Explainable Deep Learning Framework for SERS Bio-quantification

Jihan K. Zaki
Yusuf Hamied Department of Chemistry
University of Cambridge
Lensfield Rd, CB2 1EW

Jakub Tomasik
Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive, CB3 0AS

Jade A. McCune
Yusuf Hamied Department of Chemistry
University of Cambridge
Lensfield Rd, CB2 1EW

Sabine Bahn
Department of Chemical Engineering and Biotechnology
University of Cambridge
Philippa Fawcett Drive, CB3 0AS

Pietro Liò *
Department of Computer Science and Technology
University of Cambridge
15 JJ Thomson Ave, CB3 0FD
pl219@cam.ac.uk

Oren A. Scherman *
Yusuf Hamied Department of Chemistry
University of Cambridge
Lensfield Rd, CB2 1EW
oas23@cam.ac.uk

Abstract

Surface-enhanced Raman spectroscopy (SERS) is a potential fast and inexpensive method of analyte quantification, which can be combined with deep learning to discover biomarker-disease relationships. This study aims to address present challenges of SERS through a novel SERS bio-quantification framework, including spectral processing, analyte quantification, and model explainability. To this end, serotonin quantification in urine media was assessed as a model task with 682 SERS spectra measured in a micromolar range using cucurbit[8]uril chemical spacers. A denoising autoencoder was utilized for spectral enhancement, and convolutional neural networks (CNN) and vision transformers were utilized for biomarker quantification. Lastly, a novel context representative interpretable model explanations (CRIME) method was developed to suit the current needs of SERS mixture analysis explainability. Serotonin quantification was most efficient in denoised spectra analysed using a convolutional neural network with a three-parameter logistic output layer (mean absolute error=$0.15\,\mu\mathrm{M}$, mean percentage error=4.67%). Subsequently, the CRIME method revealed the CNN model to

*Corresponding author

Machine Learning and the Physical Sciences, NeurIPS 2024.

present six prediction contexts, of which three were associated with serotonin. The proposed framework could unlock a novel, untargeted hypothesis generating method of biomarker discovery considering the rapid and inexpensive nature of SERS measurements, and the potential to identify biomarkers from CRIME contexts.

1 Introduction

Surface-enhanced Raman spectroscopy (SERS) is rapidly gaining attention as a potential fast and inexpensive method of biomarker quantification. Broadly described, SERS signals capture vibrational modes unique to chemical bonds within molecules, enabling the identification and quantification of specific chemical analytes. The technique capitalizes on optical 'hot spots', i.e. localized regions of intense optical fields, created by the aggregation of noble metal nanoparticles[5, 8]. These nanoparticles offer a robust platform for in situ analysis within liquid media, rendering SERS a practical choice for broad applications. Combined with deep learning methodology, SERS could be utilized to elucidate complex biomarker-disease relationships.

There are a number of challenges that need to be solved to enable robust SERS-based analyte detection. Primarily, the reproducibility and readability of spectra is limited due to biological sample variability and noise[10]. Currently, standard practices in SERS analysis are substantially behind the state-of-the-art machine learning approaches; however, present challenges of SERS analysis could be effectively addressed with a robust computational framework. Additionally, there is a particular need for improved model explainability in SERS analysis. Current general explainability methods are capable of providing both global or individual explainations. However, SERS data often involve multiple overlapping signals and confounding factors that can influence model predictions in subtle ways. Therefore, there is a need for methods that can uncover these underlying contexts, helping to differentiate between true biomarker signals, confounders, and noise. This study aims to extend the applicability of SERS in three main directions. First, it seeks to computationally mitigate inherent biological and method-based variations in SERS. Second, it aims to explore deep learning models for more accurate targeted analyte quantification, and lastly, to propose a domain-specific model explainability method capable of identifying prediction contexts in the chemical spectral analysis.

2 Methods

2.1 Dataset Preparation

The dataset assessed consisted of 318 SERS spectra measured in a lyophilized urine medium and 364 SERS spectra measured in a water medium using a $785\,\mathrm{nm}$ laser and cucurbit[8]uril (CB[8]) spacers ($0.9\,\mathrm{nm}$) with $60\,\mathrm{nm}$ gold nanoparticles. Both urine- and water-medium samples were spiked with epinephrine, dopamine, and serotonin, with concentrations ranging from 0 to $9\,\mu\mathrm{M}$. Serotonin was used as the target analyte. This was due to the lyophilized urine medium containing varying endogenous concentrations of epinephrine and dopamine, resulting in their unknown absolute concentrations. The measured SERS spectra were one-dimensional vectors containing 937 values per spectrum, and were shortened to feature a relevant range of Raman shifts from 300 to 2000 cm^{-1}, for a total of 842 values. The specific concentrations of the neurotransmitters in each sample are presented in **Table 1** in the appendix. Prior to assessment, spectra were processed using the asymmetric least squares (ALS) algorithm [3] for baseline correction with pre-specified parameters (λ=1000, p=0.1, n=10). Following the correction, the data was normalized to intensities between 0 and 1.

2.2 Neural network models

Following baseline correction and normalization, the spectra were denoised using a denoising autoencoder. A simple denoising autoencoder was trained using full water-medium spectra consisting of 364 spectra with 937 values using an 80:20 train-test split, with the urine background samples incorporated as noise. Noisy data were generated using urine background data (Table 1: sample U), where a randomly selected measurement was overlayed to each clean spectra following baseline correction but before normalization. The quality and utility of denoised spectra was subsequently evaluated through effects on performance in quantification models.

Two model types were applied to analyse the spectra, namely the convolutional neural network (CNN) [6] and the vision transformer (ViT) [2], with custom SERS-specific layers evaluated for the CNN. Both the CNNs and the ViT model were implemented in TensorFlow. A core architecture used in all CNN models comprised of rectified linear unit (ReLu) and hyperbolic tangent (Tanh)-ReLu paired 1-D convolution layers. Three variants of the CNN were evaluated with varying additional custom layers. These included the base CNN with a linear final activation output layer (CNN_L), a CNN with a modified custom Three-Parameter Logistic (3PL) activation function (CNN_{3PL}), and a CNN model with inherent scale-adjusting capabilities through scaling layers (sCNN). The architectures of the quantification models and the custom layers are described in detail in the appendix.

Each model was evaluated in both the raw and denoised data incorporating unseen spectra and repeat spectra. Spectra were defined as repeat if separate measurements of the specific sample were used in training either the denoising autoencoder or the quantification models. Repeat spectra were split to training and validation sets with a 90:10 split, and furthermore measurements taken from an unseen serotonin free sample (sample F) were included in the validation set. The remaining unseen samples (D, and E) were included exclusively in the test set. Final spectra counts for both raw and denoised datasets consisted of 218 training spectra, with a validation set of 46, and a test set of 54 spectra.

2.3 Context representative interpretable model explanations

The reliability and explainability of the final quantification model were assessed using the Context Representative Interpretable Model Explanations (CRIME) framework, developed in this study for machine learning interpretations of data with expected contextual prediction clusters. The CRIME framework expands on the widely applied local interpretable model agnostic explanations (LIME) framework[9], by assessing model explanations through contexts. Contexts can be defined within this framework as prominent and consistent prediction explanations across a number of prediction instances. The CRIME framework attempts to identify all prediction contexts of the input data space through the latent space of a variational autoencoder (VAE) trained on the LIME predictions of all instances in the available data. The LIME predictions are flattened with regards to perturbation limits and weights prior to input and are subsequently projected to the two-dimensional latent space. Details regarding the VAE of the CRIME method are described in the appendix. Following training of the VAE, the latent space is utilized to identify context clusters representing all the possible ways in which the quantification model interprets the input data. The latent space instances are clustered into the final contexts using K-means clustering, and the latent space is visually inspected for selecting the number of clusters. Finally, a mean LIME explanation is assessed through averaging all instances in each cluster to represent the contexts. To identify the defining features of each context representation, normalized LIME feature weights are combined with mean feature values representing the spectral intensities within the context clusters. They are then set in a three-dimensional space, together with normalized feature positions, which are then further clustered into 15 clusters using K-means clustering. Following the clustering of mean spectral feature values, position z-scores, and LIME weights, the clusters are ordered according to the product of their LIME weights and spectral intensities. The five clusters with the highest score are selected to represent the regions of the spectra which contribute most to the contextual predictions.

Following the identification of the most relevant context prediction regions, the highlighted regions of the mean context spectra are assessed against measured clean spectra of the neurotransmitters known to be present in the mixture. To emphasize the explanation weights in the spectra, both the reference clean spectra and the mean context spectra are scaled according to the explanation weights in the specific feature location. To determine the cause or identity of the recognised context clusters, the final mean context indicators are compared to the weighted reference spectra using cosine similarity S_c.

2.4 Benchmarking

The utility of the denoising autoencoder was assessed by measuring performance in the raw and denoised data and additionally, comparing it with fifth order polynomial second Savitzky-Golay derivative processing with a window length of 33[4], which was previously assessed as a reference standard. The developed CNN architecture was benchmarked against simpler architectures without Tanh-ReLu pairing layers. Additionally, the best-performing quantification model was assessed against methods used previously to quantify neurotransmitter concentrations, as well as other machine

learning methods, including partial least squares regression (PLSR), random forests, support vector machines (SVM), and extreme gradient boosting (XGBoost). The primary benchmark for serotonin quantification accuracy is the previous study by Kasera *et al.* 2014. Hyperparameter tuning was determined using grid search and 3-fold cross validation. The searched parameters are included in the appendix. For comparison with CRIME, feature importance and model explainability was assessed using Logic Explained Networks (LEN)[1] and Shapley Additive Explanations (SHAP)[7]. Details of the SHAP and LEN implementation are described in the appendix.

3 Results

3.1 Quantification models

Four different neural network models were evaluated in both raw and denoised datasets. Test set results showed strong performance in the denoised dataset with the CNN_{3PL} model (mean absolute error (MAE)=$0.15\,\mu\text{M}$, mean percentage error (MPE) =4.67%) and the sCNN model (MAE=$0.11\,\mu\text{M}$, MPE=3.52%) outperforming both the ViT model (MAE=$0.30\,\mu\text{M}$, MPE=8.09%) and the CNN_L model (MAE=$0.30\,\mu\text{M}$, MPE=7.45%). Performance of all neural network models in both denoised and raw datasets is visualised in **Figure 1**.

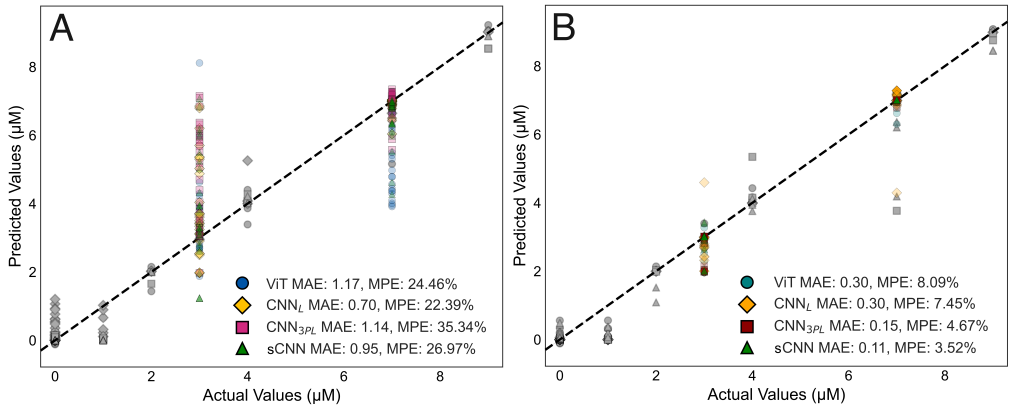

Figure 1: **Results of the final models in the validation and test sets for the four model types in both raw (A) and denoised datasets (B).** Validation set results are shown in grey, and test set results are shown in color: the linear CNN model is shown in yellow (diamond), the vision transformer model in blue (circle), the scale-adjusting CNN in green (triangle), and the three-parameter logistic output layer CNN model in red (square). The listed values were obtained from the final test set. Validation set results are presented in Table 2 in the appendix. MAE = mean absolute error, MPE = mean percentage error.

3.2 CRIME framework

The CRIME framework was fit on a VAE using the LIME explanations of the CNN_{3PL} predictions. This setting was selected due to its strong performance across both the validation and test data. Following the K-means clustering, the latent space of the VAE was clustered into six contexts. Among them, four distinct contexts were identified (contexts A, B, C, and F), as well as one intermediate context (context E), and one outlier context (context D). The mean LIME explanations for each CRIME context cluster are presented in **Figures 2A** to **2F** in the appendix. The peak-region cluster plots are visualised in **Figures 3A to 3F**. Contexts A (S_{cos}=0.87), E (S_{cos}=0.46), and F (S_{cos}=0.54) were correctly associated with serotonin, while contexts B (S_{cos}=0.98) and C (S_{cos}=0.97) were associated with dopamine and epinephrine, respectively. Complete cosine similarity values between mean CRIME context spectra and reference neurotransmitter spectra are presented for each CRIME context in **Table 3** in the appendix.

3.3 Benchmarking

The neural network models were benchmarked between different architectures and against other machine learning models, with the results summarized in **Table 4** and **Figure 4** in the appendix. The PLSR model trained with autoencoder-denoised spectra showed the best performance amongst non-neural network-based models with MAE=$0.70\,\mu\mathrm{M}$. The benchmarking results of the CNN architectures are presented in **Table 5** in the appendix. To compare the context explanations to methods of global explainability, LEN and SHAP frameworks were evaluated as a reference standard. The mean feature activations of the CNN_{3PL} model across all layers is presented in **Figure 5** in the appendix. The LEN identified logic statements explaining the four categories of serotonin concentrations with fair (0.69, no serotonin) to excellent (0.98, medium serotonin) explanation accuracy. The logic statements are visualized in **Figures 6-9** in the appendix. Peak regions near wavenumbers of 800, 1000, 1200, and 1450 cm^{-1} were consistently selected within the first-order-logic statements for all concentration ranges and were deemed relevant for serotonin concentration prediction. SHAP values were assessed for all concentration ranges separately and have been visualized on an averaged spectra in **Figure 10** in the appendix.

4 Discussion

Within the present study, a comprehensive framework of spectral quantification from complex biological media was developed. The trained denoising autoencoder improved prediction outcomes near-universally across all model types and enabled robust quantification. The assessed neural network models substantially outperformed traditional machine learning methods commonly used in the SERS domain, as well as past results[4]. The custom layers developed in the present study for the sCNN and CNN_{3PL} models significantly improved quantification performance over the ViT or the CNN_{L}. This can be attributed to the use of logistic output activation, which can often yield a better fit on assay data compared to linear activation, as near the limit of quantification an assay can become saturated. Alternatively, as the values approach the limit of detection, the linearity of the signal can deteriorate. The developed CRIME explainability method identified the spectral contexts in which the model was reliably assessing the relevant serotonin peaks, as well as contexts representing confounding factors or other sample artefacts, which were not readily observable from the outputs of the LEN or the SHAP explanations. The scalability of the method should be further evaluated, as in specific biomarker fluids such as blood, there are a significantly higher amount of analytes resulting in further noise and potential confounding factors. Similarly, as the sample size assessed increases, the number of LIME explanations increases significantly, resulting in potentially long runtimes. With further developments and validation in such datasets, the CRIME framework combined with SERS could see clinically relevant use through acting as the first step in biomarker discovery trials. The exact identification of the signalling biomarkers from SERS is challenging when global explainability methods are used for peak detection, as the spectral signals could be a result of multiple overlapping compounds. However, were the CRIME framework applied, individual target biomarkers could be identified through contexts uniquely, and subsequently assigned to the likely biomarkers through a complete library of present compounds, as well as hypothesized biomarkers. Therefore, with further development and validation, within the chemical spectral domain, the CRIME method promises to enable directly identifying disease features in biological fluids, which could be further refined into specific biomarkers through the identification of relevant contexts.

Availability of Data and Materials

The developed code for the CRIME framework can be found in the following GitHub repository: https://github.com/jkz22/CRIME

Funding

This work was supported by the Stanley Medical Research Institute (grant number: O7R-1888) by grants to SB, and by the Oskar Huttunen Foundation grant to JKZ.

References

[1] Gabriele Ciravegna, Pietro Barbiero, Francesco Giannini, Marco Gori, Pietro Lió, Marco Maggini, and Stefano Melacci. Logic explained networks. *Artificial Intelligence*, 314, 8 2021.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR 2021 - 9th International Conference on Learning Representations*, 10 2020.

[3] Paul H. C. Eilers. Parametric time warping. *Analytical Chemistry*, 76(2):404–411, 2004.

[4] Setu Kasera, Lars O. Herrmann, Jesús Del Barrio, Jeremy J. Baumberg, and Oren A. Scherman. Quantitative multiplexing with nano-self-assemblies in sers. *Scientific Reports*, 4:1–6, 10 2014.

[5] Judith Langer, Dorleta Jimenez de Aberasturi, Javier Aizpurua, Ramon A. Alvarez-Puebla, Baptiste Auguié, Jeremy J. Baumberg, Guillermo C. Bazan, Steven E.J. Bell, Anja Boisen, Alexandre G. Brolo, Jaebum Choo, Dana Cialla-May, Volker Deckert, Laura Fabris, Karen Faulds, F. Javier García de Abajo, Royston Goodacre, Duncan Graham, Amanda J. Haes, Christy L. Haynes, Christian Huck, Tamitake Itoh, Mikael Käll, Janina Kneipp, Nicholas A. Kotov, Hua Kuang, Eric C. Le Ru, Hiang Kwee Lee, Jian Feng Li, Xing Yi Ling, Stefan A. Maier, Thomas Mayerhöfer, Martin Moskovits, Kei Murakoshi, Jwa Min Nam, Shuming Nie, Yukihiro Ozaki, Isabel Pastoriza-Santos, Jorge Perez-Juste, Juergen Popp, Annemarie Pucci, Stephanie Reich, Bin Ren, George C. Schatz, Timur Shegai, Sebastian Schlücker, Li Lin Tay, K. George Thomas, Zhong Qun Tian, Richard P. van Duyne, Tuan Vo-Dinh, Yue Wang, Katherine A. Willets, Chuanlai Xu, Hongxing Xu, Yikai Xu, Yuko S. Yamamoto, Bing Zhao, and Luis M. Liz-Marzán. Present and future of surface-enhanced raman scattering. *ACS Nano*, 14:28–117, 1 2020.

[6] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2323, 1998.

[7] Scott M. Lundberg and Su In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017-December:4766–4775, 5 2017.

[8] Pilot, Signorini, Durante, Orian, Bhamidipati, and Fabris. A review on surface-enhanced raman scattering. *Biosensors*, 9:57, 4 2019.

[9] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, pages 97–101, 2 2016.

[10] Min Xiong and Jian Ye. Reproducibility in surface-enhanced raman spectroscopy. *Journal of Shanghai Jiaotong University (Science)*, 19:681–690, 12 2014.

Appendix

A Experimental Methods

All initial reagents were sourced from Alfa Aesar and Merck and were utilized in their received state unless otherwise specified. CB[8] was prepared following established literature protocols. Millipore water with a resistivity of 18 M Ω·cm was employed in all experiments, unless otherwise indicated. Fresh standard stock solutions of neurotransmitters, specifically dopamine, epinephrine, and serotonin, were prepared at varying concentrations to simulate potential interfering background analytes. Gold nanoparticles (AuNP) with a diameter of 60 nm stabilized by citrate were procured from British Biocell International. Lyophilized urine samples designated as Calibrator Lot No. 150 and Control Level II Lot No. 230 were obtained from RECIPE ClinChek-Control and were reconstituted in dilute hydrochloric acid as per the supplier's guide-lines.

Spectra for both Raman and SERS were collected with a 785 nm laser operating at 17.5 mW, using an Ocean Optics QE65000 Spectrometer. Each spectrum was acquired over a 10 s interval. AuNPs with a 60 nm diameter were first centrifuged at 12,000 rpm for 45 s, repeated twice, and 900 µL of the supernatant were removed. Subsequently, a sample preparation sequence was followed: neurotransmitters (dopamine, epinephrine, and serotonin) were first added, followed by 50 µL of the centrifuged AuNPs, then 20 µL of CB[8] at a final concentration of 20 µM, and finally, 50 µL of thawed urine, which had been initially stored on ice. An identical procedure was replicated, replacing urine with water for control experiments.

Table 1: **Added concentrations and number of spectra for all three neurotransmitters in both water and urine backgrounds.** *Samples D, E, and F were not present in the water dataset, and sample M was not present in the urine dataset. Sample U represents a baseline measurement with no added or measured concentrations of the neurotransmitters. EPI = Epinephrine, DA = Dopamine, 5-HT = Serotonin, n = number of spectra.

Sample	**EPI**	**DA**	**5-HT**	**Water**	**Urine**
	(µM)	(µM)	(µM)	(n)	(n)
A	2	0	0	25	22
B	0	2	0	17	21
C	0	0	2	22	22
D	3	0	7	0*	26
E	0	8	3	0*	28
F	7	3	0	0*	21
G	3	2	7	99	26
H	1	1	9	38	22
I	2	8	3	37	22
J	6	9	1	93	23
K	7	3	2	11	23
L	9	6	4	11	21
M	3	3	3	11	0*
U	0	0	0	58	41

B Neural networks

The denoising autoencoder was implemented in TensorFlow. The encoder comprises two dense layers. The first layer has 2×200 units and utilizes a ReLU activation function. The second layer further compresses the data into an encoding space of dimension 200, with ReLU activation. Symmetric to the encoder, the decoder also consists of two dense layers. The first layer expands the encoded data back to 2×200 dimensions using ReLU activation. Finally, the second layer reconstructs the data to its original dimension using a sigmoid activation function. The model was compiled using mean squared error (MSE) as the loss function and optimized using the Nesterov-accelerated Adaptive Moment Estimation (Nadam) optimizer. Training was conducted for 128 epochs with a batch size of 32. Both training and testing was performed on noisy data to facilitate the denoising objective.

Both the CNNs and the ViT models were implemented in TensorFlow and designed to adapt to SERS spectral data. The CNN architecture comprised sequential layers optimized for 1D convolution operations, and the core CNN architecture was used in all trained CNN models, where the initial layer is a convolutional layer featuring 8 filters and a large kernel size approximately the width of half a peak (25 values), aimed to capture broader features in the spectrum. This initial layer employs a Rectified Linear Unit (ReLU) activation function and reduces sequence length through striding. Intermediate layers employ paired hyperbolic tangent (Tanh) and ReLU activation functions, designed to capture complex patterns while maintaining non-linearity. The combination of consequential Tanh and ReLu layers is to direct the model to assess the upper half of the identified general peaks from the sweeper layer. These layers maintain the same padding to avoid changes in sequence length. The model contains two Tanh-ReLu paired layers with 2x16 and 2x32 filters respectively, with a filter size of 9. The final convolutional stage employs 64 filters with a smaller kernel size of 2 using ReLU activation, aimed to capture fine-grained details in the data. Subsequently, the data is flattened and passed through two fully connected layers employing the same Tanh to ReLu structure with 32 and 16 nodes respectively, to serve the regression task. The core architecture of the CNN models was benchmarked against similar architectures with Tanh-ReLu pairs replaced with ReLu pairs, inversed ReLu-Tanh pairs, or single ReLu layers. Each of the benchmark architectures were trained and validated as described in the methods section, and the results are summarized in **Table 5**.

The scale-adjusting CNN model was developed with two unique scaling layers implemented. These were a multi-scale assessing layer, and a local scaling layer. Both layers were utilized prior to the half-peak ReLu layer in the core CNN architecture. The multi-scale layer captures features from the input X, with three layers sized 8, 25, and 50, each with 8 filters. Each convolutional operation is defined as $C_i(X) = W_i * X + b_i$, where $*$ denotes a convolutional operation, W the weight, and b the bias. To assess the spectra at different scales simultaneously, the output of each convolutional layer is combined following the convolutional operations along the feature dimension. The local peak scaling layer in turn was developed to scale regions of the spectra which were assessed not to be relevant to the outcome variable, identified from the reference spectra of the pure compounds in water. The layer applies a set of scaling factors s_j unique to the number of pre-registered regions of interest (or non-interest), which are defined by start and end indices a_j, b_j, in the spectra. The scaling operation for each region is expressed as: $S_j(X) = X_{aj:bj} \odot s_j$, where $\odot$ denotes the element-wise multiplication. The output of the scaling operation is concatenated to reconstruct the spectra with scaled regions. The modified output layer assessed in both custom layer CNN models utilizes a Three-Parameter Logistic (3PL) activation function. The ViT architecture in turn consisted of an embedding layer with a patch size of 25 matching the CNN architectures initial layer, with a hidden size of 64 and a dropout rate of 0.1. Subsequently, the architecture consisted of 6 transformer blocks with 6 multi-head perceptron's. The transformer blocks each employed Gaussian Error Linear Unit (GELU) activation functions.

Hyperparameter tuning and architecture search for both the CNN variants and the ViT was conducted iteratively, guided by the model's performance on the validation set. Each model variant for both the CNNs and the ViT was trained 100 times, with an ensemble average used for evaluation. Both model types were optimized using the Adaptive Moment Estimation (Adam) algorithm with a learning rate of 0.001, a batch size of 64, and 256 epochs, and compiled with a mean absolute error (MAE) loss function. Additional evaluation metrics included MSE and mean percentage error (MPE). Early stopping with a patience of 64 was employed to mitigate overfitting, and model checkpoints were saved for epochs that minimized validation loss. Reproducibility was ensured by setting random seeds for TensorFlow, NumPy, and the train-test split. Of the 100 trained models in the ensemble, the model with the lowest MAE in the validation set was selected as the final model, which was assessed in the holdout test set.

Table 2: **Validation set results for neural network models.** sCNN = convolutional neural network with scaling layers and three parameter logistic (3PL) output layer, CNN_{3PL} = convolutional neural network with 3PL output layer, CNN_L = convolutional neural network with linear output layer, ViT = vision transformer.

Dataset	**CNN_{3PL}**	**CNN_L**	**sCNN**	**ViT**
Denoising Autoencoder	0.24	0.13	0.33	0.20
Raw Spectra	0.19	0.31	0.14	0.25

C CRIME variational autoencoder

The VAE architecture used in this study consisted of a simple encoder, sampler, and decoder, however the architecture can be fine-tuned depending on the individual requirements for the CRIME framework in future applications. Within the encoder of the CRIME VAE, the input data X, is transformed into the mean (μ) and logarithm of the variance ($\log(\sigma^2)$) in a proposed Gaussian distribution in the latent space through a fully connected ReLu layer with 256 nodes. During training, the outputs of the encoder are passed to a sampling layer which generates a random noise variable ϵ generated from a standard gaussian distribution, which is then transformed using the encoder outputs to draw samples z as such: $z = \mu + \sigma * \epsilon$. The decoder then applies a mirrored dense layer network to the encoder with a ReLu layer with 256 nodes, and a final output sigmoid layer with 3x842 nodes. The model was trained for 128 epochs and a batch size of 32, with the Adam optimizer with a learning rate of 0.001 and using the sum of the mean squared error, and Kullback–Leibler divergence as the loss function.

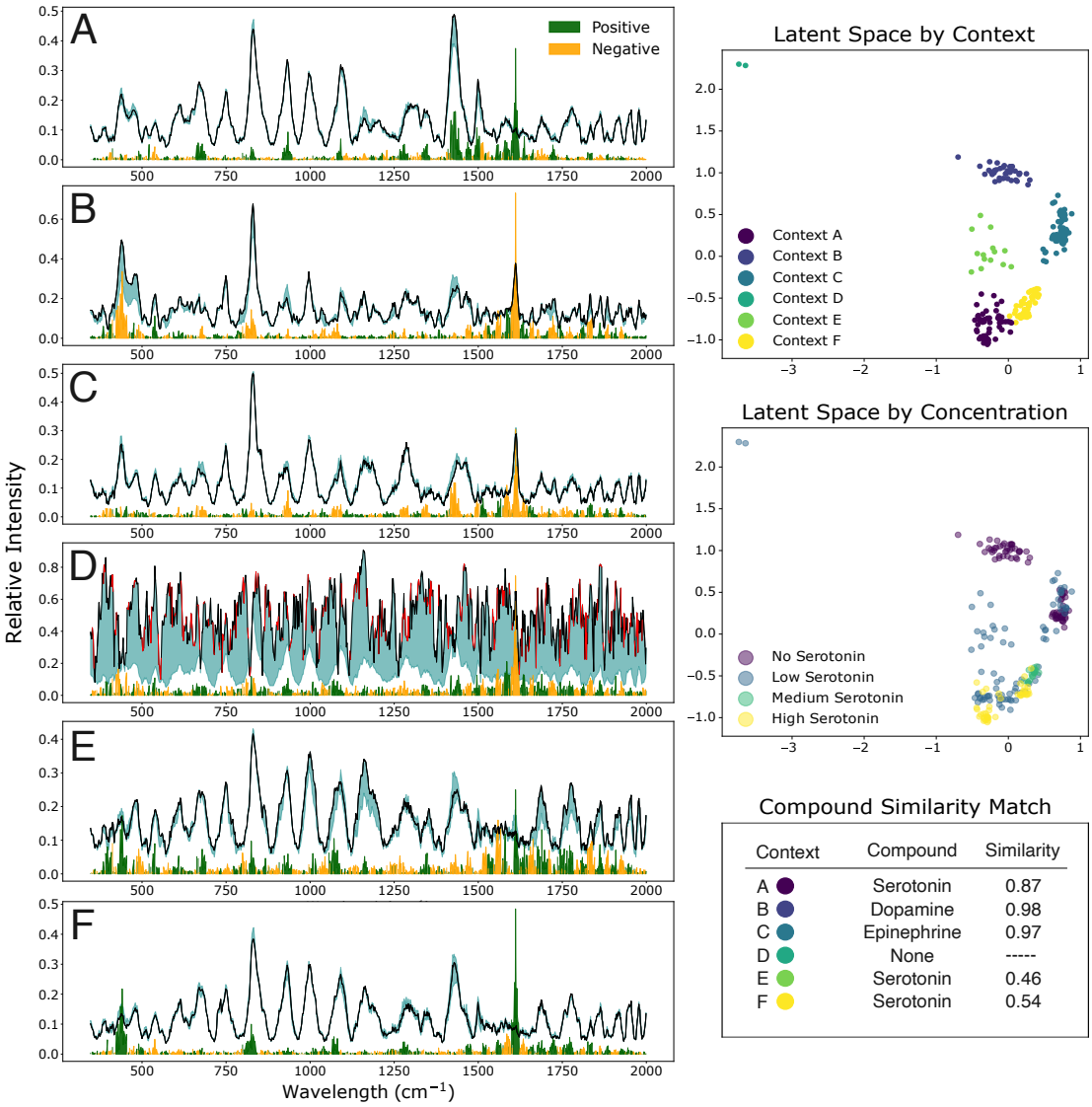

Figure 2: **Results for Context Representative Interpretable Model Explanations (CRIME) analysis.** Six distinct contexts were identified, which are visualized across mean spectra in subfigures A - F. Positive prediction weights are presented in green, negative prediction weights in yellow, and perturbation limits have been shaded in teal. Red regions in the mean spectra correspond to average perturbation limits at either the top or bottom of the feature weight range for the simplicity of the plot. Latent spaces are visualized by context and concentrations, and compound similarity matching was done using cosine similarity. The highest similarity score is presented alongside the matched compound.

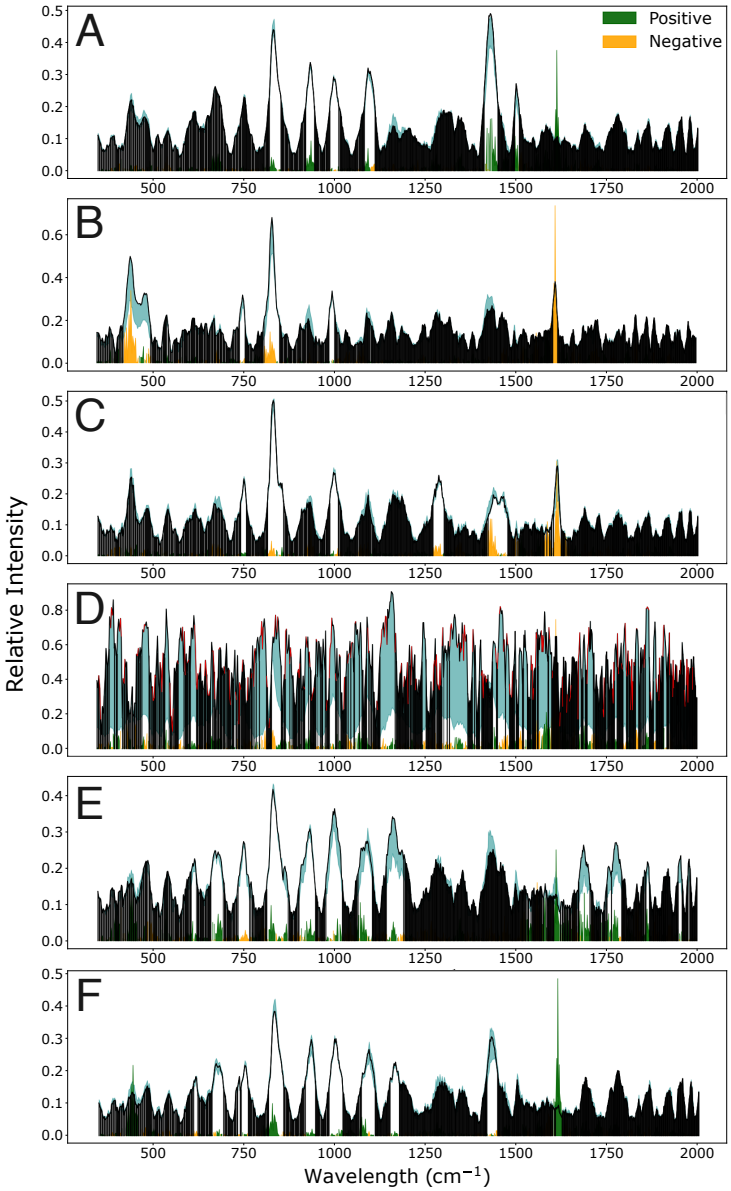

Figure 3: **Peak-region clusters of high relevance extracted from CRIME contexts for compound identification.** Context labels correspond to labels in Figure 2. Areas not relevant are marked in black. High-relevance clusters were obtained with K-means clustering of the product of peak height and LIME weights, with the top 5 largest clusters selected.

Table 3: **Cosine similarity values across explanation weighted reference spectra and explanation weighted mean context spectra.** Highest similarity values within a context cluster are bolded. X = context.

Reference compound	**A**	**B**	**C**	**D**	**E**	**F**
Serotonin	**0.87**	0.85	0.60	-0.79	**0.46**	**0.54**
Dopamine	0.05	**0.98**	0.91	-0.80	0.29	0.06
Epinephrine	0.0	0.81	**0.97**	-0.79	0.12	0.08

D Benchmarking

D.1 Quantification model benchmarks

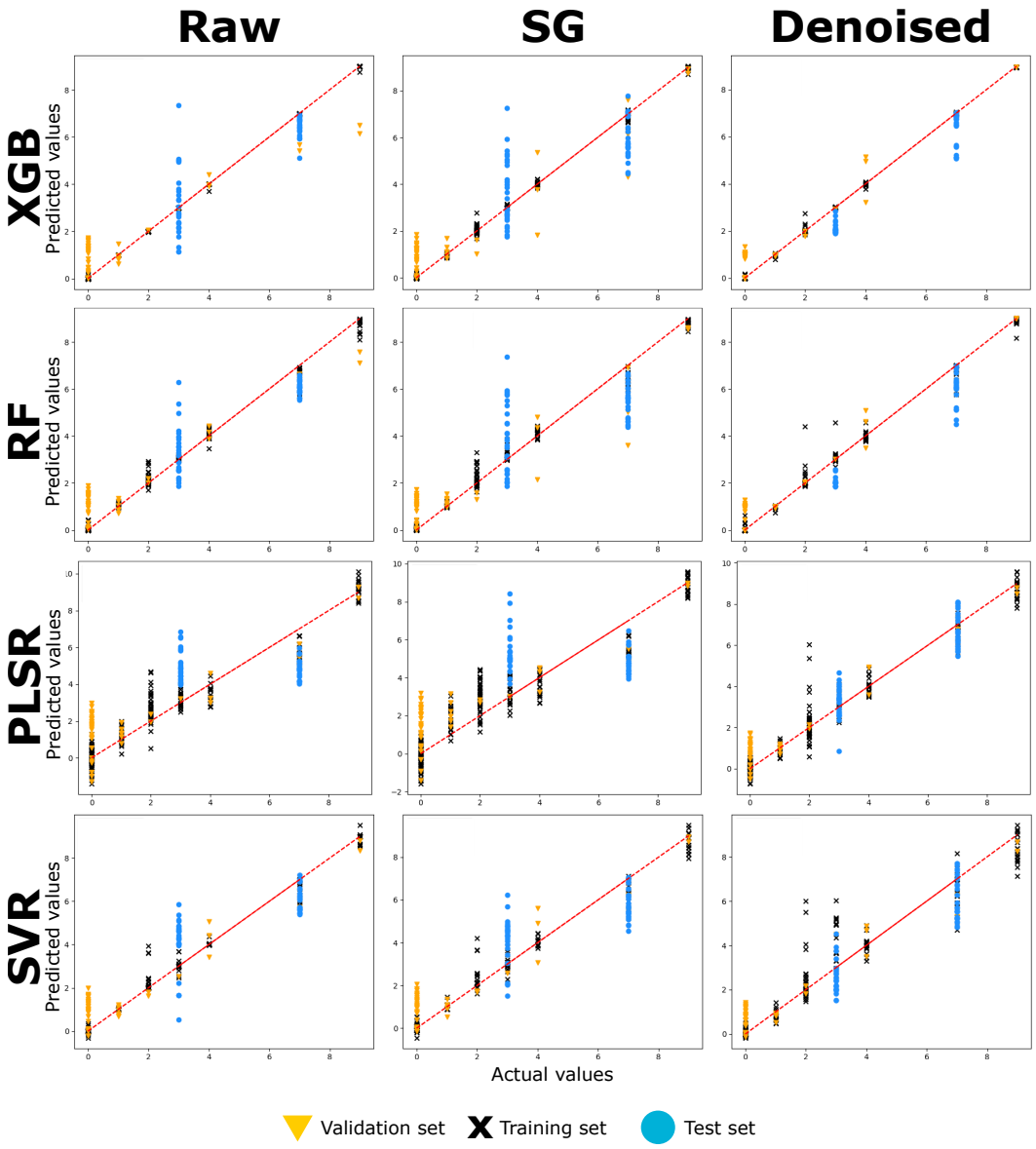

Figure 4: **Results for quantification model benchmarking.** Training set predictions are marked in black (cross), validation set predictions in orange (triangle), and test set predictions in blue (circle). XGB = extreme gradient boosting, RF = random forests, PLSR = partial least squares regression, SVM = support vector machine regression, SG = Savitzky-Golay filter.

The grids used for the hyperparameter search of the benchmark machine learning models are presented below with the final hyperparameters for the denoised models in bold.

XGBoost

Hyperparameter	Values
colsample_bytree	**0.5**, 0.7, 0.8
learning_rate	**0.01**, 0.1, 0.2, 0.3
max_depth	3, **6**, 9, 12
alpha	1, **3**, 5
n_estimators	100, 300, 600, 900, **1200**

Random Forests

Hyperparameter	Values
max_depth	1, 2, 3, 6, 7, **8**, 10
n_estimators	100, 300, 600, 900, **1200**

PLSR

Hyperparameter	Values
n_components	5, 8, **12**

SVM

Hyperparameter	Values
C	0.1, 1, 10, **50**, 100
epsilon	**0.01**, 0.1, 1
gamma	**scale**, auto

Table 4: **Comparison of test-set mean absolute errors across different machine learning models for serotonin quantification from SERS spectra.** Best performing model MAEs within each dataset has been bolded, and the two best models overall have been marked with an asterisk (*). Additionally, for baseline comparison, the mean absolute error of the previously published PLSR model has been presented. sCNN = convolutional neural network with scaling layers and three parameter logistic (3PL) output layer, CNN_{3PL} = convolutional neural network with 3PL output layer, CNN_L = convolutional neural network with linear output layer, SVM = support vector machine, RF = random forests, PLSR = partial least squares regression, XGB = extreme gradient boosting, ViT = vision transformer.

Dataset	**XGB**	**PLSR**	**RF**	**SVM**	**CNN_{3PL}**	**CNN_L**	**sCNN**	**ViT**
Denoising Autoencoder	0.78	0.70	0.93	0.96	0.15*	0.30	**0.11***	0.30
Raw Spectra	0.82	2.05	0.88	1.26	1.14	**0.70**	0.95	1.17
Savitzky-Golay	1.15	2.25	1.46	1.37	-	-	-	-
Kasera et al. 2014	-	**0.52**	-	-	-	-	-	-

Table 5: **Comparison of core CNN architectures.** All architectures were trained on denoised urine medium datasets, and validated using the holdout test set. ReLu = Rectified linear unit, Tanh = Hyperbolic tangent, MAE = Mean absolute error, MPE = Mean percentage error.

Error	**Tanh-ReLu**	**ReLu-Tanh**	**2xReLu**	**ReLu**
MAE	0.30	0.56	0.88	0.78
MPE	7.45	17.68	20.71	16.61

D.2 Logic explained networks and Shapley additive explanations

The LEN was implemented using PyTorch in python. In order to apply the LEN, the spectral input data was sectioned to discrete categories, and concept mapping was done through taking the mean of the min-max scaled feature map activations across each layer of the model. Each concept corresponded to approximately 25 x-axis points across the SERS spectrum corresponding to approximately half a peak. The LEN architecture was modified to be similar to the original model architecture, consisting of an entropy layer with 164 input nodes, a leaky ReLu layer with 32 nodes, a Tanh layer with 16 nodes, a ReLu layer with 4 nodes, and a final linear output layer. The LEN was trained using weight decay Adam as the optimizer, using binary cross entropy with logits loss as the loss function with a scaled auxiliary entropy loss at a 0.000001 multiplier. The model was trained with 5001 epochs using a learning rate of 0.0001.

SHAP calculations were done using the above-mentioned sectioned categories separately using Gradient Explainer.

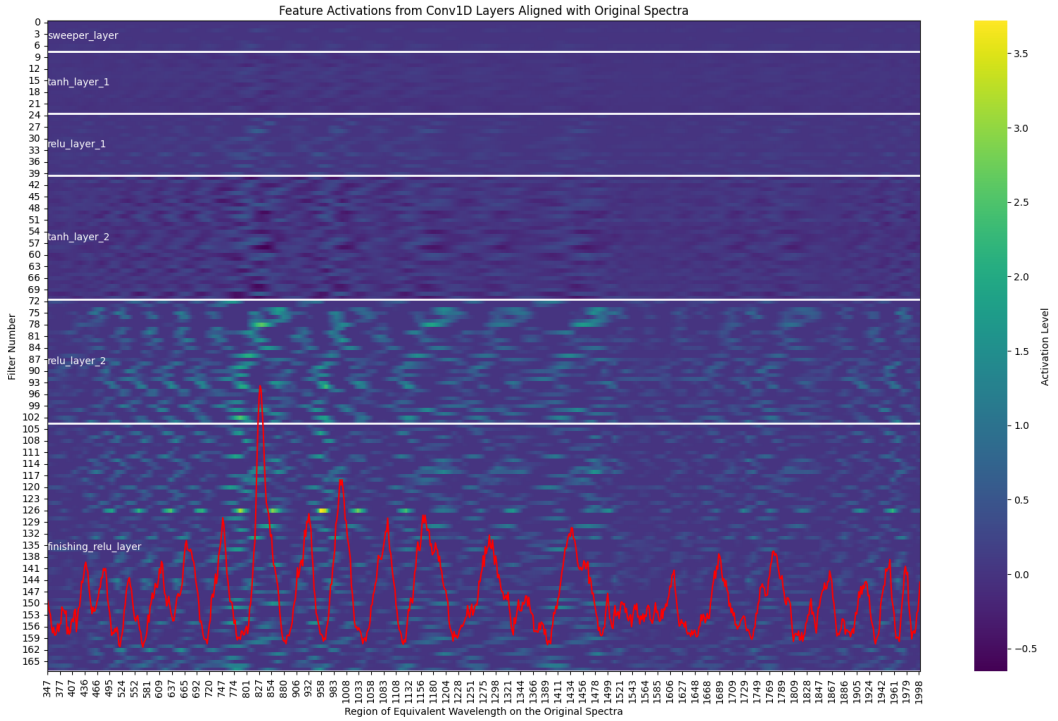

Figure 5: **Feature activation map for each convolutional layer in the CNN model overlayed with an example spectra.** SERS spectra is shown in red, and higher activations are marked with a yellow hue, with lower activations marked in blue.

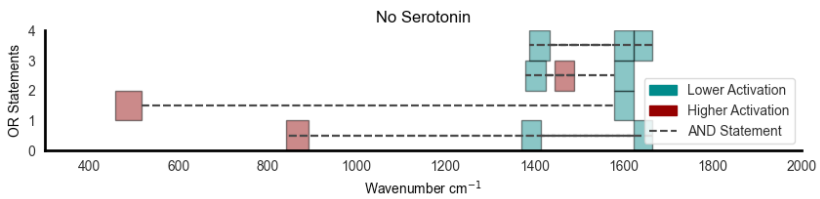

Figure 6: **LEN results visualized for samples with no serotonin concentration.** OR statements are separated vertically, and AND statements are presented level with a dashed line connecting the statements. Blue squares denote lower activation and red squares higher activation.

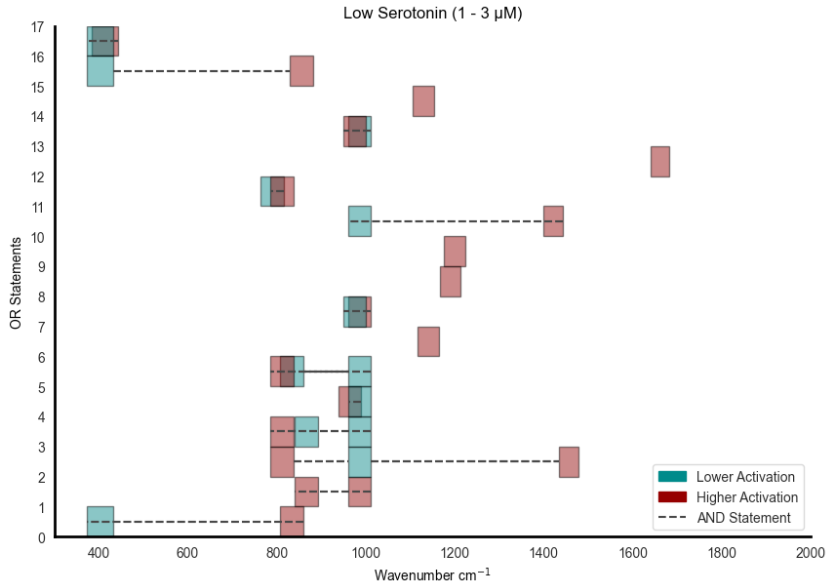

Figure 7: **LEN results visualized for low serotonin concentrations.** OR statements are separated vertically, and AND statements are presented level with a dashed line connecting the statements. Blue squares denote lower activation and red squares higher activation.

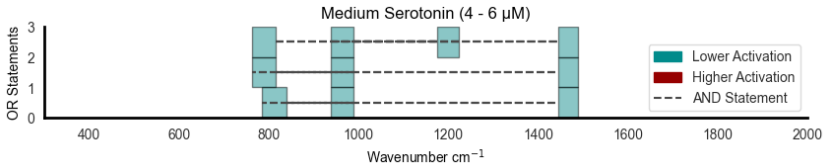

Figure 8: **LEN results visualized for medium serotonin concentrations.** OR statements are separated vertically, and AND statements are presented level with a dashed line connecting the statements. Blue squares denote lower activation and red squares higher activation.

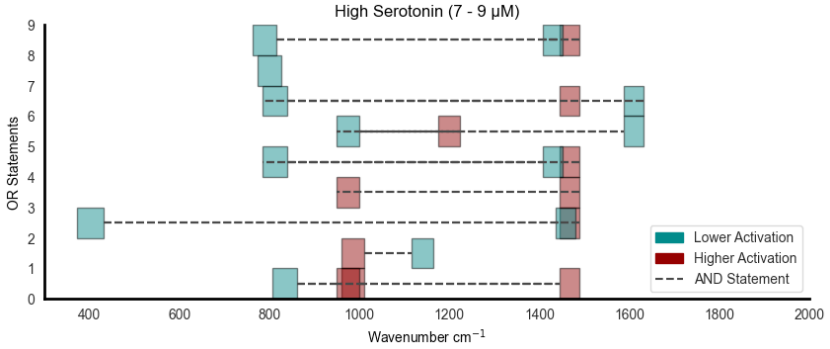

Figure 9: **LEN results visualized for high serotonin concentrations.** OR statements are separated vertically, and AND statements are presented level with a dashed line connecting the statements. Blue squares denote lower activation and red squares higher activation.

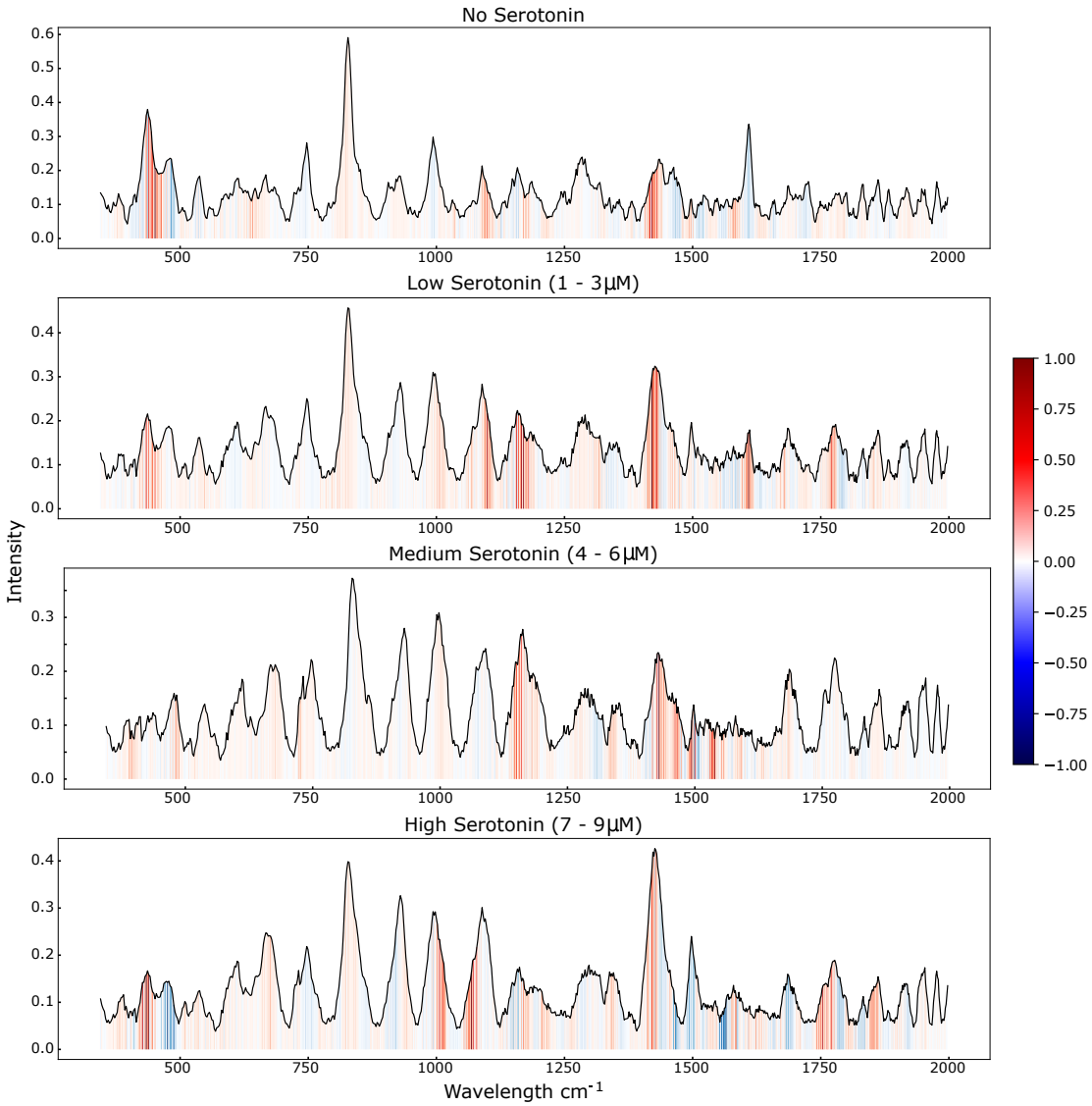

Figure 10: **Shapley additive explanations (SHAP) visualized for all serotonin concentration ranges.** Spectra shown are mean spectra across the respective concentration ranges. SHAP values were obtained using Gradient Explainer, and red areas correspond to positive SHAP values and blue areas to negative SHAP values.