Progressive and Interpretable Explanations for Machine Learning Models on Spectroscopy Data

Mingxing Zhang[1,2]

[1]School of Computer Science and Information Technology, University College Cork, Ireland
[2]Centre for Research Training in Artificial Intelligence, University College Cork, Ireland

Abstract

In recent years, machine learning models have been increasingly applied to spectroscopic datasets for chemical and biomedical analysis. For their successful adoption, particularly in clinical and safety-critical settings, professionals and researchers must be able to understand and trust the reasoning behind model predictions. However, the inherently high dimensionality and strong collinearity of spectroscopy data not only complicate model training but also undermine the stability and consistency of explanations, leading to fluctuations in feature importance across repeated training runs. To address this challenge, this study proposes an explainable machine learning pipeline that integrates feature extraction techniques, namely Sparse PCA, PLS, and MCR-ALS, with Shapley Additive exPlanations (SHAP) for post hoc interpretation, while preserving the ability to project explanations back into the original input space so that practitioners can relate them to underlying biological signals. The proposed framework supports both global and local analysis, revealing the spectral bands that drive overall model behaviour as well as the instance-specific features that influence individual predictions. To further improve the accessibility and interpretability of these results, large language models (LLMs) are used to translate visual explanations into concise, user-oriented textual summaries.

Keywords

SHAP, Spectroscopy, Medical Diagnosis, Feature Extraction

1. Introduction

Spectroscopy techniques are broadly employed in the scientific and industrial domains. Techniques such as Raman, infrared, near-infrared and nuclear magnetic resonance spectroscopy provide rich molecular fingerprint information, enabling non-destructive and label-free analysis of samples [1]. In recent years, the increased availability of high-resolution spectroscopy instruments has enabled the generation of large, complex datasets [2]. Machine learning techniques have been increasingly used to analyse and evaluate spectroscopic data for classification, regression, and anomaly detection [3].

Despite the strong predictive performance of machine learning models applied to spectroscopic data, their adoption in high-stakes domains remains limited, particularly in medical and clinical applications. In this high-risk field, the European Union Artificial Intelligence Act (EU AI Act) now imposes legally binding requirements for transparency, accountability, and meaningful human oversight in AI-driven decision-making systems [4]. At the same time, clinicians and healthcare practitioners must be able to understand and trust the reasoning behind algorithmic predictions in order to safely integrate such systems into clinical practice [5]. As a result, eXplainable Artificial Intelligence (XAI) has become a fundamental requirement for the responsible deployment of machine learning models, enabling predictions to be interpreted, validated, and justified in line with ethical and regulatory expectations.

However, achieving explainability in spectroscopic machine learning models is particularly challenging due to the intrinsic characteristics of spectroscopic data. Spectra are typically high-dimensional, comprising hundreds of features, and exhibit strong collinearity, as adjacent wavelengths originate from the same molecular vibrations and spectral peaks are inherently broad and overlapping [6]. As a result, information relevant to classification is often distributed across extended spectral regions rather than

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

123116142@umail.ucc.ie (M. Zhang)
https://github.com/StellaZhang00 (M. Zhang)

confined to isolated peaks [7]. These properties complicate both model training and post-hoc explanation: feature importance measures may become diluted across correlated variables and exhibit instability or inconsistency across repeated training runs [8]. Consequently, standard explainability techniques, when applied directly to raw spectroscopic features, may fail to provide reliable or physically meaningful interpretations. Feature extraction methods, such as Principal Component Analysis (PCA) and Partial Least Squares (PLS), are widely used in spectroscopy to project the data into a lower-dimensional space, thereby reducing noise and collinearity. However, the resulting components often lack direct biological meaning, which can limit the usefulness of XAI methods, as the explanations are then expressed in terms of latent variables rather than the original spectral features.

Therefore, there is a clear need for an explainable pipeline tailored to spectroscopic data, one that addresses the challenges of high dimensionality, strong collinearity, and distributed spectral information while still producing explanations that are stable, consistent, and physically meaningful.

2. Related Work

Since machine learning techniques are now widely applied to spectroscopic data, and prediction interpretability is required in many application domains, a growing number of explainable AI (XAI) techniques have been employed in spectroscopy to obtain meaningful explanations. A comprehensive review of existing explanatory approaches for spectroscopic signals is provided by Contreras et al. [9]. Their review indicates that relatively few XAI studies have been specifically developed for spectroscopy. Among the existing work, explanation methods can generally be categorized into four groups: activation-based, perturbation-based, gradient-based, and surrogate-based approaches, as well as various combinations of these techniques.

Rossberg et al. [10] proposed two model-specific XAI-based feature-selection methods for Raman spectroscopy, combining CNNs with Grad-CAM and Transformers with attention-derived importance scores. Their results showed that the CNN-Grad-CAM approach achieved the highest average accuracy, although no single method consistently outperformed all others across every scenario. Despite this strong predictive performance, both approaches are tied to specific model architectures, and the selected features remain mainly individual wavenumbers, which may limit physical interpretability in spectroscopy, where variables are highly correlated and meaningful information is often distributed across broader spectral regions rather than isolated features.

While Contreras et al. [11] proposed a spectral-zones-based SHAP and LIME framework for deep learning on spectral data, in which neighboring wavelengths are grouped into contiguous zones before perturbation so that the explanations emphasize interpretable spectral regions rather than single features. However, this strategy remains constrained by spatial adjacency: it can only group neighboring wavelengths and therefore cannot capture correlated but non-adjacent features that may originate from the same chemical component. As noted in the study, some biochemical signatures, such as fat-related signals, may appear in several separate spectral regions, meaning that chemically related but spatially distant variables are still treated independently.

Overall, these studies demonstrate that some form of feature grouping or segmentation is essential for explaining high-dimensional spectroscopy data. However, most of the existing approaches still rely mainly on architectural constraints or local spectral adjacency, and therefore may fail to capture the broader correlation structure of the data. In our study, we address this limitation by applying several feature extraction techniques to group correlated features into latent components, on which a classifier is subsequently trained. Explanations are then generated at the component level, enabling interpretations that are potentially more robust and meaningful than isolated feature-wise attributions.

One common approach to addressing the high dimensionality and strong inter-correlation of spectroscopic data is the use of Principal Component Analysis (PCA). PCA constructs new variables, known as principal components, along directions of maximum variance in the dataset. Because these components are orthogonal to one another, they are uncorrelated in the transformed feature space, which can make subsequent classification and explanation more stable and reliable in machine learning workflows.

In spectroscopy, PCA has also been shown to reveal systematic variation and underlying patterns in complex spectral datasets [12]. However, to improve interpretability, this study applies Sparse PCA, which introduces a sparsity-inducing penalty when estimating component directions. This encourages each component to rely on only a limited subset of features, although the resulting component directions are not guaranteed to be orthogonal.

Another widely used technique is Partial Least Squares (PLS), which extracts latent components by maximizing the covariance between the input features and the target variables. Unlike PCA, which is unsupervised and focuses only on the structure of the input data, PLS is supervised and aims to identify components that are directly relevant to the prediction task. This property makes PLS particularly attractive when the goal is not only dimensionality reduction but also improved predictive modelling.

However, a major limitation of both PCA and PLS is that, once the original spectra are transformed into latent components, the direct connection between the machine learning output and the original input variables becomes less transparent. As a result, although these techniques can improve modelling by reducing noise and collinearity, they also make it more difficult for practitioners to relate model explanations back to specific spectral signatures. Therefore, methods that allow components or explanations to be projected back onto the original spectral domain after classification are crucial for enabling meaningful integration of explainable artificial intelligence (XAI) into spectroscopic analysis.

Multivariate Curve Resolution–Alternating Least Squares (MCR-ALS) has been applied in spectroscopic studies to unmix complex spectra into physically meaningful spectral profiles, as demonstrated in [13]. In this context, model explanations may be complemented by spectral decomposition, enabling the discriminative peaks identified by the model to be examined in relation to the chemical constituents they may represent. In many cases, such decomposition can help resolve dominant underlying spectral contributions. However, MCR-ALS is inherently affected by ambiguities related to model initialisation, constraint selection, and the specified number of components, and weak or strongly overlapping constituents may remain unresolved. However, as complementary evidence that can provide practitioners with a more chemically grounded perspective on the model explanations.

3. Research Questions and Objectives

Based on the inherent challenges of spectroscopic data, we aim to address these issues by using three feature extraction techniques to group into components, applying SHAP to quantify their contributions to model predictions, and then back-projecting these contributions to the original spectral domain to obtain physically meaningful explanations.

Research Question How can XAI techniques deliver explanations for spectroscopic data that are (i) consistent across repeated model training and validation, (ii) physically meaningful with respect to underlying spectral phenomena, and (iii) fast, human-understandable and actionable in real-world spectroscopy settings?

Objectives This study aims to:

- Develop an explainable machine-learning pipeline tailored to spectroscopic data by integrating three feature extraction techniques Sparse PCA, PLS and MCR-ALS with a predictive model and post-hoc explanation methods.
- Define and apply evaluation metrics to quantify (i) the novel method we propose for assessing the stability and consistency of explanations across cross-validation folds and repeated runs, and (ii) the faithfulness of explanations to the trained model's decision process.
- MCR–ALS is also employed as a complementary interpretability tool. In addition to extracting component-based representations from the original spectra, it decomposes the data into underlying spectral profiles, which can provide chemically interpretable insights for downstream analysis.

- Investigate the use of an LLM-based prompt-engineering framework to translate local and global explanation visualizations into concise, accurate, and human-readable textual interpretations.

4. Research Plan and Methodology

To address the research question, we propose the research plan illustrated in Fig. 1. At this stage, the study draws on two datasets: cheek-mucosa Raman spectroscopy data from [14] and DRS data from [15]. The work is organised into three packages. Package 1 concerns the development of an interpretable machine-learning pipeline capable of generating both global and local visual explanations. This package is mostly complete, except for the faithfulness assessment and the inclusion of comparative experiments using PLS and MCR-ALS as alternative feature extraction techniques. Package 2 aims to compare the explanations produced by the PLS- and Sparse PCA-based pipelines with evidence derived from MCR-ALS decomposition. Package 3 aims to translate the visual outputs into clear, human-understandable, and user-oriented textual interpretations. These remaining components will be pursued in future work.

The first package comprises the following steps:

1. Feature extraction: The input spectra are transformed into a set of low-dimensional components through the application of three different feature extraction techniques, Sparse PCA, PLS, MAC-ALS.
2. Supervised prediction: the components are used as inputs to a classifier to produce class predictions; in this work, we used Random Forest (RF) and Support Vector Classifiers (SVC) to classify;
3. SHAP explainability: Shapley values are computed for the latent components to quantify their contributions to the model output;
4. Back-projection and visualisation: the Shapley values' contribution is reprojected into the original spectra as opacity to highlight the parts of the input that are more relevant to the decision. Their being higher or lower than the rest of the samples is driven by the component values, which give the color red for higher and blue for lower. Combining and overlaying these two aspects onto the original spectral shape to yield physically interpretable, feature-level explanations.
5. To assess the consistency of both global and local explanations, this study computes cosine similarity and correlation between explanation vectors obtained from models trained on different cross-validation folds. This quantifies how stable the explanations remain when the training data changes across folds. To assess faithfulness which is the future work, we will perturb the decisive components (ranked by SHAP importance) one at a time and measure the resulting drop in classification performance or the decrease in predicted class probability. If perturbing highly-ranked components causes a larger degradation than perturbing low-ranked ones, the explanations are more likely to reflect features the model truly relies on.

The second package proceeds by applying MCR-ALS to the spectra after the explanations from the Sparse PCA and PLS based models have been obtained. First, the discriminative peaks highlighted by the models are compared with the decomposed spectral profiles to investigate the biochemical constituents that may underlie these regions. Then, peaks that are assigned high importance in the model explanations but are not clearly represented in the decomposed profiles are identified. These cases are further examined as possible instances in which the predictive models rely on weak, highly overlapping, or noise-driven spectral variations that MCR-ALS cannot reliably resolve. Given the ambiguities inherent in MCR-ALS, including sensitivity to initialisation, constraint selection, and component number, its outputs are used as complementary biochemical evidence rather than as definitive ground truth.

The third package proceeds by transforming the outputs of the previous packages into textual interpretations. First, the visual explanations and decomposition results are gathered and linked across the different stages of the framework. Then, prompt engineering is used to guide large language models (LLMs) in generating concise, coherent, and user-oriented textual summaries. Finally, these summaries are produced for both intermediate outputs and the final integrated explanation, thereby

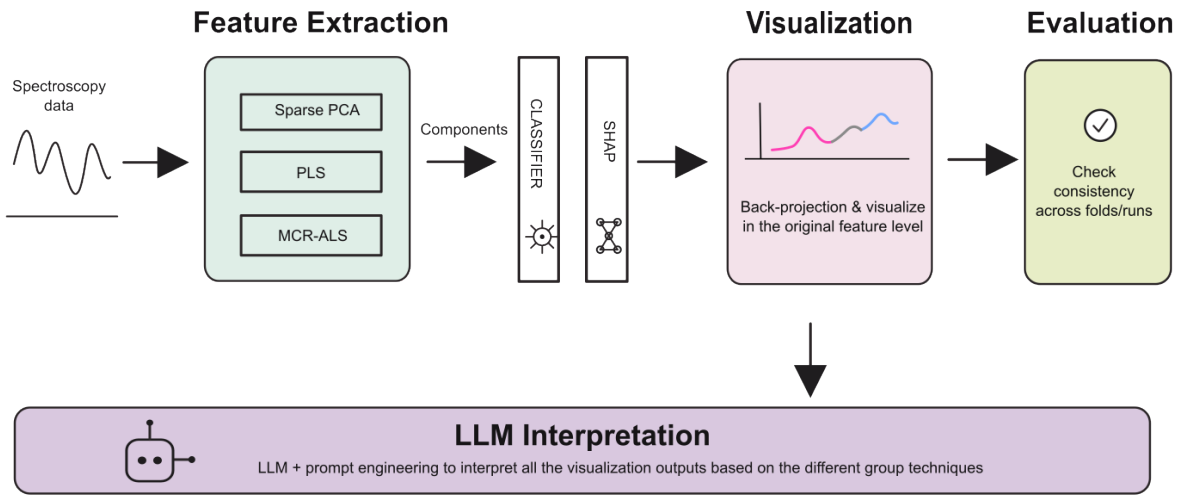

Figure 1: Overview of proposed research plan.

helping practitioners and other end users understand the combined meaning and practical significance of the results.

5. Results and Contributions to Date

To date, for both datasets, we have obtained original-space global and local explanations using the proposed interpretable Sparse PCA based pipeline. We also assessed explanation consistency against a baseline that applies SHAP directly to the original spectra without Sparse PCA. The proposed pipeline exhibits improved stability, as the component representation groups correlated spectral variables, reducing attribution dilution and fold-to-fold variability.

Figure 2 presents the global explanation for the cheek-mucosa dataset with two classes: Healthy (H) and Potentially Malignant Lesion (PML). The grey curve denotes the class-wise mean Raman spectrum. Colour encodes the class-averaged, normalised principal-component (PC) values back-projected to the original wavenumber axis, while opacity represents the class-averaged Shapley values, also back-projected to the original space. More opaque regions indicate greater importance for the model's decision; red indicates higher-than-average (positive) PC values in that region, whereas blue indicates lower-than-average (negative) values. Overall, the visualisation addresses two practically relevant questions: which spectral regions drive the classifier, and whether those regions show increased or decreased intensity relative to the expected baseline.

Figure 3 shows the local explanation for a cheek-mucosa instance that is correctly predicted as PML. The grey curve corresponds to the instance's Raman spectrum. Colour encodes the instance's normalised principal-component (PC) values, while opacity encodes the back-projected Shapley values, separated into positive and negative contributions. Positive Shapley regions indicate spectral intervals that support the PML prediction, whereas negative regions indicate intervals that oppose it. Overall, the visualisation highlights which parts of the spectrum influence the model's decision and clarifies whether their influence is supportive or inhibitory for this class. Combined with the global explanation, these local attributions enable us to assess whether the instance-level evidence aligns with the class-level patterns.

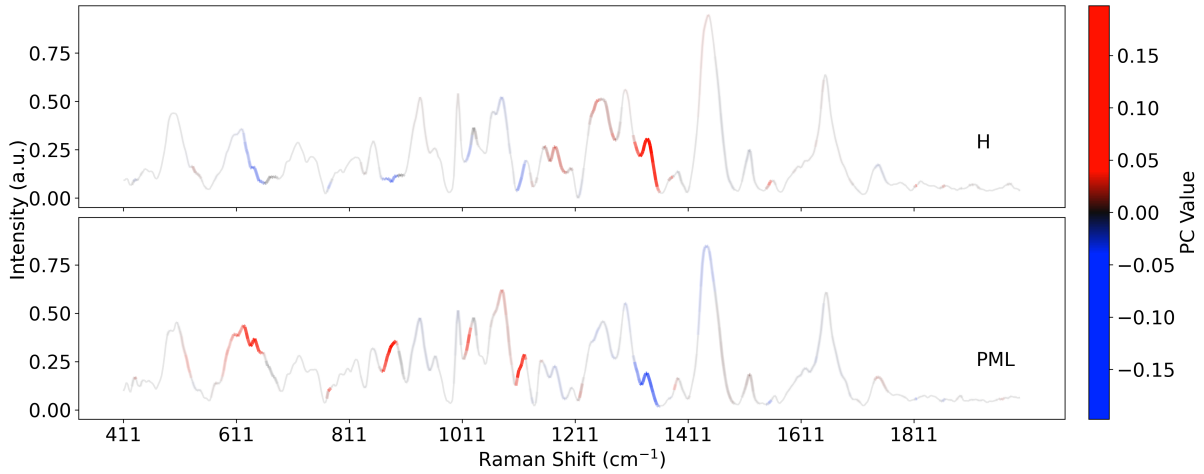

Figure 2: Global explanations for the binary Raman spectroscopy task (classes H and PML).

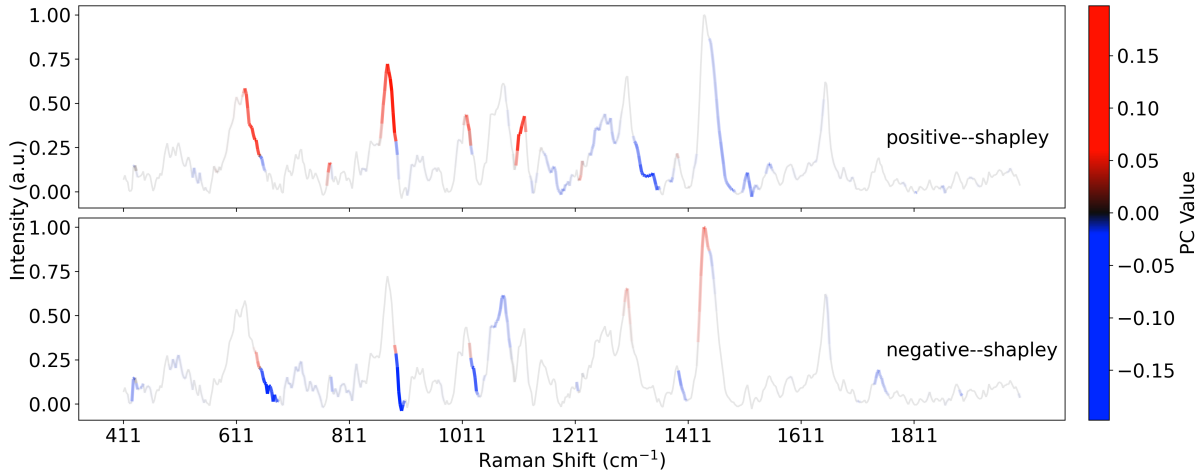

Figure 3: Local explanation for a binary classification instance correctly predicted as PML.

Table 1
Comparison of global explanation consistency and correlation across classes for Raman spectroscopy data.

Class	SHAPCA		SHAP	
	Consistency	Correlation	Consistency	Correlation
H	0.820	0.811	0.715	0.698
PML	0.842	0.835	0.759	0.747

In addition, Tables 1 and 2 summarise the reproducibility of global and local explanations for the proposed pipeline (SHAPCA) and the baseline SHAP approach, using two complementary measures: cosine similarity and Pearson correlation. Across both metrics and at both explanation levels, SHAPCA consistently outperforms baseline SHAP, indicating more stable attributions across cross-validation folds.

6. Next Research Step and Expected Final Contribution

The next stage of this research will focus on completing the remaining components of the proposed framework and extending its evaluation. First, comparative experiments using PLS and MCR-ALS will

Table 2
Comparison of local explanation consistency and correlation for Raman spectroscopy data.

	SHAPCA		SHAP	
	Consistency	Correlation	Consistency	Correlation
Local	0.622	0.615	0.573	0.568

be completed to get the explanation from different view of feature extraction approach.

The second stage will focus on the analysis of all the explanantion we get from the first stage, decomposed spectral profiles will be compared with the discriminative peaks identified by the Sparse PCA and PLS based pipelines.

A further direction of this work is the development of Package 3, which aims to translate the outputs of the framework into concise, coherent, and user-oriented textual interpretations.

More broadly, this research is expected to contribute an explainable spectroscopy analysis framework that improves the stability of feature attributions under high dimensionality and strong collinearity through the use of different feature extraction techniques, while also bridging the gap between latent space machine-learning explanations and original feature level biochemical interpretation through explanation back-projection, spectral decomposition, and language-based summarisation. The study therefore aims to advance a more comprehensive and practitioner-oriented approach to explainable artificial intelligence in spectroscopy. In the longer term, this framework could promote greater transparency, trust, and usability of machine-learning models in biomedical and chemical applications, particularly in settings where interpretability is critical.

.

Acknowledgments

This work was conducted with the financial support of Research Ireland - Taighde Éireann, under Grant Nos. 18/CRT/6223 and 12/RC/2289-P2, the latter being co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] K. A. Esmonde-White, M. Cuellar, I. R. Lewis, The role of raman spectroscopy in biopharmaceuticals from development to manufacturing, Analytical and Bioanalytical Chemistry 414 (2022) 969–991.

[2] S. F. Parker, The analysis of vibrational spectra: Past, present and future, ChemPlusChem 90 (2025) e202400461.

[3] N. Rossberg, C. L. Li, S. Innocente, S. Andersson-Engels, K. Komolibus, B. O'Sullivan, A. Visentin, Machine learning applications to diffuse reflectance spectroscopy in optical diagnosis: a systematic review, Applied Spectroscopy Reviews (2025) 1–52.

[4] E. Union, Regulation (eu) 2024/1689 of the european parliament and of the council, Official Journal of the European Union (2024).

[5] C. C. Yang, Explainable artificial intelligence for predictive modeling in healthcare, Journal of healthcare informatics research 6 (2022) 228–239.

[6] K. Guo, Y. Shen, G. A. Gonzalez-Montiel, Y. Huang, Y. Zhou, M. Surve, Z. Guo, P. Das, N. V. Chawla, O. Wiest, X. Zhang, Artificial intelligence in spectroscopy: Advancing chemistry from prediction to generation and beyond, 2025. URL: https://arxiv.org/abs/2502.09897. arXiv:2502.09897.

[7] R. Pandiselvam, V. Prithviraj, M. Manikantan, A. Kothakota, A. V. Rusu, M. Trif, A. Mousavi Khaneghah, Recent advancements in nir spectroscopy for assessing the quality and safety of horticultural products: A comprehensive review, Frontiers in Nutrition 9 (2022) 973457.

[8] A. M. Salih, Explainable artificial intelligence for dependent features: Additive effects of collinearity, 2024. URL: https://arxiv.org/abs/2411.00846. arXiv:2411.00846.

[9] J. Contreras, T. Bocklitz, Explainable artificial intelligence for spectroscopy data: a review, Pflügers Archiv-European Journal of Physiology 477 (2025) 603–615.

[10] N. Rossberg, R. Gautam, K. Komolibus, B. O'Sullivan, A. Visentin, Explainable ai-based feature selection approaches for raman spectroscopy, Diagnostics 15 (2025) 2063.

[11] J. Contreras, A. Winterfeld, J. Popp, T. Bocklitz, Spectral zones-based shap/lime: Enhancing interpretability in spectral deep learning models through grouped feature analysis, Analytical Chemistry 96 (2024) 15588–15597.

[12] S. Wiklund, Spectroscopic data and multivariate analysis: tools to study genetic perturbations in poplar trees, Ph.D. thesis, Kemi, 2007.

[13] J. Felten, H. Hall, J. Jaumot, R. Tauler, A. De Juan, A. Gorzsás, Vibrational spectroscopic image analysis of biological material using multivariate curve resolution–alternating least squares (mcr-als), Nature protocols 10 (2015) 217–240.

[14] S. Maryam, A. Benazza, E. Fahy, S. K. V. Sekar, D. US, M. Olivo, R. N. Riordain, S. Andersson-Engels, G. Humbert, K. Komolibus, et al., Liquid saliva analysis using optofluidic photonic crystal fiber for detection of oral potentially malignant disorders, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 332 (2025) 125788.

[15] C. L. Li, C. J. Fisher, K. Komolibus, H. Lu, R. Burke, A. Visentin, S. Andersson-Engels, Extended-wavelength diffuse reflectance spectroscopy dataset of animal tissues for bone-related biomedical applications, Scientific Data 11 (2024) 136.