Predict Time of Death of Rats with Artificial Intelligent through Muscle Tissue Examination

Anders Jespersen

January 2026

Abstract

The time of death of 50 rats were predicted using concentrations of 6256 molecules. These were collected with Liquid Chromatography-Mass Spectrometry (LCMS) from muscle tissue. Data were first transformed into logarithmic (x +1), which increased the variation that is explained by time of death. By filtering data with high correlation between replicated samples with Person Correlation Coefficient (PCC) and high Signal to Noise Ratios (SNR) of Quality Control (QC) samples, the noise of the data was reduced. Least Absolute Shrinkage and Selection Operator (LASSO) regression was used to develop a linear model that predicts the time of death in rats. Using the 20 most informative coefficients, the model was optimized for simplicity. This approach not only enhances interpretability but also aligns perfectly with the practical needs of forensic investigations. The final model had a high accuracy that, on average, predicts approximately 0.09 hours away from the actual time of death.

Contents

1 Introduction

The postmortem interval (PMI) is the time since death. The Estimation of PMI is critically important in human forensic work. By estimating PMI, it becomes feasible to narrow down the time frame in which the individual likely died and was placed at the location where the body was discovered (Simmons (2017)). Furthermore, once the PMI is determined, investigators can assess the alibis of persons of interest, either confirming or ruling them out based on verified whereabouts during the relevant period.

Estimating PMI can be challenging in many ways. Following death, the body undergoes a series of inevitable, irreversible, and sequential physical and chemical transformations. Although the order of these changes remains relatively constant, their pace varies considerably due to a multitude of circumstantial and environmental influences (Brooks (2016)). Today's estimate of PMI is highly dependent on investigating visual factors of the body and entomological evidence (Simmons (2017)). This means that these methods are dependent on ambient temperature, body temperature, and other factors. Changes in these factors can influence the degradation time of the body, and thereby it can be a challenge to determine PMI with these methods.

In recent years, high computational power has been available, making artificial intelligence (AI) a possible tool for forensic work (Sharma et al. (2022)). Machine Learning is a specific subset of AI that addresses this ability and enables machines and their software to learn from experience. This development in high computational machine learning techniques makes it possible to produce models where biochemical biomarker levels can predict PMI (Liu et al. (2020) & Usumoto et al. (2010)).

Earlier investigations relied on techniques such as ELISA, SDS-PAGE, and Western blotting. However, the advent of mass spectrometry (MS) in proteomics has revolutionised the field, offering both qualitative and quantitative analysis of proteins. This technology facilitates reliable protein identification and peptide sequencing (Xie et al. (2011)). An example of this is that MS data on the degradation kinetics of porcine skeletal muscle proteins have been estimated and compared with Western blotting results (Battistini et al. (2023)). In that study, the MS correlation showed a better correlation with PMI compared to Western blotting.

In this project, I will use high-dimensional Liquid Chromatography-Mass Spectrometry (LCMS) data derived from muscle tissue from rats to predict Postmortem Interval (PMI). LCMS data are instrumental in identifying and quantifying metabolites present in tissue samples (Schmid (1990)). It is hypothesized that variations in the concentrations of these metabolites over time can serve as biomarkers for PMI (Sangwan et al. (2021)).

To develop a predictive model for PMI using these LCMS data, I will employ a regularized regression approach to construct a straightforward linear model. Specifically, I will use Least Absolute Shrinkage and Selection Operator (LASSO) Regression. This technique minimizes the Residual Sum of Squares (RSS), which measures the difference between observed and predicted values, by adding a regularization term that penalizes the absolute size of the coefficients.

LASSO Regression offers an advantage over traditional least squares regression by increasing the bias of the model reducing the variance in predictions for new, unseen data. It achieves this by shrinking the coefficients of less significant features towards zero, effectively selecting only the most critical variables, and simplifying the model. LASSO regression is particularly effective for handling high dimensional data with few samples (Usumoto et al. (2010) & Vasquez et al. (2016)). To further enhance the simplicity of the model, an additional constraint was applied to limit the model to a maximum of 20 coefficients. This constraint not only simplifies the model but also aligns with the practical needs of forensic analysis, where simpler, more interpretable models are preferable. By integrating these techniques, the resulting model will be optimized for both predictive performance and interpretability, making it suitable for application in forensic contexts.

2 Materials & Methods

2.1 Data Collection

The data collection was done by Institut for Husdyr- og Veterinærvidenskab, Aarhus University (Denmark) in 2022. The rats (Sprague Dawley) weighed between 248g and 300g. A total of 50 rats were examined for muscle tissue. The rats were strangled with CO2, and molecule concentrations were measured according to

five different PMI groups. The first group was measured right after death, and the last group was measured approx. 95 hours after death.

Samples underwent an extraction process before they were analyzed using Ultra-High Performance Liquid Chromatography coupled with Quadrupole Time-of-Flight Mass Spectrometry (UHPLC-qTOF-MS). For convenience, these data will be named LCMS data in this report. The extraction process can be seen All samples were kept by -70°C before and after extraction. The data were acquired using both positive and negative electrospray ionization (ESI). Only positive ESI data was used for this project. During analysis, samples were stored at a controlled temperature of 6°C to maintain their stability and prevent degradation. At the beginning of each analytical run, a sequence of specific injections was performed to prepare and validate the system. This sequence included the injection of one blank sample, which contained only the solvent or mobile phase to check for any potential contamination and ensure that there were no interfering substances.

Following this, a system suitability control was injected to verify that the analytical system was functioning correctly and capable of producing reliable results. In Addition, four Quality Control (QC) samples were injected. These QC samples, with known quantities of analytes, were used to monitor the accuracy and consistency of the analytical process.

These initial injections were critical for system validation before proceeding with the analysis of the actual samples. After these preparatory steps, the real sample sequence began. During the run, to ensure ongoing quality control, a QC injection and a blank injection were performed regularly, specifically every 7th to 8th injection. This routine was designed to continuously monitor the the performance of the system and maintain the integrity of the analysis. Furthermore, the analysis incorporated both inter-batch and intra-batch replicates in addition to the QC samples. Inter-batch replicates involved repeated injections of the same sample across different batches or runs to check for consistency between runs. Intra-batch replicas were performed by injecting the same sample multiple times within the same batch or run to ensure precision and consistency within a single sequence of analysis.

This thorough and rigorous approach to quality control, initial validation injections, regular monitoring injections, and replicate injections, helped ensure the reliability and accuracy of the analytical results obtained using the UHPLC-qTOF-MS system.

A total of 6256 molecules were detected in the muscle tissue. For each molecule, there were 13 replicate samples and 8 QC samples. Extraction process can be found in in appendix (Appendix A, 'Extraction Process'). Raw datafile can be found in appendix (Appendix B, 'PMI Muscle Tissue').

2.2 Data Analyses

All data analysis was performed with R. Key R packages were 'tidyverse', 'dplyr', 'ggplot2', 'caret', 'glmnet' and 'umap'. Markdown of the analysis can be found in appendix (Appendix C, 'PMI Predict').

To address the inherent errors in the LCMS data, several noise reduction techniques were implemented. The workflow (Figure 1) of the analyses was to first normalise the data. It was scaled with a lot of outliers. This was done by transforming the data with the natural logarithm and adding 1 (log (x + 1)).

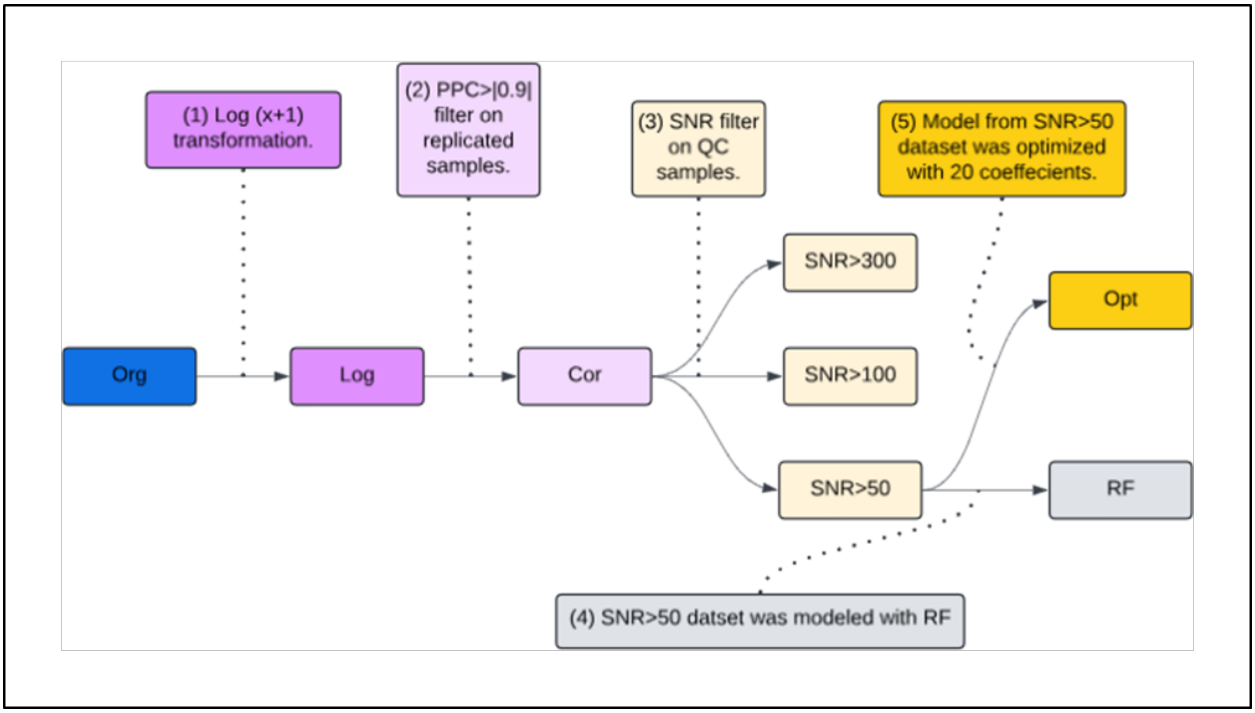

Figure 1: **Flowchart of the project**. (1) First the original data (Org) is transformed with log(x+1). (2) The log transformed dataset (Log) was then filtered with Pearson Correlation Coefficient PCC> |0.9| of replicated samples. (3) The filtered dataset (Cor) underwent additional filters with Signal to Noise Ratio SNR> 50, SNR> 100 and SNR> 300 of Quality Control (QC) samples. All six datasets were modelled with Least Absolute Shrinkage and Selection Operator (LASSO) Regression. SNR> 50 is the dataset that can be modelled with highest prediction performance. (4) The model for data with SNR> 50 was built using Random Forest (RF). When comparing the predictive performance of this non-linear model to the linear model created with LASSO regression, it was evident that a linear approach is suitable and effective. (5) The LASSO regression model of SNR> 50 was maximized with 20 coefficients to increase the interpretability.

Data were then filtered by Pearson Correlation Coefficient (PCC) of replicated samples. High-quality molecule concentration measurements are reflected by high PCC values in their replicates. Additionally, the Signal to Noise Ratio (SNR) of quality control (QC) samples was evaluated to pinpoint noise within the dataset. Molecule concentrations exhibiting a low SNR are indicative of measurements contaminated by noise.

To assess the effectiveness of noise reduction techniques, Uniform Manifold Approximation and Projection (UMAP) was used for dimensionality reduction (McInnes and John (2020)). By visualising the UMAP projections, I examined whether the samples were separated according to PMI.

After noise reduction a total of six different datasets (table 2.2.4) were modelled with LASSO regression. The model with best performance prediction was based on SNR> 50 data. This dataset is log-transformed filtered by PCC> |0.9| and SNR> 50. This dataset was then remodelled with Random Forest (RF), to conclude that a linear model is more suitable and effective than a non-linear model (Ali et al. (2012)). The SNR> 50 was then optimised for 20 coefficients to build the final model. In this section, a detailed description of the workflow is given below.

2.2.1 Normalization of data

Since data was scaled, the data were transformed with log (x+1) for all data. The plus one was added to avoid concentrations that are equal to zero from becoming negative values after the transformation. This was done to stabilise the variance between the molecule concentrations to make them more consistent and easier to interpret. To check if the distribution was normalised after the transformation, the distribution

before and after transformations was compared in histograms that included all the molecule concentration data points.

2.2.2 Person correlation coefficients

Molecule concentrations of replicated samples were compiled into a single dataset. The Person correlation coefficient (PCC) was then calculated for each molecule with equation 1:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

Where n is replicated samples, x is the first replicate and y is the second replicate. Theoretical, replicate x_i and y_i should be identical, if the measures were perfect. This will result in a PCC value of 1. For this reason, selecting molecules of a high correlation coefficient noise reduction in the dataset was made. $PPC > |0.9|$ molecules were chosen as a threshold and are the molecules in the Cor dataset (see workflow in figure 1).

2.2.3 Signal to noise ratio

Signal to Noise Ratio (SNR) were calculated between all Quality Control (QC) samples for each molecule. The SNR was calculated for each molecule with equation 2:

$$\text{SNR} = \frac{\text{signal}}{\text{noise}} = \frac{\frac{1}{n}\sum_{j=1}^{n} X_{ij}}{\sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(X_{ij} - \overline{X}_i\right)^2}} \tag{2}$$

It is expected that molecules of high quality have identical QC values. Therefore, signal is the average of all QC samples of one molecule, and noise is the standard deviation of all QC samples of the same molecule. Molecules with a lot of random variations between QC samples have low SNR value. Overall, molecules with low SNR values have a lot of noise, and these molecules are preferred to be avoided further in the analysis. Three different SNR thresholds were made. These are presented in dataset SNR> 50, SNR> 100 and SNR> 300 (see workflow in figure 1).

2.2.4 Datasets

A total of six different datasets were modelled (Table 2.2.4). Reduction of the number of features is due to the attempt at noise reduction. The trade off of having datasets with few features is that even though they have very little noise, informative molecules may have been selected out in the process of noise reduction. For this reason, serval datasets were kept further in the analysis. Notion that datasets Cor, SNR> 50, SNR> 100 and SNR> 300 are all log- transformed data that is filtered by $PPC > |0.9|$.

Dataset	Number of features
Org	6256
Log	6256
Cor	2729
SNR> 50	2030
SNR> 100	1046
SNR> 300	136

Table 1: **Datasets and number of features**. Org is the original dataset. Log is log (x+1) transformed dataset. Cor is log (x+1) transformed dataset and filtered by Pearson Correlation Coefficient PCC> |0.9| between replicated samples. SNR> 50, SNR> 100 and SNR> 300 is log (x+1) transformed, filtered by PCC> |0.9| and filtered by different threshold of Signal to Noise Ratio (SNR) of quality control (QC) samples.

Datasets were then visualized with dimensions reductions projections made by UMAP (McInnes and John (2020)). UMAP is a non-linear dimension reduction that both preserves local and global structure in high-dimensional data and efficiently captures complex relationships among data points, making it particularly well-suited for visualizing and exploring large and intricate datasets. The tuning parameter for UMAP was 2 number of components, 15 nearest neighbours were used with Euclidean distance, 200 epochs, a minimum of 0.1 in low-dimensional embedding and with Spectral initialization. UMAP1 represents the first dimension, that captures the most variation within the dataset was plotted against UMAP2 in a scatterplot. UMAP2 represents the second dimension that captures additional variation, which is not explained in UMAP1. Samples were coloured according to PMI to see how UMAP separates samples according to PMI.

2.2.5 LASSO regression

LASSO Regression modelled all six different datasets this was done in three steps (see steps in Figure 2).

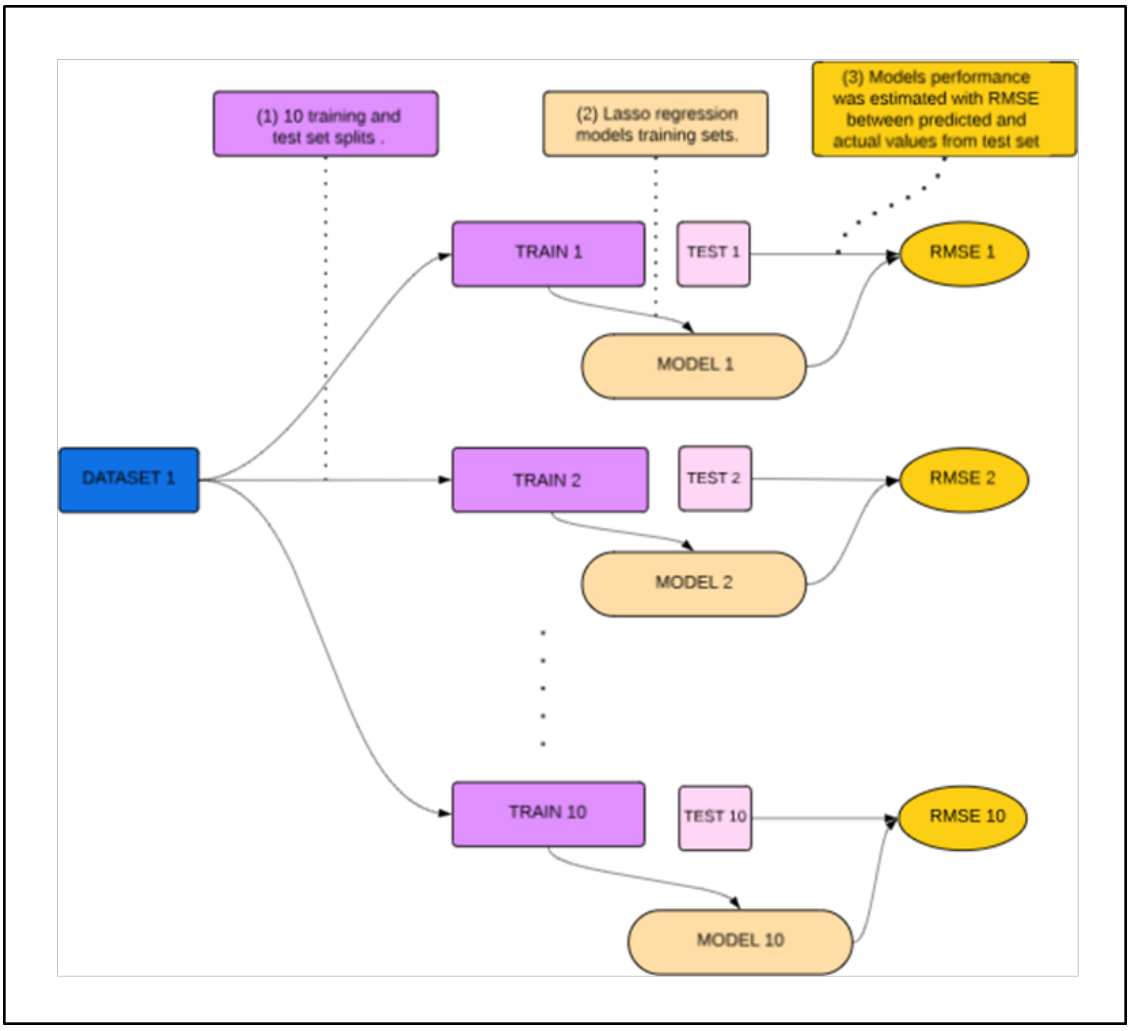

Figure 2: Flow chart of training and testing LASSO regression models. (1) Each Dataset was split into 10 different training and test sets. (2) Training sets were modeled with Least Absolute Shrinkage and Selection Operator (LASSO) Regression. The best shrinkage parameter (λ) was estimated with 10-fold repeated cross validation with 3 repeats. (3) The performance of each model was esti- mated with Root Mean Square Error (RMSE) which is the squared difference between predicted values from the model, and actual values of the test set.

1. The samples in every dataset were randomly split into a training and a test set split. 80% (42 samples) of the data was given to a training set, and 20% (8 samples) was given to the test set. The test set was later used in the test phase, to see the performance of each model. This was done with 10 different seeds, which means the random 80/20 split was done 10 times. In the end a total of 60 training sets and 60 test sets were created for all the datasets combined. The high number of seeds for each split was done to ensure more stable results of the performance of models created by each dataset.

2. LASSO regression created models to predict PMI for all 60 training sets. LASSO regression minimizes equation 3:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{3}$$

   In this equation is n is number of samples, p is number of features, β is coefficients of the regression line and λ is the shrinkage parameter. The first term of this equation is residuals sum of square (RSS). When minimizing RSS, the best linear relation between the PMI and molecule concentration is found in the training set. However, this relationship often overfits when there are relative few samples compared with number of features when it comes to predictions of the test set.

   The Second term of the equation is the regularization term, and λ is the shrinkage parameter in the equation. By increasing the value of λ, the minimization of the cost function allows coefficient to shrink all the way down to 0. A higher value of λ adds more bias to the final model, which can result in less variation be- tween PMI of the test set, and the predicted values. A too high λ value will lead to underfit.

3. The best λ was estimated for each model by a 10-fold repeated cross validation that divides the training set into 10 random folds. This was done with 3 repeats which means that the random split between the folds was repeated 3 times. This results in a total of 30 evaluations for each λ value. The best λ value was estimated by 100 different values between 0 and 0.21[1]. The best λ value of each of the training sets, is the λ that has the lowest validation error across all 30 evaluations. This validation error is Root Mean Square Error (RMSE), that is calculated with equation 4:

$$\text{RMSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4}$$

   In this equation is n number of samples, y is the actual value of PMI of the validation set, and $\hat{y}$ is the predicted value of PMI. Once the optimal λ was identified for each of the 60 models, their performance was evaluated also using the RMSE. RMSE was selected as the error metric because it provides a straightforward interpretation by maintaining the same units as the actual PMI values in the test set. This resulted in a data frame that contains 10 different RMSE values of all datasets.

   These RMSE of every dataset are the result of 10 models that have 10 different training and test set splits, with 10 different λ values. To evaluate the performance of models from each dataset, the pairwise significance between RMSEs values was estimated with the Wilcoxon Rank-Sum Test (also known as the Mann-Whitney U test). This is done by first calculating the differences between pairwise RMSE values of two datasets with equation 5:

$$D_i = x_i - y_i \tag{5}$$

   Where x is the RMSE values of the first dataset and y is the RMSE values of the second dataset. The sum of positive and negative ranks is then determined with equation 6 and equation 7:

$$T^+ = \sum_{i=1}^{n} R_i \quad \text{(sum of ranks of positive differences)} \tag{6}$$

[1] For the Org data, the best λ was found by evaluating 1000 different numbers in an interval between 0 and 3.

$$T^- = \sum_{i=1}^{n} R_i \quad \text{(sum of ranks of negative differences)} \tag{7}$$

Here is R_i in equation 6 is the $|D_i|$ if all positive D_i, and R_i in equation 7 is the $|D_i|$ for all negative D_i. The p-value is then derived from the distribution of the test statistic under the null hypothesis, which posits no difference between the RMSE values of two different datasets. It quantifies the likelihood of observing a test statistic as extreme as the minimum of either T^+ or T^- in that distribution.

It is chosen that significant level (α-level) of 0.5 results in significant differences in performance between models under each dataset. The dataset that potentially creates the best model to predict PMI, was then determined by the lowest average RMSE between the training and test splits of all six different datasets. The significant differences between RMSE values of each dataset were also considered.

2.2.6 Random Forrest

LASSO regression was made for the dataset that most accurate a predicting PMI which was SNR$>$ 50. As before the dataset was split into an 80/20 training and test set with 10 different seeds, and best λ was estimated by 10-fold repeated cross validation with 3 repeats. The performance of the models was estimated by RMSE and compared with RMSEs of Random Forest (RF) models that were created under the same dataset. RF is a tree based non-linear method in contrast with LASSO Regression creates a simple linear model (Ali et al. (2012)). This was done to see if a linear relationship is suitable for predicting PMI. Or if a non-linear model will yield significantly better predictions. The RF model was created with 500 trees and the number randomly chosen features that was selected to determine the split of each node was the total number of features divided by 3 to avoid overfitting. To see if there was a significant difference of performance between LASSO regression and RF, Wilcoxon Rank-Sum Test was utilized on RMSEs values of the two methods with a α-level of 0.05.

2.2.7 Optimize for 20 molecules

Lastly SNR$>$ 50 was then optimized for 20 coefficients, which means that this model also only considers 19 different molecule concentrations since the intercept is also a coefficient in the linear model (β_0). This was done by first finding the 20 most informative coefficients of the model. Then LASSO Regression modelled this new dataset that only contains the 20 most informative coefficients. This was again done with the dataset split into an 80/20 training and test set split with 10 different seeds. The best λ was estimated by 10-fold repeated cross validation with 3 repeats. The performance of the new optimized model was estimated by RMSE and compared with RMSEs of model that was not optimized. The Wilcoxon Rank-Sum Test was used on RMSEs values of the two models with a α-level of 0.5, to see if there is a significant difference in performance between the two models.

3 Result & Discussion

3.1 Data transformation

The data of PMI and all features was combined into one single vector. 312,850 molecule concentrations were plotted against frequence in the histogram to see if data looks normally distributed (figure 3). In terms of visualization of the original dataset, not all the molecule concentrations were included, since some molecule concentrations had high molecule concentrations $> 4 \times 105$. 18.923 molecule concentrations (6% of all data) were not included in the plot with maximum values $> 7 \times 107$ (15 values).

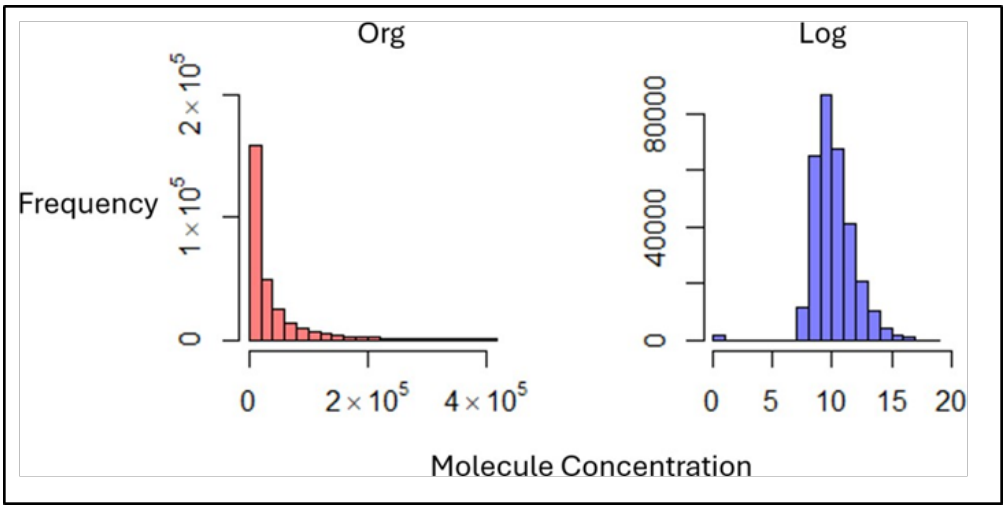

Figure 3: **Histogram of molecule concentrations measured with Liquid Chromatography Mass Spectrometry (LCMS)** Concentrations and frequency of the original dataset (right). Concentrations and frequency of the log (x +1) transformed dataset (left).

The histogram of the original dataset shows that values are scaled and screwed to the right. The values were then transformed with a log (x + 1) transformation and shown on the histogram to the right. Data gets more normalized with a peak in the middle around the average molecule concentration value of 10.1. This transformation gave rise to a more predictive accuracy of PMI and is demonstrated by looking at UMAP plot which contains the original data (Org) and plot B which contains the transformed data (Log) (figure 4, A).

Here UMAP1 contains most of the variation in the dataset, and UMAP2 explains a lot of the variation that UMAP1 does not. It is seen that either UMAP1 or UMAP2 can separate the samples according to PMI in plot A. However, after the transformation, UMAP1 clearly separates PMI values of approximate 0 from the rest of the samples in plot B. Also, UMAP2 separates most of the samples according to PMI. Overall, this means more of the variation in the dataset can be explained by PMI when data log transformed.

3.2 Person Correlation Coefficients

A total of 2699 molecule concentrations had a PCC> |0.9|. Out of 6255 molecules 3556 molecules have been selected out of the dataset, since it is believed that these molecules contain a lot of noise. This noise reduction can be demonstrated by comparing the UMAP projections of dataset of the transformed data (Log) (figure 4, B) and the transformed data and filtered by a PCC> |0.9| (Cor) (figure 4, C). In contrast to plot B, UMAP1 in plot C is not only able to separate PMI values of approximate 0 from the rest of the samples but can also separate other samples according to PMI to some extent. This is an improvement in how big a proportion of the variation in the Cor dataset can be explained by PMI compared with the Log dataset.

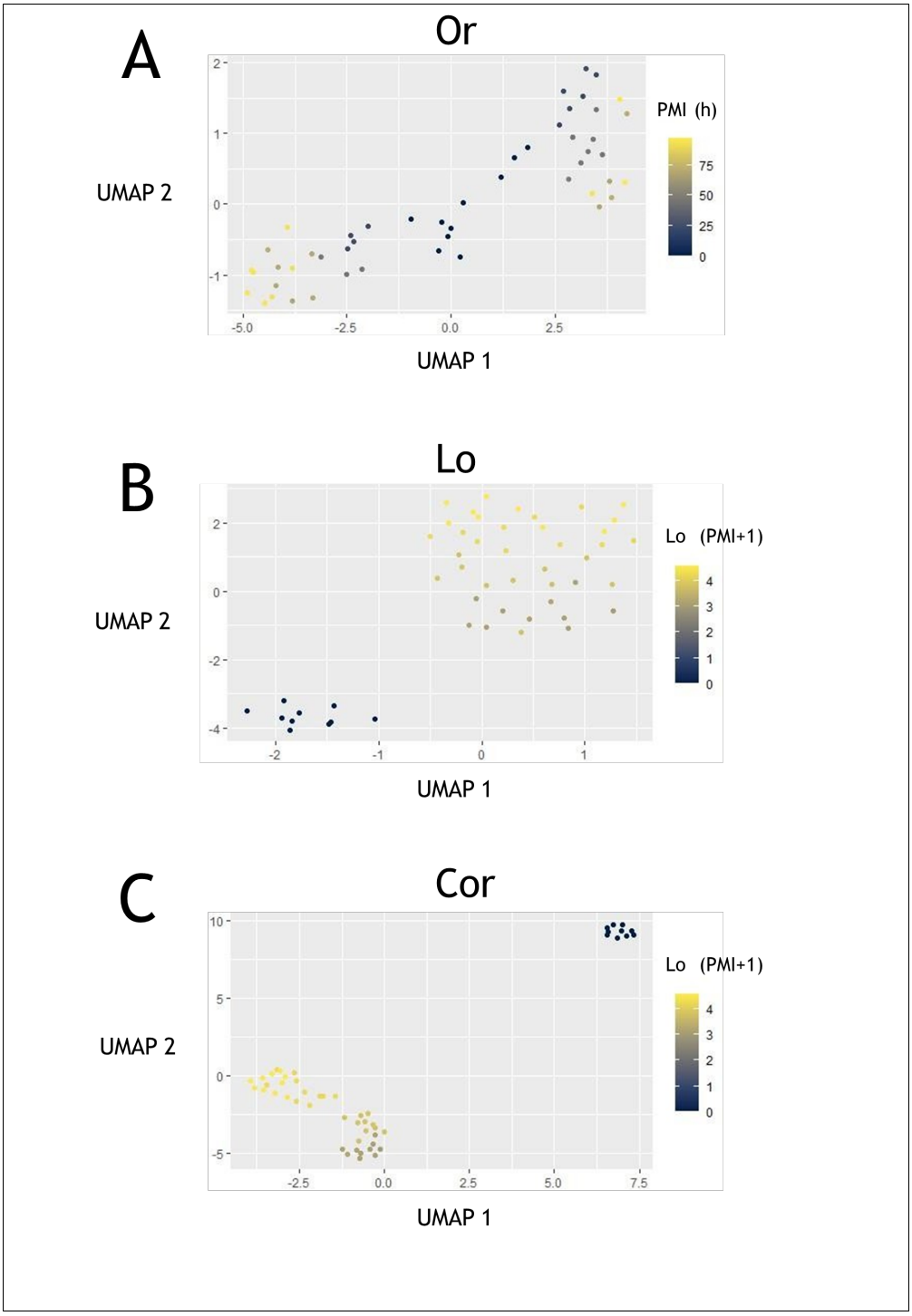

Figure 4: **Scatter plots Uniform Manifold** Approximation and Projection (UMAP). UMAP 1 contains the most variation within the dataset and UMAP 2 explains variation that UMAP 1 does not. Samples are coloured after Postmortem interval (PMI). (A) original dataset. (B) Log (x+1) transformed dataset. (C) Log (x+1) transformed and filtered by Person Correlation Coefficient $> |0.9|$ between replicated samples.

3.3 Signal to Noise Ratio

Low SNR values of the QC-samples are molecule concentrations of bad qualities. Essentially, a lower SNR indicates that the concentration of molecules is accompanied by a higher level of noise. In other words, samples with low SNR values have more noise compared to samples with high SNR values, which typically have clearer and more distinct molecular concentrations. To illustrate this, a dataset with samples that contain molecules concentration with a SNR< 30 was plotted with UMAP (figure 5). This dataset contains 137 molecules concentrations.

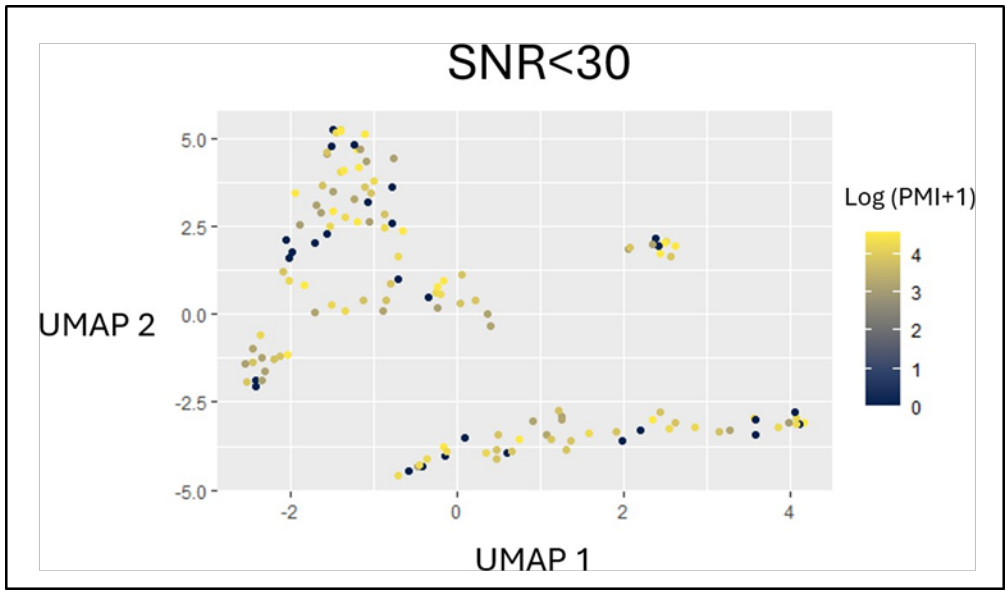

Figure 5: **Scatter plot of Uniform Manifold Approximation** and Projection (UMAP). UMAP 1 contains the most variation within the dataset and UMAP 2 explains variation that UMAP 1 does not. Samples are coloured after Postmortem interval (PMI). Data is log (x+1) transformed, filtered by Person Correlation Coefficient > |0.9| between replicated samples and filtered by Signal to Noise Ratios SNR< 30 of Quality Control samples.

In this plot either UMAP1 or UMAP2 can separate samples according to PMI, and it implies that not much of the data that have a SNR¡30 can be explained by PMI.

Three different datasets were created three different SNR filters. The first dataset (SNR> 50) was transformed data filtered by PCC> |0.9| and SNR> 50. This dataset contained 2030 molecule concentrations. The second (SNR> 100) and third dataset (SNR> 300) was also transformed and filtered by PCC> |0.9| and contained 1046 and 136 molecule concentrations respectively. A stricter SNR filter will reduce the noise in the dataset. However, there is also a risk of in- formative molecule concentrations might be filtered out, if the filter is too strict. There is no golden standard of what a good SNR filter for QC samples is. Therefore multiple datasets with multiple SNR filters were considered. To see how well the variation within the datasets with SNR filters explains PMI, UMAPs plots for all three datasets were constructed (figure 6).

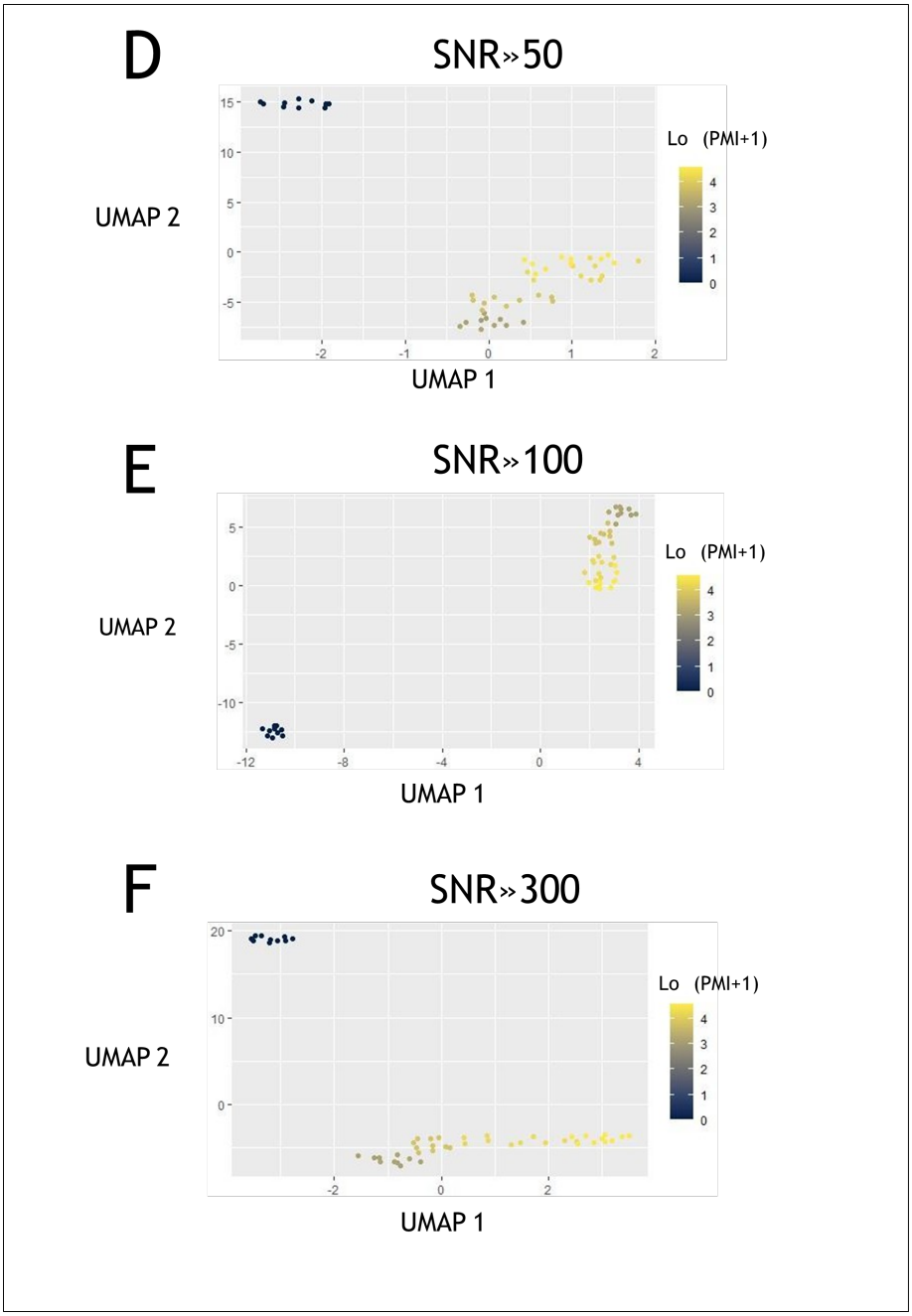

Figure 6: **Scatter plots Uniform Manifold Approximation and Projection (UMAP)** UMAP 1 contains the most variation within the dataset and UMAP 2 explains most of the variation that UMAP 1 does not. Samples are colored after Postmortem interval (PMI). (D) Log (x+1) transformed data, filtered by Person Correlation Coefficient PCC> |0.9| between replicated samples and filtered by Signal to Noise Ratio SNR> 50 for quality control (QC) samples. (E) Log (x+1) transformed, filtered by PCC> |0.9| and SNR> 100. (F) Log (x+1) transformed, filtered by PCC> |0.9| and SNR> 300.

In all three plots UMAP1 and UMAP2 separates data according to PMI to some extent. It can then be concluded that by applying SNR filters to the datasets, variation within the datasets can still be explained by PMI.

However, it is unclear whether a stricter SNR filter (SNR> 300, plot F) has significant more variation that can be explained by PMI compared with less strict SNR filter (SNR> 100 and SNR> 50, plot D and

plot E). Also, what cannot be interpreted in these plots, is that informative molecule concentrations have been filtered out during the process. For this reason, all three datasets are considered during the modelling.

3.4 LASSO regression

The linear Regularized Regression method Lasso Regression was applied to each of the six datasets with 10 different training and test split for each dataset. This resulted in 10 RMSE and 10 λ values for each dataset and mean RMSE value (table 2). RSME of the Org models was log + 1 transformed in order make values comparable to the models.

Table 2: **Root Mean Squared Error (RMSE) measured for different Least Absolute Shrinkage and Selection Operator (LASSO) models.** (except SNR> 50 (RF) which is modeled with Random Forrest (RF)). Seeds are different training and test splits. Org is original data. Log, Cor, SNR> 50, SNR> 100, SNR> 300 and SNR> 50 (Opt) is log (x+1) transformed data. Cor, SNR> 50, SNR> 100, SNR> 300 and SNR> 50 (Opt) was filtered by Person Correlation Coefficient > |0.9| between replicated samples. SNR> 50, SNR> 100, SNR> 300 and SNR> 50 (Opt) was filtered by Signal to Noise Ratios of SNR> 50, SNR> 100 and SNR> 300 respectively of Quality Control samples. SNR> 50 (Opt) was optimized for 20 coefficients.

	Seed 1	**Seed 2**	**Seed 3**	**Seed 4**	**Seed 5**	**Seed 6**	**Seed 7**	**Seed 8**	**Seed 9**	**Seed 10**	**Avg. RMSE**
Org	1.9678	2.0931	2.1285	2.0483	2.0979	2.1427	1.9270	1.7815	2.0622	1.8096	**2.0059**
Log	0.1282	0.1585	0.1981	0.1576	0.0996	0.1545	0.1850	0.1979	0.1051	0.1583	**0.1543**
Cor	0.1666	0.1536	0.1190	0.1092	0.1528	0.1143	0.1280	0.0633	0.1243	0.1180	**0.1249**
SNR> 50	0.0673	0.0792	0.1432	0.1565	0.1262	0.0767	0.1029	0.1541	0.0609	0.1612	**0.1128**
SNR> 100	0.1008	0.0959	0.1101	0.1445	0.0853	0.0748	0.1355	0.1463	0.1287	0.1144	**0.1136**
SNR> 300	0.1088	0.1361	0.1059	0.1698	0.1813	0.1174	0.1889	0.1343	0.0902	0.1756	**0.1408**
SNR> 50 **(RF)**	0.1089	0.1486	0.1015	0.2273	0.1901	0.1340	0.2077	0.0852	0.0990	0.2437	**0.1546**
SNR> 50 **(Opt)**	0.0746	0.0513	0.0539	0.0929	0.1172	0.0438	0.1199	0.0832	0.0718	0.1200	**0.0829**

The Wilcoxon rank-sum test was utilized to measure significant difference of RMSE between the models (table 3). The Log dataset has an average RMSE of 0.1543 and the models constructed from the Org dataset have an average RMSE of 2.0059 (table 2). The difference between these models is significant with a p-value of 0.0020 (table 3). This means models constructed under the Log dataset are better at predicting PMI compared with models constructed under the Org dataset. It can also be concluded that datasets filtered with PCC> |0.9| and a SNR above a certain threshold also results in improved models compared with the model constructed under the Org dataset since all these differences are significant.

Table 3: **P-values between model performance that is created by Least Absolute Shrinkage and Selection Operator (Lasso) regression.** Measured with the Wilcoxon rank-sum test. $*$ Indicates significant differences. Org is original data. Log, Cor, SNR> 50, SNR> 100, SNR> 300 and SNR> 50 (Opt) is log (x+1) transformed data. Cor, SNR> 50, SNR> 100, SNR> 300 and SNR> 50 (Opt) was filtered by Person Correlation Coefficient > |0.9| between replicated samples. SNR> 50, SNR> 100, SNR> 300 and SNR> 50 (Opt) was filtered by Signal to Noise Ratios of SNR> 50, SNR> 100 and SNR> 300 respectively of Quality Control samples. SNR> 50 (Opt) was optimized for 20 coefficients.

	Org	**Log**	**Cor**	**SNR**> 50	**SNR**> 100	**SNR**> 300	**SNR**> 50 **(RF)**
Org							
Log	0.0020*						
Cor	0.0020*	0.1309					
SNR> 50	0.0020*	0.0195*	0.5566				
SNR> 100	0.0020*	0.0098*	0.4922	0.9219			
SNR> 300	0.0020*	0.2754	0.3223	0.0488*	0.0645		
SNR> 50 **(RF)**	0.0020*	0.9219	0.2324	0.0645	0.1055	0.0840	
SNR> 50 **(Opt)**	0.0020*	0.0039*	0.0195*	0.0488*	0.0273*	0.0020*	0.0020*

By filtering log transformed data only with PCC> |0.9| (Cor) did not result in a significant improvement of the models created compared to the Log models. Even though the average RMSE of the Cor models is

less compared with the Log models the difference was non-significant. However, when data is also filtered by SNR> 50 or SNR> 100, the models become significantly improved com- pared with Log models with a p-value of 0.0195 and 0.0098 respectively. However, applying an SNR> 300 filter did not improve model performance. This could be due to the SNR> 300 filter is excluding not only noise but also informative molecular concentrations that is useful to predict PMI. I chose SNR> 50 to be the model I want to optimize. The reason for this is that I only want to optimize a single model, and SNR> 50 has a lower average RMSE compared with SNR> 100. Also, it was the only model that performed significantly better than the SNR> 300 model.

3.5 Reflecting on rather a non-linear model is suitable to predict PMI

The RF algorithm, which is a non-linear tree- based method, was used on the dataset SNR> 50. The performance of the RF model was estimated and compared the model that was created with LASSO Regression. The average RMSE for the RF model is 0.1546 which is higher than the LASSO Regression model, that has an average RMSE of 0.1128 (see table 2). For this reason, there is no indication that a non-linear model will be a better choice compared to a linear model.

3.6 Optimize the model with 20 coefficients

Models created under SNR> 50 dataset had different training and test split had variation in number of coefficients and their values, depending on the seed. I chose the model created under seed 5 since its RMSE values were closest to the mean of all seeds, to be the model I want to optimize. There was a total of 28 informative coefficients in this model. The top 20 most informative coefficients with the highest absolute values are selected to optimize the model (table 4). 10 optimized models (SNR> 50 (Opt)) with 10 different training and test split were computed.

Table 4: **Coefficient names and values of models created by Least Absolute Shrinkage and Selection Operator regression.** 20 most informative coefficient values before optimization are displayed in the middle column. These Coefficients are used to optimize the model. The values of the same coefficients after optimization are displayed in the column to the right.

Molecule name	Coefficient (β) of SNR> 50 before optimization	Coefficient (β) of SNR> 50 after optimization
(intercept)	3.1917176721	3.207356495
M257T213	0.1792286781	0.467830131
M229T59	0.1459009623	0.095819127
M286T223	0.1366735079	0.080728318
M260T343	0.1262467795	-0.114969975
M381T196	0.1200401462	-0.036741966
M229T271	0.1001003784	0.054231764
M450T223	0.0910980497	0.003508463
M409T261	0.0852952989	0.046909470
M247T234	0.0810007106	0.138308256
M124T81	0.0752021469	-0.101950850
M226T222	0.0583478371	0.012598475
M247T243	0.0573875484	0.148749244
M359T53_1	0.0567273984	0
M285T223	0.0566163863	0.109887182
M626T97	0.0520377145	0.178409306
M286T282	0.0476864628	0.010629912
M238T44	0.0206700636	0.010572796
M380T196	0.0206697170	-0.045267317
M248T79	0.0189045879	0

The coefficients values again varies between different training and test splits. Once more I chose the model that had the closet RMSE value to the mean. This was seed 8. This is the final model of the project. In this model there are 18 informative coefficients (table 4) The RMSE of this model is 0.0832. When this number is reversed log (x+1) transformed, it results in an average prediction of approximately 0.09h away from the actual PMI. The average RMSE value of the optimized models is 0.0829, which is less than the

non-optimized model which had an average RMSE value of 0.1128 (table 2). The p-value of the Wilcoxon rank-sum test between RMSE values for the optimized and non-optimized model was estimated to be 0.0488. This implies that the optimized model is a slightly significant improvement compared with the non-optimized model. Thereby a maximum of 20 coefficient of the linear model is not only a simpler model, but it also performs better at predicting PMI.

4 Conclusion

In this project noise reduction of LCMS data was successful by utilizing replicated and QC samples. It was then possible to create a linear model with few coefficients that predicts PMI with LASSO regression. This model had high predictions performance and was easy to interpret. However, the accuracy of the models varies a lot between different and training and test splits. More stable results could have been achieved with more samples. More experimental work with LCMS data from different tissues could lead to the finding of a few biomarkers that can predict PMI, which at some point can be used in forensic work. This project is an example of how models with few biomarkers can be created.

References

J. Ali, R. K., N. A., and I. M. Random forests and decision trees. *IJCSI International Journal of Computer Science*, 9(5), 2012.

A. Battistini, D. Capitanio, P. Bailo, M. Moriggi, S. Tambuzzi, C. Gelfi, and A. Piccinini. Proteomic analysis by mass spectrometry of postmortem muscle protein degradation for pmi estimation: A pilot study. *Forensic Science International*, 349:111774, 2023. doi: 10.1016/j.forsciint.2023.111774.

J. W. Brooks. Postmortem changes in animal carcasses and estimation of the postmortem interval. *Veterinary Pathology*, 53(5):929–940, 2016. doi: 10.1177/0300985816629720.

R. Liu, Y. Gu, M. Shen, H. Li, K. Zhang, Q. Wang, X. Wei, H. Zhang, D. Wu, K. Yu, W. Cai, G. Wang, S. Zhang, Q. Sun, P. Huang, and Z. Wang. Predicting postmortem interval based on microbial community sequences and machine learning algorithms. *Environmental Microbiology*, 22(6):2273–2291, 2020. doi: 10.1111/1462-2920.15000.

L. H. McInnes and James John, Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. Preprint, arXiv:1802.03426.

A. Sangwan, S. P. Singh, P. Singh, O. P. Gupta, A. Manas, and S. Gupta. Role of molecular techniques in pmi estimation: An update. *Journal of Forensic and Legal Medicine*, 83:102251, 2021. doi: 10.1016/j.jflm.2021.102251.

E. R. Schmid. Chromatography and mass spectrometry- an overview. *Chromatographia*, 30(9/10):573–576, 1990.

R. Sharma, Diksha, A. R. Bhute, and B. K. Bastia. Application of artificial intelligence and machine learning technology for the prediction of postmortem interval: A systematic review of preclinical and clinical studies. *Forensic Science International*, 340:111473, 2022. doi: 10.1016/j.forsciint.2022.111473.

T. Simmons. Post-mortem interval estimation: an overview of techniques. In *Taphonomy of Human Remains: Forensic Analysis of the Dead and the Depositional Environment*, pages 134–142. 2017. doi: 10.1002/9781118953358.ch10.

Y. Usumoto, W. Hikiji, N. Sameshima, K. Kudo, A. Tsuji, and N. Ikeda. Estimation of postmortem interval based on the spectrophotometric analysis of postmortem lividity. *Legal Medicine (Tokyo)*, 12(1):19–22, 2010. doi: 10.1016/j.legalmed.2009.09.008.

M. M. Vasquez, C. Hu, D. J. Roe, Z. Chen, M. Halonen, and S. Guerra. Least absolute shrinkage and selection operator type methods for the identification of serum biomarkers of overweight and obesity: simulation and application. *BMC Medical Research Methodology*, 16(1):154, 2016. doi: 10.1186/s12874-016-0254-8.

F. Xie, T. Liu, W. J. Qian, V. A. Petyuk, and R. D. Smith. Liquid chromatography-mass spectrometry-based quantitative proteomics. *Journal of Biological Chemistry*, 286(29):25443–25449, 2011. doi: 10.1074/jbc.R110.199703.