Journal of Raman Spectroscopy

WILEY

Journal of RAMAN SPECTROSCOPY

RESEARCH ARTICLE OPEN ACCESS

Autoencoding Raman Spectra to Predict Analyte Concentrations

Alex Poppe[1] | Charles Warren[1] | William Brooks[1] | Stuart Gibson[2] | Michael Foster[1]

[1]IS-Instruments Ltd., Tonbridge, UK | [2]School of Physics and Astronomy, University of Kent, Canterbury, UK

Correspondence: Alex Poppe (apoppe@is-instruments.com)

Received: 16 July 2024 | **Revised:** 27 May 2025 | **Accepted:** 4 June 2025

Funding: This research was funded by the Innovate UK, Eurostars—Eureka Network, grant number (E!12823).

Keywords: autoencoder | biopharmaceutical | deep UV Raman spectroscopy | machine learning regression | resonant Raman spectroscopy

ABSTRACT

Machine learning analysis has been applied to Raman data obtained in both nuclear and biopharmaceutical industrial applications. A 785-nm Raman instrument using a spatial heterodyne spectrometer (SHS) was used to acquire Raman spectra for the nuclear dataset, whilst a new deep UV resonant SHS system, featuring a 228.5-nm diode-pumped solid-state laser, was used to capture Raman spectra of biological macromolecule samples for the biopharmaceutical dataset. A key focus is on the practical challenges faced in the design of data processing tasks and machine learning architectures due to real-world limitations in data collection. A fully connected (FC) autoencoder is employed as part of a regression task, which generates predictions on analyte concentrations in mixed substances. The method was shown to outperform industry standard regression tools, principal component regression (PCR) and partial least squares (PLS) regression, each used as comparative benchmarks, by over 50% in a test of model precision across the nuclear and biopharmaceutical datasets investigated in this work. Advancements in the precision, speed and effectiveness of such tools are of critical importance in an industrial environment. This is driven by compelling motivations to reduce not only the costs associated with these processes but also to increase the quality of resulting products or to reduce the risks within industrial operations, where applicable.

1 | Introduction

Analysis of mixed chemical substances is a common problem across industries from purification within drug manufacture through to the identification of unwanted material in industrial settings. Raman instrumentation is ideally suited to this task; however, 'real world data' is often noisy and contaminated with unwanted light sources. Furthermore, these problems are often compounded by a scarcity of data. In this study, we examine the application of machine learning tools to identify and quantify chemical mixtures within two settings: nuclear clean up and biopharmaceutical drug manufacturing.

Nuclear facilities undergo a process known as postoperational clean out (POCO) at the end of an operational life cycle before decommissioning [1]. This is done to reduce potential risks and hazards, whilst simultaneously reducing the running costs associated with dismantling the redundant facility. The goal of POCO is to reduce the organic residue found on nuclear sites in vessels and pipework to an acceptable level so that they may be repurposed. This reduction is a crucial step as these sites may be inaccessible to humans due to unsafe levels of radiation in the surrounding area. If acceptable levels of reduction are not achieved, there is a heightened risk of potential fire hazards during the decommissioning process, for example, in the plasma cutting of stainless steel pipework. Therefore, being able to precisely identify and quantify concentrations of organic components holds immense importance, as this knowledge can significantly enhance cost efficiency in the safe cleanup of nuclear sites, with potential savings estimated

https://doi.org/10.1002/jrs.70003

by Sellafield Ltd. to be in the range of £10–£100 million per facility over the course of the lengthy POCO process, which can extend to upwards of 40 years.

In this work, tributyl phosphate (TBP) and odourless kerosene were chosen as 'worse-case scenario' by-products in POCO. These chemical compounds would exist within a wide range of pipes and vessels, and in a variety of forms such as bulk organics or films on aqueous surfaces, due to the use of these organic materials as 'reprocessing agents' during solvent extraction. TBP can be used as an extractant in nuclear chemistry, which serves as one of the liquid components in the separation of compounds in solution comprising two insoluble liquids [2]. This separation is based on the difference in solubility. Thus the nuclear database utilised in this work is composed of TBP dissolved in kerosene across a range of high concentrations.

With regard to the biopharmaceutical industry, drug manufacturing represents a costly undertaking. With a growing market size of over $300 billion in 2022, which is predicted to increase to over $850 billion by 2030 [3, 4], there is a strong financial incentive to improve manufacturing efficiency for drugs that are essential across numerous application areas. Such enhancements result in a decrease in waste products and an increase in drug quality due to the heightened assurance of component concentrations.

Protein biologics currently hold a position of focus in biopharmaceutical research; hence, this work includes two datasets of biomolecular compounds used in the production of monoclonal antibodies (mAbs), which are specialised drug therapies designed to target specific proteins in cancer cells through a liquid chromatography manufacturing process. The first is a macromolecule, immunoglobulin G (IgG), which is a common type of antibody found in humans [5]. The second is tryptophan, which is an amino acid used in protein synthesis [6]. Both molecules serve important roles in the human body, so they are relevant areas of study for biopharmaceutical industries in protein biologics research.

A major challenge present throughout this downstream process is a risk of protein aggregation [7–9], which affects the yield of the resulting protein products. Besides that, protein aggregates are connected to adverse immunogenicity [9]—the ability for a drug introduced into the body to produce an undesirable immune response—thus further emphasising the need for quality control during the manufacturing of mAbs. By designing monitoring tools to quantify aggregation levels, adjustments could be made to liquid chromatography processes to assist in maximising drug yield and quality.

Raman spectroscopy has been recognised as a potential monitoring tool. Nevertheless, conventional spontaneous Raman spectroscopy in the visible and near-IR wavelength regions, whilst being capable of measuring the high concentration nuclear dataset mentioned above, proves impractical for gathering high-quality data in this biopharmaceutical scenario, primarily due to significant levels of fluorescence that convolve with the Raman response to analyte proteins [10].

For the biopharmaceutical datasets, an ultraviolet resonance Raman spectroscopy (UVRRS) system [11–13], a form of RRS utilising a deep ultraviolet Raman probe laser, is used to overcome the challenge of fluorescence by operating at low wavelengths (below 250 nm). In this wavelength range, the analyte Raman and fluorescence responses become spectrally separated. In addition to this, the signal-to-noise ratio (SNR) of the resulting spectra is enhanced by two complementary factors. Firstly, as the Raman scattering intensity is proportional to λ^{-4}, the shorter wavelength laser used in the UVRRS system provides an increase in peak intensity. Secondly, as the laser wavelength is reduced, the electronic transition of many organic molecules containing conjugated structures is approached [13], which produces a resonant effect that amplifies the Raman response by several orders of magnitude [11, 14]. An additional factor to consider when measuring biopharmaceuticals within the deep UV is that samples can strongly attenuate. Typically, the Raman response is assumed to be linear with sample concentration, as attenuation can be assumed negligible (Equation 1); however, with the Deep UV sample, the exponential component of (Equation 1) becomes significant [15, 16].

$$S = \frac{L_p V(R) A O_e D_{QE} I(R) \Delta R \alpha N e^{-2\tau}}{\pi R^2} \tag{1}$$

Where S is the signal intensity, L_p is the laser power, V(R) is the overlap integral of the outgoing laser and the telescope field of view, A is the collecting telescope area, O_e is the instrument optical efficiency, D_{QE} is the detector efficiency, I(R) is the overlap integral between the outgoing and incoming beams, ΔR is the depth of the sample, α is the Raman scattering cross section of the analyte, N is the number of scattering centres, τ is the optical depth of the medium, and R is the distance to the target.

Two different Raman systems were used in this work, a spontaneous Raman spatial heterodyne spectrometer (SHS) operating at 785 nm for the nuclear industry dataset and a UVRRS SHS system operating at 228.5 nm for the biopharmaceutical industry dataset. The Raman spectra of mixed substances were measured and analysed using a machine learning feature extractor combined with a linear regression model, which was jointly trained on each dataset. Model design and implementation were a point of focus in this work, and the performance of these models was compared with industry standard methods: principal component regression (PCR) (constituting principal component analysis (PCA) and a linear regression tool) and partial least squares (PLS) regression in optimised scenarios. The efficacy of these methods is considered as potential complements to, or replacements for, existing analytical tools.

A key theme explored within this work is the effect of a small dataset size on the design and implementation of machine learning models. Small datasets are commonplace in many industrial settings [17], as a strong incentive is often required to invest time and money into producing a large dataset for exploratory research. Therefore, it is important to examine real-world applications where data availability is limited and consider ways to optimise the analysis of small datasets. Such considerations include the choice of neural network architecture and data preprocessing techniques [18], as well as the type and usage of data augmentation strategies [19].

There are a number of existing machine learning architectures trained for applications in processing mixed Raman spectra. However, the focus of these architectures is primarily on the identification of mixture components [20], or through the use of large datasets to pretrain neural networks for regression tasks [21], which is a process in contravention to the small datasets typically available in an industrial setting.

The limits of the selected data augmentation strategy are investigated by analysing the performance of the regression model based on selected modifications to the augmentation process. This test is carried out on a database of amino acid measurements from one of the biopharmaceutical datasets, which was measured at non-uniform concentration intervals. Previous measurements on these amino acids have shown them to have a Raman response that is non-linear with respect to concentration, due to higher concentration samples having a greater attenuation of the signal, particularly in the UV.

2 | Methodology

The machine learning model chosen for this regression task was a fully connected (FC) autoencoder, which was trained to fulfil the role of a non-linear data compression and feature extraction tool for the input mixture spectra. Once the autoencoder was trained, the compressed spectra were fed through a regularised linear regression model (ridge regression) in order to make predictions on the analyte concentration within each mixture. This work explores the effects of small database sizes on the choice, design and implementation of machine learning architectures, as well as the data preprocessing and augmentation techniques implemented to attempt to overcome this limitation.

2.1 | Data Acquisition and Preprocessing

The Raman spectra for the TBP chemical database were captured using a HES2000 spectrometer, an in-house setup developed at IS-Instruments. The spectrometer features an Andor iVac 316 FT detector, with 150-line/mm blazed diffraction gratings from Richardson. The 500-mW laser had a central wavelength of 785 nm. The exposure time was set to 30 s for each spectrum. A schematic for the SHS configuration [22, 23] is shown in Figure 1.

The Raman spectra for the IgG and tryptophan mixture databases were captured using a deep UV Raman spectroscopy system called Odin, which is another in-house setup developed at IS-Instruments. The spectrometer features an Andor iDus 420 FT detector, with 400-line/mm blazed diffraction gratings from Richardson, and with bespoke lenses and beam splitters designed for working at these Deep UV wavelengths. A 229-nm-long pass filter was used to suppress any laser light. A 9-mW laser was used at a central wavelength of 228.5 nm. The exposure time was set to 30 s for each spectrum. The arrangement of the Odin spectrometer is the same as the HES2000 spectrometer, as seen in Figure 1.

For the TBP nuclear dataset, there were nine concentrations measured with 128 repeats, ranging inclusively from 10% to 90% TBP dissolved in kerosene at uniform intervals. All spectra were interpolated to a wavenumber range of 100 to 2200 cm^{-1}, at a constant wavenumber resolution of 4.102 cm^{-1} using a cubic spline interpolation.

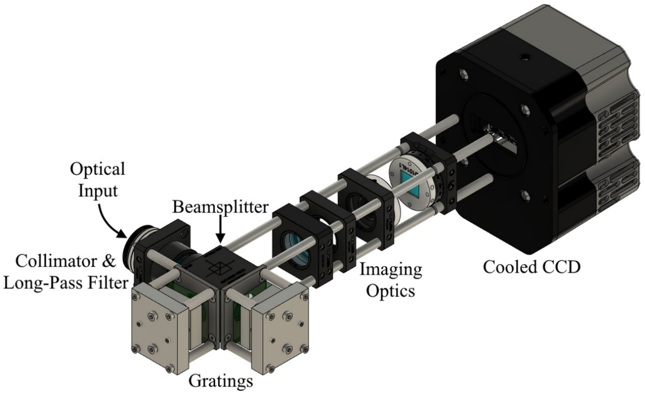

FIGURE 1 | Schematic for the SHS used to capture Raman measurements [18].

For the biopharmaceutical datasets, there were 10 repeat measurements per concentration, with 11 concentrations measured for IgG ranging from 0.1021 to 2.0173 mg mL^{-1}, and 17 concentrations measured for tryptophan ranging from 0.0127 to 5.0971 mg mL^{-1}. The biopharmaceutical datasets were interpolated using a cubic spline interpolation to separate wavenumber ranges of 600–2250 cm^{-1} at a constant wavenumber resolution of 3.223 cm^{-1} for the IgG dataset, and 570 to 2270 cm^{-1} at a constant wavenumber resolution of 2.969 cm^{-1} for the tryptophan dataset.

The interpolation resulted in spectra containing 512 bins for all three datasets. The repeat measurements for all datasets were then partitioned into the training, validation and testing datasets. This resulted in training datasets with population sizes of 918 for TBP, 110 for IgG and 170 for tryptophan. Therefore, due to the limited size of the datasets available, the application of data augmentation becomes crucial in order to introduce enough sample variability required for training a deep neural network effectively in the context of this regression task.

2.2 | Data Augmentation Strategy

Linearly weighted data augmentation was applied to increase the number of samples, and therefore the variance in each dataset, which is a vital step in training deep learning models [24]. The ability of a deep learning regression model to make accurate predictions on a target dataset is facilitated by data augmentation, which builds on the statistical diversity of the training dataset. Interstitial concentrations could be synthesised using this augmentation method, which were created from neighbouring discrete concentrations sampled from the respective raw datasets. To synthesise an augmented spectrum, three spectra were chosen at random from a data pool combining two neighbouring concentrations, which are assumed to possess a linear Raman relationship due to the small concentration difference between them. Each sample was then multiplied by a scaling

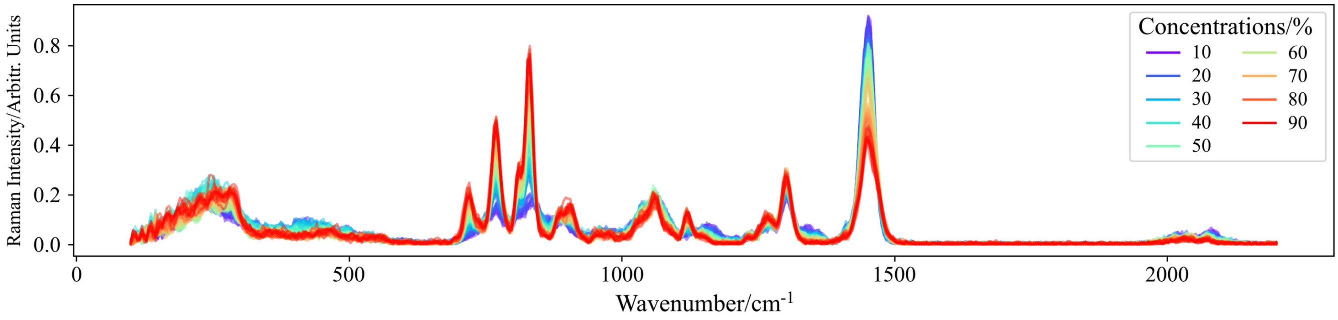

FIGURE 2 | Synthesised spectra of mixed TBP-kerosene through the entire inclusive range of concentrations.

coefficient and then linearly combined to produce the spectrum. These three coefficients were sampled from a Dirichlet distribution [25] with position concentration parameter, α, equal to 2.0 for all coefficients.

This data augmentation method allows for an arbitrary number of augmented samples to be synthesised. In total, there were 16,000 unique training samples synthesised every epoch, with 2000 fixed samples created before training for the validation and testing datasets. Once the spectra were synthesised, noise proportional to 10% of the mean intensity of each spectrum was then added to further increase sample variance, followed by L2-normalisation to remove intensity bias from the neural network during training. This method of noise addition was designed to be representative of the noise associated with the FT spectrometers used to capture each dataset for this work.

Figure 2 shows examples of mixed TBP-kerosene spectra synthesised by the data augmentation process throughout the entire concentration range. As the concentration of TBP increases with respect to kerosene, the main peaks between 500 and 1500 cm^{-1} become more intense, with the exception of the large peak before 1500 cm^{-1} that decreases with increasing TBP concentration, albeit to a nonzero count, as both TBP and kerosene have a characteristic Raman response around this wavenumber.

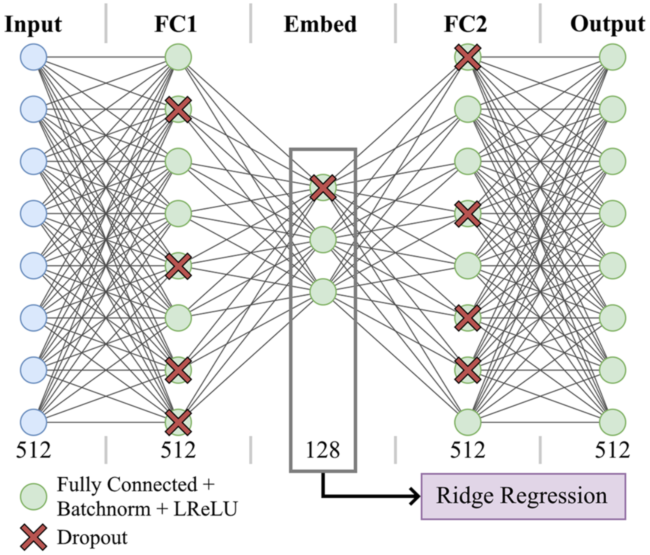

FIGURE 3 | Autoencoder block diagram. The effective size of the hidden layers during training is a fraction of each respective size (384) based on the dropout rate of 25%. Note that the batch size dimension is equal on all layers and is thus omitted. The embedding hidden layer of the trained autoencoder is used as an input vector for the ridge regression task to estimate mixture concentrations.

2.3 | Regression Model Architecture

To predict the concentration of a solution, salient features from each spectrum were extracted using an FC autoencoder, which was achieved by training the model to reconstruct input spectra. Using the embedding from this trained model, which contains the features required for reconstructing the data back to each input spectra, a separate regression model was trained for prediction. Ridge regression, a variation of linear least squares with an additional L2-normalised penalty term (typically used when data suffers from multicollinearity), was used as the regression model for this task. This regression model was fit to the low-dimensional embedded version of the same training data that was used to train the autoencoder. Similarly, the testing dataset partitioned for use by the autoencoder was used to evaluate the success of the ridge regression model, thereby evaluating the combined 'AE-Ridge' regression task.

The autoencoder used in this work contains five layers, including the input and output layers. There are two FC layers in the encoder, the last of which is the 128-unit embedding layer, which was used as an input for the decoder. The decoder mirrors the architecture of the encoder. The output of each hidden layer was normalised using batch normalisation, followed by a Leaky ReLU activation function with slope coefficient, α, of 0.3. Dropout was used as the last layer in each hidden block with a dropout rate of 0.25 to regularise the model during training. The model depth and size for each layer were determined through a grid search optimisation, minimising the mean squared error (MSE) loss. A block diagram for the AE-Ridge model is shown in Figure 3.

The model was trained for 5000 epochs, with a static learning rate of 0.01 and a batch size of 200 spectra. The loss function used to train the autoencoder was the MSE loss between the input and reconstructed spectra, and the Adam optimisation algorithm was used—with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-7}$—to adjust the model parameters during training. All trainable layers were regularised using L2 weight decay with a regularisation factor, γ, of 0.1, and clipnorm [26] was used to clip the calculated gradients to the maximum L2-norm value. Once the autoencoder was trained, concentration estimates were made by training the ridge regression model, which used the embedding vector of the autoencoder as input data and the synthesised concentrations as output target labels, as shown in Figure 3.

2.4 | Influence of Non-Uniform Concentrations on Dataset Design

For the data augmentation strategy, there is an assumption of linear scaling between neighbouring samples with small differences in concentration. Due to this, two tryptophan datasets were defined: one retaining the 5 mg mL^{-1} sample, termed Trypt-5; and the other discarding it, termed Trypt-2. This resulted from the closest neighbour to the 5 mg mL^{-1} sample being sufficiently separated in concentration as to reduce model performance should the former sample be retained. Figure 4 and Figure S1 show examples of synthesised spectra produced by the data augmentation process for the IgG and Trypt-5 datasets. The performances of these models are evaluated in Subsection 3.3, followed by an exploratory test to overcome performance issues in the AE-Ridge model trained on the Trypt-5 dataset—owing to decreased uniformity of measured concentrations—by modifying the data augmentation strategy.

3 | Results and Discussion

The performance of the combined AE-Ridge model is evaluated against industry standard regression tools PCR and PLS regression in Subsection 3.1 for the TBP nuclear dataset. This comparison is made using performance metrics commonly found in industrial settings: the coefficient of determination (R^2), the 95% prediction interval (PI), and the limit of detection (LoD). The complete set of results is provided for all regression tools, trained both with and without data augmentation, demonstrating the increase in performance of the combined AE-Ridge regression tool, aided by the data augmentation technique designed for this regression task, over the industry standard used within this work. Subsection 3.2 covers the results and evaluation of the biopharmaceutical datasets using the same metrics. Lastly, Subsection 3.3 investigates a modification made to the data augmentation strategy to account for the non-uniform nature of sampled concentrations in the tryptophan dataset, specifically by taking the natural logarithm of the discrete (unaugmented) concentrations before synthesising interstitial spectra, showing a drastic increase in the performance of regression models trained on both the Trypt 2 and Trypt-5 datasets.

3.1 | Results and Comparison to Industry Standard Methods for the Nuclear Dataset

The performance of the AE-Ridge regression model was compared against two standard regression methods used for concentration prediction: PCR and PLS regression. The latter of which is widely used in chemometrics and other similar areas of spectroscopic data processing [27–29]. PCR had two distinct processes: PCA for dimensionality-reduction and ridge regression, chosen to match the AE-Ridge regression model for a more comparable test, to make predictions on the concentrations of each sample.

Although it is typical to determine an appropriate number of latent components to retain based on a user-specified threshold, such as the first N latent components to contain at least 95% of the explained variance, the AE-Ridge regression model was compared to the best case scenario from both the PCR and PLS regression tasks to set a high benchmark. Sets of both models were trained based on the full range of retained latent components, meaning from one latent component to the maximum number of original components, 512. The model that achieved the lowest 95% PI score was selected as the best model, individually for each regression method and sample dataset, to be compared to the respective AE-Ridge model.

For a complete evaluation, all three regression methods were tested on the TBP datasets, both with and without data augmentation applied. The purpose of this was to determine whether the increase in performance in the AE-Ridge model was due to the non-linear features extracted by the neural network or through the data augmentation strategy. This was done in addition to selecting the optimal number of latest components to retain for each regression model. The results of these tests are shown in Table 1 for the TBP dataset.

The results of the concentration predictions on the TBP dataset using the three regression methods showcase the increase in performance of the AE-Ridge machine learning regression model over the industry standard alternatives when combined with data augmentation. Without the use of data augmentation, the AE-Ridge regression model still outperformed the

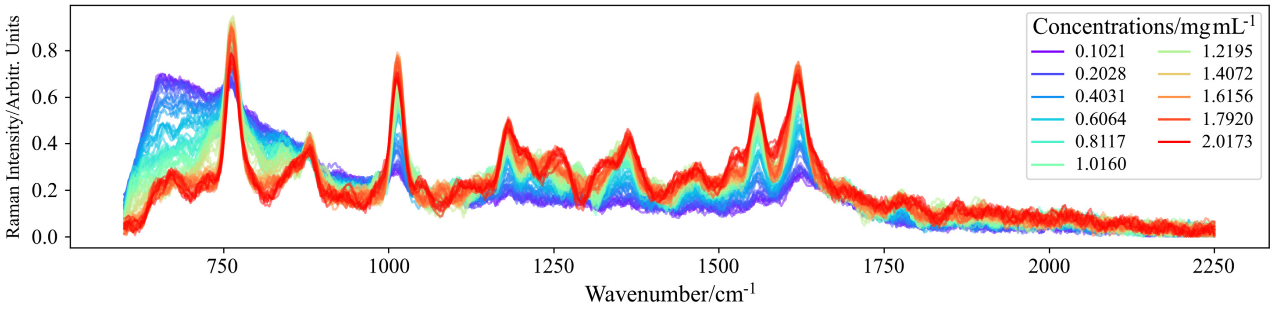

FIGURE 4 | Synthesised mixture spectra of IgG through the entire range of concentrations between approximately 0.1 to 2 mg mL^{-1}. The lower concentrations feature a large, broad water peak before 750 cm^{-1} that is suppressed as the concentration increases. Conversely, the intensity of the main IgG peaks increase as the concentration increases.

TABLE 1 | Results of evaluating the three regression methods trained on the TBP dataset, both with (augmented) and without (discrete) data augmentation.

	Discrete				Augmented			
Method	***N***	R^2	**95%PI (%)**	**LoD (%)**	***N***	R^2	**95%PI (%)**	**LoD (%)**
PCR	321	0.9909	±4.92	1.59	284	0.9964	±3.11	0.80
PLS	8	0.9918	±4.67	1.55	17	0.9963	±3.13	0.79
AE-Ridge	128	**0.9924**	**±4.51**	**1.48**	128	**0.9984**	**±2.04**	**0.75**

Note: The value *N* represents the number of latent components each dataset had after dimensionality-reduction. The best results are shown in bold.

TABLE 2 | Results of evaluating the three regression methods trained on the three mixtures datasets without data augmentation.

		Discrete			
Dataset	**Method**	***N***	R^2	**95%PI/mg mL^{-1}**	**LoD/mg mL^{-1}**
IgG	PCR	5	**0.9872**	**±0.1404**	0.1148
	PLS	3	0.9798	±0.1763	0.1326
	AE-Ridge	128	0.9804	±0.1738	**0.0838**
Trypt-2	PCR	66	0.7998	±0.6216	0.2790
	PLS	3	0.7917	±0.6340	0.2689
	AE-Ridge	128	**0.8383**	**±0.5585**	**0.1563**
Trypt-5	PCR	86	**0.5263**	**±1.6940**	0.8634
	PLS	3	0.5161	±1.7122	0.9000
	AE-Ridge	128	0.4253	±1.8658	**0.7834**

Note: The value *N* represents the number of components each dataset had after dimensionality-reduction. The best results are shown in bold.

alternative approaches in that category, although the results are closely comparable. It should be noted that all regression models benefitted from the use of data augmentation; however, the AE-Ridge model saw the greatest gain in performance. In particular, the value of the 95% PI metric for the AE-Ridge model trained with data augmentation achieved approximately 50% better performance over both PCR and PLS regression, which achieved similar results.

3.2 | Results and Comparison to Industry Standard Methods for the Biopharmaceutical Datasets

The complete set of permutations (mixture datasets, regression models, and data processing procedures) was tested. As before, the AE-Ridge model was retrained and reevaluated on the unaugmented mixtures datasets to fulfil this requirement. Tables 2 and 3 demonstrate the advantage of machine learning in the ability to learn complex, non-linear relationships. For Raman measurements in the IgG, Trypt-2 and Trypt-5 datasets, the peak heights did not scale linearly with concentration, primarily due to attenuation of the Raman signal at DUV wavelengths. The ability of AE-Ridge to learn that non-linearity has resulted in substantial improvements in the 95% PI score.

The necessity for a representative data augmentation procedure is also highlighted, as the AE-Ridge model substantially outperforms the PCR and PLS regression models on all three datasets. Interestingly, data augmentation improved the performance of all regression models across all datasets, with the exception of PLS regression trained on the Trypt-5 dataset that decreased in R^2 and 95% PI performance. This could be attributed to the inability of the data augmentation procedure to accurately synthesise interstitial concentrations between measured concentrations with a large separation, as in the case for the 2- and 5-mg mL^{-1} samples, but retaining an improvement to the LoD alongside the other regression models.

For the IgG dataset, the AE-Ridge model was outperformed by PCR on the discrete data with the exception of LoD, but succeeded PCR in all metrics when trained on augmented data—particularly on the 95% PI metric, which shows an approximate 50% improvement. With regard to the tryptophan datasets, the performance of all regression models was lower than the IgG counterparts owing to the increased non-uniformity of the measured concentrations. When trained on the Trypt-2 dataset, the AE-Ridge model outperformed both PCR and PLS regression regardless of the use of data augmentation—though the performances of all models were improved by its inclusion. Whereas, when trained on the Trypt-5 dataset, the AE-Ridge model only surpassed the other methods with the inclusion of the data augmentation procedure. This suggests that the AE-Ridge model is limited in the ability to learn non-uniform separations in data. The results demonstrate that data augmentation enables machine learning models to gain a more significant improvement in performance than both PCR and PLS regression. Additionally, the lack of data augmentation on the Trypt-5 discrete dataset, and the subsequent poor performance, emphasises the fact that most machine learning models—in particular deep learning models—are data hungry.

3.3 | Effects of Modifying Data Augmentation Process on Model Performance

The effects of data augmentation when applied to machine learning are showcased in Figure 5, in combination with the results in Tables 2 and 3 in the previous section. The range of predicted concentrations is both more accurate and precise due to data augmentation, which benefits from the near-uniform separation of measured concentrations.

As mentioned, the data augmentation procedure benefits from samples being measured at uniformly distributed concentrations. The trendlines plotted for both the discrete and augmented variant models highlight the improvement to each regression model with the exclusion of the outlier sample at around 5 mg mL^{-1} (see Figure S2). Additionally, regardless of the data augmentation procedure implemented, the raw data must be faithful to the concentration that it is labelled to represent. As the data augmentation procedure synthesises interstitial concentrations in a linear fashion, any flaws in the raw data are represented alongside the desired distinguishing features. This effect is seen in Figure S2b as the data points around the 1-mg mL^{-1} sample concentration feature a low-frequency oscillation. The AE-Ridge model trained on the Trypt-2 augmented dataset displays a minor plateauing of predicted concentrations beyond approximately 1.5 mg mL^{-1}, suggesting that the regression model has learned a data representation based partially off of a simple peak-height estimation, alongside the non-linear nature of the signal intensity.

The data augmentation method employed benefits greatly from the raw dataset being measured at uniformly spaced

TABLE 3 | Results of evaluating the three regression methods trained on the three mixtures datasets, both with data augmentation.

		Augmented			
Dataset	**Method**	***N***	***R*²**	**95%PI/mg mL^{-1}**	**LoD/mg mL^{-1}**
IgG	PCR	6	0.9890	±0.1303	0.0340
	PLS	14	0.9872	±0.1406	0.0287
	AE-Ridge	128	**0.9951**	**±0.0871**	**0.0179**
Trypt-2	PCR	311	0.9162	±0.4007	0.1119
	PLS	16	0.8588	±0.5219	0.1176
	AE-Ridge	128	**0.9758**	**±0.2126**	**0.0534**
Trypt-5	PCR	11	0.6503	±1.4555	0.3945
	PLS	29	0.4821	±1.7713	0.4470
	AE-Ridge	128	**0.7221**	**±1.2975**	**0.3090**

Note: The value *N* represents the number of components each dataset had after dimensionality-reduction. The best results are shown in bold.

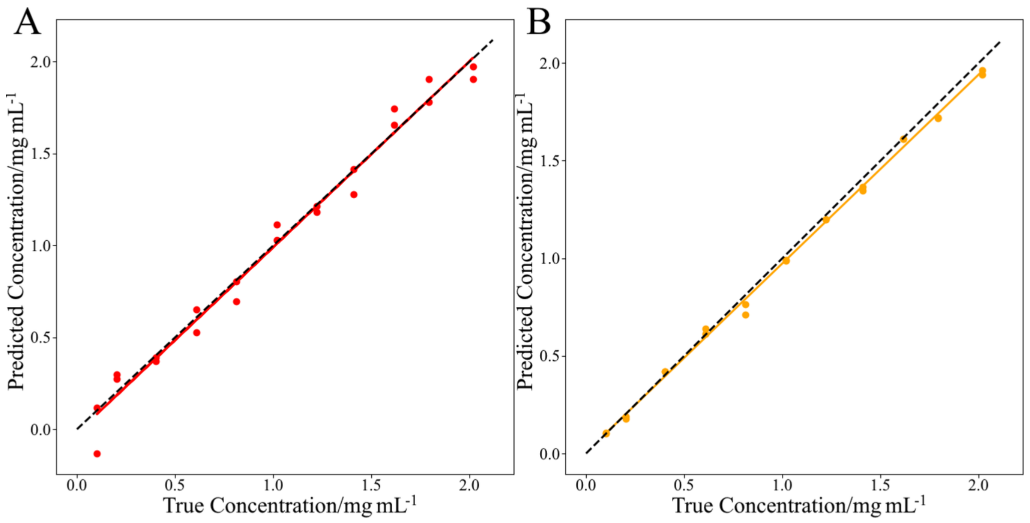

FIGURE 5 | Concentration predictions using the AE-Ridge regression model on the IgG dataset both with and without data augmentation. The dashed black lines show the identity line, and the orange and red lines show the trendlines for the plotted data points. Regression metrics are shown in Tables 2 and 3. (a) IgG discrete model test predictions. (b) IgG augmented model test predictions.

concentrations. The effect of this is demonstrated in Subsection 3.2, Tables 2 and 3 by the improved model performance of the IgG dataset in comparison to the two tryptophan datasets that possess notably worse data uniformity. An exploratory test was performed to improve model performance on both the Trypt-2 and Trypt-5 augmented datasets through a modification to the data augmentation process.

The respective models were retrained on the same data as before, but with the natural logarithm applied to all associated labels (concentrations) at the data augmentation stage, which were used throughout the training and inference processes, and only exponentiated back to the true values in order to convert concentration predictions into the desired units. This conversion shifted the measured concentrations closer towards uniform intervals (see Figure S3), which had the effect of greatly improving model performance, owing to the benefit of uniform data sampling to the data augmentation strategy.

Table 4 shows an improvement to the Trypt-2 augmented dataset, and a drastic improvement to the Trypt-5 dataset, in terms of the R^2 metric, by utilising the natural logarithm of the data labels at the data augmentation stage. The LoD performance saw an approximate 20% improvement on the Trypt-2 dataset and a 30% improvement on the Trypt-5 dataset, due to the increased data uniformity. However, the 95% PI metric remained in close proximity between 'standard' and 'logarithmic' versions of each tryptophan regression model, despite the large improvement to the R^2 value. Due to the mathematical nature of this procedure and the increased density of low concentration samples that were measured, the resulting models are more accurate at predicting samples with a lower concentration.

As the concentration of a sample increases, the model begins to diverge in its predictive capability, as can be seen in Figure 6, in which the line of best fit diverges from the identity line at higher concentrations (though this effect is harder to see in Figure 6b owing to the difference in scale). The nature of this divergence would explain the retention of the high 95% PI values. Despite this, the results demonstrate a clear improvement in the performance of the tested regression models owing to the increased uniformity in the sample data—placing the performance of the 'logarithmic' Trypt-2 model, with data augmentation, closer in line with that of the IgG counterpart model (see Table 3). This method may also allow for the incorporation of the outlier sample around 5 mg mL^{-1}, providing that the reduction in model performance is acceptable in the bounds of the particular regression task, which heavily depends on the importance of the outlier samples.

This result emphasises the advantage to the data augmentation strategy, and by extension, its positive impact on the performance of regression models, of measuring samples at uniform concentration intervals. Alternatively, specialised adaptations to online processing techniques, like the one outlined here, can alleviate the need for such rigid data collection prerequisites, which would otherwise impose impractical constraints from an industrial standpoint.

TABLE 4 | Results of training the Trypt-2 and Trypt-5 datasets by taking the natural logarithm of the labels (concentrations), in comparison to the standard values. The values for each metric were exponentiated back to mg mL^{-1} for the comparison. The value N represents the number of components each dataset had after dimensionality-reduction. The best results are shown in bold.

	Standard labels				Logarithmic labels			
Dataset	***N***	***R*²**	**95%PI/mg mL^{-1}**	**LoD/mg mL^{-1}**	***N***	***R*²**	**95%PI/mg mL^{-1}**	**LoD/mg mL^{-1}**
Trypt-2	128	0.9758	±0.2126	0.0534	128	0.9932	±0.1835	0.0456
Trypt-5	128	0.7221	±1.2975	0.3090	128	0.9764	±1.2666	0.2346

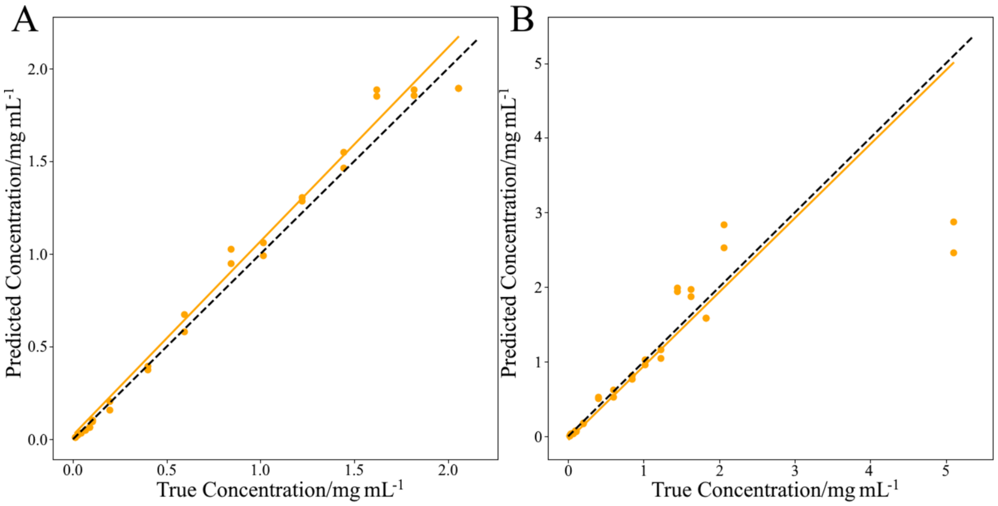

FIGURE 6 | Concentration predictions using the AE-Ridge regression model on the Trypt-2 and Trypt-5 augmented datasets, trained with logarithmic concentration labels. The dashed black lines show the identity line, and the orange lines show the trendlines for the plotted data points. Regression metrics are shown in Table 4. (a) Trypt-2 logarithmic model predictions. (b) Trypt-5 logarithmic model predictions.

4 | Conclusion

A competitive machine learning regression tool was developed that outperformed two industry standard tools, PCR and PLS regression, which were used as comparative benchmarks within this work. By separately training the AE-Ridge regression model on datasets from both nuclear and biopharmaceutical industries, it was proven that the model could effectively adapt across a wide variety of data, with different scales for the dependent variable. Sample concentrations in the nuclear database were represented as a decimal percentage, and hence bound between 0 and 1, whereas sample concentrations in the biopharmaceutical databases were measured in units of $\mathrm{mg\,mL^{-1}}$, which is an unbounded scale ranging from 0 to positive infinity.

In addition to different dependent variables, the databases differed in other aspects, such as varying magnitudes in concentration, different Raman spectroscopy measurement methods (spontaneous and RRS) and analyte molecule sizes. In particular, amplified non-linear Raman spectra were measured for both biopharmaceutical databases due to a combination of RRS and the size of the macromolecules, which caused greater sample attenuation, meaning that the exponential term of the LIDAR equation [15, 16] could no longer be ignored like it can for spontaneous Raman spectroscopy.

The AE-Ridge model achieved an approximate 50%–90% improvement on the 95% PI metric between the TBP with odourless kerosene nuclear dataset and IgG and tryptophan biopharmaceutical datasets. In terms of R^2 performance, values ranged between 0.975 and 0.995 for all datasets when trained on synthesised data and with modifications to the sampling strategy (specifically for the tryptophan datasets). Lastly, the AE-Ridge model improved upon the LoD performance of the other regression models by approximately 5% for the nuclear dataset and substantially greater at 30%–110% between the biopharmaceutical datasets.

These performance gains were facilitated through the use of a data augmentation strategy to increase sample variance during training. A linear relationship was assumed between neighbouring sample concentrations for the data augmentation strategy, which proved effective at predicting the concentrations of the TBP and odourless kerosene dataset due to the linear Raman response and the uniform intervals of concentration measurements.

As for the biopharmaceutical datasets, the AE-Ridge model performed better on all metrics on the IgG dataset over the tryptophan dataset, primarily owing to greater sample uniformity. This is verified by large performance gains in the tryptophan data when the sampling strategy used in data augmentation was modified to provide a more uniform dataset. There was an approximate 2% increase in the R^2 metric on the Trypt-2 dataset versus a 35% increase on Trypt-5, and the LoD further improved by 17% on Trypt-2 versus a 32% increase on Trypt-5, showing that the modification naturally improves datasets with worse sample uniformity to a greater extent than those with better.

Careful consideration must be given to the nature of any modifications made to the data augmentation sampling strategy, as the accuracy of the Trypt-5 trendline begins to diverge at greater concentrations as a result of the exponential used in the modification. This result shows potential for the inclusion of outlier samples within an arbitrary dataset without adversely affecting the performance of standard samples. However, modifications to either the regression model or the data augmentation technique would be required in order to improve predictions made on the outlier samples themselves. Although the non-linear Raman response from the organic macromolecules used in the biopharmaceutical datasets may place a limit on such potential.

Whilst each regression model was trained on datasets of liquid samples, this method could be applied to any mixed sample, such as gases or mixed solid particles—given small enough particle sizes that homogeneity could be assumed. There is already a precedent set for including multiple states of matter inside chemical databases within POCO procedures, as it is common to encounter residual organics in the form of bulk substances or vapours. Consequently, there is a need for future research aimed at adapting to such multiphase databases.

The improvements made by the AE-Ridge model would translate to tangible benefits for both the nuclear and biopharmaceutical industries and to other industrial applications not seen in this work. For the nuclear sector, in particular POCO, the increased certainty of sample concentrations over competing industry standard regression tools would enable cost reductions in the context of risk assessments. This could also contribute to maximising drug quality and yield in the biopharmaceutical industry, as improvements in quality control measurements during the manufacturing of mAbs would help to prevent adverse effects resulting from undetected protein aggregation, and improved yield would translate into reduced costs in the rapidly growing market.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data are not available for public access.

References

1. NDA, "Nuclear Decommissioning Authority Strategy," 2016.

2. C. Zidi, R. Tayeb, M. B. S. Ali, and M. Dhahbi, "Liquid–Liquid Extraction and Transport Across Supported Liquid Membrane of Phenol Using Tributyl Phosphate," *Journal of Membrane Science* 360, no. 1–2 (2010): 334–340.

3. T. J. Seymour, P. Crawford, and D. M. Ecker, *The Therapeutic Monoclonal Antibody Product Market* (BioProcess International, 2020).

4. "Biopharmaceuticals Market Size to Hold USD 856.1 Bn by 2030". BioSpace, 2022. Accessed: 13th June 2024.

5. V. Irani, A. J. Guy, D. Andrew, J. G. Beeson, P. A. Ramsland, and J. S. Richards, "Molecular Properties of Human IgG Subclasses and Their Implications for Designing Therapeutic Monoclonal Antibodies Against Infectious Diseases," *Molecular Immunology* 67, no. 2 (2015): 171–182.

6. Z. Wei, J. Feng, H. Y. Lin, et al., "Identification of a Single Tryptophan Residue as Critical for Binding Activity in a Humanized Monoclonal Antibody Against Respiratory Syncytial Virus," *Analytical Chemistry* 79, no. 7 (2007): 2797–2805.

7. A. A. Bana, N. Sajeev, S. Halder, H. A. Masi, S. Patel, and P. Mehta, “Comparative Stability Study and Aggregate Analysis of Bevacizumab Marketed Formulations Using Advanced Analytical Techniques,” *Heliyon* 9, no. 9 (2023): e19478, https://doi.org/10.1016/j.heliyon.2023.e19478.

8. K. L. Zapadka, F. J. Becher, A. L. dos Gomes Santos, and S. E. Jackson, “Factors Affecting the Physical Stability (Aggregation) of Peptide Therapeutics,” *Interface Focus* 7, no. 6 (2017): 20170030, https://doi.org/10.1098/rsfs.2017.0030.

9. A. A. K. Bana, P. Mehta, and K. A. K. Ramnani, “Physical Instabilities of Therapeutic Monoclonal Antibodies: A Critical Review,” *Current Drug Discovery Technologies* 19.6 (2022): 1–11.

10. M. K. Shukla, P. Wilkes, N. Bargary, et al., “Identification of Monoclonal Antibody Drug Substances Using Non-Destructive Raman Spectroscopy,” *Spectrochimica Acta, Part A: Molecular and Biomolecular Spectroscopy* 299 (2023): 122872, https://doi.org/10.1016/j.saa.2023.122872.

11. M. Foster, W. Brooks, P. Jahn, J. Hedberg, A. Andersson, and L. Ashton, “Demonstration of a Compact Deep <>UV</scp> Raman Spatial Heterodyne Spectrometer for Biologics Analysis,” *Journal of Biophotonics* 15, no. 7 (2022): e202200021, https://doi.org/10.1002/jbio.202200021.

12. L. Ashton and R. Goodacre, "Application of Deep UV Resonance Raman Spectroscopy to Bioprocessing," *European Pharmaceutical Review* 16 (2011): 3.

13. E. Smith and G. Dent, *Modern Raman Spectroscopy: A Practical Approach*, 2nd ed. (John Wiley & Sons, Ltd, 2005).

14. M. D. Morris and D. J. Wallan, "Resonance Raman Spectroscopy," *Analytical Chemistry* 51 (1979): 2.

15. M. Foster, M. Wharton, W. Brooks, M. Goundry, C. Warren, and J. Storey, "Remote Sensing of Chemical Agents within Nuclear Facilities using Raman Spectroscopy," *Journal of Raman Spectroscopy* 51 (2020): 12.

16. S. P. Burton, M. A. Vaughan, R. A. Ferrare, and C. A. Hostetler, "Separating Mixtures of Aerosol Types in Airborne High Spectral Resolution Lidar Data," *Atmospheric Measurement Techniques* 7 (2014): 2.

17. J. J. Faraway and N. H. Augustin, “When Small Data Beats Big Data,” *Statistics & Probability Letters* 136 (2018): 142–145.

18. J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon, and S. J. Gibson, “Deep Convolutional Neural Networks for Raman Spectrum Recognition: A Unified Solution,” *Analyst* 142 (2017): 4067–4074.

19. Y. Wu, J. Liu, Y. Wang, S. J. Gibson, M. Osadchy, and Y. Fang, "Reconstructing Randomly Masked Spectra Helps DNNs Identify Discriminant Wavenumbers," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46 (2024): 3845–3861, ISSN 0162-8828.

20. L. Pan, P. Pipitsunthonsan, C. Daengngam, S. Channumsin, S. Sreesawet, and M. Chongcheawchamnan, “Identification of Complex Mixtures for Raman Spectroscopy Using a Novel Scheme Based on a New Multi-Label Deep Neural Network,” *IEEE Sensors Journal* 21, no. 9 (2021): 10834–10843, https://doi.org/10.1109/JSEN.2021.3059849.

21. W. J. Thrift and R. Ragan, “Analytical Chemistry,” 91 (2019): 21.

22. J. Harlander, R. J. Reynolds, and F. L. Roesler, “Spatial Heterodyne Spectroscopy for the Exploration of Diffuse Interstellar Emission Lines at Far-Ultraviolet Wavelengths,” *Astrophysical Journal* 396 (1992): 730, https://doi.org/10.1086/171756.

23. M. J. Foster, J. Storey, and M. A. Zentile, “Spatial-Heterodyne Spectrometer for Transmission-Raman Observations,” *Optics Express* 25, no. 2 (2017): 1598, https://doi.org/10.1364/OE.25.001598.

24. A. Mikołajczyk and M. Grochowski, "Data Augmentation for Improving Deep Learning in Image Classification Problem," *IIPhDW* (2018): 117–122.

25. D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, 2003).

26. R. Pascanu, T. Mikolov, and Y. Bengio, "On the Difficulty of Training Recurrent Neural Networks," *PMLR* 28 (2013): 3.

27. Y. Yu, Y. Lin, C. Xu, et al., "Label-Free Detection of Nasopharyngeal and Liver Cancer using Surface-Enhanced Raman Spectroscopy and Partial Lease Squares Combined with Support Vector Machine," *Biomedical Optics Express* 9 (2018): 12.

28. R. M. Balabin and E. I. Lomakina, “Support Vector Machine Regression (SVR/LS-SVM)—An Alternative to Neural Networks (ANN) for Analytical Chemistry? Comparison of Nonlinear Methods on Near Infrared (NIR) Spectroscopy Data,” *Analyst* 136 (2011): 1703–1712.

29. L. Zhang, Q. Li, W. Tao, B. Yu, and Y. Du, “Quantitative Analysis of Thymine With Surface-Enhanced Raman Spectroscopy and Partial Least Squares (PLS) Regression,” *Analytical and Bioanalytical Chemistry* 398 (2010): 1827–1832.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.