RamanSPy: Augmenting Raman Spectroscopy Data Analysis with AI

Dimitar Georgiev [1 2] **Simon Vilms Pedersen** [2] **Ruoxiao Xie** [2] **Álvaro Fernández-Galiana** [2] **Molly M. Stevens** [2] **Mauricio Barahona** [3]

Abstract

Raman spectroscopy is a non-destructive and label-free chemical analysis technique, which plays a key role in the analysis and discovery cycle of various branches of life and physical sciences. Recently, there has been a marked increase in the adoption of machine learning techniques in Raman spectroscopic analysis. Nonetheless, progress in the area is still impeded by the lack of software, methodological and data standardisation, and the ensuing fragmentation and lack of reproducibility of analysis workflows thereof. To address these issues, we introduce *RamanSPy*, an open-source Python package for Raman spectroscopic data analysis, which supports day-to-day tasks, integrative analyses, the development of methods and protocols, and the integration of advanced data analytics. *RamanSPy* is highly modular, not tied to a particular technology or data format, and can be readily interfaced with the burgeoning ecosystem for data science, statistical analysis and machine learning in Python. *RamanSPy* is hosted at `https://github.com/barahona-research-group/RamanSPy`, supplemented with extended online documentation, available at `https://ramanspy.readthedocs.io`, that includes tutorials, example applications, and details about the real-world research applications presented in this paper.

[1]Department of Computing & UKRI Centre for Doctoral Training in AI for Healthcare, Imperial College London, London, United Kingdom, SW7 2AZ [2]Department of Materials, Department of Bioengineering & Institute of Biomedical Engineering, Imperial College London, London, United Kingdom, SW7 2AZ [3]Department of Mathematics, Imperial College London, London, United Kingdom, SW7 2AZ. Correspondence to: Molly M. Stevens <m.stevens@imperial.ac.uk>, Mauricio Barahona <m.barahona@imperial.ac.uk>.

Accepted at the 5th AI for Science Workshop at ICML 2024. Copyright 2024 by the author(s).

1. Introduction

Raman spectroscopy (RS) is a powerful sensing modality based on inelastic light scattering, which provides qualitative and quantitative chemical analysis with high sensitivity and specificity (Colthup, 2012). RS yields a characterisation of the vibrational profile of molecules, which can help elucidate the composition of chemical compounds, biological specimens and materials (McCreery, 2005; Shipp et al., 2017; Fernandez-Galiana et al., 2023). In contrast to most conventional technologies for (bio)chemical characterisation (e.g., staining, different omics, fluorescence microscopy and mass spectrometry), RS is both label-free and non-destructive, thereby allowing the acquisition of rich biological and chemical information without compromising the structural and functional integrity of probed samples. This advantage has enabled a broad range of applications of RS across biomedical and pharmaceutical research (Smith et al., 2016; Shipp et al., 2017; Cialla-May et al., 2017; Wang et al., 2018; Vankeirsbilck et al., 2002; Auner et al., 2018; Kong et al., 2015), materials science (Weber & Merlin, 2013; Kumar, 2012), environmental science (Halvorson & Vikesland, 2010; Ong et al., 2020), and others (Pang et al., 2016; Chalmers et al., 2012; Terry et al., 2022).

An area of topical interest is the frontier of Raman spectroscopy, chemometrics and artificial intelligence (AI), with its promise of more autonomous, flexible and data-driven RS analytics (Pan et al., 2022; Luo et al., 2022; Lussier et al., 2020). There has been a recent surge in the adoption of AI methods in Raman-based research (Fernandez-Galiana et al., 2023), with applications to RS now spanning domains as broad as the identification of pathogens and other microbes (Ho et al., 2019; Lu et al., 2020; Yan et al., 2021; Wang et al., 2021); the characterisation of chemicals, including minerals (Carey et al., 2015), pesticides (Zhu et al., 2021) and other analytes (Han & Ram, 2020; Akpolat et al., 2020); the development of novel diagnostic platforms (Ralbovsky & Lednev, 2020; Talari et al., 2019; Heng et al., 2021; Zhang et al., 2022); as well as the application of techniques from computer vision for denoising and super-resolution in Raman imaging (Horgan et al., 2021).

Yet, progress in the area is still hindered by practical factors stemming from the restrictive, functionally disparate,

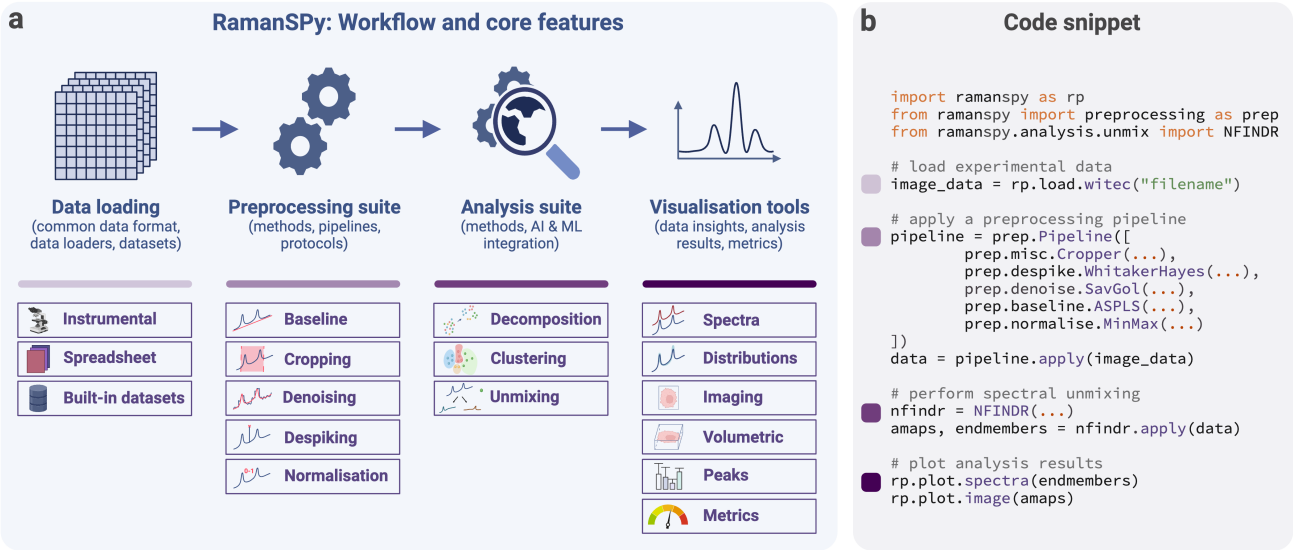

Figure 1. Core infrastructure of *RamanSPy*. a, *RamanSPy* provides a comprehensive library of standardised, simple-to-use procedures for data loading, preprocessing, analysis and visualisation organised within a modular and extensible architecture. b, An example workflow use case in *RamanSPy*: Raman data is loaded, preprocessed and analysed in a few lines of code.

and highly encapsulated nature of current commercial software for RS data analysis. RS data analysis often operates within proprietary software environments and data formats, which have induced methodological inconsistencies and reduced cross-platform and integrative efforts, with growing concerns around reproducibility (Byrne et al., 2016; Tanwar et al., 2021; Barton et al., 2022; Möller et al., 2017; Ntziouni et al., 2022). These restrictions have hampered the development, validation and deployment of emerging AI-based technologies for RS.

In response to these challenges, we have developed *RamanSPy* - a modular, open-source framework for Raman Spectroscopy analytics in Python. *RamanSPy* is designed to systematise day-to-day workflows, enhance algorithmic development, integration and reproducibility, and accelerate the adoption of AI technologies into the RS field.

2. Core infrastructure of RamanSPy

RamanSPy is based on a modular, object-oriented programming infrastructure comprising a comprehensive collection of pre-defined tools for RS data analysis, which streamlines the analysis life cycle and reduces computational barriers to RS analytics (Figure 1).

Data loading and management. The framework adopts a scalable array-based data representation based on *NumPy* (Harris et al., 2020), which accommodates different spectroscopic modalities, including single-point spectra, Raman imaging data, and volumetric scans. Experimental data can be loaded through custom loaders built into *RamanSPy* or through standard tools available in Python. The data representation functions as a common data container that facilitates the integrative analysis of data across setups and vendors, independent of instrumental origin and acquisition modality, and defines the interface between RS data and analysis tools within and beyond *RamanSPy*.

Preprocessing, analysis and visualisation. On top of its data management infrastructure, *RamanSPy* provides an extensive toolbox for preprocessing, analysis and visualisation. The preprocessing suite includes techniques for denoising, baseline correction, cosmic spike removal, normalisation and background subtraction, among others. Likewise, the analysis toolbox includes modules for decomposition (useful for dimensionality reduction), clustering and spectral unmixing. *RamanSPy* also includes a set of data visualisation tools, intended to facilitate routine visualisation and exploratory analysis. All these modules are organised into a common class structure, which standardises their application across projects and datasets to facilitate transferable analysis workflows. Note that this suite is highly flexible and designed to cater to a wide range of requirements, applications and user profiles.

Automated pipelining of spectral preprocessing protocols. Due to a lack of standardisation and frameworks that streamline the preprocessing of RS data (Byrne et al., 2016), researchers tend to utilise variable preprocessing protocols, often dispersed across different software systems (Rozenstein et al., 2014; Alshdaifat et al., 2021). This significantly affects reproducibility and validation, especially in the con-

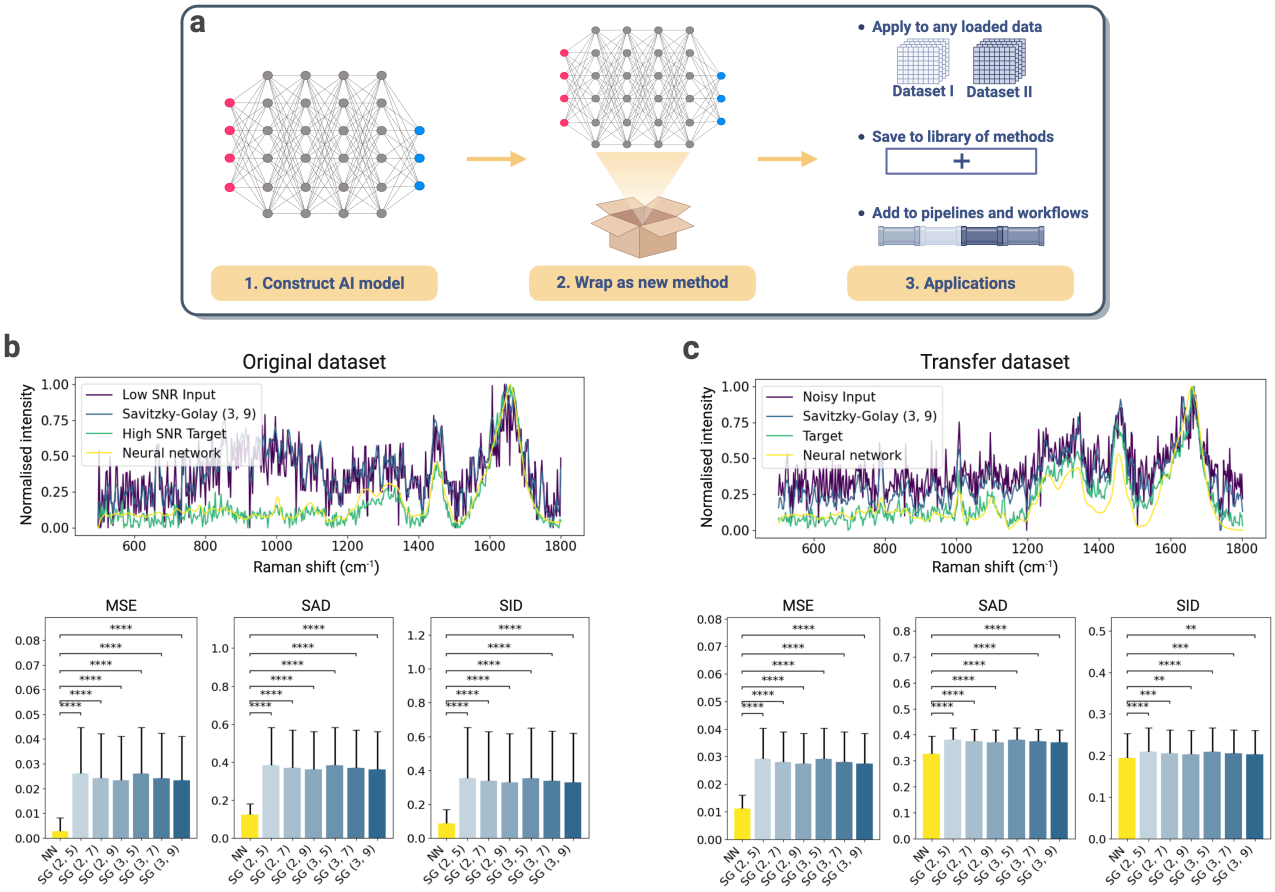

Figure 2. RamanSPy interfaces with AI/ML Python frameworks to create new methods for RS analysis. a, *RamanSPy* allows users to incorporate AI/ML models seamlessly into pipelines created within the platform. b, A pre-trained 1D ResUNet deep-learning denoiser (Horgan et al., 2021) is integrated as a preprocessing module within *RamanSPy* to investigate its performance against the Savitzky-Golay (SG) filter (Savitzky & Golay, 1964). top, Denoising of a spectrum from Horgan et al. (2021), where the low-SNR (purple) is the input and the high-SNR (green) is the target. The data is denoised with an SG filter of polynomial order 3 and kernel size 9, SG(3, 9) (blue), and with the implemented deep-learning denoiser (yellow). bottom, The results on the test set from Horgan et al. (2021) ($n = 12694$) show that the deep-learning denoiser outperforms six SG filters across three performance metrics (MSE, SAD, SID). Error bars represent one standard deviation around the sample mean. Statistical significance measured with a two-sided Wilcoxon signed-rank test with adjustment for multiple comparisons based on Benjamini-Hochberg correction (Benjamini & Hochberg, 1995) (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$). c, Same analysis on unseen data from Kallepitis et al. (2017) ($n = 1600$). The input (purple) corresponds to data contaminated with added noise and the target (green) to the original data. In this case, the deep-learning denoiser only shows an improvement for MSE.

text of AI model development.

To facilitate the creation of reproducible protocols, *RamanSPy* incorporates a pipelining infrastructure, which systematises the process of creating, customising and executing preprocessing pipelines. Users can use a specialised class, which defines a generic, multi-layered preprocessing procedure, to assemble pipelines from selected built-in preprocessing modules or other in-house methods. To reduce overhead, constructed pipelines are designed to function exactly as any single method, i.e., they are fully compatible with the rest of the modules and data structures in the package. Furthermore, pipelines can be easily saved, reused and shared (e.g., upon publication) to foster the development of a shared repository of preprocessing protocols. As a seed to this repository, *RamanSPy* provides a library of assembled preprocessing protocols (custom pre-defined, or adapted from the literature (Bergholt et al., 2016)), which users can access and exploit.

3. AI and Raman Spectroscopy: Bridging the gap with *RamanSPy*

To help accelerate the adoption of AI technologies for RS analysis, *RamanSPy* is endowed with a permeable archi-

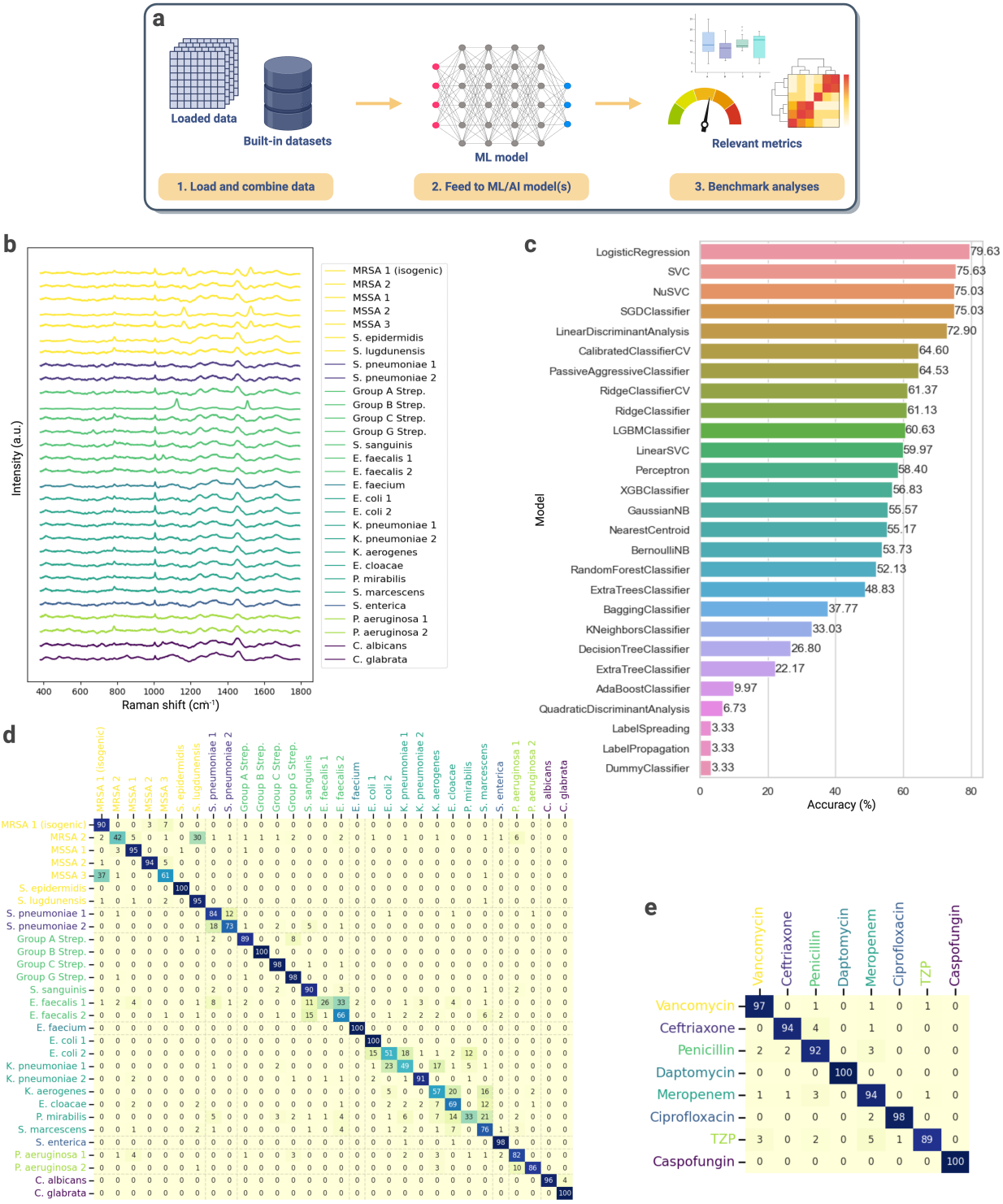

Figure 3. RamanSPy as a suite for algorithmic development. a, Data representations in *RamanSPy* are compatible with the Python AI/ML ecosystem, allowing data flow from *RamanSPy* to *scikit-learn* (Pedregosa et al., 2011), *PyTorch* (Paszke et al., 2019), *tensorflow* (Abadi et al., 2016), etc. *RamanSPy* is also equipped with standard datasets and relevant metrics to support model development and validation. b-e, Benchmarking ML classification models on the task of bacteria identification using Raman spectra from Ho et al. (2019). b, Mean Raman spectra of each bacterial species in the dataset used for training. Spectra are min-max normalised to the range 0–1 for visualisation purposes. c, Benchmarking results of 28 ML models. The best accuracy was achieved by the logistic regression classifier. d-e, Confusion matrices for the best species-level (d) and antibiotic-level (e) classifier with accuracies of 79.63% and 94.63%, respectively.

tecture that streamlines the interface between Raman spectroscopic data and the burgeoning machine learning (ML) ecosystem in Python. This is complemented by access to built-in datasets and performance metrics to further support the development and testing of new models and algorithms. We show below two examples of *RamanSPy*'s capabilities for ML integration and validation.

AI integration. First, *RamanSPy* allows the seamless integration of methods from standard AI/ML frameworks in Python (e.g., *scikit-learn* (Pedregosa et al., 2011), *PyTorch* (Paszke et al., 2019) and *tensorflow* (Abadi et al., 2016)) as tools for RS analysis (Figure 2a). As an illustration of how custom methods can be integrated into analysis pipelines within *RamanSPy*, we use our package to construct a deep learning denoiser based on the one-dimensional ResUNet model - a fully convolutional UNet neural network with residual connections, presented in Horgan et al. (2021). To do this, we simply wrap within *RamanSPy* the pre-trained neural network[1] as a custom denoising method. Once wrapped, the denoiser is automatically compatible with the rest of *RamanSPy* and can be readily employed for different applications. For instance, we replicated the results in Horgan et al. (2021), and show in Figure 2b that the application of this deep-learning denoiser to the low signal-to-noise ratio (SNR) test set from Horgan et al. (2021) consistently outperforms the commonly-used Savitzky-Golay filter (Savitzky & Golay, 1964). This is quantified by various metrics also coded within *RamanSPy* (e.g., mean squared error (MSE), spectral angle distance (SAD) (Kruse et al., 1993) and spectral information divergence (SID) (Chang, 1999)), which we use to measure the performance of each denoising method by comparing denoised signals against the provided high SNR data, which act as a reference.

Importantly, applying this pipeline to new data only involves changing the data source. Taking advantage of this transferability, we test the denoiser on unseen volumetric Raman data from another cell line (THP-1 (Kallepitis et al., 2017)), to which we added Gaussian noise. For these data, Figure 2c shows improved denoising performance according to the MSE metric, which is dependent on normalisation and scale, but with lower significance according to scale-invariant metrics available in *RamanSPy*. This example emphasises the sensitivity of algorithms to data shifts and the importance of incorporating robust validation criteria based on the unique requirements of each application. The collection of metrics that *RamanSPy* provides is intended to serve as a starting resource to test performance according to different objectives.

[1]Model was trained on spectra from MDA-MB-231 breast cancer cells (Horgan et al., 2021). Deposited by authors at: https://github.com/conor-horgan/DeepeR.

AI interoperability. Secondly, the data management backbone of *RamanSPy* ensures a direct data flow to the rest of the Python ecosystem, i.e., data can be loaded, preprocessed, and analysed in *RamanSPy* and then exported to conduct further modelling and analysis elsewhere (Figure 3a). As an example application, we perform AI-based bacteria identification using Raman measurements from 30 bacterial and yeast isolates as provided in Ho et al. (2019) (Figure 3b). After loading and visualising the spectra with *RamanSPy*, we interface the data with the *lazypredict* Python package (Pandala), which allows us to directly benchmark 28 ML classification models (including logistic regression, support vector machines and decision trees) on the task of predicting the bacterial species from a given spectrum. The models were first trained on the provided fine-tuning dataset (100 spectra per isolate) and then tested on the unseen test set of the same size. Our benchmarking analysis in Figure 3c finds logistic regression as the best-performing model, achieving a classification accuracy of 79.63% on the species-level classification task (Figure 3d), and 94.63% for antibiotic treatment classification (Figure 3e).

Dataset suite for model evaluation. To further assist the process of testing and evaluating new computational approaches, *RamanSPy* provides access to a library of curated datasets collated from existing literature for different tasks (e.g., classification, denoising, Raman imaging). With this library, we aim to seed the growth of a common repository of RS datasets that helps reduce barriers to data access, especially for ML teams with limited access to RS instruments (Luo et al., 2022). The dataset library in *RamanSPy* already includes data loaders for Raman data from bacterial species (Ho et al., 2019), cell lines (Horgan et al., 2021; Kallepitis et al., 2017), COVID-19 samples (Yin et al., 2021; 2020), multi-instrument Surface Enhanced Raman Spectroscopy (SERS) measurements of adenine samples (Fornasaro et al., 2020), wheat lines (ŞEN et al., 2023), minerals (Lafuente et al., 2015), and will continue to be expanded. Recognizing the potential benefits that synthetic and surrogate datasets can provide for the generation of controlled ground truths for algorithmic validation, we have also integrated the synthetic Raman data generator described in Georgiev et al. (2024) within *RamanSPy*.

4. Conclusion

In this paper, we have introduced *RamanSPy* - a computational framework for integrative Raman spectroscopic data analysis aimed at overcoming the limitations of currently available commercial software tools in terms of accessibility, flexibility and reproducibility, and facilitating the adoption and validation of advanced AI technologies for next-generation RS analysis.

The codebase of *RamanSPy* is fully open-source and disseminated under a permissive license that allows for unrestricted use, adaptation, and extension, including for commercial purposes. It is supplemented with extended online documentation containing a comprehensive selection of tutorials and example applications, as well as further information about the analyses presented in this paper. We believe this will be critical for the continuous development of the platform and its adoption across different scientific domains, including biomedical research, chemistry, and materials science, among others. Future directions include the expansion of our suite of built-in methods, tools and datasets; the incorporation of cutting-edge AI technologies into the framework as the field progresses; and the integration of the package into experimental setups and other software solutions.

Data availability

All data used in this article are previously published open-access data that have been deposited by the respective authors online. Instructions on how to access, download, and load the data sets provided in *RamanSPy* are available in the documentation at https://ramanspy.readthedocs.io/en/latest/datasets.html.

Code availability

The codebase of *RamanSPy* is open-source and hosted on GitHub at https://github.com/barahona-research-group/RamanSPy. The package can be installed via pip using 'pip install ramanspy'. Documentation, including detailed tutorials and examples, is available at https://ramanspy.readthedocs.io. The scripts used to produce the analysis results presented in this paper are also provided as executable Jupyter Notebook examples at https://github.com/barahona-research-group/RamanSPy/tree/3dd2c1e09420c5ac473a72ebd6ed06a91c30a85c/paper_reproducibility and as part of the documentation of *RamanSPy* at https://ramanspy.readthedocs.io/en/latest/auto_examples/index.html.

Acknowledgements

D.G. is supported by UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1]. S.V.P. gratefully acknowledges support from the Independent Research Fund Denmark (0170-00011B). R.X. and M.M.S. acknowledge support from the Engineering and Physical Sciences Research Council (EP/P00114/1 and EP/T020792/1). A.F.-G. acknowledges support from the Schmidt Science Fellows, in partnership with the Rhodes Trust. M.M.S. acknowledges support from the Royal Academy of Engineering Chair in Emerging Technologies award (CiET2021\94). M.B. acknowledges support by the Engineering and Physical Sciences Research Council under grant EP/N014529/1, funding the EPSRC Centre for Mathematics of Precision Healthcare at Imperial College London, and under grant EP/T027258/1. The authors thank Dr Akemi Nogiwa Valdez for proofreading and data management support. Figures were assembled in BioRender.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. TensorFlow: A System for Large-Scale Machine Learning. In *Osdi*, volume 16, pp. 265–283. Savannah, GA, USA, 2016.

Akpolat, H., Barineau, M., Jackson, K. A., Akpolat, M. Z., Francis, D. M., Chen, Y.-J., and Rodriguez-Saona, L. E. High-Throughput Phenotyping Approach for Screening Major Carotenoids of Tomato by Handheld Raman Spectroscopy Using Chemometric Methods. *Sensors*, 20(13): 3723, 2020.

Alshdaifat, E., Alshdaifat, D., Alsarhan, A., Hussein, F., and El-Salhi, S. M. F. S. The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance. *Data*, 6(2):11, 2021.

Auner, G. W., Koya, S. K., Huang, C., Broadbent, B., Trexler, M., Auner, Z., Elias, A., Mehne, K. C., and Brusatori, M. A. Applications of Raman spectroscopy in cancer diagnosis. *Cancer and Metastasis Reviews*, 37(4): 691–717, 2018.

Barton, B., Thomson, J., Diz, E. L., and Portela, R. Chemometrics for Raman Spectroscopy Harmonization. *Applied Spectroscopy*, 76(9):1021–1041, 2022.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bergholt, M. S., St-Pierre, J.-P., Offeddu, G. S., Parmar, P. A., Albro, M. B., Puetzer, J. L., Oyen, M. L., and Stevens, M. M. Raman spectroscopy reveals new insights into the zonal organization of native and tissue-engineered articular cartilage. *ACS central science*, 2(12): 885–895, 2016.

Byrne, H. J., Knief, P., Keating, M. E., and Bonnier, F. Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells. *Chemical Society Reviews*, 45(7):1865–1878, 2016.

Carey, C., Boucher, T., Mahadevan, S., Bartholomew, P., and Dyar, M. Machine learning tools for mineral recognition and classification from Raman spectroscopy. *Journal of Raman Spectroscopy*, 46(10):894–903, 2015.

Chalmers, J. M., Edwards, H. G. M., and Hargreaves, M. D. (eds.). *Infrared and Raman Spectroscopy in Forensic Science*. Wiley Online Library, 2012.

Chang, C.-I. Spectral information divergence for hyperspectral image analysis. In *IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No. 99CH36293)*, volume 1, pp. 509–511. IEEE, 1999.

Cialla-May, D., Zheng, X.-S., Weber, K., and Popp, J. Recent progress in surface-enhanced Raman spectroscopy for biological and biomedical applications: from cells to clinics. *Chemical Society Reviews*, 46(13):3945–3961, 2017.

Colthup, N. *Introduction to infrared and Raman spectroscopy*. Elsevier, 2012.

Fernandez-Galiana, A., Bibikova, O., Pedersen, S., and Stevens, M. Fundamentals and applications of Raman-based techniques for the design and development of active biomedical materials. *Advanced Materials*, 2023. doi: 10.1002/adma.202210807.

Fornasaro, S., Alsamad, F., Baia, M., Batista de Carvalho, L. A., Beleites, C., Byrne, H. J., Chiadò, A., Chis, M., Chisanga, M., Daniel, A., et al. Surface enhanced Raman spectroscopy for quantitative analysis: results of a large-scale European multi-instrument interlaboratory study. *Analytical chemistry*, 92(5):4053–4064, 2020.

Georgiev, D., Fernández-Galiana, A., Pedersen, S. V., Papadopoulos, G., Xie, R., Stevens, M. M., and Barahona, M. Hyperspectral unmixing for Raman spectroscopy via physics-constrained autoencoders. *arXiv preprint arXiv:2403.04526*, 2024.

Halvorson, R. A. and Vikesland, P. J. Surface-enhanced Raman spectroscopy (SERS) for environmental analyses, 2010.

Han, N. and Ram, R. J. Bayesian modeling and computation for analyte quantification in complex mixtures using Raman spectroscopy. *Computational Statistics & Data Analysis*, 143:106846, 2020.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Heng, H. P. S., Shu, C., Zheng, W., Lin, K., and Huang, Z. Advances in real-time fiber-optic Raman spectroscopy for early cancer diagnosis: Pushing the frontier into clinical endoscopic applications. *Translational Biophotonics*, 3 (1):e202000018, 2021.

Ho, C.-S., Jean, N., Hogan, C. A., Blackmon, L., Jeffrey, S. S., Holodniy, M., Banaei, N., Saleh, A. A., Ermon, S., and Dionne, J. Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nature communications*, 10(1):1–8, 2019.

Horgan, C. C., Jensen, M., Nagelkerke, A., St-Pierre, J.-P., Vercauteren, T., Stevens, M. M., and Bergholt, M. S. High-Throughput Molecular Imaging via Deep-Learning-Enabled Raman Spectroscopy. *Analytical chemistry*, 93 (48):15850–15860, 2021.

Kallepitis, C., Bergholt, M. S., Mazo, M. M., Leonardo, V., Skaalure, S. C., Maynard, S. A., and Stevens, M. M. Quantitative volumetric Raman imaging of three dimensional cell cultures. *Nature communications*, 8(1):1–9, 2017.

Kong, K., Kendall, C., Stone, N., and Notingher, I. Raman spectroscopy for medical diagnostics—from in-vitro biofluid assays to in-vivo cancer detection. *Advanced drug delivery reviews*, 89:121–134, 2015.

Kruse, F. A., Lefkoff, A., Boardman, J., Heidebrecht, K., Shapiro, A., Barloon, P., and Goetz, A. The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. *Remote sensing of environment*, 44(2-3):145–163, 1993.

Kumar, C. S. *Raman spectroscopy for nanomaterials characterization*. Springer Science & Business Media, 2012.

Lafuente, B., Downs, R. T., Yang, H., and Stone, N. The power of databases: The RRUFF project. In *Highlights in mineralogical crystallography*, pp. 1–30. De Gruyter, 2015.

Lu, W., Chen, X., Wang, L., Li, H., and Fu, Y. V. Combination of an artificial intelligence approach and laser tweezers Raman spectroscopy for microbial identification. *Analytical chemistry*, 92(9):6288–6296, 2020.

Luo, R., Popp, J., and Bocklitz, T. Deep Learning for Raman Spectroscopy: A Review. *Analytica*, 3(3):287–301, 2022.

Lussier, F., Thibault, V., Charron, B., Wallace, G. Q., and Masson, J.-F. Deep learning and artificial intelligence

methods for Raman and surface-enhanced Raman scattering. *TrAC Trends in Analytical Chemistry*, 124:115796, 2020.

McCreery, R. L. *Raman spectroscopy for chemical analysis*. John Wiley & Sons, 2005.

Möller, S., Prescott, S. W., Wirzenius, L., Reinholdtsen, P., Chapman, B., Prins, P., Soiland-Reyes, S., Klötzl, F., Bagnacani, A., Kalaš, M., et al. Robust cross-platform workflows: how technical and scientific communities collaborate to develop, test and share best practices for data analysis. *Data Science and Engineering*, 2(3):232–244, 2017.

Ntziouni, A., Thomson, J., Xiarchos, I., Li, X., Bañares, M. A., Charitidis, C., Portela, R., and Lozano Diz, E. Review of Existing Standards, Guides, and Practices for Raman Spectroscopy. *Applied Spectroscopy*, 76(7):747–772, 2022.

Ong, T. T., Blanch, E. W., and Jones, O. A. Surface Enhanced Raman Spectroscopy in environmental analysis, monitoring and assessment. *Science of The Total Environment*, 720:137601, 2020.

Pan, L., Zhang, P., Daengngam, C., Peng, S., and Chongcheawchamnan, M. A review of artificial intelligence methods combined with Raman spectroscopy to identify the composition of substances. *Journal of Raman Spectroscopy*, 53(1):6–19, 2022.

Pandala, S. R. Lazy Predict help build a lot of basic models without much code and helps understand which models works better without any parameter tuning. URL https://github.com/shankarpandala/lazypredict.

Pang, S., Yang, T., and He, L. Review of surface enhanced Raman spectroscopic (SERS) detection of synthetic chemical pesticides. *TrAC Trends in Analytical Chemistry*, 85:73–82, 2016.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in neural information processing systems*, 32, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Ralbovsky, N. M. and Lednev, I. K. Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning. *Chemical Society Reviews*, 49(20):7428–7453, 2020.

Rozenstein, O., Paz-Kagan, T., Salbach, C., and Karnieli, A. Comparing the effect of preprocessing transformations on methods of land-use classification derived from spectral soil measurements. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6):2393–2404, 2014.

Savitzky, A. and Golay, M. J. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

ŞEN, A., Kecoglu, I., Ahmed, M., Parlatan, U., and Unlu, M. Differentiation of advanced generation mutant wheat lines: Conventional techniques versus Raman spectroscopy. *Frontiers in Plant Science*, 14, 2023.

Shipp, D. W., Sinjab, F., and Notingher, I. Raman spectroscopy: techniques and applications in the life sciences. *Advances in Optics and Photonics*, 9(2):315–428, 2017.

Smith, R., Wright, K. L., and Ashton, L. Raman spectroscopy: an evolving technique for live cell studies. *Analyst*, 141(12):3590–3600, 2016.

Talari, A. C., Rehman, S., and Rehman, I. U. Advancing cancer diagnostics with artificial intelligence and spectroscopy: identifying chemical changes associated with breast cancer. *Expert review of molecular diagnostics*, 19(10):929–940, 2019.

Tanwar, S., Paidi, S. K., Prasad, R., Pandey, R., and Barman, I. Advancing Raman spectroscopy from research to clinic: Translational potential and challenges. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 260:119957, 2021.

Terry, L. R., Sanders, S., Potoff, R. H., Kruel, J. W., Jain, M., and Guo, H. Applications of surface-enhanced Raman spectroscopy in environmental detection. *Analytical Science Advances*, 3(3-4):113–145, 2022.

Vankeirsbilck, T., Vercauteren, A., Baeyens, W., Van der Weken, G., Verpoort, F., Vergote, G., and Remon, J. Applications of raman spectroscopy in pharmaceutical analysis. *TrAC Trends in Analytical Chemistry*, 21(12):869–877, 2002. ISSN 0165-9936. doi: https://doi.org/10.1016/S0165-9936(02)01208-6.

Wang, L., Liu, W., Tang, J.-W., Wang, J.-J., Liu, Q.-H., Wen, P.-B., Wang, M.-M., Pan, Y.-C., Gu, B., and Zhang, X. Applications of Raman spectroscopy in bacterial infections: principles, advantages, and shortcomings. *Frontiers in Microbiology*, 12:683580, 2021.

Wang, W.-T., Zhang, H., Yuan, Y., Guo, Y., and He, S.-x. Research progress of Raman spectroscopy in drug analysis. *AAPS PharmSciTech*, 19(7):2921–2928, 2018.

Weber, W. H. and Merlin, R. *Raman scattering in materials science*, volume 42. Springer Science & Business Media, 2013.

Yan, S., Wang, S., Qiu, J., Li, M., Li, D., Xu, D., Li, D., and Liu, Q. Raman spectroscopy combined with machine learning for rapid detection of food-borne pathogens at the single-cell level. *Talanta*, 226:122195, 2021.

Yin, G., Li, L., Lu, S., Yin, Y., Su, Y., Zeng, Y., Luo, M., Ma, M., Zhou, H., Yao, D., Liu, G., and Lang, J. Data and code on serum Raman spectroscopy as an efficient primary screening of coronavirus disease in 2019 (COVID-19). *Figshare*, 2020. doi: 10.6084/m9.figshare.12159924.v1.

Yin, G., Li, L., Lu, S., Yin, Y., Su, Y., Zeng, Y., Luo, M., Ma, M., Zhou, H., Orlandini, L., et al. An efficient primary screening of COVID-19 by serum Raman spectroscopy. *Journal of Raman Spectroscopy*, 52(5):949–958, 2021.

Zhang, L., Li, C., Peng, D., Yi, X., He, S., Liu, F., Zheng, X., Huang, W. E., Zhao, L., and Huang, X. Raman spectroscopy and machine learning for the classification of breast cancers. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 264:120300, 2022.

Zhu, J., Sharma, A. S., Xu, J., Xu, Y., Jiao, T., Ouyang, Q., Li, H., and Chen, Q. Rapid on-site identification of pesticide residues in tea by one-dimensional convolutional neural network coupled with surface-enhanced Raman scattering. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 246:118994, 2021.