to perform reasonably well when applied in real time, we have considered the use of alternative database search algorithms capable of quick database searches. For example, MSFragger[73] achieves this by using fragment ion indexing to allow speedy database search and peptide confidence scoring. Although MSFragger performs well in terms of speed, one limitation of the tool is the memory storage necessary for the pre-processing and indexing of the protein database. This results in large database index files that would need to be stored on the appropriate computer receiving MS files in real-time.

## 5.5 Algorithmic Improvements

Further development on certain aspects of our algorithm may yield improved algorithmic performances. This can be achieved by addressing the issue of peptides that have been

incorrectly excluded from analysis, the accuracy of our peptide retention time predictions, and further improvements regarding our protein identification confidence assessment.

Our algorithm analyzes MS data as they are acquired and uses these data to assess if a protein is confidently identified. Peptides from proteins identified with high confidence are added to the dynamic exclusion list to prevent the analysis of these peptides. However, our algorithm can incorrectly exclude a peptide from analysis when assessing if the mass spectrometer should analyze or exclude a given *m/z* value. A peptide can be incorrectly excluded if there exists a distinct peptide with a similar mass and retention time that has been added to the dynamic exclusion list. As discussed in the results section, the peptide retention time prediction performed by RTCalc[66] and other retention time prediction tools struggle to accurately estimate peptide retention times due to differences in relative protein abundance, chromatographic separation, and MS instrument variation. Retention time prediction tools are typically capable of predicting the order in which peptides elute off the liquid chromatography column, however most tools struggle to accurately predict the exact time at which a peptide will indeed enter the mass spectrometer. It is for this reason we decided to adaptively calculate the retention time of a peptide with our retention time prediction correction algorithm. Referenced in Figure 6, after retention time correction, performance improved from 0.57% to 48.2% of peptides within 2 minutes of their predicted retention time, however a large portion of incorrectly predicted peptide retention time remains. The use of other retention time prediction tools can be used in place of RTCalc, for example SSRCalc.[74,75]

RTCalc uses artificial neural networks to estimate a peptide's retention time based on its amino acid sequence. The parameters of the activation functions used by the tool were provided with the software. The accuracy of our peptide retention time estimation may be achieved by

retraining the neural network with data from MS experiments similar to the biological sample to be analyzed with our pipeline with similar length of MS experiment and chromatographic separation. If the training dataset used for RTCalc is more specific to the dataset tested, better accuracy of peptide RT prediction could potentially be achieved. However, this may not be practical for MS experiments with a limited set of samples available.

Improvements in how our algorithm decides which peptides to include and exclude from analysis by the mass spectrometer can also be implemented by a soft-thresholding approach for exclusion of a peptide around its predicted retention time. Currently, the algorithm excludes a mass from MS2 spectrum acquisition if it is on the dynamic exclusion list and within a certain time range of its predicted retention time. Instead of using a strict cut-off time window around its retention time, the algorithm could instead exclude a peptide more frequently the closer it is to its retention time.

The performance of our algorithm could also be improved by considering supplemental features for assessing protein identification confidence. Some features available from the Comet database search tool not used by our classifier include Delta CN,[19] Delta CN*,[59] and E-value.[64] Delta CN describes the difference between the first and second candidates of a theoretical peptide spectrum to an observed spectrum, while Delta CN* describes the difference between the first and fifth candidates. The E-value is the expectation value of the confidence score based on the score distribution of all confidence scores. The addition of some of these features and their cross-features to the classifier may yield a better performance. While adding extra features increases the chance of overfitting to the training dataset, the number of features remain small, and therefore, this risk remains minimal

Additionally, performance improvement may be implemented with a different classification technique. Currently, we use a logistic regression classifier for estimating protein identification confidence. We chose a logistic regression classifier over other classification techniques in machine learning because its computation is fast, can be adjusted with a probability threshold, and is easy to interpret. However, there could be other classification techniques that may yield better performances. One such algorithm is the support vector machine. This algorithm fits a decision boundary between its positive and negative training sets, maximizing the difference between the decision boundary and each set. This algorithm would result in a classification of a protein as confidently identified or not as opposed to a probabilistic model we have from the logistic regression classifier. With the logistic regression classifier, we can fine-tune the probability threshold based to control how stringent we want the classifier to perform. The support vector machine algorithm could be advantageous as it would remove the requirement for a user-defined probability threshold.

## 5.6 Future Directions

Our goal for this project was to develop an algorithm capable of identifying a similar number of proteins using fewer MS resources compared to conventional MS data acquisition strategies. While the confident identification of proteins in a given sample is important for understanding its role in a biological context, the quantification of proteins in different cellular conditions can yield further insight.

In MS-based proteomics, the mass spectrometer can estimate the relative abundance of proteins in a sample based on the acquired mass spectra. This can be achieved based on the number of MS2 spectra acquired for a peptides corresponding to a given protein, called spectral counting, or by inferring its quantification based on the change in intensity of the *m/z* as it varies

over time in the acquired MS1 spectra for peptides corresponding to a given protein, called extracted ion chromatography.[76] The spectral counting method relies on the assumption that peptides from more abundant proteins will be analysed more often than less abundant proteins. However, since our algorithm guides MS data acquisition and excludes the analysis of MS2 spectra from confidently identified proteins, this is inconsistent to the spectral count assumption and thus spectral counting cannot be used as a proxy for relative protein quantification. However, Extracted Ion Chromatography, can still be used since it relies on MS1 spectra. The quantification is calculated based on the area under the curve of the intensity of a given peptide's chromatogram. Peptides of higher abundance will have a greater area under the curve under its chromatogram across its retention times. Nevertheless, the variance of the intensity measurements for the different peptides forming a protein can vary significantly. It is therefore often necessary to quantify multiple peptides of a protein to obtain an accurate quantification value for it. Once again, proteins that are more abundant typically meet this requirement for accurate quantification. However, this is often not the case for proteins of lower abundance, which have much less peptides identified in a typical MS experiment. Our approach could therefore be used to prevent the acquisition of data from peptides of proteins that are deemed confidently quantified to allow the mass spectrometer to allow the acquisition of data from peptides of less abundant proteins and favour, in turn, their quantification.

In addition to the algorithmic improvements previously discussed, our machine learning guided dynamic exclusion can be utilized with a specific research goal in mind. Our algorithm currently saves MS resources by dynamically excluding peptides from proteins identified with high confidence. Future iterations of this method may include the reverse of a dynamic exclusion list, instead utilizing a peptide inclusion list indicating peptides from a set of peptides of interest.

By excluding from analysis peptides from proteins identified with high confidence, the saved MS resources can be repurposed to analyze peptides from such an inclusion list. In this way, our algorithm can guide MS data acquisition towards a specific set of peptides of interest. These peptides of interest may include peptides that are unique to certain protein sequences, pathways or species. Indeed, many peptides analyzed via MS are peptides that belong to more than one protein. For example, in a mouse lung-tissue, these shared peptides can comprise up to 50% of the peptides in a protein database.[77] This is due to a high number of homologous proteins and protein isoforms.[38] As a result, these peptides are less informative for protein identification than peptides that are unique to exactly one protein, pathway, or species. The identification of these uniquely identifying peptides has the potential to increase the number of unambiguous proteins or pathways identified in an MS experiment. The identification of these protein or pathway identifying peptides could increase the confidence that a given molecular pathway is active in a given MS experiment. Finally, in metaproteomics MS experiments are performed on samples containing a number of distinct microbial species. It is difficult to confidently identify a specific species if no unique peptides of that species are detected. The identification of peptides specific to a given species could therefore improve the bacterial species coverage of metaproteomics experiments and provide a better understanding of their composition.

# 6. References


1. Desiere F. The PeptideAtlas project. *Nucleic Acids Res*. 2006;34(90001):D655-D658. doi:10.1093/nar/gkj040
2. Lane L, Argoud-Puy G, Britan A, et al. neXtProt: a knowledge platform for human proteins. *Nucleic Acids Res*. 2012;40(D1):D76-D83. doi:10.1093/nar/gkr1179
3. Schwenk JM, Omenn GS, Sun Z, et al. The Human Plasma Proteome Draft of 2017: Building on the Human Plasma PeptideAtlas from Mass Spectrometry and Complementary Assays. *J Proteome Res*. 2017;16(12):4299-4310. doi:10.1021/acs.jproteome.7b00467
4. Kelleher NL. A cell-based approach to the human proteome project. *J Am Soc Mass Spectrom*. 2012;23(10):1617-1624. doi:10.1007/s13361-012-0469-9
5. Huttlin EL, Bruckner RJ, Paulo JA, et al. Architecture of the human interactome defines protein communities and disease networks. *Nature*. 2017;545(7655):505-509. doi:10.1038/nature22366
6. Krogan NJ, Cagney G, Yu H, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*. 2006;440(7084):637-643. doi:10.1038/nature04670
7. Breitkreutz A, Choi H, Sharom JR, et al. A global protein kinase and phosphatase interaction network in yeast. *Science*. 2010;328(5981):1043-1046. doi:10.1126/science.1176495
8. Rauniyar N, Subramanian K, Lavallée-Adam M, Martínez-Bartolomé S, Balch WE, Yates JR. Quantitative Proteomics of Human Fibroblasts with I1061T Mutation in Niemann-Pick C1 (NPC1) Protein Provides Insights into the Disease Pathogenesis. *Mol Cell Proteomics*. 2015;14(7):1734-1749. doi:10.1074/mcp.M114.045609
9. Subramanian K, Rauniyar N, Lavalleé-Adam M, Yates JR, Balch WE. Quantitative Analysis of the Proteome Response to the Histone Deacetylase Inhibitor (HDACi) Vorinostat in Niemann-Pick Type C1 disease. *Mol Cell Proteomics*. 2017;16(11):1938-1957. doi:10.1074/mcp.M116.064949
10. Witze ES, Old WM, Resing KA, Ahn NG. Mapping protein post-translational modifications with mass spectrometry. *Nat Methods*. 2007;4(10):798-806. doi:10.1038/nmeth1100
11. Nørregaard Jensen O. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Curr Opin Chem Biol*. 2004;8(1):33-41. doi:10.1016/j.cbpa.2003.12.009
12. Brönstrup M. Absolute quantification strategies in proteomics based on mass spectrometry. *Expert Rev Proteomics*. 2004;1(4):503-512. doi:10.1586/14789450.1.4.503
13. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature*. 2003;422(6928):198-207. doi:10.1038/nature01511
14. Swaney DL, Wenger CD, Coon JJ. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J Proteome Res*. 2010;9(3):1323-1329. doi:10.1021/pr900863u
15. Tsiatsiani L, Heck AJR. Proteomics beyond trypsin. *FEBS J*. 2015;282(14):2612-2626. doi:10.1111/febs.13287
16. Gosetti F, Mazzucco E, Zampieri D, Gennaro MC. Signal suppression/enhancement in high-performance liquid chromatography tandem mass spectrometry. *J Chromatogr A*.

2010;1217(25):3929-3937. doi:10.1016/j.chroma.2009.11.060
17. Taylor PJ. Matrix effects: the Achilles heel of quantitative high-performance liquid chromatography–electrospray–tandem mass spectrometry. *Clin Biochem*. 2005;38(4):328-334. doi:10.1016/j.clinbiochem.2004.11.007
18. Hunt DF, Yates JR, Shabanowitz J, Winston S, Hauer CR. Protein sequencing by tandem mass spectrometry. *Proc Natl Acad Sci*. 1986;83(17):6233-6237. doi:10.1073/pnas.83.17.6233
19. Eng JK, McCormack AL, Yates JR. 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(11):976-989. doi:10.1016/1044-0305(94)80016-2
20. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*. 2004;20(9):1466-1467. doi:10.1093/bioinformatics/bth092
21. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. 1999;20(18):3551-3567. doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2
22. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods*. 2007;4(3):207-214. doi:10.1038/nmeth1019
23. Elias JE, Gygi SP. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. In: ; 2010:55-71. doi:10.1007/978-1-60761-444-9_5
24. The M, Tasnim A, Käll L. How to talk about protein-level false discovery rates in shotgun proteomics. *Proteomics*. 2016;16(18):2461-2469. doi:10.1002/pmic.201500431
25. Aggarwal S, Yadav AK. False Discovery Rate Estimation in Proteomics. In: ; 2016:119-128. doi:10.1007/978-1-4939-3106-4_7
26. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem*. 2002;74(20):5383-5392. http://www.ncbi.nlm.nih.gov/pubmed/12403597.
27. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Anal Chem*. 2003;75(17):4646-4658. doi:10.1021/ac0341261
28. Brosch M, Yu L, Hubbard T, Choudhary J. Accurate and Sensitive Peptide Identification with Mascot Percolator. *J Proteome Res*. 2009;8(6):3176-3181. doi:10.1021/pr800982s
29. Spivak M, Weston J, Bottou L, Käll L, Noble WS. Improvements to the Percolator Algorithm for Peptide Identification from Shotgun Proteomics Data Sets. *J Proteome Res*. 2009;8(7):3737-3745. doi:10.1021/pr801109k
30. The M, MacCoss MJ, Noble WS, Käll L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J Am Soc Mass Spectrom*. 2016;27(11):1719-1727. doi:10.1007/s13361-016-1460-7
31. Liu H, Sadygov RG, Yates JR. A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Anal Chem*. 2004;76(14):4193-4201. doi:10.1021/ac0498563
32. Berg M. Reproducibility of LC-MS-based protein identification. *J Exp Bot*. 2006;57(7):1509-1514. doi:10.1093/jxb/erj139
33. Kreimer S, Belov ME, Danielson WF, et al. Advanced Precursor Ion Selection Algorithms for Increased Depth of Bottom-Up Proteomic Profiling. *J Proteome Res*.

2016;15(10):3563-3573. doi:10.1021/acs.jproteome.6b00312
34. Hsieh EJ, Bereman MS, Durand S, Valaskovic GA, MacCoss MJ. Effects of column and gradient lengths on peak capacity and peptide identification in nanoflow LC-MS/MS of complex proteomic samples. *J Am Soc Mass Spectrom*. 2013;24(1):148-153. doi:10.1007/s13361-012-0508-6
35. Nägele E, Vollmer M, Hörth P, Vad C. 2D-LC/MS techniques for the identification of proteins in highly complex mixtures. *Expert Rev Proteomics*. 2004;1(1):37-46. doi:10.1586/14789450.1.1.37
36. Vollmer M, Nägele E, Hörth P. Differential proteome analysis: two-dimensional nano-LC/MS of E. coli proteome grown on different carbon sources. *J Biomol Tech*. 2003;14(2):128-135. http://www.ncbi.nlm.nih.gov/pubmed/14676311.
37. Nägele E, Vollmer M, Hörth P. Improved 2D nano-LC/MS for proteomics applications: a comparative analysis using yeast proteome. *J Biomol Tech*. 2004;15(2):134-143. http://www.ncbi.nlm.nih.gov/pubmed/15190086.
38. Zhao H, Creese AJ, Cooper HJ. Online LC-FAIMS-MS/MS for the Analysis of Phosphorylation in Proteins. *Methods Mol Biol*. 2016;1355:241-250. doi:10.1007/978-1-4939-3049-4_16
39. Kim B, Araujo R, Howard M, Magni R, Liotta LA, Luchini A. Affinity enrichment for mass spectrometry: improving the yield of low abundance biomarkers. *Expert Rev Proteomics*. 2018;15(4):353-366. doi:10.1080/14789450.2018.1450631
40. Hodge K, Have S Ten, Hutton L, Lamond AI. Cleaning up the masses: Exclusion lists to reduce contamination with HPLC-MS/MS. *J Proteomics*. 2013;88:92-103. doi:10.1016/j.jprot.2013.02.023
41. Zhang Y, Wen Z, Washburn MP, Florens L. Effect of Dynamic Exclusion Duration on Spectral Count Based Quantitative Proteomics. *Anal Chem*. 2009;81(15):6317-6326. doi:10.1021/ac9004887
42. Holman SW, McLean L, Eyers CE. RePLiCal: A QconCAT Protein for Retention Time Standardization in Proteomics Studies. *J Proteome Res*. 2016;15(3):1090-1102. doi:10.1021/acs.jproteome.5b00988
43. McQueen P, Spicer V, Rydzak T, et al. Information-dependent LC-MS/MS acquisition with exclusion lists potentially generated on-the-fly: Case study using a whole cell digest of Clostridium thermocellum. *Proteomics*. 2012;12(8):1160-1169. doi:10.1002/pmic.201100425
44. Erickson BK, Mintseris J, Schweppe DK, et al. Active Instrument Engagement Combined with a Real-Time Database Search for Improved Performance of Sample Multiplexing Workflows. *J Proteome Res*. 2019;18(3):1299-1306. doi:10.1021/acs.jproteome.8b00899
45. Schweppe DK, Eng JK, Bailey D, et al. Full-featured, real-time database searching platform enables fast and accurate multiplexed quantitative proteomics. *bioRxiv*. January 2019:668533. doi:10.1101/668533
46. Hebert AS, Thöing C, Riley NM, et al. Improved Precursor Characterization for Data-Dependent Mass Spectrometry. *Anal Chem*. 2018;90(3):2333-2340. doi:10.1021/acs.analchem.7b04808
47. Wichmann C, Meier F, Winter SV, Brunner A-D, Cox J, Mann M. MaxQuant.Live enables global targeting of more than 25,000 peptides. *bioRxiv*. January 2018:443838. doi:10.1101/443838
48. Ma B. Novor: real-time peptide de novo sequencing software. *J Am Soc Mass Spectrom*.

2015;26(11):1885-1894. doi:10.1007/s13361-015-1204-0
49. Fan F, Xiong J, Wang G. On Interpretability of Artificial Neural Networks. 2020.
50. Hooker S, Erhan D, Kindermans P-J, Kim B. A Benchmark for Interpretability Methods in Deep Neural Networks. In: Wallach H, Larochelle H, Beygelzimer A, d\textquotesingle Alché-Buc F, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc.; 2019:9737-9748. http://papers.nips.cc/paper/9167-a-benchmark-for-interpretability-methods-in-deep-neural-networks.pdf.
51. Huang L-T, Gromiha MM, Ho S-Y. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics*. 2007;23(10):1292-1293. doi:10.1093/bioinformatics/btm100
52. Valdes G, Luna JM, Eaton E, Simone CB, Ungar LH, Solberg TD. MediBoost: a Patient Stratification Tool for Interpretable Decision Making in the Era of Precision Medicine. *Sci Rep*. 2016;6(1):37854. doi:10.1038/srep37854
53. Bramer M. Avoiding overfitting of decision trees. *Princ data Min*. 2007:119-134.
54. Schaffer C. Overfitting avoidance as bias. *Mach Learn*. 1993;10(2):153-178.
55. Han H, Jiang X. Overcome support vector machine diagnosis overfitting. *Cancer Inform*. 2014;13(Suppl 1):145-158. doi:10.4137/CIN.S13875
56. Yongqiao Wang, Shouyang Wang, Lai KK. A new fuzzy support vector machine to evaluate credit risk. *IEEE Trans Fuzzy Syst*. 2005;13(6):820-831. doi:10.1109/TFUZZ.2005.859320
57. Jovanovic M, Radovanovic S, Vukicevic M, Van Poucke S, Delibasic B. Building interpretable predictive models for pediatric hospital readmission using Tree-Lasso logistic regression. *Artif Intell Med*. 2016;72:12-21. doi:10.1016/j.artmed.2016.07.003
58. Martin-Barragan B, Lillo R, Romo J. Interpretable support vector machines for functional data. *Eur J Oper Res*. 2014;232(1):146-155. doi:10.1016/j.ejor.2012.08.017
59. Tabb DL, McDonald WH, Yates JR. DTASelect and Contrast: Tools for Assembling and Comparing Protein Identifications from Shotgun Proteomics. *J Proteome Res*. 2002;1(1):21-26. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2811961/.
60. Cociorva D, L. Tabb D, Yates JR. Validation of Tandem Mass Spectrometry Database Search Results Using DTASelect. *Curr Protoc Bioinforma*. 2006;16(1). doi:10.1002/0471250953.bi1304s16
61. Käll L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods*. 2007;4(11):923-925. doi:10.1038/nmeth1113
62. Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat Biotechnol*. 2004;22(2):214-219. doi:10.1038/nbt930
63. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*. 2008;24(21):2534-2536. doi:10.1093/bioinformatics/btn323
64. Eng JK, Jahan TA, Hoopmann MR. Comet: An open-source MS/MS sequence database search tool. *Proteomics*. 2013;13(1):22-24. doi:10.1002/pmic.201200439
65. UniProt: a hub for protein information. *Nucleic Acids Res*. 2015;43(D1):D204-D212. doi:10.1093/nar/gku989
66. Pedrioli PGA. Trans-Proteomic Pipeline: A Pipeline for Proteomic Analysis. In: Hubbard SJ, Jones AR, eds. *Proteome Bioinformatics*. Totowa, NJ: Humana Press; 2010:213-238.

doi:10.1007/978-1-60761-444-9_15
67. Lundgren DH, Hwang S-I, Wu L, Han DK. Role of spectral counting in quantitative proteomics. *Expert Rev Proteomics*. 2010;7(1):39-53. doi:10.1586/epr.09.69
68. Mant, C. T.; Burke, T. W. L.; Hodges RS. A rapid, simple approach to predicting the effects of varying run parameters on reversed-phase gradient elution profiles of peptides. *LC GC*. 1994;12(5):396-404.
69. Shibue M, Mant CT, Hodges RS. Effect of anionic ion-pairing reagent hydrophobicity on selectivity of peptide separations by reversed-phase liquid chromatography. *J Chromatogr A*. 2005;1080(1):68-75. doi:10.1016/j.chroma.2005.03.035
70. Guo D, Mant CT, Hodges RS. Effects of ion-pairing reagents on the prediction of peptide retention in reversed-phase high-resolution liquid chromatography. *J Chromatogr A*. 1987;386:205-222. doi:10.1016/S0021-9673(01)94598-4
71. Amodei D, Egertson J, MacLean BX, et al. Improving Precursor Selectivity in Data-Independent Acquisition Using Overlapping Windows. *J Am Soc Mass Spectrom*. 2019;30(4):669-684. doi:10.1007/s13361-018-2122-8
72. Martens L, Hermjakob H, Jones P, et al. PRIDE: The proteomics identifications database. *Proteomics*. 2005;5(13):3537-3545. doi:10.1002/pmic.200401303
73. Kong AT, Leprevost F V, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nat Methods*. 2017;14(5):513-520. doi:10.1038/nmeth.4256
74. Krokhin O V. Sequence-Specific Retention Calculator. Algorithm for Peptide Retention Prediction in Ion-Pair RP-HPLC: Application to 300- and 100-Å Pore Size C18 Sorbents. *Anal Chem*. 2006;78(22):7785-7795. doi:10.1021/ac060777w
75. Spicer V, Yamchuk A, Cortens J, et al. Sequence-Specific Retention Calculator. A Family of Peptide Retention Time Prediction Algorithms in Reversed-Phase HPLC: Applicability to Various Chromatographic Conditions and Columns. *Anal Chem*. 2007;79(22):8762-8768. doi:10.1021/ac071474k
76. Wong JWH, Cagney G. An Overview of Label-Free Quantitation Methods in Proteomics by Mass Spectrometry. In: ; 2010:273-283. doi:10.1007/978-1-60761-444-9_18
77. Jin S, Daly DS, Springer DL, Miller JH. The Effects of Shared Peptides on Protein Quantitation in Label-Free Proteomics by LC/MS/MS. *J Proteome Res*. 2008;7(1):164-169. doi:10.1021/pr0704175

# 7. Contributions of Collaborators

Zhibin Ning generated the mass spectrometry datasets used in the training and testing of the logistic regression classifier as well as those used for the simulated mass spectrometry data acquisition evaluation. Zhibin also helped with writing sections 3.1-3 in the thesis, pertaining to mass spectrometry dataset generation. Nora Wong provided code for a prototype of the software simulating mass spectrometry data acquisition using the described heuristic exclusion algorithm without restricting exclusion to a confined retention time.