Retention Time and Solvent Concentration Prediction for an Automated Peptide Manufacturing Platform

by

Benjamin Russell

B.S. Mechanical Engineering, Northeastern University (2015)

Submitted to the Department of Mechanical Engineering
in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Manufacturing

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
September 2020

Author . . . . . .
Department of Mechanical Engineering
August 14, 2020

Certified by . . . . . .
Brian Anthony
Principal Research Scientist, Mechanical Engineering

Accepted by. . . . . .
Nicolas Hadjiconstantinou
Chairman, Department Committee on Graduate Theses

"Essentially all models are wrong. But some are useful."

George Box

Retention Time and Solvent Concentration Prediction
for an Automated Peptide Manufacturing Platform

by

Benjamin Russell

Submitted to the Department of Mechanical Engineering
on August 14, 2020 in Partial Fulfillment of the
Requirements for the Degree of

Master of Engineering in Manufacturing

Abstract

This thesis investigates statistical and machine learning techniques for the regression-based prediction of peptide retention time and solvent concentration using amino acid physio-chemical properties and historical test data from liquid chromatography and mass spectroscopy (LC-MS) testing. This research was performed alongside the team at Mytide Therapeutics, in Boston, MA between May and August 2020.

Mytide delivers high-purity custom peptides on rapid timelines enabled by their novel robotic manufacturing system. This system automates and connects the disparate processes involved in manufacturing peptides. Through prior work Mytide has built a database of peptide LC-MS testing data. These results are leveraged to make predictions of solvent concentration at the retention time for specific peptides, that is in turn used to generate methods for their purification process on a per peptide basis. These optimized methods replace a general time-intensive solvent gradient. Implementation of these models cut the operating time of their purification process by 53%, while maintaining the required resolution of UV chromatogram data. Implementation of this workflow increases the throughput of their purification machine, while also reducing solvent used by 37%.

Thesis Supervisor: Brian Anthony

Title: Principal Research Scientist, Mechanical Engineering

Acknowledgments

This thesis would not have been possible without support from several individuals.

Thank you,

Professor David Hardt - for your advisement, sage advice, and continued support through the duration of this project.

Mytide Team - Chase, Dale, Kevin, and Liam. For your mentorship and patience. You were all instrumental in acquainting me with the technology and processes that enabled this thesis.

Mytide Team-mates - Abby and Liudi. For your help through the project. It was great working alongside you both.

Jose Pacheco - For your career advice and co-ordination in setting up this experience.

Family – For your continuous love and support.

Alyssa – For listening. You are my everything.

Contents

List of Figures

List of Tables

List of Equations

1 Introduction

This is an industry project thesis, completed alongside the team at Mytide Therapeutics at their headquarters in Boston, MA. The goal of this project is to reduce testing time in their manufacturing process and increase confidence in molecular composition through predictor models. This work lays out the tools and methodologies used for analyze the data, build and train regression models, generate new methods based on the predictions, and how these new methods can improve and monitor peptide-based molecule purification retention time in the manufacturing process. This chapter will provide a background on peptides, the processes and technologies used to manufacture them, a machine learning (ML) primer, and an overview of the manufacturing process specific to Mytide Therapeutics.

1.1 Peptide Introduction

Peptides are naturally occurring polymers composed of a chain of amino acids. In terms of size, peptides lie between small molecules and proteins. Peptides are between 2 and 50 amino acids in length, while proteins begin at a length of 50. Structurally, peptides and proteins are similar as they are held together with amide bonds in between amino acids. Both natural and synthetic peptides can have therapeutic value (the first application of a therapeutic peptide was that of insulin in 1922 [1]), and their use has increased over time. As a result, commercial production of peptides is now a growing segment of the biotechnology industry.

Peptides are signaling molecules that bind to specific receptors in the body and trigger some form of intracellular effect. They fill critical roles in the human body and can act as hormones, neurotransmitters, and anti-infectives. In fact, more than 7000 naturally occurring peptides have been identified [2]. Synthetic peptides represent engineered peptides that are not naturally occurring, but often based on sequences of naturally occurring peptides. The amino acid chain can be outfitted with a variety of synthetic modifications which have led to increasing use of peptides for therapeutic applications.

Peptide characteristics such as their high bioactivity, high specificity, and low toxicity make them valuable for therapeutics. Their advantages and disadvantages are summarized in Table 1, adopted from Raffery [4].

Advantages	Disadvantages
High Potency: Allows for lower dose requirements	**Poor Metabolic Stability:** Repeated dosing often necessary
High Selectivity: Lower Side Effects	**Poor Membrane Permeability:** Typical dose is injected
Broad range of biological targets	**High Production Cost**
Low Toxicity: Low accumulation in tissue over time	Tendency to aggregate, breakdown by hydrolysis and/or oxidation
Discoverable at peptide and/or nucleic acid levels	Low Biostability

Table 1: Therapeutic Peptide Advantages and Disadvantages

Amino acid representation uses a standard alphabet seen in Table 2. There are both 3-letter and 1-letter abbreviations for each amino acid. To denote peptide sequences, a string of the individual amino acids is represented. For example, the peptide Acyl Carrier Protein (ACP), a difficult to synthesize peptide used in evaluating manufacturing protocols, is abbreviated VQAAIDYING.

Name	3-Letter	1-Letter	Name	3-Letter	1-Letter
Alanine	Ala	A	Leucine	Leu	L
Arginine	Arg	R	Lysine	Lys	K
Asparagine	Asn	N	Methionine	Met	M
Aspartic acid	Asp	D	Phenylalanine	Phe	F
Cysteine	Csy	C	Proline	Pro	P
Glutamine	Gln	Q	Serine	Ser	S
Glutamic Acid	Glu	E	Threonine	Thr	T
Glycine	Gly	G	Tryptophan	Trp	W
Histidine	His	H	Tyrosine	Tyr	Y
Isoleucine	Ile	I	Valine	Val	V

Table 2: Amino Acid Abbreviations

1.2 Peptide Market

Advances in computational power, manufacturing process improvement, and measurement specificity in recent decades give greater confidence and more willingness to develop pioneering peptide-based therapeutics. As of March 2017, there have been 68 peptides approved for therapeutic use in the United States, Europe, and/or Japan [3]. Figure 1 shows the makeup of 484 peptides. Of these peptides, 54% were discontinued, 12% were approved by the FDA, while 32% are still in

development. Active peptides are those in research and development ranging from pre-registration with the FDA to Phase III clinical trials, the last step in the drug development timeline.

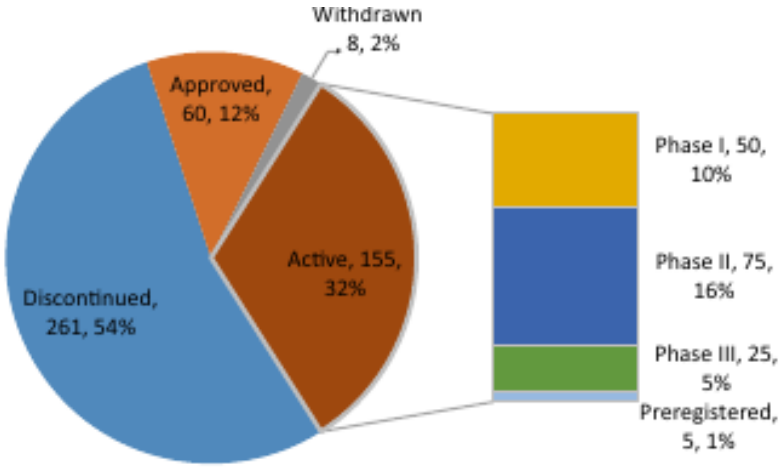


Figure 1: Peptide Development up to 2017. Adapted from [3]

Applications for the approved peptides in therapeutics include treatment for diseases such as prostate and breast cancer, Type 2 diabetes, hypertension, and HIV [4]. Within the overall peptide market, there are several areas garnering significant research. These areas include using peptides as antimicrobials (AMP's) and vaccines. In 2010, PROVENGE, a therapeutic vaccine, became the first widely marketed peptide for cancer therapy in its treatment for prostate cancer [4]. The development of peptides has steadily climbed through the previous decades. In the 1970's, an average of one peptide per year entered clinical trials. Ten and twenty years later that number increased to five and ten peptides per year respectively [5] .

1.3 Mytide Therapeutics

Mytide Therapeutics was founded in June 2018 with a multi-disciplinary core of MIT engineers, scientists, and business professionals. Mytide manufactures custom peptide-based molecules at its headquarters in Boston, MA, using their novel robotic platform, "Rapid Automated Computation, Coupling, Cleavage, and Chromatography Execution", (RAC4E). This platform leverages artificial intelligence and robotics to rapidly produce peptide molecules.

Mytide's platform is developed to serve as both a business-to-customer (B2C) and business-to-business (B2B) service. Their customers currently span academic institutes, pharmaceutical companies, and biotechnology laboratories. The developed technology platform enables differentiation of Mytide

from their competitors in the peptide manufacturing space as they can reliably produce peptides that are difficult to make in a short period of time.

1.4 Peptide Synthesis

Peptide synthesis allows for the creation of synthetic peptides. The discovery of solid-phase peptide synthesis (SPPS) in 1963 by R.B. Merrifield improved peptide synthesis production rate, and is widely viewed as one of the most important methodologies in chemistry and biology [6]. The fundamental premise of SPPS is that amino acids can be assembled into a peptide of any desired sequence one at a time. The first amino acid is built upon an insoluble support, that is then cleaved (removed) with a reagent at the conclusion of the peptide sequence. Using the peptide supporting structure as the base allows for the removal of purification and isolation steps in between each amino acid in the sequence. SPSS remains the accepted method for peptide synthesis to date and improvements have been made in previous decades on this fundamental principle.

At Mytide, Fmoc[1] synthesis is used to ensure proper coupling between amino acids by protecting the reactive areas of the amino acid reagents to prevent incorrect bonding. Fmoc synthesis is the industry standard, and dates to its origin in 1978 [7]. When the peptide is built stepwise, one amino acid at a time, there are a few operations performed before coupling the next amino acid. During synthesis, a base, commonly piperidine, is used to remove (deprotect) these areas so they can bond with the next amino acid in the sequence.

Synthesis of peptides is not perfect, and problems can occur. Common issues include truncation, deletion, duplication, and formation of by-products. Some of these issues can be attributed to incomplete removal of the deprotection group. Example deletion and duplication errors of a sample peptide sequence are represented by Table 3.

Desired Sequence	VQAAIDYING
Aspartic Acid Duplication	VQAAIDDYING
Aspartic Acid Deletion	VQAAIYING

Table 3: Peptide Deletion and Duplication Example

[1] fluorenylmethoxycarbonyl protecting group (Fmoc)

In the desired sequence Aspartic Acid (D) should be the sixth amino acid in the chain. With a duplication this amino acid is repeated, back-to-back. For a deletion that specific amino acid is missing from the sequence entirely.

1.5 LC-MS Overview

Liquid chromatography (LC) and mass spectroscopy (MS) techniques are widely used in the pharmaceuticals industry for identifying specific and sensitive measurements of chemical compounds in a solution. LC is a method used for separating a mixture where components of interest, in this case peptides, are dissolved in a liquid called the mobile phase. This mobile phase is often a mixture of water and a polar solvent such as acetonitrile (ACN). This mobile phase is passed through a chromatography column, packed with a fine powder, known as the stationary phase. Chromatography columns vary in length, diameter, packing material, and packing particle size. All of which are chosen through careful experimentation based on the application. Advances in column manufacturing and packing material, UV detector sensitivity, and flow pump accuracy have all contributed towards making this a technology used for modern analytical drug discovery [8].

The underlying principle of LC is separation of a mixture based on properties of the different components in the overall mixture when exposed to a solvent. The sample, along with this solvent, is pushed through a column with liquid flow generated by pumps that operate at pressures between 400-500 bar. In reversed-phase liquid chromatography (RPLC), which is the method used at Mytide, the percentage of the polar solvent (ACN) is increased throughout the duration of the test according to a gradient profile that is experimentally determined. An example of a gradient is the orange trace in Figure 2. The gradient is a representation of the percentage of ACN throughout the duration of the LC run. In most RPLC applications the solvents that makeup the gradient are water and ACN.

Physio-chemical properties of the components in the solution cause varying flow velocities through the packed column. This occurs due to a difference in polarity between the non-polar stationary phase and polar mobile phase. Particles in the mobile phase that are similar in polarity to the column will be strongly attracted causing a longer retention time. Retention time is the point in time that a compound exits the column. To ensure all compounds are removed from the column, the end of the test includes a step to 95% ACN for five minutes. This step is critical as it prevents compounds from previous runs from interfering with future production.

The percentage of ACN can be visualized by the orange trace in Figure 2. Over the length of the run, the ACN percentage linearly increases from a low to high concentration, this is aptly referred to

as a linear solvent gradient. Another technique in separating solutions is the use of isocratic gradients. These gradients have no linear slope throughout the test, just step changes, however this method is not commonly used for peptides. That said, this thesis only explores linear gradients.

Figure 2 illustrates the components of a LC-MS machine. The solvents are stored in separate containers and are pulled by in-line pumps through an injector into the column. Control of the separate solvent pumps allows the ratio of ACN to water to vary over the length of the test. To begin, the sample solution is loaded into the column with the injector. Next, the solvent pumps push the mobile phase through the injector and into the column at the ratio prescribed by the test method. As the different compounds elute (emerge) from the column they are passed first through a UV detector and then a mass spectrometer. The UV detector, which measures at a wavelength of 214nm and 254nm, common for this application, relates magnitude of the absorbance of light to the concentration of the compound passing through at that time. In essence, it informs the quantity of the product you have made. After the UV detector, these separated compounds move to a mass spectrometer that measures what compound was made, further detailed below. When combined these two technologies are powerful as they can precisely measure what compounds were made and how much of each compound was made. Finally, the aliquot and solution are passed to a waste container.

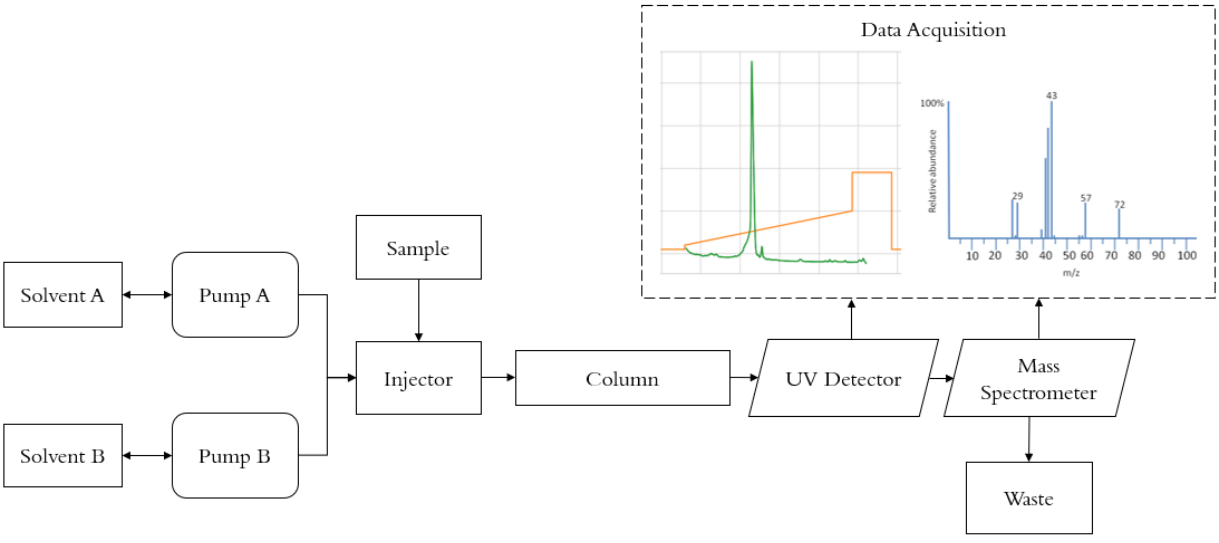


Figure 2: Reversed Phase High Pressure Liquid Chromatography Component Overview

To build a new LC method the gradient needs to be defined. This is done by specifying the ACN % at specific times throughout the test. The gradient is a series of lines that have determined length and slope. Selecting times and ACN %'s for the points that define the gradient requires an understanding of the system and peptide being separated. These points also present a tradeoff between resolution (peak fidelity) and test time.

If a solution contains several compounds that co-elute, meaning they have very similar retention times, a low slope is desired since it provides better separation in time between peaks and allows the compound's UV signature to be captured individually, and not overlap. Conversely, choosing times and solvent concentrations that map to a steep slope means that over that period there will be a high change in the ACN %, causing co-elution, making the chromatogram difficult to interpret.

Directly following LC is MS, a technique that converts molecules into ions by imparting an electrical charge. The mass to charge ratio of this charged ion is measured by a proportional electrical current and plotted in a mass spectrum. The makeup of the molecule can then be identified by correlating identified masses with known masses of elements.

When viewing a mass spectrum there are several key items to consider including the mass to charge ratio, relative abundance, and base peak. Measured on the y-axis, the base peak is the most intense or tallest peak in the spectrum, and indicates the ion with the greatest measured quantity, or abundance. Other identified peaks will have an abundance percentage relative to the base peak. Measured on the x-axis is the mass to charge (m/z) ratio. This is a ratio of the ion's mass to its ionized charge, a result of the MS process. Overall, MS allows for incredibly precise measurements of the molecular weight of individual compounds that can then be compared to the theoretical weights to determine what was produced.

The analytical LC-MS system used at Mytide is from Agilent Technologies and comprises a 1260 Infinity II HPLC system and an InfinityLab MSD (mass-selective detection) module.

Figure 3: LC-MS System. At left, LC module. At right, MSD module. [9]

1.6 Purification Overview

The goal in purification is to obtain a solution that meets the quality and purity requirements for the product. This is accomplished by minimizing impurities from incorrect synthesis of the peptide. Purification runs the crude synthesized peptide through a liquid chromatography step and separates the sample into a collection of vials that are synchronized with the data from the UV chromatogram. Because of the hydrophobic interaction in the chromatography column, components of the sample (including the target peptide) will separate at different retention times. After a set time, which is determined by the flow rate of the process and vial size, the vial that collects the sample will be automatically replaced with a new vial by the machine. Mytide uses a Teledyne AccQPrep HP150, that runs this operation and catalogs the continuous UV chromatogram as well as the vial that is in place at each time step. Inspecting this data allows the chemist to select vials with a high presence of the desired peptide as noted by the UV chromatogram.

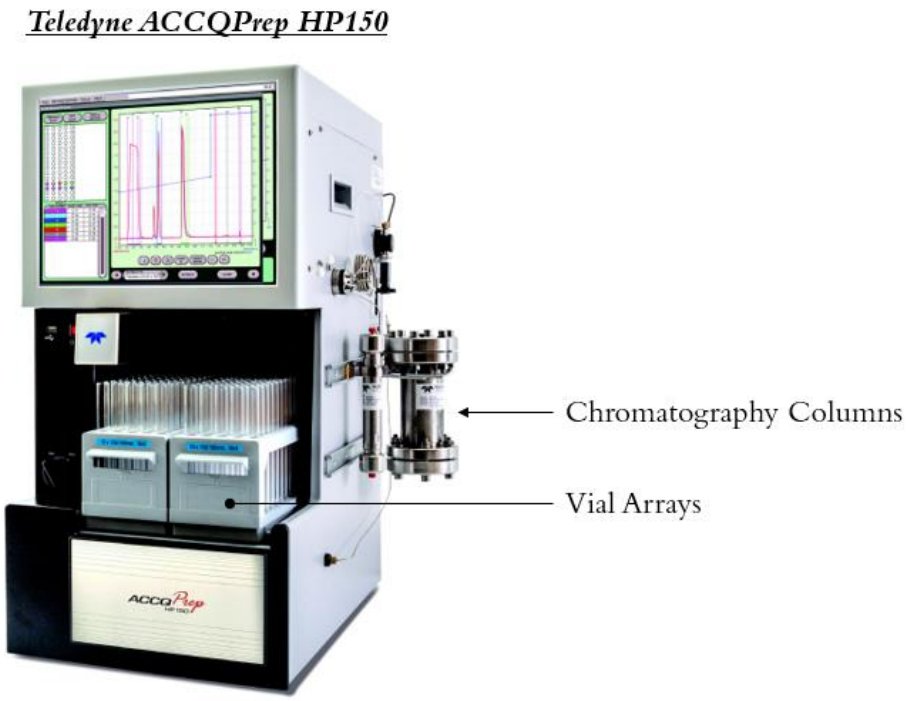


Figure 4: Teledyne AccQPrep HP150 RPLC [10]

There are a few key differences that should be noted between the analytical LC-MS system and the preparative (purification) LC system. First, the analytical system contains a mass spectrometer, while the preparative system does not. The purpose of the analytical system is to measure what was made and the purity - which requires the spectrometer. The preparative system uses the same chromatography technique but does not have a spectrometer. It runs at higher flow rates and is used for separation of the peptide into discrete vials, and not for analysis.

1.7 Mytide Process Overview

Peptide manufacturing consists of distinct processes, each requiring custom built equipment, laboratory instrumentation, and operating parameters. Mytide's overall process is highlighted in Figure 5. This thesis focuses on the LCMS verification and purification steps in their process, shown with yellow parallelograms.

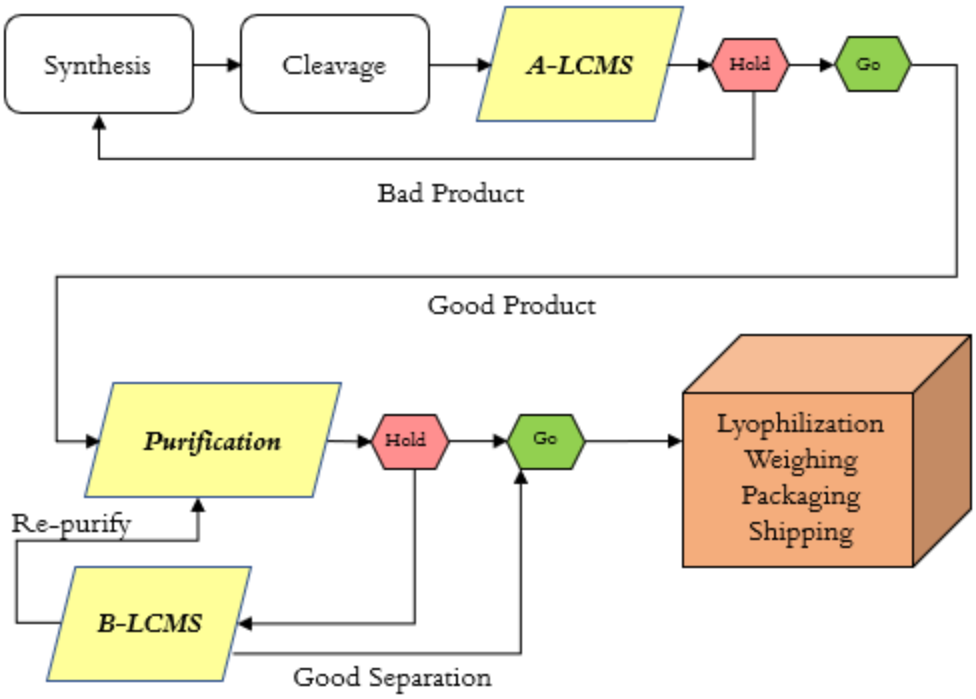


Figure 5: Mytide Manufacturing System Diagram

The Mytide process can be defined with the following information.

Synthesis: In this process the peptide is built on a resin base using continuous flow reactors, one amino acid at a time with successive fmoc protection and deprotection steps.

Cleavage: In this process the resin from the synthesized product is removed so the solution can be prepped for testing and purification.

A-LCMS: Analytical testing for verification of proper synthesis. This is performed on a sample of synthesized product, known as an aliquot.

Purification: In this process the solution is separated into individual vials via LC to remove impurities.

B-LCMS: In this process selected vials from the purification process are analytically verified for correct product.

Lyophilization, weighing, packaging, and shipping are all required processes that transform the liquid solution into solid form for shipping to the customer.

Shown in Figure 5 are two LCMS verification points. The A-LCMS test provides a look at the output of the peptide synthesis, after analytical cleavage and is measured on the analytical LCMS system. This

test provides determination that the subsequent steps in the process are worth pursuing based on the detected compounds and mass abundance. The A-LCMS test is performed on a small sample of the overall solution, known as an aliquot, which measures 5µL. The B-LCMS test is a further verification and ensures that the peptide meets purity and volume requirements. B-LCMS tests are performed after the purification process on each of the selected vials.

Figure 6 shows the sample output of an A-LCMS test conducted on a peptide. The orange line with increasing slope indicates the gradient and the blue line is the UV chromatogram at 210nm. The percentage of ACN increases linearly with time from 2 to 12 minutes with a starting percentage of 5% to a maximum of 65%. The solvent wash can also be seen in the step increase between 13 and 15 minutes, as the percentage of ACN is 95%.

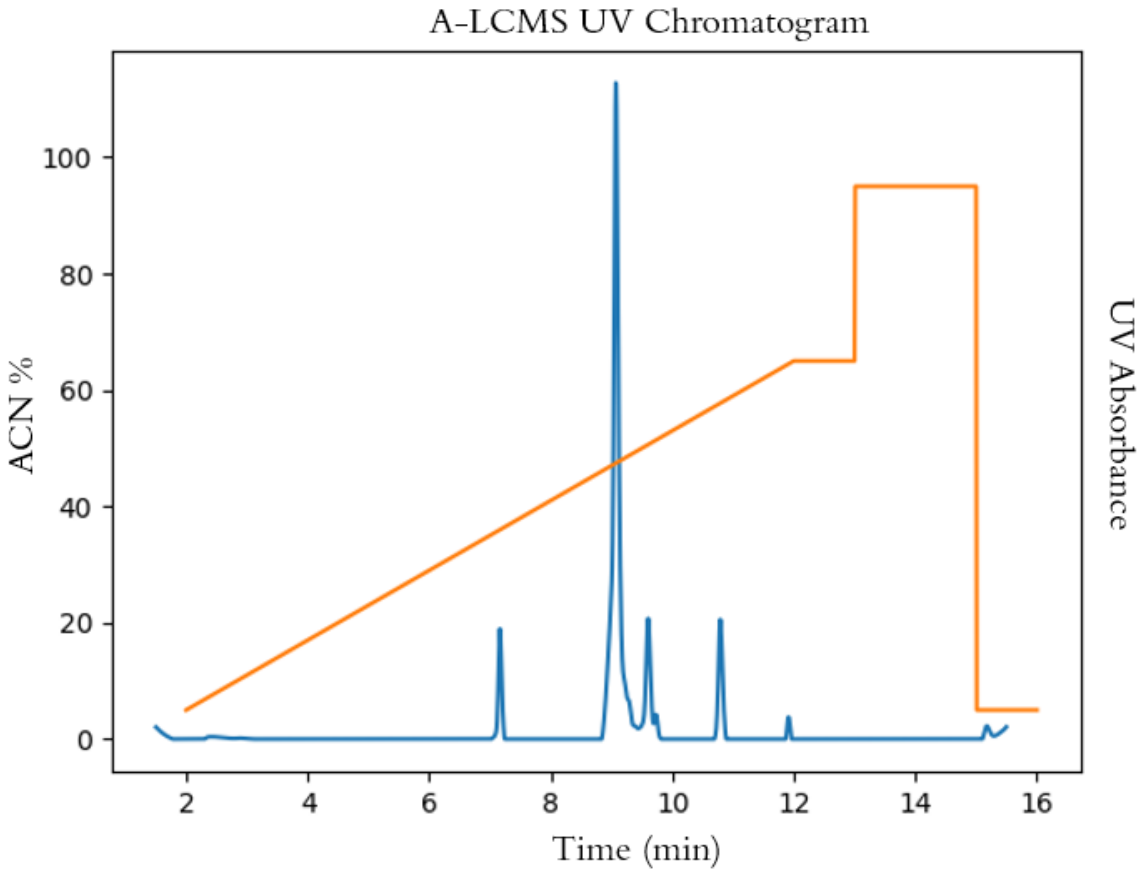


Figure 6: Liquid Chromatography Retention Time Plot with Gradient (orange) and UV absorbance (blue)

The MS test for this sample yields the output shown in Figure 7. The x-axis is the mass per charge (m/z), while the y-axis indicates the abundance. The color-coding present on the marks is from an internal tool developed at Mytide that compares the measured mass from the theoretical mass of the

peptide. If they are similar a green mark is produced. Red and yellow marks indicate matches of a different sequence or unidentified compounds.

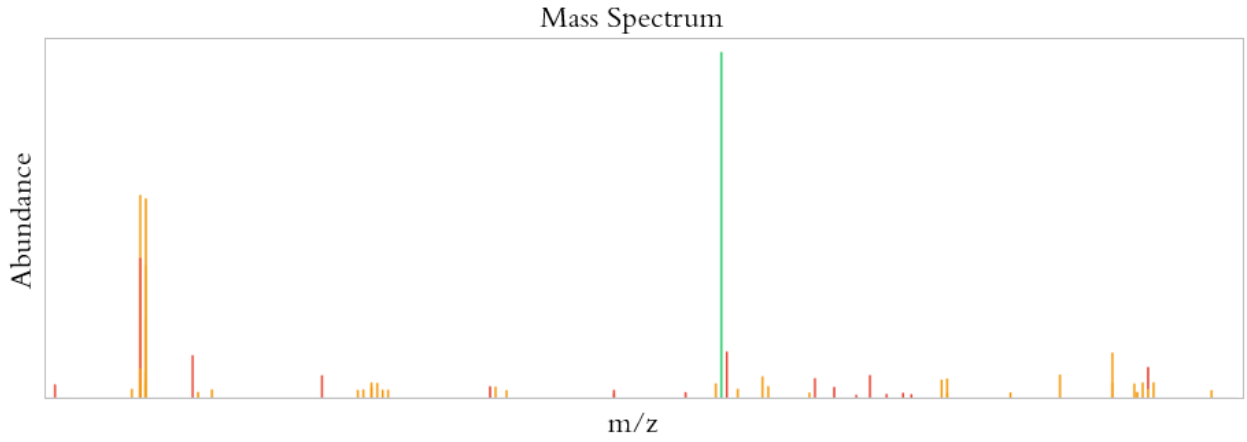


Figure 7: Mass Spectroscopy Spectrum

There is a time delay between the LC and MS measurements as the solution needs to flow through a length of tubing to reach the MS machine for the second measurement. Internal calibration testing at Mytide was performed to find the offset value. Using this offset the LC and MS data points can be overlaid for a complete picture of the peptide testing.

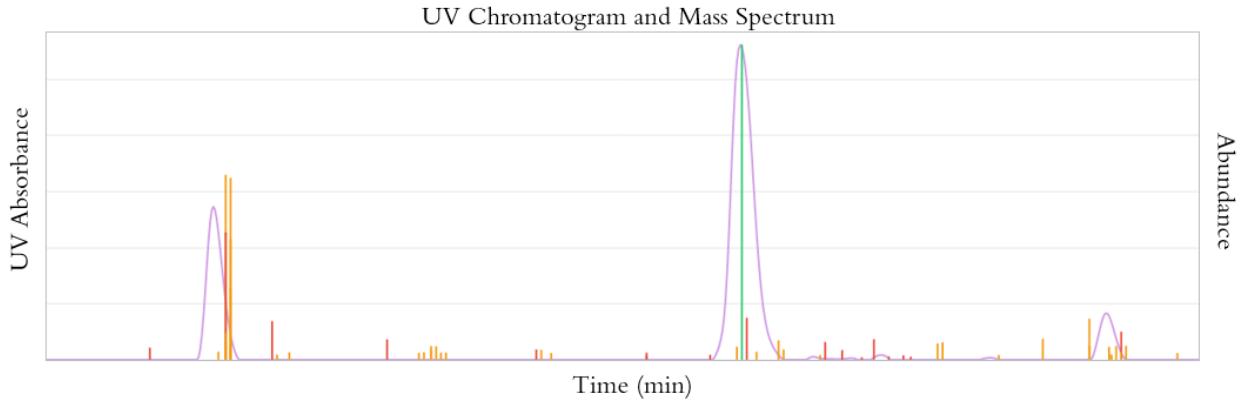


Figure 8: Overlaid LC-MS Results - Mytide Platform

As previously discussed in the purification overview, the purpose is to isolate the volume of material that contains the peptide of interest and exclude byproducts. Figure 9 illustrates how the process changes the collection vial over the duration of the test. Each vertical line in the center of the plot indicates the start of a new vial. The machine that runs purification will correlate the vial number, retention time, and UV chromatogram data.

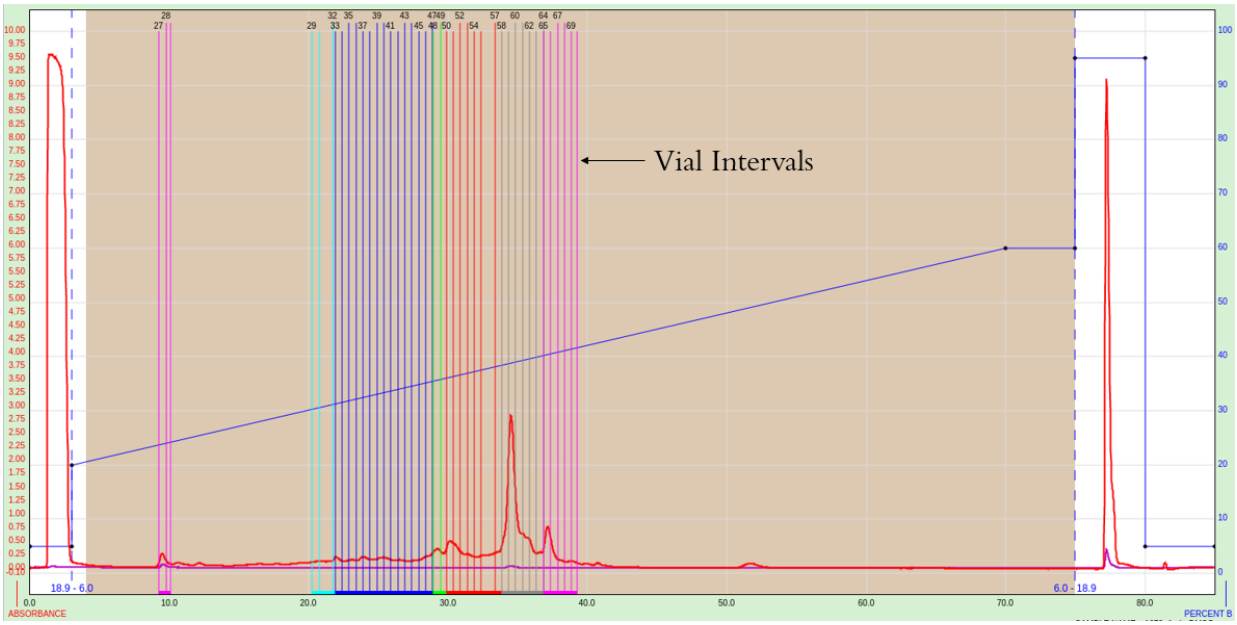


Figure 9: Purification Plot with Vial Markings

Figure 10 zooms in on the peak of interest from the test in Figure 9. The solution captured in vials 58 to 60 (numbering at the top of the plot) correspond with the highest peak, and therefore likely the target peptide. In this case, an aliquot from each of the vials (58 to 60 - 3 total vials) would then be re-tested on the analytical LCMS machine with a B-LCMS test to further verify the solution. Pending positive results for all vials, they are combined and moved forward in the process.

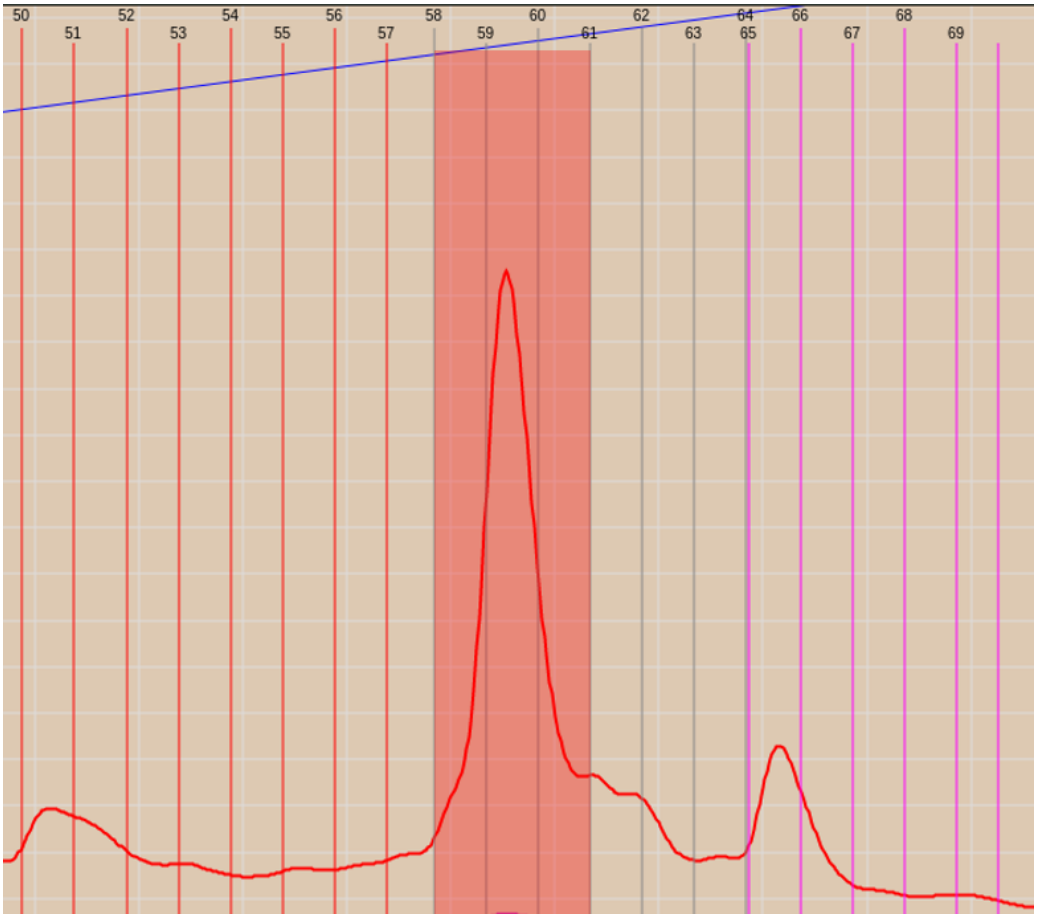


Figure 10: Purification Plot Magnified on Peak showing Vial Division. Vials corresponding to the target peptide are colored with a translucent red fill.

Figure 11 shows the LC and MS data for a purified peptide. Comparing this to Figure 8 shows the effect of the purification process. There are no peaks outside of the target peptide, shown by the peak at approximately 9.3 minutes. This is a result of the purification process and removing the material that does not correspond to the target compound.

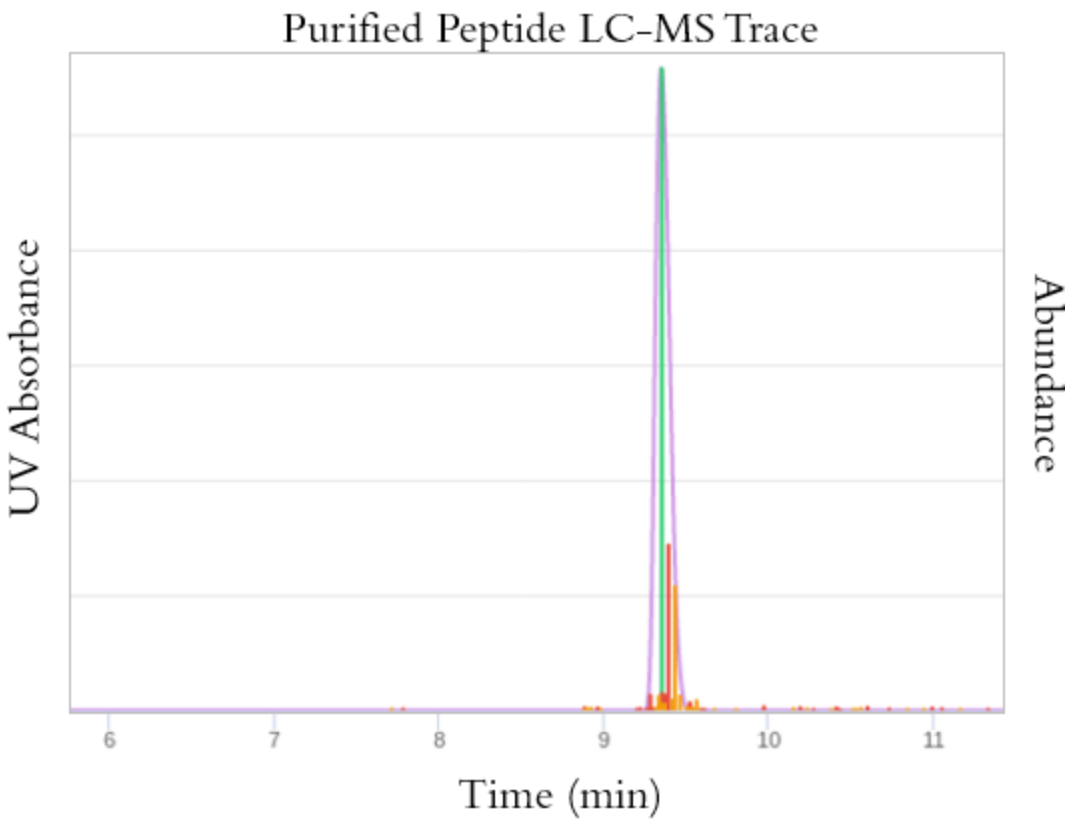


Figure 11: Purified Peptide LC-MS Data - Mytide Platform. The target peptide is identified by a large green bar and high UV Absorbance. There are no other peaks indicating the target peptide was isolated.

1.8 Machine Learning Primer

The following section will introduce machine learning concepts that are used for the models developed in this research.

In predictive modeling datasets are broken into inputs (features) and outputs (targets). Algorithm performance is characterized in the ability to map between the two. ML algorithms can be classified into supervised, unsupervised, and reinforcement learning methods. In supervised models there is access to both inputs and labelled outputs. In contrast, unsupervised models do not have labelled outputs and need to infer the output, while reinforcement learning models use reward functions to penalize bad "actions" and reward "good actions".

In our case, the experimental data records actual retention time and input parameters, therefore the type of method is supervised learning. Algorithms can further be divided into classification and regression types. In this case, retention time is a continuous variable, therefore literature on this topic

use regression-based algorithms. Classification models on the other hand predict discrete output classes based on the model inputs.

Supervised machine learning models are constructed by training the algorithm on a set of labelled data, appropriately titled the training set. Labelled data implies that there is a known output for each sample. During training, the algorithm will attempt to estimate the true test value by minimizing the training loss function (TLF). In regression, common error loss functions include mean square error (L2 loss) and mean absolute error (L1 loss). In these equations y_i is the true target value and $\tilde{y}_i$ is the predicted target value for each individual sample in the test set, composed of a total of n datapoints.

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n}$$

Equation 1: Mean Square Error

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \tilde{y}_i|}{n}$$

Equation 2: Mean Absolute Error

The model is evaluated on a test set, where the target is not known to the algorithm. These two datasets are split from the overall dataset, typically in a ratio of 70-80% training data and 20-30% testing data. Performance of the model can be measured in a variety of ways; however, it is always a measure of the difference between true value and predicted value of the testing data.

Bias and variance are essential to understanding how a model will generalize to new data. Bias represents how far the average prediction of the model is from the true value. When the average error is large the model contains a high bias. Models with high bias tend to under-fit, meaning the model ignores input features and oversimplifies. On the other hand, variance is a measure of the uncertainty, or spread, of the estimates. Models with high variance typically have good performance on the training data, but poor performance on the test set. In this case, too much weight is placed on the features and the model is overfit.

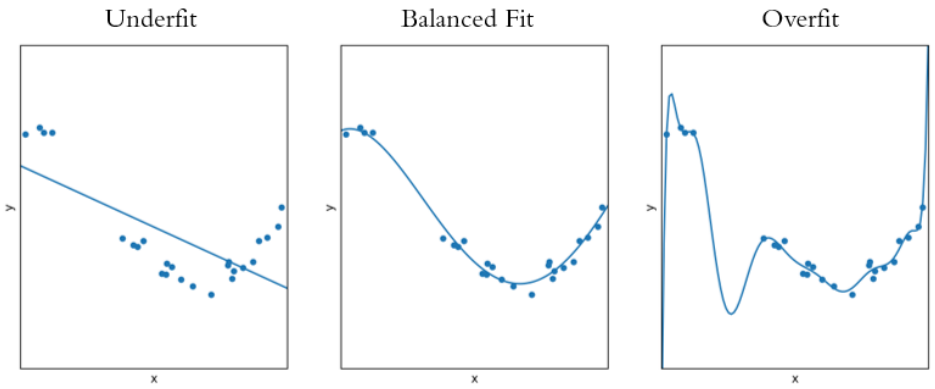


Figure 12: Model Fitting Balance

A few more final key principles to cover include hyperparameters and cross-validation. Dr. Raschka defines hyper-parameters and why they are important. "Almost every machine learning algorithm comes with a large number of settings that we, the machine learning researchers and practitioners, need to specify. These tuning knobs, the so-called hyperparameters, help us control the behavior of machine learning algorithms when optimizing for performance, finding the right balance between bias and variance. Hyperparameter tuning for performance optimization is an art in itself, and there are no hard-and-fast rules that guarantee best performance on a given dataset." [11]

As stated, each algorithm has parameters that can be tuned to optimize the model for a given application. Tuning hyperparameters and comparing the model results can be summarized as a process known as model selection.

Cross validation is a technique often used that helps prevent overfitting of models, particularly in problems with limited data. Fundamentally, cross validation is a resampling procedure that allows for all data samples to be tested. A common technique is k-fold cross validation. In this method the training data is divided into a specified number of folds. A fold refers to a portion of the training set that includes the inputs and outputs. For example, if the selected cross validation study uses five folds (k=5), each fold would contain 20% of the training data. To cross validate, the model is trained on k-1 of these folds, with the last fold used as a validation set, used for testing the trained algorithm and evaluating the error of that fold. This process is repeated for all folds which allows each datapoint to be tested on. Dr. Raschka has made a figure that visualizes this process [11].

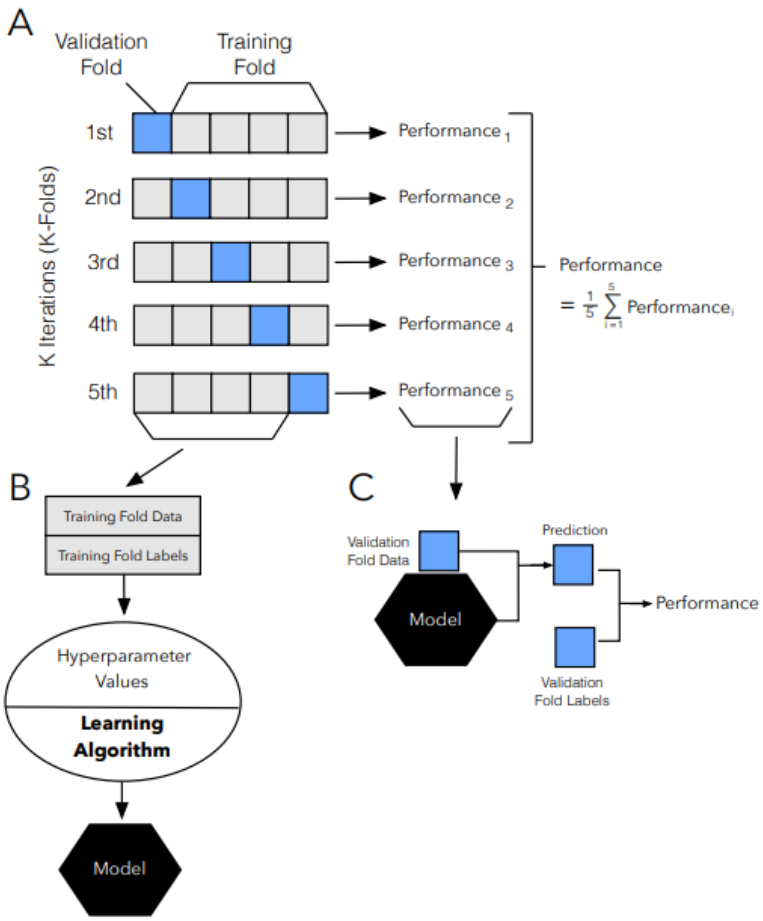


Figure 13: K-fold Cross Validation Process. Adopted from [11]

Figure 13 shows a five-fold cross validation model. After the training data is split into five folds, it is trained and validated on each fold separately. The indexing blue box shows how the validation set is composed of different samples at each fold. Ultimately, the overall performance of the model is an average of the performance of the individual folds.

1.9 Related Work

There have been a host of models and machine learning (ML) architectures developed on liquid chromatography retention time prediction. Recent models, such as DeepLC [12], DeepRTplus [13], LsRP [14], and ELUDE [15] use publicly available databases composed of thousands of samples to train and measure performance of the designed predictor. While newer models exploit the latest advances in neural networks and computer hardware for massive data processing, chromatography prediction actually dates back to the early 1950's when Knight researched simple peptides using paper chromatography [16].

Each approach incorporates different features of their choosing. One work uses an artificial neural network (ANN) for prediction using peptide descriptors such as the length, sequence, hydrophobicity, and nearest-neighbor amino acid [17], while others use only the amino acid sequence information, coded with SMILES strings, a notation used for describing the structure of chemicals, and run that through neural networks [12]. The following table summarizes their respective features and model types for retention time prediction.

Model	Input Features	Model Type	Training [Testing] Peptide Count
DeepLC	Amino Acid Atomic Composition	Convolutional Deep Learning	Multiple datasets totaling: 361,892
ELUDE	Physical and Chemical Descriptors	SVR, radial kernel	1674 [1683]
DeepRTplus	Sequence	Convolutional Deep Learning	Eight datasets totaling: 312,840 [34,765]
METLIN [18]	Molecular Fingerprint and Descriptors	Deep Neural Network	60,029 [20,009]
LsRP	Sequence and amino acid composition	SVR, radial kernel	5619 [5618]

Table 4: Related Work Comparison

These works provide justification for the features used in the predictor models generated in this thesis. The deep learning models presented have complex architectures that can glean out differences in retention time from little input information. Conversely, the models generated in this thesis use additional input features for each sample as the dataset is currently of smaller size, more similar in design to ELUDE.

2 Team and Problem Statement

2.1 Thesis Project Team and Contributions

This document serves as partial fulfillment of the requirements set forth by the Masters of Engineering in Advanced Manufacturing & Design. The period between May and August 2020 was spent working with Mytide Therapeutics on several phases of their production processes. Working as a team of three, individual projects were carried forth under advisement of Professor David Hardt.

The author, worked on predictive modelling of LC-MS testing based on peptide descriptors and purification process time reduction. Liudi Yang focused on synthesis anomaly detection and peptide purity prediction [19], while Abigail Campbell implemented a machine vision system for the robotic platform for in-process inspection [20].

2.2 Problem Statement

This body of work can be separated into two phases:

1. A-LCMS retention time prediction from peptide properties
2. Purification process time reduction through solvent concentration prediction.

Phase 1:

The prediction of A-LCMS retention time allows for generation of new testing methods that can be used to improve efficiency of LCMS testing. Predicted A-LCMS retention time is also useful for Mytide as it gives increased confidence to new results based on historical data. Furthermore, these predictions can be used as inputs into the purification solvent concentration predictor in lieu of actual test results, and to design more efficient analytical test methods.

Phase 2:

Implementation of models that predict the concentration of solvent at which a peptide elutes from a chromatography column during liquid chromatography creates opportunities to significantly shorten testing time in purification. Typically, a conservative gradient is used in generalized peptide

manufacturing to ensure the peptide solution of interest will be measured with good resolution and without peak smearing, albeit at the expense of testing time.

The motivation for this phase can be best visualized in Figure 14, which shows a typical peptide purification run. The orange shaded areas can be considered wasted testing time, as the peptide of interest is not eluting within these regions. By knowing the ACN % at which the peptide will elute from the column, and therefore corresponding retention time, the shaded areas can be eliminated from the run. This will reduce the overall test length and wasted solvents. The width of the white section where the peptide of interest elutes can be adjusted based on the confidence of the predicted value.

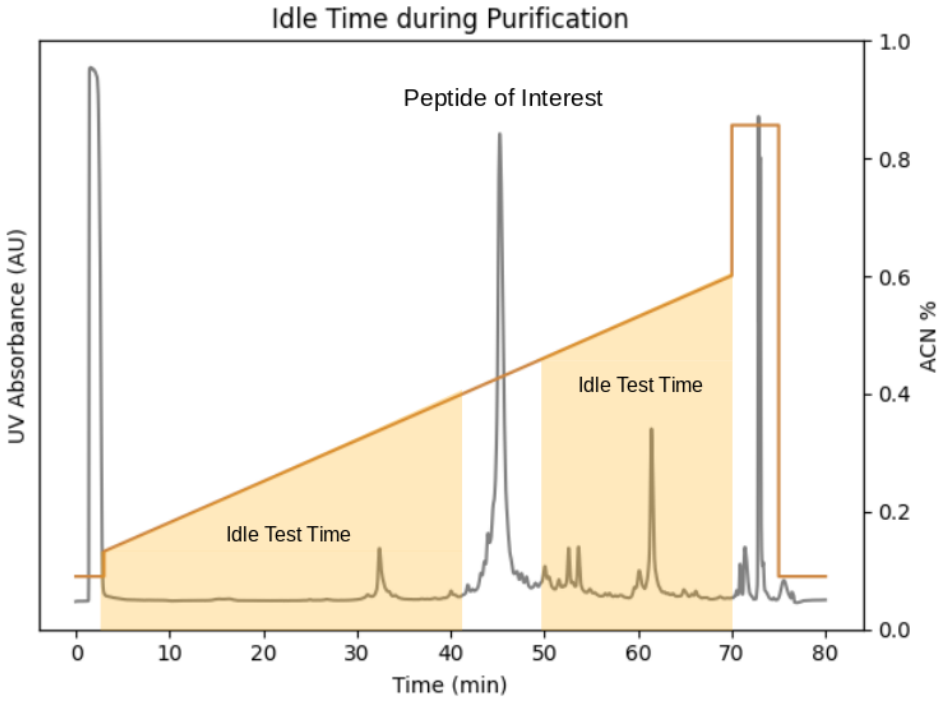


Figure 14: Typical Purification Run with Idle Regions

3 Methodology

The purpose of this chapter is to introduce the methods used to gather and analyze various data, describe the parameters for purification and analytical chromatography, and explain chosen machine learning models as well as methods for optimization and evaluation.

3.1 Data Collection

Mytide stores their manufacturing data in a cloud database that can be accessed via graphical user interface (GUI), or through an application programming interface (API). The database contains all data output from peptide manufacturing runs including LC-MS data, synthesis data, purification data, and peptide metadata. Metadata includes information on modifications and the sequence.

To view manufacturing data of a single peptide it is straightforward to enter the specific peptide ID, navigate to the manufacturing data website and view the information. However, for large-scale data processing and parsing of peptide information, accessing the data through a HTTPS[2] REST[3] API[4] is a must. REST API's organize HTTP calls by calling predefined functions on data that is located at specific URL addresses. There are three main types of REST API requests: GET, POST, and DELETE. GET requests pull information from a site, while POST requests send information to a site.

Python scripts were developed to interface with the manufacturing data on the cloud servers. For a variety of reasons such as poor synthesis or experimental runs, not all peptides have purification and LCMS data. For that reason, a script is needed to analyze all peptides and filter to only the specific ones that contain A-LCMS and B-LCMS data. Using this down-selected bank of peptide ID's, peptide properties and manufacturing data can then be pulled own and used for analysis. A view of the workflow in data collection can be seen in Figure 15.

[2] Hypertext Transfer Protocol Secure (HTTPS) is an encrypted communications protocol commonly for accessing information on the Internet

[3] Representational State Transfer (REST) is a software architecture style used for creating and organizing Internet resources for access by computer systems.

[4] An Applications Programming Interfaces (API) is a set of programming code used as an intermediary and allows software to communicate with other software.

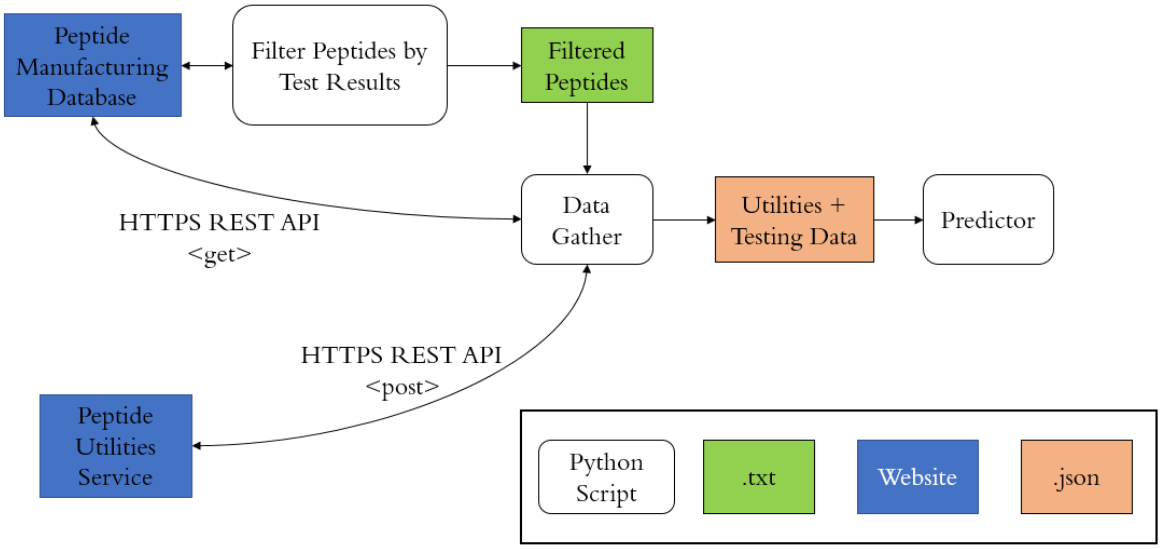


Figure 15: Data Collection Flow Chart

This figure illustrates how Python scripts were used to initially filter all experiments down to a set of filtered peptides which had the testing criteria that was required. Using this text file as an input, the data gathering script was used to access both the manufacturing database for testing results, and then access the peptide utilities service, further explained below, to gain a complete dataset of the peptide properties. This dataset was then used for building and training the predictor models.

3.2 Peptide Utilities Service

The peptide utilities service is a tool developed at Mytide that calculates peptide-specific properties when fed a properly formatted input. The input needs to contain the peptide sequence in either 3-letter or 1-letter abbreviation and the n-term and c-term modifications. Using this information, a variety of physio-chemical properties are calculated. A full list of the properties can be seen in Table 5.

Property Name	Description
Gravy	Measure of the peptide's hydrophobicity based on an amino acid hydropathy scale [21]
Instability	Measure of instability of the sequence [22]
Bulkiness	Ratio of sequence side chain volume to length. Gives average cross section of the peptide sequence [23]
Polarity	Electric force due to the side chain acting on its immediate surroundings [23]
Secondary Structures	Returns a list of the fraction of amino acids which tend to be in Helix, Turn, or Sheet structures. [24]
Sequence Length	Number of amino acids in peptide sequence
Isoelectric Point	pH value where the peptide has no electrical charge
Amino Acid Composition	Total count of each individual amino acid in the sequence. Amino acid percentages are derived from this property
Monoisotopic Mass	Sum of the mass for the primary isotope of each atom in the peptide (measured in daltons)
Synthesis Difficulty	Coefficient representing difficulty of synthesis [25]
N-term[5]	Modifications to the N-terminus are one-hot encoded
C-term[6]	Modifications to the C-terminus are one-hot encoded

Table 5: Physio-Chemical Peptide Properties and Definitions:

Further properties can be extracted from this base list including the relative percentage of individual amino acid types in the peptide. This is found by taking the count for each individual amino acid divided by the peptide length. This adds twenty additional features to the predictor.

3.3 A-LCMS Experimental Parameters and Equipment

Mytide has two unique chromatography columns and several different LC protocols that are used depending on the peptide. Approximately 12% (90 total) of all peptides are run on a C3 column, and the remainder (661 total) on the C18 column. The LC-MS C18 column is a narrow bore column

[5] The N-term (terminus) is the start of a polypeptide and refers to the end residue of the peptide often containing an amine group.
[6] The C-term (terminus) is the end of a polypeptide, typically terminated by a free carboxyl group (-COOH). These endings can be modified to alter structure, properties, and function of the peptide.

with a 2.1mm inner diameter and 50mm length. The particle type is fully porous, with a size of 1.8 μm. The C3 column is a wide bore column with a 2.1mm inner diameter and 150mm length.

Other parameters of concern are the solvents, flow rate, temperature, and gradient profile. All A-LCMS experiments keep these parameters alike. The exact conditions are:

1. Flow rate: 0.5 mL/min
2. Solvent A: H_2O / 0.1% FA[7]
3. Solvent B: ACN / 0.1% FA
4. Column Temperature: 25C

The analytical test's standard solvent gradient is defined by eight time points through the duration of the test with a ratio of solvent A and solvent B at each point. The solvent concentration at any two points can be found via linear interpolation.

Time	Solvent A	Solvent B
0 min	95%	5%
2 min	95%	5%
12 min	35%	65%
13 min	35%	65%
13.01 min	5%	95%
15 min	5%	95%
15.01 min	95%	5%
16 min	95%	5%

Table 6: Analytical LC Gradient Profile

3.4 Purification Experimental Parameters

Purification parameters are currently adjusted by the chemistry team based on intuition and review of A-LCMS testing. However, these parameters are all captured and available for retrieval given the peptide ID. Table 7 is a typical purification gradient. The most common gradient profile total length is 85 minutes, with the main increase in ACN % occurring between 3 minutes and 70 minutes.

[7] Formic Acid

Time	Solvent A	Solvent B
0 min	95%	5%
5 min	95%	5%
5 min	90%	10%
70 min	35%	65%
70 min	95%	5%
75 min	95%	5%
80 min	5%	95%
85 min	5%	95%

Table 7: Purification LC Gradient Profile

The flow rate for purification ranges from 6 to 18.9 mL/min. Typically, the first five minutes (loading phase) and last 10 minutes (high solvent wash phase) are run at 18.9 mL/min, while the linear gradient is run at 6 mL/min. Column temperature and solvents remain the same as A-LCMS testing. C4 (like the C3 column) and C18 columns are used during purification, and it also carries over that the C18 column is heavily favored for most peptides.

The purification C18 column is 50mm in length, with an inner diameter of 30mm. The particle substrate is composed of 5μm spherical silica beads, and the pore size is 100 Å. The C4 column has an inner diameter of 19mm and length of 150mm. The particle substrate is 10μm spherical beads, and the pore size is 300 Å.

Another variable that influences the purification results is the solvent used for reconstituting (recon) the peptide. This solvent is added to the peptide to bring it into solution before it pumped into the chromatography column. The most used recon solvent is solvent 1 (87% usage).

Figure 16 illustrates the importance of including the column and purification recon solvent in the predictor. Column type, indicated by marker size, shows the C4 column has a higher ACN% at retention time for the purification process compared to the LC-MS process. The addition of solvent 3 causes high purification ACN% at retention time, while solvent 2 causes a lower ACN% at purification compared to the LCMS test.

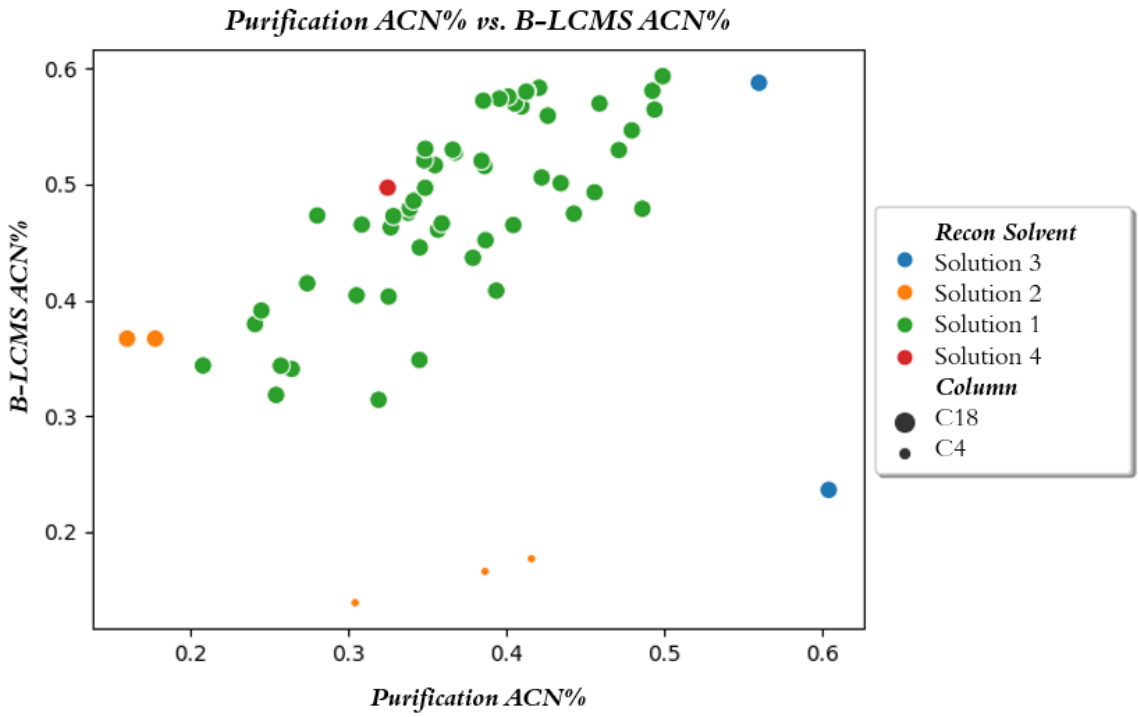


Figure 16: Scatterplot between Purification ACN% and B-LCMS ACN%

3.5 By-Product Data Generation (A-LCMS Predictor)

From the beginning of this project, it was foreseen that there would be complications in effectively training a machine learning model with the limited number of peptides. To increase the dataset size for the LC-MS retention time prediction, processing and filtering was performed on the by-products generated during synthesis and measured during LC-MS testing. This allows an increase in the datapoints as modifications like deletions and duplications can also be captured and analyzed.

For the final product Mytide only delivers the correct peptide based on LC-MS results. However, the testing process gathers and stores data on by-products as well, including retention time and the modified sequence. The MS test outputs what are called "hits". When a hit is correct, as visualized in Figure 7, the theoretical mass of the target peptide matches the measured mass, and is noted in green. The red marks indicate peptides that are identified, but do not match the target peptide.

These red marks can be used for additional data points. Based on the measured mass, the test will output what modification occurred to the target peptide as well as what the modified peptide sequence is.

For example, the top hits in Figure 17 show the correct peptide as rank 1 as it was the most abundant. The second hit is a modification, where the fourth amino acid Lysine was deleted. The corresponding MS chart visualizes this hit and the Lysine deletion.

Top Hits

Detected targets and other most intense masses, sorted by descending intensity. 2 rows per hit.

Rank	Name	Sequence	Mass	Abn	Rel Abn (%)	RT (min)
	M/Z's (real m/z, target m/z, charge state, isotope)					
1	**Correct**	Gly-Ile-Gly-Lys-Phe-Leu-Lys-Lys-Ala-Lys-Lys-Phe-Gly-Lys-Ala-Phe-Val-Lys-Ile-Leu-Lys-Lys	2477.1898	5235704	**100**	**9.37**
	(496.4, 496.1338, 5H+, M); (496.4, 496.3344, 5H+, 13C1); (496.4, 496.5351, 5H+, 13C2); (620.2, 619.9153, 4H+, M); (620.2, 620.1661, 4H+, 13C1); (620.2, 620.4169, 4H+, 13C2); (826 826.2177, 3H+, M); (826.5, 826.5522, 3H+, 13C1); (826.6, 826.8866, 3H+, 13C2); (1238.9, 1238.8227, 2H+, M); (1239.3, 1239.3244, 2H+, 13C1)					
2	**del:Lys4**	Gly-Ile-Gly-Phe-Leu-Lys-Lys-Ala-Lys-Lys-Phe-Gly-Lys-Ala-Phe-Val-Lys-Ile-Leu-Lys-Lys	2349.0168	1113504	**21.27**	**8.75**
	(588.1, 587.8915, 4H+, M); (588.1, 588.1424, 4H+, 13C1); (588.1, 588.3932, 4H+, 13C2); (783.8, 783.5194, 3H+, M); (783.8, 783.8539, 3H+, 13C1); (1174.9, 1174.7752, 2H+, M); (1175.3, 1175.2769, 2H+, 13C1); (1175.7, 1175.7786, 2H+, 13C2)					
3	**del:Gly1**	Ile-Gly-Lys-Phe-Leu-Lys-Lys-Ala-Lys-Lys-Phe-Gly-Lys-Ala-Phe-Val-Lys-Ile-Leu-Lys-Lys	2420.1382	294719	**5.63**	**8.85**
	(605.8, 605.6599, 4H+, M); (605.8, 605.9107, 4H+, 13C1); (606, 606.1616, 4H+, 13C2); (807.4, 807.2106, 3H+, M); (807.4, 807.545, 3H+, 13C1); (1210.4, 1210.312, 2H+, M); (1210.8, 1210.8136, 2H+, 13C1); (1211.3, 1211.3153, 2H+, 13C2); (1221.3, 1221.3029, HNa+, M)					

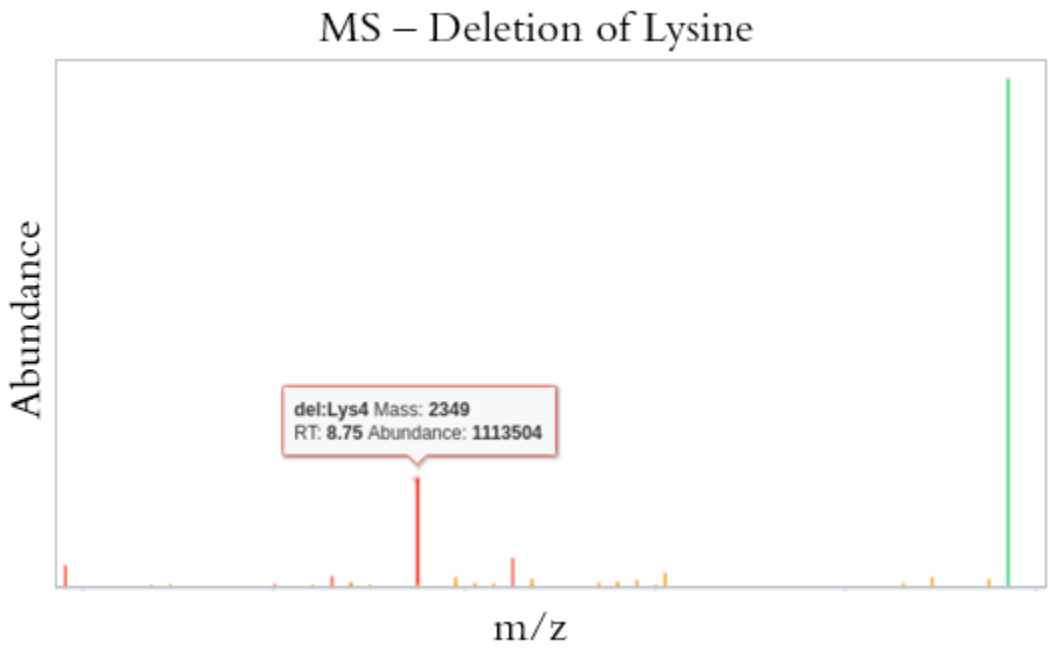


Figure 17: Top Hits of an LC-MS run and corresponding MS chart (Pexiganan). The deletion of Lysine in the lower plot can also be seen in the Top Hits write-up (Rank 2).

For each LC-MS run there can be potentially dozens of by-product hits, all with varying quantities and compositions. Filtering of these by-product hits is needed to produce clean, reliable data, which can be fed into the prediction model.

This filtering eliminates hits with the following criteria:

1. Relative abundance < 10%
2. Duplicate sequences for each peptide ID
3. Less than 3 identified charge states
4. Formulations of:
    a. Glutarimide[8] (glu)
    b. Aspartimide[9] (asp)
5. Fmoc modifications: fmoc not cleaved from N-term
6. Truncations

Sequences with formulations of glutarimide, aspartimide, truncations, and fmoc modifications were removed as they exhibit retention times that are not indicative of standard peptide sequences without these alterations. The identified charge state filter is used as an added mechanism to improve confidence in the sequence identified by the mass spectrometer. It is common for properly identified sequences to have several unique charge states. The threshold was set at 3 after reviewing data for high probability hits.

For the test shown in Figure 17 the top two hits would be used in the predictor, while the third hit would not, as the third hit contains a maximum abundance of less than 10%. A threshold of 10% was chosen as the hits under this amount are variable and unreliable. Including by-products in the predictor adds 277 additional unique peptides from an original database of 384 unique peptide runs. The final count used in the predictor is made up of 661 unique peptide sequences.

3.6 Input Feature Preprocessing

For both the A-LCMS predictor and purification ACN% predictor, a log transformation was used on the monoisotopic mass. This was done for interpretability as the monoisotopic mass averages several thousand while all other variables were an order of magnitude less.

---

[2] Glutaramide and [9] Aspartimide formations are highly sequence dependent and occur during fmoc removal or peptide coupling.

The input features for the purification ACN% predictor were scaled and normalized with the SciKit-Learn's[10] function Standard Scaler. This function standardizes features by removing the mean and scaling to the individual feature's variance. This process was performed on the training data, and then applied to the testing data. The test data was transformed using the standard deviation and mean from the training data. This is critical so that the features from the test set do not bias the training data. For example, if the test set were combined with the training set and the data was normalized and scaled together, information from the test set would leak into pre-processing of the training data, which is done before building a model. This is not a realistic scenario, and not allowed when you are deploying models around new data.

$$x' = \frac{x - \mu}{\sigma}$$

Equation 3: Data Normalization via Standard Scaler. Here μ represents the mean and σ represents the standard deviation for all samples of that feature.

One hot encoding was performed on all categorical variables, which included: n-term, c-term, column type, and recon solvent. One hot encoding transforms strings into binary values, where a new column is added for each type of variable for that feature. Only one of the N-types of variables will receive a 1 for each sample. This is needed as machine learning models cannot intake strings as features. It is also essential as there are several variables in this dataset that are not numerical.

Suppose there are three recon solvents used in testing. If each of the three sample peptides had a different recon solvent, there would be 3 columns as inputs, where each peptide receives a 1 for the solvent that was used in that test. By using binary values instead of having one column (recon) and an increasing discrete number for each unique type of, further weight is not added to higher values, which would influence the regression models.

Peptides	recon_Solvent1	recon_Solvent2	recon_Solvent3
Peptide 1	**1**	0	0
Peptide 2	0	**1**	0
Peptide 3	0	0	**1**

Table 8: One Hot Encoding Example

Preprocessing also included reviewing the dataset for missing information and filling in the null values. This was most needed for the column, recon solvent, and gradient parameters. The recon solvent and column variables are parsed from manually entered fields. While we humans are incredibly adept and versatile, our hand-entered inputs on test equipment do not always follow standard operating

[10] SciKit-Learn is a machine learning library built for the Python programming language.

procedures. These instances were amended. On a small number of occasions, the gradient parameters were not populated when the script attempted to fetch their values and returned null values. For these instances, the data was populated after manual review and intervention.

3.7 Selected Algorithms

The predictors used for in this thesis include: Ridge, LASSO, ElasticNet, Random Forest, Single Regression, and Multi-Variable Regression models. For each of the individual models, the corresponding function in Sci-Kit Learn was used. This package was chosen as it is well documented, easily implementable, and suitable for this application. RMSE was used as the metric for evaluation of model performance.

Ordinary least squares (OLS) linear and multi-variable regression methods were initially used to understand the predictive strength of single variables as well as all the model features. Equation 4 calculates the residual sum of squares (RSS) used in linear regression. Here, n represents the number of distinct datapoints, β are the regression coefficients, y_i is the test value, $\hat{y}_i$ is the predicted value, p represents the number of variables, and x_{ij} represents the j-th variable for the i-th observation.

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2$$

Equation 4: Residual Sum of Squares

Next, regularized models, including Ridge, LASSO, and ElasticNet, were evaluated. These models are L1[11] and L2[12] regularized models. More specifically, LASSO uses L1 regularization, Ridge uses L2 regularization, and ElasticNet uses a combination of both L1 and L2 regularization. Regularization is applied in the form of a penalty term (alpha), that is applied in addition to the typical residual sum of squares loss function to minimize the number of variables used in the prediction. Regularization helps prevent an overfit model.

[11] L1 regularization adds a penalty to the loss function equal to the absolute value of magnitude of the coefficients

[12] L2 regularization adds a penalty to the loss function equal to the squared value of magnitude of the coefficients

$$LASSO = RSS + \alpha \sum_{j=1}^{P} |\beta_j|$$

Equation 5: Lasso Regularization

$$Ridge = RSS + \alpha \sum_{j=1}^{P} {\beta_j}^2$$

Equation 6: Ridge Regularization

$$ElasticNet = RSS + l_1 \alpha \sum_{j=1}^{P} |\beta_j| + \frac{1 - l_1}{2} \alpha \sum_{j=1}^{P} {\beta_j}^2$$

Equation 7: Elastic Net Regularization

In the regularized model equations alpha is the penalty coefficient that when increased forces the coefficients to zero, thereby removing that feature from the predictor.

In ElasticNet, l_1 is a mixing term. With l_1 = 1, the model is equivalent to Ridge Regression, conversely, when l_1 = 0, the model is equivalent to LASSO regression. These terms were optimized using a parameter sweep to find the solution that gave the lowest RMSE on the evaluation data.

Model	Ridge	LASSO	Elastic Net
Hyperparameters	Alpha: [10^-2 to 10, 20 points]	Alpha: [10^-5 to 1, 20 points]	Alpha: [10^-5 to 1, 20 points]
			L1 Ratio: [0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99, 1]

Table 9: Regularized Model Parameter Search

Figure 18 visualizes how a parameter sweep works for one variable and one model type. In this case, the R^2 value is the scoring function. The train and test fit values of a LASSO model are plotted against different values for the penalty term, alpha. Here it can be seen that the alpha term of .01 gives the highest testing R^2 value. This penalty term would then be used on the evaluation data to validate the model's tuned performance. Using this method across all the selected models allows a comparison in performance to be made.

For models with many hyperparameters, this method still applies. However, each parameter is changed one at a time, which leads to an exhaustive and time intensive search. Luckily, functions in

SciKit-Learn do the heavy lifting, and all that is needed are the ranges for the hyperparameters that are intended to optimize for, and the number of points within that range of interest in testing.

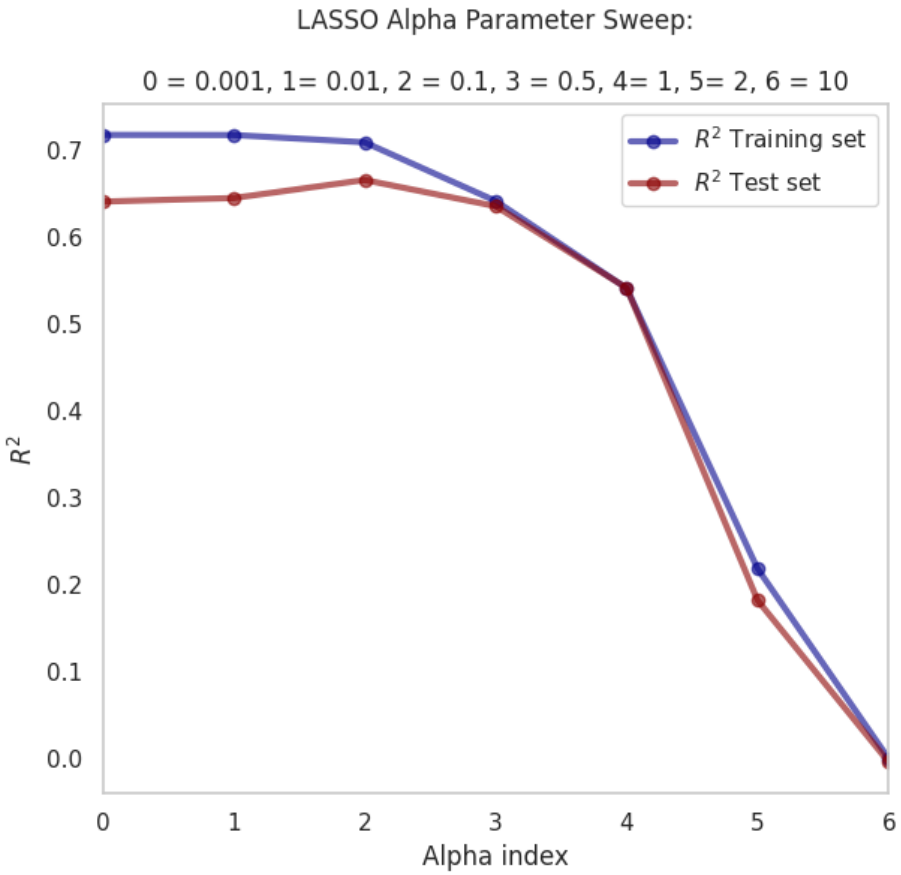


Figure 18: LASSO Alpha Parameter Sweep

The last model type evaluated was the Random Forest Regressor, which is a decision tree-based model. Random forest algorithms build many individual decision trees based on the model features and take the mean of them all for a final predicted value. This helps decrease the variance of the model. The construction of these trees is governed by hyperparameters that were optimally found using a randomized grid search. This randomized grid search used the parameters and areas in Table 10 for tuning.

Model	Random Forest	Parameter Purpose
Hyperparameters	Estimators [10 to 2500, 100 points]	Number of Trees in forest
	Max Features [Auto, Sqrt, Log2]	Number of features to consider when splitting a node
	Max Depth: [10 to 1100, 50 points]	Maximum Depth of Each Tree
	Min Samples per Split: [2, 5, 10, 15, 20]	Minimum number of samples required to split an internal node
	Min Samples per Leaf: [1, 2, 4, 10, 15]	Minimum number of samples required at each leaf node
	Bootstrap: [True, False]	If bootstrap samples are used when building trees

Table 10: Random Forest Hyperparameter Grid Search Variables

4 Results

This chapter is organized into the two phases presented in the problem statement: A-LCMS retention time prediction (Sections 4.1 through 4.3) and the purification ACN% prediction (4.4 through 4.5). The purpose is to summarize and present results from all steps of the project including dataset analysis and preliminary testing, model selection, and final implemented results.

4.1 LC Retention Time Normality

Over the past 8 months Mytide has used Pexiganan as an internal testing and validation peptide sequence for the process. This sequence, tested 21 separate times, can be used for checking retention time normality. When evaluating data, both normal probability plots and histograms are effective in visualizing the distribution of the retention times. Figure 19 shows a Q-Q plot for the retention time as compared to a normal distribution. The data indicates it is derived from a normal distribution as the samples are aligned closely with a linear fit.

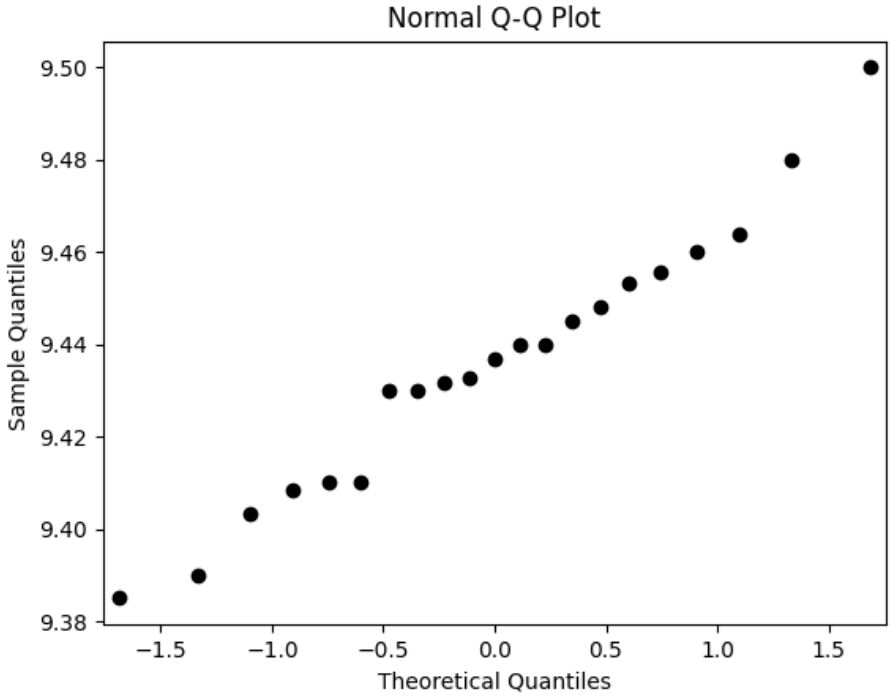


Figure 19: Pexiganan LC Retention Time Probability Plot

The histogram, below, also shows a normal distribution centered around 9.44 minutes. The standard deviation of these retention times is 0.028 minutes, or 1.68 seconds. Based on this information, it can be understood that the retention time of a single peptide sequence is repeatable and precise. As the predictor will use inputs of varied peptide sequences, it will not be more accurate than this measure.

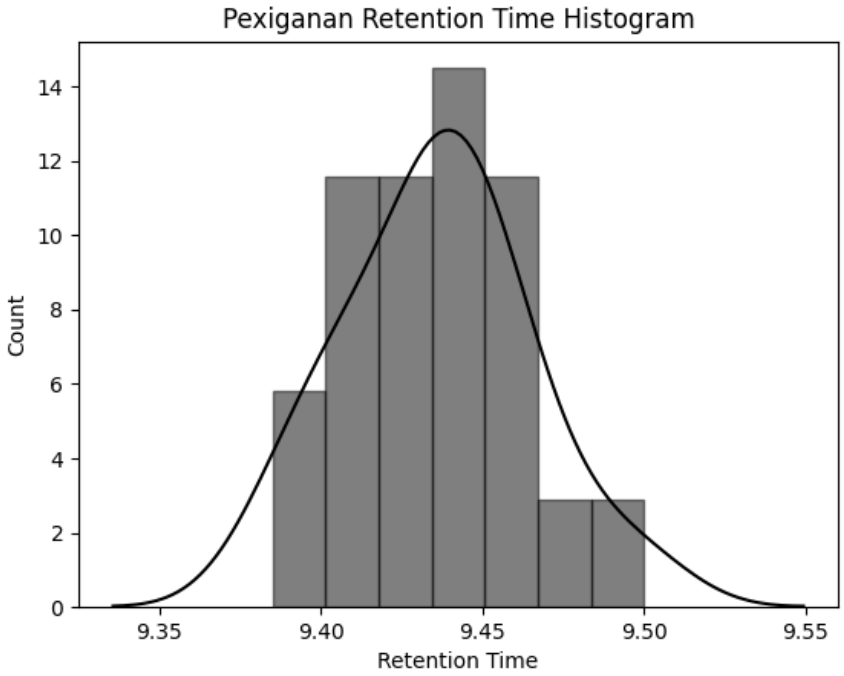


Figure 20: Pexiganan LC Retention Time Histogram

4.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) summarizes the process of visualizing data to better understand the relationships between independent variables and between the independent variables and dependent variable. With EDA, a variety of techniques can be used to maximize insight into the dataset, uncover underlying structure, identify outliers, and extract important variables [26].

A starting point is the histogram in Figure 21 which summarizes all retention times for the dataset. There is clearly a maximum at just less than 9 minutes, and approximately equal occurrences plus or minus one minute from the peak.

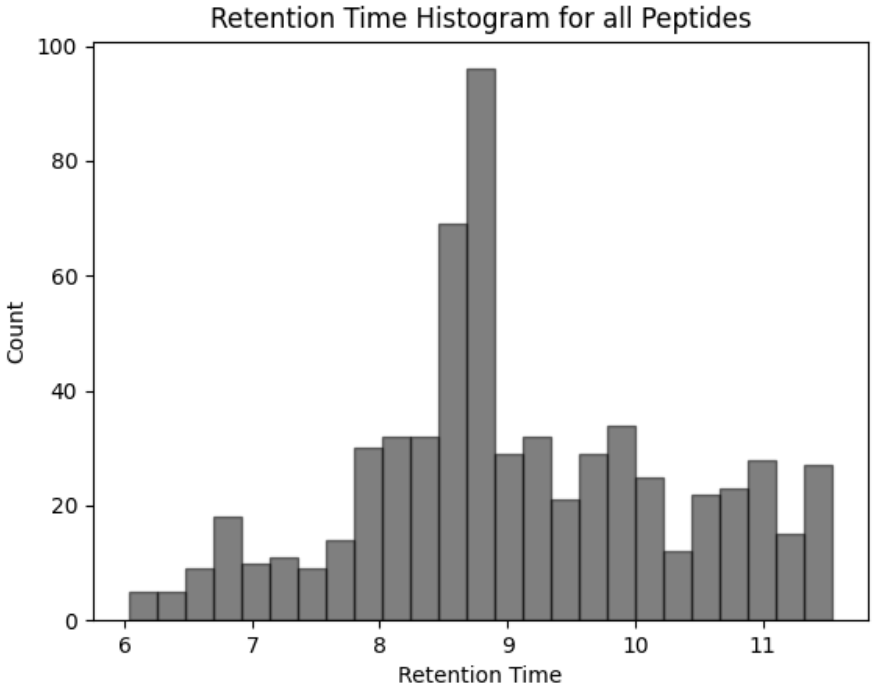


Figure 21: Retention Times for All Peptides

To establish a basic understanding between the input features and between the inputs and the retention time, a correlation matrix is used. In this case, the correlation matrix uses the Pearson correlation coefficient (PCC) to compare individual features and measure the strength of their linear association. A color map based on the PCC provides another context for quickly scanning the variable relationships. This is shown in the Figure 22.

$$r = \frac{\sum(x_i - \bar{x})\sum(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

Equation 8: Pearson Correlation Coefficient. Here $\bar{x}$ and $\bar{y}$ are the sample means of two arrays of values. When r = 1 there is a perfect correlation, r = -1, a perfectly inverse relationship, lastly, r = 0 indicates no correlation.

Reviewing the correlation coefficients revealed that several variables are highly correlated. These include amino acid length and monoisotopic mass (.94), refractivity and bulkiness (.78), synthesis difficulty and percentages of turn secondary structures in the peptide (.96), and hydrophilicity and polarity (.80). The variables that indicate the percentage of each individual amino acid (22 total) were excluded from this figure for readability.

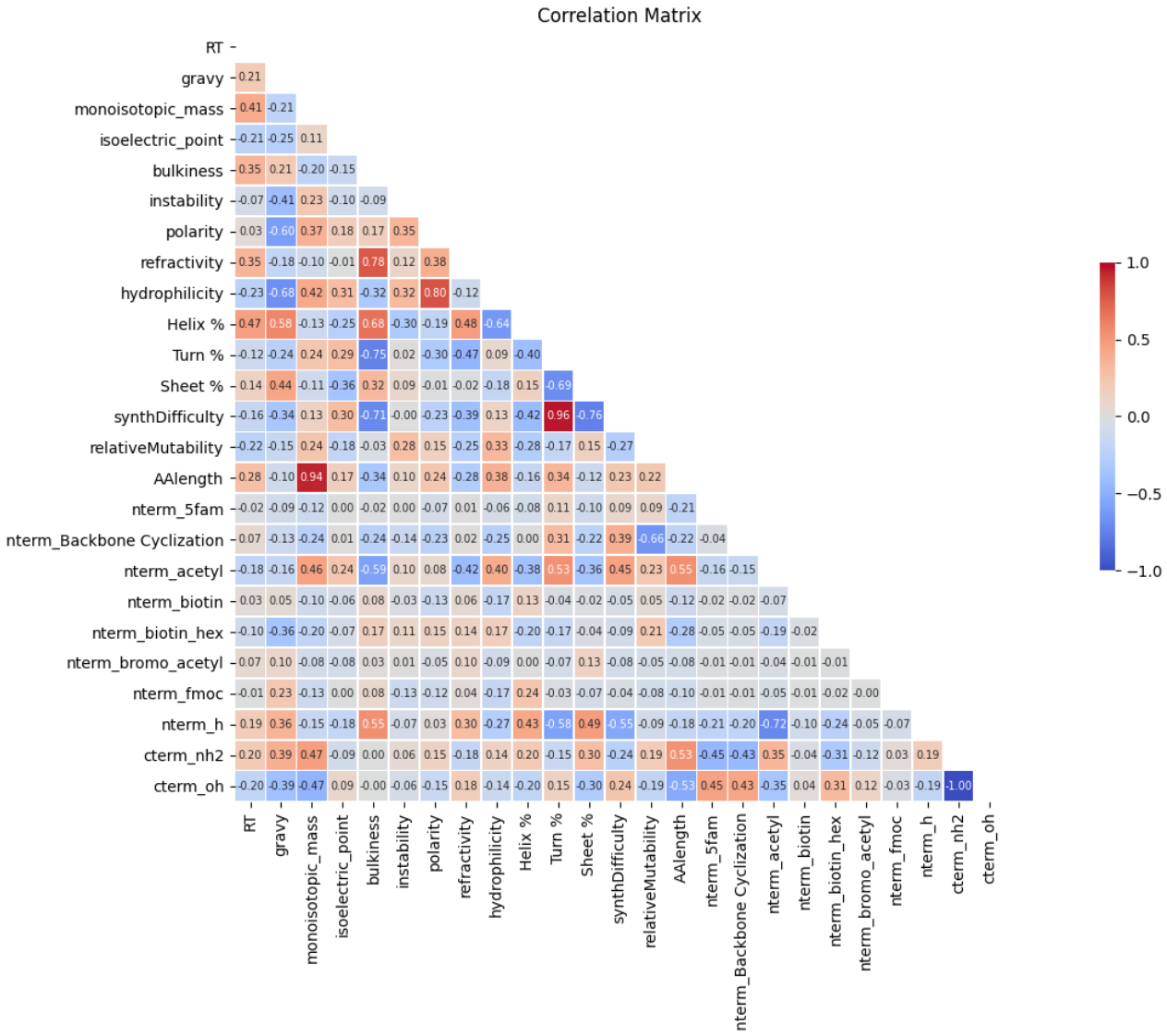


Figure 22: Correlation matrix between input variables for LCMS prediction

The correlation matrix also includes the independent variable, retention time (RT). The top six correlated variables to retention time include: Helix % (0.47), monoisotopic mass (0.41), refractivity and bulkiness (0.35), peptide length (0.28), and gravy score (0.21).

Swarm plots overlaid on boxplots show the range and distribution of the retention time for the varied assortment of peptide modifications on the N-terminus and C-terminus. In these plots, the black points represent the actual test values. They are spaced apart for visibility of their distribution over the retention time. For C-terminus modifications, the NH2 modification is more common and spans the entire retention time range while skewed to longer retention times.

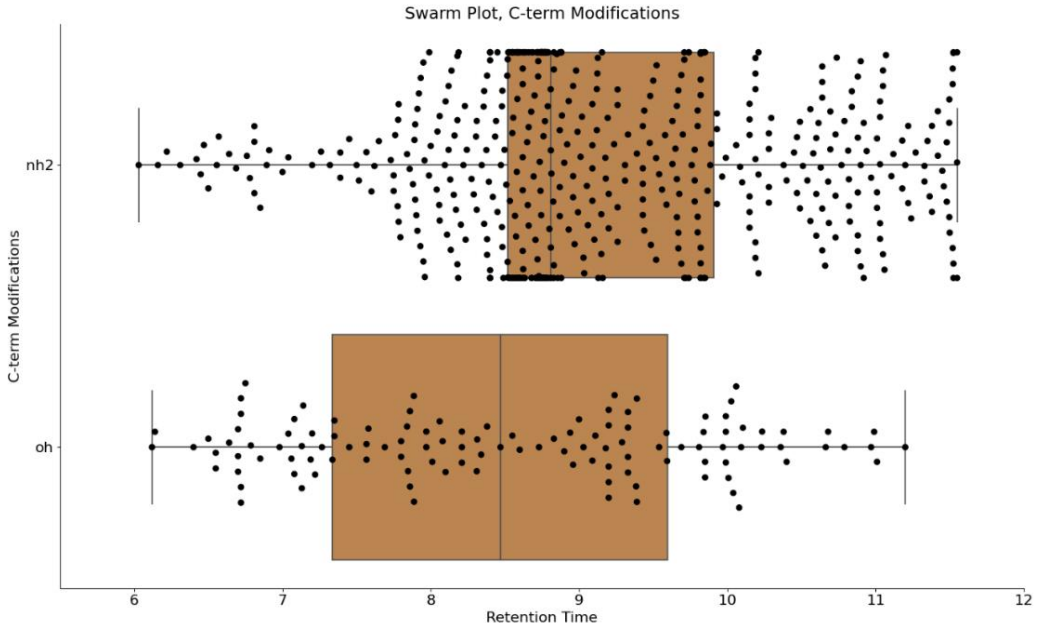


Figure 23: Swarmplot and Boxplot of C-Terminus Modifications

N-terminus modifications show the most common modification to be H, and least common is Bromo Acetyl.

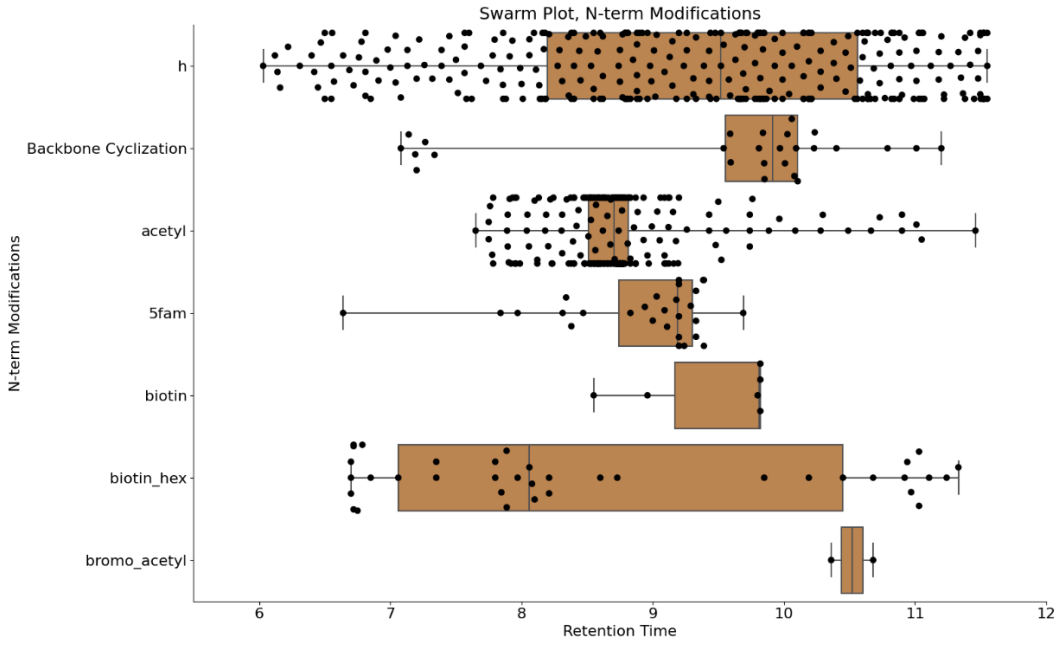


Figure 24: Swarmplot and Boxplot of N-Terminus Modifications

Outlier analysis was performed with boxplots to visualize if there were any samples that were fundamentally different than the rest of the population. The first dependent variable that was investigated was the peptide length. It can be identified in Figure 25, that there are ten samples with sequence lengths significantly longer than the bulk of the population (clusters at 85 and 100). These samples were omitted from the study as the chromatography technique cannot be well represented by such long peptides as they exhibit slightly different interactions with the column. The other dependent variables were also analyzed with boxplots, however no other discoveries were made.

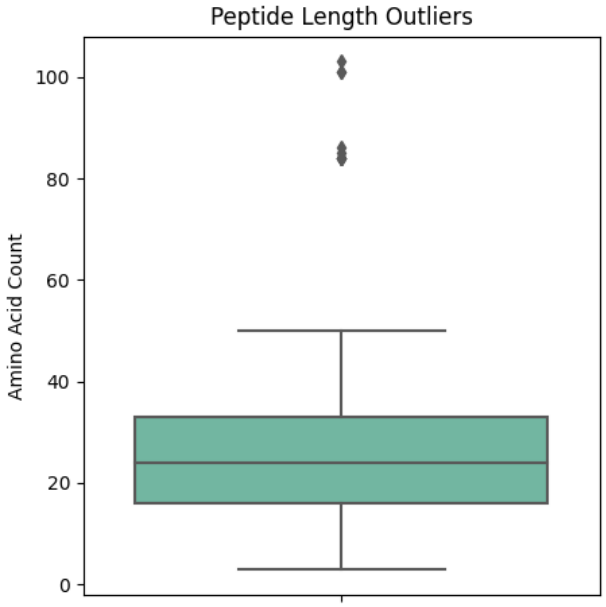


Figure 25: Peptide Length Outliers. Outliers clusters at 85 and 100 were found. These peptides were omitted due to their large difference in overall length to that of commonly run peptides.

4.3 A-LCMS Predictor: Model Comparison

Ordinary Least Squares (OLS) regression was used as a first pass estimator for the A-LCMS retention time. Using 10-fold cross validation, the boxplot captures the RMSE for each of the 10 folds, across five variables. These variables were selected as they are the most correlated with the retention time as found in the correlation matrix in Figure 22.

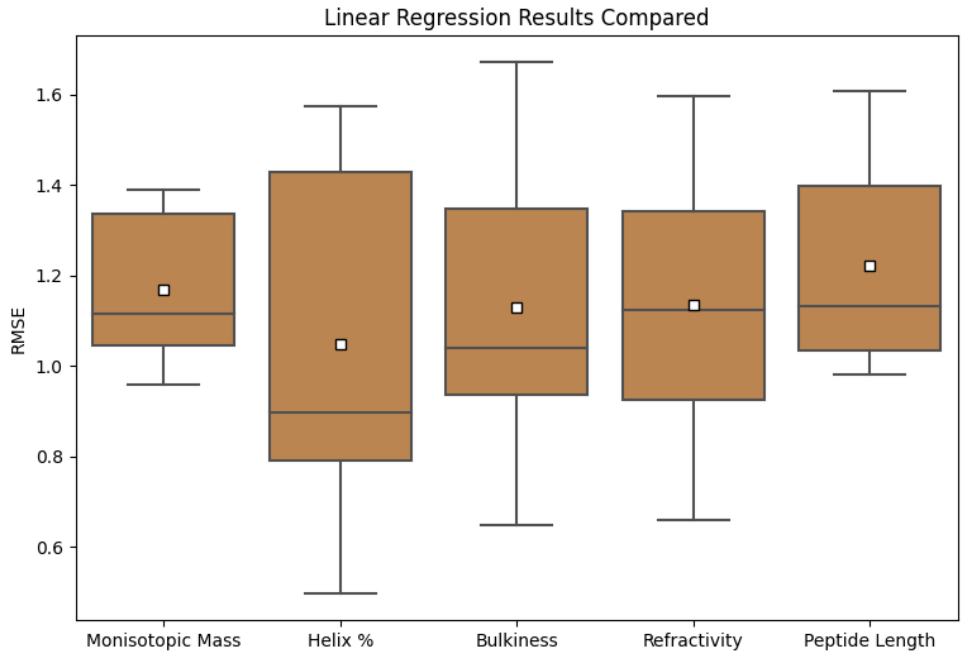


Figure 26: Linear Regression Cross Validated Results Compared. The mean RMSE is shown with the white box near the center of the individual boxplots. The results for each variable are comparable, landing around 1.1 minutes. While the Helix % variable has a slightly better RMSE than other factors, its range is larger.

Next, more detailed models were considered. Below, multi-variate, LASSO, Ridge, and Elastic Net model results are summarized. These results all produce models that perform better than a standard linear regression model.

Test Set Metrics	Random Forest	Ridge	Lasso	Elastic Net	Multivariate
RMSE	0.55	0.72	0.72	0.72	0.74
R^2	0.78	0.65	0.66	0.65	0.63

Figure 27: A-LCMS Predictor Model Results Summary

The Random Forest model gives the lowest RMSE and highest fit of the model types. For this model, a randomized hyperparameter grid search with K-fold cross-validation was performed tuning the parameters in Table 10. As a result, the following parameters were found to be optimal:

- Number of Estimators: 437
- Minimum Samples per Split: 2
- Minimum Samples per Leaf: 1
- Max Features, sqrt
- Max Depth: 922
- Bootstrap: False

A plot of the predicted versus actual retention times of the test set can be seen in Figure 28. The dark orange points are the training test samples, while the lighter shaded points along the diagonal line are the training samples.

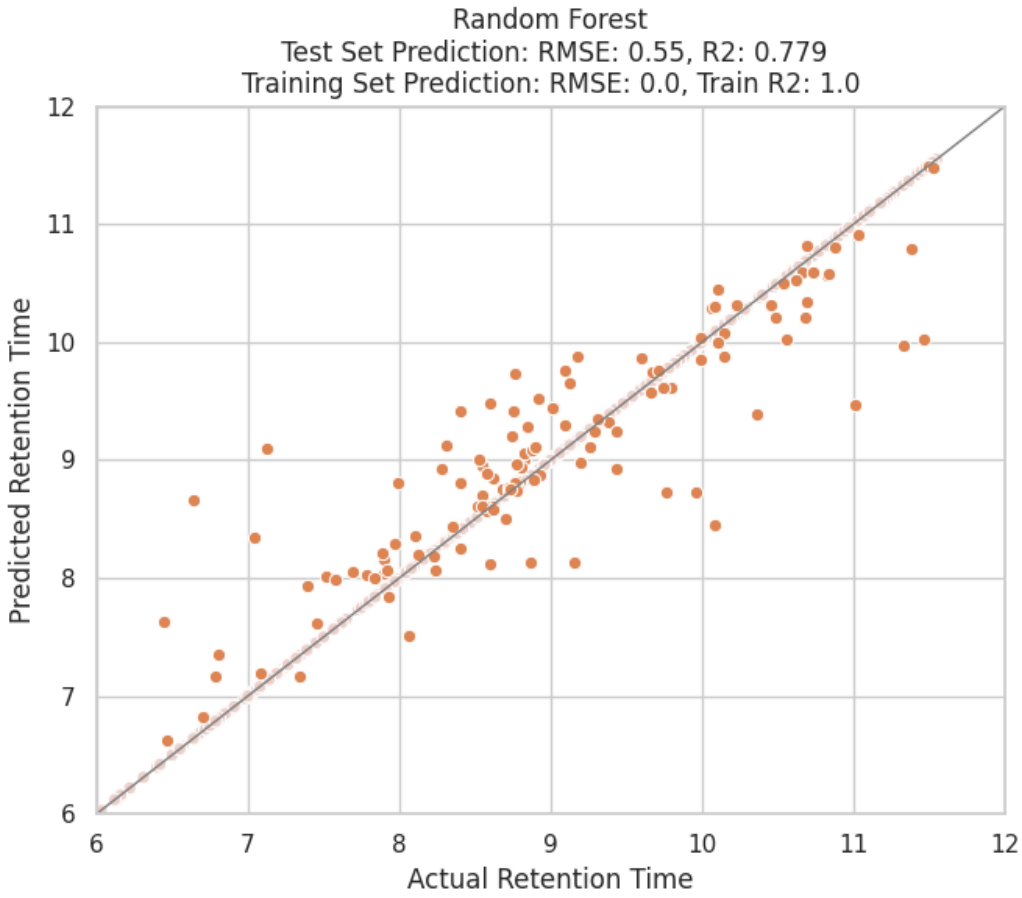


Figure 28: Random Forest with Hyperparameter Optimization

Examining the residuals showed a random distribution throughout the range of the retention time, with no trends indicating non-linearity. The residuals are symmetric about 0, no glaring outliers were revealed.

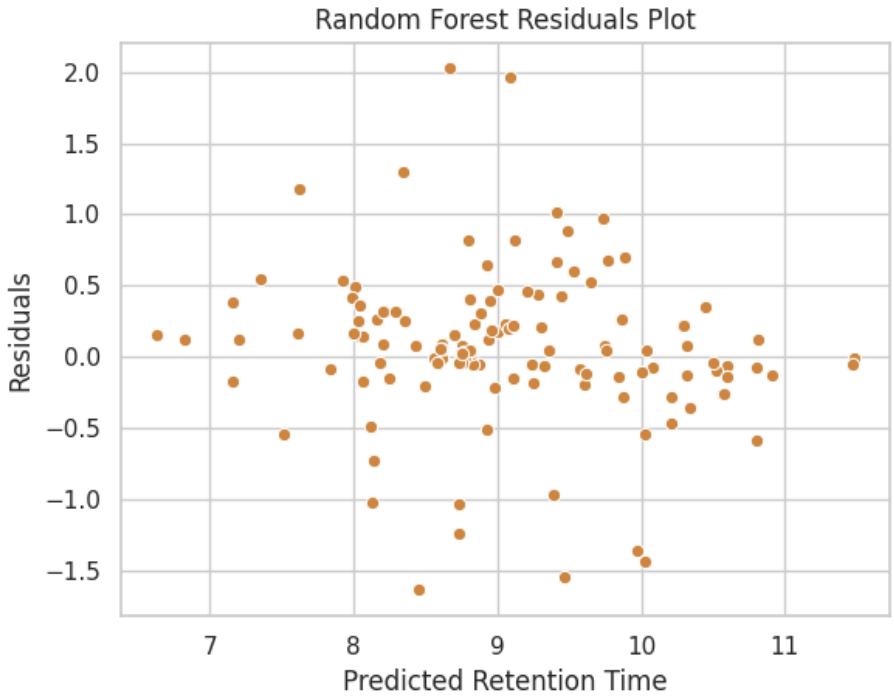


Figure 29: Random Forest Residual vs. Fits Plot

Feature importance is a property within the Random Forest scikit-learn algorithm that measures the features that best explain the predicted output. In general, the higher the number for each feature, the more important. The top ten features are shown in Table 11. One observation here is that four of the top five most important features are identical to the features most correlated to the retention time as shown in Figure 22.

Feature	Importance
Monoisotopic Mass	0.095
Helix %	0.069
Peptide Length	0.068
Bulkiness	0.051
L%	0.044
Sheet %	0.038
Refractivity	0.037
F%	0.034
Gravy	0.034
Relative Mutability	0.033

Table 11: Random Forest Feature Importance

4.4 Purification ACN% Predictor Deployment Workflow

Deployment of a machine learning model involves taking inputs, running the inputs through a pipeline, and using the output to make production or business decisions on new data. Deployment of an end-to-end machine learning model requires separate testing and training processes. This is critical as the testing data will not be always on hand and can come at a point in the future. Having separate training and evaluation models also prevent the need to re-train the model each time there is new data. Figure 30 illustrates the breakout of files and file types used for the purification ACN% predictor.

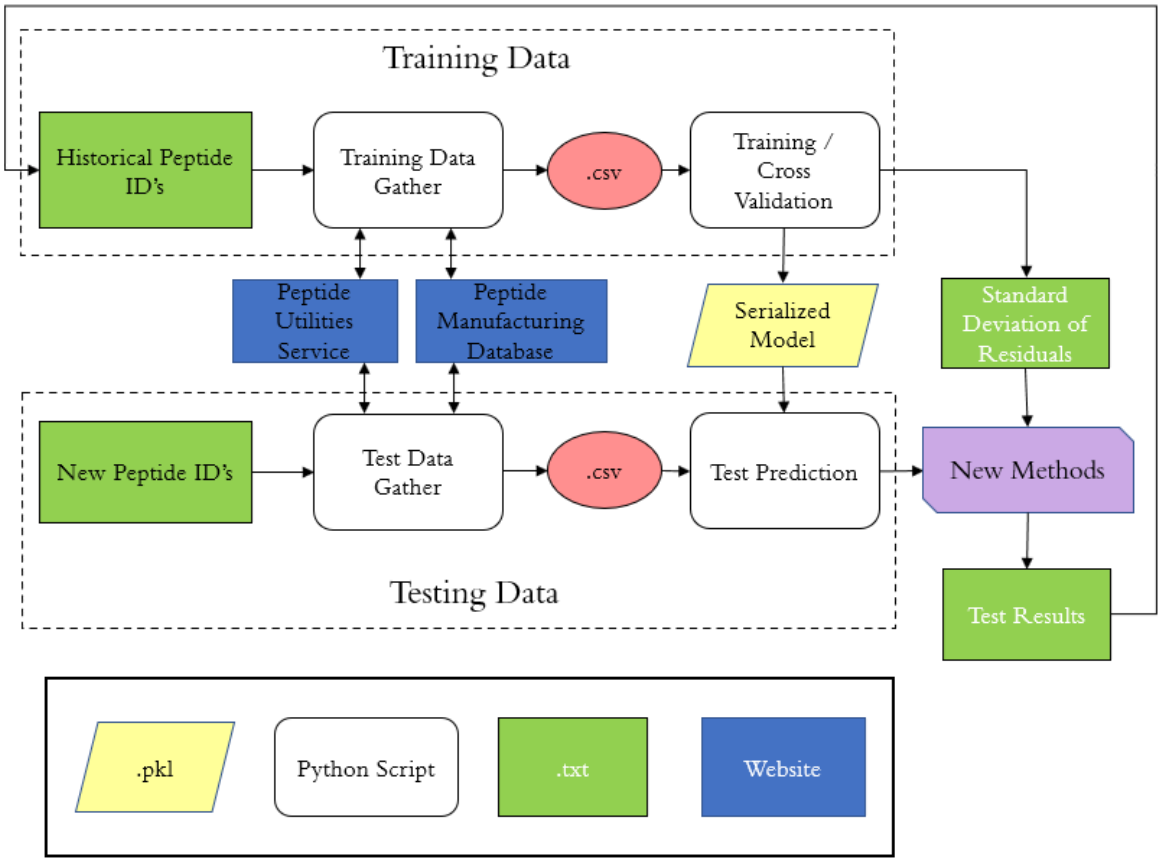


Figure 30: Machine Learning Workflow for Purification ACN% Prediction

The training data process takes in a text file containing peptide ID's, gathers properties using REST API calls to the Peptide Utilities API and test results server, and outputs a .csv containing all needed testing results, gradient parameters, and physio-chemical properties. This matrix of input features is used as the input for model training and selection.

In this case, the model selection and validation were performed with 5-fold cross validation on the training data. More specifically, 70% of the training data searched for optimal hyperparameters of four different model types (Ridge, Lasso, ElasticNet, and Random Forest) using the criteria in Table 9 and Table 10. The remaining 30% was used to evaluate and select the best performing model. Table 12 shows the results of these four model types.

Evaluation of the ACN% predictor was based on four model types: Ridge, ElasticNet, LASSO, and Random Forest. Shown in this table are RMSE results for each model type with the optimized hyperparameters. Random state is a seed specifier in Python used in the train, test splitting function. Specifying a random state value maintains the same values in the testing and training set. By fixing the state for each model type the samples tested with each algorithm are the same. Overall, we can see that ElasticNet and LASSO perform similarly, while the LASSO model has a standard deviation on the RMSE of all random states which is slightly better.

		Ridge	ElasticNet	LASSO	Random Forest
Optimal Hyper-parameters	alpha	603	0.0723	0.0085	2000 estimators
	L1	N/A	0.1	N/A	Max Depth: 200
		RMSE			
Random State	1	0.074	0.047	0.049	0.0419
	2	0.088	0.067	0.068	0.0708
	3	0.069	0.068	0.066	0.0697
	4	0.07	0.052	0.052	0.0495
	5	0.084	0.059	0.06	0.068
	6	0.069	0.045	0.045	0.0483
	7	0.109	0.089	0.088	0.0944
	8	0.089	0.068	0.066	0.0739
	9	0.069	0.06	0.059	0.0649
	10	0.076	0.057	0.056	0.0695
	mean	0.08	0.061	0.061	0.065
	std	0.013	0.013	0.012	0.015

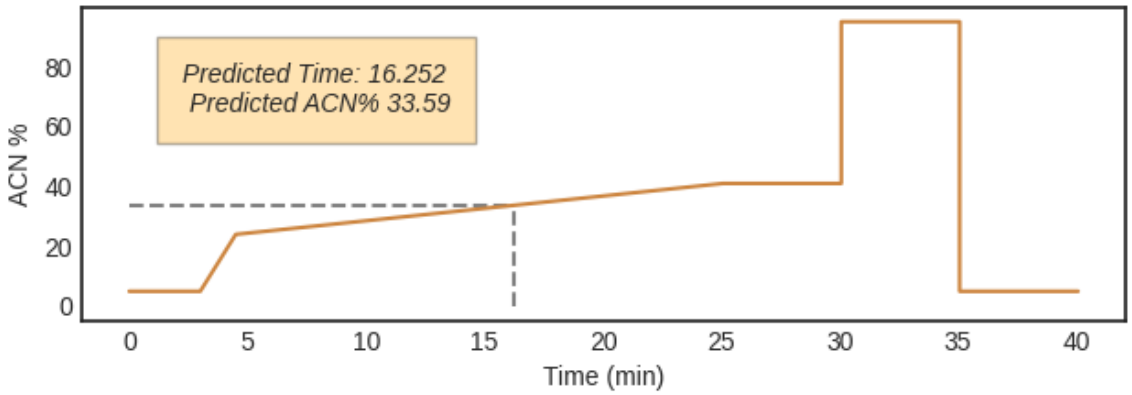
Table 12: Purification ACN% Model Comparison

Using the best performing model, found to be the LASSO model, a serialized version is output with Pickle, a Python module. "Pickling" the model encodes all aspects of the machine learning training pipeline that can include one hot encoding requirements, scaling and normalization parameters for preprocessing, and regression coefficients, into a byte stream that can be saved to disk or sent to another file. In this application, the model is "unpickled", or decoded, in the test prediction file.

The testing data follows a similar process to the training data as the model features need to match the training set features. After the serialized model is decoded, the testing set features are imported into the evaluation file, and predictions are made. These predictions are then used to generate new methods for the purification process.

Model improvement is the final step in the machine learning workflow. This is done by taking the output of the test results and feeding them back into the model. Without this feedback, the model training dataset would not grow, and the predictive power would remain stagnant.

Figure 31 illustrates a generated method based on the prediction of the machine learning algorithm. The algorithm's target (predicted ACN%), is translated into a predicted time using the slope and the time value of point 2. The ACN% for the start and stop of the linear gradient are found using the standard deviation of prediction residuals from the training set.



fc0a07c7-0645-4427-a6df-417d8c20eb03
Slope: 0.008
Loading / Wash Flow: 18.9, Linear Gradient Flow: 6

Points	Start Time	Duration	ACN %
1	3	1.5	5.0
2	4.5	20.57	24.0
3	25.07	5.0	41.0
4	30.07	0.0	41.0
5	30.07	5.0	95.0
6	35.07	0.0	95.0
7	35.07	5.0	5.0

Figure 31: Generated Gradient from Purification ACN% Prediction. In this plot the dotted gray horizontal line represents the predicted ACN% for the specific peptide, the vertical dotted line represents the predicted time, and the orange line is the purification gradient. The points which form this gradient are presented in a tabular format for programming the purification machine.

This format gives the user programming the purification machine a table of time values and ACN % values to key in so a peptide-specific gradient can be run. The line plot visualizes this table for added convenience. Last, the flow rate is included for specification in the programming of the purification machine, while the slope is shown for convenience.

4.5 Purification ACN% Predictor Implementation

Now that the method output has been introduced at a high level, it is critical to know how the predicted value was found. Figure 32 shows the relationship between actual and predicted values for the best performing LASSO model. These points are from the 30% test split on the training data, also known as the validation data.

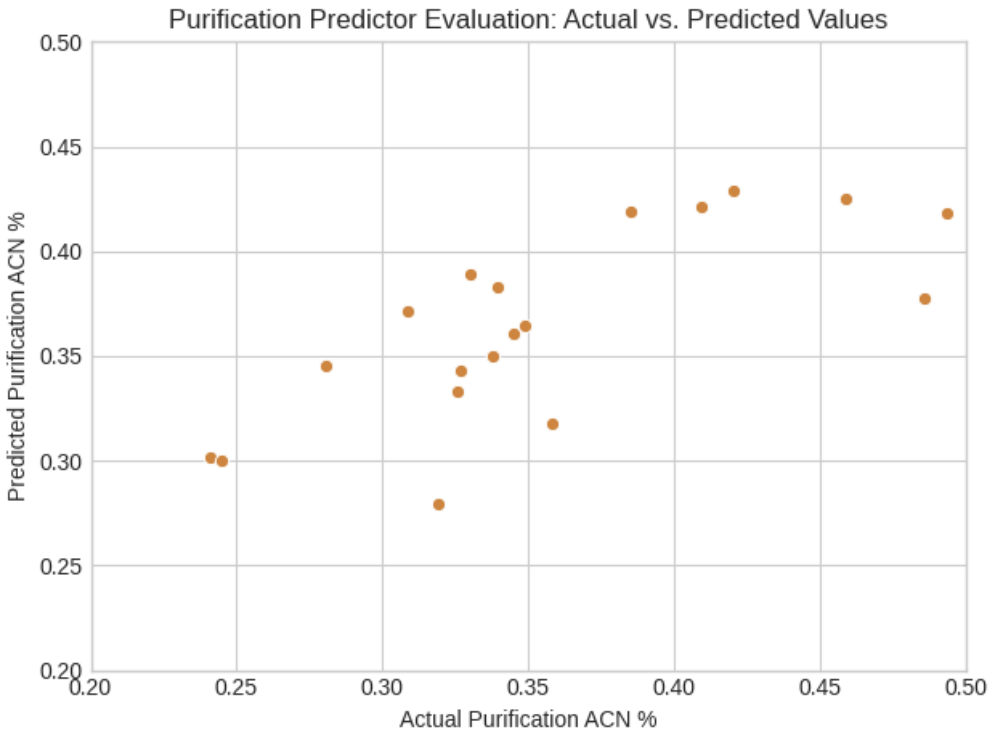


Figure 32: Purification ACN % Actual vs Predicted Values

A look at the residuals reveal no trends or non-linearity, and while there is limited data, the histogram (Figure 34) does trend towards a normal distribution. The standard deviation (σ) of the prediction residuals is 0.047. This value is critical as it is used as an input for the method generation function. Within the method generation function this standard deviation is multiplied by a safety factor that sets the boundaries for the solvent concentration start and end points of the linear gradient. Multiplying by a safety factor increases the range of the solvent window around the predicted concentration and is a safety mechanism to protect against differences between the actual and predicted values of the concentration.

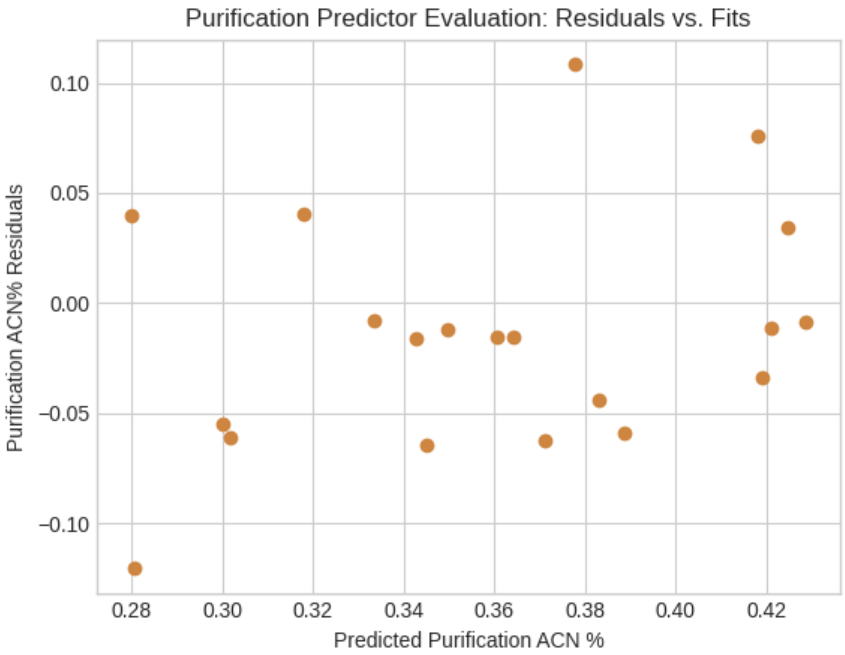


Figure 33: Purification ACN % - Residuals vs. Fits

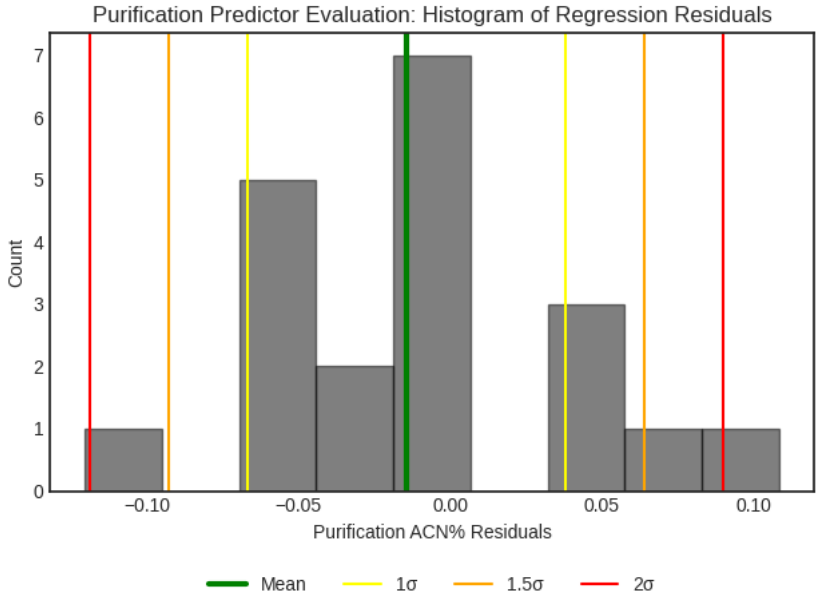


Figure 34: Purification ACN% - Histogram of Residuals

In Figure 35, the solvent window uses this safety factor to increase the range of the solvent across the linear gradient. This parameter as well as the slope of the linear gradient have a large and direct influence on the overall test time. For initial experiments, the slope and window were conservative

to verify the peptide eluted during the linear gradient, and not during the initial loading or final washing phases of the gradient.

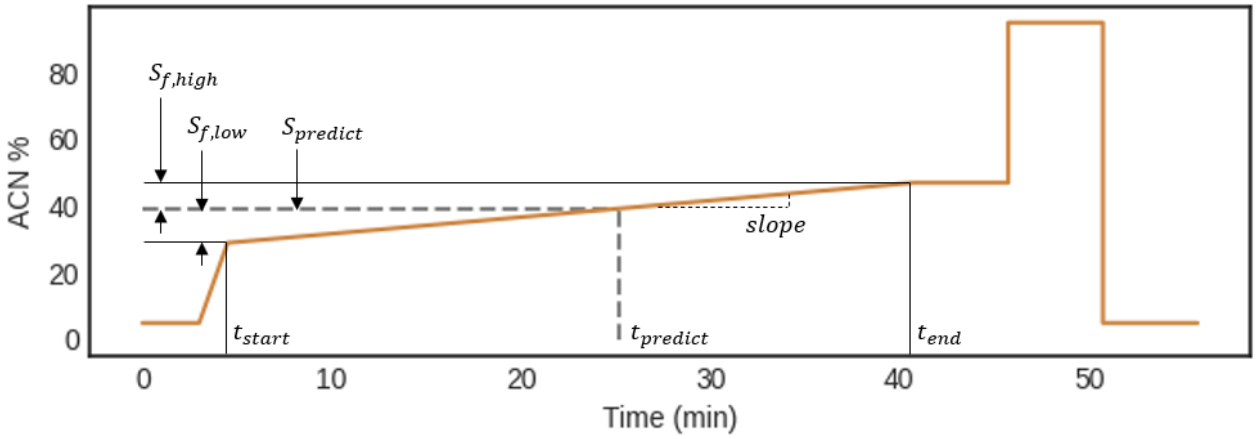


Figure 35: Gradient Window Development from Standard Deviation of Residuals

The following formulas were used within the method generation function to produce the upper and lower bounds of the linear gradient, prediction time, and end time. The lower bound uses a higher safety factor as it was learned experimentally that peptides that elute early in the gradient have poor resolution. This is due to the aggressive increase in solvent concentration up to the start of the linear gradient. With this aggressive increase the compounds which elute along that steep slope will elute at the same time, crowding the chromatogram near the potential target.

$$S_{f,high} = 1.5\sigma$$

Equation 9: Lower Tolerance for Purification Prediction

$$S_{f,low} = 2\sigma$$

Equation 10: Upper Tolerance for Purification Prediction

$$t_{predict} = \frac{(S_{predict} - S_{f,low})}{slope} + t_{start}$$

Equation 11: Purification Prediction Time Interpolation

$$t_{end} = \frac{(S_{f,high} - S_{f,low})}{slope} + t_{start}$$

Equation 12: End of Linear Gradient Time Interpolation

4.6 Purification ACN% Predictor Testing

The predictor was initially tested on 12 peptides split over two days. For these tests, the slope of the linear gradient was set to 0.5 ACN % / min which was the slope used in the historical dataset. The predicted percentage of ACN ranged from 34.53 to 41.70. Overall, the model overpredicted by an average of 5% ACN. However, it is encouraging that the standard deviation of this error was low at 1.2.

Date	Sample	Actual ACN%	Predicted ACN	Error	Solvent Start	Solvent End	slope
7.22	Sample 1	33.05	38.76	**5.71**	26.00	47.00	0.51
7.22	Sample 2	35.50	39.97	**4.47**	27.00	48.00	0.51
7.22	Sample 3	33.00	38.07	**5.07**	25.00	46.00	0.51
7.22	Sample 4	33.95	38.83	**4.88**	26.00	47.00	0.51
7.22	Sample 5	30.43	36.89	**6.46**	24.00	45.00	0.51
7.22	Sample 6	28.07	35.21	**7.14**	22.00	43.00	0.51
7.23	Sample 7	38.65	41.70	**3.04**	31.00	49.00	0.50
7.23	Sample 8	35.66	40.27	**4.60**	30.00	48.00	0.50
7.23	Sample 9	35.00	39.15	**4.15**	29.00	47.00	0.50
7.23	Sample 10	33.11	38.00	**4.89**	28.00	46.00	0.50
7.23	Sample 11	35.82	39.83	**4.01**	29.00	48.00	0.52
7.23	Sample 12	27.81	34.53	**6.72**	24.00	42.00	0.50
			Average	**5.09**			
			Standard Deviation	**1.21**			

Table 13: Initial Test of Purification Predictor

Next, the model was re-trained with these 12 additional datapoints. An offset of 5% ACN was also added into the model based on the initial results. This was done by modifying the predictor and subtracting 5% ACN from the predicted value, then using this value to generate the method. Following the model re-training, a batch of 11 more peptides were run. In this batch the model underpredicted (actual ACN% was higher than predicted ACN%) by an average of 2.5 % ACN.

Date	Sample	Actual ACN%	Predicted ACN	Error	Solvent Start	Solvent End	slope
7.28	Sample 13	37.63	33.14	**-4.49**	24.00	40.00	0.78
7.28	Sample 14	32.66	31.41	**-1.25**	22.00	38.00	0.78
7.28	Sample 15	35.94	33.59	**-2.35**	24.00	41.00	0.83
7.28	Sample 16	31.36	30.68	**-0.68**	21.00	38.00	0.83
7.28	Sample 17	33.60	31.86	**-1.74**	22.00	39.00	0.83
7.28	Sample 18	36.66	33.00	**-3.67**	24.00	40.00	0.78
7.28	Sample 19	34.32	32.10	**-2.22**	23.00	39.00	0.78
7.28	Sample 20	32.88	30.89	**-2.00**	21.00	38.00	0.83
7.28	Sample 21	35.15	32.60	**-2.55**	23.00	40.00	0.83
7.28	Sample 22	35.26	33.07	**-2.19**	24.00	40.00	0.78
7.28	Sample 23	37.72	33.59	**-4.13**	24.00	41.00	0.83
			Average	**-2.48**			
			Standard Deviation	**1.18**			

Table 14: Re-trained Model, Purification Predictor Results

As more training data is entered into the model the predictor will be able to better generalize to test data. This can be seen by the improved performance between the first test set and second test set. If the 5% offset were removed from the second set, the average error would be 2.52. Between the first and second set of tests the prediction error was cut in half because more data was entered into the model for training. For future runs, it was decided to remove the offset.

4.7 Purification ACN% Predictor Slope

The slope of the linear gradient has a measurable effect on the total test time. At a slope of 0.5 ACN% / min, a linear gradient that runs 17%, the typical value for the developed predictor, will take 34 minutes. If the slope is increased to 0.8 ACN% / min, that same gradient will take 21.25 minutes. Increasing the slope is desirable provided that the vials collected contain separated compounds. To verify resolution at the increased slope, a peptide was separated into two samples. Both identical samples were run on gradients of increasing slope equal to 0.5 and 0.8 ACN% / min. All other parameters for purification remained the same. After purification, two vials for each slope were analytically verified by B-LCMS for purity. Table 15 shows the purity for each of the two vials for both runs. The purity is calculated by taking the area underneath the UV chromatogram that

corresponds to the correct product and dividing that by the total area under the UV chromatogram, which includes by-products.

Fraction	Slope	Product Area (210 nm)	Purity
73	0.5	1807	98.80%
74	0.5	3277.8	99.00%
103	0.8	3394.7	97.50%
104	0.8	2417	94.30%

Table 15: B-LCMS Purity Verification for Increased Gradient Slope

The fractions that contain most of the material are each of comparable quality, indicating no detriment to chromatographic performance at the more aggressive slope.

4.8 Cost and Time Analysis

A working short-term goal at Mytide is to produce 12 peptides per day. Given the existing purification test time at Mytide of 85 minutes and allotting 1 hour for purification sample setup and gradient programming, the maximum throughput on the purification machine is 16 peptides per day. This is assuming a 24-hour day, which is reasonable considering this machine does not need supervision once it is setup. Using this same logic, with the implemented gradients presented in this thesis, the maximum throughput is increased to 34 peptides per day for a single machine.

This is significant as Mytide is aggressively ramping production and intends on scaling production up to 144 peptides per day. Considering this information, more machines will be required to maintain the required throughput.

Gradient	1 Machine	3 Machines	5 Machines	7 Machines	9 Machines
85min	16	48	81	113	146
40min	34	103	172	241	310

Table 16: Peptide limit per day as a function of the gradient and machine count

In fact, 9 machines running at 23 hours per day are required to meet that demand with the original gradient. Optimized gradients reduce that requirement to 5 machines. Reducing machine count pushes the capital expense out further (helpful for a startup), reduces the inventory of columns, vials and solvent used, and minimizes the number of components susceptible to failure.

Another way in which the optimized gradient impacts the company's bottom line is through a reduced use of ACN solvent. The original 85-minute gradient uses a total of 326 mL of ACN, while the optimized gradient uses 205 mL (37% reduction). ACN is relatively low-cost, however in volume

production, significant savings can be realized. Purchase orders for ACN show that 16L is approximately $200. From this, we can calculate a per peptide solvent cost of $4.07 and $2.56, for the 85-minute and 40-minute gradients, respectively.

While the near-term savings at 12 peptides per day is $18.12, at 144 peptides per day the cost savings in ACN alone is $217.44 per day. This amounts to monthly savings of $362.4 and $4,348.8 respectively, assuming 20 working days per month.

5 Conclusions and Further Work

5.1 Conclusions

In summary, this thesis used machine learning models and workflows to predict analytical LCMS retention times based on physio-chemical peptide properties and purification solvent percentage using peptide properties and prior analytical test result. Implementation of the purification gradient predictor provided a 53% reduction in purification test time. To visualize the overall impact, Figure 36 overlays the predicted gradient (in red) with a historical gradient (in gray).

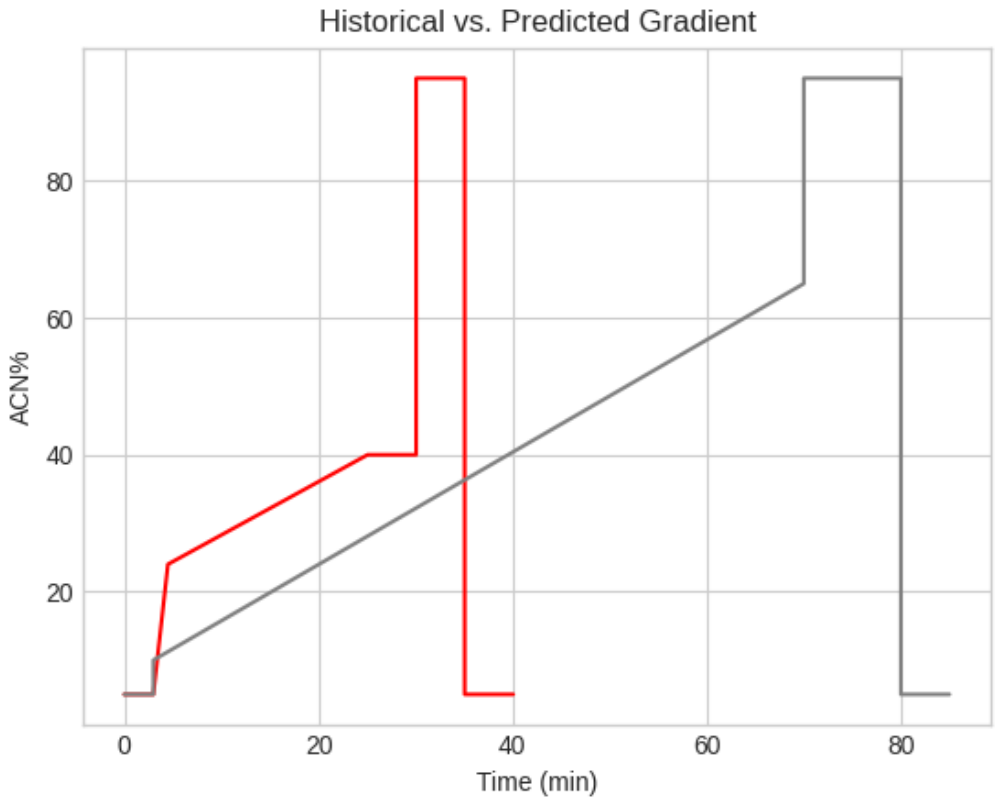


Figure 36: Predicted vs. Historical Gradient Comparison

Use of optimized gradients will increase Mytide's purification capacity to 34 peptides per day with the same equipment. Future production will be also be aided with optimized gradients as the company will require fewer machines to meet the projected demand for their system. In addition, these machines will reduce the amount of solvent used in the purification process by 37%, a savings of roughly $4348.8 per month at 144 peptides per day.

5.2 Further Work

Several areas of research came up in the implementation of the purification predictor that could increase its usability and further reduce the purification test time. Additionally, other parts in the manufacturing process could be optimized using a similar methodology.

Usability

In current form the model requires the user to run a series of Python scripts sequentially. First, to train the model (only required once, or when more data is available), second, to gather data on the test peptides, and third, to generate the new methods. These new methods are stored as pdf's in a local folder, which are then accessed and referenced when manually inputting the gradient on the purification machine. While all files are stored with Git[13], there is a learning curve for use. There are several ways this can be optimized.

1. Updating the workflow of sequential scripts into one function.
    a. Doing this would allow the user to add the test peptide ID's and run one script. They could then check the output folder for the new methods. This is of low complexity but given the timeframe of the project it was not prioritized.
2. Generation of a microweb application with use of a tool such as Flask[14].
    a. This point builds upon the first. The whole workflow could be deployed to a web application that would enable any individual to generate methods without first setting up their computer with the required scripts. This would make the whole process easier for the employees at Mytide, with some upfront development work required.
3. Change the method output structure to manually upload to the purification machine.
    a. In the current form the saved methods need to be referenced and then manually inputted into the purification machine by an operator. This process takes approximately 2 minutes per method and is prone to human error. This process could be replaced if the new gradient were saved in a format acceptable to the machine, with a naming structure noting the peptide. This is of moderate complexity as it involves interfacing with the purification machine firmware over the wireless

[13] Git is a distributed version-control system used during software development to track changes. Codebases are stored in repositories, which can be cloned (downloaded) from different computers and edited.
[14] Flask is an open-source, lightweight web framework that provides tools and features to create web applications

network. That said, this should be considered high priority. As Mytide continues to scale production, more time will be needed to program the purification machine.

Further Reduction in Test Time

1. Adjust safety factor for similar peptides
    a. The current method uses the same safety factor for the upper and lower bound on the solvent range for the linear gradient for all test peptides. It is currently set at $+1.5\sigma$, -2σ. The test method could be further improved by decreasing the size of this window. This could be accomplished by implementing a lookup function which takes the test peptide sequence and modifications, and scans through the training database for identical peptides. If there is a match, this window could be constrained as the predictor is likelier to predict an accurate value.
    b. The safety factor windowing approach can be further expanded with use of a similarity factor in place of an identical match. For example, if the peptide to be tested has an identical length, n-term and c-term, and some proportion of specific amino acids as one that was previously trained upon, there is likely high similarity between peptides. Creation of an index which quantifies these relationships would be the first step in implementing this strategy.
2. Remove isocratic hold (See Figure 37)
    a. The isocratic hold is added as a safety mechanism for late eluting peptides. If the predicted ACN% was too low, this gives added assurance the peptide will come out of the column before the test moves to the washing phase. As more peptides are added to the model, and more confidence in the model is gained, this section can be removed saving 5 minutes of test time.
3. Remove or reduce low % ACN flush at end of run (See Figure 37)
    a. The low ACN % flush is done at the end of the run for 5 minutes to ensure the column is removed of all compounds. However, when two samples are run in series, the loading phase also has a 5% ACN wash period at the start of the run. Some additional testing can be done to verify that the compounds are fully removed from the column after the high % ACN wash. With these results, the low % ACN wash can be reduced or removed, saving up to 5 minutes of test time.

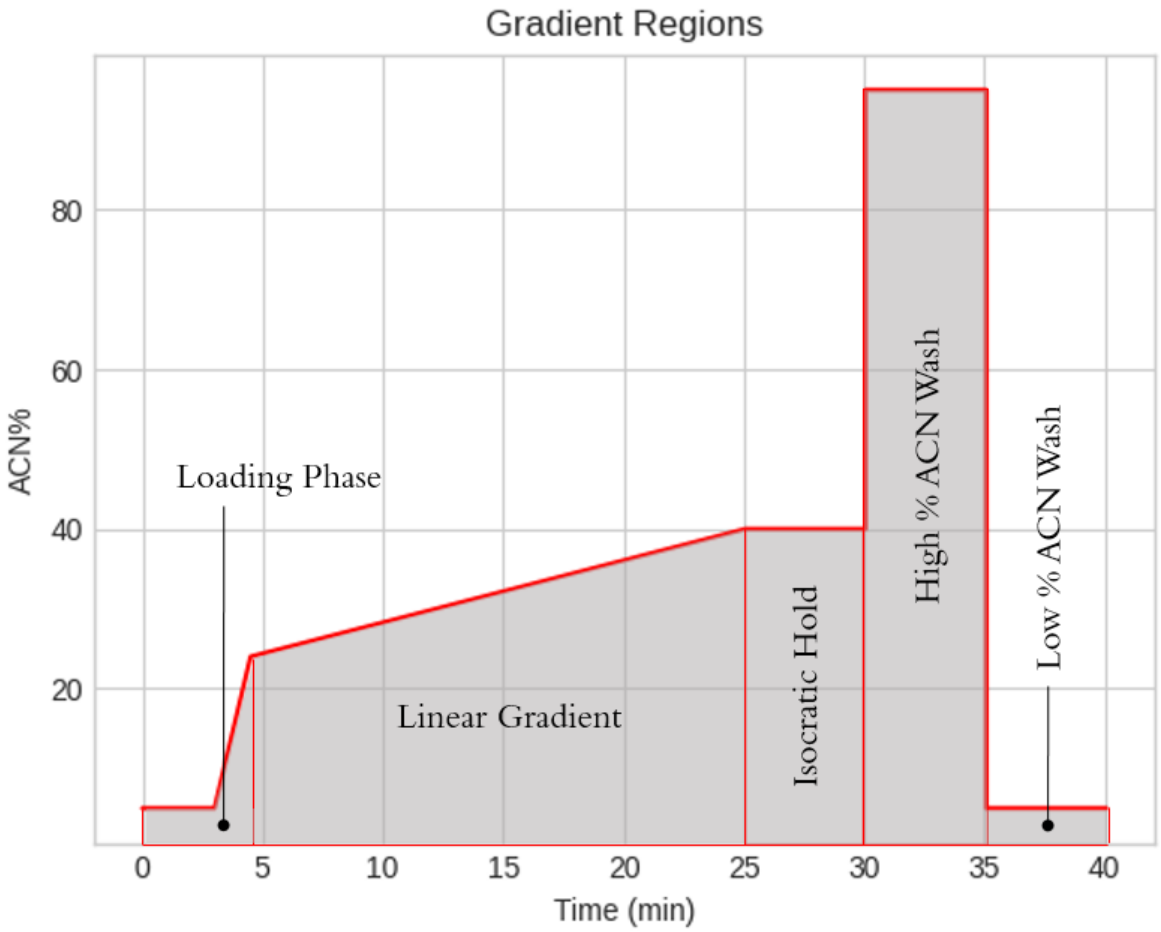


Figure 37: Purification Gradient Regions

Model Improvements and Additions

Throughout the peptide manufacturing process Mytide uses liquid chromatography including the: A-LCMS, B-LCMS, and purification process steps. As presented, this work implements the model for purification. That said, this methodology can be applied for generating methods for both the A-LCMS and B-LCMS tests.

A-LCMS

For implementation in A-LCMS testing a similar method can be used, albeit with a few changes. Per the results of Phase 1 in this thesis, an A-LCMS retention time can be predicted with accuracy of 0.55 min using only physio-chemical properties. Using this prediction, in addition to the other peptide properties as features for a new method generator, the A-LCMS test which takes 16 minutes, can also be optimized.

B-LCMS

From the purification process there are typically 3 to 4 vials that contain the compound of interest. Each of these vials are run on the B-LCMS test, which takes 16 minutes each time. This test is opportune for new method development as well. In addition to the peptide properties used in the purification predictor, the purification gradient parameters and peak time can be added as additional features.

Adding other potential areas for optimization such as the A-LCMS and B-LCMS tests will require a well-developed workflow and seamless integration with the purification machine, as described in the usability subsection. These improvements will minimize the amount of time operators need to interact with the software and free up their responsibilities to help them make the next therapeutic, one high-purity peptide at a time.

Bibliography

[1] "Bliss M. Banting, Best, and Collip's accounts of the discovery of insulin," *Bull Hist Med.,* vol. 56, pp. 554-568, 1982.

[2] K. F. a. T. Hoffmann, "Peptide Therapeutics: Current Status and Future Directions," *Drug Discovery Today,* no. 20, pp. 122-128, 2015.

[3] D. M. Lau JL, "Therapeutic peptides: Historical perspectives, current development trends, and future directions," *Bioorg Med Chem.,* pp. 2700-2707, 2018.

[4] John Rafferty, "Peptide Therapeutics and the Pharmaceutical Industry: Barriers Encountered Translating from the Laboratory to Patients," *Current Medicinal Chemistry,* vol. 23, no. 37, pp. [4231 - 4259], 2016.

[5] R. Lax, "The future of peptide development in the pharmaceutical industry," *PharManufacturing: The International Peptide Review,* vol. 2, pp. 10-15, 2010.

[6] R. Merrifield, "Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide," *Journal of the American Chemical Society,* 1963.

[7] R. Sheppard, "The Fluorenylmethoxycarbonyl Group in Solid Phase Synthesis," *Journal of Peptide Science,* vol. 9, pp. 545-552, 2003.

[8] K. &. S. R. &. S. M. Unger, "Particle Packed Columns and Monolithic Columns in High-Performance Liquid Chromatography-Comparison and Critical Appraisal," *Journal of Chromatography,* vol. 1184, pp. 393-415, 2008.

[9] "Analytical HPLC Systems 1260 Infinity II LC System," Agilent, [Online]. Available: https://www.agilent.com/en/product/liquid-chromatography/hplc-systems/analytical-hplc-systems/1260-infinity-ii-lc-system. [Accessed 11 08 2020].

[10] "ACCQPrep® HP150," Teledyne ISCO, [Online]. Available: https://www.teledyneisco.com/en-us/accqprep. [Accessed 11 08 2020].

[11] S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," University of Wisconsin–Madison. Department of Statistics, 2018.

[12] R. G. N. H. L. M. S. D. Robbin Bouwmeester, "DeepLC can predict retention times for peptides that carry as-yet unseen modifications," *Unpublished,* 2020.

[13] Y. R. J. Y. Z. R. H. Y. a. S. L. Chunwei Ma, "Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning," *Analytical Chemistry,* pp. 10881-10888, 2018.

[14] S. L. W. C. Y. Z. P. Y. X. L. Wenyuan Lu, "Locus-specific Retention Predictor," *Nature Scientific Reports,* vol. 7, 2017.

[15] S. A. F. J. e. a. Moruz L, "Chromatographic retention time prediction for posttranslationally modified peptides," *Proteomics,* vol. 12, pp. 1151-1159, 2012.

[16] C. Knight, "Paper Chromatography of Some Lower Peptides," *Journal of Biological Chemistry,* 1951.

[17] L. J. K. B. Y. M. E. M. E. F. S. Konstantinos Petritis, "Improved Peptide Elution Time Prediction for Reversed-Phase Liquid Chromatography by Incorporating Peptide Sequence Information," *Analytical Chemistry,* vol. 78, 2006.

[18] C. G. E. B. J. R. M.-B. Xavier Domingo-Almenara, "The METLIN small molecule dataset for machinelearning-based retention time prediction," *Nature Communications,* vol. 10, 2019.

[19] L. Yang, "Product Purity Prediction and Anomaly Detection for an Automated Peptide Manufacturing Platform," Unpublished Master's Thesis, MIT, Cambridge, MA, 2020.

[20] A. Campbell, "Machine Vision System for In-Process Inspection on an Automated Peptide Manufacturing Platform," Unpublished Master's Thesis, MIT, Cambridge, MA, 2020.

[21] K. J. a. D. RF, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology,* vol. 157, pp. 105-132, 1982.

[22] K. Guruprasad, "Correlation between stability of a protein and its dipeptide composition: A novel approach for predicting in vivo stability of a protein from its primary sequence," *Protein Engineering,* vol. 4, pp. 155-161, 1991.

[23] J. Zimmerman, "The Characterization of Amino Acid Sequences in Proteins by Statistical Methods," *Journal of Theoretical Biology,* vol. 21, pp. 170-201, 1968.

[24] S. S. K. &. N. V. G. Indraneel Majumdar, "PALSSE: A program to delineate linear secondary structural elements from protein structures," *BMC Bioinformatics,* vol. 6, no. 202, 2005.

[25] R. d. .. Milton, "Predicton of difficult sequences in solid-phase peptide synthesis," *J. Am. Chem. Soc.,* vol. 112, no. 16, pp. 6039-6046, 1990.

[26] "NIST/SEMATECH e-Handbook of Statistical Methods," Updated April 2012. [Online]. Available: https://doi.org/10.18434/M32189. [Accessed 29 06 2020].

[27] B. M. Thomas Hamelryck, "PDB file parser and structure class implemented in Python," *Bioinformatics,* vol. 19, no. 17,22, pp. 2308-2310, 2003.