Enhanced baseline correction for Raman spectroscopy using a hybrid deep learning approach

Vu Duong[1*], Dang Cong Vinh[2], Nguyen Trong Hieu[2], Vu Tien Dung[2], Pham Hong Minh[1]

[1]*Institute of Physics, Vietnam Academy of Science and Technology, 10 Dao Tan Street, Giang Vo Ward, Hanoi, Vietnam*
[2]*University of Science, Vietnam National University - Hanoi, 334 Nguyen Trai Street, Thanh Xuan Ward, Hanoi, Vietnam*

Received 28 November 2024; revised 4 February 2025; accepted 8 April 2025

Abstract:

This research introduces an enhanced baseline correction method for Raman spectroscopy, combining a hybrid deep learning approach with traditional techniques such as polynomial fitting, Gaussian functions, and other nonlinear components. The proposed method significantly improves the signal-to-noise ratio (SNR), achieving up to a tenfold increase over raw spectra and outperforming conventional algorithms such as Imodpoly (polynomial fitting) and AirPLS (Penalised least squares). With a processing time of just 1.07 seconds, the method is well-suited for real-time applications in portable Raman spectroscopy systems. This improvement is critical in Raman spectroscopy, where background noise often obscures weak spectral features, making a high SNR essential for accurate chemical analysis. The rapid processing capability allows for immediate correction of spectral data, ensuring efficient and accurate analysis in practical applications. Thus, this hybrid approach establishes itself as a robust and effective solution for real-time Raman spectroscopy.

Keywords: **baseline correction, deep learning, Raman spectroscopy.**

Classification numbers: **1.2, 1.3, 2.3**

1. Introduction

Raman spectroscopy, a powerful analytical technique [1-3], is widely used for material characterisation and chemical analysis but suffers from weak signals that are easily obscured by background noise [4]. Accurate baseline correction is crucial; however, conventional manual methods are time-consuming, subjective, and hinder data standardisation. This necessitates a fully automated, physically realistic approach, especially for building reliable Raman databases [5]. Such automation ensures consistent data quality, enhances analytical reliability, and promotes widespread adoption, particularly in high-throughput and real-time applications.

While numerous baseline correction methods exist, they often suffer from limitations. Traditional polynomial fitting approaches (e.g., Imodpoly [6]) require manual parameter tuning and struggle with complex baselines. AirPLS [7], though automated, may still necessitate adjustments. More sophisticated techniques like MCR [8, 9] and PCA [10] can be computationally intensive or risk losing spectral information. These shortcomings highlight the need for more robust and automated solutions.

Deep learning approaches offer promise [11-14], but many methods neglect the baseline's physical origins, compromising spectral characteristics that are crucial for standardised databases. While some prioritise speed, they often sacrifice accuracy and physical fidelity. Our approach prioritises the preservation of these physical characteristics, which are essential for reliable analysis and database creation, even with slightly longer processing times. This focus on physical realism distinguishes our work from purely data-driven approaches. Building on S. Dong, et al. (2024) [15]'s significant advance in physically informed baseline modelling, we propose a novel hybrid solution that combines traditional fitting with deep learning. Our method incorporates adaptive, rule-based adjustments, achieving significant improvements in the SNR and fully automated baseline subtraction. This approach preserves the reliability and physical characteristics of Raman spectra, which are essential for standardised databases. Specifically, we leverage neural networks to define basis functions for accurate, automated baseline correction, achieving a prediction time of 0.3 seconds that is suitable for real-time applications and handheld systems.

2. Methods

In this study, we utilised our custom Python code for Raman spectroscopy denoising, which is publicly available on GitHub [16].

[*]Corresponding author: Email: duongvu@iop.vast.vn

2.1. Hybrid model

Raman spectra comprise the Raman signal, baseline, and noise. This study focuses on removing the baseline, which arises from distinct physical phenomena, including fluorescence (often modelled as exponential decay), Rayleigh scattering (pronounced at low wavenumbers), and linear backgrounds (typically arising from instrumental factors). Our model uses basis functions that reflect these physical origins, with additional components included when necessary to minimise residual error. The convolutional neural network (CNN) predicts the presence probabilities (α) of each basis function using a softmax output (see Table 1). These probabilities guide the subsequent rule-based fitting procedure. The rule-based procedure uses these probabilities to adjust the weights of the basis function, minimising the deviation between the input spectrum and the reconstructed baseline. This physically motivated approach preserves Raman peak integrity. Iteratively, the model adjusts the probabilities ($\alpha_1,\alpha_2,\ldots,\alpha_k$) and basis function parameters, reintroducing the partially corrected spectrum to the network for refinement. The process continues until the change in deviation falls below a threshold (early stopping) [7], ensuring convergence.

Table 1. Summary of the architecture for the convolutional neural network.

Layer (type)	Output shape
Input layer	(None, 880)
Reshape	(None, 880, 1)
Conv1D	(None, 876,16)
Average Pooling1D	(None, 438, 16)
Flatten	(None, 7008)
Dense	(None, 100)
Dense	(None, 3)

2.2. Data set

The training and evaluation datasets were constructed from the liquid and powder mixture dataset published by X. Fan, et al. (2023) [17]. This dataset includes six chemical samples: acetonitrile, ethanol, methanol, polyacrylamide, sodium acetate trihydrate, and sodium carbonate. The raw spectra were first cleaned and pre-processed to remove noise and baseline artefacts. To improve model robustness and simulate realistic baseline variations, we augmented the dataset by adding synthetic baselines and noise. Specifically, polynomial functions (up to degree 3) were used to model background and stray light effects, while exponential functions simulated fluorescence and Rayleigh scattering. Random Gaussian noise was also added to reflect system and environmental disturbances. This augmentation resulted in a total of 15,000 training samples and 3,200 evaluation samples, each containing the Raman signal, synthetic baseline, added noise, and ground-truth labels indicating the baseline components.

For real-world evaluation, we used an ethanol (Merck) sample, serving as a standard test case. A total of 180 raw Raman spectra were collected using the LabRAM HR Evolution (Horiba) system under diverse measurement conditions, including three excitation wavelengths (785, 633, and 532 nm), varying laser power levels, and different ethanol concentrations. This dataset was used to validate the generalisation and practical effectiveness of our baseline correction model.

2.3. Evaluation parameters

The SNR was used to assess the effectiveness of baseline correction. SNR, defined as the ratio of Raman signal strength to background noise, was calculated as follows:

$$SNR = \frac{I_{Peak} - I_{BG}}{I_{BG}} \quad (1)$$

where I_{Peak} is the mean intensity of the Raman peak, and I_{BG} is the mean intensity at nearest valley expected to be free of Raman peaks.

Figure 1 shows an ethanol Raman spectrum acquired using a 532 nm excitation laser (black open circles). Two peaks were evaluated: 878 cm^{-1} (C-H stretch) and 2924 cm^{-1} (C-H stretch). The mean intensity of the peaks around 2924 cm^{-1} (grey area) was used as a reference, and the mean intensity of the 1600-2500 cm^{-1} region was used as another reference. The raw spectrum's SNR was 1.9. The average SNR across all spectra provided a global effectiveness measure, independent of experimental conditions, allowing for SNR comparison that quantifies overall signal improvement.

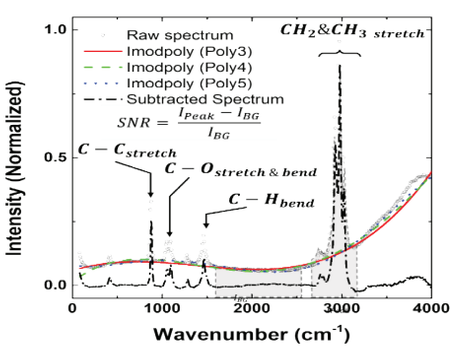

Fig. 1. The raw Raman spectrum of ethanol, with baseline fitting using the Imodpoly method. Polynomial orders of 3 (red solid line), 4 (green dashed line), and 5 (blue dotted line) are applied, illustrating the effect of increasing polynomial order on baseline correction.

The structural similarity index (SSIM) measures the similarity between the corrected spectrum and the ground truth in terms of structure, contrast, and luminance. For one-dimensional signals, SSIM is calculated as follows:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (2)$$

where μ_x, μ_y represent the means of signals x and y, respectively; σ_x^2, σ_y^2 denote their variances; and σ_{xy} is covariance between x and y. The constants C_1, C_2 are included to stabilise the division and prevent numerical instabilities when the denominators are close to zero.

The root mean square error (RMSE) is defined as the square root of the mean squared error (MSE), providing a measure of the average magnitude of the error between the predicted and true values. It retains the same unit as the original signal and is calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

where y_i is the true value, $\hat{y}_i$ the predicted value, and n is the total number of data points.

3. Results and discussion

To ensure robust performance, we evaluated the proposed method on both synthetic and real-world datasets. The synthetic set consists of 15,000 training and 3,200 evaluation spectra derived from six chemical compounds, while the real-world dataset comprises 180 ethanol spectra collected under varying experimental conditions, including different excitation wavelengths, laser power levels, and concentrations.

Table 2. Signal-to-noise ratio of the Raman spectrum after baseline subtraction.

Methods	SNR_{878}	SNR_{2924}
Raw	1.90	4.69
Poly3	25.35	25.50
Poly4	7.44	32.63
Poly5	43.07	53.41

For benchmarking purposes, we compared our method with Imodpoly, implemented using the PyBaseline library in Python [18], across multiple polynomial orders. The corresponding SNR values, calculated according to Equation 1, are presented in Table 2. Although baseline subtraction generally leads to increased SNR, indicating a better fit, the SNR values also vary noticeably with the choice of polynomial order. This variability is particularly evident for narrow Raman peaks and underscores the difficulty of optimising traditional baseline correction approaches.

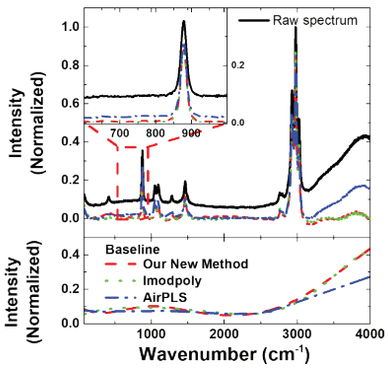

Fig. 2. Baselines fitted by automated tools: imodpoly (green dotted line), AirPLS (blue dash-dotted line), and our method (red dashed line).

Figure 2 shows baseline subtraction results using our fully automated process, compared to the AirPLS algorithm [9] (blue line). While AirPLS, based on least squares optimisation, effectively fits relatively flat regions, achieving a 30-fold increase in SNR_{2924}, it struggles with rapid signal variations, particularly in low-vibrational energy regions. For example, the improvement in SNR_{878} increased only twofold (Table 3), highlighting a key limitation of AirPLS.

Table 3. Signal-to-noise ratio of the Raman spectrum after fully automated baseline removal process.

Methods	SNR_{878}	SNR_{2924}
Raw	1.90	4.69
AirPLS	5.41	101.3
Our method	38.22	128.09

SSIM: Structural similarity index; RMSE: Root mean square error.

Table 4. Quantitative comparison of baseline correction methods using SSIM, RMSE, and processing time.

Methods	SSIM	RMSE	Processing time
Imodpoly	0.98	0.85	0.002
AirPLS	0.96	0.79	0.007
Our method	0.99	0.85	1.07

From Table 4, our proposed baseline correction method achieves superior structural preservation, as reflected in the highest SSIM value of 0.99, along with a competitive RMSE of 0.85. Although the RMSE of 0.85 is slightly higher compared to some existing methods, such as AirPLS (0.79), this can be attributed in part to the presence of synthetic random noise added during spectral generation to better simulate real-world measurement conditions. This deliberate inclusion of noise improves model robustness and generalisability but also marginally increases absolute error metrics like RMSE. Despite this, the method remains well-suited

for real-time Raman spectroscopy, thanks to its strong balance between accuracy, adaptability, and efficiency. The average processing time is 1.07 seconds, including both neural network inference and rule-based fitting. While the latter step contributes to the runtime, it enhances the model's ability to generalise across diverse spectra through dataset augmentation. In practical settings with well-characterised noise, this latency can be significantly reduced through pre-calibration and parallel processing, without compromising correction quality.

4. Conclusions

In conclusion, we present a novel, fully automated hybrid baseline correction method for Raman spectroscopy that combines deep learning with traditional signal processing. This approach leverages the speed of traditional methods and the adaptability of deep learning, achieving superior baseline correction and noise removal compared to conventional techniques. The integration of exponential and polynomial basis functions enables accurate and flexible baseline corrections, effectively capturing diverse Raman baseline profiles.

The method was validated on both synthetic data from six representative chemical compounds and real-world ethanol spectra collected under varied experimental conditions. Evaluation on the ethanol Raman dataset demonstrated an average SNR improvement of up to tenfold, significantly outperforming AirPLS in terms of accuracy and reliability. While the processing time of 1.07 seconds is slightly longer than that of purely data-driven methods, the gains in accuracy and signal fidelity are crucial for precise peak identification and for building reliable Raman databases. Future work will focus on further enhancing SNR and expanding applicability to more complex spectral challenges.

CRediT author statement

Vu Duong: Conceptualisation, Methodology, Writing original draft, Writing - Reviewing and Editing, Project administration; Dang Cong Vinh: Investigation, Data curation, Coding, Model development, Writing - Reviewing and Editing; Nguyen Trong Hieu: Validation, Formal analysis, Discussion; Vu Tien Dung: Discussion, Supervision; Pham Hong Minh: Funding acquisition, Supervision.

ACKNOWLEDGEMENTS

This research was funded by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number NCUD.01-2019.14.

COMPETING INTERESTS

The authors declare that there is no conflict of interest regarding the publication of this article.

REFERENCES

[1] E. Smith, G. Dent (2019), "Modern Raman spectroscopy", *Modern Raman Spectroscopy*, 256pp, DOI: 10.1002/9781119440598.

[2] S.M. Asiala, N.C. Shand, K. Faulds, et al. (2017), "Surface-enhanced, spatially offset Raman spectroscopy (SESORS) in tissue analogues", *ACS Appl. Mater. Interfaces*, **9(30)**, pp.25488-25494, DOI: 10.1021/acsami.7b09197.

[3] M.S. Bergholt, K. Lin, J. Wang, et al. (2016), "Simultaneous fingerprint and high-wavenumber fiber-optic Raman spectroscopy enhances real-time *in vivo* diagnosis of adenomatous polyps during colonoscopy", *J. Biophotonics*, **9(4)**, pp.333-342, DOI: 10.1002/jbio.201400141.

[4] W. Min, X. Gao (2023), "Raman scattering and vacuum fluctuation: An Einstein-coefficient-like equation for Raman cross sections", *J. Chem. Phys.*, **159(19)**, DOI: 10.1063/5.0171382/2921424.

[5] R.T. Vulchi, V. Morgunov, R. Junjuri, et al. (2024), "Artifacts and anomalies in Raman spectroscopy: A review on origins and correction procedures", *Molecules*, **29(19)**, DOI: 10.3390/molecules29194748.

[6] J. Zhao, H. Lui, D.I. Mclean, et al. (2007), "Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy", *Appl. Spectrosc.*, **61(11)**, pp.1225-1232, DOI: 10.1366/000370207782597003.

[7] Z.M. Zhang, S. Chen, Y.Z. Liang (2010), "Baseline correction using adaptive iteratively reweighted penalized least squares", *Analyst*, **135(5)**, pp.1138-1146, DOI: 10.1039/b922045c.

[8] A. de Juan, R. Tauler (2006), "Multivariate curve resolution (MCR) from 2000: Progress in concepts and applications", *Crit. Rev. Anal. Chem.*, **36(3-4)**, pp.163-176, DOI: 10.1080/10408340600970005.

[9] M. Dadashi, H. Abdollahi, R. Tauler (2012), "Maximum likelihood principal component analysis as initial projection step in multivariate curve resolution analysis of noisy data", *Chemom. Intell. Lab. Syst.*, **118**, pp.33-40, DOI: 10.1016/j.chemolab.2012.07.009.

[10] C. Guyon, T. Bouwmans, E. Zahzah (2012), "Robust principal component analysis for background subtraction: Systematic evaluation and comparative analysis", *Principal Component Analysis*, DOI: 10.5772/38267.

[11] J. Shen, M. Li, Z. Li, et al. (2022), "Single convolutional neural network model for multiple preprocessing of Raman spectra", *Vib. Spectrosc.*, **121**, DOI: 10.1016/j.vibspec.2022.103391.

[12] J. Wahl, M. Sjödahl, K. Ramser (2020), "Single-step preprocessing of Raman spectra using convolutional neural networks", *Appl. Spectrosc.*, **74(4)**, pp.427-438, DOI: 10.1177/0003702819888949.

[13] M.T. Gebrekidan, C. Knipfer, A.S. Braeuer (2021), "Refinement of spectra using a deep neural network: Fully automated removal of noise and background", *J. Raman Spectrosc.*, **52(3)**, pp.723-736, DOI: 10.1002/jrs.6053.

[14] N. Iqbal (2022), "DeepSeg: Deep segmental denoising neural network for seismic data", *IEEE Trans. Neural Netw. Learn. Syst.*, **34(7)**, pp.3397-3404, DOI: 10.1109/TNNLS.2022.3205421.

[15] S. Dong, Y. Liu, H. Yu, et al. (2024), "An iterative curve-fitting baseline correction method for Raman spectra driven by neural network", *Appl. Spectrosc.*, **78(1)**, pp.111-119, DOI: 10.1177/00037028231212941.

[16] Ving2k2 (2024), *RamanDenoise*, GitHub, https://github.com/Ving2k2/RamanDenoise, accessed 2 August 2024.

[17] X. Fan, Y. Wang, C. Yu, et al. (2023), "A universal and accurate method for easily identifying components in Raman spectroscopy based on deep learning", *Anal. Chem.*, **95(11)**, pp.4863-4870, DOI: 10.1021/acs.analchem.2c03853.

[18] D. Erb (2024), "Pybaselines: A Python library of algorithms for the baseline correction of experimental data", DOI: 10.5281/zenodo.10676584.