RamanSPy: Augmenting Raman Spectroscopy Data Analysis with AI

Dimitar Georgiev — *Imperial College London*

Simon Vilms Pedersen — *Imperial College London*; *Present: University of Southern Denmark*

Ruoxiao Xie — *Imperial College London*; *Present: University of Liverpool*

Álvaro Fernández-Galiana — *Imperial College London*; *Present: University of Oxford*

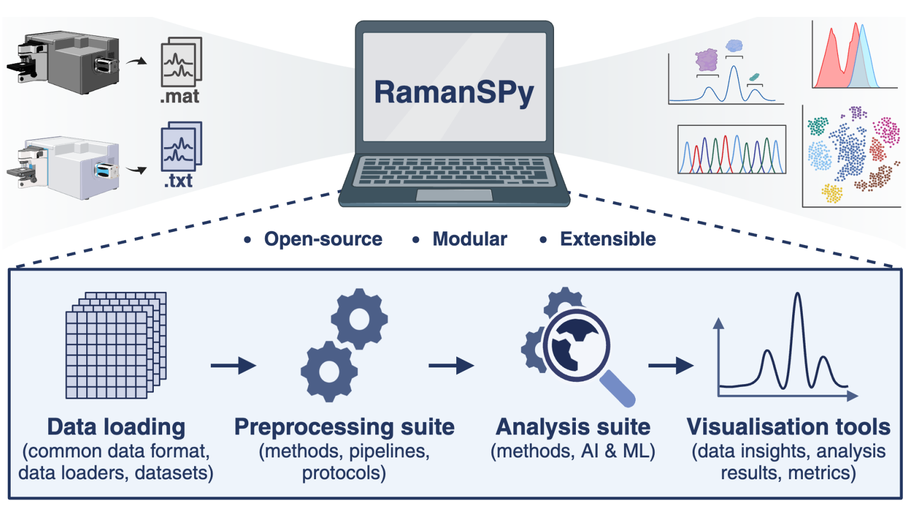
Molly M. Stevens — *Imperial College London*; *Present: University of Oxford*

Mauricio Barahona — *Imperial College London*

TL;DR

We introduce RamanSPy - an open-source **Python package for Raman spectroscopy analytics**, designed to systematise **day-to-day workflows**, enhance **algorithmic development**, **integration** and **reproducibility**, and accelerate the **adoption of AI** technologies in the field. We showcase the core features of RamanSPy through real-world research applications.

Project website ***Codebase*** ***Full paper***



Background

Raman spectroscopy (RS) is a powerful sensing modality based on inelastic light scattering, which enables non-destructive and label-free chemical analysis. As such, RS plays a key role in the analysis and discovery cycle of various branches of life and physical sciences.

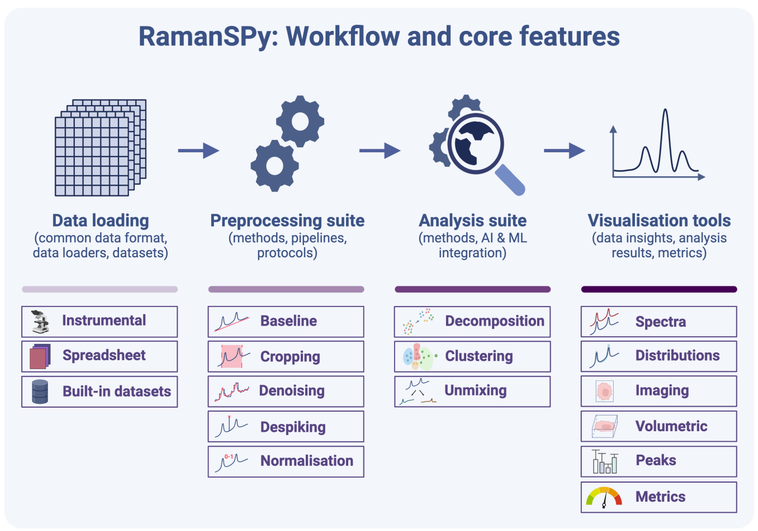
An area of topical interest is the frontier of **Raman spectroscopy, chemometrics and artificial intelligence (AI)**, promising more autonomous, flexible and data-driven RS analytics.

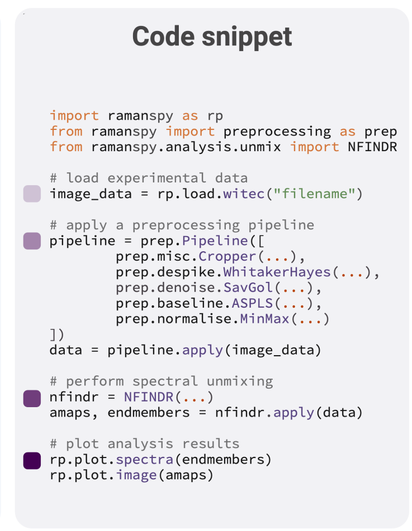
Yet, progress in the area is still impeded by the **lack of software, methodological and data standardisation**, and the ensuing **fragmentation** and **lack of reproducibility** of analysis workflows thereof.

Core infrastructure of RamanSPy

To overcome these challenges, we developed an open-source Python package called RamanSPy.

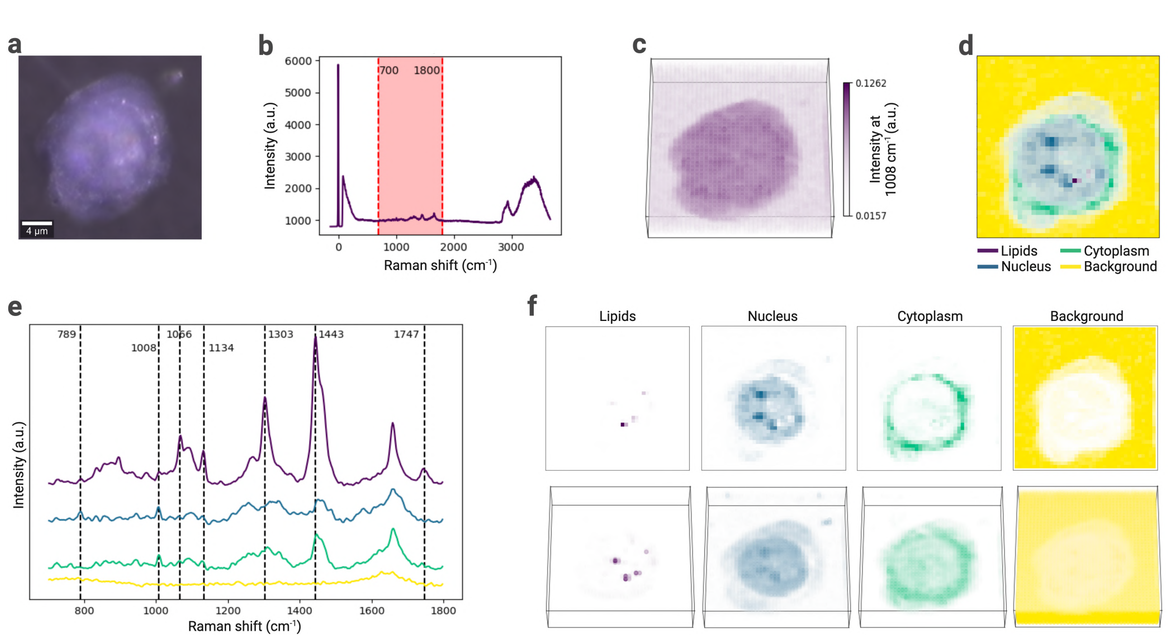
RamanSPy is based on a **modular** infrastructure which comprises a **comprehensive collection of built-in tools** for RS data analysis, including methods for data loading, preprocessing, analysis, and visualisation (*left*). This toolbox streamlines the analysis life cycle and reduces computational barriers to RS analytics (*right*).





```python
import ramanspy as rp
from ramanspy import preprocessing as prep
from ramanspy.analysis.unmix import NFINDR

# load experimental data
image_data = rp.load.witec("filename")

# apply a preprocessing pipeline
pipeline = prep.Pipeline([
        prep.misc.Cropper(...),
        prep.despike.WhitakerHayes(...),
        prep.denoise.SavGol(...),
        prep.baseline.ASPLS(...),
        prep.normalise.MinMax(...)
])
data = pipeline.apply(image_data)

# perform spectral unmixing
nfindr = NFINDR(...)
amaps, endmembers = nfindr.apply(data)

# plot analysis results
rp.plot.spectra(endmembers)
rp.plot.image(amaps)
```

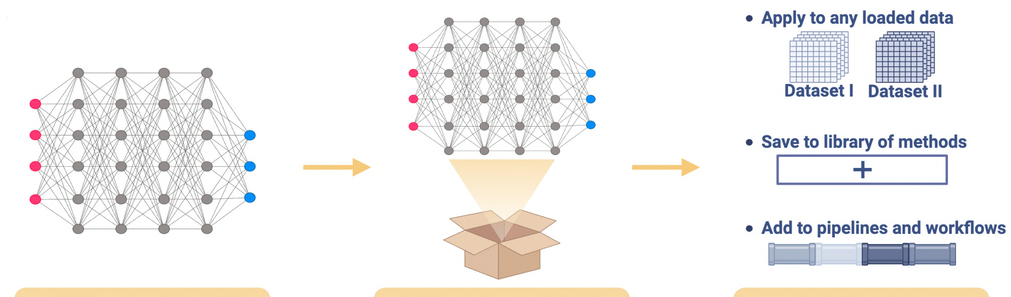
Real-world application: We use RamanSPy to analyse a volumetric Raman imaging scan of a human leukaemia monocytic cell[1] and study the morphology of the cell in a label-free manner.



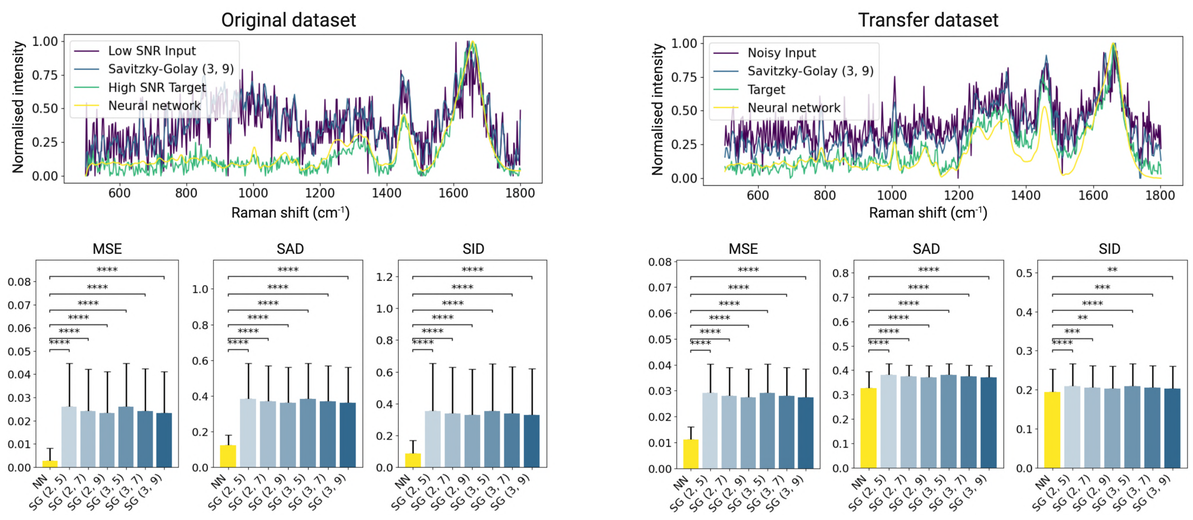
a, Bright-field image of the cell. **b**, Exemplar measurement. **c**, Volumetric slice across the 1008 cm^{-1} band (characteristic of proteins) after preprocessing. **d**, Reconstruction of the cell derived via hyperspectral unmixing, together with the corresponding endmembers (**e**) and fractional abundances (**f**).

AI integration for next-generation Raman analytics

AI integration: RamanSPy is endowed with a permeable architecture that streamlines the integration of methods from standard frameworks for data science, statistical analysis, and machine learning in Python into pipelines and workflows within the package.



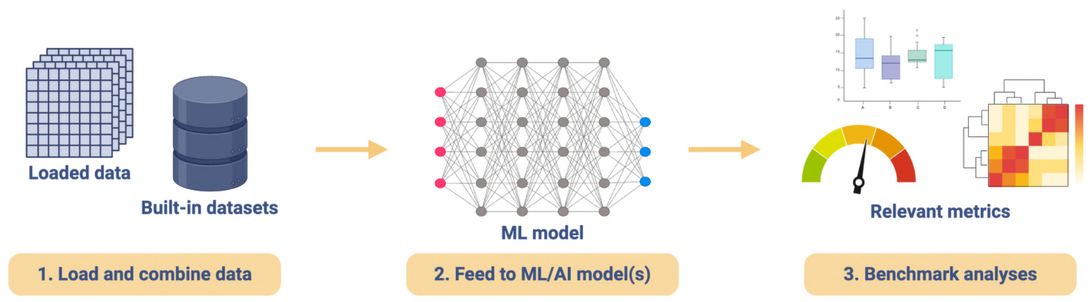
Real-world application: Using RamanSPy, we construct a deep-learning denoiser based on a pre-trained 1D ResUNet model[2]. We test the neural network (NN) denoiser against standard Savitzky–Golay (SG) filters[3] on: the original Raman data from breast cancer cells[2] (*left*); and unseen Raman data from a human leukaemia monocytic cell[1] with added Gaussian noise (*right*).



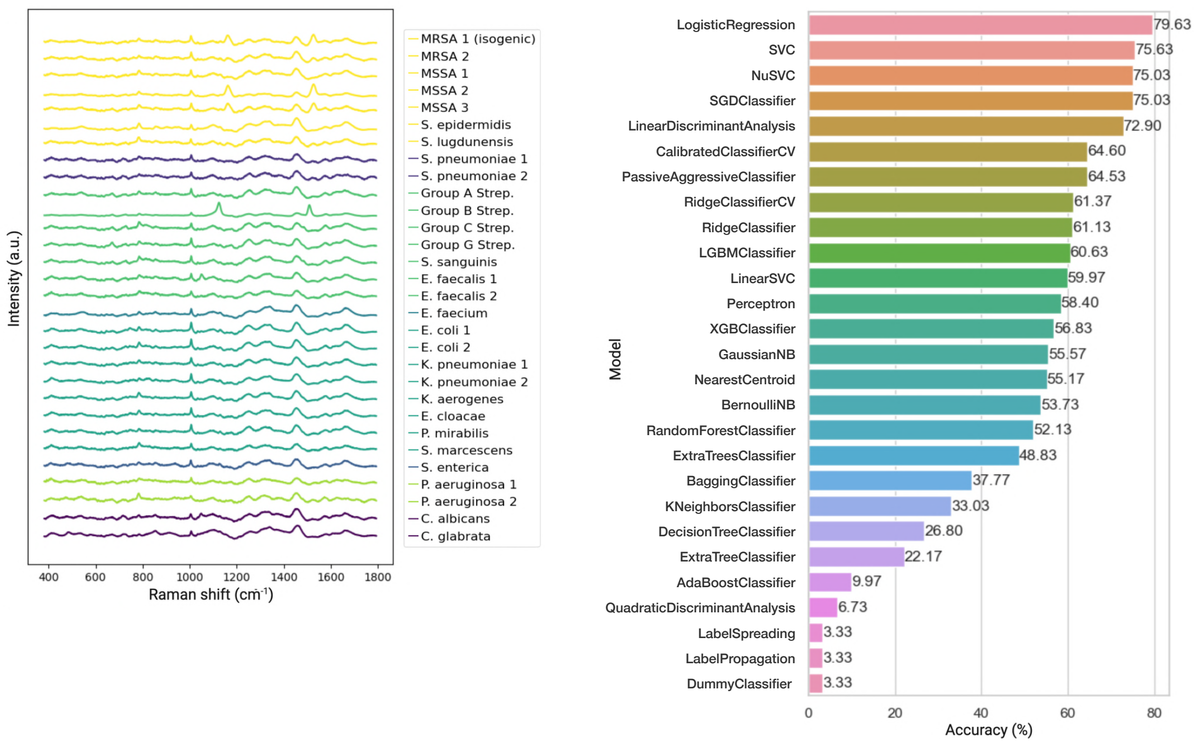
MSE:=mean squared error; SAD:=spectral angle distance[4]; SID:=spectral information divergence[5]

RamanSPy as a suite for model development

AI validation: RamanSPy provides access to a library of curated datasets and performance metrics. This suite lays the foundations of a common repository of RS datasets that reduces barriers to data access and supports model development and validation.



Real-world application: We use one of the datasets built into RamanSPy, which comprises Raman measurements from 30 bacterial and yeast isolates[6] (*left*), to benchmark 28 machine learning (ML) classification models (including logistic regression, support vector machines, and decision trees) on the task of bacteria identification. Our benchmarking analysis finds logistic regression as the best-performing model, achieving a classification accuracy of 79.63% (*right*).



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. Figures were assembled in BioRender.

References

[1] Kallepitis C, Bergholt MS, Mazo MM, Leonardo V, Skaalure SC, Maynard SA, Stevens MM. *Nature Communications* **2017**, 8, 14843.
[2] Horgan CC, Jensen M, Nagelkerke A, St-Pierre JP, Vercauteren T, Stevens MM, Bergholt MS. *Analytical Chemistry* **2021**, 93, 15850–15860.
[3] Savitzky A, Golay MJ. *Analytical Chemistry* **1964**, 36, 1627-1639.
[4] Kruse FA, Lefkoff AB, Boardman YJ, Heidebrecht KB, Shapiro AT, Barloon PJ, Goetz AF. *Remote Sensing of Environment* **1993**, 44, 145–163.
[5] Chang CI. *IEEE International Geoscience and Remote Sensing Symposium* **1999**, 1, 509–511.
[6] Ho CS, Jean N, Hogan CA, Blackmon L, Jeffrey SS, Holodniy M, Banaei N, Saleh AA, Ermon S, Dionne J. *Nature Communications* **2019**, 10, 1–8.