

**Analysis of Raman Spectra With a Machine Learning Approach for Improved  
Quantification of Microcystin-LR**

Honors Research Thesis

Presented in partial fulfillment of the requirements for graduation with *honors distinction* in  
Chemistry in the undergraduate colleges of The Ohio State University

Chase Fensore

The Ohio State University

April 2022

Advisors: Dr. Heather Allen, Departments of Chemistry and Biochemistry and Pathology, Dr.

Xia Ning, Departments of Biomedical Informatics and Computer Science and Engineering

Thesis Committee Members: Dr. Heather Allen, Departments of Chemistry and Biochemistry  
and Pathology, Dr. Xia Ning, Departments of Biomedical Informatics and Computer Science and  
Engineering

## Abstract

Cyanobacterial harmful algal blooms (cyanoHABs) have increased in prevalence in recent years, threatening 0.5% of potable water on earth. Microcystins, a class of toxins that can be produced in a cyanoHAB, are particularly harmful to humans and ecosystems alike. Microcystin-LR (MC-LR) is among the most toxic and the most common microcystins. Though biosensors have recently shown impressive capabilities in detecting microcystins, exhibiting high sensitivity, selectivity, and portability, traditional Raman spectroscopy combined with modern machine learning may still offer a path to sensitive, portable detection of MC-LR. The objective of this project focuses on evaluating the efficacy of three machine learning algorithms at detecting MC-LR in water at concentrations near the EPA's benchmark limit of 1  $\mu\text{g/L}$ . Raman spectra were collected from MC-LR dilutions in water at concentrations ranging from 0.001 to 6.0  $\mu\text{g/L}$ . A sample size of  $n=1000$  Raman spectra was achieved, and spectral preprocessing methods including background subtraction of water, z-score feature normalization, and baseline removal were employed. Regression models to predict MC-LR concentration in water from Raman spectral data were built using three machine learning algorithms: kernel support vector machine for regression (SVR), regression deep neural network (DNN), and partial least squares regression (PLSR). These three models are compared using mean-square-error (MSE) and mean-absolute-error (MAE) to evaluate their efficacy for predicting MC-LR concentrations in the range of 0.001 to 6.0  $\mu\text{g/L}$ . After validation of the models on test data ( $n=200$ ), MSE values were found to increase in the order of PLSR without feature normalization (0.191) < PLSR with feature normalization (0.199) < SVR (0.432) < DNN (3.155).

## **Acknowledgements**

I would like to express gratitude to everyone who supported me in my undergraduate research pursuits and the completion of this undergraduate thesis. I would like to thank Dr. Heather Allen for welcoming me into her chemistry and biochemistry research group when I was a second year computer science student, and for affording me opportunities to work on numerous exciting projects. I would also like to thank Dr. Xia Ning for serving as my co-advisor and oral defense committee member, and for offering guidance in the early stages of this project.

I would like to thank Abigail Enders and Nicole North for their outstanding guidance, mentorship, and patience throughout my time in the Allen Lab.

Finally, I would also like to thank my parents, Alex and Susan, and my partner Bhavya for each of their support of my undergraduate study and my research goals.

## Table of Contents

<b>Abstract</b>	<b>1</b>
<b>Acknowledgements</b>	<b>2</b>
<b>Introduction</b>	<b>4</b>
<b>Methods</b>	<b>7</b>
<b>Results and Discussion</b>	<b>13</b>
<b>Conclusion</b>	<b>31</b>
<b>References</b>	<b>32</b>

## Introduction

Water safety monitoring is an area of increasing urgency throughout the world. In the United States, 68% of people get their drinking water from public water systems which are supplied by surface water sources such as rivers and lakes.<sup>1</sup> In response to climate change, cyanobacterial harmful algal blooms (cyanoHABs) have increased in prevalence, and the toxins produced by the algae pose a threat to the 0.5% of potable water on earth.<sup>2,3</sup> Microcystins (MCs) are a diverse class of hepatotoxins produced by these cyanoHABs; over 90 MCs have been identified, and microcystin-LR (MC-LR) is the most prevalent and among the most toxic.<sup>4</sup>

Monitoring concentration of MC in water is paramount to public health and safety. Detection is a difficult task because the safe concentration threshold is low and over 90 variants of MCs exist. In 2019, the EPA recommended that ambient water swimming advisories be issued when MC-LR concentration reaches 8  $\mu\text{g/L}$ .<sup>5</sup> In drinking water, the maximum accepted MC-LR concentration is 1  $\mu\text{g/L}$  according to the World Health Organization (WHO) and 1.6  $\mu\text{g/L}$  for adults, according to the EPA.<sup>5,6</sup>

Biosensors have recently shown impressive capabilities in detecting MCs, exhibiting high sensitivity, selectivity, and portability.<sup>7</sup> For example, in 2019 a surface-enhanced Raman scattering (SERS) spectroscopic immunosensor was able to achieve a limit of detection of 0.014  $\mu\text{g/L}$  and a linear dynamic detection range of 0.01 to 100  $\mu\text{g/L}$  with respect to MC-LR, orders of magnitude better than the commercial enzyme-linked immunosorbent assay (ELISA) test.<sup>8</sup> However, biosensors are still in development and require testing in various field settings.<sup>7</sup>

As an alternative to the recent successes of biosensors, traditional Raman spectroscopy may still offer a path to sensitive detection of MC-LR. Raman spectroscopy is based on a phenomenon called Raman scattering (inelastic scattering of photons). A laser with a fixed

wavelength (which is not absorbed by the sample) is directed at the sample causing the reflected photons to increase or decrease in energy. This energy shift is measured by a detector after filtering reflected light, which gives information about the vibrational modes of the sample.<sup>9</sup> In Raman spectroscopy the phenomenon of most interest is Stokes Raman scattering, where reflected photons increase their vibrational energy state ( $\Delta\nu > 0$ ) through a virtual state transition. In other words, Raman spectroscopy provides a direct measure of the vibrational energies of the substance being sampled, providing a unique chemical fingerprint in the form of a spectrum. Further, Raman-based methods overall offer distinct advantages over other types of vibrational spectroscopy, such as traditional IR spectroscopy, because water has weak Raman-scattering, allowing samples to be directly analyzed in aqueous form. Raman scattering is typically very weak compared to Rayleigh scattering, but the use of edge or notch filters for laser rejection can enhance the detection of inelastically scattered photons. Other techniques to enhance detection of Raman scattering have been employed, such as SERS and drop coating deposition Raman (DCDR), both of which are more sensitive than traditional Raman spectroscopy, but both require more complex sample preparation, which is not practical for field monitoring of MCs.<sup>6,8</sup> Still, traditional Raman spectroscopy is most often associated with determination of molecular structure and qualitative analysis, and is not typically associated with quantitative analysis reaching low detection limits because weak inelastic scattering of laser photons may only result in very subtle changes in spectra, which elementary fitting methods are unable to detect.<sup>10</sup>

However, when combined with recent advances in machine learning (ML) and deep learning spectral analysis techniques, it may be possible to use Raman spectroscopy to achieve detection of MC-LR at concentrations comparable with those of DCDR or SERS, which would

suggest the combination of Raman and modern ML methods could be useful for detection of MC at concentrations near the EPA's lowest limit of 1  $\mu\text{g/L}$ .<sup>5</sup> With a sensitive enough spectral analysis technique, Raman spectroscopy may be able to offer a reliable approach to monitor low concentrations of MCs in solution.

### **Significance**

The development of a rapid, non-destructive approach to detect even low concentrations of MC in water would enable effective management and control of these toxins.<sup>7</sup> A detection approach that meets these criteria is crucial in order to prevent associated health risks — such as headaches, fever, diarrhea, abdominal pain, nausea and vomiting — from drinking or swimming in water containing MCs.<sup>11</sup> Although traditional statistical data analysis methods used with Raman spectroscopy are not suited for detecting low concentrations of MC in water, the Raman spectroscopic experimental technique meets other criteria for a suitable MC detection method because it is rapid, requires minimal sample preparation, is non-destructive, and can even be applied using portable Raman spectrometers.

By applying modern machine learning techniques to Raman spectra collected across various concentrations of MC-LR in solution, there is potential to improve the sensitivity of Raman data analysis to low concentrations of MC-LR. In an improvement from basic statistical analysis methods, certain multivariate analysis methods excel at processing Raman and IR spectral data because of their ability to more accurately differentiate between subtle spectral changes with respect to concentration which may be overlooked with other techniques.<sup>12,13</sup> A more advanced ML algorithm is better poised to distinguish between different concentrations of an analyte by learning minor changes in the peak intensity and peak shape of the Raman spectra at different analyte concentrations. Beyond MC detection, using machine learning to improve

Raman's limit of detection could apply to many other challenges such as determining illicit substance concentration, disease diagnoses, and determining aqueous food ingredient concentration.<sup>14-16</sup>

## **Methods**

### **Overview**

Raman spectra were collected from solutions of MC-LR diluted in water at 100 concentrations within range 0.001 to 6.0  $\mu\text{g/L}$  according to the procedure described below. After Raman spectra acquisition, a sample size of  $n=1000$  Raman spectra was achieved. All spectra were pre-processed via background subtraction of the average water spectrum calculated throughout data collection. Three machine learning models were then individually trained and tested on collected spectra according to a 80/20 train-test split: kernel support vector machine for regression (SVR), regression deep neural network (DNN), and partial least squares regression (PLSR). Further feature selection experiments were performed to compare the mean-square-error (MSE) of each model.

### **Procedure**

#### *Sample Preparation*

MC-LR was purchased commercially (Enzo Life Sciences) and prepared in aqueous solutions at concentrations ranging from 0.001 to 6.0  $\mu\text{g/L}$  (Table A). To create the stock solution of 10  $\mu\text{g/L}$  MC-LR, Milli-Q water was repeatedly pipetted into the purchased sample vial containing 500  $\mu\text{g}$  of solid MC-LR, the vial was sonicated for 2 minutes, then the contents were aspirated to attain a from-vial solution concentration of 500  $\mu\text{g/L}$  MC-LR. The stock solution of



10  $\mu\text{g/L}$  was then prepared from this from-vial solution, and used to prepare 100 sample vials by aliquoting of Milli-Q water and stock solution according to the volumes in Table B and Table C. The resulting 100 sample vials prepared at concentrations described in Table A were stored at refrigerator temperature while not in use.

*Raman Spectral Acquisition*

A 532 nm polarized Raman spectrometer with a 600 blaze grating was used to collect all spectra, and a sample size of  $n=1,000$  was achieved. Exposure time was set to 0.1 second, with each recorded spectrum averaged among 20 exposures. All sample vials were removed from the refrigerator before sample collection to allow for equilibration to room temperature. To collect 10 averaged spectra for each concentration represented in Table A, the corresponding sample vial was shaken for 20 seconds, then  $\sim 2$  mL of solution was transferred from the corresponding sample vial into the sample cuvette. The cuvette was rinsed twice with Milli-Q water between each sample collection. Also, 10 water spectra were collected every 5 samples in order to form an average water spectrum for background subtraction. Each spectrum was saved in csv format as a 1 x 1339-dimensional array, where each entry corresponds to an intensity at a wavenumber.

**Table A.** Sample concentration distribution for prepared vials of MC-LR in water ( $\mu\text{g/L}$ ).

	A	B	C	D	E	F	G	H	I	J
1	0.001	0.015	0.15	0.65	0.8	1.1	2.1	3.1	4.1	5.1
2	0.002	0.02	0.2	0.66	0.85	1.2	2.2	3.2	4.2	5.2
3	0.003	0.03	0.25	0.67	0.9	1.3	2.3	3.3	4.3	5.3
4	0.004	0.04	0.3	0.68	0.91	1.4	2.4	3.4	4.4	5.4

5	0.005	0.05	0.35	0.69	0.93	1.5	2.5	3.5	4.5	5.5
6	0.006	0.06	0.4	0.7	0.95	1.6	2.6	3.6	4.6	5.6
7	0.007	0.07	0.45	0.71	0.97	1.7	2.7	3.7	4.7	5.7
8	0.008	0.08	0.5	0.72	0.98	1.8	2.8	3.8	4.8	5.8
9	0.009	0.09	0.55	0.73	0.99	1.9	2.9	3.9	4.9	5.9
10	0.01	0.1	0.6	0.75	1	2	3	4	5	6

**Table B.** Volume matrix ( $\mu\text{g/L}$ ) of 10  $\mu\text{g/L}$  MC-LR stock solution to prepare sample vials.

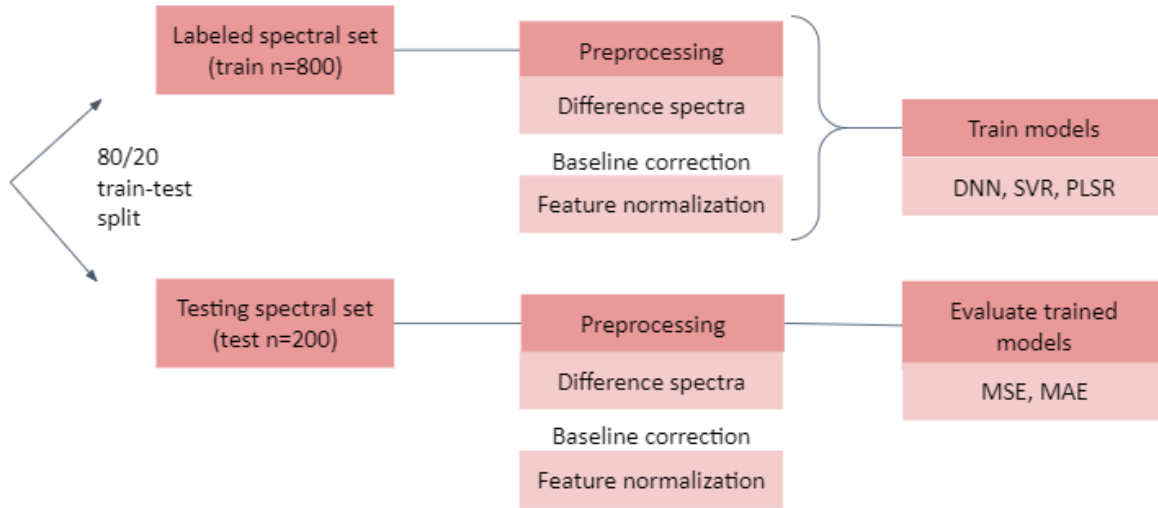
	A	B	C	D	E	F	G	H	I	J
1	0.002	0.03	0.3	1.3	1.6	2.2	4.2	6.2	8.2	10.2
2	0.004	0.04	0.4	1.32	1.7	2.4	4.4	6.4	8.4	10.4
3	0.006	0.06	0.5	1.34	1.8	2.6	4.6	6.6	8.6	10.6
4	0.008	0.08	0.6	1.36	1.82	2.8	4.8	6.8	8.8	10.8
5	0.01	0.1	0.7	1.38	1.86	3	5	7	9	11
6	0.012	0.12	0.8	1.4	1.9	3.2	5.2	7.2	9.2	11.2
7	0.014	0.14	0.9	1.42	1.94	3.4	5.4	7.4	9.4	11.4
8	0.016	0.16	1	1.44	1.96	3.6	5.6	7.6	9.6	11.6
9	0.018	0.18	1.1	1.46	1.98	3.8	5.8	7.8	9.8	11.8
10	0.02	0.2	1.2	1.5	2	4	6	8	10	12

**Table C.** Volume matrix ( $\mu\text{g/L}$ ) of Milli-Q water used to prepare sample vials.

	A	B	C	D	E	F	G	H	I	J
1	19.998	19.97	19.7	18.7	18.4	17.8	15.8	13.8	11.8	9.8
2	19.996	19.96	19.6	18.68	18.3	17.6	15.6	13.6	11.6	9.6
3	19.994	19.94	19.5	18.66	18.2	17.4	15.4	13.4	11.4	9.4
4	19.992	19.92	19.4	18.64	18.18	17.2	15.2	13.2	11.2	9.2
5	19.99	19.9	19.3	18.62	18.14	17	15	13	11	9
6	19.988	19.88	19.2	18.6	18.1	16.8	14.8	12.8	10.8	8.8
7	19.986	19.86	19.1	18.58	18.06	16.6	14.6	12.6	10.6	8.6
8	19.984	19.84	19	18.56	18.04	16.4	14.4	12.4	10.4	8.4
9	19.982	19.82	18.9	18.54	18.02	16.2	14.2	12.2	10.2	8.2
10	19.98	19.8	18.8	18.5	18	16	14	12	10	8

### Regression Machine Learning Modeling

Three regression models were built: kernel support vector machine for regression (SVR), regression deep neural network (DNN), and partial least squares regression (PLSR) model. The average water spectrum was subtracted from all train and test spectra to produce difference spectra. Then, baseline removal was applied to each spectrum using the IModPoly algorithm to conduct multi-polynomial fitting.<sup>17</sup> Second degree polynomials were used for fitting with IModPoly. Finally, feature normalization to ensure a zero mean and unit variance was explored with one model for each algorithm. To prevent data leaking, features were normalized separately for the train and test datasets.



**Figure 1:** Flowchart of preprocessing methods and regression model construction.

Models were trained and evaluated with an 80/20 train-test split using the spectra and their associated MC-LR concentrations as labels. Prediction performance of the three models in the range 0.001 to 6.0  $\mu\text{g/L}$  MC-LR was evaluated using MSE of the withheld test data ( $n=200$ ), with  $n=1,000$  being the number of spectra and  $y_i$  and  $\hat{y}$  being predicted and observed values, respectively.<sup>15</sup> I hypothesized that the DNN would outperform the kernel SVM, which would outperform the PLSR algorithm, with MSE values increasing in order of  $\text{DNN} < \text{SVR} < \text{PLSR}$ . Mean-absolute-error (MAE) was also used for validation of each model on withheld test data.

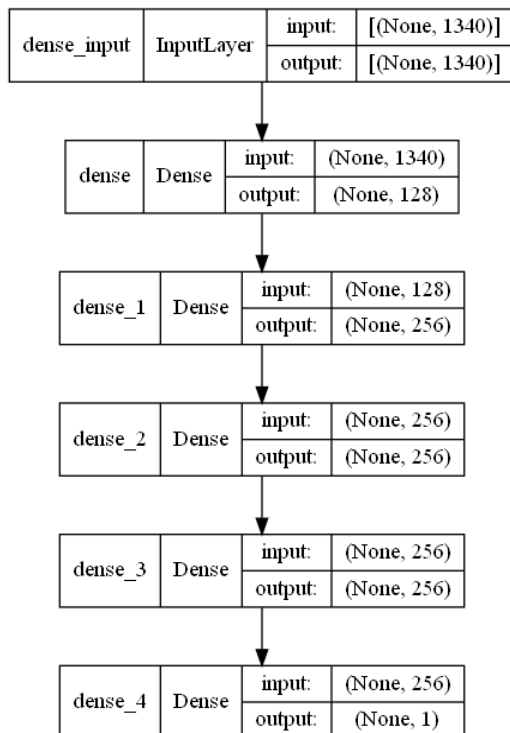
$$MSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y} - y_i)^2}{n}}$$

**Formula 1.** Mean-square-error (MSE) formula, where  $\hat{y}$  is the predicted value and  $y_i$  is the true value.

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

**Formula 2.** Mean-absolute-error (MAE) formula, where  $\hat{y}$  is the predicted value and  $y_i$  is the true value.

A deep neural network for regression was built using *Keras* with the architecture detailed in Figure 1, and ReLu activation was used for all hidden layers. A bias term was appended to each example and initialized to 0 to extend each 1 x 1339-dimensional spectrum to 1 x 1340. The MSE loss function was used for training and the Adam optimizer was used for stochastic gradient descent. The batch size was chosen to be 32, a kernel initializer of type normal was chosen for each layer, the model was trained for 150 epochs, and early stopping was employed when validation loss did not improve between epochs. An 80/20 train-test split was conducted before training and evaluating the model. No feature normalization was conducted for the DNN model.



**Figure 2.** DNN model architecture for training on full Raman spectrum.

A kernel SVR was built using *scikit-learn* using the radial basis function (RBF) kernel. An 80/20 train-test split was conducted before fitting and evaluating the model. Additionally features were normalized to ensure a zero mean and unit variance. To prevent data leaking, features were normalized separately for the train and test datasets.

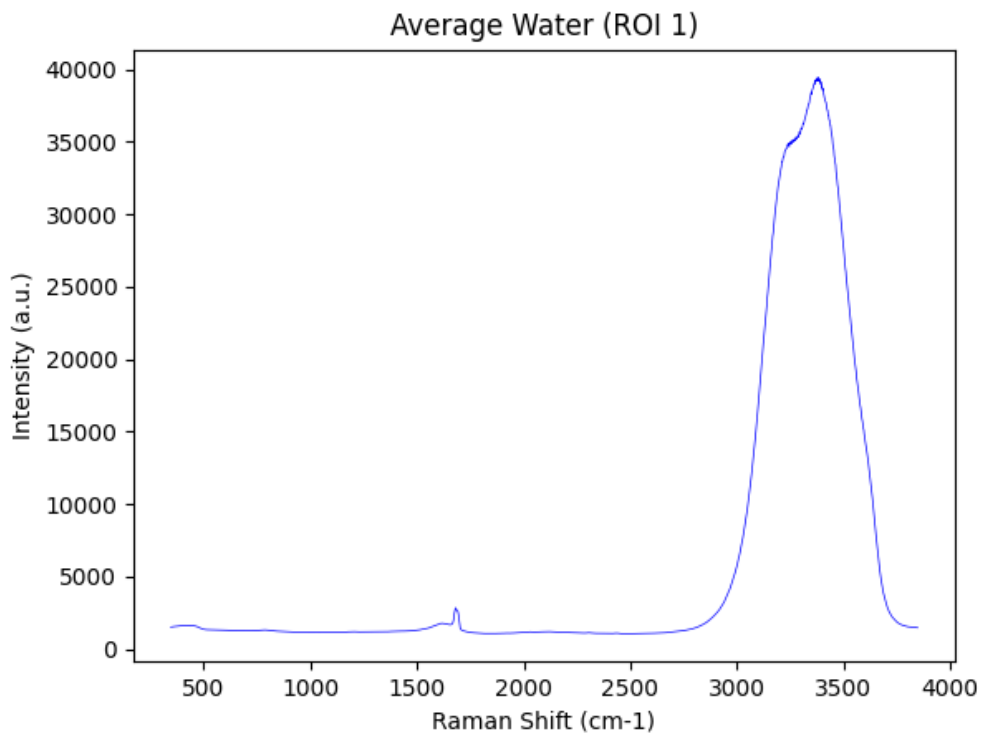
A PLSR model was built using *scikit-learn*. K-fold cross-validation (k=10) was conducted for 3 epochs on training data after an 80/20 train-test split in order to perform component analysis, building components which are correlated with MC-LR concentration changes.<sup>10</sup> The number of partial least squares components used was determined before fitting based on which optimal number of components minimized MSE during k-fold cross-validation of training data (k=10). Additionally, features were normalized to ensure a zero mean and unit

variance. To prevent data leaking, features were normalized separately for the train and test datasets.

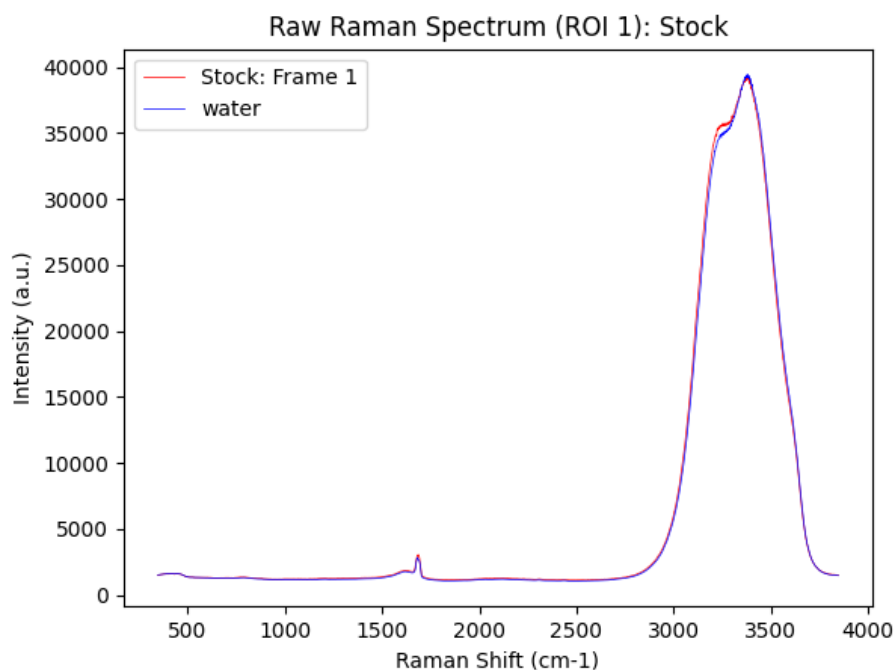
## Results and Discussion

### Water Spectra and Raw MC-LR Spectra

An average Raman spectrum of water was obtained by averaging water spectra collected every 50 spectra (n=210) in order to account for day-to-day environmental differences potentially affecting the background.



**Figure 3.** Averaged water spectrum built from water spectra collected every 5 vials.



**Figure 4.** Stock solution Raman spectrum (10  $\mu\text{g/L}$ ) overlaid with the average water Raman spectrum.

## Evaluation of Machine Learning Regression Models

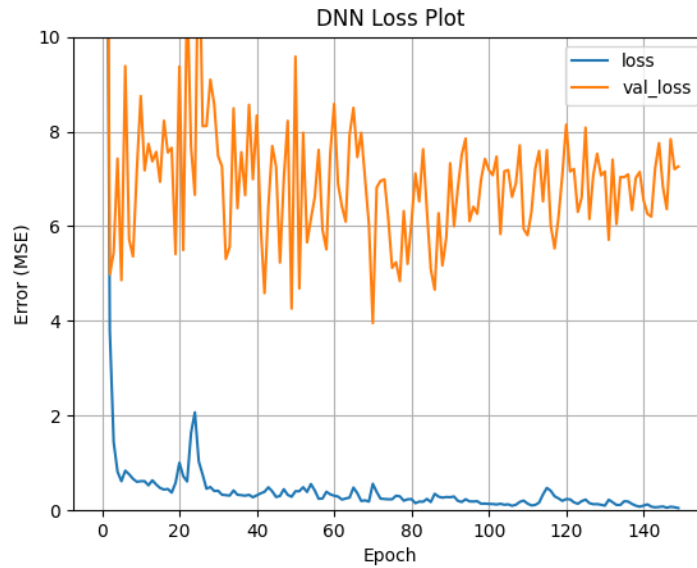
### *DNN*

For the DNN model, using difference spectra, baseline subtraction, but unnormalized features for train and test data, an average test MSE of 3.155 and average MAE of 1.497 were achieved averaging among three separately trained models (Figure 6). The loss plot of a trained model shows oscillation of the validation loss while the training loss decreases as epochs progress (Figure 5). To prevent overfitting, early stopping was used, where the model selected for evaluation on test data used the weights at the last checkpoint with minimal validation loss.

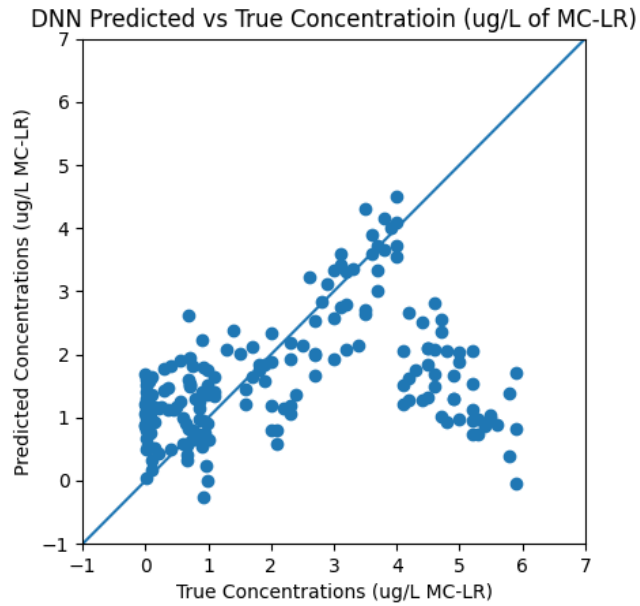
After evaluating the DNN models on test data ( $n=200$ ), concentrations appear to be consistently overpredicted in the true concentration range 0.001 - 1  $\mu\text{g/L}$ , and underpredicted by



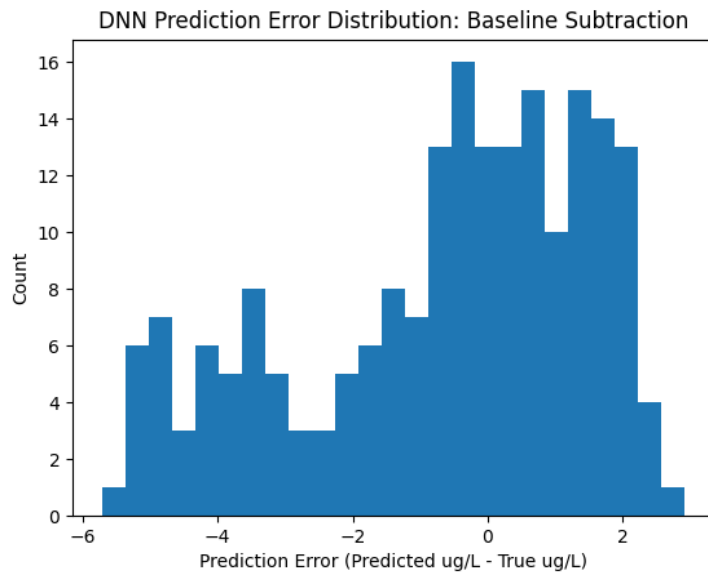
5 - 6  $\mu\text{g/L}$  in the true concentration range 4 - 6  $\mu\text{g/L}$ . However, the model appears to predict concentrations with low error in the range 1 - 4  $\mu\text{g/L}$  (Figure 6).



**Figure 5.** Regression DNN loss plot during training, where loss indicates training loss, and val\_loss indicates validation loss.



**Figure 6.** Prediction scatter plot of test data (n=200) from the regression DNN model.

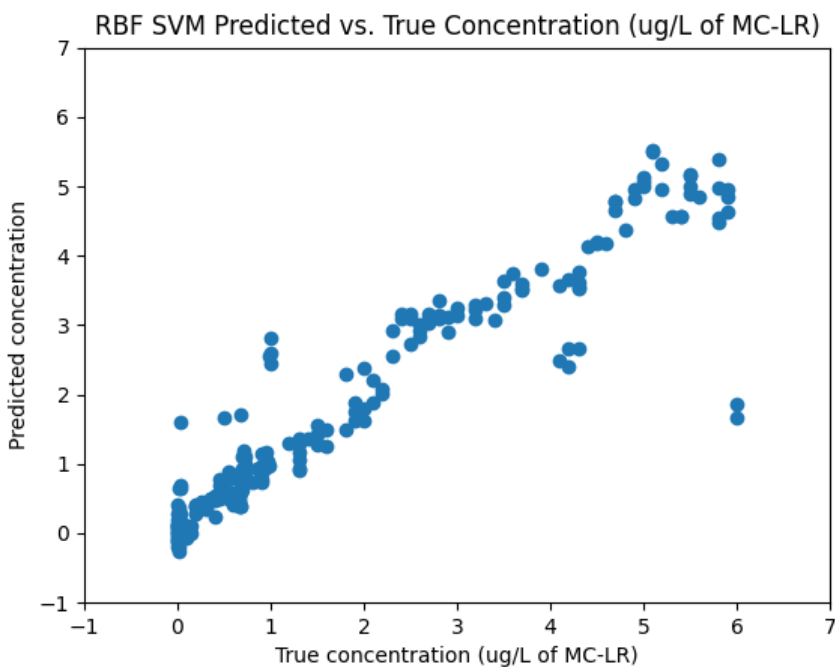


**Figure 7.** Prediction error histogram of test data (n=200) from the regression DNN model.

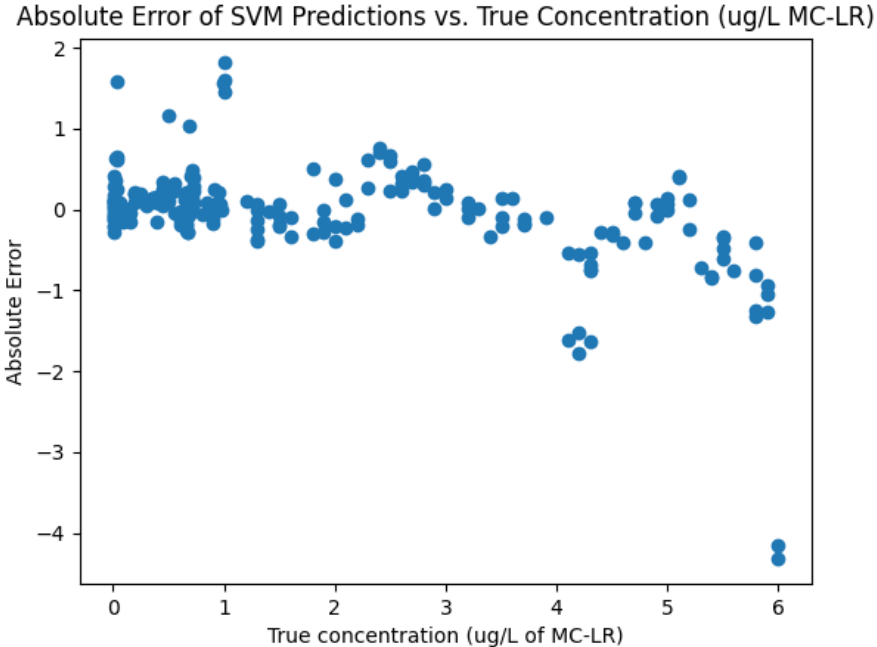
## SVR

For the RBF kernel SVR model, test MSE of 0.432 and MAE of 0.371 were achieved by using difference spectra, baseline removal, and normalized features separately for train and test data (Figure 8).

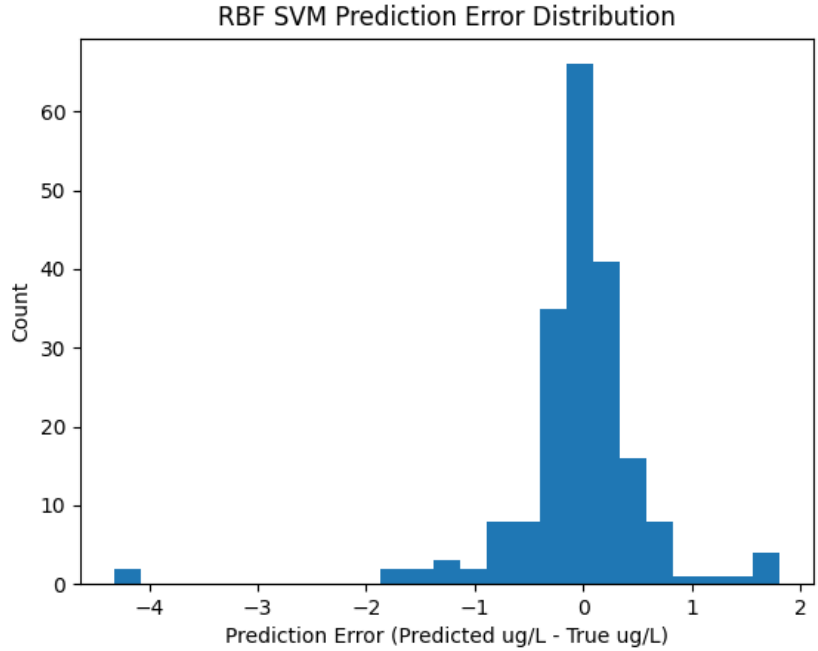
After evaluating on test data (n=200), concentrations appear to be predicted with moderate error in the true concentration range 0 - 1  $\mu\text{g/L}$  and somewhat underpredicted in the range 4 - 6  $\mu\text{g/L}$ . However, the SVR model appears to predict concentrations with low error in the range 0.001 - 4  $\mu\text{g/L}$  (Figure 9). For a small number of test spectra with true concentration around 6  $\mu\text{g/L}$  MC-LR, concentration was significantly underpredicted (Figure 8).



**Figure 8.** Prediction scatter plot of test data (n=200) from the RBF SVR model trained on spectra preprocessed with baseline removal and normalized features.



**Figure 9.** Absolute error scatter plot of test data (n=200) from the RBF SVR model trained on spectra preprocessed with baseline removal and normalized features.

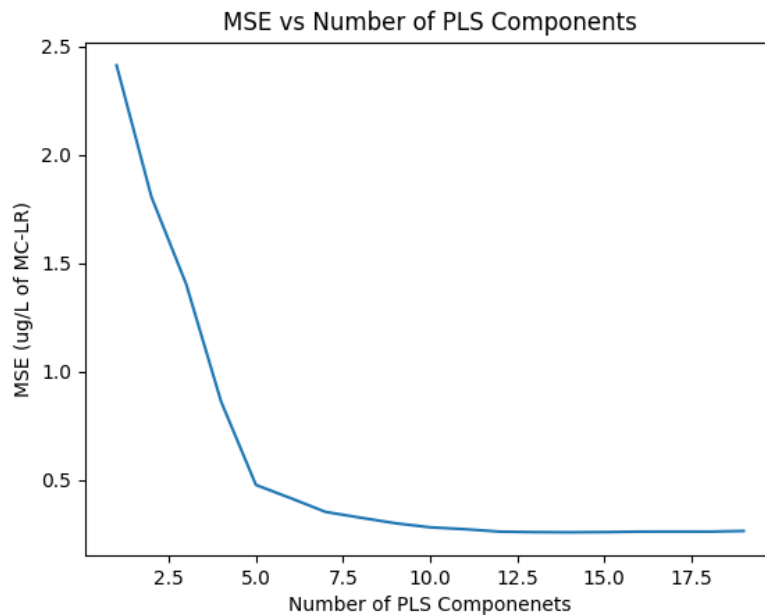


**Figure 10.** Prediction error histogram of test data (n=200) from the RBF SVR model trained on spectra preprocessed with baseline removal and normalized features.

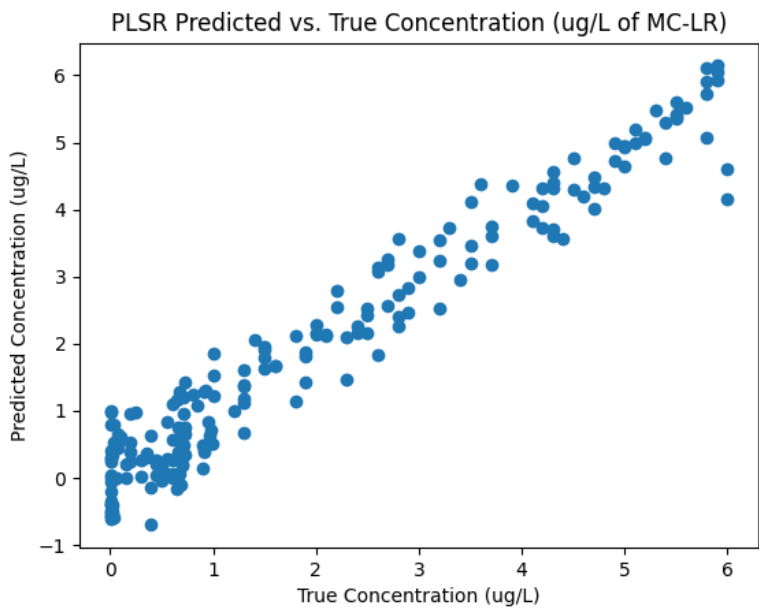
### *PLSR*

Using difference spectra, baseline subtraction, and scaled features for train and test data, test MSE of 0.199 and MAE of 0.353 were achieved using 16 partial least squares components (Figure 12). The optimal number of PLS components was determined by comparing MSE of PLSR models with an increasing number of PLS components, and choosing the number of components which minimized MSE on the training dataset (Figure 11). When features were not scaled when evaluating the test data, a test MSE of 0.191 and MAE of 0.332 was recorded (Figure 13).

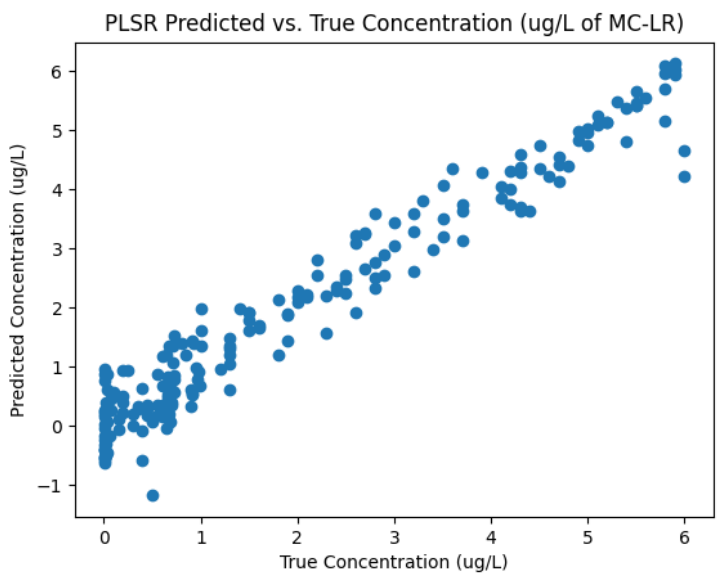
After evaluating the model on test data (n=200), concentrations appear to be predicted with low error in the true concentration range 0.001 - 5  $\mu\text{g/L}$  when baseline removal and feature normalization was applied to the data. Also, concentrations were mildly underpredicted by 1.5 - 2  $\mu\text{g/L}$  for 2-3 test examples in the range 5.5 - 6  $\mu\text{g/L}$ . When baseline removal was applied to data but not feature normalization was not, concentrations of 2 test examples with true concentrations between 0.001 - 0.5  $\mu\text{g/L}$  were underpredicted by 1 - 1.5  $\mu\text{g/L}$  (Figure 17).



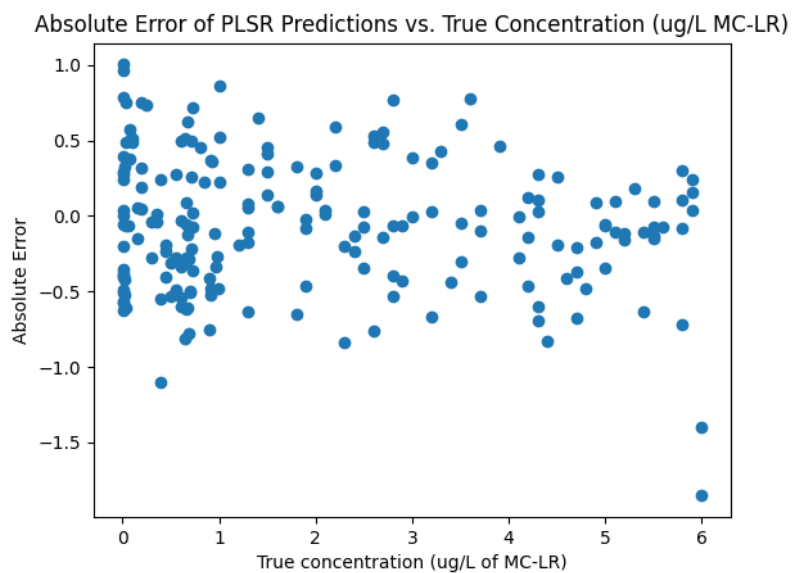
**Figure 11.** Mean-square-error (MSE) of PLSR models vs. number of PLS components used.



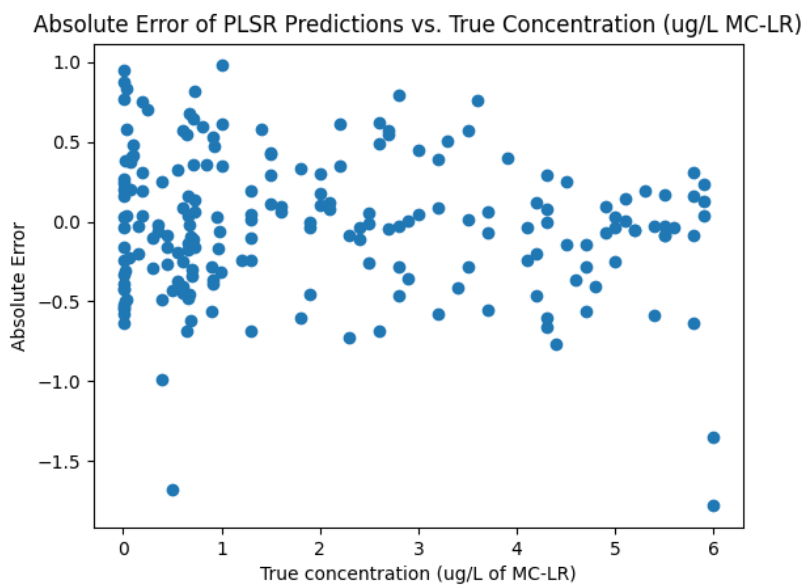
**Figure 12.** PLSR prediction scatter plot of test data (n=200) using baseline removal and normalized features.



**Figure 13.** PLSR prediction scatter plot of test data (n=200) using baseline removal and normalized features.

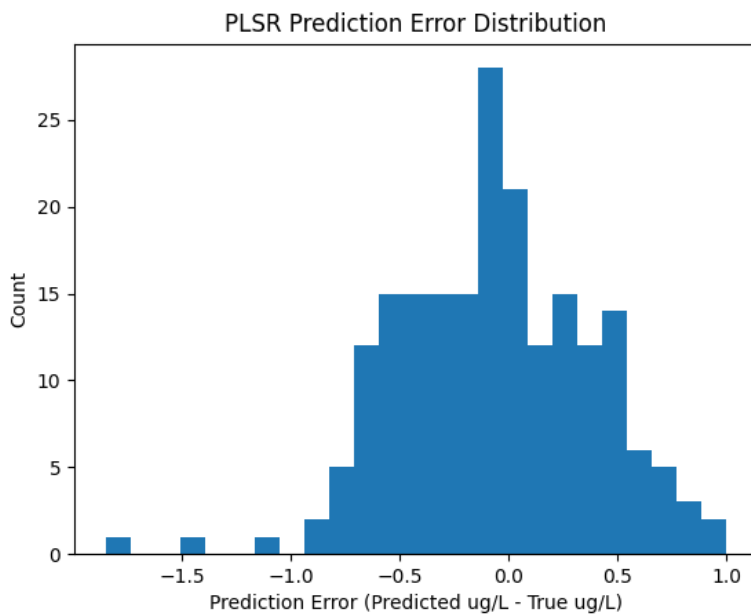


**Figure 14.** PLSR test errors using baseline removal and normalization of features.

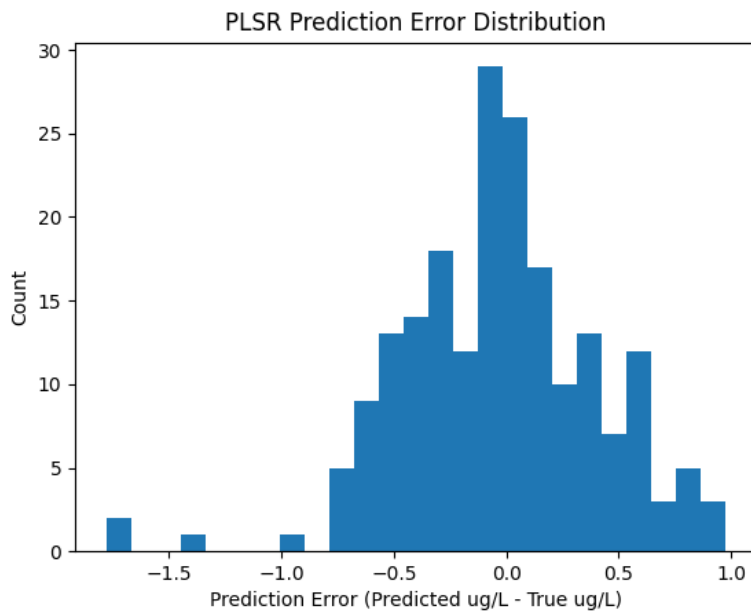


**Figure 15.** PLSR test errors using baseline removal without normalization of features





**Figure 16.** PLSR error histogram using baseline removal and feature normalization.




**Figure 17.** PLSR error histogram using baseline without feature normalization.

## Comparison of Machine Learning Models

It was hypothesized that for prediction of MC-LR concentration on withheld test data in the range 0.001 to 6.0  $\mu\text{g/L}$ , the DNN would outperform the kernel SVM, which would outperform PLSR models, with MSE values increasing in order of DNN < SVR < PLSR. After evaluating each trained model on test data (n=200), MSE values were found to increase in the order of PLSR without feature normalization (0.191) < PLSR with feature normalization (0.199) < SVR (0.432) < DNN (3.155). These results do not support the hypothesis for MSE of the three models.

Difference Spectra:  
raw - avg\_water



Model	Baseline Removal	Z-Score Normalization
DNN	3.155, 1.497	–
RBF SVR	–	0.432, 0.371
PLSR	0.199, 0.353	0.191, 0.332

**Table D.** Performance of trained models on test data (n=200) with varying spectral preprocessing methods. Metrics here are recorded as mean-square error (MSE), mean-absolute-error (MAE).

When predicting concentration of MC-LR in water in the range 0.001 to 6.0  $\mu\text{g/L}$ , PLSR models with background subtraction of water and baseline removal were found to perform best compared to kernel SVR and DNN models. Superior performance of PLSR models may be due

to the use of PLS components to effectively select combinations of features which are correlated with MC-LR concentration.<sup>10</sup> In contrast, SVR and DNN models were built and evaluated using all features in each 1 x 1339 dimensional Raman spectrum, which likely incorporated redundant spectral information into the models and impaired performance.

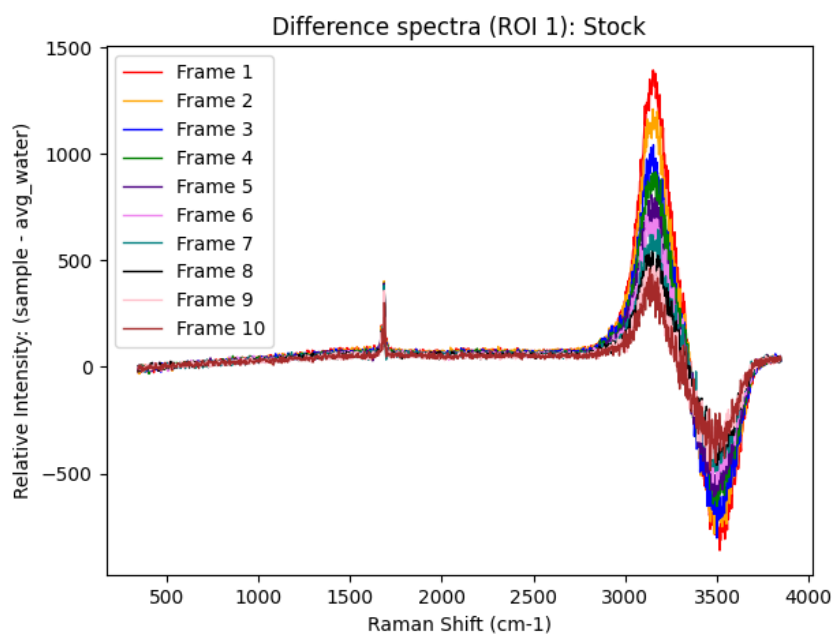
Among all models evaluated, the PLSR model with baseline removal and normalization performed most consistently among the entire concentration range 0.001 - 6  $\mu\text{g/L}$ , with minor underpredictions by 1.5 - 2  $\mu\text{g/L}$  in the true concentration range 5.5 - 6  $\mu\text{g/L}$  (Figure 14). The PLSR model with baseline removal but not normalization also performed consistently, but suffered from minor underprediction error by 1 - 1.5  $\mu\text{g/L}$  in the true concentration range 0.001 - 0.5  $\mu\text{g/L}$  (Figure 15).

When considering the potential efficacy of models for MC-LR monitoring, DNN and SVR models may benefit from dimensionality reduction of spectra to improve prediction of high concentrations of MC-LR, as both models underpredicted concentration for 2-3 test examples (Figures 7, 10). In particular, underprediction errors of high concentrations of MC-LR above the EPA limits of 1  $\mu\text{g/L}$  for drinking water and 1.8  $\mu\text{g/L}$  for swimming water have potential to be especially harmful if made in practice, therefore future work exploring additional preprocessing methods or data augmentation for DNN and SVR methods is suggested. Experimental error during sample preparation or collection of Raman spectra is another potential explanation for these underpredictions observed at high concentrations by DNN and SVR models. As a result, future work is suggested to replicate sample preparation and collection of Raman spectra.

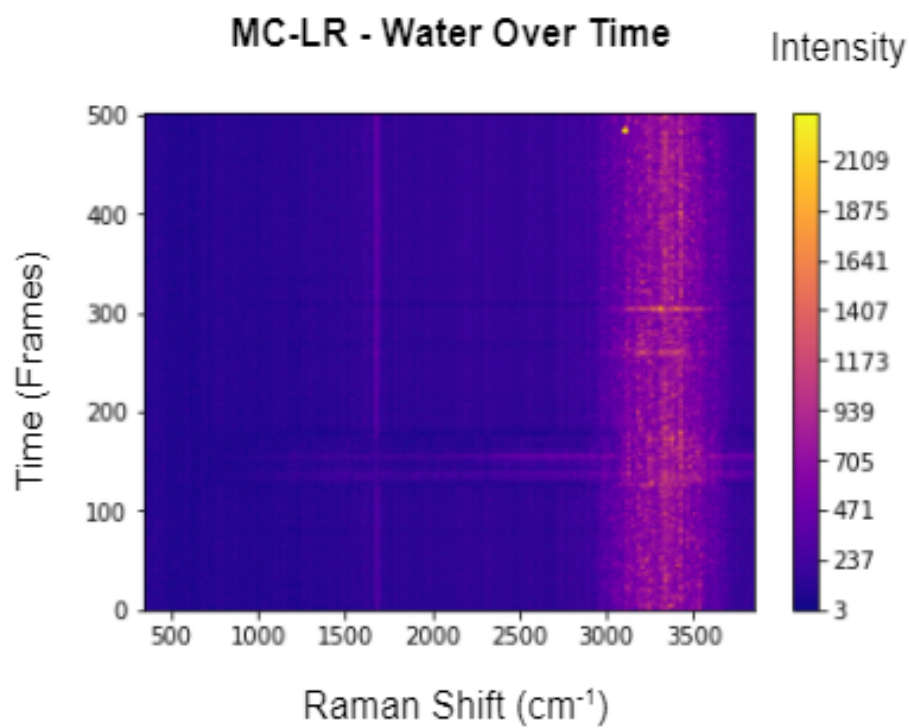
## **Deposition Study**

After observing peak shrinkage as acquisition of frames for each concentration progressed, potential MC-LR deposition in water was studied in the stock solution of 10  $\mu\text{g/L}$  (Figure 18). Within the ten frames collected for each concentration, consistent peak shrinkage and decrease in peak intensity were observed in the symmetric -OH stretching region around 3410  $\text{cm}^{-1}$  as spectral acquisition for a single sample progressed. For example, this can be observed among the difference spectra of the stock (10.0  $\mu\text{g/L}$ ) in Figure 18, where the subtracted peaks around 3100 and 3500  $\text{cm}^{-1}$  appear to shrink as collection progresses from frame 1 to 10 over time. Similar patterns of peak shrinking in the -OH stretching region were observed during collection of spectra with sample concentrations in the range 0.001 - 6.0  $\mu\text{g/L}$ .

Potential MC-LR deposition in water was further studied in the stock solution of 10  $\mu\text{g/L}$ , and a heatmap in Figure 19 shows intensity (a.u.) at each Raman shift wavenumber ( $\text{cm}^{-1}$ ) over 500 frames. Although the heat map did not confirm significant deposition of the MC-LR through a decrease in intensity over time around 3410  $\text{cm}^{-1}$  (Figure 18), future work is suggested to compare the peak intensity over time for MC-LR among known solvents for MC-LR including 100% ethanol, methanol, and dimethylsulfoxide (DMSO).<sup>18</sup>



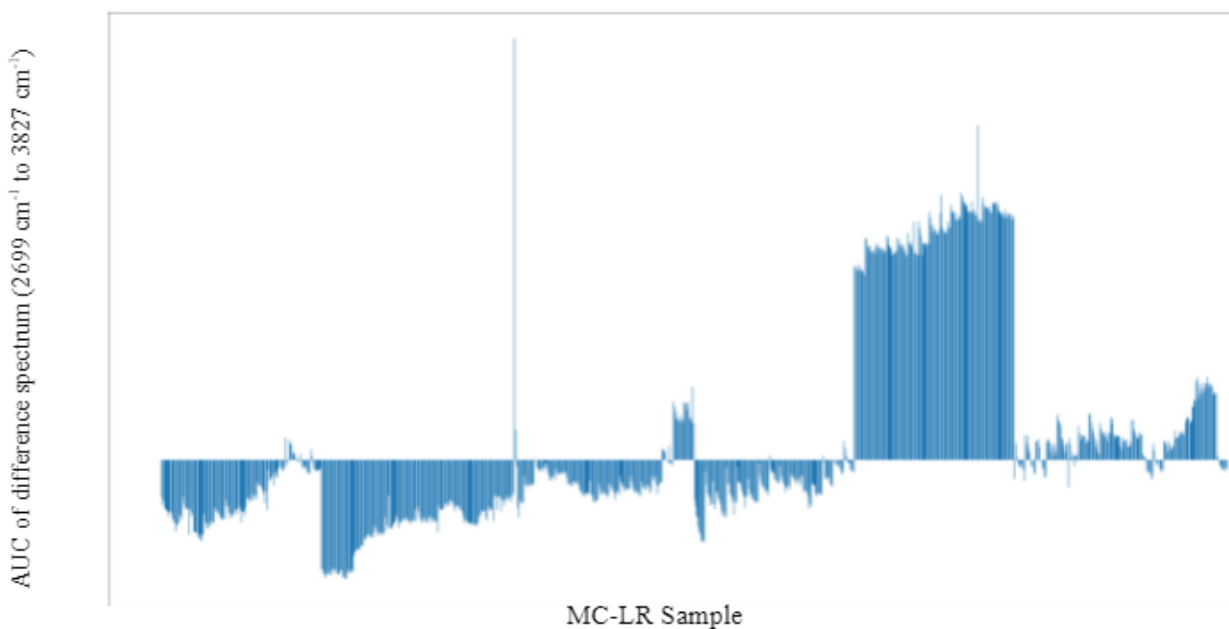
**Figure 18.** Difference spectra for 10 frames of MC-LR stock (10.0  $\mu\text{g/L}$ ).



**Figure 19.** Heat map for Raman intensity of MC-LR stock over 500 frames (10  $\mu\text{g/L}$ ).

### Analysis of Variance (ANOVA) for Area Under the Curve of Difference Spectra

All 1,000 Raman difference spectra of MC-LR in water were cropped to the region  $2699\text{ cm}^{-1}$  to  $3827\text{ cm}^{-1}$ , which contains the peak for the OH stretching mode in the Raman spectrum of water.<sup>19</sup> The area under the curve (AUC) of each cropped spectrum was then calculated using Simpson's rule. The resulting AUC was plotted for each of the 1,000 samples, representing 100 separate concentrations (Figure 20). Several outliers were observed among these AUC values, which could indicate the presence of experimental error, especially for samples F1 through H10.



**Figure 20.** Area under the curve (AUC) 1000 samples of MC-LR difference spectra cropped to the region  $2699\text{ cm}^{-1}$  to  $3827\text{ cm}^{-1}$ .

In order to determine if the spectra of MC-LR in water differed significantly among different concentration groups, ANOVA testing was conducted on the AUC values for the

cropped regions calculated above. There were 100 groups used, each group corresponding to a sample concentration in the range 0.001 - 6.0  $\mu\text{g/L}$  from Table A. After conducting a one-way ANOVA test with an alpha value of 0.01, we reject the null hypothesis and conclude that the mean AUC is significantly different among some Raman spectra for the various concentrations of MC-LR ( $p$  was approximately equal to 0). However, because outliers in AUC potentially due to experimental error were observed, a second ANOVA test was conducted excluding samples F1 through H10. After conducting this one-way ANOVA test with 80 groups and an alpha value of 0.01, we still reject the null hypothesis and conclude that the mean AUC is significantly different among some Raman spectra for the various concentrations of MC-LR ( $p = 1.3\text{E-}184$ ). Therefore, a statistically significant difference in AUC of MC-LR Raman difference spectra in the region  $2699\text{ cm}^{-1}$  to  $3827\text{ cm}^{-1}$  was found after excluding samples F1 through H10 where experimental error is assumed to have inflated the AUC.

Two chemical explanations are proposed for this statistically significant difference in AUC among concentrations. First, ion pairing of MC-LR with water could occur if MC-LR molecules ionize and pair with water molecules via the polar carboxylic acids, amino and amido functions on MC-LR.<sup>20</sup> Second, the hydrophobicity of MC-LR may also lead to micelle formation, causing perturbations to the Raman spectrum of water as the concentration of MC-LR in solution increases. Studies have found that Raman spectroscopy is sensitive to aggregation dynamics and hydration shell effects, suggesting that the Raman spectrum of water could be impacted by either of the above processes as the concentration of MC-LR in water increases.

## Conclusion

After using Raman spectra of MC-LR in water to predict concentrations of MC-LR in the range 0.001 - 6.0  $\mu\text{g/L}$ , findings from validation of machine learning models on withheld Raman spectra do not support the hypothesis that a DNN would outperform the kernel SVM, which would outperform the PLSR models, with MSE values hypothesized to increase in order of DNN < SVR < PLSR. After validation of the models on test data (n=200), MSE values were found to increase in the order of PLSR without feature normalization (0.191) < PLSR with feature normalization (0.199) < SVR (0.432) < DNN (3.155). Additionally, MAE values, which indicate the average absolute prediction error ( $\mu\text{g/L}$ ), were found to increase in the order of PLSR without feature normalization (0.332) < PLSR with feature normalization (0.353  $\mu\text{g/L}$ ) < SVR (0.371  $\mu\text{g/L}$ ) < DNN (1.479  $\mu\text{g/L}$ ). Based on the results of ANOVA testing, which indicated that a statistically significant difference in AUC of MC-LR Raman difference spectra in the OH stretching region 2699  $\text{cm}^{-1}$  to 3827  $\text{cm}^{-1}$  was found, two potential chemical explanations for this change in AUC among concentrations were proposed: ion pairing of MC-LR with water, and MC-LR hydrophobicity leading to micelle formation. Future work is suggested to repeat collection of Raman spectra to obtain a larger sample size with higher signal-to-noise ratio spectra and to verify that AUC in the OH stretching region changes as a function of MC-LR concentration. Additionally, future work is suggested to augment the sample size of Raman spectra to greater than n = 1,000 in order to retrain and avoid overfitting of the regression DNN models.



## References

- (1) EPA. FACTOIDS: Drinking Water and Ground Water Statistics for 2007, 2008.
- (2) Baker, B.; Aldridge, C.; Omer, A. *Water: Availability and Use*, 2016.
- (3) Francy, D. S.; Brady, A. M. G.; Stelzer, E. A.; Cicale, J. R.; Hackney, C.; Dalby, H. D.; Struffolino, P.; Dwyer, D. F. Predicting Microcystin Concentration Action-Level Exceedances Resulting from Cyanobacterial Blooms in Selected Lake Sites in Ohio. *Environ. Monit. Assess.* **2020**, *192* (8), 513. <https://doi.org/10.1007/s10661-020-08407-x>.
- (4) Schmidt, J.; Wilhelm, S.; Boyer, G. The Fate of Microcystins in the Environment and Challenges for Monitoring. *Toxins* **2014**, *6* (12), 3354–3387. <https://doi.org/10.3390/toxins6123354>.
- (5) EPA. Recommended Human Health Recreational Ambient Water Quality Criteria or Swimming Advisories for Microcystins and Cylindrospermopsin, 2019.
- (6) Pham, T.-L.; Utsumi, M. An Overview of the Accumulation of Microcystins in Aquatic Ecosystems. *J. Environ. Manage.* **2018**, *213*, 520–529. <https://doi.org/10.1016/j.jenvman.2018.01.077>.
- (7) Massey, I. Y.; Yang, F. A Mini Review on Microcystins and Bacterial Degradation. *Toxins* **2020**, *12* (4), 268. <https://doi.org/10.3390/toxins12040268>.
- (8) Li, M.; Paidi, S. K.; Sakowski, E.; Preheim, S.; Barman, I. Ultrasensitive Detection of Hepatotoxic Microcystin Production from Cyanobacteria Using Surface-Enhanced Raman Scattering Immunosensor. *ACS Sens.* **2019**, *4* (5), 1203–1210. <https://doi.org/10.1021/acssensors.8b01453>.
- (9) Jones, R. R.; Hooper, D. C.; Zhang, L.; Wolverson, D.; Valev, V. K. Raman Techniques: Fundamentals and Frontiers. *Nanoscale Res. Lett.* **2019**, *14* (1), 231. <https://doi.org/10.1186/s11671-019-3039-2>.
- (10) Pelletier, M. J. Quantitative Analysis Using Raman Spectrometry. *Appl. Spectrosc.* **2003**, *57* (1), 20A–42A. <https://doi.org/10.1366/000370203321165133>.
- (11) Kordasht, H. kholafazad; Hassanpour, S.; Baradaran, B.; Nosrati, R.; Hashemzaei, M.; Mokhtarzadeh, A.; la Guardia, M. de. Biosensing of Microcystins in Water Samples; Recent Advances. *Biosens. Bioelectron.* **2020**, *165*, 112403. <https://doi.org/10.1016/j.bios.2020.112403>.
- (12) Horton, R. B.; Duranty, E.; McConico, M.; Vogt, F. Fourier Transform Infrared (FT-IR) Spectroscopy and Improved Principal Component Regression (PCR) for Quantification of Solid Analytes in Microalgae and Bacteria. *Appl. Spectrosc.* **2011**, *65* (4), 442–453. <https://doi.org/10.1366/10-06122>.
- (13) O'Connell, M.-L.; Ryder, A. G.; Leger, M. N.; Howley, T. Qualitative Analysis Using Raman Spectroscopy and Chemometrics: A Comprehensive Model System for Narcotics Analysis. *Appl. Spectrosc.* **2010**, *64* (10), 1109–1121. <https://doi.org/10.1366/000370210792973541>.
- (14) Auner, G. W.; Koya, S. K.; Huang, C.; Broadbent, B.; Trexler, M.; Auner, Z.; Elias, A.; Mehne, K. C.; Brusatori, M. A. Applications of Raman Spectroscopy in Cancer Diagnosis. *Cancer Metastasis Rev.* **2018**, *37* (4), 691–717. <https://doi.org/10.1007/s10555-018-9770-9>.
- (15) Lin, M.; Wu, Y.; Rohani, S. Simultaneous Measurement of Solution Concentration and Slurry Density by Raman Spectroscopy with Artificial Neural Network. *Cryst. Growth Des.* **2020**, *20* (3), 1752–1759. <https://doi.org/10.1021/acs.cgd.9b01482>.
- (16) Rahimi, N.; Price, A.; Ghasempour, A.; Pan, X.; Ndi, F. Determination of Extremely Low Concentration of Sucrose in Aqueous Solution by Raman Spectroscopy. In *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XVII*; Farkas, D. L., Leary, J. F., Tarnok, A., Eds.; SPIE: San Francisco, United States, 2019; p 20.

- <https://doi.org/10.1117/12.2510078>.
- (17) Zhao, J.; Lui, H.; McLean, D. I.; Zeng, H. Automated Autofluorescence Background Subtraction Algorithm for Biomedical Raman Spectroscopy. *Appl. Spectrosc.* **2007**, *61* (11), 1225–1232. <https://doi.org/10.1366/000370207782597003>.
  - (18) Pendleton, P.; Schumann, R.; Wong, S. H. Microcystin-LR Adsorption by Activated Carbon. *J. Colloid Interface Sci.* **2001**, *240* (1), 1–8. <https://doi.org/10.1006/jcis.2001.7616>.
  - (19) Sun, Q. The Raman OH Stretching Bands of Liquid Water. *Vib. Spectrosc.* **2009**, *51* (2), 213–217. <https://doi.org/10.1016/j.vibspec.2009.05.002>.
  - (20) Rivasseau, C.; Martins, S.; Hennion, M.-C. Determination of Some Physicochemical Parameters of Microcystins (Cyanobacterial Toxins) and Trace Level Analysis in Environmental Samples Using Liquid Chromatography. *J. Chromatogr. A* **1998**, *799* (1–2), 155–169. [https://doi.org/10.1016/S0021-9673\(97\)01095-9](https://doi.org/10.1016/S0021-9673(97)01095-9).