IMPERIAL

From Photons to Phenotypes: Decoding Organoid Biochemistry via Deep Learning-Enhanced Raman Spectroscopy

Dimitar Dimitrov Georgiev

Supervisors:

Prof. Mauricio Barahona

Prof. Dame Molly M. Stevens

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Department of Computing, Imperial College London

Abstract

Three-dimensional neural organoids have emerged as powerful *in vitro* models for studying human brain development, disease and drug response. Yet their analysis remains constrained by standard imaging and characterisation techniques, which are invasive, require exogenous labelling and offer limited multiplexing. There is a pressing need for advanced analytical and imaging approaches that can interrogate spatial organisation and dynamics in intact organoid systems under physiologically relevant settings, without prior knowledge of specific molecular targets or the use of external labels.

To address this challenge, this thesis develops computational methods for non-invasive, label-free biochemical analysis using Raman spectroscopy (RS). First, it introduces RamanSPy – an open-source Python package for RS data analysis designed to improve standardisation, reproducibility, accessibility and interoperability. RamanSPy provides a comprehensive library of tools for spectroscopic analysis that supports day-to-day tasks, integrative analyses, method and protocol development, and the integration of advanced data analytics. Next, hyperspectral unmixing algorithms for RS based on autoencoder neural networks are developed and systematically validated using synthetic and experimental benchmark datasets. Unmixing autoencoders consistently outperform standard methods, achieving more accurate qualitative and quantitative analysis and enabling enhanced volumetric Raman imaging of a monocytic cell.

Building on these tools, a Raman imaging platform for unsupervised, spatially resolved biochemical analysis in neural organoids is established. This approach enables non-invasive, label-free mapping of cellular and subcellular structures in both cryosectioned and intact organoids. Using this platform, volumetric imaging of a neural rosette within a neural organoid is demonstrated. Furthermore, changes in biochemical composition during early developmental stages in intact neural organoids are interrogated, revealing spatiotemporal variations in lipids, proteins and nucleic acids.

Overall, this work expands the methodological and computational toolkit for RS chemometrics and establishes a versatile framework for high-content, label-free (bio)chemical phenotyping with broad applications in organoid research, cell biology, disease modelling and drug discovery.

Statement of Originality

I hereby certify that this thesis and the research to which it refers are the product of my own work. Any ideas, contributions or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with standard referencing practices.

Dimitar Georgiev

Copyright

© The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

This thesis marks the end of an incredible journey. Like most PhDs, it was far from smooth, shaped by many personal and professional challenges. Yet through it all, it has brought immense growth, joy and meaning. Without a doubt, it has been one of the most rewarding experiences of my life, and one I will always cherish. None of it would have been possible without the enormous support, guidance and inspiration of many. It truly takes a village – and I have been fortunate to be part of an exceptional one.

First and foremost, I would like to express my deepest gratitude to my supervisors, Professor Mauricio Barahona and Professor Dame Molly Stevens. Thank you for the opportunity and the freedom to pursue a project that excited me every day, and for your unwavering support throughout the journey. This work would simply not have been possible without your help. Thank you for believing in my ideas, even when I struggled to, and for helping me grow – both as a researcher and a person. Your guidance shaped this project, and me, in lasting ways.

I am also sincerely grateful to Dr Ruoxiao Xie, Dr Álvaro Fernández-Galiana, Dr Simon Vilms Pedersen and Dr Daniel Reumann for their mentorship and invaluable contributions to this project. Your advice and feedback have been critical to this work, and I have learned so much from each of you. I truly hope that – despite my countless late-night questions and endless drafts – you enjoyed working with me as much as I enjoyed working with you.

To all members – past and present – of the Stevens and Barahona groups across Imperial and Oxford: thank you for the science, the discussions and the memories. I have been incredibly lucky to work alongside such talented, inspiring and supportive people. You made every day in the office and lab not only stimulating but genuinely fun. I am truly grateful to everyone whose input, questions or feedback helped shape this work. Special thanks to the TEA time (later SofTEAs) subgroup within the Stevens group for countless helpful discussions and suggestions. Many thanks as well to the Stevens group's outstanding technical and support staff, whose work behind the scenes ensured everything ran smoothly. A very special thanks goes to Dr Akemi Nogiwa Valdez for her support with manuscript preparation, data management and so much more throughout

the past four years.

Beyond my own PhD project, I have also been fortunate to contribute to many other truly transformative projects. I want to thank the many collaborators I have had the privilege to work with – far too many to name here.

I would like to specifically acknowledge a few people whose experimental and technical contributions were essential to this work: Dr Álvaro Fernández-Galiana and Georgios Papadopoulos (sugar mixtures and Raman spectroscopy); Dr Ruoxiao Xie, Dr Daniel Reumann and Xiaoyu Zhao (neural organoid culture and Raman imaging); Dr Ruoxiao Xie and Dr Daniel Reumann (immunostaining and fluorescence microscopy); Ellena O'Keeffe (immunostaining of a neural organoid cryosection); and Kevin Giraldo Rodriguez (Raman imaging of a neural organoid cryosection).

I gratefully acknowledge the support of UK Research and Innovation through the UKRI Centre for Doctoral Training in AI for Healthcare (AI4Health CDT) at Imperial College London (Grant No. EP/S023283/1), without which this work would not have been possible. A special thanks to Britta Ross, Safa Abu-Sway, Dr Ahmed Fetit, Professor Aldo Faisal and the entire CDT team for creating such a vibrant and supportive community. The conferences I had the opportunity to attend, along with all the CDT events, remain the most memorable and formative experiences of my PhD.

To my family – thank you for the love and care that have surrounded me throughout my whole life. Your support has meant the world to me, and I am deeply grateful for all of it. I would like to thank my parents in particular – for everything they have given me, and for everything they gave up so that I could pursue this path. Thank you for devoting so much of your lives so I could build mine. You have made all of this possible, and for that, I am forever grateful. I dedicate this thesis to you.

Last but by no means least, I would like to thank Katrin – for your love, patience and unwavering support. Thank you for being by my side through every high and low. You made the hard moments easier and the good ones unforgettable. Thank you for encouraging me when I doubted myself, and for sharing in the joy of every milestone, big and small. Thank you for reminding me every day of what truly matters, and for helping me grow into the person I am today. I am deeply, endlessly grateful.

Contents

1 Introduction **16**

1.1 Thesis aim and objectives 17

1.2 Thesis outline 18

1.3 Publications 18

2 Background **21**

2.1 Chemical analysis with Raman spectroscopy 21

2.1.1 Light-matter interactions and the Raman effect 21

2.1.2 Raman spectroscopy techniques 27

2.1.3 Applications of Raman spectroscopy 31

2.1.4 Data processing and interpretation in Raman spectroscopy 34

2.1.5 Chemical analysis via Raman spectroscopy and chemometrics . . 42

2.2 Neural organoids: Modelling the human brain in a dish 51

2.2.1 The need for better biological models of the human brain 51

2.2.2 From 2D to 3D neural cell cultures 55

2.2.3 Neural organoids 57

2.2.4 Applications of neural organoids in brain research 60

2.2.5 Organoid imaging and analysis 63

3 RamanSPy: An open-source Python package for integrative Raman spectroscopy data analysis **71**

3.1 Introduction . . . 71
3.2 RamanSPy – a platform for general Raman spectroscopy analytics . . . 72
3.2.1 Core features of RamanSPy . . . 72
3.2.2 Raman cell imaging via hyperspectral unmixing using RamanSPy 74
3.3 RamanSPy enables automated pipelining of spectral preprocessing protocols 75
3.3.1 Preprocessing pipelining in RamanSPy . . . 76
3.3.2 Pipelining illustration . . . 77
3.3.3 Computational efficiency . . . 77
3.4 RamanSPy facilitates AI integration and validation of next-generation Raman data analytics . . . 78
3.4.1 AI integration: Deep learning denoising . . . 78
3.4.2 AI interoperability: Deep learning bacteria classification . . . 80
3.4.3 Dataset suite . . . 82
3.5 Conclusion . . . 82

4 Hyperspectral unmixing for Raman spectroscopy via physics-constrained autoencoders 84

4.1 Introduction . . . 84
4.2 Raman unmixing with autoencoder neural networks . . . 85
4.2.1 Unmixing autoencoders . . . 85
4.2.2 Autoencoder design . . . 87
4.2.3 Model training and evaluation . . . 90
4.2.4 Baseline methods for comparison . . . 91
4.2.5 Implementation . . . 91
4.3 Benchmarking unmixing autoencoders on synthetic Raman mixtures . . . 91

4.3.1 Generating synthetic Raman mixtures 91
4.3.2 Synthetic datasets for validation 93
4.3.3 Benchmark results 95
4.3.4 Computational efficiency 98
4.4 Validation of unmixing autoencoders on experimental Raman data from sugar mixtures 99
4.4.1 Data collection 100
4.4.2 Preprocessing and analysis 101
4.4.3 Benchmark results 102
4.5 Application of unmixing autoencoders to biological data: volumetric Raman imaging of a THP-1 cell 106
4.5.1 Spectral preprocessing 106
4.5.2 Comparative unmixing analysis 106
4.6 Conclusion 108

5 Label-free biochemical imaging and timepoint analysis of neural organoids via deep learning-enhanced Raman microspectroscopy 109

5.1 Introduction 110
5.2 Deep learning-enhanced Raman microspectroscopy 110
5.3 Methods and Materials 112
5.3.1 Organoid culture and preparation 112
5.3.2 Immunofluorescence staining and microscopy 114
5.3.3 Raman microspectroscopy 114
5.3.4 Spectral preprocessing of Raman imaging data 115
5.3.5 Hyperspectral unmixing analysis 115

5.4 Label-free imaging in intact cerebral organoids via deep learning-enhanced Raman microspectroscopy . . . 117

5.5 Experimental validation and optimisation against fluorescence microscopy 121

5.5.1 Qualitative validation using paired fluorescence microscopy and Raman imaging data from a neural organoid cryosection . . . 121

5.5.2 Pipeline optimisation, ablation studies and comparison with standard unmixing methods . . . 123

5.5.3 Raman imaging of a second neural organoid cryosection . . . 124

5.5.4 Quantitative validation against fluorescence microscopy . . . 127

5.5.5 Robustness to experimental noise . . . 129

5.6 Biochemical imaging in intact neural organoids at subcellular resolution . 131

5.7 Volumetric *in situ* imaging of a neural rosette within an intact neural organoid . . . 133

5.8 Tracing changes in biochemical composition during early neurodevelopment in intact neural organoids . . . 135

5.9 Conclusion . . . 141

6 Discussion **142**

6.1 Thesis achievements . . . 142

6.2 Broader significance . . . 143

6.3 Limitations . . . 144

6.4 Future research directions . . . 145

6.5 Closing remarks . . . 148

Bibliography **149**

A Classical derivation of Raman scattering **186**

B RamanSPy: Supplementary information **188**

B.1 List of features in RamanSPy 189

B.2 RamanSPy manual 194

C Unmixing Raman autoencoders: Supplementary information **201**

D Organoid imaging: Supplementary information **208**

List of Figures

2.1 Molecular fingerprinting via Raman scattering – schematic overview . . . 24

2.2 Schematic of a spontaneous Raman spectroscopy system 28

2.3 Overview of common spectral preprocessing steps in Raman spectroscopy 37

2.4 Overview of key application areas of neural organoids 61

3.1 General Raman spectroscopy workflow and core features of RamanSPy . 73

3.2 Morphological analysis of a THP-1 cell via spectral unmixing with RamanSPy 75

3.3 Spectral preprocessing pipelining in RamanSPy 76

3.4 RamanSPy interfaces with AI/ML Python frameworks to create new methods for RS analysis 79

3.5 RamanSPy as a suite for algorithmic development 81

4.1 Hyperspectral unmixing for Raman spectroscopy using autoencoder neural networks 86

4.2 Generation of synthetic Raman mixtures 92

4.3 Key measures and statistics of benchmark datasets used for validation of unmixing autoencoders 94

4.4 Similarity analysis of the benchmark datasets used for validation of unmixing autoencoders 96

4.5 Benchmarking autoencoders on synthetic Raman mixtures 98

4.6 Computational efficiency analysis 99

4.7 Experimental validation on Raman spectroscopy data from sugar mixtures 102

4.8 Endmember estimates on high signal-to-noise ratio sugar data 104

4.9 Endmember estimates on low signal-to-noise ratio sugar data 105

4.10 Volumetric Raman imaging of a THP-1 cell 107

5.1 Schematic diagram of developed organoid imaging pipeline. 111

5.2 Schematic diagram of organoid culture 113

5.3 Deep learning-enhanced Raman microspectroscopy enables label-free *in situ* biochemical imaging in intact neural organoids in 2D and 3D 118

5.4 Collection of paired Raman imaging and fluorescence microscopy data from neural organoid cryosections for validation 120

5.5 Raman imaging of a neural organoid cryosection produces reconstructions aligned with fluorescence microscopy 122

5.6 Effect of endmember initialisation on autoencoder performance 123

5.7 Effect of training loss on autoencoder performance 125

5.8 Comparative analysis of different chemometric methods on Raman imaging data from a neural organoid cryosection 126

5.9 Hyperspectral unmixing analysis of a second neural organoid cryosection 127

5.10 Autoencoder-based unmixing enables enhanced morphochemical Raman imaging of neural organoid section as validated against fluorescence microscopy 128

5.11 Comparative analysis of different chemometric methods in the presence of higher degrees of experimental noise 130

5.12 Biochemical imaging in intact neural organoids at subcellular resolution . 132

5.13 Volumetric *in situ* biochemical imaging of a neural rosette within an intact neural organoid 134

5.14 Timepoint analysis of intact neural organoids: Study design 136

5.15 Preparation of timepoint Raman data for quantitative analysis 137

5.16 Tracing changes in biochemical composition during early neurodevelopment in intact neural organoids 138

5.17 Abundance images of the two most abundant endmembers across timepoint Raman spectroscopy scans 140

6.1 Hyperspectral unmixing analysis of broadband CARS imaging data from a cancer cell 147

C.1 Full unmixing results obtained with VCA + NNLS on THP-1 cell data . 205

C.2 Full unmixing results obtained with Dense AE model on THP-1 cell data 206

C.3 Full unmixing results obtained with Deep Dense AE model on THP-1 cell data 207

D.1 Full unmixing results for low-resolution volumetric Raman imaging scan of an intact day-20 neural organoid 210

D.2 Full unmixing results for Raman imaging scan of a day-30 neural organoid cryosection 211

D.3 Full unmixing results for Raman imaging scan of a second day-30 neural organoid cryosection 212

D.4 Full unmixing results for high-resolution Raman imaging scan of an intact day-10 neural organoid 213

D.5 Full unmixing results for volumetric Raman imaging scan of a neural rosette within an intact day-20 neural organoid 214

D.6 Full unmixing results for timepoint Raman imaging data 215

D.7 Comparative analysis on broadband CARS imaging data from a cancer cell 216

List of Tables

2.1 Comparison of commonly used techniques in cellular and tissue imaging . 70

4.1 Comparison of autoencoders and standard unmixing techniques 88

B.1 Main features of RamanSPy (October 2025) 189

C.1 Benchmark results on synthetic Raman mixture datasets 202

C.2 Benchmark results on Raman spectroscopy data from sugar solutions . . 203

C.3 Benchmark results on Raman spectroscopy data from sugar solutions – endmember similarity based on the Pearson correlation coefficient 204

D.1 Primary antibodies used for immunofluorescence staining 209

D.2 Secondary antibodies used for immunofluorescence staining 209

Chapter 1

Introduction

The ability to probe and visualise the inner workings of cells and tissues has shaped modern biology, driving some of the most transformative advances across the biological and medical sciences. Imaging and characterisation technologies have been instrumental in revealing how cells divide, differentiate and assemble into complex tissues, how these processes are disrupted in disease, and how therapies may one day repair what nature has broken. To this end, a broad array of techniques, ranging from histological staining and immunofluorescence to omics-based profiling, has been developed, each offering distinct insights across spatial and molecular scales[1–6]. Nonetheless, **a fundamental challenge persists: most available methods are inherently invasive and/or require molecular labelling**.

These constraints become particularly limiting as scientific research shifts toward complex, three-dimensional (3D) living tissue models, such as organoids. Over the past decade, neural organoids derived from pluripotent stem cells have emerged as powerful biological models of the developing human brain, recapitulating key features of early tissue development and organisation[7–10]. Compared to two-dimensional (2D) neural cell culture models, neural organoids exhibit enhanced complexity and three-dimensional structure, providing a more advanced representation of tissue architecture[7,10], cellular diversity[11,12] and functional network formation[13]. Consequently, neural organoids have spurred new opportunities for studying human brain development, disease and drug response *in vitro*[14–17]. Yet, despite growing interest, organoid research remains constrained by the limitations of conventional methods for tissue imaging and characterisation.

There is a pressing need for techniques that can interrogate spatial organisation and dynamics in intact organoid systems under physiologically relevant settings, without prior knowledge of specific molecular targets or the use of exogenous labels.

Raman spectroscopy (RS), a technique within the field of vibrational spectroscopy, offers a promising alternative for biochemical analysis. Unlike fluorescence-based imaging, RS provides molecular contrast through the analysis of the intrinsic inelastic light scattering of molecules, thereby enabling non-invasive, label-free chemical characterisation with minimal sample preparation[18–20]. Consequently, RS has become a ubiquitous analytical technique across the life and physical sciences[21–31], with increasing use in analysing biological specimens such as cells and tissues[32–37].

Nonetheless, the application of RS to complex biological systems such as organoids remains challenging due to significant practical, methodological and analytical issues. At the core of these is the analysis of RS data itself[38,39]. Measurements from heterogeneous biological samples yield noisy, highly convoluted spectra that are difficult to interpret for qualitative and quantitative analysis[40–42]. Conventional chemometric approaches often lack the sensitivity and fidelity needed to extract biologically meaningful chemical information in such settings. While machine learning (ML) has emerged as a powerful tool for chemometric analysis in RS, current methods are largely limited to constrained, supervised tasks and are therefore not well-suited to biological contexts where ground truth is rarely available[43]. Moreover, their integration into RS workflows remains technically challenging as the field continues to face broader challenges in accessibility, reproducibility and interoperability, stemming from the lack of robust software infrastructure, analysis frameworks and data standardisation[44]. Addressing these limitations is essential for realising the potential of RS for organoid characterisation and supporting its adoption across biomedical research more broadly.

1.1 Thesis aim and objectives

Thesis aim: The aim of this thesis is to advance the methodological and computational toolkit of Raman spectroscopy and enable non-invasive, label-free biochemical imaging in intact neural organoids to investigate early human brain development.

In particular, the thesis sets out to achieve the following objectives:

- Improve standardisation, reproducibility and interoperability in RS data analysis.
- Streamline the interface between RS data and machine learning frameworks.
- Develop methods for more accurate RS-based chemical imaging and analysis.
- Achieve subcellular-resolution, label-free imaging in intact neural organoids using non-invasive RS.

- Investigate neuroarchitecture and spatiotemporal variations in biomolecular composition during early neurodevelopment in neural organoids.

1.2 Thesis outline

The remainder of this thesis is structured as follows:

Chapter 2 provides essential background on Raman spectroscopy and organoid models. It reviews current methodologies and limitations in RS analysis and existing approaches for imaging and characterising organoids. In particular, it highlights the key challenges addressed in subsequent chapters.

Chapter 3 introduces RamanSPy, an open-source Python package for integrative RS data analysis. It outlines the design, architecture and core functionalities of RamanSPy, and demonstrates its utility through real-world applications in cell biology and microbiology.

Chapter 4 presents a framework for RS chemometric analysis based on hyperspectral unmixing autoencoders. Multiple AE models for unmixing are developed and benchmarked against standard methods on synthetic and experimental datasets.

Chapter 5 establishes an integrated imaging platform for non-invasive, label-free organoid analysis, combining Raman microspectroscopy with deep learning-based unmixing. The chapter describes the development and validation of the pipeline and investigates RS-based biochemical imaging and timepoint analysis in neural organoids.

Chapter 6 summarises the key contributions of the thesis and reflects on their implications for RS analysis, organoid imaging and computational spectroscopy. It also outlines future directions for extending the tools and frameworks developed.

1.3 Publications

This thesis is based on the following three manuscripts – two published and one submitted for publication – listed in the order in which they are featured:

1) **RamanSPy: An open-source Python package for integrative Raman spectroscopy data analysis.**

   Dimitar Georgiev, Simon Vilms Pedersen, Ruoxiao Xie, Àlvaro Fernàndez-Galiana, Molly M. Stevens, Mauricio Barahona.
   Analytical Chemistry 2024, 96(21), 8492–8500.

This work has also been presented at the following conferences (in reverse chronological order):

- *BioMedEng24 Conference, poster presentation, London, UK, 2024*
- *AI for Science Workshop at International Conference on Machine Learning (ICML), poster presentation, Vienna, Austria, 2024*
- *Machine Learning for Life and Material Science Workshop at International Conference on Machine Learning (ICML), poster presentation, Vienna, Austria, 2024*

2) **Hyperspectral unmixing for Raman spectroscopy via physics-constrained autoencoders.**

Dimitar Georgiev, Àlvaro Fernàndez-Galiana, Simon Vilms Pedersen, Ruoxiao Xie, Georgios Papadopoulos, Molly M. Stevens, Mauricio Barahona.
Proceedings of the National Academy of Sciences of the United States of America (PNAS) 2024, 121(45) e2407439121.

This work has also been presented at the following conferences (in reverse chronological order):

- *AI4X Conference, oral presentation, Singapore, 2025*
- *RamanFest Conference, poster presentation (Best PhD Poster award), Paris, France, 2024*
- *BioMedEng24 Conference, oral presentation, London, UK, 2024*
- *AI for Science Workshop at International Conference on Machine Learning (ICML), oral presentation, Vienna, Austria, 2024*
- *Machine Learning for Life and Material Science Workshop at International Conference on Machine Learning (ICML), poster presentation, Vienna, Austria, 2024*
- *Joint CDT Conference on AI for Healthcare (CAI4H), poster presentation, Edinburgh, UK, 2024*
- *Joint CDT Conference on AI for Healthcare (CAI4H), poster presentation, York, UK, 2023*

3) **Label-free biochemical imaging and timepoint analysis of neural organoids via deep learning-enhanced Raman microspectroscopy.**

Dimitar Georgiev, Ruoxiao Xie, Daniel Reumann, Xiaoyu Zhao, Àlvaro Fernàndez-Galiana, Mauricio Barahona, Molly M. Stevens. (submitted).
Preprint available on bioRxiv: 10.1101/2025.09.29.679057, 2025.

This work has also been presented at the following conferences (in reverse chronological order):

- *Imageomics Workshop at Conference on Neural Information Processing Systems (NeurIPS), poster presentation, San Diego, USA, 2025*
- *AI Virtual Cells and Instruments Workshop at Conference on Neural Information Processing Systems (NeurIPS), poster presentation, San Diego, USA, 2025*
- *AI for Science Workshop at Conference on Neural Information Processing Systems (NeurIPS), poster presentation, San Diego, USA, 2025*
- *Joint CDT Conference on AI for Healthcare (CAI4H), poster presentation, York, UK, 2025*
- *Learning Meaningful Representations of Life Workshop at International Conference on Learning Representations (ICLR), poster presentation, Singapore, 2025*

Chapter 2

Background

To situate the contributions of this thesis within the broader scientific landscape, relevant background and existing literature are first reviewed. In particular, the principles of Raman spectroscopy and its application in chemical analysis are discussed in Section 2.1, followed by an overview of neural organoids as models of early human neurodevelopment in Section 2.2. Throughout, key challenges and open questions that motivate the research presented in this thesis are highlighted.

2.1 Chemical analysis with Raman spectroscopy

Raman spectroscopy is introduced in this section, beginning with the fundamentals of light–matter interactions and Raman scattering (Section 2.1.1). This is followed by an overview of instrumentation and Raman modalities (Section 2.1.2), key application areas of RS (Section 2.1.3), the Raman data analysis pipeline (Section 2.1.4), and chemometric strategies for chemical analysis using RS (Section 2.1.5).

2.1.1 Light-matter interactions and the Raman effect

The interaction between light and matter has long been a central topic of human inquiry – from early philosophical debates about the nature of light, to the development of geometric optics in the 17th century, classical electromagnetism in the 19th, and quantum mechanics in the 20th. Throughout this progression, the study of light-matter interactions has remained fundamental to our understanding of the physical world, offering deep insight into the structure and dynamics of matter at atomic and molecular scales, and underpinning much of modern physics, chemistry and materials science.

Beyond advancing theory, the study of light-matter interactions has also led to important experimental breakthroughs. Since these interactions depend on the internal energy states

and structural features of materials, they can be exploited to extract detailed physical and chemical information. This principle underlies a wide range of spectroscopic techniques and imaging modalities that use light to probe the structural, electronic and vibrational properties of atoms and molecules.

Overview of light-matter interactions

Light can interact with matter in several distinct ways, each governed by different physical mechanisms and yielding distinct experimental observables[45,46]. These interactions are commonly grouped into the following categories:

- **Reflection.** Light is redirected at the boundary of a material due to differences in refractive index between media. Reflection is governed by macroscopic surface and dielectric properties, and underlies optical systems such as mirrors, reflective coatings and interferometers.
- **Transmission.** Light passes through a material, possibly changing direction (refraction) due to differences in refractive index between media. Transmission is characteristic of optically transparent or semi-transparent materials and is fundamental to techniques such as brightfield microscopy and ultraviolet-visible (UV-Vis) transmission spectroscopy.
- **Absorption.** Light transfers energy to a material, exciting electrons or vibrational modes to higher energy states. Absorption reflects the material's energy structure and forms the basis of methods such as UV-Vis, near-infrared (NIR) and infrared (IR) spectroscopy.
- **Emission.** Excited systems release energy by emitting photons as they return to lower energy states. Emission processes include fluorescence, phosphorescence and spontaneous or stimulated emission, and underpin techniques such as luminescence spectroscopy and fluorescence microscopy.
- **Scattering.** Light is deflected due to interactions with the material. Scattering can be:
  - **Elastic scattering** – the energy (frequency) of the scattered photon remains the same. Key examples include Rayleigh scattering (from particles much smaller than the wavelength) and Mie scattering (from particles comparable in size to the wavelength).
  - **Inelastic scattering** – the photon exchanges energy with internal excitations of the material, resulting in a frequency shift. Key examples include Raman scattering (interactions with vibrational modes, basis of Raman spectroscopy)

and Brillouin scattering (interactions with acoustic phonons, used in elasticity and viscoelastic imaging).

This thesis focuses on Raman spectroscopy as an analytical technique for chemical analysis based on inelastic Raman scattering.

The discovery of the Raman effect

The inelastic scattering of light by molecules was first predicted in 1923 by Austrian physicist Adolf Smekal, who suggested that scattered light could exhibit frequency shifts due to interactions with molecular vibrations[47]. The phenomenon was experimentally observed in 1928 by Indian physicists Sir Chandrasekhara Venkata Raman and his collaborator Sir Kariamanikkam Srinivasa Krishnan. By illuminating various liquids with sunlight and analysing the scattered light, they identified faint spectral lines at shifted frequencies – clear evidence of inelastic scattering[48]. Almost simultaneously, the phenomenon was independently observed by Soviet physicists Grigory Landsberg and Leonid Mandelstam, who reported similar frequency-shifted scattering in crystals[49]. Although their publication appeared slightly later than Raman and Krishnan's, their work is generally recognised as an independent discovery.

These parallel findings marked the discovery of what later became known as Raman scattering (or the Raman effect), and laid the foundation for Raman spectroscopy as an analytical technique that exploits inelastic light scattering to probe molecular vibrations and structure. In recognition of this discovery, Sir Raman was awarded the Nobel Prize in Physics in 1930.

Stokes and anti-Stokes Raman scattering

As previously noted, inelastic Raman scattering arises from the energy exchange between incident photons and vibrational modes of irradiated molecules. Through these interactions, photons may lose or gain energy, resulting in a shift in their frequency (since energy is proportional to frequency and inversely proportional to wavelength). Depending on the direction of energy transfer, two types of inelastic Raman scattering can occur – **Stokes Raman scattering** (photon loses energy → longer wavelength) and **anti-Stokes Raman scattering** (photon gains energy → shorter wavelength) (Figure 2.1A).

This process can be described in two complementary ways: classically, as the modulation of the induced dipole by vibrational motion of molecules (Appendix A); and quantum

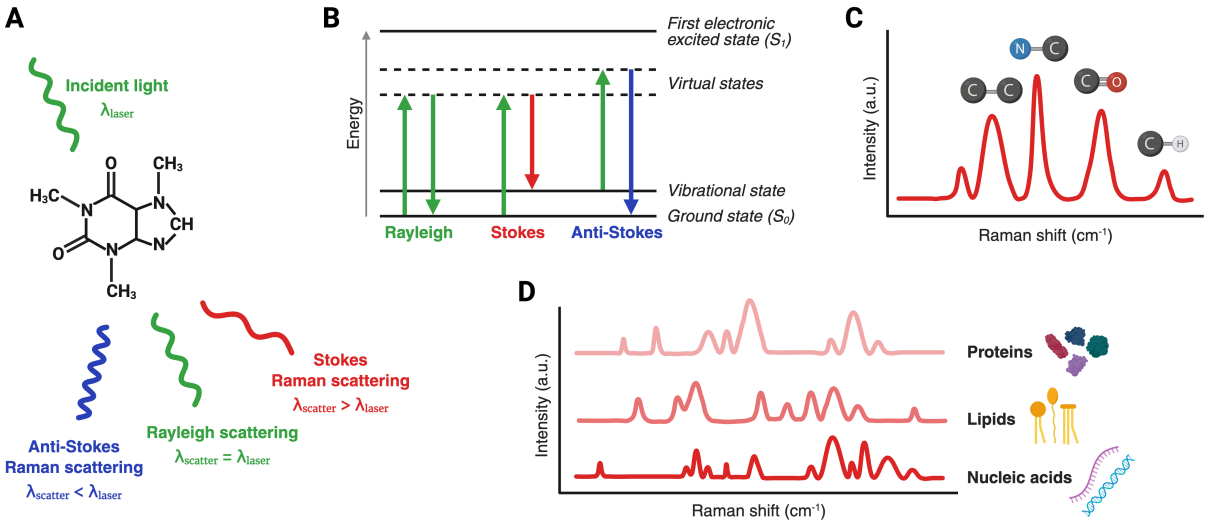


Figure 2.1: Molecular fingerprinting via Raman scattering – schematic overview. A, Illustration of light scattering from molecules, including elastic scattering (Rayleigh) and inelastic Raman scattering (Stokes and anti-Stokes). **B,** Energy level diagram depicting the ground state, virtual states and vibrational states involved in Raman scattering. **C,** Raman scattering is typically recorded as a one-dimensional spectrum showing the number of scattered photons as a function of Raman shift, representing a superposition of the vibrational modes of the sampled molecule(s). **D,** Raman spectra reveal the characteristic vibrational signatures of molecules, serving as molecular fingerprints. Displayed vibrational modes and Raman spectra are for illustrative purposes only. Created in BioRender.com.

mechanically, as a second-order light-matter interaction in which a photon briefly excites a system to a short-lived virtual state before relaxation to a vibrational level different from the initial one. In the discussion below, the quantum-mechanical description, which provides a more intuitive and illustrative framework, is adopted.

At thermal equilibrium, most molecules occupy their vibrational ground state, as described by the Boltzmann distribution[18]. Upon irradiation with incident light, such molecules can interact with photons and get promoted to an excited virtual state (Figure 2.1B). Because this state is unstable, the system rapidly relaxes to a more stable configuration. In most cases, molecules return to their original ground state without energy transfer, resulting in scattering of a photon with the same energy (same wavelength) as the incident light. This process is known as elastic scattering and is by far the dominant type of light scattering. In a much smaller fraction of interactions – only about one in 10^6–10^8 incident photons[50] – the molecule relaxes to an intermediate excited vibrational state rather than the ground state. This results in the emission of a photon with slightly less energy (longer wavelength) than the incident photon. This represents Stokes Raman scattering, a process orders of magnitude weaker than Rayleigh scattering.

Conversely, anti-Stokes Raman scattering occurs when a molecule, which is already in an excited vibrational state (e.g. due to thermal population), is further excited to a

higher-energy virtual state upon irradiation with incident light. Again, because this state is unstable, the molecule rapidly relaxes to a lower, more stable vibrational state. Typically, this is the vibrational ground state, and a photon with higher energy (shorter wavelength) than the incident light is emitted.

At room temperature, most molecules are at thermal equilibrium and occupy their vibrational ground state. Thus, Stokes scattering is far more probable, and therefore more intense, than anti-Stokes scattering under typical biological or ambient conditions. For this reason, spontaneous Raman spectroscopy typically relies on the detection of Stokes-shifted photons, offering a higher signal-to-noise ratio (SNR). Raman spectroscopy modalities are discussed in more detail in Section 2.1.2.

Molecular fingerprinting with Raman spectroscopy

The frequency shift of Raman-scattered photons provides a direct measure of the characteristic vibrational modes of molecules, encoding valuable information about their structure and bonding. As such, Raman scattering can be exploited as a tool to probe and characterise chemical composition. This, in essence, is the core idea of Raman spectroscopy.

In RS, samples are illuminated with monochromatic light – typically from a laser – and the scattered light is collected over time. While most scattered photons are elastically scattered, a small fraction emerge at slightly different energies due to Raman scattering. These scattered photons are analysed to determine their frequency shift relative to the laser source $\Delta\tilde{\nu}$, defined as

$$\Delta\tilde{\nu} = \tilde{\nu}_{\text{exc}} - \tilde{\nu}_{\text{scat}} = \frac{1}{\lambda_{\text{exc}}} - \frac{1}{\lambda_{\text{scat}}}. \tag{2.1}$$

This shift is commonly referred to as the **Raman shift** and is conventionally reported in wavenumbers (cm^{-1}). While the absolute wavelength of the scattered light depends on the excitation source, the Raman shift itself is independent of it. This makes Raman spectra directly comparable across different instruments and excitation wavelengths.

Raman scattering events are recorded as a one-dimensional spectrum of photon counts (proportional to intensity) versus their Raman shift (Figure 2.1C). By collecting Raman-scattered photons over an extended integration time, peaks corresponding to specific Raman-active vibrational modes gradually build up in the spectrum. More dominant vibrational modes produce stronger signals, while weaker modes appear as less intense peaks. A vibrational mode is Raman-active if it induces a change in the molecular polarisability, i.e. if the derivative of the polarisability α with respect to the vibrational

coordinate Q is non-zero ($\partial\alpha/\partial Q \neq 0$; see Appendix A). Physically, this means that vibration distorts the molecular electron cloud and alters the induced dipole moment. The combination of peak positions and intensities forms a spectral fingerprint of the sample, which can be interpreted to infer and characterise its molecular composition.

Advantages of Raman spectroscopy

As an analytical technique, Raman spectroscopy offers several key advantages. These are summarised below, with a focus on biological applications.

- **High chemical specificity.** Vibrational modes are highly specific to the chemical structure of molecules, enabling molecular fingerprinting and qualitative analysis with high specificity (Figure 2.1D).
- **Non-destructive and non-invasive.** RS typically uses low-power laser excitation and non-resonant interactions, making it inherently non-destructive. This is particularly beneficial for sensitive samples such as cells and tissues, where RS enables chemical analysis without compromising structural or functional integrity. This also facilitates *in situ* and real-time capability.
- **Label-free with minimal sample preparation.** Unlike techniques such as immunofluorescence microscopy, RS is inherently label-free, requiring no dyes, stains or external markers to detect molecular signatures. Furthermore, unlike methods such as IR spectroscopy, mass spectrometry, X-ray diffraction or electron microscopy, RS generally requires little to no sample preparation, enabling direct analysis of solids, liquids, and gases without extensive processing.
- **(Semi-)quantitative.** Under non-resonant, far-from-absorption conditions, the intensity of spontaneous Stokes Raman scattering scales with the relative concentration of analytes, allowing (semi-)quantitative chemical analysis. Nevertheless, intensity also depends on a number of experimental factors, including excitation wavelength, optical geometry, collection efficiency, orientation/polarisation effects and matrix/environmental influences. This is why careful calibration, preprocessing and multivariate analysis for reliable quantitative comparisons (see Section 2.1.4).
- **Minimal interference from water.** RS is largely unaffected by water, making it highly suitable for aqueous solutions and biological samples with high water content. In contrast, techniques like infrared spectroscopy are often hindered by strong water absorption.
- **Microscale spatial resolution and depth profiling.** Coupled with confocal microscopy, RS provides submicron spatial resolution and enables depth profiling,

allowing 2D and 3D chemical imaging in complex (biological) specimens. This is discussed in more detail in the next section.

Together, these features establish RS as a powerful analytical tool for non-destructive, label-free chemical analysis of molecular compounds, biological specimens and materials, providing in-depth insight into their structure and function[18–21,23,32,33,42,51–53].

2.1.2 Raman spectroscopy techniques

Although RS is conceptually simple, its practical implementation presents significant technical challenges. As discussed earlier, Raman scattering is an intrinsically weak process, with only a tiny fraction of incident photons undergoing inelastic scattering. Detecting these rare events requires highly sensitive instrumentation and carefully optimised optical systems. As a result, it took decades of technological progress in laser sources, optics, detectors and data processing before RS became a practical tool for chemical analysis. In particular, nearly 50 years passed before the technique could be reliably applied to complex biological samples such as cells and tissues[42].

Today, modern RS systems integrate a range of advanced optical and electronic components to enable precise measurement of Raman signals. Depending on the application, these systems can be configured in various forms, from compact benchtop spectrometers to automated Raman microscopes capable of spatially resolved chemical imaging. In addition to standard spontaneous RS, several specialised techniques have also emerged to enhance sensitivity, selectivity or acquisition speed. These include resonance Raman spectroscopy (RRS), coherent anti-Stokes Raman scattering (CARS) and surface-enhanced Raman scattering (SERS), among others. A brief overview of these RS techniques is provided in the following sections.

Spontaneous Raman spectroscopy

Spontaneous Raman spectroscopy is the most common and widely used implementation of RS. It refers to the standard approach where inelastic scattering occurs naturally, without any external amplification or nonlinear optical effects. A typical spontaneous RS system consists of several core components that work together to excite the sample and detect the weak Raman signal. This includes:

- **Laser source.** A stable, monochromatic light source, usually a diode or solid-state laser, excites the sample. The excitation wavelength is carefully chosen to maximise Raman scattering efficiency while minimising sample fluorescence.

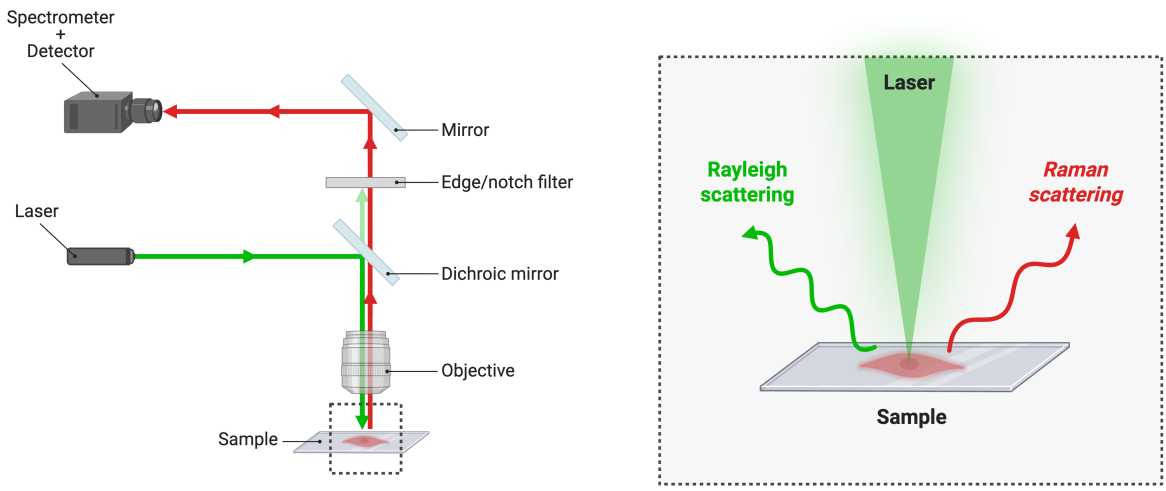


Figure 2.2: Schematic of a spontaneous Raman spectroscopy system. The diagram illustrates the primary hardware components involved in a spontaneous RS system, including the laser source, optical elements, objective and spectrometer. Arrows indicate the path of excitation light and the collection of inelastically scattered Raman photons from the sample. Created in BioRender.com.

- **Optical components.** Lenses, mirrors and beam splitters direct and focus the laser onto the sample with high spatial precision. High numerical aperture objectives are often used to achieve strong focusing and efficient signal collection.
- **Sample stage.** The sample is positioned on a stage, typically under a microscope, for point measurements or scanning. Motorised, computer-controlled stages enable automated Raman mapping and 2D and 3D imaging.
- **Edge or notch filters.** These filters reject the intense elastically scattered light while transmitting the weaker Raman-shifted light.
- **Spectrometer.** The filtered Raman signal is dispersed by a diffraction grating within the spectrometer, separating the scattered light by wavelength. The spectral resolution of the grating determines the ability to distinguish closely spaced vibrational modes.
- **Detector.** Sensitive detectors, such as charge-coupled devices (CCDs), record the dispersed signal to produce the Raman spectrum. Deep-cooled detectors are typically used to minimise noise and allow detection of extremely weak Raman signals.

Figure 2.2 provides a graphical representation of a typical spontaneous RS system, highlighting the key hardware components discussed above. The modular architecture of RS systems allows for flexible configurations that can be customised to accommodate specific sample types and experimental requirements.

In addition to the hardware components, RS systems often include dedicated control software that ensures system integration and manages key settings such as laser power, acquisition parameters and stage movement (e.g. for Raman imaging; see next section).

Raman imaging

Beyond single-point measurements, Raman spectroscopy can also be extended to capture spatially resolved chemical information in the form of hyperspectral images – a modality known as Raman imaging (or Raman mapping). In this approach, Raman signals are acquired at multiple locations across a sample, generating a hyperspectral data cube in which each pixel represents the measured Raman response at the corresponding spatial position. The result is a chemical map that encodes the spatial distribution of different molecular species present in the sample. This is particularly valuable in biological and materials science applications, where spatial heterogeneity is often critical to understanding structure-function relationships.

Raman imaging can be performed in several different ways, including:

- **Point-scanning.** A tightly focused laser is used to scan the sample point by point, acquiring a full Raman spectrum at each location. This is the most common acquisition method and offers high spectral and spatial resolution, albeit at the cost of longer acquisition times.
- **Line-scanning.** A line-shaped laser beam is used to illuminate a strip of the sample at a time, simultaneously collecting spectra along the whole line. The sample is then scanned perpendicularly to build a two-dimensional image, offering a good balance between speed and resolution.
- **Wide-field imaging.** The entire sample area is illuminated at once, and Raman signals are collected in parallel using spectrally selective detection methods such as tunable filters. Although less common in spontaneous RS due to low signal strength, this approach can be effective in more advanced Raman modalities.

For more information about these three imaging techniques, the reader is referred to the review by Schlücker et al.[54].

Raman imaging can also be extended into three dimensions to generate volumetric Raman images – a technique known as volumetric Raman imaging. This is typically accomplished by acquiring a series of 2D Raman images at different focal planes along the depth (z-axis) of a sample. Individual imaging slices are then combined to construct a 3D chemical map of the scanned area.

To achieve depth resolution, a confocal Raman system is often required. In confocal Raman microscopy, a pinhole is integrated to block out-of-focus light and enhance axial resolution and enable optical sectioning. Confocal pinholes are most commonly used in point-scanning systems for high-resolution or volumetric imaging of complex, heterogeneous samples.

Advanced Raman modalities

Spontaneous RS provides a powerful and versatile analytical platform for chemical analysis with high molecular specificity. However, it has multiple practical limitations. These include its intrinsically weak signal intensity, slow acquisition speeds and potential interference from background fluorescence.

To overcome these issues and improve the performance of RS, a range of advanced RS techniques has been developed, including:

- **Resonance Raman Spectroscopy (RRS).** In RRS, the excitation wavelength is tuned to match an electronic absorption band of the molecule of interest[55–57]. This resonance condition selectively enhances specific vibrational modes, resulting in signal amplification of up to 10^3–10^6 times compared to spontaneous Raman scattering. RRS is particularly useful for targeting specific chromophores, pigments and biomolecules with strong electronic transitions.
- **Surface-Enhanced Raman Scattering (SERS).** SERS techniques exploit the plasmonic properties of nanostructured metallic surfaces (e.g. silver or gold), which can amplify the electromagnetic field near the surface through localised surface plasmon resonances, and thus enhance the Raman signal by several orders of magnitude[58,59]. As a result, SERS enables highly sensitive detection and is widely used in trace analysis and biosensing applications[60–63].
- **Coherent Raman Scattering (CRS).** CRS methods use multiple pulsed laser beams to generate a coherent Raman signal that is several orders of magnitude stronger than spontaneous Raman scattering[64–67]. This results in significantly faster acquisition times, enabling video-rate imaging[68]. Common CRS methods include Coherent Anti-Stokes Raman Scattering (CARS) and Stimulated Raman Scattering (SRS).

While advanced techniques can offer significant improvements in sensitivity, specificity and acquisition speed, these methods also come with limitations. This includes increased technical complexity, higher costs and more limited accessibility. As a result, despite substantial progress, many of these techniques are still not widely available in routine

laboratory settings. Indeed, spontaneous RS remains the most widely used and accessible modality, especially for general-purpose chemical and biological analysis. For a more detailed discussion of advanced modalities and their use in biological research, including other variants such as tip-enhanced Raman (TERS), spatially offset Raman spectroscopy (SORS) and transmission Raman spectroscopy (TRS), the reader is referred to the works by Butler et al.[21], Cialla-May et al.[69], Auner et al.[70].

In this work, spontaneous Raman spectroscopy was used throughout. Thus, unless otherwise specified, the term "Raman spectroscopy" refers to spontaneous Raman spectroscopy.

2.1.3 Applications of Raman spectroscopy

The versatility of RS, along with its non-destructive and label-free nature, has positioned it as an important analytical technique across a wide range of applications in both research and industry. While the primary focus of this thesis is on cell biology, many of the analytical tools and approaches presented here can be readily applied across other disciplines as well. For this reason, an extended overview of key application domains is provided below.

Cell biology

A primary application of RS is in cell biology, where it provides a powerful tool for biochemical analysis and imaging of cells and tissues[36]. A particular advantage of RS in this context is its ability to investigate cellular composition, dynamics and function in native environments, without the need for external labels or destructive preparation.

At the cellular level, Raman imaging has been employed to map the spatial distribution of key biomolecules such as nucleic acids, proteins and lipids, thereby enabling the visualisation of major organelles and cell morphology in both fixed and live cells[71–73]. Volumetric Raman imaging has further enabled three-dimensional reconstructions of entire cells, facilitating both qualitative and quantitative analyses across cell types[74]. Furthermore, Raman imaging has been valuable for assessing cell responses to external stimuli, including drugs[75] and drug delivery vehicles[76,77], toxins[78] and environmental stressors[79]. To improve specificity, synthetic Raman tags have been developed for chemical labelling and multiplexed imaging of defined molecular targets, including key biomolecules, drugs and metabolites[22]. In addition to imaging analyses, RS has been an effective method for single-cell analysis, with downstream applications ranging from cell typing and pheno-

typing to drug screening, cell cycle progression and differentiation[80,81].

Beyond individual cells, RS has been extensively applied to tissue samples, providing detailed, label-free histological and pathological analysis[82]. In biological research, Raman imaging has been used to study tissue morphogenesis[83], elucidate disease aetiology[84], and monitor the distribution and pharmacological response of therapeutic compounds within tissues[85,86]. To enable rapid imaging across larger tissue areas, coherent Raman techniques such as CARS and SRS have also been employed[87–89].

With recent advances in stem cell biology, there is growing interest in extending Raman imaging to three-dimensional organoid models to monitor development, biochemical organisation and drug response in physiologically relevant systems. Nonetheless, despite some initial demonstrations[90], the application and validation of Raman imaging in this context remain in their early stages. This is discussed in more detail in Section 2.2.5.

Medicine and clinical research

RS has also attracted considerable interest in medicine and clinical research.

A main application area of RS in this domain has been diagnostics, with the promise of enabling minimally invasive testing, rapid and portable readouts, and more sensitive early detection[25]. In particular, RS has been explored for the detection of biomarkers across a range of diseases, including cancer[91,92], metabolic disorders[93] and infectious diseases[94]. To this end, RS has been applied directly to liquid biopsies and body fluids[95,96], as well as integrated into paper-based assays[97], microfluidic chips[98] and sensor platforms[99]. In microbiology, RS has further emerged as a powerful tool for the rapid identification of bacteria and other pathogens[100,101].

In addition to this, RS-based techniques are increasingly explored in surgical and intraoperative settings[27,102,103]. Handheld probes, fibre-optic endoscopes and real-time Raman imaging systems have been developed to delineate tumour margins[70,104], enhance biochemical contrast[105] and guide resections during procedures. Moreover, Raman imaging approaches for label-free histopathology have been developed for applications in tissue classification, molecular subtyping and tumour analysis[82,106,107].

Chemical and pharmaceutical sciences

Owing to its molecular sensitivity and specificity, RS has become a powerful tool across the chemical and pharmaceutical sciences, with applications ranging from molecular identification to real-time process monitoring.

In chemical research, RS is routinely used to characterise organic and inorganic compounds in solid, liquid and gaseous forms. Common applications include identifying reaction intermediates and products[108], detecting impurities and adulterants[109], and analysing chemical mixtures[110,111]. Its non-invasive, non-contact nature makes RS particularly valuable for *in situ* monitoring of chemical reactions[112,113], enabling kinetic and mechanistic studies across a range of domains, including catalysis[114], polymerisation[115] and corrosion[116]. In addition, its sensitivity to functional groups, bond types and molecular symmetry has further positioned RS as an effective approach for discriminating closely related species such as regioisomers, stereoisomers, conformers and polymorphs[117–119].

These capabilities also extend to pharmaceutical sciences, where RS has been integrated at multiple stages of the drug development pipeline[29]. In early phases, it has been applied to product formulation and structural analysis[120], pharmacokinetic studies of active pharmaceutical ingredients[121], degradation and stability testing[122], and solid-state screening[123]. RS has also been widely used as a process analytical technology during manufacturing, facilitating monitoring and quality control[124,125]. More recently, Raman-based probes have been explored in anti-counterfeiting for the authentication of medicines and packaging[126].

Materials science and engineering

RS has also established itself as a key characterisation technique in materials science and engineering, offering insights into structural, chemical and mechanical properties of a broad range of systems.

In crystalline solids, RS has been widely used to evaluate phase composition, crystallinity and internal stresses, and analyse polymorphs and lattice dynamics. Furthermore, it has been a valuable tool for assessing defects, layer thickness and doping levels in carbon-based materials, such as graphene and carbon nanotubes[127]. Beyond this, RS has been applied in the study of semiconductors, ceramics, polymers and thin films. Advanced SERS-based techniques have further enabled ultra-sensitive characterisation of nanoscale features, plasmonic interactions and catalytic interfaces, with applications in energy storage, surface chemistry and heterogeneous catalysis[62].

Beyond inorganic systems, RS is also increasingly applied to biomaterials and engineered scaffolds to study mineralisation processes, protein-matrix interactions and synthetic-biological interfaces, and guide the rational design of scaffolds, implants and composite biomaterials with tailored structural and biochemical properties[28].

Environmental sciences

Furthermore, RS has become an important tool across environmental sciences. An area of high topical interest has been the detection and assessment of pollutants in food and water, with RS used for trace analysis, contamination screening and quality control[30,128,129]. Common targets include heavy metals[130], organic residues[131], microplastics[132], pesticides[133], and other harmful compounds. RS has also been applied in agriculture, from detecting pesticides and other contaminants[134,135], to assessing crop health[136] and plant stress[137]. Similarly, chemical analysis with RS has been extended to atmospheric research to analyse aerosol composition, water vapour and trace gases[138,139]. Beyond these applications, Raman-based techniques have been used to characterise mineral phases in soils and sediments, with applications spanning mineralogy[140], pollution[141], archaeology[142], and even planetary research[143].

Forensics and security

Lastly, RS has found broad applications in forensics and security. Owing to its chemical sensitivity, RS has emerged as a valuable tool for trace detection of illicit drugs, explosives and chemical warfare agents[31,144]. To this end, handheld Raman spectrometers have been deployed at security checkpoints for counter-terrorism and homeland security applications, providing rapid on-site analysis[145]. In forensic investigations, RS has been used for trace analysis and source attribution of material evidence, including inks, dyes, fibres, paints and glass fragments[146].

Growing concerns around counterfeiting have also led to an increased interest in Raman-based authentication technologies[147]. Recent advances include the development of anti-counterfeiting probes based on Raman-active physical unclonable functions, in the form of barcodes[148], inks[149] and nanostructured materials[150,151]. These approaches have been explored for safeguarding pharmaceuticals[152], banknotes[153] and other high-value products and consumer goods.

2.1.4 Data processing and interpretation in Raman spectroscopy

Raman spectroscopy generates rich, high-dimensional datasets that capture detailed molecular information about samples. This wealth of information, however, brings its own challenges. Extracting meaningful insights from experimental RS data often requires extensive method development and optimisation across the entire data lifecycle, from experimental design and data acquisition to spectral preprocessing and downstream data analysis[39,41,154].

This section outlines the key stages of the RS data analysis pipeline and their significance, particularly in the context of biological applications.

Data collection

Preparation for RS analysis begins at the point of data acquisition, well before any preprocessing or data interpretation. A range of instrumental parameters and experimental settings can influence the quality of the collected data and therefore warrant careful consideration. These include excitation wavelength, laser power, integration time, and number of accumulations, among others. Each of these factors can significantly affect signal strength, signal-to-noise ratio, the presence of artefacts, and the overall reproducibility of measurements – directly impacting subsequent preprocessing, analysis and interpretation. As such, acquisition conditions must be carefully controlled and optimised to minimise the effect of artefacts, noise and other sources of variability, guided by the specific requirements of the intended application. The key acquisition parameters and the rationale behind their selection are briefly discussed below.

Excitation wavelength. One of the first parameters to consider is the wavelength of the excitation laser. This choice can influence a range of optical factors, including signal strength, fluorescence background, sample penetration depth and spatial resolution. Shorter wavelengths, such as 532 nm, generally produce stronger Raman signals but are more likely to induce unwanted fluorescence. In contrast, longer wavelengths, such as 785 nm, tend to suppress fluorescence but typically result in weaker Raman signals. Therefore, selecting the optimal excitation wavelength requires balancing signal intensity against background interference, and should be informed by the specific properties of the sample and the requirements of the application.

Laser power. Another key consideration is the laser power, which can be adjusted using neutral density filters. Generally, increasing the laser power improves signal strength and the signal-to-noise ratio. However, excessive power can also cause local heating and damage the sample. This trade-off is critical when working with sensitive samples, such as cells, tissues and other biological specimens. As a rule of thumb, laser power is typically set as high as possible without causing any observable photodamage during acquisition.

Integration time and accumulations. Two other important and closely related parameters that directly influence data quality are the integration time and the number of accumulations per point. Integration time determines the total photon count collected; therefore, longer integration times inherently improve SNR. Similarly, acquiring multiple accumulations at each point and averaging them helps reduce random noise and mitigate artefacts. Nonetheless, increasing both integration time and the number of accumulations

extends the overall acquisition time and can intensify photobleaching or thermal effects, particularly during imaging or volumetric scans, which may span hours or even days (see next paragraph). Optimal settings are typically determined empirically, balancing data quality against practical constraints such as acquisition speed and sample stability.

Spatial resolution and sampling step size. When it comes to Raman imaging, image quality is governed by two main factors: spatial resolution and sampling step size. The spatial (lateral and axial) resolution is primarily determined by the optical setup, particularly the objective lens, laser wavelength and confocal aperture settings. These parameters define the smallest volume from which a Raman signal can be effectively collected. In contrast, the step size refers to the distance between adjacent sampling points (pixels) in the image and is user-defined.

As with other acquisition parameters, there is an inherent trade-off between image quality and total acquisition time. Smaller step sizes and higher resolution improve spatial discrimination but significantly increase the number of sampling points and therefore the overall acquisition time. For example, consider an imaging setup with two accumulations and 1 second integration time per point – i.e. approximately 2 seconds per pixel, excluding overhead such as motor movement, autofocus routines and data processing. Under these settings, a scan covering a 100×100 µm area at a 1 µm step size (10 000 pixels) would require over 5.5 hours. Extending this to a volumetric scan with 10 z-slices at 1 µm spacing increases the total to 100 000 pixels, or more than 55 hours.

This example highlights how acquisition times in conventional spontaneous Raman spectroscopy, particularly for high-resolution or large-volume scans, can quickly become impractical and raise concerns about sample degradation. As a result, although high-resolution imaging is almost always desirable, reduced spatial resolution and larger step sizes are often necessary to balance detail with time and sample stability constraints. Faster acquisition can be achieved using advanced techniques such as SRS and CARS, as discussed in Section 2.1.2.

Sample preparation. The final aspect discussed here is sample preparation, which plays a crucial role in ensuring reliable and reproducible measurements. Depending on the application, several factors must be considered, including sample type, physical dimensions, and the choice of mounting medium, among others. In Raman imaging specifically, a key consideration is maintaining sample stability throughout potentially long acquisitions, as any movement or degradation can introduce artefacts and reduce signal quality. This is especially important for biological samples, which may undergo structural changes over time due to dehydration, photodamage or environmental fluctuations. To mitigate these effects, biological samples are typically chemically fixed before imaging to preserve

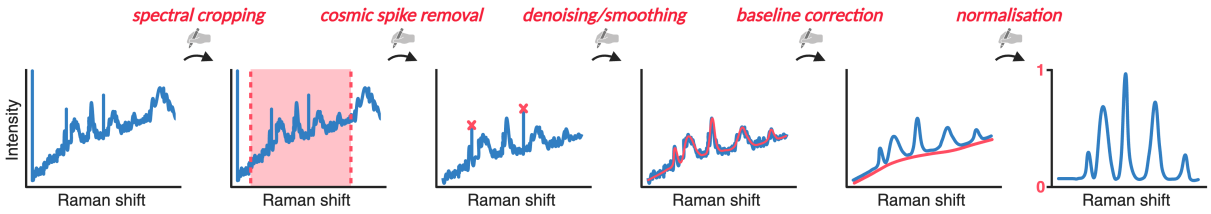


Figure 2.3: Overview of common spectral preprocessing steps in Raman spectroscopy. The diagram outlines a typical sequence of preprocessing procedures applied to raw Raman spectra, including spectral cropping, cosmic spike removal, denoising, baseline correction and normalisation. These steps enhance spectral quality and improve the reliability of downstream analysis. Created in BioRender.com.

their structure and mounted using appropriate supports such as holders, coverslips or hydrogels. Although live-cell or tissue Raman imaging is technically feasible, it is rarely performed routinely due to the need for precise environmental control (e.g. temperature, CO_2, humidity), as well as the requirement for specialised setups and/or advanced Raman modalities to enable faster acquisition.

Spectral preprocessing

But even under ideal acquisition conditions, experimental RS measurements often still contain significant artefacts and are not directly usable for downstream analysis. The high sensitivity of RS makes it susceptible to various sources of non-specific signal and noise, such as baseline drifts, fluorescence backgrounds, cosmic spikes, instrumental noise and other variations[155,156]. These artefacts, abundant in most biological settings, can severely obscure relevant spectral features and compromise subsequent analysis and interpretation. Consequently, spectral preprocessing is a critical component of any RS analysis workflow, aimed at correcting artefacts, enhancing signal quality and ensuring comparability across samples. This typically involves a structured pipeline involving a series of preprocessing steps applied in sequence to clean and standardise the data (see Figure 2.3).

An overview of the main preprocessing steps and methods in RS is provided below. It is important to note that there is no universally applicable preprocessing strategy; the optimal sequence depends on the specific application and varies considerably across the literature. This section follows the order used in the analyses presented in this thesis, which aligns with established best practices and published guidelines[21,39,42,154,157,158].

Spectral cropping. The first step in most preprocessing pipelines is to crop the spectra to exclude regions that are irrelevant for subsequent analysis. This can not only improve the reliability of consecutive preprocessing steps, such as baseline correction, denoising and normalisation, but also reduce dimensionality for downstream modelling.

For instance, the region around $0\,\mathrm{cm}^{-1}$ (i.e. the Rayleigh line) and the adjacent detector-suppressed zone are typically removed early, as they contain no useful Raman information and can interfere with preprocessing. In many biological applications, the so-called silent region (∼1800–2800 cm^{-1}), which lacks significant Raman features in biological samples and is often dominated by noise, is also excluded. Biological analyses may further be constrained to the fingerprint region (∼600–1800 cm^{-1}), which contains rich molecular information. Depending on the sample type, the high-wavenumber CH/OH-stretching region (∼2800–3300 cm^{-1}) may also be retained.

Spectral interpolation. When spectra are collected using different spectral axes or resolutions, spectral interpolation may be necessary to align them onto a common wavenumber axis. This step is essential for integrating data across samples or instruments, and must be done carefully to avoid introducing artefacts, especially around sharp peaks or in noisy regions. In practice, this is typically achieved using linear, spline or cubic interpolation methods.

Cosmic spike removal. Another common preprocessing step in RS is the removal of cosmic spikes caused by high-energy particles (e.g. cosmic rays) hitting the detector during data acquisition. These artefacts appear as sharp, non-reproducible intensity spikes that are unrelated to the underlying biochemistry and can affect both visual inspection and model performance. The likelihood of cosmic spikes naturally increases with longer integration times. Removing them early in the preprocessing pipeline helps prevent interference with other steps and reduces the risk of mistaking them for true spectral features during analysis. Most despiking methods rely on the distinctive shape and intensity of these spikes and detect them using outlier-based criteria such as z-scores[159] or wavelet transform methods[160]. Once identified, the affected points are typically removed and replaced through interpolation based on nearby values.

Wavenumber calibration. Instrumental and environmental variability can compromise the reproducibility of the wavenumber axis in recorded Raman spectra. Beyond inter-instrument differences, factors such as instrumental drift, optical misalignment and environmental fluctuations (e.g. changes in temperature or humidity) may introduce small but meaningful shifts. To correct for these and ensure that peak positions reflect true chemical shifts and that spectra collected at different times or on different instruments remain comparable, wavenumber calibration is often performed using standard reference materials with well-known, sharp spectral features, such as silicon or polystyrene. This can be done before data acquisition (e.g. via the instrument's built-in control software) or retrospectively during preprocessing.

Intensity calibration. Additionally, variations in detector sensitivity, laser power or

wavelength-dependent instrument response can affect the relative intensity values in recorded Raman spectra. To ensure that intensities remain comparable across measurements, samples and instruments – especially in quantitative analyses – intensity calibration may be applied using a known reference material or a calibrated broadband light source. This typically involves generating an instrument response function and adjusting measured spectra accordingly. Whether full intensity calibration is required depends on the experimental design, the characteristics of the data collected and the degree of technical variability. In many applications, simpler normalisation approaches are used to compensate for intensity variations due to laser power, integration time or exposure settings.

Denoising. Another major consideration in RS is random noise. RS data can be affected by various sources of noise, including electronic detector noise, photon shot noise and environmental interference. These noise contributions can be further exacerbated under low signal conditions – for instance, when acquisition times are short or analyte concentrations are low. Denoising (or smoothing) is a common preprocessing step, aimed at improving SNR and enhancing spectral quality for downstream analysis. Popular methods include Savitzky-Golay filtering[161], least-squares-based smoothing[162], Gaussian kernel smoothing and wavelet-based denoising. Note that denoising must be carefully optimised to avoid distorting or removing altogether true spectral features.

Baseline correction. In addition to random noise, RS spectra can also contain broad baseline distortions, most commonly from background signals such as fluorescence or stray light. These signals are generally unrelated to the Raman-active vibrational modes of interest but can obscure true peaks and complicate downstream analysis, particularly in quantitative or comparative studies. To address this, baseline correction is typically applied during preprocessing. A variety of algorithms have been developed for this purpose, based on polynomial fitting[163,164], least squares[165–167] and other curve-fitting approaches[168,169]. The choice of method and parameters should be carefully tailored to the specific data and application, as both overcorrection and undercorrection can introduce artefactual distortions.

Background removal. Beyond fluorescence-related baselines, Raman spectra may also contain background signals from external sources such as substrates, mounting media or other experimental components. To improve specificity and isolate the true Raman signal, background removal may be applied by subtracting a reference spectrum collected from the source of background (e.g. a blank substrate or buffer-only control). Care must be taken to ensure that the reference spectrum is representative and properly aligned to avoid introducing additional artefacts.

Normalisation. As previously mentioned, normalisation or other forms of scaling can be applied to reduce non-specific variability across instruments, samples or acquisition settings. This step may also be useful when preparing data for statistical modelling or machine learning, where the relative scaling of input features can affect performance. Common normalisation methods include vector normalisation (L2 norm), min-max scaling, standard normal variate (SNV), and area under the curve (AUC) normalisation, among others. In situations where relative signal intensities carry important biological or chemical meaning, such as concentration differences across samples, global scaling may be preferred. This approach allows spectra to be normalised while preserving overall intensity trends across the dataset.

Outlier removal. Even after multiple preprocessing steps, some spectra may still contain significant artefacts, noise or other anomalies. These low-quality spectra can distort statistical models or introduce bias, particularly in data-driven analyses. Therefore, outlier removal may be performed to enhance the robustness and reliability of downstream analysis. Although manual inspection and filtering may be feasible for certain applications involving smaller datasets, automated outlier detection methods are often required. Using these, spectra that deviate substantially from the rest of the dataset are flagged based on quality control metrics, statistical thresholds or multivariate techniques. Common approaches include Hotelling's T^2 test, Mahalanobis distance[170] or custom thresholding. Regardless of the method, care must be taken to ensure that meaningful variation is not inadvertently discarded.

Data analysis

While carefully optimised preprocessing can significantly improve signal quality and data standardisation, the interpretation of the resulting data remains a significant challenge. The high dimensionality of Raman spectra, combined with the complexity and volume of data, makes manual inspection impractical and often unreliable.

To this end, a wide range of statistical and computational tools is employed to uncover underlying patterns and enable interpretation. Depending on the specific goals of the study, these may include techniques for dimensionality reduction, feature extraction, clustering, classification and regression. An overview of common analysis techniques in RS is provided below.

Univariate analysis. Univariate analysis is a commonly used approach in RS, typically involving the comparison of intensity values at specific peaks or wavenumbers. While limited in scope, it can be a valuable tool for exploratory or visualisation purposes, helping to highlight potential biomarkers and investigate trends or differences between

study groups. To assess the statistical significance of observed differences, tests such as t-tests or ANOVA are often applied[171].

Dimensionality reduction. Given the high dimensionality of RS spectra, dimensionality reduction is a common step in RS analysis, aimed at reducing the number of input variables while preserving as much relevant information as possible. This can be useful for exploratory analysis, noise reduction, feature extraction, visualisation and downstream modelling. The most widely used method is principal component analysis (PCA)[172–174], which derives the set of orthogonal components that capture the maximal variance of the dataset. Non-linear approaches such as t-distributed stochastic neighbour embedding (t-SNE)[175] and uniform manifold approximation and projection (UMAP)[176] are also frequently used to visualise local structures or potential clusters in the data.

Clustering. When uncovering underlying structure or subpopulations in the data is the primary goal, more specialised clustering techniques can be applied. Common clustering algorithms in RS include k-means[177,178], hierarchical clustering[179] and density-based approaches such as DBSCAN[180], designed to group spectra in an unsupervised manner based on their similarity. This can be particularly useful for exploratory analysis, visualisation and segmentation tasks. Note that the choice of algorithm, distance metric and parameters (e.g. number of clusters) can significantly influence clustering results and should thus be selected with the specific data structure and study objectives in mind.

Predictive modelling. Supervised learning methods are also commonly used in RS to build predictive models that can map spectral features to known outcomes. Depending on the task, these outcomes may be categorical (classification) – such as disease status or sample type – or continuous (regression), such as analyte concentration or phenotypic properties. Common classification algorithms include logistic regression[181], linear discriminant analysis (LDA)[182], support vector machines (SVM)[183], and random forests[184], among others. On the other hand, popular regression methods include linear regression and partial least squares regression (PLSR)[185,186]. In recent years, neural networks and deep learning models have gained significant traction for both classification and regression, demonstrating substantial improvements in accuracy and robustness, particularly for more complex tasks or when working with large datasets.

Decomposition and spectral unmixing. In many RS applications, especially when dealing with complex, heterogeneous samples, recorded spectra represent mixtures of signals from different chemical or biological sources. Decomposition methods such as non-negative matrix factorisation (NMF)[187] can be used to identify the underlying components and estimate their relative contributions. Unlike PCA, NMF imposes non-negativity constraints, enabling a more interpretable, part-based decomposition that

aligns well with the inherently non-negative nature of RS data. Depending on the context, more specialised unmixing algorithms may also be applied. Spectral unmixing is explored in more detail in Section 2.1.5.

Software tools in Raman spectroscopy

Despite its importance and widespread use across the life and physical sciences, RS continues to face significant practical and methodological challenges, which hinder its broader deployment and the advancement of RS chemometrics.

Central to this problem is the restrictive, functionally disparate and highly encapsulated nature of current software for RS data analysis. Much of RS data processing and analysis still relies on proprietary, closed-source software environments, which vary significantly in data formats, functionality and design[44]. As a result, data workflows remain fragmented and tightly coupled to specific software, leading to methodological inconsistencies, duplicated effort and reduced cross-platform and integrative efforts, with growing concerns around reproducibility[38,40,188,189]. The constraints of proprietary software have also hindered the integration of emerging AI technologies into RS analysis, limiting the implementation of specialised procedures beyond what is natively supported. Consequently, researchers increasingly resort to developing in-house scripts, resulting in workflows dispersed across software tools, custom in-house scripts and ad hoc pipelines, further adding to methodological fragmentation and irreproducibility[39].

As new hardware, software and data acquisition RS technologies continue to emerge[190,191], there is a pressing need for an integrated RS data analysis environment, which facilitates the development of pipelines, methods and applications, and bolsters the use of RS in industry and research.

2.1.5 Chemical analysis via Raman spectroscopy and chemometrics

As a characterisation technique, the primary objective of RS analysis is to extract meaningful chemical information from collected spectral data. This is particularly important in applications beyond direct predictive modelling – e.g. biomarker discovery, material characterisation, disease diagnosis – where explicit chemical insight is essential for scientific interpretation, validation and downstream decision-making. In this context, the information of interest may be: (i) **qualitative** – i.e. what analytes are present in a

sample, and/or (ii) **quantitative** – i.e. how much of each analyte is present.

To this end, a wide range of chemometric methods have been applied to RS data. This section reviews key chemometric approaches in RS – from classical techniques to more recent machine learning-based methods, along with their limitations, and introduces hyperspectral unmixing as a framework for unsupervised chemometric analysis.

Standard methods for Raman chemometrics

As discussed in Section 2.1.4, RS analysis encompasses a range of mathematical, statistical and computational techniques. Many of these methods are applied for chemometric purposes to extract qualitative and quantitative chemical information from Raman spectra. Common approaches can be broadly grouped as univariate or multivariate, depending on how they handle spectral features.

Univariate methods focus on analysing individual peaks or spectral bands. They are typically used in simpler, well-characterised systems, where specific spectral features can be reliably linked to known molecular vibrations or target compounds. Qualitative analysis often involves detecting the presence or absence of characteristic peaks using thresholding or peak detection algorithms. For quantitative assessments, peak height or integrated area is usually measured and related to concentration through pre-established calibration curves. In some cases, analysis can be extended to multiple peaks by comparing ratios between peak intensities or areas.

The main advantages of univariate methods are their simplicity and interpretability. However, they become less effective in more complex systems, such as biological samples, where peaks from multiple sources may overlap and vary significantly due to noise, baselines and other artefacts. As noted earlier, instrumental and experimental variation can also cause peak shifts, further complicating univariate analysis. Lastly, these methods typically require prior knowledge of target peaks and strong domain expertise, which limits their applicability in exploratory, target-agnostic or highly heterogeneous settings.

Multivariate methods, by contrast, consider multiple spectral features simultaneously, often leveraging the full spectrum. This allows them to capture global patterns and correlations within the data, making them generally more effective for complex samples where peaks overlap or interact.

When reference spectra for target compounds are available, either from separately collected measurements or spectral libraries, a common approach is to model measured spectra as linear combinations of these references using methods such as non-negative

least squares (NNLS)[192]. This enables estimation of the relative contribution of each target component. However, this approach has two limitations. First, it assumes that the relevant target components are known in advance. Second, the fitting process can be highly sensitive to noise, baseline variation and other artefacts. This can be particularly problematic in highly heterogeneous biological systems such as cells and tissues, where it is often infeasible to obtain a comprehensive set of reference spectra that fully captures the diversity of biochemical components. Moreover, spectral signatures observed in complex biological matrices can differ substantially from those measured in pure solutions, further complicating interpretation.

In cases where quantitative labels are available, such as known concentrations or class labels (e.g. sample type, disease state), supervised multivariate methods like partial least squares (PLS) regression and its classification variant, PLS-DA, are widely used. These methods learn a mapping between spectral features and labelled outputs, which can be used to predict the composition of new, unseen spectra. However, this also supposes a fixed set of target compounds or classes and the availability of labelled compositional data.

In fully unsupervised settings, where neither reference spectra nor labels are available, dimensionality reduction techniques such as PCA and NMF are commonly used to infer dominant spectral components directly from the data. Quantitative information is then extracted from component loadings or scores, or by combining these components with regression methods such as NNLS. However, this approach also presents challenges. First, qualitative interpretation typically requires post hoc matching to known spectra, often through expert curation or spectral similarity measures (e.g. hit-quality index, HQI) and library searches. Second, both the extracted components and their estimated abundances can be highly susceptible to sample heterogeneity, experimental noise and artefacts.

General challenges in Raman chemometrics of biological samples

Despite the significant advances in RS and the wide range of chemometric techniques available, the analysis of RS data remains a major challenge, particularly in the context of biological systems. Raman measurements from complex, heterogeneous samples, such as cells, tissues and organoids, often yield noisy, highly convoluted spectra containing overlapping spectral features arising from a diverse mix of molecular species coexisting and interacting at micro- to nanoscales. This complexity hinders the qualitative and quantitative investigation of RS measurements[40,41]. Many established methods for chemical analysis struggle to extract meaningful and interpretable information under such conditions, limiting the effectiveness of RS for biological imaging and analysis.

Another challenge lies in the sheer volume of data produced by modern RS experiments. High-resolution mapping, volumetric imaging and time-point studies routinely generate thousands to millions of spectra per experiment. Each spectrum is inherently high-dimensional, capturing hundreds to thousands of wavenumbers. The resulting data volume and complexity make manual inspection impractical and often unreliable.

Addressing these issues requires not only more advanced analytical techniques that can identify meaningful spectral patterns associated with biological states or processes, but also software tools and workflows capable of handling noisy, high-dimensional data in a reproducible and scalable manner.

Deep learning approaches for Raman chemometrics

More recently, deep learning methods have emerged as a powerful framework for the analysis of RS data, gaining considerable traction over the past few years[28,193–197]. Promising more autonomous, flexible and data-driven analytics, deep neural networks been applied across a wide range of domains, spanning the identification of pathogens and other microbes[100,198–200], the characterisation of chemicals, including minerals[201], pesticides[202] and other analytes[203,204], the development of novel diagnostic platforms[205–208], and the application of techniques from computer vision for denoising and super-resolution in Raman imaging[209]. Common approaches include multilayer perceptrons (MLPs)[210], convolutional neural networks (CNNs)[211,212] and recurrent neural networks (RNNs)[213], each offering distinct advantages in how they model spectral structure in RS data. More recently, transformer-based architectures have gained significant traction, enabling more effective modelling of long-range dependencies across spectral features and often outperforming alternative methods on benchmark tasks[214].

An area of topical interest in RS is the application of deep learning methods to chemometric analysis. In recent years, a wide range of neural network architectures has been explored for qualitative and quantitative analysis of RS data, aiming to accurately infer the chemical composition of samples directly from spectral measurements[43]. However, a major challenge remains.

> Despite their advantages over classical chemometric techniques, most deep learning approaches in Raman spectroscopy share a key limitation: they are developed and evaluated in fully supervised settings. This significantly limits their applicability in real-world scenarios, where ground-truth information about sample composition is often unavailable.

In supervised settings, models are learned using ground-truth annotations for predefined classification or regression tasks. For instance, approaches for compositional characterisation of Raman measurements often rely on explicit information about the presence (or absence) of particular target species[215] or known concentrations[212,214,216–218], which is provided during model training. These tasks may take various forms, including binary classification (e.g. species A present or not), multi-class discrimination (e.g. species A vs. species B), or quantitative estimation (e.g. 10 µL vs. 20 µL concentration levels).

However, such annotations are rarely available in practice, particularly in domains involving complex or poorly characterised samples. In the context of this study, the precise molecular composition of organoid tissues is generally unknown and cannot be feasibly determined at scale under varying experimental conditions using orthogonal techniques. While partial compositional information can be obtained through complementary characterisation techniques such as fluorescence microscopy[81,219–221], the resulting annotations are typically limited to a few targets, leaving many molecular species unlabelled. They also tend to provide categorical rather than quantitative supervision, compressing subtle biochemical variation important for downstream analysis. In addition, cross-modal co-registration is often non-trivial due to differences in imaging parameters, sample preparation and fields of view.

More fundamentally, most supervised models are inherently trained on a fixed corpus of chemical species or concentration classes/ranges, limiting their generalisability to new or previously unseen components. Furthermore, enumerating all relevant species in complex, heterogeneous samples, such as cells or tissues, is generally not feasible. This poses a significant barrier to deploying such models in open-world or discovery-driven applications, where both the presence and concentration of chemical species may vary unpredictably and are often not known in advance.

These challenges underscore the need for unsupervised, general-purpose chemometric approaches capable of learning and extracting meaningful information from unlabelled RS data without known target substances and concentrations. Developing such techniques is essential to realise the potential of RS for label-free chemical characterisation.

Unsupervised chemometric analysis via hyperspectral unmixing

To address these limitations, this thesis develops methods for fully unsupervised Raman imaging and analysis based on the framework of hyperspectral unmixing.

Hyperspectral unmixing (also known as (hyper)spectral deconvolution or multivariate curve resolution) refers to the task of decomposing complex spectral measurements into their constituent chemical components[222,223]. It offers a general framework for qualitative and quantitative analysis, aiming to identify the underlying species present in a sample and estimate their relative proportions.

Raman spectra can be represented as vectors $\mathbf{x} \in \mathbb{R}_+^b$, whose components correspond to the intensity of inelastically scattered light binned over b wavelength or wavenumber bands. In hyperspectral unmixing, these measurements are treated as the result of an underlying mixing of n 'pure' components, defined by their Raman signatures $\mathbf{m}_i \in \mathbb{R}_+^b$, $i = 1, \ldots, n$ (**endmembers**), and their respective proportions $\{\alpha_i\}_{i=1}^n$, $\alpha_i \in \mathbb{R}_+$ (**abundances**).

Methods for hyperspectral unmixing aim to resolve mixed measurements $\mathbf{x}$ by identifying the individual components present $\{\mathbf{m}_i\}_{i=1}^n$ (**endmember identification**) and/or quantifying their proportions $\{\alpha_i\}_{i=1}^n$ (**abundance estimation**). This problem can be approached under two settings:

- **Non-blind unmixing:** the endmember spectra $\mathbf{m}_i$ are known in advance (e.g. from a reference database), and only the abundances need to be estimated.
- **Blind unmixing:** both the endmembers and their abundances are unknown and must be inferred from the observed spectral data.

This thesis focuses on blind unmixing, which is a more general framework that does not require reference spectra or labelled training data and is therefore well-suited to exploratory scenarios where prior knowledge of sample composition is limited or unavailable. Nonetheless, the proposed techniques are readily adaptable to the simpler problem of non-blind unmixing, as discussed in Chapter 4.

Note that in blind hyperspectral unmixing, the number of endmembers n to extract is a key hyperparameter. Choosing n directly influences the calculated abundance maps, as underestimating or overestimating the number of endmembers can lead to mixing of distinct spectral components or splitting of single components across multiple endmembers.

A major challenge in unmixing is that it is, in general, an **ill-posed and unidentifiable problem**[224]. In the blind setting, there is insufficient information to uniquely recover the underlying components and their abundances without additional assumptions. This is due to the lack of prior knowledge about the mixing process, the number and nature of the components (endmembers), and the structure of the spectral data. Achieving accurate unmixing in a data-driven manner – i.e. directly from observed measurements, without manual curation or prior domain knowledge – remains a central challenge in RS.

To make the problem more tractable, it is common to impose a specific mixing model that governs how the observed spectra are generated from the underlying components. The simplest and most widely used formulation is the **linear mixing model (LMM)**, where measurements are assumed to be a linear combination of the endmembers:

$$\mathbf{x} = M\boldsymbol{\alpha} = \sum_{i=1}^{n} \alpha_i \mathbf{m}_i, \tag{2.2}$$

where $M = \begin{bmatrix} \mathbf{m}_1 & \mathbf{m}_2 & \cdots & \mathbf{m}_n \end{bmatrix}$ is an $b \times n$ non-negative matrix containing the n endmember signatures, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_n)^\intercal$ is an $n \times 1$ vector storing the corresponding abundances. A random noise term $\boldsymbol{\epsilon} \in \mathbb{R}^b$ may also be added to Equation 2.2 to model stochastic variations. The abundances α_i are constrained to be non-negative (i.e. the **abundance non-negativity constraint (ANC)**, $\alpha_i \geq 0, \forall i$), and are forced to sum to 1 when corresponding to proportions (i.e. the **abundance sum-to-one constraint (ASC)**, $||\boldsymbol{\alpha}||_1 = 1$).

The LMM is a good approximation when endmember species are spatially well-separated with respect to the focal volume, and complex light interactions that may cause non-linear signal contributions can be neglected[225]. However, when non-linear interactions become significant, more intricate models are required[226]. To represent phenomena such as multiple scattering events, topographic variances and shadowing effects, the LMM has been extended in remote sensing to more complex variants[226,227], including intimate mixture models[228], bilinear models[229–231], multilinear models[232], or post-nonlinear models[233], among others. A popular bilinear mixing model is the Fan model[230]:

$$\mathbf{x} = \sum_{i=1}^{n} \alpha_i \mathbf{m}_i + \sum_{k=1}^{n} \sum_{\substack{l=1, \\ l \neq k}}^{n} \alpha_k \mathbf{m}_k \odot \alpha_l \mathbf{m}_l, \tag{2.3}$$

where $\odot$ is the Hadamard product. However, accounting for non-linear mixing interactions increases the complexity and computational cost of unmixing[226,227], an issue of especial relevance in RS where datasets (e.g. imaging/volumetric scans) are typically larger than in remote sensing. Hence, despite its limitations, the LMM remains a cornerstone of hyperspectral unmixing in most practical settings[222,224].

For more information about hyperspectral unmixing, the reader is pointed to previous works by Keshava and Mustard[225] and Li et al.[222].

Standard methods for hyperspectral unmixing

The linear mixing assumption allows the development of tractable approaches for unmixing. Popular approaches, originated in remote sensing, include N-FINDR[234] and vertex

component analysis (VCA)[235] for endmember identification, and NNLS[192] and fully constrained least squares (FCLS)[236] for abundance estimation[222,237]. N-FINDR and VCA are geometric methods based on the concept of a simplex in Euclidean space. N-FINDR exploits the fact that, under Equation 2.2, endmembers represent vertices of a simplex spanning the data, and operates by iteratively finding a set of points (endmembers) that maximise the volume of the simplex they form. In contrast, VCA finds endmembers by projecting the data onto directions orthogonal to the subspace spanned by previously found endmembers and identifying new endmembers as the farthest points in these directions, effectively constructing a simplex that encompasses all data points. In both methods, the number of endmembers to extract is specified *a priori* by the user. Once endmember signatures M are derived, optimisation-based algorithms such as NNLS and FCLS are employed to estimate the abundances $\boldsymbol{\alpha}$ for a given spectrum $\mathbf{x}$ by minimizing the reconstruction error between the observed data and the model $\min_{\boldsymbol{\alpha}} \|M\boldsymbol{\alpha} - \mathbf{x}\|^2$. NNLS imposes the ANC, whereas FCLS imposes both the ANC and ASC.

However, these methods have several significant limitations[225,238]. First, they are inherently restricted to linear unmixing and thus fail to capture potential non-linear interactions that can be present in real-world Raman spectra. They are also highly sensitive to common artefacts, such as dark noise, baseline fluctuations and cosmic spikes, which are often misidentified as genuine endmember components rather than as outliers. Moreover, approaches like N-FINDR and VCA assume that the underlying chemical constituents are present as 'pure pixels' in the dataset (i.e. individual pixels that contain only a single component). Yet, this assumption does not hold in many practical scenarios, particularly in highly heterogeneous samples such as biological tissues. Finally, algorithms like FCLS can become computationally prohibitive when applied to large datasets, limiting their practicality for high-resolution imaging or volumetric scans.

Unmixing autoencoders

To address these limitations, this thesis frames the problem of hyperspectral unmixing as an autoencoder learning problem.

Autoencoder (AE) neural networks have recently emerged as a framework for hyperspectral unmixing in remote sensing, spurred by the availability of standardised benchmark datasets for model evaluation (e.g. *Urban*, *Samson*, *AVARIS Cuprite*)[239–243]. Yet, despite initial explorations[244,245], the utility of unmixing AEs for RS data remains largely unexplored.

Autoencoders are a family of (deep) neural network models consisting of two sub-networks (*encoder* and *decoder*) connected sequentially[246]. To illustrate how they can be adapted for hyperspectral unmixing, consider an AE model consisting of an encoder $\mathcal{E} : \mathbb{R}^b \to \mathbb{R}^m$ that maps an input spectrum $\mathbf{x}$ to a latent representation $\mathbf{z} = \mathcal{E}(\mathbf{x})$ in a lower-dimensional space, typically $m \ll b$. This latent vector is then passed through a decoder $\mathcal{D} : \mathbb{R}^m \to \mathbb{R}^b$ to reconstruct the original input:

$$\widehat{\mathbf{x}} = \mathcal{D}(\mathbf{z}) = \mathcal{D}(\mathcal{E}(\mathbf{x})). \tag{2.4}$$

Notice that the decoder $\mathcal{D}$ can be understood as playing the role of a mixing function on the representation $\mathbf{z}$. For instance, consider a linear decoder $\mathcal{D}_{\text{Lin}}$ consisting of a single linear layer defined by a $b \times m$ weight matrix W, then:

$$\widehat{\mathbf{x}} = \mathcal{D}_{\text{Lin}}(\mathbf{z}) = W\mathbf{z}\,. \tag{2.5}$$

It follows from the formulation of the LMM (Equation 2.2) that the latent representations $\mathbf{z}$ resemble the abundances $\boldsymbol{\alpha}$, the weight matrix W resembles the matrix of endmembers M, and the dimensionality m of the latent space defines the number n of endmembers to learn. To reinforce the physical interpretation of unmixing into the AE learning, relevant constraints, such as the non-negativity of W and the non-negativity (ANC) or sum-to-one constraint (ASC) of $\mathbf{z}$, are enforced by applying appropriate choices of activation functions and penalties during training.

This framework can accommodate non-linear mixture models through the design of the decoder. For instance, the bilinear Fan model (Equation 2.3) can be implemented by extending $\mathcal{D}_{\text{Lin}}$ to account for the additional bilinear terms:

$$\widehat{\mathbf{x}} = \mathcal{D}_{\text{Bilin}}(\mathbf{z}) = W\mathbf{z} + \sum_{k=1}^{m} \sum_{\substack{l=1,\\ l\neq k}}^{m} z_k \mathbf{w}_k \odot z_l \mathbf{w}_l, \tag{2.6}$$

where z_k, z_l are components of $\mathbf{z}$, and $\mathbf{w}_k, \mathbf{w}_l$ are column vectors of W. Similarly, one can devise decoders suited for other mixture models[243,247,248], or adopt a general decoder that learns the underlying mixing model in a more data-driven manner, at the cost of interpretability of the extracted endmembers and abundances. Finally, note that the AE unmixing framework can be directly adapted for non-blind unmixing by fixing the weight matrix W in the decoder to a given set of predefined endmembers.

2.2 Neural organoids: Modelling the human brain in a dish

Building on the discussion of Raman spectroscopy, this section introduces neural organoids. It begins by outlining major challenges in brain research and highlighting the need for *in vitro* models of the human brain (Section 2.2.1), followed by a brief historical overview of the cell culture advances that led to the development of three-dimensional organoids (Section 2.2.2). A detailed discussion of neural organoids is then presented, including how they are generated *in vitro* and why they represent a significant advancement in biomedical research (Section 2.2.3). Key applications of neural organoids are reviewed (Section 2.2.4), and current limitations in imaging and analysis techniques are examined, motivating the application of RS for organoid characterisation and analysis (Section 2.2.5).

2.2.1 The need for better biological models of the human brain

The human brain is the most complex organ in the human body. Weighing approximately three pounds, it comprises billions of neurons and trillions of synaptic connections, organised into an intricate hierarchy of specialised yet interconnected regions[249–251]. Over millions of years of evolution, the human brain has developed into a defining feature of the human species, underpinning human cognition, behaviour and many vital physiological functions. As such, understanding how the brain works has long been a fundamental question in biology, as well as a focal point of broader philosophical inquiry[252,253].

Central to this pursuit is the development of tools that allow us to probe the brain's structure, function and development[254]. Such tools are essential not only for uncovering the biological basis of cognition and behaviour, but also for studying how the human brain develops, identifying what goes awry in complex neurological and psychiatric disorders, and informing the design of effective therapeutic strategies.

Challenges in studying the developing human brain

Yet, studying the human brain remains a major scientific challenge, marked by significant technical, ethical and practical constraints. One of the primary challenges is the inherent inaccessibility of the brain. Direct investigation is rarely possible: *in vivo* studies in humans are largely infeasible, live brain tissue is generally inaccessible, and post-mortem samples provide only static snapshots, often lacking the dynamic context required to study development or disease progression.

These limitations are even more pronounced in the context of early human brain development. This period, which occurs during embryogenesis, involves rapid and tightly regulated processes, such as progenitor proliferation, neuronal differentiation, migration and circuit formation, which establish the foundation for brain architecture and function[255]. Probing the cellular and molecular mechanisms underlying these events is essential for advancing our understanding of brain development and developmental disorders, and for designing more targeted therapies and drug candidates. However, access to embryonic or fetal brain tissue is highly restricted due to ethical and legal constraints[256,257], and the ability to interrogate dynamic processes in these tissues across spatial and temporal scales within physiologically relevant environments remains limited[258]. As a result, many aspects of early human brain development and the developmental diseases associated with it remain poorly understood.

Experimental models in brain research

In the absence of direct access to the human brain, neuroscience research has long relied on alternative experimental models, including 2D cell cultures and a range of animal systems. These models have been instrumental in advancing our understanding of neural development, brain structure and function, and the mechanisms underlying neurological disease[256,259–262]. The most widely used experimental systems and their relevance to human brain research are briefly discussed below.

- **2D cell cultures.** Two-dimensional cell culture systems, including immortalised cell lines, primary neural cultures and stem cell-derived neural cells, are widely used in neuroscience due to their accessibility and ease of manipulation[263–267]. While they lack the complexity of *in vivo* models, 2D cultures enable high-throughput genetic, molecular and pharmacological studies, making them well suited for investigating cell-intrinsic processes such as cell proliferation, differentiation and signalling.
- ***Ex vivo* tissues.** Building on cellular models, *ex vivo* preparations such as acute brain slices, organotypic cultures and biopsies serve as valuable intermediate systems in neuroscience research[268–271]. Positioned between *in vitro* and *in vivo* approaches, these models – typically derived from freshly collected animal tissue and, more rarely, from human surgical or postmortem brain samples – offer improved physiological relevance compared to dissociated cultures, while retaining a degree of experimental accessibility. They preserve key structural and functional features of the nervous system, including native cytoarchitecture, cellular diversity and aspects of the local microenvironment, and enable high-resolution, controlled investigations of cellular, synaptic and circuit-level processes.
- **Rodents.** Going beyond cellular and tissue preparations, whole-animal models of-

fer the complexity of intact, living systems. Rodents, particularly mice and rats, remain the most widely used animal models in neuroscience, critical for studying brain structure, function, disease processes and drug response[272–275]. The availability of advanced genetic engineering techniques and well-established methodologies for neuroimaging, behavioural and electrophysiological studies further supports their widespread use[276,277].

- **Larger mammals and non-human primates.** Larger mammals, including pigs, dogs and ferrets, are also employed as experimental model systems in neuroscience, recapitulating important anatomical, functional and social characteristics relevant to the human brain[278–280]. Non-human primates, such as marmosets, lemurs and macaques, extend this further, providing models for neuroscience research and disease modelling that more closely resemble aspects of human brain evolution, structure and function[281–284].

Each of these experimental model systems comes with distinct strengths and limitations, and the choice of model depends heavily on the application and the biological question being addressed. In many cases, a combination of models may be used. For instance, drug discovery often involves initial testing in cell cultures to assess basic efficacy, mechanisms of action and potential cytotoxicity, followed by toxicity and dosing studies in rodents and/or larger animals, before progressing to clinical trials in humans[285,286].

Limitations of standard models

While 2D cell culture and animal model systems have been instrumental tools in advancing neuroscience, they present several important limitations. These are discussed below, focusing on human brain development.

- **Limited structural and cellular complexity in 2D cell cultures.** Two-dimensional monolayer cultures lack the three-dimensional architecture, spatial organisation and cellular diversity that define the human brain. As a result, they fail to capture key features of brain biology, including region-specific interactions, layered structures, complex circuitry, and diverse cell-cell and cell-matrix interactions[287–289]. Without the microenvironmental and topological complexity of native tissue, these systems offer limited relevance for modelling tissue morphogenesis and cellular heterogeneity, both of which are crucial for understanding early brain development and its disruption in disease.
- **Inherent differences between human and animal biology.** While animal models remain essential to biomedical research, they often fail to replicate critical aspects of human brain development and physiology[257,290]. This leads to significant

differences in brain size, structure, cellular diversity, developmental timing, gene regulation and immune function[291–295]. For instance, key hallmarks of human brain development, including neocortical expansion, the presence of outer radial glia and prolonged cortical neurogenesis, are absent or poorly represented in widely used species such as rodents and small mammals[296–298]. Furthermore, many neurological and neurodevelopmental disorders, such as Parkinson's disease, autism spectrum disorder and schizophrenia, lack analogues in common animal models and cannot be adequately studied[299,300].

- **Limited accessibility and viability of *ex vivo* tissue models.** *Ex vivo* human tissues are scarce, difficult to obtain and highly variable due to differences in genetic background, environmental exposures and sample handling. Furthermore, these tissues often represent static snapshots, generally of late-stage states, which are not representative of processes *in vivo*[301]. While animal-derived slices are more accessible, they may not fully capture human-specific features. As a result, studying complex, dynamic biological events, including early developmental trajectories, progressive disease mechanisms and longitudinal functional changes using *ex vivo* models remains challenging.
- **Practical and ethical constraints.** Access to human subjects and tissues for research, particularly embryonic or fetal brain tissue, is extremely limited due to ethical, regulatory and practical considerations[302–305]. Animal studies also face significant constraints, specifically in the context of invasive experiments, or those involving primates or transgenic models[256,257,284,306]. To ensure ethical standards are maintained, stringent guidelines governing the use of animals in research have been established[307].

As a result, our ability to study the human brain and neurobiology remains limited. Key aspects of brain development, including cellular dynamics, structural organisation, tissue morphogenesis and disease mechanisms, are still largely inaccessible using standard models. This constrains research into the fundamental biology of brain formation, function and developmental disorders. It also contributes to the poor translation of experimental findings to clinical contexts, with estimates suggesting that approximately 90% of candidate drugs ultimately fail in human trials[308–310] – a pressing challenge amid the growing global burden of neurological disorders[311–313].

In recognition of these limitations, regulatory bodies such as the U.S. Food and Drug Administration (FDA) have launched initiatives to reduce animal testing in preclinical safety studies, encouraging the development and adoption of more predictive, human-relevant models[314].

Developing alternative experimental neuroscience models that are accessible, experimentally tractable and more accurately recapitulate the unique features of the developing human brain – along with the tools required to study them – is critical for advancing our understanding of brain development and disease, improving preclinical success rates, and accelerating the discovery of effective therapies.

2.2.2 From 2D to 3D neural cell cultures

Despite their limitations, 2D cell cultures remain one of the most accessible and widely used tools in biological research, offering a flexible platform to investigate cellular processes under controlled conditions.

Two-dimensional neural cell cultures and their limitations

The origins of cell culture date back to the early 20th century, with pioneering work on tissue dissociation and the cultivation of mammalian cells outside the organism that laid the foundation for modern *in vitro* systems[315–317]. In the decades since, cell culture developed into an indispensable research tool, becoming deeply embedded in experimental workflows across molecular biology, biomedicine, pharmacology and regenerative medicine.

Among the various platforms available, two-dimensional culture systems, where dissociated cells are plated onto flat substrates and grow as adherent monolayers, remain the most widely used. Owing to their simplicity, reproducibility, scalability and compatibility with imaging and molecular assays, 2D cultures have become central to biological research, serving as key models for studying cell proliferation, differentiation, cell-cell interactions, signalling pathways, disease mechanisms, therapeutic responses, and other fundamental aspects of cell biology. In neuroscience, 2D culture systems have facilitated the investigation of various aspects of neural development and function, including cell regulation and differentiation[318], network formation[319], neurotransmitters[320], disease mechanisms[321] and pharmacological response[322]. Early work relied heavily on primary neural cultures derived from rodent brain tissue and immortalised cell lines engineered to proliferate indefinitely (e.g. SH-SY5Y, PC12), which provided accessible and scalable models for *in vitro* research[323]. Despite contributing to many discoveries, these systems still exhibited significant limitations, including species-specific differences, restricted differentiation potential and limited physiological relevance compared to human neural tissue.

A major advance in cell culture came with the isolation of **pluripotent stem cells**

(PSCs), which have the capacity to differentiate into cell types derived from all three germ layers. Mouse embryonic stem cells (ESCs) were first established in the early 1980s[324,325], followed by human ESCs in 1998[326]. These breakthroughs opened up new avenues for research in developmental biology, disease modelling and regenerative medicine, yet the use of ESCs remained ethically and legally contentious, primarily due to the destruction of embryos[327,328]. This limitation was largely overcome with the generation of induced pluripotent stem cells (iPSCs) by Shinya Yamanaka and Kazutoshi Takahashi in 2006, who managed to reprogram somatic cells into a pluripotent state using specific transcription factors[329,330]. iPSCs offer a renewable, patient-specific and more ethically viable alternative to ESCs, and have since become central to a wide range of research and translational applications, including in neuroscience[331].

Many early neural differentiation protocols from PSCs relied on 2D monolayer cultures[332]. An alternative approach involved forming **embryoid bodies (EBs)**, in which mouse and human ESCs self-assemble into 3D aggregates when cultured in suspension. EBs recapitulate key aspects of early embryogenesis, demonstrating lineage emergence and cellular self-organisation *in vitro*, giving rise to derivatives of all three germ layers[333–335]. Using this approach, Zhang et al. showed that human ESCs can produce neural progenitors, neurons and glia, and spontaneously form neural rosettes resembling the early neural tube[336]. To enhance lineage specificity and region-specific neural induction, many studies explored the use of different morphogen cues[337–339].

Nevertheless, despite the use of 3D EBs, most of these methods remained fundamentally two-dimensional, with EBs serving as a transient priming step before being plated onto adherent substrates where neural induction and differentiation proceeded in monolayers. Consequently, many of the inherent limitations of 2D cultures, such as restricted spatial architecture, cell-cell interactions and microenvironmental gradients, were largely unaddressed.

Towards three-dimensional neural cell cultures

There was a growing consensus that advancing beyond 2D monolayer cultures was essential to develop models that more accurately recapitulate the structural and functional complexity of the human brain *in vivo*[340]. While EBs introduced some degree of three-dimensionality, more sustained and developmentally relevant 3D culture systems were required.

One of the earliest approaches to establishing neural-specific 3D cultures was the **neurosphere assay**, developed by Reynolds and Weiss in 1992[341]. In this method, neural stem or progenitor cells isolated from fetal or adult brain tissue were cultured in suspen-

sion with mitogenic factors, resulting in the formation of free-floating 3D cell aggregates capable of differentiating into neurons, astrocytes and oligodendrocytes. While the assay proved useful for studying basic properties of neural stem cells, neurospheres were heterogeneous in size and cellular composition and lacked spatial organisation, regional identity and developmental relevance.

To overcome these limitations, researchers turned to PSCs, aiming to exploit their self-organising potential. Building on EB-based protocols, Reubinoff et al. demonstrated that human ESCs could generate neural progenitors that formed 3D aggregates expressing basic neural markers, though without organised tissue architecture[342]. In a subsequent study, Eiraku et al. showed that patterned neural induction of mouse ESC-derived EBs produced neuroepithelial tissues with apical-basal polarity and a neural plate-like structure[343]. Maintained entirely in suspension, these aggregates provided further evidence that PSCs can self-organise into structured neural tissues in three dimensions. These PSC-derived 3D aggregates, broadly called **neural spheroids**, enabled enhanced spatial patterning and cell-cell interactions, leading to the emergence of early brain-like features not achievable in 2D cultures. Later studies demonstrated more precise regional specification through patterned neural induction, resulting in spheroids with distinct forebrain subdomains[344,345].

These studies marked an important conceptual shift toward 3D models as more physiologically relevant systems for studying human brain development. Despite these advances, however, early spheroid models remained limited in complexity and cellular diversity. Although more complex models, termed **organoids**, had been developed for several organs by the early 2010s, including the intestine[346], pituitary[347], retina[348] and thyroid[349], a comprehensive 3D model of the developing human brain had yet to be achieved.

2.2.3 Neural organoids

Generation of neural organoids

This line of research culminated in the landmark study by Lancaster et al. in 2013, who reported the development of three-dimensional **neural organoids*** derived from human PSCs[7]. These organoids, which reached several millimetres in size, developed multiple

*The tissues described in the study by Lancaster et al.[7] were originally termed "cerebral organoids" – from the Latin *cerebrum*, meaning brain. Since then, the field has adopted broader and more anatomically precise terminology[350]. To avoid confusion with specific brain regions like the cerebellum and align with current nomenclature, the term "neural organoids" is used throughout this thesis when referring to PSC-derived brain organoids, particularly when regional identity has not been definitively characterised.

discrete yet interconnected brain region-like domains, including retinal tissue, forebrain, choroid plexus, and midbrain/hindbrain structures. They recapitulated key aspects of early human cortical development, forming ventricular zone-like structures, rudimentary cortical plate-like layers, abundant outer radial glia and various subtypes of mature and functional cortical neurons. Interestingly, the authors achieved this by modifying the culture conditions rather than applying external patterning factors. In particular, they embedded cell aggregates in Matrigel to support epithelial polarity and lumen formation, and used spinning bioreactors to improve nutrient and oxygen diffusion. These conditions, intended to promote the intrinsic self-organisation of PSCs, appeared to activate innate developmental programs, resulting in structures resembling early corticogenesis. This approach is referred to as unguided differentiation (or self-patterning), and the resulting tissues as unguided neural organoids.

In the same study, Lancaster et al. also demonstrated the potential of neural organoids for modelling human neurodevelopmental disorders[7]. Using patient-derived human iPSCs, they generated organoids that recapitulated primary microcephaly – a condition that had been difficult to model in mice. Organoids derived from microcephaly patients exhibited reduced overall size and premature neuronal differentiation, providing insight into potential mechanisms underlying the disease phenotype *in vivo*. This work served as proof of concept that PSC-derived neural organoids can model patient-specific developmental and pathological features *in vitro*.

Since then, various alternative protocols have been established. In addition to unguided differentiation approaches that rely on intrinsic cues, many guided differentiation methods have been developed, using external patterning with morphogens to produce more homogeneous tissues with defined regional identities. In particular, protocols have been created for generating region-specific neural organoids, such as optic cup[348], forebrain[351], midbrain[352], hindbrain[353] and striatum[354]. These regionally patterned organoids have also been used in the context of assembloid systems, in which pre-patterned neural organoids with distinct regional identities are fused to model interactions between different brain areas[9,355–357]. Advances on the bioengineering side, such as the use of tailored biomaterials, scalable culture systems and protocols for extended long-term growth, have further enhanced the maturation, cellular diversity and architectural complexity of neural organoids[358–362]. Organoid culture systems have also since been extended to a range of organs beyond the nervous system, including the kidney[363], stomach[364], lung[365], and heart[366], among others. In addition, more comprehensive embryo-like models have been developed, such as gastruloids[367,368] and embryonic trunk-like structures containing somites[369].

Together, these advances have broadened the structural complexity and functional ca-

pabilities of neural organoids, positioning them as versatile systems for studying human neurodevelopment and neurobiology, modelling neurological diseases, and assessing therapeutic responses *in vitro*. For a more comprehensive historical overview of organoid research and the evolution of neural organoid technologies, the reader is referred to the works by Lancaster and Knoblich[8], Pașca[9], Simian and Bissell[370], Koo et al.[371], Corrò et al.[372], Lancaster[373].

In Chapter 5 of this thesis, unguided neural organoids generated following the original protocol by Lancaster et al.[7] are analysed.

Advantages of neural organoids

While they remain imperfect representations of the human brain, neural organoids provide a powerful experimental platform that addresses many of the limitations of 2D cell cultures and animal models[8–10,15,374]. The main advantages of neural organoids in this context are summarised below.

- **3D architecture.** Neural organoids exhibit a spatially organised, three-dimensional architecture that more closely reflects the structure of the human brain *in vivo* compared to conventional 2D monolayer cultures. This structural context supports cell-cell and cell-matrix interactions and establishes more physiologically relevant gradients of oxygen, nutrients and morphogens, guiding processes such as cellular differentiation, neuronal migration and cortical tissue organisation[7,287–289,351].
- **High cellular diversity.** Neural organoids are capable of generating a wide variety of neural cell types, including progenitors, excitatory and inhibitory neurons, astrocytes and oligodendrocyte precursors[11,12]. This level of cellular diversity is typically lacking in 2D cultures and poorly represented in animal models, where species-specific differences influence lineage specification and maturation[296].
- **Human-specificity and patient-specific modelling.** Neural organoids derived from human PSCs capture aspects of human brain development that are difficult to model in animals, including the presence of abundant outer radial glia, prolonged progenitor proliferation, expanded cortical layer and genetic regulation[7,12,351,375–378]. In addition, organoids generated from patient-derived iPSCs offer genetically tailored, patient-specific models of neurological disease, enabling personalised investigation of genetic variants, molecular pathologies and differential drug responses to advance precision neuroscience[7,357,379–381].
- **Access to early human brain development.** Neural organoids recapitulate hallmark features of early human neurodevelopment, including the formation of

ventricular zones, cortical layering and radial glial scaffolds, which are largely absent in 2D systems and only partially represented in animal models with divergent developmental programs[7,10,15,377]. They also reproduce sequential developmental events such as neuroepithelial rosette formation, migration of neurons along radial glia and laminar specification[8]. Over extended culture, organoids also establish synaptic connections and display evolving patterns of spontaneous electrophysiological activity, providing access to circuit-level biology and drug responses in a human-specific context[13,382].

- **Experimental scalability and flexibility.** Organoids can be generated from diverse PSC lines and produced at scale, enabling high-throughput studies across individuals and disease cohorts[383]. When combined with gene editing and drug screening technologies, they provide a flexible platform for functional genomics, perturbation assays and therapeutic testing – capabilities that are often limited in animal models or poorly represented in traditional 2D cultures[384–386].
- **Ethical advantages.** While some ethical considerations remain as organoid technology advances, neural organoid models address many limitations associated with animal and human research, providing an accessible experimental platform for studying early human brain development and developmental diseases in physiologically relevant contexts, which reduces the use of human embryonic or fetal tissue[387–389].

Overall, neural organoids recapitulate key aspects of the organisation, dynamics and heterogeneity of the developing human brain *in vivo*, providing an accessible and physiologically relevant platform for studying human brain development, neurological disorders and potential therapeutic interventions.

2.2.4 Applications of neural organoids in brain research

The advantages of neural organoids over traditional models have established them as a valuable addition to the neuroscience and biomedical research toolkit, enabling a wide range of applications in developmental biology, disease modelling, drug screening and beyond. This section reviews key applications of neural organoids across these domains.

Although the primary focus of this thesis is on brain development, related applications in other areas are also briefly discussed for completeness, particularly given the potential for the tools developed here to be applied in these contexts as well.

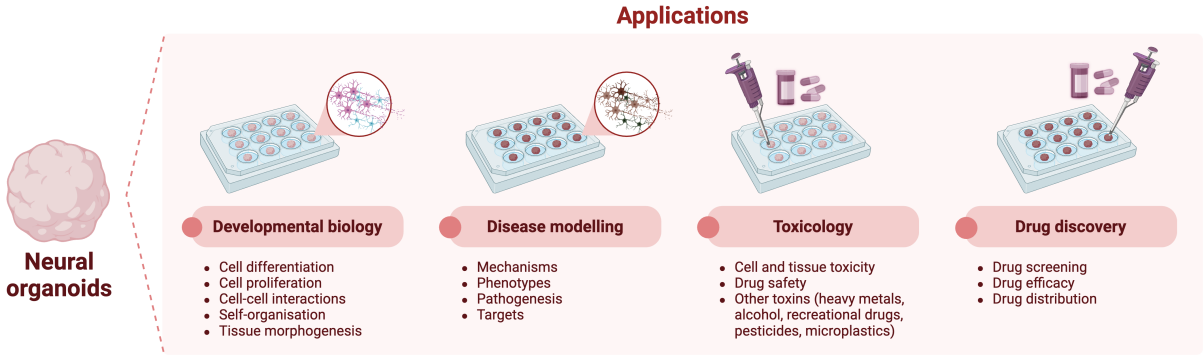


Figure 2.4: Overview of key application areas of neural organoids. The figure highlights major applications of neural organoid models across several domains, including developmental biology, disease modelling, neurotoxicity testing and drug discovery. Created in BioRender.com.

Developmental biology

One of the primary applications of neural organoids – and the main focus of this thesis – is in developmental biology, offering access to developmental structures, stages and processes that are otherwise inaccessible in 2D cultures and animal or human embryos[8,10,15,377].

Neurogenesis, patterning and morphogenesis. Neural organoids have provided insight into the molecular and mechanical processes driving cortical morphogenesis, including the formation of radial units, laminar structure and regional patterning. Single-cell sequencing has revealed human-specific features of early neurogenesis and the intrinsic transcriptional programs that guide cortical architecture[378,390]. Studies using regionally patterned organoids have further shown how morphogen gradients specify neuronal subtypes across distinct brain areas[391]. Structural analyses have also demonstrated how cytoskeletal tension, mechanical forces and epithelial dynamics shape lumen formation, folding and progenitor zone architecture, influencing cell fate transitions and lineage spatialisation[392].

Cell type diversity, lineage trajectories and signalling. Single-cell sequencing and lineage tracing in neural organoids have enabled the reconstruction of developmental trajectories and the identification of diverse neural and glial populations[12,378,393–395]. Neural organoids have also been applied to investigate cell-cell interaction and signalling and their downstream effects on cortical morphogenesis, regional patterning, cytoarchitecture and function in both health and disease[396–398].

Network formation and circuits. Calcium imaging, multi-electrode arrays and electrophysiology have been used to characterise the emergence of spontaneous activity, synaptic connectivity and functional network dynamics in neural organoids[399,400]. These

studies have revealed oscillatory activity and coordinated firing patterns reminiscent of early *in vivo* development[13], and how they are altered in disease states such as Rett syndrome[401]. Assembloid models have further enabled the study of long-range axonal projections and interregional connectivity[355], and the impact of disease[357] and drug treatments[402] on circuit function.

Disease modelling

Disease modelling represents another major application of neural organoids, offering a human-specific experimental system to investigate the aetiology and pathogenesis of a wide range of neurological disorders[403–405].

Neurodevelopmental disorders. Naturally, neural organoids have become a valuable tool for studying neurodevelopmental disorders[406,407]. By using organoids derived from patient-specific human iPSCs or genetically modified cell lines, researchers can investigate how disease affects key developmental processes such as cell proliferation, differentiation and cortical organisation. This approach has been applied to a variety of conditions, including microcephaly[7], autism spectrum disorder[379,408], Miller-Dieker syndrome[409], Timothy syndrome[357], Rett syndrome[410], and different forms of epilepsy[411].

Neurodegenerative diseases. Advances in long-term culture have enabled the use of neural organoid systems to model late-onset neurodegenerative diseases, offering valuable insights into the cellular and molecular mechanisms underlying conditions such as Alzheimer's disease[412], Parkinson's disease[380], Huntington's disease[413] and amyotrophic lateral sclerosis[414]. These models enable the study of key pathological processes, such as protein aggregation, neuronal loss, synaptic dysfunction and neuroinflammation, which drive disease progression and are often difficult to capture using traditional models[289,415].

Infectious diseases. Neural organoids have been applied to the study of neurotropic infections[416]. During the Zika virus epidemic, organoid-based models played a significant role in uncovering how viral infection disrupts brain development, with infected organoids exhibiting impaired neurogenesis, increased apoptosis of neural progenitors and microcephaly-like phenotypes[358,417–419]. Similar approaches have been used to model infections caused by viruses such as SARS-CoV-2[420,421], cytomegalovirus[422] and HSV-1[423], providing insights into viral entry, innate immune responses and tissue damage in a human-specific context.

Cancer. In oncology, neural organoids serve as a valuable platform for modelling brain cancers[424]. Organoids derived from patient tumours or engineered with oncogenic mutations retain key histological and molecular features, providing insights into tumour ini-

tiation, progression, heterogeneity, invasion, and interactions with the tumour microenvironment[425–427]. Glioblastoma is among the most extensively modelled brain cancers using neural organoid systems[428,429].

Drug discovery and toxicology

Neural organoids have emerged as promising platforms for preclinical drug screening and toxicological testing, with the potential to reduce reliance on animal models and improve translational success rates[430,431].

Drug discovery. Neural organoids are widely used in drug discovery, offering a human-relevant platform for evaluating pharmacological effects and screening therapeutic compounds[432–434]. Organoid-based systems have been applied to test small molecules[435], biologics[436] and gene therapies[437] across a range of neurological conditions. In the context of precision medicine, patient-derived organoids can serve as personalised avatars to predict individual treatment responses and guide therapeutic decisions[438–441]. As the field advances, organoids are being integrated into high-throughput screening pipelines combined with technologies such as single-cell transcriptomics, automated imaging and machine learning, with the aim of establishing them as robust tools for preclinical drug development[442–445].

Toxicology. Neural organoids provide a physiologically relevant, human-specific model for neurotoxicity testing, regulatory toxicology and safety pharmacology, enabling the evaluation of on-target and off-target effects of environmental toxins, pharmaceuticals and industrial chemicals on neural development and function[446,447]. For example, organoid-based assays have been used to investigate the effects of heavy metals[448], pesticides[449], alcohol[450], microplastics[451] and recreational and prescription drugs[402,452,453], allowing detection of subtle disruptions in neural progenitor proliferation, differentiation, migration and synaptic activity.

2.2.5 Organoid imaging and analysis

A critical aspect of working with organoids is the ability to visualise and characterise them. Many of the applications described in Section 2.2.4 rely on effective imaging and analysis techniques, which are essential for interrogating morphogenesis, tissue structure and spatial organisation, disease states and drug responses. Note that imaging in 3D organoid systems is inherently more challenging than in conventional 2D cell monolayers, due to increased optical scattering, structural complexity and sample thickness.

This section outlines key categories of microscopy and profiling techniques used in organoid research, focusing on morphological analysis. Each method is briefly introduced, along with its advantages and limitations in the context of (neural) organoid imaging and analysis.

Brightfield microscopy

Brightfield (BF) microscopy is one of the most fundamental and widely used imaging modalities in biological research[454]. This technique relies on the transmission of white light through a specimen, with image contrast arising from differences in light absorption, scattering and refractive index, typically associated with variations in sample thickness and density. This enables label-free visualisation of general structure and morphology, making BF microscopy a simple, fast, cost-effective and non-invasive technique, particularly well-suited for high-throughput imaging workflows and live imaging applications.

BF microscopy is a valuable tool in organoid research for assessing major morphological features such as overall shape, size and symmetry. In neural organoids, for example, it can capture important developmental features, such as neuroepithelial expansion, cyst formation and the emergence of neural tube-like structures. As an imaging technique, it has been widely applied across a variety of spheroid and organoid models to monitor morphogenesis[455], assess culture quality[456], and evaluate disease phenotypes[7,457] or responses to drug treatments[458,459]. Given its non-invasiveness and scalability, BF microscopy is a promising tool for high-throughput, real-time analysis of live organoids in large-scale longitudinal studies and drug screening platforms.

In recent years, the integration of machine learning with BF microscopy has gained traction, enabling scalable, automated organoid tracking, segmentation, phenotyping and morphological quantification,[460–463]. To facilitate the development and benchmarking of such methods, Schröter et al. have recently released a dataset containing 1400 BF images of neural organoids derived from iPSCs from healthy and diseased patients[464].

Limitations. Despite its utility, BF microscopy has several key limitations. Relying on transmitted light alone, it lacks molecular specificity and provides limited contrast, particularly in transparent samples with subtle optical differences like cells and tissues. While it enables visualisation of general outlines and gross morphology, it is challenging to resolve fine structural features such as individual cells or subcellular organelles within multicellular cultures using BF microscopy. Moreover, due to reduced recovery of transmitted light with increasing tissue thickness, brightfield microscopy is largely restricted to surface or near-surface imaging, limiting access to internal structures in complex 3D tissues such as organoids.

While more advanced techniques, such as phase contrast microscopy, differential interference contrast (DIC) microscopy and quantitative phase imaging (QPI), have been developed and offer important improvements, many of the core limitations of the aforementioned limitations remain.

Fluorescence microscopy

To enhance molecular contrast and specificity, **fluorescence microscopy (FM)** techniques have been developed and have become essential tools for the analysis of 2D and 3D cellular models[465–468]. These approaches utilise fluorescent molecules, known as fluorophores, to selectively label cellular components of interest. Upon excitation by light of a specific wavelength, these fluorophores emit fluorescence at a longer emission wavelength, enabling the visualisation of target structures with high sensitivity and specificity.

Several labelling strategies exist, each relying on different types of fluorophores and offering unique advantages depending on the biological question and sample preparation[3,469,470]. One of the simplest methods involves **fluorescent dyes**. These are small molecules that can permeate cells or tissues and bind to specific cellular components, and can be applied to both live and fixed samples. Common examples in cellular biology include DAPI, a DNA-binding dye primarily used to stain cell nuclei, and lipophilic dyes such as DiO, which label cellular membranes[471]. Another widely used technique is **immunofluorescence (IF) staining**, which employs fluorophore-conjugated antibodies designed to bind specific target molecules with higher specificity[472]. However, this approach typically requires chemical fixation and permeabilisation of cell membranes to enable antibody penetration, making it generally incompatible with live-cell imaging. **Genetically encoded fluorescent proteins**, such as green fluorescent protein (GFP) and its derivatives, are also widely employed in live-cell imaging[473]. These fluorescent tags are genetically fused to proteins of interest and expressed during cell development or differentiation, enabling real-time monitoring of protein dynamics, localisation and interactions within living cells.

Fluorescence microscopy has been instrumental in advancing neural organoid research[474,475]. IF staining combined with FM on fixed, cryosectioned samples remains a gold-standard technique for characterising morphological features, tissue architecture, cell identity, layer organisation, polarity and differentiation states within organoids. Emerging techniques for high-throughput cryoprocessing and tissue analysis have further extended this approach, enabling more efficient analysis of spatiotemporal changes in neural marker expression during early neurodevelopment[476]. Advances in optical clearing techniques have enabled volumetric 3D imaging of whole organoids, allowing the imaging and assess-

ment of internal structures in intact samples[477]. Beyond standard widefield and confocal modalities, light-sheet fluorescence microscopy has gained traction, enabling high-speed, volumetric 3D imaging of intact neural organoids with micron-scale resolution[478,479]. Live imaging capabilities are also expanding, with light-sheet and other high-speed volumetric microscopy techniques being used to monitor long-term morphodynamics and neuronal development in live organoids[480–483].

Limitations. Despite its strengths, fluorescence microscopy has several inherent limitations. First, while protein labelling is generally well-established, detecting other biomolecular classes, such as carbohydrates, lipids and various small molecules (e.g. drugs and metabolites), often requires more specialised probes or labelling strategies, or is not possible altogether. Additionally, IF staining involves multiple washing and permeabilisation steps, which can disrupt cellular architecture and compromise membrane integrity, potentially leading to artefactual results[484]. These steps may also displace small molecules such as drugs, metabolites and other transient compounds, obscuring their true spatial localisation *in situ* within cells and tissues. While other labelling strategies, such as fluorescent protein tagging, genetic modifications and dye-based staining, may help avoid these washing steps, they remain label-based (i.e. targets must be known and available), and issues with labelling non-proteins persist. Furthermore, regardless of the labelling strategy, the intrinsic spectral overlap of fluorescent markers constrains multiplexing, typically to four targets in standard fluorescence microscopy[485], or seven in more advanced techniques[486,487]. Moreover, genetic modifications used in live-cell imaging to express fluorescent protein reporters can be costly, time-consuming, and may induce poorly understood side effects, such as protein aggregation or disruption of ion transport dynamics.

While advanced techniques such as total internal reflection[488], light-sheet[489], two-photon excitation[490] and super-resolution[491] fluorescence microscopy address some of these challenges – offering improved resolution, greater imaging depth and reduced phototoxicity – they require specialised equipment, complex sample preparation and increased costs. Moreover, the fundamental limitations of fluorescence labelling described above remain.

Other analysis techniques

In addition to brightfield and fluorescence microscopy, a range of other analytical techniques are also widely used in tissue analysis, depending on the research context. The most representative methods are briefly outlined below.

Hematoxylin and Eosin (H&E) staining. H&E staining is a routine method in histopathology for assessing tissue morphology[492]. By differentially staining nuclei and cytoplasm, it enables evaluation of overall architecture and cellular organisation. In

neural organoids, H&E staining has been used to assess processes such as tissue morphogenesis[493], protein aggregation in models of neurodegeneration[494], viral infections[495] and tumour formation[425]. However, this technique is destructive, limited to endpoint analysis and lacks chemical specificity beyond general histological features.

Immunohistochemistry (IHC). IHC is an antibody-based method for tissue imaging, closely related to IF and FM. Depending on the readout, it can use chromogenic substrates (as in classical histopathology) or fluorescent labels (as in immunofluorescence), enabling highly specific detection of proteins and molecular markers within tissue sections[1]. In organoid research, IHC is commonly applied to fixed sections to validate structural and molecular features, but it remains invasive, requires fixation and sectioning, and is thus limited to endpoint analysis.

Bulk and single-cell omics (e.g. transcriptomics, proteomics, metabolomics). Bulk omics represent a category of techniques based on high-throughput molecular profiling, where RNA, proteins or metabolites are quantified across entire samples[6]. More recently, these methods have been extended to the single-cell level, probing molecular environments within individual cells[496]. These approaches provide rich biochemical profiles of global molecular states and responses, offering insights into pathways and cellular composition at the population level. In the context of neural organoids, an integrated single-cell transcriptomic cell atlas of human neural organoids has been recently developed, comprising data from more than 1.7 million cells[497]. Despite their strengths, omics approaches are inherently destructive, lack spatial resolution, and cannot directly image chemical morphology within tissues.

Spatial omics (e.g. spatial transcriptomics). Spatial omics extend bulk profiling by preserving positional information, for instance using spatially resolved capture or molecular tagging strategies[4,5]. In this way, molecular maps are reconstructed across tissue sections that capture aspects of both chemical and structural organisation, and link molecular states to spatial context. In neural organoids, spatial transcriptomics has been used to study mechanisms underlying processes such as morphogenesis[392], cell lineage[394] and patterning[498]. Nonetheless, spatial omics methods remain limited in spatial resolution compared to microscopy, and are destructive, costly and technically demanding.

Others. Other techniques, including electron microscopy (for ultrastructural detail)[499] and mass spectrometry imaging (for spatial metabolomics)[500], have also been applied to tissues, including neural organoids[501,502]. Each of these methods provides unique insights into structural or molecular features, but they are generally invasive, technically demanding, and limited in throughput, restricting their routine use.

Overall, while these techniques provide valuable insights, they are often destructive, require complex and costly preparation, and offer limited spatial or chemical resolution. Hence, in practice, brightfield and fluorescence microscopy remain the most widely used methods for general structural and molecular characterisation in cell biology and organoid research.

Toward Raman imaging in neural organoids

While brightfield and fluorescence microscopy provide valuable spatial and morphological insights, BF reveals little about molecular and functional states, and FM is invasive and requires labelling. These limitations highlight the need for non-invasive, label-free imaging methods that can probe the composition of intact 3D organoid systems *in situ*. In this context, RS offers a promising alternative, enabling biochemical imaging without fluorescent labels, stains or extensive sample preparation (Table 2.1).

In neurodevelopment, RS has been applied to profile neural cells at various stages of differentiation. Yet, most studies to date have focused on 2D cultures of individual cell types with RS measurements acquired at a single-cell level – e.g. distinguishing neural stem cells from neurons by identifying distinct cell-specific spectral signatures[503–507]. The application of RS to complex 3D cultured tissues containing multiple cell types, such as neural organoids, remains challenging due to their high biochemical heterogeneity, which hinders signal unmixing and interpretation. Recently, Bruno et al. explored the use of supervised ML to classify RS spectra from cortical organoids at four predefined maturation time points (weeks 6, 12, 16 and 20)[508]. However, their approach focused on single-point prediction and did not explore the 3D architecture, cellular interactions and functional complexity intrinsic to neurodevelopment. So far, the use of RS for spatially resolved analysis of neural organoid structure and maturation remains unexplored.

LaLone et al. recently reported a technique based on RS for 3D biochemical imaging and drug distribution analysis in liver organoids[90]. The proposed methodology relied on the acquisition of 'tissue phantoms' – i.e. calibrated reference spectra from biomolecules of interest, which were used to guide the estimation of chemical concentration maps across the scanned area using ordinary least squares with non-negativity constraint[192]. This enabled the detection and visualisation of key biomolecules and drugs *in situ* in whole liver organoids. Nonetheless, the approach has two limitations. First, it assumes that target biomolecules are known, thereby restricting potential applications and undermining the label-free capabilities of RS. Second, the performance and robustness of ordinary least squares for chemical estimation can be compromised by complex mixtures, experimental noise and artefacts, and variations introduced during data calibration and preprocessing,

leading to inaccurate qualitative and quantitative analysis.

Further advances in software, data processing and chemometrics are essential to establishing Raman spectroscopy as an analytical tool for non-invasive, label-free imaging and analysis in organoid research.

Table 2.1: Comparison of commonly used techniques in cellular and tissue imaging.

	Invasive	Label-free	Sample prep.	Sample	Lateral resolution[a]	z-sectioning	Limitations
Brightfield microscopy	No	Yes	Minimal	Live/Fixed	∼200–300 nm	Yes	Low contrast; no molecular specificity; depth limits; scattering sensitivity
Fluorescence microscopy[b]	Yes	No	Medium/ Extensive	Live/Fixed	∼200–300 nm	Yes	Requires labels (invasive); sample preparation; depth limits; limited multiplexing; often sectioning; photobleaching/phototoxicity
H&E staining	Yes	No	Extensive	Endpoint	∼200–300 nm	No	Destructive; 2D sections only; limited chemical specificity
Bulk/Single-cell omics	Yes	N/A	Extensive	Endpoint	N/A	N/A	Destructive; loss of spatial context (bulk) or tissue architecture (single-cell); costly; sample preparation
Spatial omics	Yes	N/A	Extensive	Endpoint	∼10–55 µm (capture) ∼200–300 nm (FISH/ISS)	N/A	Destructive; sectioning; costly; resolution-throughput trade-offs; sample preparation
Raman spectroscopy	No	Yes	Minimal	Live/Fixed	∼200–300 nm	Yes	Weak signal; autofluorescence; long acquisitions; depth limits; data processing and analysis

Abbreviations: H&E = hematoxylin and eosin; 2D = two-dimensional; N/A = not applicable; FISH = fluorescence *in situ* hybridization; ISS = *in situ* sequencing.

[a]Reported resolutions are approximate diffraction-limited maxima under typical high-NA confocal imaging conditions ($\lambda \approx 488$–$532\,\mathrm{nm}$). Practical resolution may be lower depending on sample scattering, aberrations or specific system configurations.

[b]Here, fluorescence microscopy refers to conventional diffraction-limited techniques (e.g. widefield, confocal). Super-resolution modalities are not included.

Chapter 3

RamanSPy: An open-source Python package for integrative Raman spectroscopy data analysis

This chapter is based on the following peer-reviewed manuscript:

Georgiev, D., Pedersen, S. V., Xie, R., Fernández-Galiana, Á., Stevens, M. M., & Barahona, M. (2024). RamanSPy: An open-source Python package for integrative Raman spectroscopy data analysis. *Analytical Chemistry*, 96(21), 8492-8500.

Published by American Chemical Society under the CC BY 4.0 license.
© 2024 The Authors.

Author contributions. I conceived and led the project, and developed the software, analysis pipelines and supporting tools, including the accompanying documentation website. I performed all experiments and data analyses, making use of previously published datasets. My co-authors – Dr Simon Vilms Pedersen, Dr Ruoxiao Xie, Dr Álvaro Fernández-Galiana, Prof. Dame Molly Stevens and Prof. Mauricio Barahona – contributed to supervision, conceptualisation, experimental design and interpretation of results. I wrote the initial draft of the manuscript and revised it with input from all co-authors.

3.1 Introduction

As discussed in Section 2.1.4, RS data analysis remains fragmented, with workflows scattered across proprietary environments, in-house scripts and ad hoc pipelines. The lack of integrated software tools, interoperability and standardised formats has led to method-

ological inconsistencies, reduced reproducibility and slowed the adoption of emerging computational approaches in RS.

These limitations have not only presented broader challenges for the field but also directly impacted the present work and the efforts to advance organoid imaging with RS. They hindered the construction of preprocessing pipelines capable of handling the abundant artefacts in spectra from 3D organoid samples, complicated the integration of imaging data across experiments, and constrained the development of new methods for more accurate chemical analysis.

In response to these challenges, this chapter presents RamanSPy – a modular, open-source framework for integrated 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 novel AI technologies into the RS field. First, RamanSPy serves as a platform for general-purpose RS analytics supporting the RS data life cycle by providing a suite of ready-to-use modules for data loading, preprocessing, analysis and visualisation. By design, these functionalities are not tied to any specific technology or data type, thereby allowing integrative and transferable cross-platform analyses. Second, RamanSPy addresses challenges in data preprocessing by facilitating the compilation of reproducible pipelines to streamline and automate preprocessing protocols. Third, RamanSPy helps bridge the gap between RS data and state-of-the-art AI technologies within the extensive ML ecosystem in Python. This is further complemented by direct access to relevant datasets, preprocessing protocols and performance metrics.

3.2 RamanSPy – a platform for general Raman spectroscopy analytics

3.2.1 Core features of RamanSPy

RamanSPy is based on a modular, object-oriented programming (OOP) infrastructure comprising a comprehensive collection of built-in tools for RS data analysis, which streamlines the analysis life cycle and reduces computational barriers to standard analyses (Figure 3.1). The framework adopts a scalable array-based data representation that 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 defines the interface between

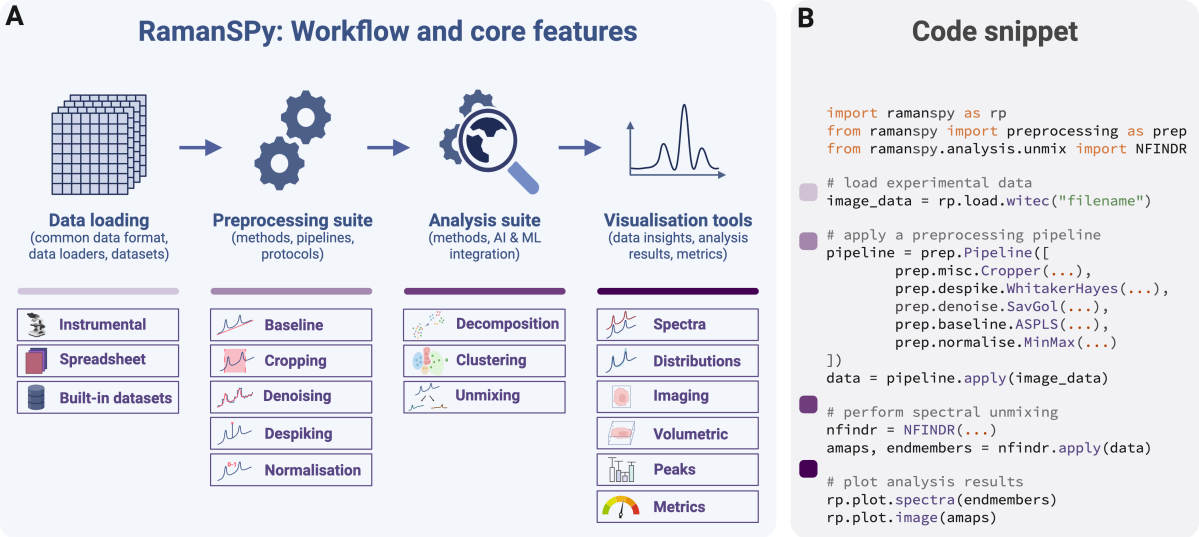


Figure 3.1: General Raman spectroscopy workflow and core features of RamanSPy. A, RamanSPy supports the RS data analysis life cycle via a modular, loosely coupled architecture. RS data is parsed to a common data representation format, which is interfaced with preprocessing, analysis and visualisation tools within RamanSPy. The core features of RamanSPy include a comprehensive library of standardised, simple-to-use procedures for data loading, preprocessing, analysis and visualisation. These modules are flexible and allow for the incorporation of new techniques and in-house methods. For complete information about the modules available in RamanSPy, refer to the documentation at https://ramanspy.readthedocs.io. **B,** An example workflow use case in RamanSPy: Raman data is loaded, preprocessed and analysed in a few lines of code. Created in BioRender.com. Reproduced from Georgiev et al.[509], published by American Chemical Society, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

RS data management and manipulation within RamanSPy, facilitating the integrative analysis of data across setups and vendors when appropriate (e.g. measurements from patient samples acquired by different laboratories), independent of instrumental origin and acquisition modality.

On top of that, RamanSPy provides an extensive toolbox for preprocessing, analysis and visualisation. The preprocessing suite includes techniques for spectral cropping, cosmic spike removal, denoising, baseline correction, 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. On the one hand, RamanSPy enables the compilation of different workflows by using built-in tools as "building blocks" to streamline repetitive tasks and routine analyses. On the other hand, the package also

retains the flexibility to develop and integrate custom methods as new "blocks" within the package for more advanced applications.

An extended list of the features implemented in RamanSPy is provided in Table B.1 in Appendix B.1.

3.2.2 Raman cell imaging via hyperspectral unmixing using RamanSPy

The core features of RamanSPy are showcased through a real research application. Specifically, using RamanSPy, the spectral unmixing analysis by Kallepitis et al.[74] was replicated, where volumetric Raman imaging scans of human leukaemia monocytic (THP-1) cells were collected and analysed using methods from chemometrics to investigate their composition in a label-free manner. For this demonstration, the first scan (scan *'001'*) was used, which was exported as MATLAB files from the *WITec Project FIVE* software and loaded into RamanSPy using built-in data loaders (Figure 3.2A).

After loading, data were preprocessed using a protocol consisting of: (1) spectral cropping to the 700–1800cm^{-1} region; (2) cosmic rays removal with the algorithm proposed by Whitaker and Hayes[159], with a kernel size $m = 3$ and z-value threshold $\tau = 8$; (3) denoising with a Savitzky-Golay filter[161] of polynomial order 3 and kernel size 7; (4) baseline correction with asymmetric least squares[165] with $\lambda = 10^6$; and (5) global Min-Max normalisation to the interval $[0, 1]$. Using the visualisation tools in the package, data quality was inspected (Figure 3.2B) and initial exploratory analysis was performed by examining intensity profiles across prominent wavenumber bands (Figure 3.2C).

After preprocessing, spectral unmixing was applied in RamanSPy using (i) N-FINDR[234] for endmember detection (number of endmembers set to 5) and (ii) FCLS[236] for component quantification. This process was exploited to demix signal contributions from different cellular components and study their morphological organisation within the THP-1 cell. Following the peak assignment and characterisation in the original paper, endmember components related to cytoplasm (band 1008 cm^{-1}), nucleic acid (band 789 cm^{-1}), lipids (bands 1066, 1134, 1303, 1443, and 1747 cm^{-1}), and the background (Figure 3.2E) were distinguished. Finally, abundance reconstructions based on the selected four endmembers were produced with RamanSPy, which were examined on a single-layer level and across the entire volume to localise cellular organelles within the cell (Figures 3.2D and 3.2F, respectively).

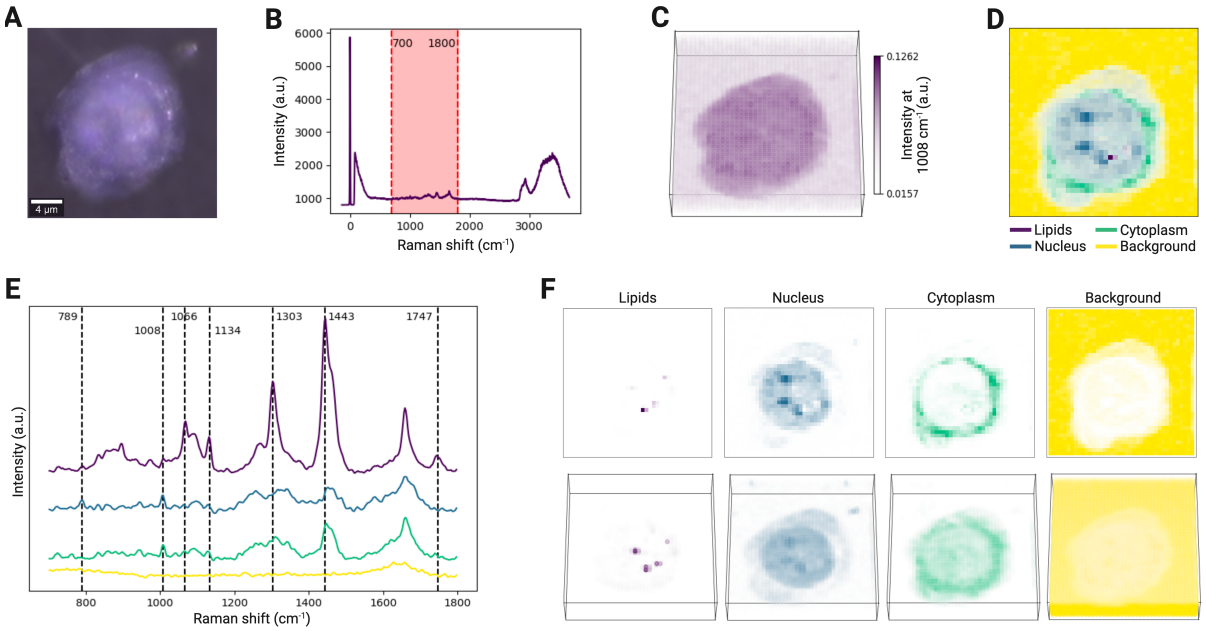


Figure 3.2: Morphological analysis of a THP-1 cell via spectral unmixing with RamanSPy. **A,** Brightfield image of a THP-1 cell. The same cell was also volumetrically imaged by RS. Brightfield image and raw RS data are from a previous work by Kallepitis et al.[74]. **B,** An example spectrum from the raw volumetric Raman data (taken from the centre region of the layer in **D**). The fingerprint region (700–1800 cm^{-1}), shaded in red, was used for the analysis. **C,** Exploratory analysis: a volumetric slice across the 1008 cm^{-1} band (characteristic of proteins) after preprocessing. **D–F,** Spectral unmixing analysis reveals the distribution of components within the cell: lipids (violet), nucleus (blue), cytoplasm (green), and background (yellow). **D,** Merged reconstruction of the sixth depth layer (10 in total) of the THP-1 cell determined via spectral unmixing. **E,** Four endmembers derived with N-FINDR[234] characterised via peak assignment. **F,** Fractional abundance maps calculated with FCLS[236] – *top,* single slice ($z = 6$); *bottom*, entire volume. Peak assignments follow Kallepitis et al.[74]. Assembled in BioRender.com. Reproduced from Georgiev et al.[509], published by American Chemical Society, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

3.3 RamanSPy enables automated pipelining of spectral preprocessing protocols

As highlighted in Section 2.1.4, experimental RS data are susceptible to various sources of non-specific signals and artefacts (e.g. cosmic rays, autofluorescence background, noise, variability in instrumentation), which can severely affect downstream analyses. Preprocessing is therefore a critical step in any spectroscopic analysis workflow[41,154]. Yet, due to a lack of standardisation and frameworks that streamline the preprocessing of RS data[40], researchers tend to utilise variable preprocessing protocols, often dispersed across different software systems, thus affecting reproducibility and validation[510,511].

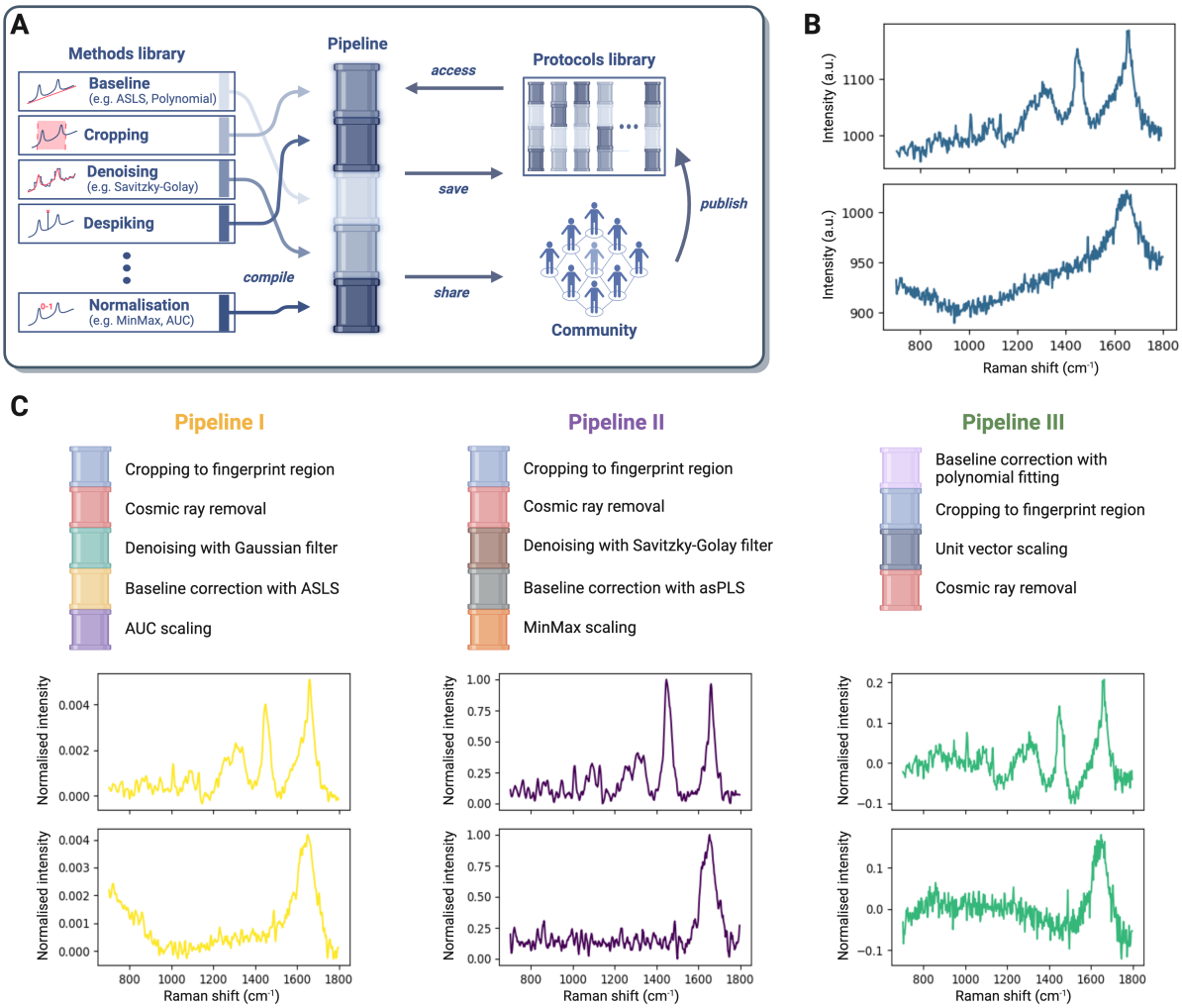


Figure 3.3: Spectral preprocessing pipelining in RamanSPy. A, RamanSPy automates the construction, customisation and execution of multilayered preprocessing procedures via pipelining. Users can assemble built-in and in-house methods into complete preprocessing pipelines, which are fully compatible with data integrated within RamanSPy and can be saved, reused and shared. RamanSPy also provides access to a library of preprocessing pipelines. **B,** Two raw spectra from the THP-1 data by Kallepitis et al.[74] are used to compare the effect of different preprocessing pipelines. **C,** The results of three preprocessing pipelines built within RamanSPy, demonstrating the need for transparency and reproducibility. Note on preprocessing methods: fingerprint region is 700–1800 cm^{-1}; ASLS – asymmetric least squares[165]; asPLS – adaptive smoothness penalised least squares[512]; AUC – area under the curve; cosmic rays removed with the algorithm by Whitaker and Hayes[159]. Created/assembled in BioRender.com. Reproduced from Georgiev et al.[509], published by American Chemical Society, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

3.3.1 Preprocessing pipelining in RamanSPy

To facilitate the creation of reproducible protocols, RamanSPy incorporates a pipelining infrastructure, which systematises the process of creating, customising and executing preprocessing pipelines (Figure 3.3A). Users can use a specialised class, which defines a generic, multilayered preprocessing procedure, to assemble pipelines from selected built-in

preprocessing modules or other in-house methods. To reduce overhead, the 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 repository of preprocessing protocols. As a seed to this repository, RamanSPy provides a library of assembled preprocessing protocols (custom predefined, or adapted from the literature[513]), which users can access and exploit.

3.3.2 Pipelining illustration

To illustrate the pipelining functionalities of RamanSPy, three representative preprocessing protocols were constructed. Within RamanSPy, this entails compiling selected methods in the desired order of execution and applying them out of the box to data loaded into the platform. The three pipelines were used to preprocess Raman cell imaging data from Kallepitis et al.[74] (Figure 3.3B). Note how the three pipelines yield substantially different results (Figure 3.3C), reinforcing the importance of transparency and reproducibility in the selection of preprocessing protocols. These three pipelines were added to the repository of protocols in RamanSPy.

3.3.3 Computational efficiency

As an indication of the computational efficiency of RamanSPy for such preprocessing tasks, the computational efficiency of a representative preprocessing protocol (*Pipeline I*) was profiled on a standard laptop (MacBook Air laptop with Apple M2 chip, 8-core CPU, 10-core GPU, and 16-core Neural Engine). As shown in Figure 3.3C, *Pipeline I* consisted of: (1) spectral cropping to the 700–1800cm^{-1} region; (2) cosmic rays removal with the algorithm proposed by Whitaker and Hayes[159]; (3) denoising with a Gaussian filter; (4) baseline correction with asymmetric least squares[165]; and (5) pixel-wise normalisation based on setting the area under the curve to 1.

The pipeline was applied to synthetic data of three sizes: 1000, 10000 and 100000 spectra, where each spectrum was generated by sampling 1500 values from a uniform distribution over $[0, 1)$. The experiments showed that it takes ∼0.25 seconds to preprocess 1000 spectra with this protocol; ∼2.5 seconds for 10000 spectra; and ∼20–25 seconds for 100000 spectra.

3.4 RamanSPy facilitates AI integration and validation of next-generation Raman data analytics

To help accelerate the adoption of AI technologies for RS analysis, RamanSPy is endowed with a permeable architecture that streamlines the interface between RS data and the burgeoning 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. Below, two examples of RamanSPy's capabilities for ML integration and validation are discussed.

3.4.1 AI integration: Deep learning denoising

First, RamanSPy allows the seamless integration of standard Python AI/ML methods (e.g. from *scikit-learn*[514], *PyTorch*[515] and *tensorflow*[516]) as tools for RS analysis (Figure 3.4A).

As an illustration of how custom methods can be integrated into analysis pipelines within RamanSPy, the package was used to construct a deep-learning denoiser based on the one-dimensional ResUNet model – a fully convolutional UNet neural network with residual connections – developed by Horgan et al.[209]. To do this, the pre-trained neural network (trained on spectra from MDA-MB-231 breast cancer cells) was simply wrapped within RamanSPy 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, the analysis by Horgan et al.[209] was replicated, showing in Figure 3.4B that the application of this deep-learning denoiser to the low SNR test set provided by the authors consistently outperforms the commonly used Savitzky-Golay filter[161]. This is quantified by various metrics also coded within RamanSPy (e.g. mean squared error (MSE), spectral angle distance (SAD)[517] and spectral information divergence (SID)[518]), which measured the performance of each denoising method by comparing denoised signals against the provided high SNR data, which acted as a reference.

Importantly, applying this pipeline to new data involves changing only the data source. Taking advantage of this transferability, the denoiser was tested on unseen volumetric Raman data from another cell line (THP-1, data by Kallepitis et al.[74]), degraded with added Gaussian noise. The denoising analysis was performed on the middle depth layer of the volumetric scan (fifth layer out of 10). This layer consisted of a 40×40 image scan, i.e. 1600 spectra. To be consistent with the paper by Horgan et al.[209], the same preprocessing protocol was applied. Namely, the *WITec Project FIVE* software was

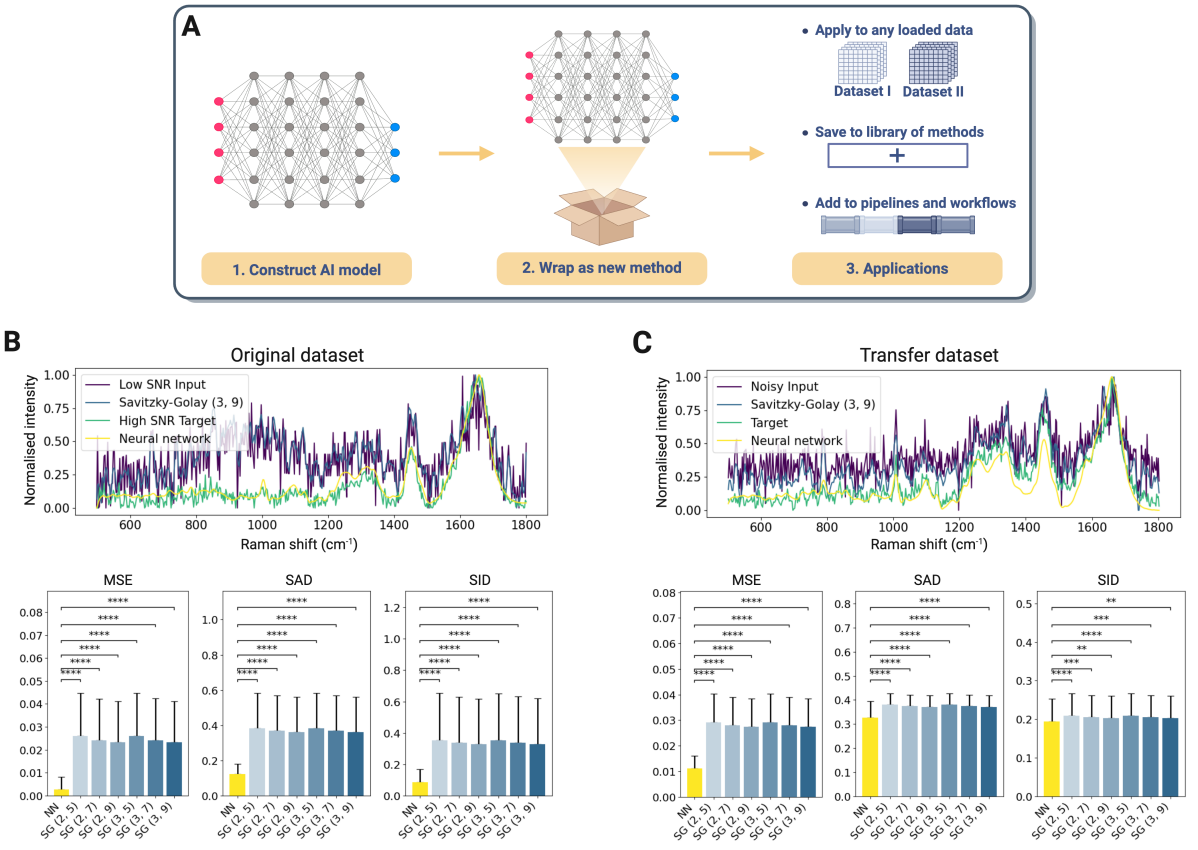


Figure 3.4: RamanSPy interfaces with AI/ML Python frameworks to create new methods for RS analysis. A, RamanSPy allows users to incorporate AI/ML models seamlessly into pipelines created within the platform. **B,** A pre-trained 1D ResUNet deep-learning denoiser from a previous work by Horgan et al.[209] is integrated as a preprocessing module within RamanSPy to investigate its performance against the Savitzky-Golay (SG) filter[161]. *Top*, Denoising of a spectrum from cell data by Horgan et al.[209], 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 provided by Horgan et al.[209] ($n = 12694$) show that the deep-learning denoiser outperforms six SG filters across three performance metrics (MSE, SAD, SID). **C,** Analysis on unseen data by Kallepitis et al.[74] ($n = 1600$). The input (purple) corresponds to data contaminated with added noise and the target (green) corresponds to the original data. In this case, the deep-learning denoiser only shows an improvement for MSE. Error bars represent one standard deviation of the sample mean. Statistical significance measured with a two-sided Wilcoxon signed-rank test with adjustment for multiple comparisons based on Benjamini-Hochberg correction[519] (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$). Created/assembled in BioRender.com. Reproduced from Georgiev et al.[509], published by American Chemical Society, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

used to crop the data to the region 500–1800cm^{-1}, followed by baseline correction using the "*shape*" method with $\alpha = 500$. To assess the performance of the deep-learning denoiser, low SNR spectra were created by adding Gaussian noise to the original spectra by Kallepitis et al.[74]. Each spectrum was Min-Max-normalised to the range 0–1, and Gaussian noise with a standard deviation $\sigma = 0.15$ was added. This resulted in spectra of similar noise levels to those in the data by Horgan et al.[209]. These noisy samples were used as the input to the model, and the uncontaminated data were taken as ground-truth

targets. The performance of the neural network denoiser was benchmarked against six Savitzky-Golay filters[161]. To make all models comparable, and to correct for potential artefacts of how the model was trained originally in the paper by Horgan et al.[209], all denoising metrics were computed after Min-Max-normalisation of the denoised outputs of each denoiser to the range 0–1 again.

For these data, Figure 3.4C 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, also 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. It should be noted that this denoiser was used solely to exemplify the integration and testing capabilities of RamanSPy, and was not included as a built-in denoiser in the package.

3.4.2 AI interoperability: Deep learning bacteria classification

Second, 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 3.5A).

As an example application, AI-based bacteria identification was performed using Raman measurements from 30 bacterial and yeast isolates from a previous study by Ho et al.[100]. After loading and visualising the spectra with RamanSPy, shown in Figure 3.5B, the data were interfaced with the *lazypredict* Python package[520], which directly benchmarked 28 ML classification models in Python (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 fine-tuning dataset (100 spectra per isolate) and then tested on the unseen test set of the same size. The benchmarking analysis in Figure 3.5C finds logistic regression as the best-performing model, achieving a classification accuracy of 79.63% on the species-level classification task (Figure 3.5D) and 94.63% for antibiotic treatment classification (Figure 3.5E).

*Here, "data shift" refers to a change in the statistical distribution between training and test datasets, not to a spectral (Raman) shift.

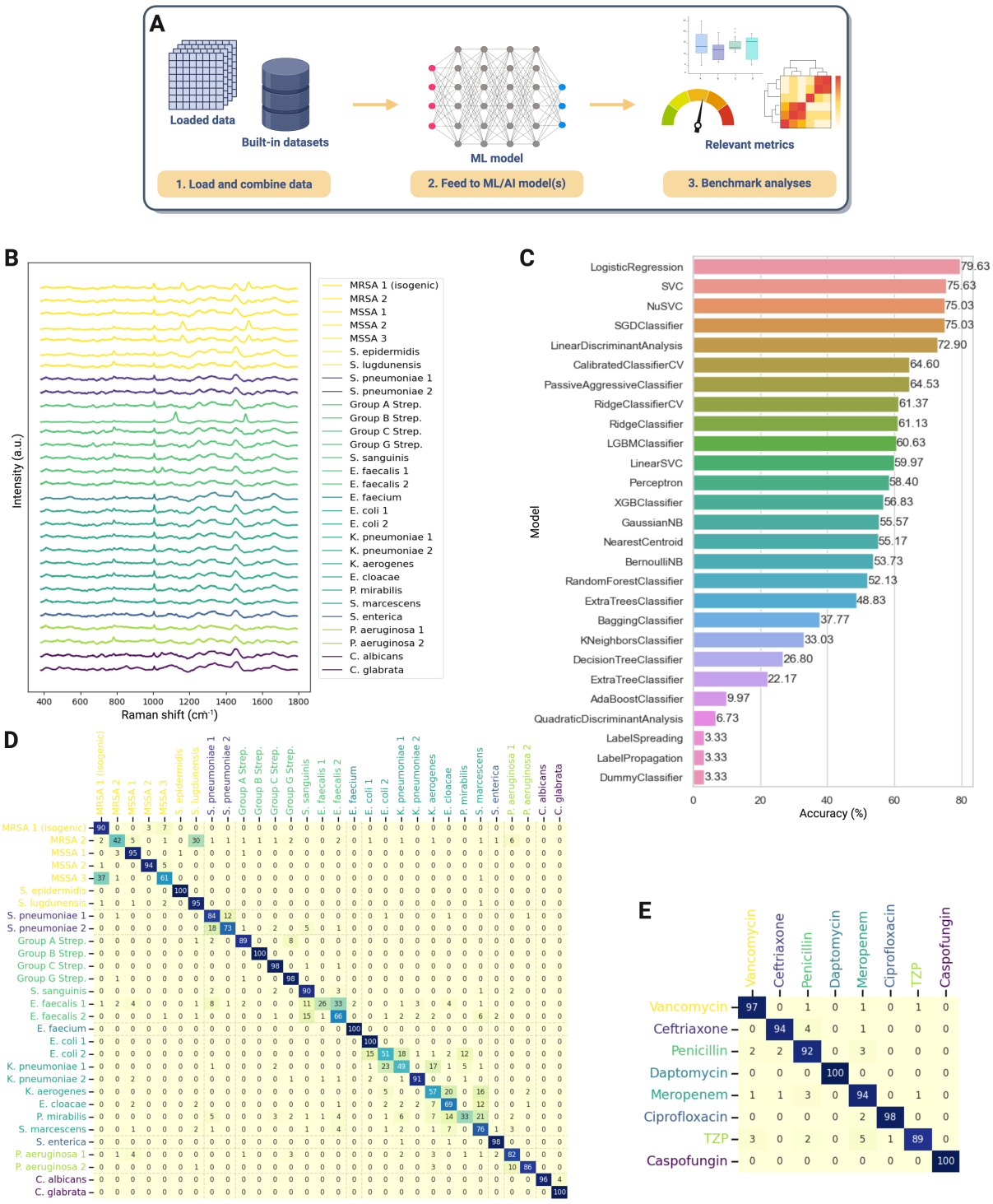


Figure 3.5: RamanSPy as a suite for algorithmic development. **A,** Data representations in RamanSPy are compatible with the Python AI/ML ecosystem, allowing data flow from RamanSPy to frameworks such as *scikit-learn*[514], *PyTorch*[515] and *tensorflow*[516]. RamanSPy is also equipped with standard 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 by Ho et al.[100]. **B,** Mean Raman spectra of each bacterial species in the dataset used for training. Spectra are MinMmax normalised to the range 0–1 for visualisation purposes. **C,** Benchmarking results of 28 ML models. The best accuracy was achieved using 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. Created/assembled in BioRender.com. Reproduced from Georgiev et al.[509], published by American Chemical Society, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

3.4.3 Dataset suite

To further assist the process of testing and evaluating new computational approaches, RamanSPy provides access to a library of curated datasets for different tasks (e.g. classification, denoising, Raman imaging). This library aims 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[195]. The dataset library in RamanSPy already includes data loaders for Raman data from bacterial species[100], cell lines[74,209], COVID-19 samples[521,522], adenine samples[523], wheat lines[524], and minerals[525], and will continue to be expanded. Recognising the potential benefits that synthetic and surrogate datasets can provide for the generation of controlled ground truths for algorithmic validation, the synthetic Raman data generator described in Chapter 4 was also integrated in RamanSPy.

3.5 Conclusion

This chapter introduced RamanSPy – a computational framework for integrative RS data analysis, aimed at overcoming some of the limitations of currently available commercial software tools in terms of accessibility, flexibility, reproducibility and advanced data analysis capabilities. RamanSPy offers a comprehensive collection of tools for spectroscopic analysis designed to systematise the RS data analysis life cycle, reducing typical overheads of analysis workflows and improving methodological standardisation. The package also lays the foundations of a common repository of standardised methods, protocols and datasets, which users can readily access and exploit. Furthermore, RamanSPy is fully compatible with frameworks for data science and ML in Python, thereby facilitating the adoption and validation of advanced AI technologies for next-generation RS analysis.

RamanSPy is fully open-source and disseminated under a permissive license that allows for unrestricted use, adaptation and extension, including commercial purposes. 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.

Data availability

All data used in this chapter are previously published open-access data that have been deposited by the respective authors online. Instructions on how to access, download

and load the datasets 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.

More information about how to install and use RamanSPy is provided in Appendix B.2.

Chapter 4

Hyperspectral unmixing for Raman spectroscopy via physics-constrained autoencoders

This chapter is based on the following peer-reviewed manuscript:

Georgiev, D., Fernández-Galiana, Á., Vilms Pedersen, S., Papadopoulos, G., Xie, R., Stevens, M. M., & Barahona, M. (2024). Hyperspectral unmixing for Raman spectroscopy via physics-constrained autoencoders. *Proceedings of the National Academy of Sciences*, 121(45), e2407439121.

Published by PNAS under the CC BY 4.0 license. © 2024 The Authors.

Author contributions. I conceived and led the project, and developed the software, analysis pipelines and supporting tools. I performed all experiments and data analyses, using benchmark datasets created by me (synthetic mixtures), Dr Álvaro Fernández-Galiana and Georgios Papadopoulos (sugar mixtures), as well as using previously published data (cell imaging). Co-authors – Dr Álvaro Fernández-Galiana, Dr Simon Vilms Pedersen, Dr Ruoxiao Xie, Prof. Dame Molly Stevens and Prof. Mauricio Barahona – contributed to supervision, conceptualisation, experimental design and interpretation of results. I wrote the initial draft of the manuscript and revised it with input from all co-authors.

4.1 Introduction

Building on the software infrastructure introduced in Chapter 3, this chapter turns to the next key challenge in applying RS to neural organoids, which lies in the inherent complex-

ity of analysing and interpreting Raman data from heterogeneous biological systems such as cells, tissues and organoids. As discussed in Section 2.1.5, conventional chemometric techniques often fail to extract accurate qualitative and quantitative chemical information from the noisy and convoluted measurements typical of such samples. While deep learning has emerged as a promising alternative for RS chemometrics, most existing approaches remain limited to constrained, supervised tasks and are thus not well-suited to biological contexts where ground truth is rarely available.

Section 2.1.5 presented hyperspectral unmixing as a powerful framework for unsupervised chemical analysis, aimed at decomposing mixed spectra by identifying the individual components present (endmember identification) and quantifying their proportions (abundance estimation) (Figure 4.1A). Yet, as detailed in the section, standard unmixing algorithms, such as N-FINDR, VCA, NNLS and FCLS, also face significant limitations. Specifically, these methods are restricted to linear mixing; lack robustness to data artefacts abundant in RS data (e.g. dark noise, baseline variations, cosmic spikes); rely on additional assumptions (e.g. endmembers present as 'pure pixels' in the data); and are computationally demanding for large datasets (e.g. imaging and volumetric Raman raster scans).

To address these issues, this chapter presents a framework for Raman unmixing based on autoencoder (AE) neural networks. Leveraging the proposed approach, a suite of AE architectures for hyperspectral unmixing is developed and systematically evaluated against conventional unmixing methods, using synthetic and experimental benchmark datasets generated as part of this chapter. The results demonstrate that unmixing AEs offer improved accuracy, robustness and efficiency compared to standard unmixing methods. Furthermore, the applicability of autoencoders to complex biological settings is explored, achieving enhanced volumetric Raman imaging of a monocytic cell.

4.2 Raman unmixing with autoencoder neural networks

4.2.1 Unmixing autoencoders

Section 2.1.5 introduced how the problem of hyperspectral unmixing can be formulated as an autoencoder learning problem. Again, autoencoders are a family of (deep) neural network models consisting of two sub-networks (*encoder* and *decoder*) connected sequentially[246]. The encoder $\mathcal{E} : \mathbb{R}^b \to \mathbb{R}^m$, where $m \ll b$, transforms input data $\mathbf{x}$ to a lower-dimensional latent space representation $\mathbf{z} = \mathcal{E}(\mathbf{x})$, which the decoder $\mathcal{D} : \mathbb{R}^m \to \mathbb{R}^b$ uses to produce reconstructions $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z})$ of the original input. AE models are typically

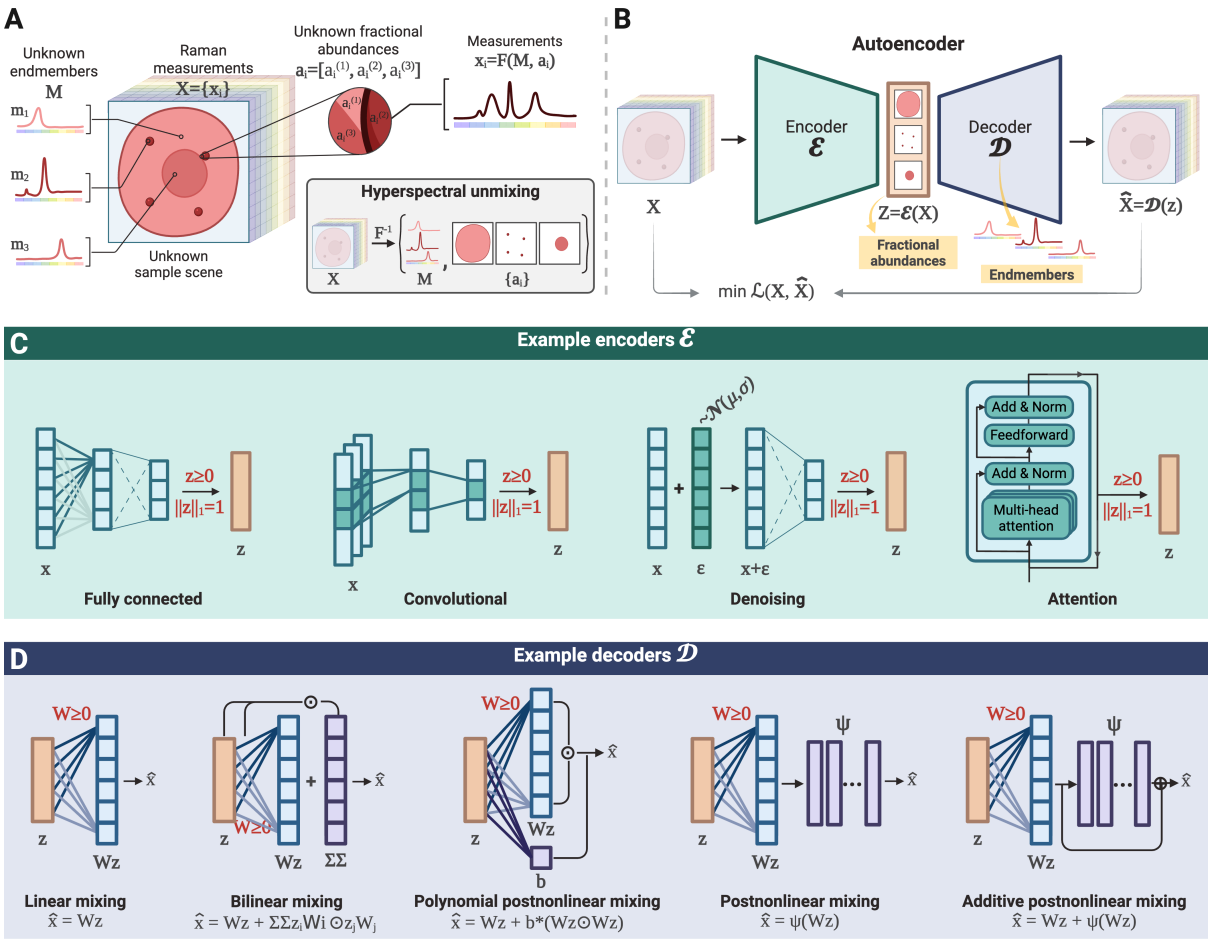


Figure 4.1: Hyperspectral unmixing for Raman spectroscopy using autoencoder neural networks. a, Illustration of hyperspectral unmixing. A set of Raman measurements $X = \{\mathbf{x}_i\}$, where each $\mathbf{x}_i$ is a spectrum, are collected from a sample of interest. The spectra $\mathbf{x}_i$ are viewed as mixtures of a set of Raman spectral signatures M (*endmembers*), related to the chemical species present, and their respective *abundances* $\mathbf{a}_i$. Hyperspectral unmixing aims to infer the set of unknown underlying endmembers M and abundances $\mathbf{a}_i$ from the set of measurements X, thereby obtaining qualitative and quantitative compositional information about the sample. **b,** Hyperspectral unmixing as a self-supervised autoencoder learning problem: the decoder $\mathcal{D}$ learns endmember signatures during training and the encoder $\mathcal{E}$ learns to derive the corresponding abundances. **c,** Encoders can accommodate different concepts from representation learning, such as convolutional layers and attention, to improve feature extraction and provide more accurate and robust unmixing. **d,** Decoders can be structured to model different linear and non-linear mixing models. Physics-inspired constraints are indicated in red in **c** and **d**. Created in BioRender.com. Reproduced from Georgiev et al.[526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

trained in a *self-supervised* manner by minimising a loss function $\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}})$ that measures the discrepancy between the input $\mathbf{x}$ and the reconstruction $\hat{\mathbf{x}}$ (e.g. the mean squared error (MSE)); hence, no ground-truth target values are provided during training. As the training of the model proceeds, the encoder progressively learns a latent representation that captures the most salient features of the input data, whereas the decoder learns how to recover the data back from the latent representation.

This dual functionality can be harnessed to design autoencoders for hyperspectral unmixing: the latent representations $\mathbf{z} = \mathcal{E}(\mathbf{x})$ can be interpreted as abundances (with respect to the input spectrum $\mathbf{x}$), and the decoder $\mathcal{D}(\cdot)$ acts as a mixing function on these representations by encoding endmember signatures and other interactions. Hence, AE models learn to perform an 'unmixing' where the decoder identifies endmember signatures and the encoder quantifies the abundances of these learned endmembers in the input spectrum (Figure 4.1B). To guide the learning, physical constraints are incorporated into the AE architecture to reflect the nature of hyperspectral unmixing, e.g. non-negativity of endmembers and abundances, and sum-to-one abundances.

Compared to standard methods for unmixing, autoencoders are more effective at capturing complex, noisy and non-linear relationships in the data, and do not rely on assumptions such as the presence of 'pure pixels'. An extended discussion about these and other advantages of unmixing AEs is provided in Table 4.1. Moreover, AEs offer a more flexible and versatile framework for unmixing, as illustrated by the wide design space of encoder and decoder architectures.

On the one hand, the learning of physical and biochemical features in the encoder can be enhanced by adopting strategies from representation learning, such as convolutional layers to capture spectral and/or spatial correlations among neighbouring bands and/or pixels[527–529], or attention mechanisms to model long-range dependencies[530] (Figure 4.1C). In addition, sparsity, part-based learning and denoising objectives can be enforced during training to enhance explainability and robustness[531–535].

On the other hand, the design of the decoder allows for flexible modelling of input data, specifically to account for various mixture models, e.g. linear, bilinear and post-nonlinear (Figure 4.1D)[243,247,248]. This is akin to introducing an inductive prior with respect to the mixture model directly via the AE architecture. Furthermore, the decoder can be pre-initialised with a set of endmembers (e.g. an endmember library or signatures derived using methods such as VCA), or readily adapted to non-blind unmixing by fixing certain parameters to predefined endmember signatures.

4.2.2 Autoencoder design

To assess the effectiveness of AEs as a framework for RS unmixing, a collection of AE models was developed and evaluated, each defined by a specific encoder and decoder.

Table 4.1: Comparison of autoencoders and standard unmixing techniques. Reproduced from Georgiev et al.[526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

	Standard unmixing methods (e.g. VCA+NNLS)	**Autoencoders**
Assumptions	Standard methods rely on strict assumptions - e.g. endmembers present in the data as 'pure pixels'; fixed mixture model; number of endmembers is known.	Autoencoders can facilitate data-driven unmixing without many strong assumptions about the data and task.
Mixture model	A specific model is assumed *a priori*. Usually, a linear mixing only, as alternatives can become too complex and computationally expensive.	Autoencoders can be designed to capture various mixture models, including non-linear mixtures. This can be enforced by appropriate architectural constraints or learned in a data-driven manner.
Input modality & spatial correlations	Spectra are processed individually. If imaging or volumetric data are provided, these are typically unfolded and used pixel-wise, thereby discarding any spatial correlations.	Autoencoders can be extended to utilise the spatial correlations available in imaging or volumetric data by incorporating (2D, 3D, and/or 4D) convolutional layers.
Unmixing process	Typically, two separate algorithms are applied.	Autoencoders simultaneously extract both endmembers and fractional abundances by default. They can also be adjusted for non-blind unmixing by fixing a pre-defined endmember matrix in the decoder.
Robustness to artefacts	Performance is usually extremely dependent on the quality of the data, as well as on preprocessing, which can vary greatly from application to application.	Autoencoders can incorporate feature selection blocks and training loss to promote invariance to scaling, baselines, and noise, making them more robust to artefacts and outliers.
Variable spectral axis	Observations are assumed to share a common spectral axis, and the spectral axis is discarded. This impedes the integration of data across experimental setups unless they share the same spectral axis.	The axis can be integrated as an input to the model (e.g. as a positional encoding), allowing the direct integration of diverse data sources.
Scalability	Methods can become computationally prohibitive for larger datasets. More advanced options are intractable in real-world applications involving imaging or volumetric Raman scans.	With modern software and hardware, deep neural networks are extremely scalable and parallelisable.
Number of endmembers	Typically, fixed *a priori*.	Normally fixed *a priori*, too, but can potentially be learned by introducing sparsity and information-theoretic criteria in the latent space, for instance.
Extensibility	Most conventional techniques are based on specialised optimisation-based algorithms tailored for the particular unmixing task (e.g. mixture model, scene type, number of endmembers), which makes their extension challenging.	Easy extension and adaptation. One can readily expand unmixing autoencoders into more advanced variants, as well as integrate models into established AI and ML pipelines.
Downstream applications	Conventional techniques are generally constrained within hyperspectral unmixing, with algorithms only outputting the derived unmixing results.	Autoencoders can be used as feature extractors in a variety of modelling and predictive downstream tasks, such as classification, clustering and anomaly detection. They can also leverage transfer learning techniques to improve performance in situations with limited labelled data by utilising unmixing autoencoders for self-supervised pre-training.
Data requirements	Standard methods typically require the presence of endmember signatures as 'pure pixels' in the dataset, and their effectiveness depends on the availability of these signatures.	AEs do not necessarily require 'pure pixel' signatures but may need larger datasets with representative mixtures for effective learning.
Interpretability	Standard methods are based on explicit mathematical models, making the results generally interpretable and traceable back to the original input features.	The internal representations learned by AEs are typically challenging to interpret and can represent complex, abstract relationships that are challenging to reverse-engineer or directly correlate with physical properties.
Hyperparameters	The primary hyperparameter in standard methods is the number of endmember signatures to consider, which simplifies the tuning process.	AEs model training can involve numerous hyperparameters (e.g. number of layers, size of layers, learning rate, training epochs, batch size), requiring extensive tuning and optimisation for effective performance.

Encoder design

Four types of encoders were considered, encompassing a variety of architectures, from traditional dense layers to contemporary convolutional and attention mechanisms: 1) an encoder consisting of fully connected layers (*Dense*); 2) an encoder with a 1D convolutional feature extractor block, followed by a fully connected part (*Convolutional*); 3) a transformer-based encoder using multi-head attention (*Transformer*); and 4) a transformer-based encoder with a 1D convolutional feature extractor (*Convolutional Transformer*).

Dense AE. This autoencoder employs an encoder comprising 2 fully connected (or dense) layers. The first layer projects spectra of dimension b to hidden features of dimension 128 (Leaky ReLU activation with a slope of 0.02), which the second layer further projects to latent representations of dimension n (n is the number of endmembers to extract). In the *Deep Dense AE* model used in the analysis of the THP-1 cell, the number of hidden layers is increased to five, comprising 512, 256, 128, 64 and 32 neurons, respectively, before the final layer of size n.

Convolutional AE. This model extends the *Dense AE* by adding a convolutional block before the dense layers. The convolutional block consists of two layers of 1D convolutions connected in parallel, each comprising 16 filters of size 3 and 16 filters of size 5 (ReLU activation; input padded with zeroes). The outputs from these two layers are concatenated and merged (channel-wise) via a 2-dimensional dense layer to produce representations of dimension b, which are then fed to the *Dense* encoder described above.

Transformer AE. In this transformer-based encoder, input spectra are first projected to features of size 32 through a fully connected layer, and then fed to a transformer encoder layer comprising a multi-head attention block with 2 attention heads of size 32[536], followed by two fully connected layers expanding the features to size 64 (ReLU activation) and condensing back to 32 (no activation). Layer normalisation[537] and dropout (10%)[538] are applied after the multi-head attention block and the fully connected layers. The output of the transformer block is then channelled into the last fully connected layer of size n.

Convolutional Transformer AE. In this model, the *Transformer AE* architecture is extended with the same convolutional block used in the *Convolutional AE*, here added before the transformer-based encoder block.

Decoder design

The two types of decoders investigated were: 1) a decoder designed for linear unmixing, and 2) a decoder designed for bilinear unmixing. The linear unmixing decoder architecture consists of a single fully connected layer using the identity activation function without bias (see Equation 2.5). The bilinear Fan decoder shares the same architecture as the linear decoder but also incorporates the additional bilinear interaction terms (see Equation 2.6).

Physics-inspired constraints

Abundance constraints were applied through the choice of latent space activation functions. To enforce both the abundance non-negativity constraint (ANC) and abundance sum-to-one constraint (ASC), a softmax activation function was applied in the final layer of each encoder. When only ANC was used, the activation function was changed to a 'softly-rectified' hyperbolic tangent function given by $\frac{1}{\gamma}log(1 + e^{\gamma * tanh(x)})$, with $\gamma = 10$, customly designed to ensure abundances are between 0 and 1 but do not necessarily add up to one. To ensure the non-negativity of endmembers, the weight matrix of the linear layer in the decoder was constrained by clipping negative values to zero during training.

4.2.3 Model training and evaluation

The autoencoders were trained in a self-supervised manner by minimising a loss based on the spectral angle divergence (SAD)[517] that measures the cosine similarity between input and reconstructed spectra (see Equation 4.2). Training was performed using the Adam optimiser with a learning rate of 0.001[539].

In the cases where ground-truth information is available (synthetic data and sugar mixtures), predictive accuracy was quantified according to two measures: (i) the mean squared error (MSE) between ground-truth and predicted abundances $\boldsymbol{\alpha}$ and $\widehat{\boldsymbol{\alpha}}$:

$$\text{MSE}(\boldsymbol{\alpha}, \widehat{\boldsymbol{\alpha}}) = \frac{1}{n}||\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}||_2, \tag{4.1}$$

with n representing the number of endmembers extracted, and (ii) the SAD between ground-truth and predicted endmembers $\mathbf{m}_i$ and $\widehat{\mathbf{m}}_i$:

$$0 \leq \text{SAD}(\mathbf{m}_i, \widehat{\mathbf{m}}_i) = \arccos\left(\frac{\mathbf{m}_i \cdot \widehat{\mathbf{m}}_i}{\|\mathbf{m}_i\|_2 \, \|\widehat{\mathbf{m}}_i\|_2}\right) \leq \pi. \tag{4.2}$$

Each evaluation was performed after first matching the derived and ground-truth endmembers (and corresponding abundances) via the Hungarian algorithm with SAD as the objective to minimise. When the number of extracted endmembers n was higher than the number of ground-truth endmembers n_{true}, only the n_{true} best-matched estimated endmembers (i.e. those with the lowest SAD scores) and their corresponding abundances were used to compute the final performance metrics.

Note that each dataset was analysed independently by fitting separate models from scratch.

4.2.4 Baseline methods for comparison

The performance of the developed AE models was compared to conventional unmixing approaches: N-FINDR and VCA for endmember extraction, followed by NNLS or FCLS for abundance estimation. PCA was also evaluated, which, despite not being designed for unmixing, is commonly used in applications. PCA results were omitted from the main body of the chapter as they exhibit substantially lower performance (see Appendix C for full results).

4.2.5 Implementation

Analyses were performed in Python, using *tensorflow*[516] for autoencoder model development and training, and RamanSPy (Chapter 3) for unmixing with conventional methods, data loading and management, preprocessing and plotting.

4.3 Benchmarking unmixing autoencoders on synthetic Raman mixtures

The performance of the AE architectures was first benchmarked on synthetic datasets created for validation.

4.3.1 Generating synthetic Raman mixtures

To this end, a custom data generator was developed that produces synthetic Raman mixtures with different characteristics (e.g. number and type of endmembers, abundance profiles, mixture model, data artefacts) with full knowledge of the underlying endmembers

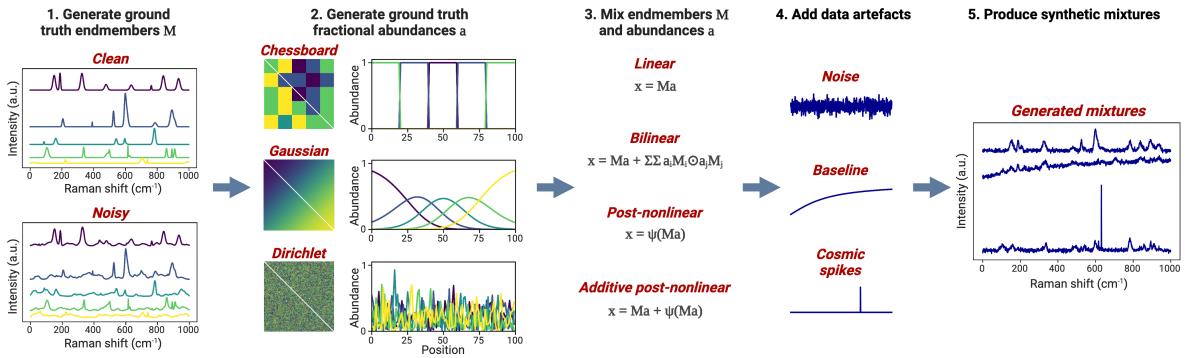


Figure 4.2: Generation of synthetic Raman mixtures. Schematic diagram of the synthetic data generator developed to create mixtures with known ground truth for benchmarking unmixing methods. Created in BioRender.com. Reproduced from Georgiev et al. [526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

and abundances (Figure 4.2).

Generating endmembers

For each synthetic dataset, n endmembers were first generated spanning b spectral bands. For the scope of this work, $n = 5$ and $b = 1000$. Each endmember $\mathbf{m}_i \in \mathbb{R}_+^b$ was created by a superposition of a set of $n_{\text{peaks},i}$ Gaussian peaks of different amplitude, width and location, randomly sampled as follows. The number of peaks was sampled from a discrete uniform distribution $n_{\text{peaks},i} \sim \mathcal{U}(5, 9)$. Each peak p was described by $p = h_p \sigma_p \sqrt{2\pi}\, \mathcal{N}(b_p, \sigma_p)$, where $\mathcal{N}(\cdot)$ represents a Gaussian distribution. The height of the peak was defined as $h_p = h_1 \cdot h_2$, where $h_1 = 1 + 5\, h_\beta$ with $h_\beta \sim \text{Beta}(1, 3)$ and $h_2 \sim \mathcal{U}(0.1, 1)$. The centre of the peak was sampled from $b_p \sim \mathcal{U}(10, b - 10)$, and the width of the peak was defined as $\sigma_p = w_p \sigma$, with $\sigma \sim \mathcal{U}(0.1, 1)$.

Two types of endmembers were created: *clean* and *noisy*. For the former, peaks with $w_p = 1$ were produced. For the latter, *clean* endmembers were augmented by adding $n^{\text{small}}_{\text{peaks},i} \sim \mathcal{U}(50, 99)$ smaller peaks sampled with $h_1 = 1/3$ and $w_p = 2$, thus making *noisy* endmembers better resemble experimental Raman signatures.

Generating abundances

For visualisation purposes, abundance profiles were presented in the form of two-dimensional scenes comprising $H \times W$ pixels, where each pixel represents abundance vector $\boldsymbol{\alpha} \in \mathbb{R}_+^n$. Here, $H = W = 100$, resulting in 10000 spectra per scene/dataset. In the simplest scene (*Chessboard*), the scene was divided into 20×20 square patches, each containing a single

randomly assigned endmember (i.e. all 400 pixels in each patch are the same one-hot vector). The second scene (*Gaussian*) consisted of n Gaussian functions equally spaced along the diagonal of the scene. After each pixel was normalised to comply with the ASC, abundance profiles were obtained with varying levels of overlap of components. The last abundance scene (*Dirichlet*) corresponded to a highly mixed scene, where each pixel was individually sampled from an n-dimensional Dirichlet distribution, producing a random mixture of all endmembers. Note that the abundance profile of each pixel in all three scenes complied with both ANC and ASC.

Mixing model

Having generated a set of endmembers and an underlying abundance scene, mixed data measurements $\mathbf{x} \in \mathbb{R}^b$ were created based on a selected mixing model. In this study, linear mixtures (Equation 2.2) and bilinear mixtures based on the Fan model (Equation 2.3) were considered.

Adding data artefacts

Finally, data artefacts (noise, baseline, cosmic spikes) were optionally added to create more realistic synthetic Raman spectra. Here, Gaussian noise $\boldsymbol{\epsilon} \in \mathbb{R}^b$ was added to each spectrum, with independent and identically distributed components $\epsilon_i \sim \mathcal{N}(0, \sigma_N)$. Further, a baseline signal $\mathbf{B} = h_B \arctan(\pi[1 : b]/b) \in \mathbb{R}^b$ was added to each spectrum with probability p_B. Finally, with probability p_S, a cosmic spike of intensity $S \sim h_S U(0.75, 1.25)$ was added to each spectrum at a band $b_S \sim \mathcal{U}\{2, b-2\}$. For all experiments, the generation parameters were set to: $\sigma_N = 0.1$, $p_B = 0.25$ $h_B = 2$, $p_S = 0.1$, $h_S = 5$.

4.3.2 Synthetic datasets for validation

Using the data generator, 11 types of synthetic datasets of variable complexity were produced, based on four mixture scenarios over three abundance scenes. In order of complexity, the four mixture scenarios were: 1) a linear mixture with *clean* endmembers and no data artefacts (*ideal*); 2) a linear mixture with *clean* endmembers, but contaminated with artefacts representing dark noise, baseline variations and cosmic spikes (*+artefacts*); 3) a linear mixture with *noisy* endmembers (i.e. containing additional smaller noise peaks) and artefacts (*+realistic*); and 4) a bilinear mixture based on the Fan model[230] with *noisy* endmembers and artefacts (*+bilinear*). For each of the four mixture scenarios, three dataset variants were generated (two for the *+bilinear* scenario since no bilinear

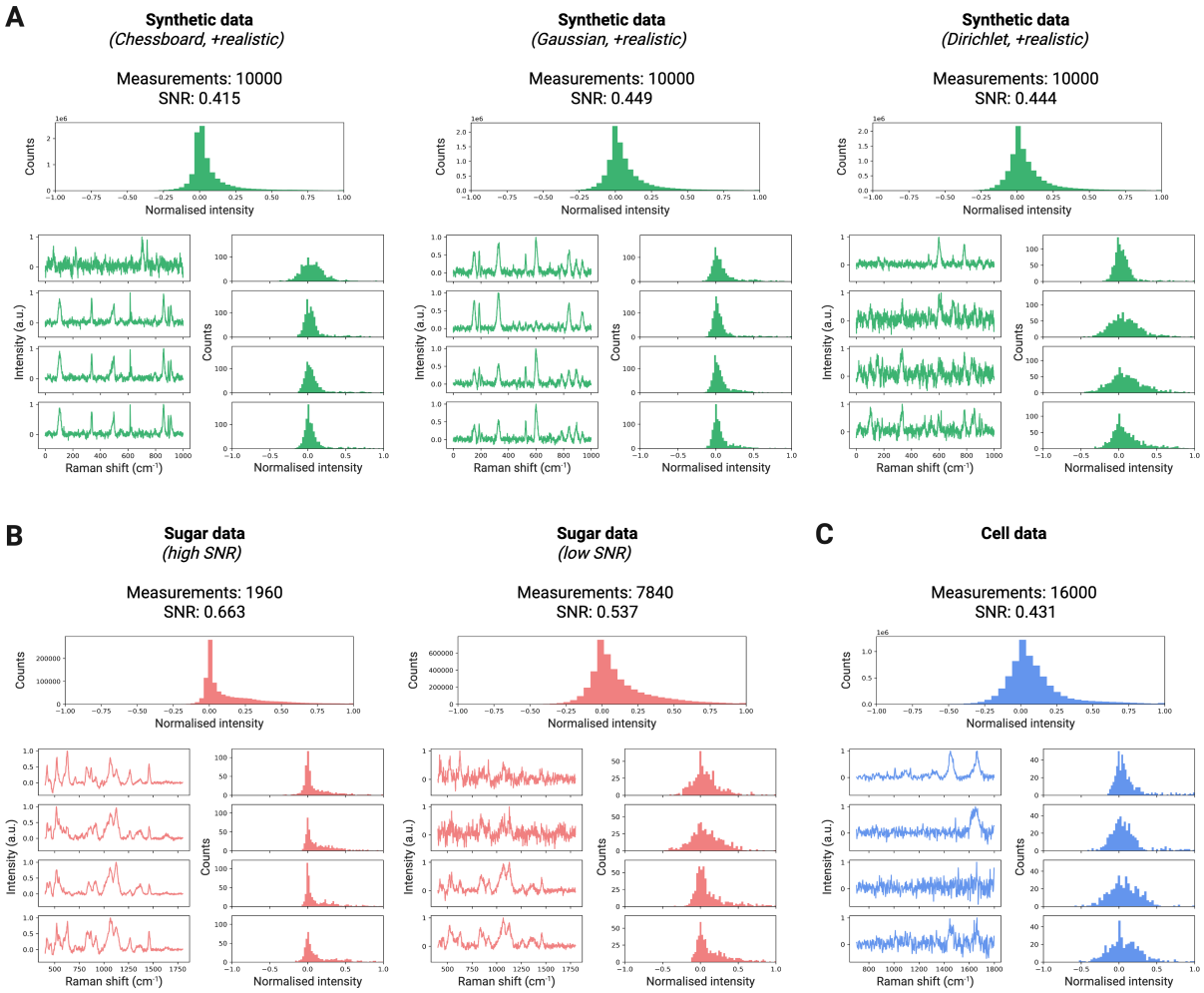


Figure 4.3: Key measures and statistics of benchmark datasets used for validation of unmixing autoencoders. Summary information for datasets used for validation – **A,** three synthetic datasets; **B,** the two experimental datasets from sugar solutions with reference spectra; and **C,** the volumetric cell data from Kallepitis et al.[74]. Each panel provides the number of measurements present in the dataset and the mean SNR of the spectra, a histogram (50 bins) of all intensity values in the dataset, and 4 randomly selected spectra from the dataset alongside their individual intensity histograms (50 bins). Assembled in BioRender.com. Reproduced from Georgiev et al.[526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

interactions are present in the *Chessboard* scene) based on custom 100×100 abundance scenes. This produced 10k spectra per dataset, organised into two-dimensional scenes for visualisation purposes. In increasing level of mixing, the scenes considered were: 1) a scene comprising well-separated patches, each containing a single species (*Chessboard* scene); 2) a semi-mixed scene given by a Gaussian mixture of species (*Gaussian* scene); and 3) a highly-mixed scene where each pixel represents a random sample of species drawn from a Dirichlet distribution (*Dirichlet* scene). Therefore, the synthetic datasets covered increasingly complex scenarios, from the *ideal Chessboard* dataset that is trivial for conventional methods, to noisier, more complex mixtures containing different types of artefacts.

Representativeness of synthetic datasets

To assess the representativeness of the synthetic datasets generated by the data generator, analyses were performed comparing statistical and similarity metrics between the synthetic datasets and the experimental data (sugar solutions, cell line) used for validation in this chapter. Before analysis, data were baseline corrected using adaptive smoothness parameter penalised least squares[512] and normalised by dividing each spectrum by its maximum value. Experimental data were also cropped to the fingerprint region 700–1800 cm^{-1} before preprocessing.

Figures 4.3 presents summary statistics for each of the three data types, including the number of measurements, mean SNR, dataset intensity histograms, and randomly selected spectra with their corresponding histograms. The panels illustrate that synthetic datasets (all *+realistic* scenarios) display comparable properties to real data. Dataset similarity was further assessed using the Wasserstein-1 distance between spectral histograms and multidimensional scaling (MDS) analysis. Figure 4.4 shows that generated synthetic datasets, especially the *+artefacts* and *+realistic* scenarios, align well with real Raman measurements, as indicated by low spectral distances and overlapping MDS projections.

Recognising that synthetic data can serve as a valuable validation, particularly when ground truth data is unavailable or non-existent, the data generator developed here was incorporated into RamanSPy, enabling the generation of diverse synthetic mixtures with known ground-truth endmembers and abundances. Nevertheless, it is important to exercise caution when extrapolating findings from synthetic data alone. Although the synthetic data generation accounts for various sources of variability (e.g. dark noise, cosmic spikes, baseline variations), synthetic data inherently provide an idealised scenario. Still, synthetic datasets can provide a useful initial validation step during method development, which can then be further supported by subsequent experiments using real data, as done in this chapter.

4.3.3 Benchmark results

The generated synthetic datasets enabled the evaluation of unmixing performance across different methods and data settings.

Each experiment was performed on 5 datasets and 5 model initialisations using different random seeds, resulting in $5 \times 5 = 25$ replicates per evaluation, or 1650 experiments in total: 6 models (2 conventional, 4 AEs) $\times$ 11 dataset variants $\times$ 25 replicates. Random seeds were kept the same across mixture scenarios to allow direct comparison. The latent

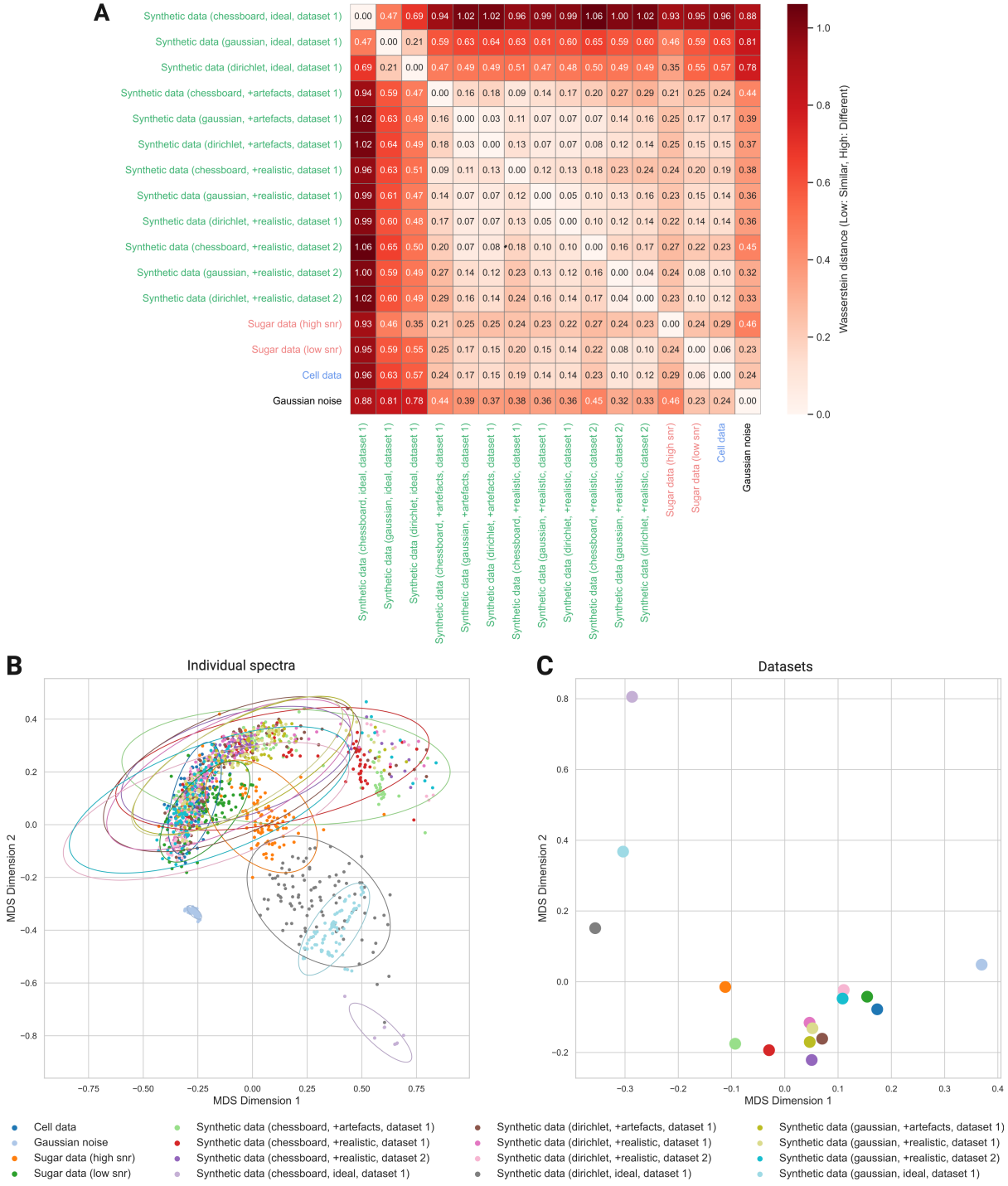


Figure 4.4: Similarity analysis of the benchmark datasets used for validation of unmixing autoencoders. A, Wasserstein-1 distance between the dataset-wide intensity histograms (50 bins) of the synthetic and experimental datasets used. 'Gaussian noise' represents a dataset that contains 10000 'spectra' each comprising 1000 values independently sampled from a normal distribution with mean 0 and standard deviation of 0.5, added for reference. **B–C,** Multidimensional scaling (MDS) analysis of the calculated Wasserstein distances between the histograms of: 100 randomly selected spectra from each considered dataset (**B**); and dataset-wide histograms (**C**). Ellipses in **B** represent 90% confidence regions for each group based on a multivariate Gaussian distribution model, showcasing the spread and overlap of the data points. Assembled in BioRender.com. Reproduced from Georgiev et al. [526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

dimensionality m of each AE model was set to 5 for the *ideal* mixture scenario, and 6 for the other mixture scenarios with data artefacts. Both ANC and ASC were enforced for all experiments on synthetic data. AE models were trained for 10 epochs.

To illustrate the evaluation process, Figure 4.5A shows a comparison of the six unmixing methods on an example dataset (*Chessboard* scene, *+artefacts* scenario). Using the ground truth, derived endmembers and abundances were evaluated both qualitatively and quantitatively. In the following, the results across all synthetic datasets are discussed.

Linear mixtures

First, the results on the nine dataset variants created through the linear mixture scenarios (1–3) are presented. These data comply with the linear mixing assumption of conventional methods and, for consistency, the AE models tested here were equipped with a decoder for linear unmixing. Figure 4.5B summarises the performance of the six models (two conventional and four AEs) across the nine dataset variants, with experiments performed over 5 distinct datasets and 5 model initialisations for each variant (refer to Table C.1 in Appendix C for numeric performance metrics). The discrepancy between ground truth and derived endmembers was measured using SAD, while the discrepancy between ground truth and derived abundances was measured using MSE, as discussed in Section 4.2.3.

Results show that the AE models consistently outperform the two conventional methods, providing more accurate endmembers and abundances across virtually all scenarios and abundance scenes. The AEs recover the performance of the conventional methods on the simple *ideal Chessboard* datasets, and the improvement in AE performance becomes increasingly prominent for mixture scenarios with higher levels of noise and data artefacts.

Non-linear mixtures

Next, the benchmark analysis on synthetic data generated using a non-linear mixture model (i.e. *+bilinear* scenario) is presented. The results are displayed in Figure 4.5C, where AEs were equipped with a decoder specific to the bilinear mixture model by merely adapting the decoder architecture. Once again, the results demonstrate that all four AE models provide a substantial improvement in unmixing accuracy compared to methods like N-FINDR+FCLS and VCA+FCLS for both endmember and abundance estimation.

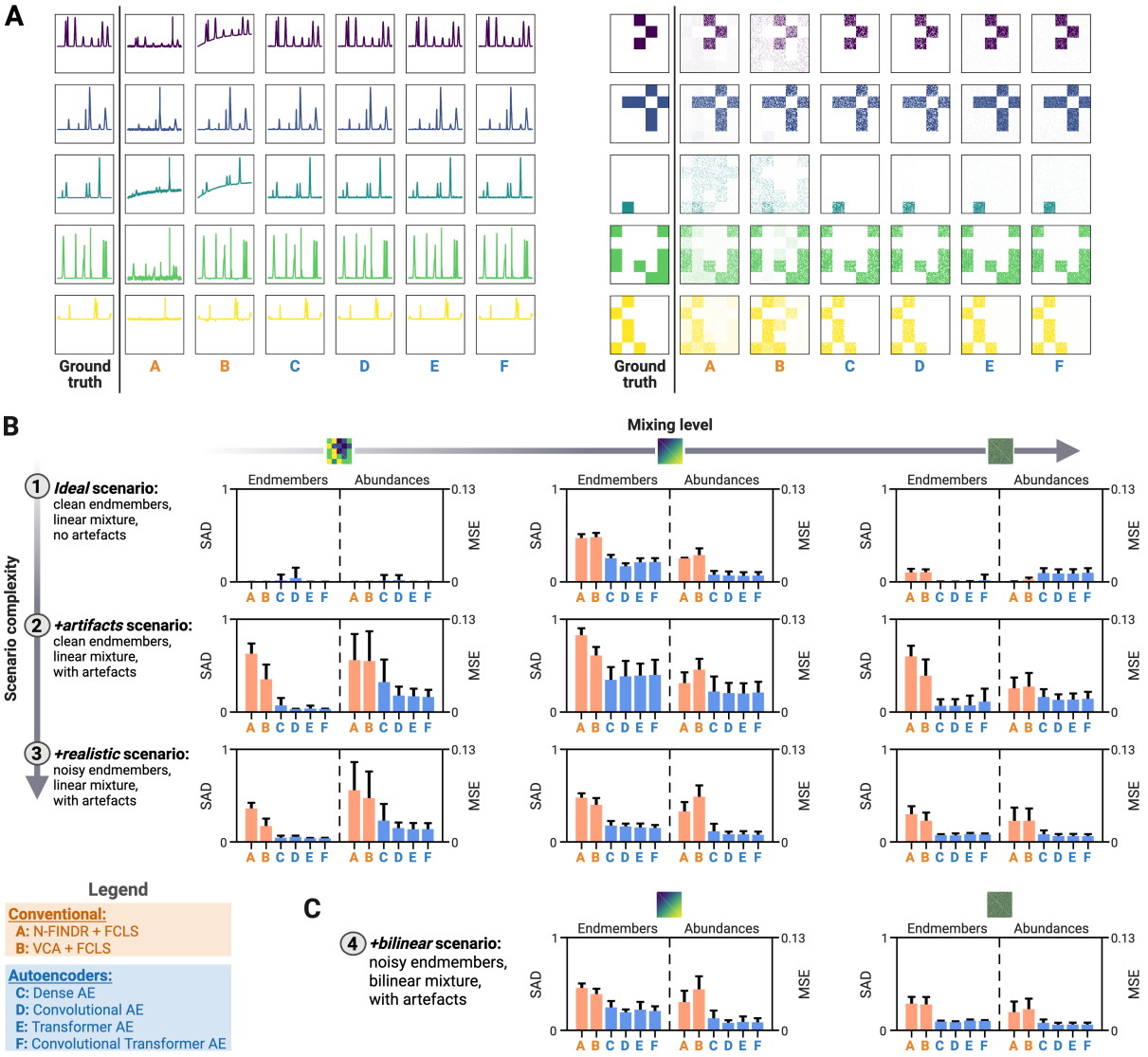


Figure 4.5: Benchmarking autoencoders on synthetic Raman mixtures. A, Representative unmixing results – endmembers (*left*) and abundances (*right*) – obtained by each of the six algorithms considered when applied to an example synthetic dataset (*Chessboard* scene, *+artefacts* scenario). **B–C** Summary of the unmixing performance of the six algorithms on 11 types of synthetic datasets, which span variable mixing levels and complexity. Linear mixtures are shown in **B**, and bilinear mixtures in **C**. Confidence intervals are given by one standard deviation around the sample mean (each bar represents $n = 25$ evaluations: 5 datasets with 5 model repetitions each). Assembled in BioRender.com. Reproduced from Georgiev et al.[526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

4.3.4 Computational efficiency

The computational complexity and scalability of unmixing methods can become a significant bottleneck in real-world applications, particularly for imaging and volumetric Raman scans, which can contain hundreds of thousands of spectra.

To examine this issue, the computational cost of the four developed AE models (linear

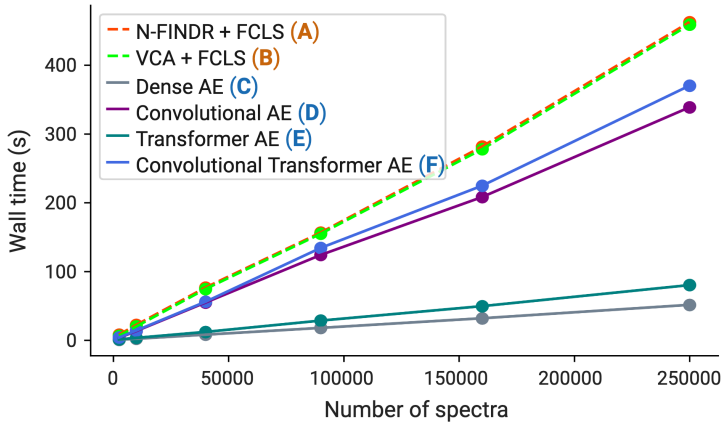


Figure 4.6: Computational efficiency analysis. The computational cost (measured as wall time) of autoencoders and conventional methods on synthetic datasets (*Chessboard* scene, *+artefacts* scenario) as the number of spectra is increased. Each dot represents the average across 3 evaluations (confidence intervals based on one standard deviation are small and not visible to the eye). AE models are equipped with decoders for linear unmixing. Labelled in BioRender.com. Reproduced from Georgiev et al. [526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

decoders) and the two conventional methods was profiled on synthetic datasets (*ideal* scenario, *Chessboard* scene) of increasing sizes, from 2500 to 250000 spectra. The number of endmembers to extract was set to $n = 5$ for all methods. For each experiment, 3 separate evaluations were performed, measuring the wall time of each method (including the full training time for autoencoders). All experiments were conducted on a MacBook Air laptop with an Apple M2 chip (8-core CPU, 10-core GPU, and 16-core Neural Engine). Only CPU computation was employed throughout to ensure a fair comparison with traditional methods, which, by design, do not utilise GPU acceleration.

Figure 4.6 shows that all four AE models are faster than standard methods such as N-FINDR+FCLS and VCA+FCLS, providing efficient unmixing even without utilising GPU acceleration and parallel processing, which can further enhance their performance.

4.4 Validation of unmixing autoencoders on experimental Raman data from sugar mixtures

To validate the unmixing performance of AEs on real experimental data, benchmark analyses were performed on data from a library of 240 sugar mixtures prepared in-house with four types of sugar (glucose, sucrose, fructose, maltose) at different concentrations (see Figure 4.7A).

4.4.1 Data collection

Preparation of sugar solutions

1 mol/L solutions of each type of sugar (sucrose, fructose, maltose, and glucose) were prepared by dissolving the appropriate weight of sugar into 40 mL of ultrapure distilled water (Invitrogen™ – UltraPure™ DNase/RNase-Free Distilled Water). The weights of sugars dissolved were 13.83 g for sucrose (Thermo Scientific Chemicals – Sucrose, 99%), 7.279 g for fructose (Thermo Scientific Chemicals – D-Fructose, 99%), 15.171 g for maltose (Thermo Scientific Chemicals – D-(+)-Maltose monohydrate, 95%) and 7.279 g for glucose (D-(+)-Glucose, AnalaR NORMAPUR® analytical reagent). All solutions were mixed and vortexed in standard 50 mL centrifuge tubes until no solute was visible.

Sugar mixtures were prepared in standard 96-well plates, with a volume of 375 µL per well. A full factorial experiment was performed comprising 4 volume levels for each sugar (0 µL, 30 µL, 75 µL and 120 µL), filled with distilled water where necessary. Discarding the mixtures exceeding the volume of the well and the one that contains no sugar, 240 distinct mixtures were prepared. In addition, 5 extra 'pure' solutions (i.e. 375 µL of water, sucrose, fructose, maltose, or glucose) were prepared and used to extract reference spectra for each chemical species. This resulted in a total of 245 wells distributed in three standard 96-well plates. Mixtures were stirred using standard 200 µL pipettes before spectral acquisition to ensure good mixing.

Credits: Sugar mixtures were prepared by Dr Álvaro Fernández-Galiana and Georgios Papadopoulos.

Raman measurements from sugar solutions

Spectra were acquired using a custom Raman microspectroscopy platform designed for high-throughput analysis, known as B-Raman. This platform is based on the Thorlabs Cerna® and features the BWTek BRM-785-0.55-100-0.22-SMA laser excitation source and the Ibsen EAGLE Raman-S spectrometer. The instrument was calibrated using an Argon wavelength calibration source (AR-2 – Ocean Insight) reference lamp before data collection. The excitation wavelength was 785 nm and the power incident to the samples was 36.3 mW. The Raman scattering was collected in reflection via a Leica N PLAN 10x/0.25 objective with 0.25 numerical aperture. The raw spectra were acquired over the spectral wavenumber range of 142–3684.8 cm^{-1}.

Spectra were measured from the centre (horizontal) of each well at a fixed depth that was established to provide the highest signal. To evaluate unmixing performance under

different signal-to-noise ratio (SNR) conditions, two sets of data were collected from each well. Specifically, high and low SNR measurements were acquired using integration times of 5 seconds and 0.5 seconds, respectively. Several measurements were collected from each well, resulting in a total of 240 solutions × 2 measurements × 4 repetitions = 1920 high-SNR measurements (1960 with reference spectra); and 240 solutions × 8 measurements × 4 repetitions = 7680 low-SNR measurements (7840 with reference spectra).

Credits: RS data from sugar mixtures were collected by Dr Álvaro Fernández-Galiana and Georgios Papadopoulos.

4.4.2 Preprocessing and analysis

Each sugar dataset was preprocessed using the following protocol: 1) cropping to the region 400–1800 cm^{-1}; 2) baseline correction with adaptive smoothness parameter penalised least squares (asPLS)[512] with smoothing parameter $\lambda = 10^5$, differential matrix of order 2, maximum iterations set to 100, and exit criteria with tolerance $t = 0.001$; 3) global vector normalisation, where each observation was divided by the highest magnitude observed in the data. Baseline removal is important to ensure models extract relevant features (i.e. characteristic peaks) as opposed to merely capturing the trend.

Next, hyperspectral unmixing was performed to estimate the content of each mixture, i.e. types of sugar and their concentrations. Ground-truth endmember signatures were estimated by taking the median (band-wise) of the reference spectra (40 in high SNR setup, and 160 in low SNR setup) collected from the 5 additional wells containing pure solutions (Figure 4.7B). Ground-truth abundances were determined by calculating the ratio of the components present based on the experimental concentrations used to prepare each mixture. As with the synthetic data above, the performance of the four AE models (linear decoders) was benchmarked against N-FINDR+FCLS and VCA+FCLS. These experimental datasets with ground truth were used to systematically evaluate unmixing algorithms under typical experimental artefacts, such as baseline shifts, environmental noise and cosmic spikes.

Hyperspectral unmixing was performed following similar training and evaluation protocols to those employed for the synthetic data, with the number of endmembers to extract set to $n = 5$. The number of epochs was increased to 15 for low SNR data and 50 for high SNR data, given the more limited number of spectra collected. Moreover, an additional MSE term was incorporated in the training loss $\mathcal{L}$ of autoencoders on the high SNR data:

$$\mathcal{L}(\mathbf{x}, \widehat{\mathbf{x}}) = \mathrm{SAD}(\mathbf{x}, \widehat{\mathbf{x}}) + \lambda\, \mathrm{MSE}(\mathbf{x}, \widehat{\mathbf{x}}), \tag{4.3}$$

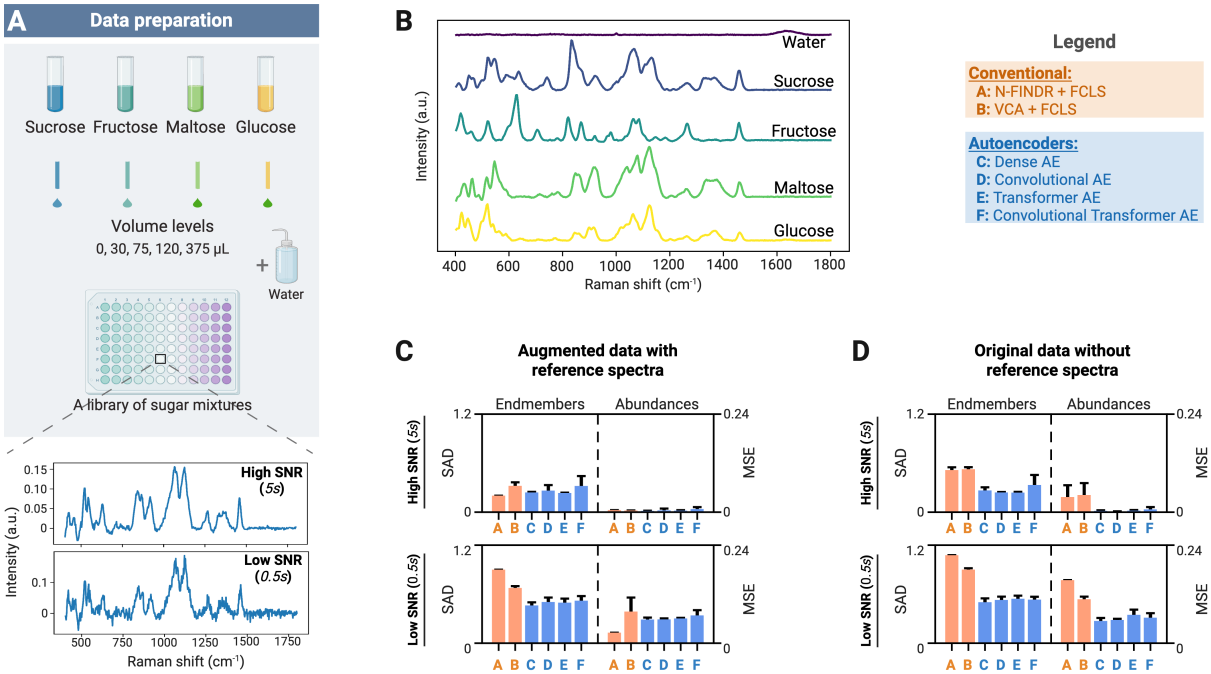


Figure 4.7: Experimental validation on Raman spectroscopy data from sugar mixtures. **A,** Schematic diagram of sugar mixture preparation. Two sets of Raman measurements are acquired from a library of 240 sugar mixtures – high and low signal-to-noise ratio (SNR) data collected using integration times of 5 s and 0.5 s, respectively. Raw RS data acquired by Dr Álvaro Fernández-Galiana and Georgios Papadopoulos. **B,** Ground-truth endmember signatures were estimated by taking the median of high SNR reference spectra measured from pure solutions containing a single sugar species. **C–D,** Summary of unmixing performance for: **C,** idealised scenario where spectra collected from pure solutions are included; and **D,** original data without the spectra from pure solutions (i.e. no 'pure pixels'). Confidence intervals are given as one standard deviation around the sample mean ($n = 5$). Created/assembled in BioRender.com. Reproduced from Georgiev et al. [526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

with $\lambda = 1000$. This term breaks the invariance to scale and leads to better abundance estimation given the weak water endmember (see Table C.2 in Appendix C for full results and numeric performance metrics). The standard SAD loss was used for experiments on low SNR. Each experiment was repeated for 5 model initialisations.

4.4.3 Benchmark results

Results on data with reference spectra

First, an idealised scenario was considered, purposefully devised to favour conventional methods, whereby endmembers were present in the data. To simulate this, the additional reference spectra measured from pure solutions were also included in the total data, serving as 'pure pixels'. As expected, when such 'pure pixels' are available, conventional methods (NFINDR+FCLS, VCA+FCLS) performed comparably to AEs on clean, high

SNR data (Figure 4.7C). Yet, AEs already provided improved performance in low SNR regimes.

Results on data without reference spectra

In many experimental applications, however, the underlying endmembers are not present in the data and cannot be separately obtained (e.g. target-agnostic applications or unknown species). To consider such cases, the analysis was repeated using the original datasets without including the additional reference spectra from pure solutions. The results in Figure 4.7D demonstrated that, in such situations, AEs substantially outperform conventional methods in both low and high SNR settings. Qualitative examination further supports these findings, with standard methods failing to extract accurate endmember signatures, particularly in low SNR data where outlier spectra with significant artefacts were often selected as endmembers (Figures 4.8 and 4.9).

Further performance evaluation data using an alternative endmember distance based on Pearson's correlation coefficient (PCC) is presented in Table C.3 in Appendix C. The additional results show an even more pronounced improvement in endmember estimation accuracy.

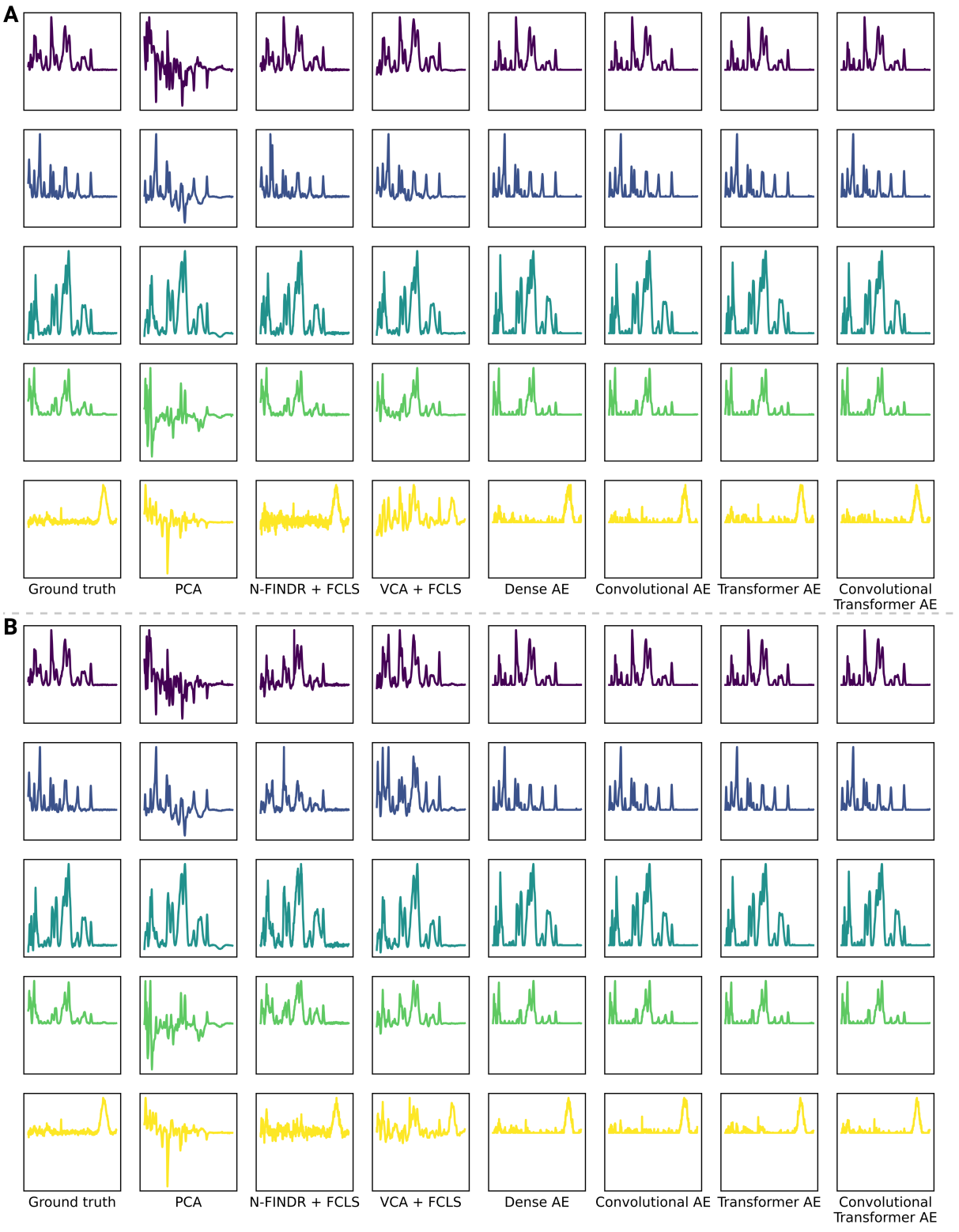


Figure 4.8: Endmember estimates on high SNR sugar data. Qualitative comparison of derived endmembers on the high SNR sugar datasets. **A,** Data with reference spectra. **B,** Data without reference spectra. Endmembers are scaled such that their maximum intensity is equal to 1 for visualisation purposes. x-axes represent the Raman shift region 400–1800 cm^{-1}, and y-axes, which are shared for each row, represent normalised intensity (a.u.). Raw RS data acquired by Dr Álvaro Fernández-Galiana and Georgios Papadopoulos. Assembled in BioRender.com. Reproduced from Georgiev et al. [526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.



Figure 4.9: Endmember estimates on low SNR sugar data. Qualitative comparison of derived endmembers on the low SNR sugar datasets. **A,** Data with reference spectra. **B,** Data without reference spectra. Endmembers are scaled such that their maximum intensity is equal to 1 for visualisation purposes. x-axes represent the Raman shift region 400–1800 cm^{-1}, and y-axes, which are shared for each row, represent normalised intensity (a.u.). Raw RS data acquired by Dr Álvaro Fernández-Galiana and Georgios Papadopoulos. Assembled in BioRender.com. Reproduced from Georgiev et al. [526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

4.5 Application of unmixing autoencoders to biological data: volumetric Raman imaging of a THP-1 cell

As an application to biological research, the unmixing autoencoders were used to analyse a low SNR volumetric RS raster scan of a human leukaemia monocytic (THP-1) cell by Kallepitis et al.[74] (Figure 4.10A). Using RS and chemometric techniques, the composition of the cell can be probed to study its morphology in a non-destructive, label-free manner.

4.5.1 Spectral preprocessing

The volumetric Raman scan of the THP-1 cell was collected using an integration time of 0.3 seconds, and comprises a $z = 1, \ldots, 10$ stack of ten 40×40 raster scans, organised into a single volumetric hypercube for analysis[74]. The data were loaded into RamanSPy and preprocessed before unmixing using the following protocol: 1) spectral cropping to the fingerprint region 700–1800 cm^{-1}; 2) cosmic spike removal using the algorithm by Whitaker and Hayes[159] with kernel of size 3 and z-value threshold of 8; 3) denoising with Savitzky-Golay filter[161] using a cubic polynomial kernel of size 7; 4) baseline correction using Asymmetric Least Squares (AsLS)[165] with smoothing parameter $\lambda = 10^6$, penalizing weighting factor $p = 0.01$, differential matrix of order 2, maximum iterations set to 50, and exit criteria with tolerance threshold of $t = 0.001$; and 5) global Min-Max normalisation to the interval $[0, 1]$.

4.5.2 Comparative unmixing analysis

After loading and preprocessing the data using RamanSPy, hyperspectral unmixing was conducted with: 1) VCA+NNLS – as in the original paper; 2) *Dense AE* – the simplest and most computationally efficient AE model developed here; and 3) *Deep Dense AE* – an extension of *Dense AE* with a deeper encoder with five layers. Unmixing was performed following the same AE training protocol as in other analyses, with the number of training epochs set to 20 and the number of endmembers to extract to $n = 20$. Here, the constraint that abundances must sum to one was discarded.

The endmembers derived by each of the three methods were characterised via peak assignment to identify biochemical species present in the scanned cell. Out of those 20, the 5 endmembers deemed most biologically relevant were displayed following the peak assignment as per the original paper by Kallepitis et al.[74]. For VCA+NNLS, two of the five obtained endmembers corresponded to the same cell organelle, namely cytoplasm, and were visualised using the same colour in the merged reconstruction displayed in Fig-

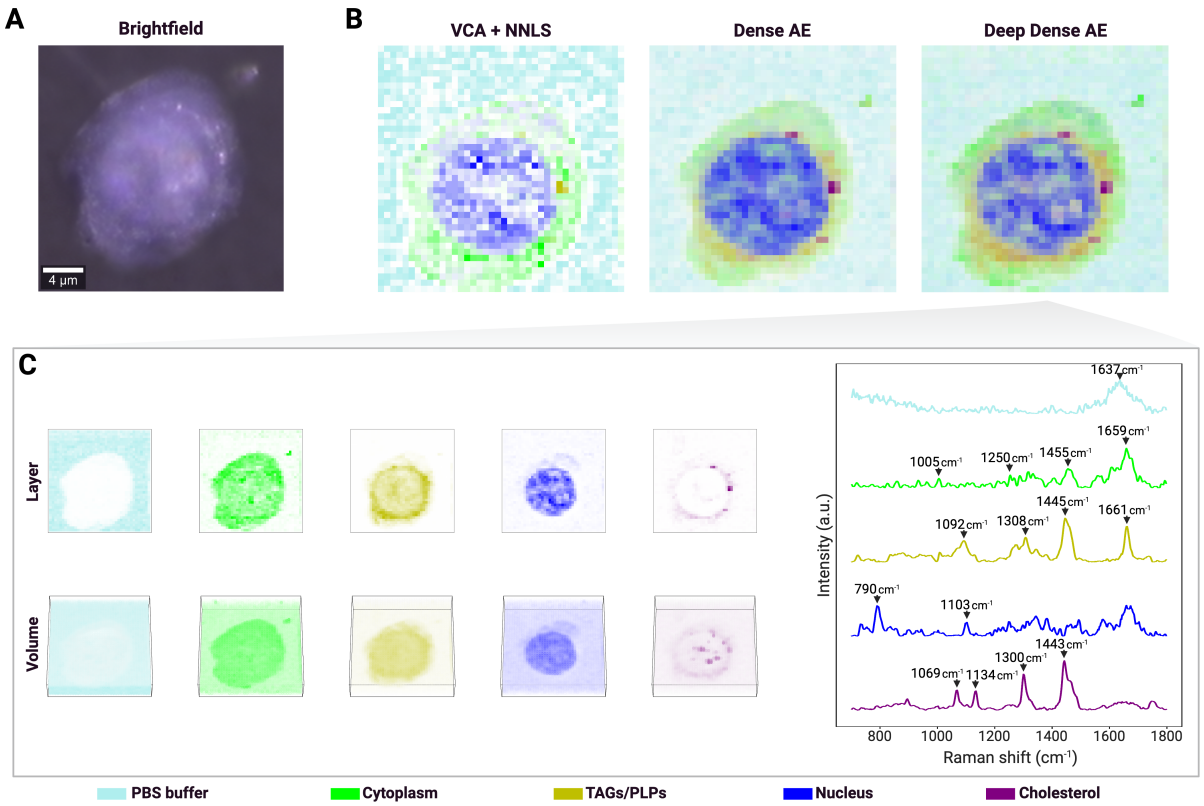


Figure 4.10: Volumetric Raman imaging of a THP-1 cell. A, Brightfield image of the studied THP-1 cell. **B,** Reconstruction of a cross-section of the cell (layer $z = 7$) obtained by overlaying the abundances derived by: VCA+NNLS and the developed *Dense AE* and *Deep Dense AE* models. **C,** Detailed overview of the unmixing results obtained with the *Deep Dense AE* model, displaying the spatial distribution (layer and volume) of the individual abundances (*left*) and the associated endmember signatures (*right*). Abundance maps are normalised for consistent visualisation. Brightfield image and raw RS data are from a previous study by Kallepitis et al.[74]. Assembled in BioRender.com. Reproduced from Georgiev et al.[526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

ure 4.10B. The observed species were related to deoxyribonucleic acid (DNA), proteins, triglycerides (TAGs), phospholipids (PLPs) and cholesterol esters. Specifically, cell organelles were determined based on the following peaks: PBS buffer - 1637 cm^{-1} (water peak); cytoplasm - 1005 cm^{-1} (phenylalanine), 1250 cm^{-1} (Amide III), 1659 cm^{-1} (Amide I) and 1445 cm^{-1} (CH deformations of proteins and lipids); TAGs/PLPs - 1092 cm^{-1} (C–C stretching), 1308 cm^{-1} (CH_2 twists), 1445 cm^{-1} (CH deformation), and 1661 cm^{-1} (C=C stretching); nucleus/DNA - 790 cm^{-1} (symmetric phosphodiester stretch and ring breathing modes of pyrimidine bases) and 1103 cm^{-1} (symmetric dioxy-stretch of the phosphate backbone); cholesterol - 1069 cm^{-1} and 1134 cm^{-1} (cholesteryl stearate), 1300 cm^{-1} (CH_2 twists), and 1443 cm^{-1} (CH deformation)[32,74,540].

Figure 4.10B shows the reconstructions of the cell created by overlaying selected abundances derived by each method, revealing the spatial organisation of key cellular or-

ganelles, including the nucleus, cytoplasm, lipid bodies and membranes. Although direct comparisons are challenging due to the lack of ground truth, the unmixing results of the AE models are aligned with the original findings by Kallepitis et al.[74], albeit with more distinct endmember signatures and well-defined abundance features (see Figures C.1–C.3 in Appendix C for full unmixing data). The *Deep Dense AE* model provides cleaner endmember signatures that enable more precise spectral and compositional information (Figure 4.10C). Notably, unlike the original VCA+NNLS approach, the developed AEs detected cholesterol, an important functional and structural component in cells that plays a key role in membrane fluidity and stability, signalling pathways and immune response[541–543].

Finally, it is noted that the unsupervised autoencoder models developed here do not utilise any ground truth information during training, in contrast with previous supervised machine learning approaches for Raman cell imaging and characterisation, where specific target concentrations are used as labels to train the model (e.g. using separately acquired fluorescence images)[81,219–221].

4.6 Conclusion

This chapter has presented an autoencoder-based methodology for hyperspectral unmixing in RS, which was validated on a wide array of synthetic and real experimental datasets. The results demonstrate that autoencoders are adept at handling diverse mixture scenarios and exhibit robustness against data artefacts, offering an effective, versatile and efficient framework for RS unmixing.

Data availability

The original datasets presented in this work, both synthetic data and experimental data from sugar mixtures, are available at https://doi.org/10.5281/zenodo.10779223. Previously published data from a THP-1 cell used in this work is available at https://doi.org/10.5281/zenodo.256329 (scan '001').

Code availability

The code and associated scripts developed for this work are available at https://github.com/barahona-research-group/raman-unmixing-aes. The synthetic data generator is also available within RamanSPy (https://github.com/barahona-research-group/RamanSPy).

Chapter 5

Label-free biochemical imaging and timepoint analysis of neural organoids via deep learning-enhanced Raman microspectroscopy

This chapter is based on the following manuscript, submitted for publication:

Georgiev, D.*, Xie, R.*, Reumann, D., Zhao, X., Fernández-Galiana, Á., Barahona, M., & Stevens, M. M. (submitted). Label-free biochemical imaging and timepoint analysis of neural organoids via deep learning-enhanced Raman microspectroscopy. *Preprint available on bioRxiv: 10.1101/2025.09.29.679057*, 2025.

*D.G. and R.X. contributed equally to this work.

Author contributions. I designed and optimised the computational pipeline, and developed all software and supporting tools. I performed all experiments and data analyses, making use of Raman imaging data collected by my co-authors Dr Ruoxiao Xie, Dr Daniel Reumann and Xiaoyu Zhao, who also cultured the neural organoids used in this study. Dr Ruoxiao Xie and Dr Daniel Reumann performed fluorescence microscopy. Dr Ruoxiao Xie, Dr Daniel Reumann, Dr Álvaro Fernández-Galiana, Prof. Mauricio Barahona and Prof. Dame Molly Stevens – contributed to supervision, conceptualisation, experimental design and interpretation of results. I wrote the manuscript, with input from all authors.

5.1 Introduction

As discussed in Section 2.2, neural organoids provide powerful *in vitro* models of the developing human brain, offering an experimental platform for studying brain development, disease and drug response. However, their analysis remains constrained by the limitations of standard imaging methods based on fluorescence microscopy, which are invasive, require exogenous labelling and offer limited multiplexing (Section 2.2.5). To fully exploit the potential of organoids, new approaches are needed that enable non-invasive, label-free imaging of intact samples under physiologically relevant conditions. Despite the potential of RS in this context, its application to neural organoids has thus far been largely unexplored due to persisting challenges in RS data analysis and chemometrics.

The unmixing autoencoders introduced in Chapter 4 enables fully unsupervised, label-free biochemical analysis with improved accuracy compared to standard chemometric techniques, as validated on synthetic datasets, sugar mixtures and cell imaging data. However, it remains an open question whether this framework can be applied effectively in more complex experimental settings involving intact neural organoids, which pose greater heterogeneity and structural complexity than the samples used previously.

Building on the work described in Chapters 3 and 4, this chapter presents a pipeline for non-invasive, label-free organoid imaging based on spontaneous Raman microspectroscopy and unsupervised autoencoder-based hyperspectral unmixing analysis. The proposed approach achieves improved biochemical imaging compared to standard methods for RS analysis, as validated against fluorescence microscopy, and enables detailed morphochemical imaging and spatially resolved analysis with subcellular resolution in both cryosectioned and intact neural organoids. Using the platform, volumetric *in situ* imaging of a neural rosette structure within a neural organoid is demonstrated – a key morphological feature of the developing nervous system. Furthermore, changes in biochemical composition during early developmental stages in intact neural organoids are interrogated, revealing spatiotemporal variations in lipids, proteins and nucleic acids.

5.2 Deep learning-enhanced Raman microspectroscopy

First, the technical methodology behind the organoid imaging pipeline based on Raman microspectroscopy combined with autoencoder-based hyperspectral unmixing is discussed. Figure 5.3 provides an overview of the proposed approach.

Raman measurements were collected from fixed neural organoid samples by performing point-wise imaging scans across predefined regions of interest. Volumetric scans were

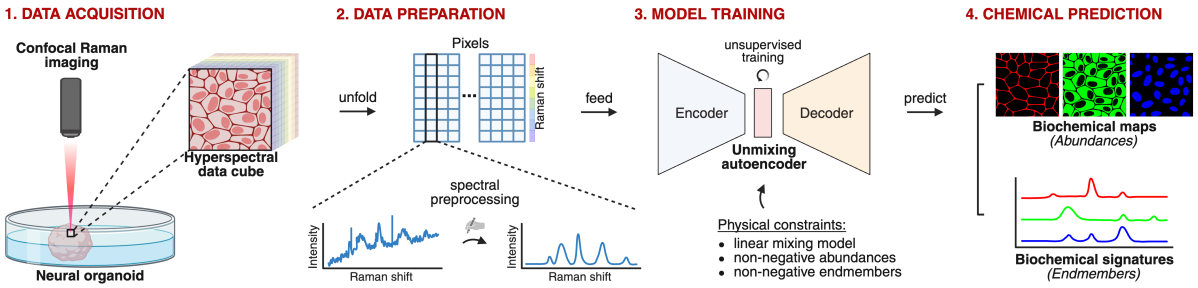


Figure 5.1: Schematic diagram of developed organoid imaging pipeline. The pipeline combines Raman microspectroscopy and deep learning-enhanced hyperspectral analysis. Created in BioRender.com.

generated by acquiring sequential scans along the axis perpendicular to the imaging plane (e.g. sampling at different depths along the z-axis). Acquired data were preprocessed to remove non-specific signal contributions and spectral artefacts, such as cosmic spikes, autofluorescence baselines and dark noise (see Section 5.3). This was done using the RamanSPy Python toolbox (Chapter 3).

After preprocessing, hyperspectral unmixing was performed to derive constituent spectral components (endmembers) and quantify their relative contributions (abundances) in each measurement. This was achieved using an autoencoder neural network model designed for hyperspectral unmixing, following the approach presented in Chapter 4. Briefly, the autoencoder architecture included: (i) an encoder responsible for abundance estimation, which comprised a multi-branch spectral convolutional block followed by a five-layer fully connected network, and (ii) a one-layer decoder responsible for endmember identification. Relevant physical constraints were incorporated in the model and the representation learning process, namely, constraining the learnt endmembers and abundances to be non-negative through the use of appropriate activation functions and weight clipping, and constraining the mixing process to linear mixing through the use of a single-layer linear decoder. Models were trained on individual scans or datasets in an unsupervised manner by minimising a reconstruction loss based on the weighted sum of the mean squared error and spectral angle divergence[517] between input spectra and output reconstructions.

As noted in Section 2.1.5, the number of endmembers to extract from a given dataset is a key hyperparameter in blind unmixing, and its choice strongly affects the resulting abundance maps. To guide this selection, the number of endmembers for each scan/dataset was determined using principal component analysis (PCA)[172,173] (see Section 5.3.5).

After model training, learnt endmember signatures and their corresponding abundances were derived, generating multi-channel abundance images/volumes with channels rendering chemical maps of the distribution of learnt endmembers. The obtained endmembers

and abundances were examined to filter out components linked to non-tissue signals, such as background and noise. The remaining tissue-specific endmembers were characterised via peak assignment, whereby prominent peaks were linked to chemical bonds to identify essential biomolecular species, such as nucleic acids, proteins, lipids and carbohydrates. Peak identification was guided by established reference materials[33]. The corresponding abundance maps were visualised as morphochemical images showing the spatial distribution of the obtained chemical components to study organoids' morphology and biochemical composition.

Further technical details regarding data acquisition, preprocessing and unmixing analysis are presented in the following section.

5.3 Methods and Materials

5.3.1 Organoid culture and preparation

Neural organoids were grown using a protocol for unguided differentiation based on previous work from Lancaster et al.[7] (Figure 5.2). Human embryonic stem cells (WA09, WiCell) were cultured feeder-free on hES-qualified Matrigel (354277, Corning) coated plates in a 5% CO_2 incubator at 37°C. mTeSR1 medium (STEMCELL Technologies) was used for daily cell maintenance according to manufacturer instructions.

To generate neural organoids, cells were dissociated into single cells using Accutase (STEMCELL Technologies), resuspended in mTeSR1 supplemented with 50 µm Y-27632 (S1049, Selleck Chemicals) and 1% (v/v) Penicillin/Streptomycin (15140122, Thermo Fisher Scientific), and then added to an ultra-low-binding 96-well plate (Corning) at 9000 cells/150 µm per well. On day 3, 100 µm of the media was removed and 150 µm fresh mTeSR1 containing 1% (v/v) Penicillin/Streptomycin was added. From day 5 onwards, media was changed to neural induction (NI) media, composed of Dulbecco's Modified Eagle Medium (DMEM)/F12 (Thermo Fisher Scientific) supplemented with 1 µm/mL heparin (Sigma-Aldrich), 1% (v/v) Penicillin/Streptomycin, 1%(v/v) GlutaMAX (Thermo Fisher Scientific), 1% (v/v) N2 supplement (Thermo Fisher Scientific), and 1% (v/v) MEM-NEAA (Thermo Fisher Scientific). On day 10, organoids were embedded in Matrigel (354234, Corning) and transferred to 10 cm petri dishes. From day 12 to day 18, the culture medium was switched to Improved-A medium consisting of a 1 : 1 mixture of DMEM/F12 and Neurobasal Medium (Thermo Fisher Scientific) supplemented with 1% (v/v) Penicillin/Streptomycin, 1% (v/v) B27 supplement without vitamin A (Thermo Fisher Scientific), 1% (v/v) GlutaMAX, 0.5% (v/v) N2 supplement, 1/4000 (v/v) insulin

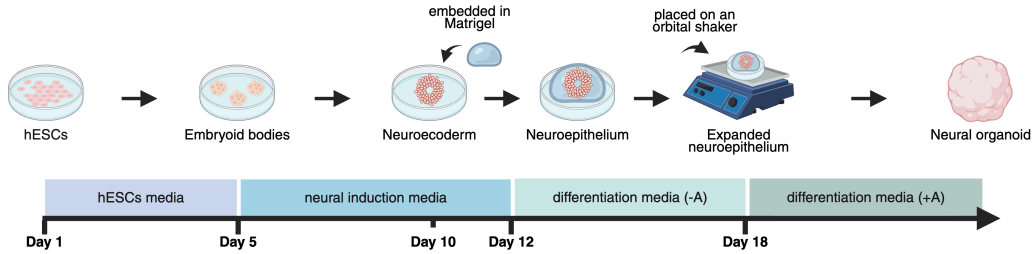


Figure 5.2: Schematic diagram of organoid culture. Neural organoids were grown using a protocol for unguided differentiation based on previous work from Lancaster et al.[7]. Created in BioRender.com.

(I9278, Sigma-Aldrich), 0.5% (v/v) MEM-NEAA, and 1 mg/mL sodium bicarbonate. Organoids were fed with fresh Improved-A medium every 3 days – i.e. on days 12 and 15. From day 18 onwards, the petri dishes were placed on an orbital shaker, and the culture medium was replaced with Improved+A medium consisting of a 1 : 1 mixture of DMEM/F12 and Neurobasal Medium (Thermo Fisher Scientific) supplemented with 1% (v/v) Penicillin/Streptomycin, 2% (v/v) B27 supplement with vitamin A (Thermo Fisher Scientific), 1% (v/v) GlutaMAX, 0.5% (v/v) N2 supplement, 1/4000 (v/v) insulin (I9278, Sigma-Aldrich), 0.5% (v/v) MEM-NEAA, 1 mg/mL sodium bicarbonate, and 0.07 mg/mL vitamin C.

Organoids were fixed at days 5, 10, 20, 30 and 40 in 4% (w/v) paraformaldehyde (PFA) overnight at 4°C, then rinsed three times with PBS (Thermo Fisher Scientific). Fixed organoids were then immobilised in 1% (w/v) agarose for Raman imaging.

To prepare the organoid slices for Raman imaging or immunostaining, the organoid tissues were cryosectioned. Briefly, fixed organoid tissues were first incubated in 30% sucrose at 4°C for a day on a roller mixer, then transferred into Tissue-Tek O.C.T. Compound (Sakura Finetek USA), gently swirled, and incubated for 5 minutes. The tissues were then frozen in a cryomold filled with Tissue-Tek O.C.T. For the two cryosections prepared for RS imaging, this was achieved using the technique for high-throughput cryoprocessing previously reported by Reumann et al.[476]. Compound to form cryoblocks containing the frozen organoids. The cryoblocks were subsequently sectioned into 20 µm slices using a CryoStar NX70 Thermo Cryostat, and collected on positively charged cryoslides. The tissue sections were dried for at least 3 hours before being stored at -20°C or used for downstream analysis (fluorescence microscopy and Raman microspectroscopy).

Credits: Organoid culture was performed by Dr Ruoxiao Xie, Dr Daniel Reumann and Xiaoyu Zhao.

5.3.2 Immunofluorescence staining and microscopy

For immunostaining, microscopy slices were washed in PBS for 5 minutes to wash off residual OCT, then permeabilised and blocked in PBS containing 0.3% (v/v) Triton X-100 (Thermo Fisher Scientific) and 5% (w/v) BSA (Thermo Fisher Scientific) for 0.5–1.5 hours at room temperature. Primary and secondary antibodies were diluted in PBS containing 0.1% (v/v) Triton X-100 and 5% (w/v) BSA. Primary antibody solutions were incubated with the slices overnight at 4°C, followed by three rinses in PBS, and three washes with PBS-T (PBS+0.01% TX100) for 10–15 minutes on a slow rotating orbital shaker at room temperature. Secondary antibody solutions were then added and incubated for 2 hours at room temperature. Extended information about the primary and secondary antibodies used for immunostaining is provided in Table D.1 and Table D.2 in Appendix D, respectively. Cell membranes were labelled using DiO (BT60011, Biotium) – a fluorescent lipophilic dye – which was added to the secondary antibody solution. DAPI (4',6-diamidino-2-phenylindole) (Thermo Fisher Scientific), diluted 1 : 1000 in PBS, was used to stain nuclei for 7 minutes. After this, slides were rinsed three times in PBS, and washed twice in PBS-T and once in PBS for 10–15 minutes on a slowly rotating orbital shaker. Stained organoid slices were then mounted using fluorescent mounting media (S302380-2, Agilent DAKO) or FluoroSave (Millipore). Fluorescent images were acquired using a Leica SP8 confocal microscope and a Nikon TI-2 widefield fluorescent microscope, and processed in FIJI.

Credits: Immunofluorescence staining and microscopy were performed by Dr Ruoxiao Xie and Dr Daniel Reumann. Ellena O'Keeffe assisted with immunostaining of one organoid cryosection.

5.3.3 Raman microspectroscopy

Raman spectroscopy measurements were performed using a confocal Raman microscope (alpha300 R, WITec GmbH) equipped with a 532 nm solid-state laser (WITec GmbH). Laser power at the sample plane was set to about 48 mW for organoid cryosections or 32 mW for intact samples. Backscattered light was collected through a water immersion objective – 20× (Zeiss W N-Achroplan, N.A.= 0.5) for organoid cryosections or 63× (Zeiss W Plan-Apochromat, N.A.= 1) for intact samples, and delivered via a 10 µm single-mode silica fiber to an imaging spectrograph (Newton, Andor Technology Ltd) with a 600 groove/mm grating and a thermoelectrically cooled CCD detector (−60°C). The system provided a spectral resolution of $\sim$ 2–3 cm^{-1} over the wavenumber range 0–3670 cm^{-1}. Spectra were acquired with pixel sizes in x and y and step sizes in z ranging from 1 µm to 10 µm, and an integration time of 0.5 seconds.

Credits: RS data collection was performed by Dr Ruoxiao Xie, Dr Daniel Reumann and Xiaoyu Zhao. Kevin Giraldo Rodriguez assisted with RS imaging of one organoid cryosection.

5.3.4 Spectral preprocessing of Raman imaging data

Acquired Raman spectroscopy measurements were exported as MATLAB files using the WITec Project FIVE Suite software and loaded in Python using RamanSPy. Following this, each scan was individually preprocessed using RamanSPy as follows: (1) spectral cropping whereby wavenumbers less than 300 cm^{-1} were removed; (2) cosmic rays removal, using the algorithm reported by Whitaker and Hayes[159], with a kernel size $m = 3$ and z-value threshold $\tau = 6$; (3) denoising with a third-order Whittaker smoother[162] with $\lambda = 10^3$; (4) baseline correction with asymmetric least squares[165] with $\lambda = 10^5$; and (5) vector normalisation. Normalisation was applied on a pixel-wise basis for the collected timepoint data to account for depth-dependent signal attenuation. Global vector normalisation was applied to the rest of the data, where intensity values were divided by the ℓ^2-norm of the spectrum with the highest norm in the given scan. Additionally, wavenumber calibration was applied to the two cryosection scans by calculating the median position of the laser peak (Rayleigh scattering) and setting it to zero.

5.3.5 Hyperspectral unmixing analysis

Chemometric analysis was performed using a custom autoencoder neural network model designed for blind hyperspectral unmixing, following the approach developed in Chapter 4.

The encoder comprised two separate blocks applied sequentially. The first part was a multi-branch convolutional block comprising four parallel convolutional layers with kernel sizes of 5, 10, 15 and 20, designed to capture patterns at multiple spectral scales. Each convolutional layer contained 32 filters with ReLU activation, He initialisation[544] and 'same' padding. Batch normalisation[545] and dropout with a rate of 0.2[538] were applied to each convolutional layer to improve training stability and generalisation. The outputs of the four convolutional layers were merged channel-wise through a fully connected layer to yield an output of dimension matching that of the input spectrum. The rationale behind this was to transform intensity values into representations that capture local spectral features (e.g. peak shape, width, local neighbourhood) and thus promote better generalisability. The second part of the encoder was a fully connected dimensionality reduction block, applied to learn patterns between the learnt spectral features. This

block comprised a series of fully connected layers of sizes 256, 128, 64 and 32 with He initialisation and ReLU activation. Batch normalisation and dropout (rate of 0.5) were also applied at each fully connected layer. The block was followed by a final fully connected layer (Xavier uniform initialisation[546]) that reduced the final 32 features to a latent space of size n. The number n was treated as a hyperparameter that encodes the number of endmembers to extract, with latent representations treated as abundance fractions. To improve interpretation, non-negativity was enforced in the latent space using a 'softly-rectified' hyperbolic tangent function $f(x) = \frac{1}{\gamma} log(1 + e^{\gamma * tanh(x)})$, with $\gamma = 10$, as in Chapter 4. This ensured that latent values were constrained to the range $(0, 1)$ in accordance with the abundance non-negativity constraint, but does not enforce abundances to sum to one (abundance sum-to-one constraint).

The decoder consisted of a single fully connected layer (Xavier uniform initialisation) mapping the n-dimensional latent space representations back to the original spectral dimension b. In this layer, a linear activation was used and no bias term. Under this setup, the decoder mimics linear unmixing, where endmember signatures are encoded in the learnt weight matrix of the decoder layer, as discussed in Chapter 4. A non-negative kernel constraint was used to enforce the non-negativity of endmember signatures using weight clipping during training. To accelerate model training, the weight matrix in the linear decoder layer was initialised before training using endmember signatures derived with vertex component analysis[235].

Autoencoder models were trained for 5 epochs using the Adam optimiser[539] with a batch size of 32. The learning rate was set to 0.001 for the two cryosection scans and subcellular day-10 scan, or 0.003 for all other data. An exponential learning rate decay was used to ensure stable convergence, applying a factor of 0.9 to the learning rate after each epoch. The training loss $\mathcal{L}$ was defined as a weighted sum of the mean squared error (MSE) and spectral angle divergence (SAD) between input spectra $\mathbf{x}$ and output reconstructions $\hat{\mathbf{x}}$:

$$\mathcal{L}(\mathbf{x}, \hat{\mathbf{x}}) = \mathrm{SAD}(\mathbf{x}, \hat{\mathbf{x}}) + \lambda\, \mathrm{MSE}(\mathbf{x}, \hat{\mathbf{x}}), \tag{5.1}$$

with a weighting factor $\lambda = 1000$. Similarly to Chapter 4, MSE and SAD were defined as:

$$\mathrm{MSE}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{b} ||\mathbf{x} - \hat{\mathbf{x}}||_2, \tag{5.2}$$

with b denoting the dimension of $\mathbf{x}$ (i.e. the number of spectral bands), and

$$\mathrm{SAD}(\mathbf{x}, \hat{\mathbf{x}}) = \arccos\left(\frac{\mathbf{x} \cdot \hat{\mathbf{x}}}{\|\mathbf{x}\|_2 \, \|\hat{\mathbf{x}}\|_2}\right). \tag{5.3}$$

The number of endmembers to derive was estimated individually for each scan/dataset

of interest using principal component analysis. Specifically, the cumulative explained variance was computed as a function of the number of principal components and manually selected the number of components at the point where the increase in explained variance visually plateaued.

Autoencoder model implementation, training and evaluation steps were carried out using TensorFlow[516]. Analysis with baseline chemometric methods was performed using RamanSPy.

5.4 Label-free imaging in intact cerebral organoids via deep learning-enhanced Raman microspectroscopy

The capabilities of the platform were investigated using Raman imaging data from neural organoids cultured in-house, as discussed in Section 5.3. Brightfield microscopy revealed spherical neural organoid samples reaching about 1–2 mm in size after 20 days in culture (Figure 5.3A). Furthermore, fluorescence microscopy showed characteristic morphology with neural rosettes readily visible on day 20 (Figure 5.3B).

As a demonstration of the Raman imaging platform, volumetric Raman imaging was performed on a tissue region within an intact day-20 neural organoid. Specifically, a z-stack was acquired with an axial step size of 10 µm in z (organoid depth), with each layer representing an 80×80 imaging scan with a lateral pixel size of 5 µm in the x-y plane. Five consecutive layers along the z-axis were used for unmixing analysis, covering tissue areas up to approximately 70 µm deep into the sample. Figure 5.3C shows the intensity profile at 1659 cm^{-1} of one of the five layers after preprocessing, a band associated with proteins (amide I) and lipids (C=C stretching), along with an exemplar tissue spectrum before and after preprocessing. Although the overall outline of the organoid was visible, finer structures were difficult to resolve using univariate analysis of single Raman bands. Note that a slightly more aggressive baseline correction was intentionally applied; while this caused minor overcorrection of the water O–H stretching peaks in the high wavenumber region (3000–3500 cm^{-1}), it facilitated more accurate reconstruction of smaller, more relevant peaks in the fingerprint region (300–1800 cm^{-1}) during downstream unmixing, likely due to the weighting effect of the reconstruction loss.

To address this issue, multivariate hyperspectral unmixing was performed using the developed deep learning-enhanced Raman microspectroscopy pipeline. The analysis yielded six tissue-specific spectral endmembers, which were characterised via peak assignment into major biomolecular classes based on characteristic Raman vibrations in the fingerprint

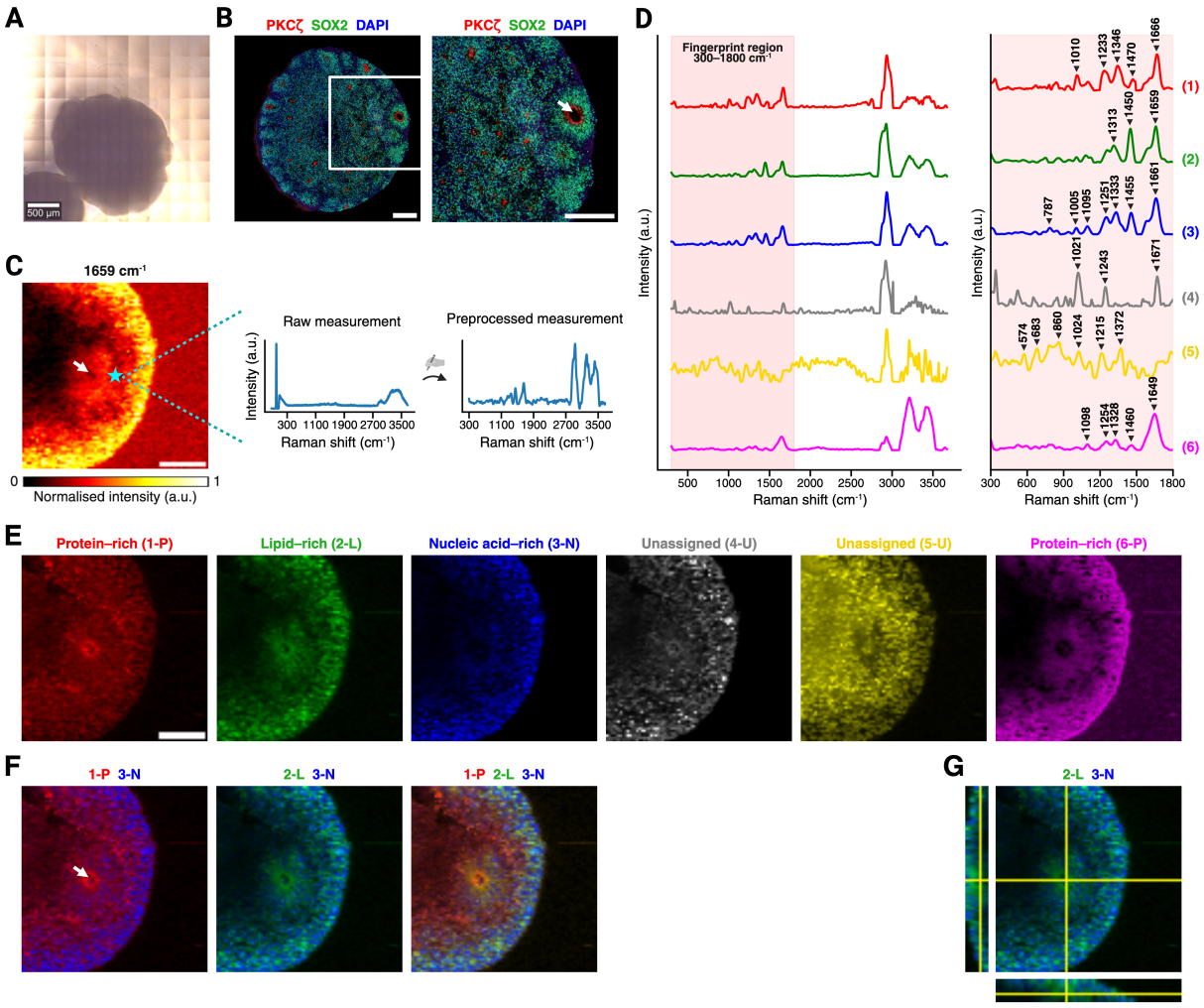


Figure 5.3: Deep learning-enhanced Raman microspectroscopy enables label-free *in situ* biochemical imaging in intact neural organoids in 2D and 3D. A, Brightfield image of a representative day-20 neural organoid sample. Scale bar: 500 µm. Image acquired by Dr Ruoxiao Xie. **B,** Fluorescence microscopy image of a representative day-20 neural organoid sample – full organoid sample (*left*), and a zoomed-in region (*right*), with staining markers for apical domains (PKCζ), neural progenitor cells (SOX2) and nuclei (DAPI). Arrow indicates the lumen of a present neural rosette. Scale bar: 200 µm. Image acquired by Dr Ruoxiao Xie. **C,** Snapshot of collected volumetric Raman imaging data from an intact day-20 neural organoid, showing the intensity profile at the 1659 cm^{-1} band (proteins and lipids) of a selected z-layer (*left*), and a representative tissue spectrum before and after preprocessing (*right*). Arrow indicates the lumen of a present neural rosette. Scale bar: 100 µm. Raw RS data acquired by Dr Ruoxiao Xie and Xiaoyu Zhao. **D–G,** Results of hyperspectral unmixing analysis. **D,** Derived tissue-specific endmember signatures – whole wavenumber range (*left*) and fingerprint region only (*right*). Endmembers scaled for visualisation purposes (maximum intensity set to 1). **E–F,** Corresponding abundance maps for the same z-layer as in **C**, including single-channel images (**E**) and representative multi-channel images (**F**). Arrow indicates the lumen of the neural rosette shown in **F**. Scale bar: 100 µm. **G,** Volumetric abundance reconstruction of two selected endmembers linked to lipid- and nucleic acid-rich species. Assembled in BioRender.com.

region 300 cm^{-1} to 1800 cm^{-1} (Figure 5.3D; full unmixing data are available in Figure D.1 in Appendix D). The corresponding abundance maps were visualised as morphochemical images of the scanned organoid sample (Figure 5.3E–F). Endmember 1 (red) exhibited

a spectral signature characteristic of protein species, with prominent peaks at 1010 cm^{-1} (phenylalanine), 1233 cm^{-1} (amide III), 1346 cm^{-1} (CH deformation), 1470 cm^{-1} (CH_2 bending, C=N stretching) and 1666 cm^{-1} (amide I). The abundance map showed broad distribution throughout the organoid tissue, likely corresponding to cytoplasmic regions. Slightly increased intensity was observed around a structure resembling the apical domain of a neural rosette, possibly indicating membrane-associated protein structures. Endmember 2 (green) was characterised by strong lipid-associated peaks at 1313 cm^{-1} (CH_2/CH_3 twisting), 1450 cm^{-1} (CH_2 bending) and 1659 cm^{-1} (C=C stretching), suggesting the presence of lipid-rich compartments such as cellular membranes and lipid bodies. This component was similarly broadly distributed across the tissue, with increased signal in the outer layers of the sample, and around the apical domain of the neural rosette, forming a radially oriented pattern. Endmember 3 (blue) displayed main nucleic acid vibrations, including 787 cm^{-1} (O–P–O stretching) and 1095 cm^{-1} (PO_2^- groups), alongside protein-related peaks at 1005 cm^{-1}, 1251 cm^{-1}, 1455 cm^{-1} and 1661 cm^{-1}. This is consistent with cell nuclear regions, which is further supported by visual inspection when overlaid with the abundance maps of endmembers 1 and 2. Endmembers 4 (grey) and 5 (yellow) presented less conventional profiles, suggesting more complex biomolecular mixtures. Endmember 4 included peaks at 1021 cm^{-1} (glycogen), 1243 cm^{-1} (proteins) and 1671 cm^{-1} (proteins/lipids), and its spatial distribution overlapped with that of endmember 3. Endmember 5 displayed a noisier spectrum, with higher abundance within the interior of the organoid. Prominent peaks were associated to proteins (574 cm^{-1}, tryptophan/cytosine; 860 cm^{-1}, C–C stretching; 1215 cm^{-1}, amide III) and carbohydrates (1024 cm^{-1}, glycogen; 1372 cm^{-1}, CH deformation). Lastly, endmember 6 (magenta) showed a broadly distributed component, with bands related to proteins (1254 cm^{-1}, 1328 cm^{-1}, 1460 cm^{-1} and 1649 cm^{-1}) and nucleic acids (1098 cm^{-1}).

Beyond 2D imaging, the approach also enables biochemical analysis in three dimensions. To illustrate this, a volumetric rendering of the imaged organoid was generated by overlaying the two derived components related to lipid-rich (endmember 2) and nucleic acid-rich (endmember 3) species as markers of nuclear and membrane regions, respectively. The reconstruction, displayed in Figure 5.3G, allowed the visualisation and analysis of the three-dimensional organisation and architecture of the imaged organoid tissue, including that of the present neural rosette (rosette lumen indicated with an arrow in Figure 5.3C). The derived organoid image using RS showed characteristic morphology, correlated with fluorescence microscopy (Figure 5.3B), yet was achieved without exogenous labelling or tissue sectioning. This demonstrates that the Raman-based imaging platform can effectively delineate organoid composition *in situ* and generate 2D and 3D morphochemical maps of key biochemical species in intact neural organoid samples in a fully label-free manner. Building on this proof of concept, the chapter proceeds to higher-resolution

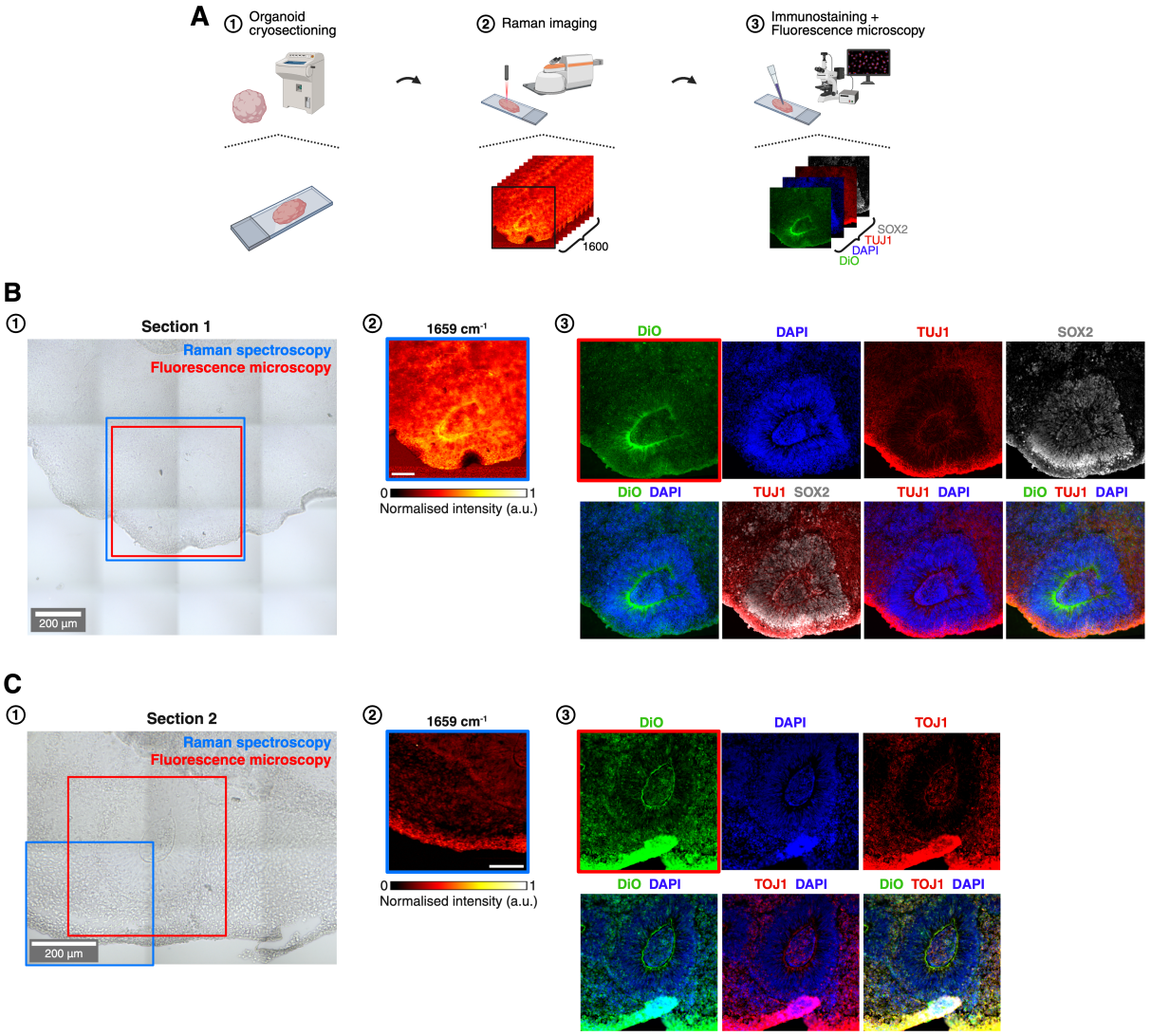


Figure 5.4: Collection of paired Raman imaging and fluorescence microscopy data from neural organoid cryosections for validation. A, Schematic diagram of validation strategy, which involved the collection of paired Raman imaging and fluorescence microscopy data from neural organoid cryosections. **B–C,** Collected data from two day-30 neural organoid cryosections. *left*, Brightfield image of the sectioned neural organoid tissue region imaged using Raman spectroscopy (blue) and fluorescence microscopy (red). *middle*, Raman intensity profile at the 1659 cm^{-1} band (linked to proteins and lipids). *right*, Fluorescence microscopy images with markers for nuclei (DAPI), membranes (DiO), neural progenitors (SOX2) and neurons (TUJ1). Scale bars: 100 µm. Brightfield images, fluorescence microscopy data and raw RS data acquired by Dr Daniel Reumann. Created/assembled in BioRender.com.

Raman imaging in the following sections.

5.5 Experimental validation and optimisation against fluorescence microscopy

To further validate that the method produces accurate morphochemical organoid reconstructions, its performance against fluorescence microscopy and standard methods for hyperspectral RS analysis was evaluated. To this end, defined tissue regions within two day-30 neural organoid cryosections were imaged using confocal Raman microspectroscopy, followed by immunostaining and fluorescence microscopy (Figure 5.4). Raman imaging was performed first to avoid potential signal interference from staining molecules. To ensure morphological diversity, areas containing both distinct neural rosette structures and surrounding neural tissue were selected for downstream imaging. Raman imaging was performed over areas of 600 µm×600 µm and 400 µm×400 µm from the two respective sections, with a pixel size of 2 µm. Subsequently, the organoid sections were stained for nuclei (DAPI), cell membranes (DiO), neural progenitor cells (SOX2) and neurons (TUJ1), and fluorescence microscopy images were taken to validate and correlate with organoid morphology predicted by RS.

5.5.1 Qualitative validation using paired fluorescence microscopy and Raman imaging data from a neural organoid cryosection

Using the autoencoder-based pipeline, the acquired RS data from the first organoid section were analysed to infer its biochemical composition. Figure 5.5A shows the four tissue-specific components extracted, which were associated with contributions from lipids, nucleic acids and proteins (full unmixing data are available in Figure D.2 in Appendix D). Endmember 1 (green) presented a lipid-rich spectral signature with distinct peaks at 1307 cm^{-1}, 1455 cm^{-1}, and 1663 cm^{-1}. Its abundance was consistent with membrane structures, including cellular membranes and the apical domain of the visible neural rosette. The signal of this endmember showed a high degree of visual overlap with DiO and TOJ1. Endmember 2 (blue) displayed a nucleic acid-rich profile with a prominent band at 783 cm^{-1}, alongside vibrations linked to proteins (1338 cm^{-1}, 1472 cm^{-1} and 1668 cm^{-1}). Strong glass signal was also observed around 570 cm^{-1} and 1100 cm^{-1}. The abundance map revealed well-defined, spherical compartments, closely resembling cell nuclei, with increased density within the present neural rosette structure. Endmembers 3 (red) and 4 (grey) shared similar spectral signatures, dominated by protein-associated peaks around ∼1010 cm^{-1}, ∼1100 cm^{-1}, ∼1265 cm^{-1}, ∼1320 cm^{-1}, ∼1460 cm^{-1}, and ∼1670 cm^{-1}. Both endmembers were generally broadly distributed across the sample. Endmembers 3 showed increased signal at the apical domain of the present rosette, po-

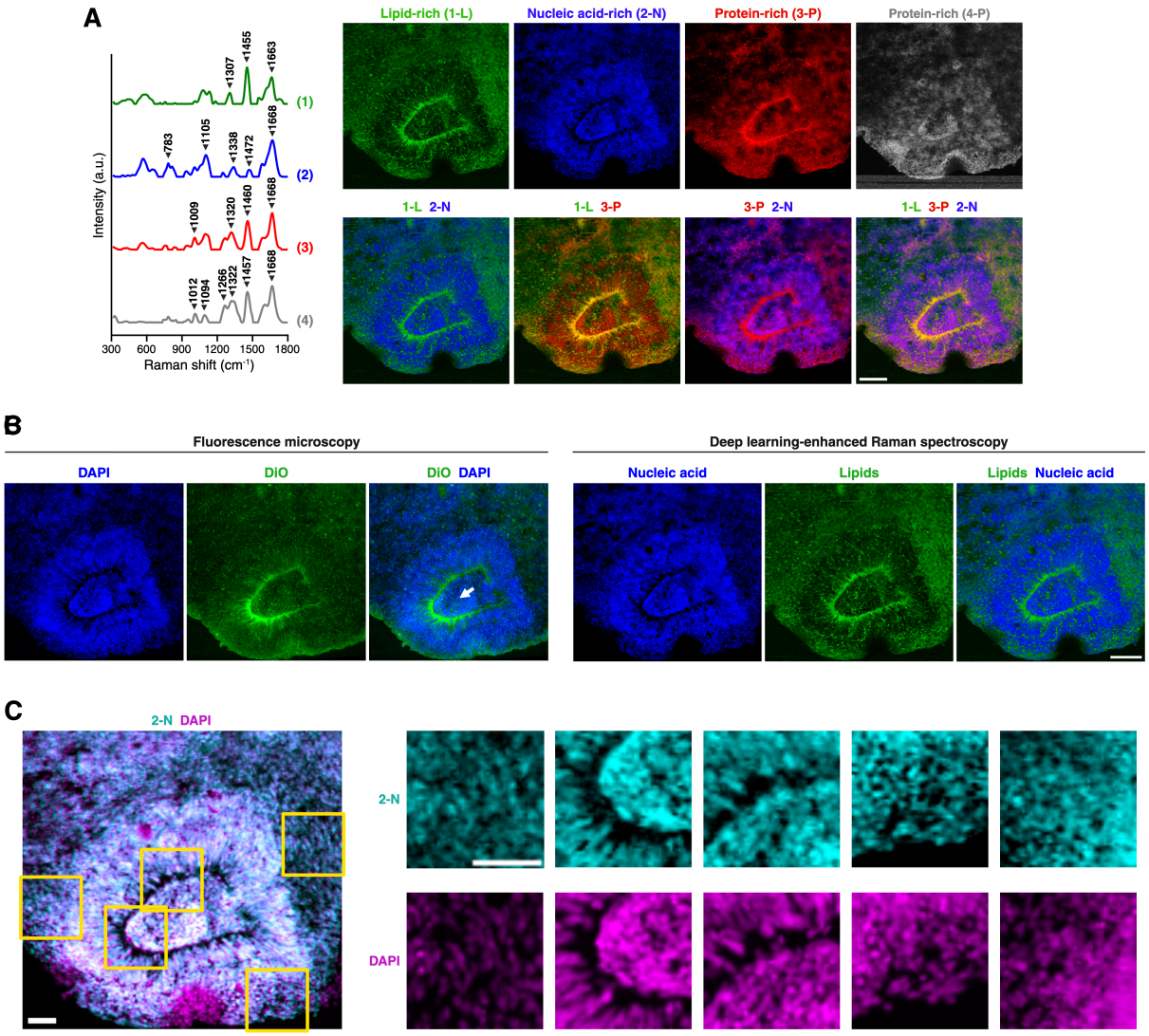


Figure 5.5: Raman imaging of a neural organoid cryosection produces reconstructions aligned with fluorescence microscopy. A, Result of hyperspectral unmixing analysis on data from a day-30 neural organoid cryosection. *left,* Derived tissue-specific endmember signatures. Endmembers scaled for visualisation purposes (maximum intensity set to 1). *right,* Corresponding abundance maps. Scale bar: 100 µm. **B,** Side-by-side comparison of fluorescence microscopy images of DAPI (nuclei) and DiO (membranes) (*left*), and RS-based abundance reconstructions of nucleic acid and lipids derived with the developed autoencoder pipeline (*right*). Arrow indicates the lumen of the present neural rosette. Scale bar: 100 µm. **C,** Overlaid images of DAPI and endmember associated with nucleic acids (2-N). Scale bars: 25 µm. Fluorescence microscopy images and raw RS data acquired by Dr Daniel Reumann. Assembled in BioRender.com.

tentially suggesting structural proteins. The abundance of endmember 4 was increased within the rosette structure.

Figure 5.5B shows a side-by-side comparison between the predicted abundance maps of the two endmembers related to nucleic acid-rich and lipid-rich species and the corresponding fluorescence microscopy images of nuclei (DAPI) and cell membranes (DiO), respectively. The RS imaging pipeline accurately reconstructed individual nuclei, aligning with

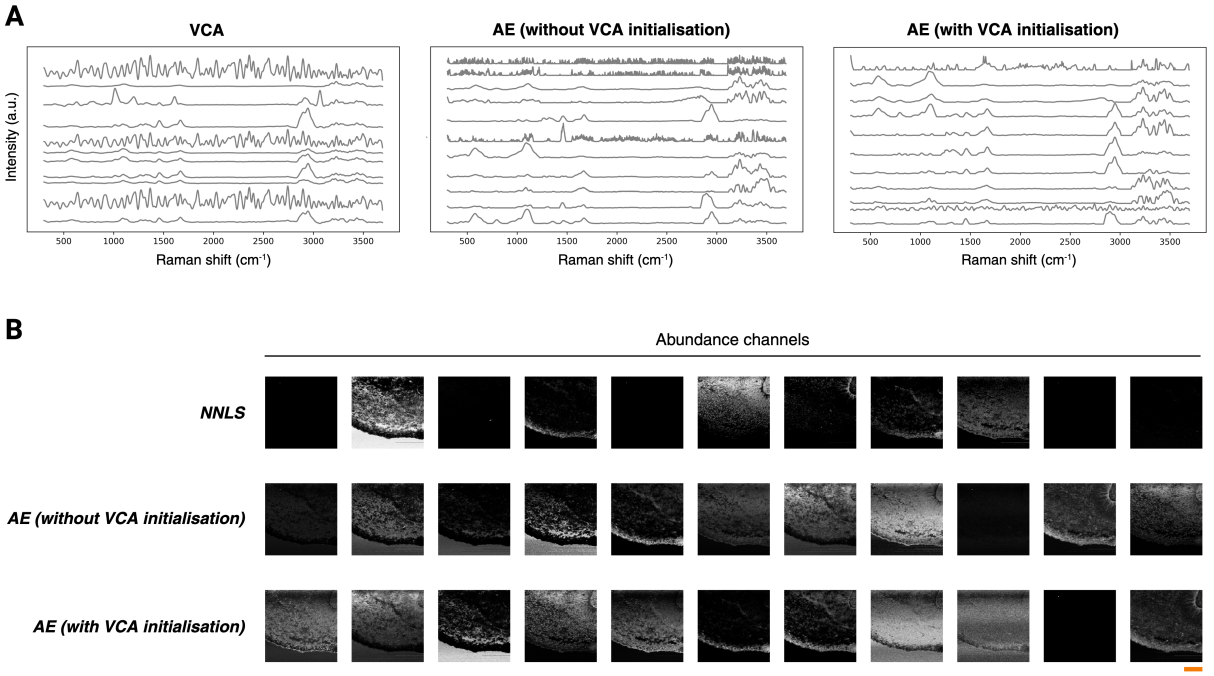


Figure 5.6: Effect of endmember initialisation on autoencoder performance. Unmixing results obtained with: (i) vertex component analysis (VCA), (ii) the developed autoencoder model without VCA initialisation, and (iii) the autoencoder model with VCA initialisation. **A,** Derived endmembers. **B,** Derived abundances. Abundance images were normalised for visualisation purposes. Scale bar: 100 µm (bottom right, marked in orange). Raw RS data acquired by Dr Daniel Reumann. Assembled in BioRender.com.

the DAPI staining across the imaged tissue (Figure 5.5C). It also effectively captured important membrane features, including both fine cellular membranes and prominent structures such as the apical domain of the visible neural rosette. Qualitatively, strong correspondence between the lipid-rich component obtained with RS and the DiO signal was observed around the apical domain of the rosette and with the TUJ1 signal outside the rosette (see Figure 5.4B).

5.5.2 Pipeline optimisation, ablation studies and comparison with standard unmixing methods

Having shown that the method accurately reconstructs key cellular components in cryosectioned neural organoid samples, the data collected from the second organoid section were used to perform a series of comparative analyses evaluating different configurations of the autoencoder-based pipeline. These included endmember initialisation strategies (Figure 5.6) and variations in training losses (Figure 5.7). Based on these results, initialisation with vertex component analysis (VCA)[235] was found to be beneficial despite the added

computational cost, yielding cleaner endmember spectra and more accurate abundance maps. Additionally, the custom loss function combining MSE and SAD, defined in Equations 5.1–5.3, was selected for subsequent experiments, as it provided a good balance between reconstruction fidelity and spectral discrimination.

Following this, the unmixing performance of the developed approach was compared against a range of standard chemometric methods for multivariate analysis (Figure 5.8). This included methods for dimensionality reduction (PCA, NMF), clustering (k-means), and hyperspectral unmixing (N-FINDR and VCA for endmember identification, with NNLS for abundance estimation). PCA yielded a single tissue-specific component – its first principal component – that effectively captured the overall tissue architecture. The remaining components displayed convoluted abundances with high uniform background signals and negligible biochemical specificity. In contrast, NMF identified three distinct endmembers, resembling major molecular species such as protein-rich, lipid-rich and nucleic acid-rich components. However, it was unable to resolve finer biochemical features beyond these primary categories. K-means clustering showed limited biochemical interpretability, segmenting the data into different tissue regions, possibly due to intensity variation instead of meaningful compositional components. More specialised unmixing techniques such as N-FINDR+NNLS exhibited serious susceptibility to artefacts, producing multiple abundance maps with isolated "noisy" pixels. Beyond one general tissue-level component, N-FINDR+NNLS failed to resolve distinct cellular structures such as nuclei or membranes. VCA+NNLS demonstrated improved performance relative to N-FINDR+NNLS, producing five distinct abundance maps that captured features such as nuclei, the apical domain of the imaged neural rosette and potentially cytoplasmic regions. Nonetheless, some artefacts and shading inconsistencies persisted – for instance, within the nuclear component (channel 6), which displayed a strong signal in the upper field that markedly diminished toward the bottom. By comparison, the developed AE offered improved robustness to artefacts and noise and produced more discriminative and interpretable abundance maps, enabling better-delineated imaging reconstruction of organoid morphology. The derived components are characterised in the following section.

5.5.3 Raman imaging of a second neural organoid cryosection

Figures 5.9A shows the six tissue-specific endmembers derived by the AE pipeline, which were characterised as lipid-, nucleic acid- and protein-rich species (full unmixing data are available in Figure D.3 in Appendix D). Visual examination of the estimated abundance maps revealed major cellular and subcellular features, including membranes, nuclei, cytoplasmic regions and apical domains of neural rosettes (Figure 5.9B). Endmember 1 (green) exhibited a lipid-rich profile with prominent peaks at 1313 cm^{-1}, 1452 cm^{-1},

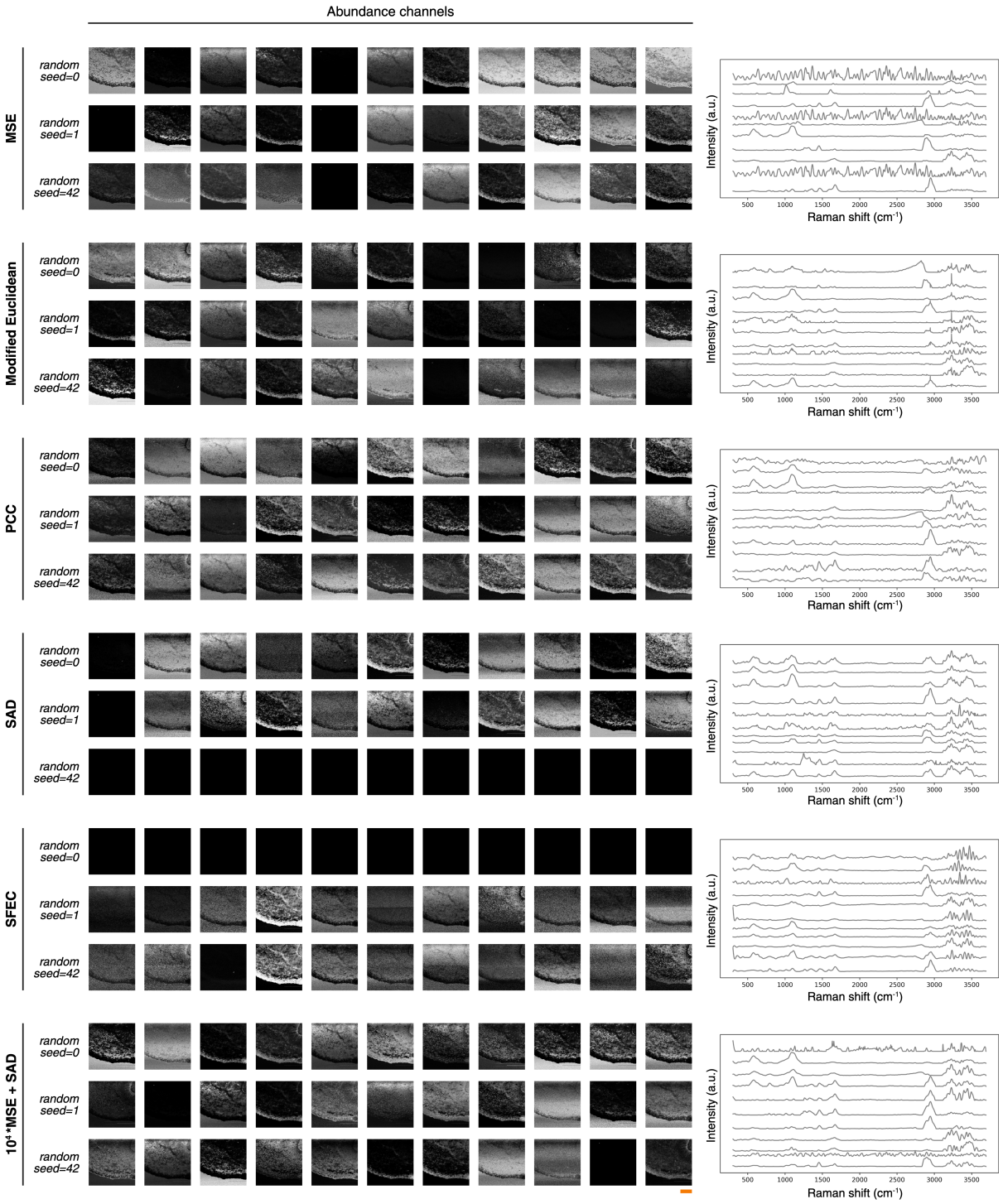


Figure 5.7: Effect of training loss on autoencoder performance. Losses tested include: mean squared error (MSE), modified Euclidean[547], Pearson's correlation coefficient (PCC), spectral angle divergence (SAD)[517], squared first-difference Euclidean cosine (SFEC)[548], and a custom loss based on a combination of MSE and SAD. Abundance images were normalised for visualisation purposes. Endmember signatures displayed on the right correspond to the analyses performed with a random seed set to 42. Scale bar: 100 µm (bottom right, marked in orange). Raw RS data acquired by Dr Daniel Reumann. Assembled in BioRender.com.

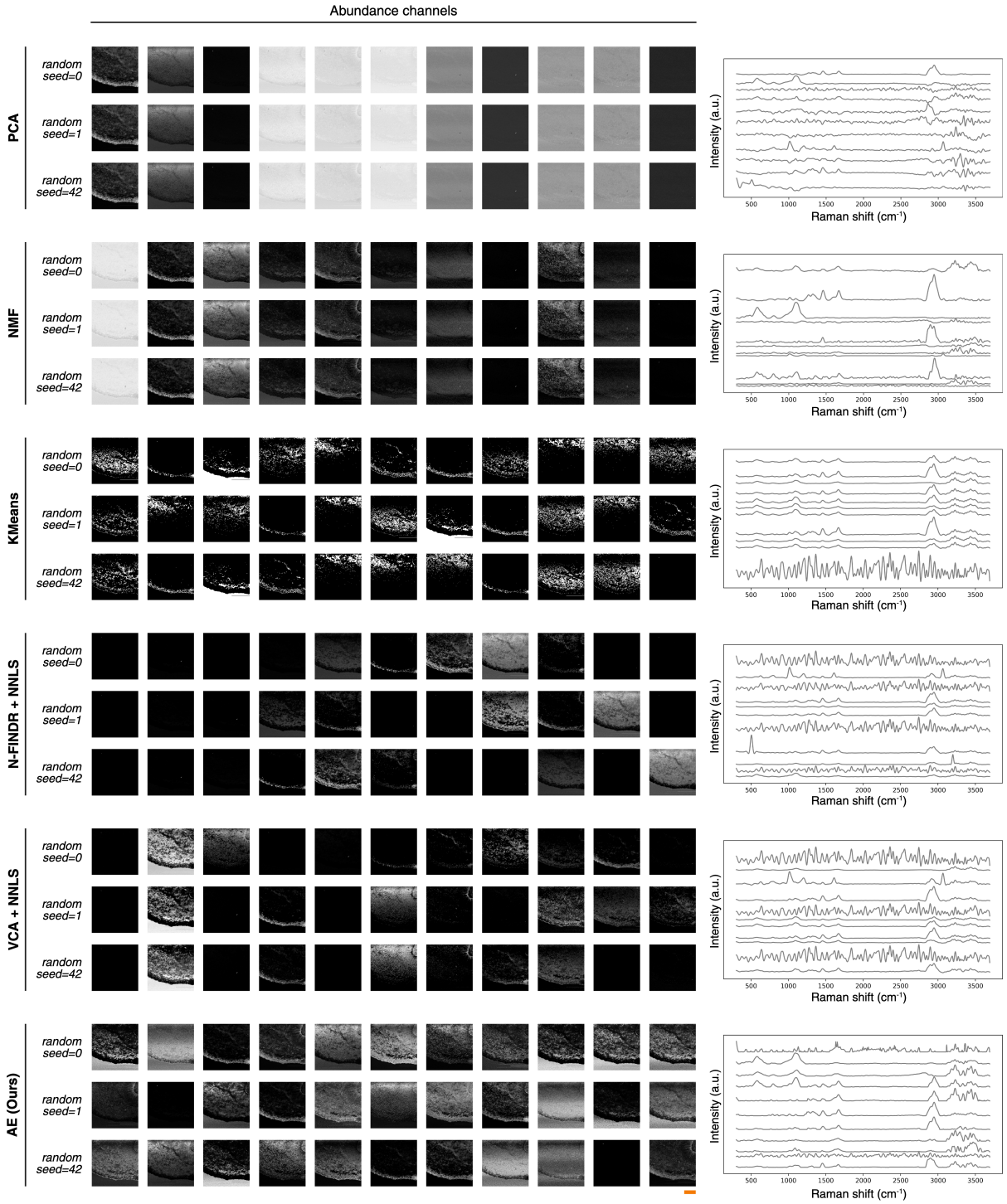


Figure 5.8: Comparative analysis of different chemometric methods on Raman imaging data from a neural organoid cryosection. Methods include: principal component analysis (PCA), non-negative matrix factorisation (NMF), k-means clustering, N-FINDR + non-negative least squares (NNLS), vertex component analysis (VCA) + NNLS, and the developed autoencoder (AE) pipeline. Abundance images were normalised for visualisation purposes. Endmember signatures displayed on the right correspond to the analyses performed with a random seed set to 42. Scale bar: 100 µm (bottom right, marked in orange). Raw RS data acquired by Dr Daniel Reumann. Assembled in BioRender.com.

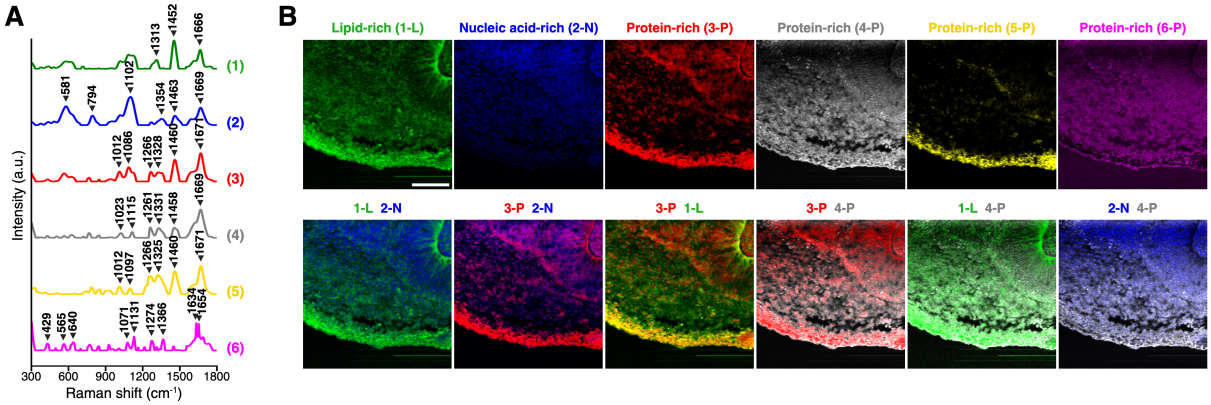


Figure 5.9: Hyperspectral unmixing analysis of a second neural organoid cryosection. Analysis of a day-30 neural organoid cryosection (*Section 2*). **A,** Derived tissue-specific endmember signatures. Endmembers scaled for visualisation purposes (maximum intensity set to 1). **B,** Corresponding abundance maps. Scale bar: 100 µm. Raw RS data acquired by Dr Daniel Reumann. Assembled in BioRender.com.

and 1666 cm^{-1}. The corresponding abundance map presented morphological architecture consistent with membrane-like structures, including individual cellular membranes and the apical domain of the present neural rosette. Endmember 2 (blue) displayed a strong nucleic acid band at 794 cm^{-1}, alongside several vibrations linked to protein species (1354 cm^{-1}, 1463 cm^{-1} and 1669 cm^{-1}), as well as bands attributed to the glass slide at (581 cm^{-1} and 1102 cm^{-1}). The abundance map revealed well-defined, spherical compartments, closely resembling cell nuclei. Endmembers 3 (red), 4 (grey) and 5 (yellow) exhibited similar spectral signatures, dominated by protein-associated peaks around ∼1015 cm^{-1}, ∼1100 cm^{-1}, ∼1265 cm^{-1}, ∼1330 cm^{-1}, ∼1460 cm^{-1}, and ∼1670 cm^{-1}. Endmember 4 showed a broad spatial distribution, likely corresponding to cytoplasmic or general background tissue signal. Endmembers 3 and 5, which showed a more pronounced 1460 cm^{-1} band, and an increased signal near the edge of the tissue section, likely due to tissue curling, as well as the basal domains of the present neural rosettes. Endmember 6 (magenta) presented a noisier signature, with prominent protein-associated peaks at 1131 cm^{-1} (C–N stretching), 1274 cm^{-1}, and 1366 cm^{-1} (tryptophan). Its spatial distribution showed substantial overlap with that of Endmember 4.

5.5.4 Quantitative validation against fluorescence microscopy

To quantitatively assess the performance of the autoencoder-based pipeline, the image similarity between Raman-based reconstructions and fluorescence microscopy images was evaluated in an overlapping organoid region. To this end, the two modalities were spatially aligned and the Pearson's correlation coefficient (Pearson) and structural similar-

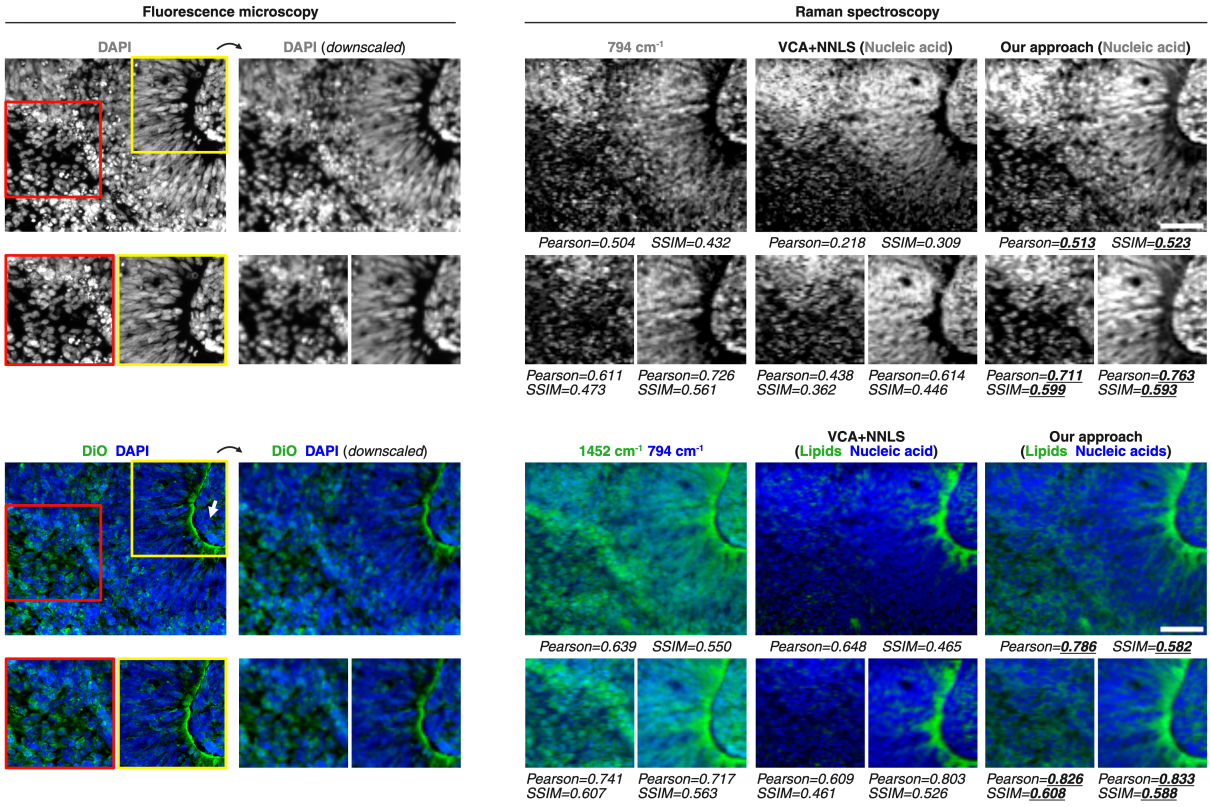


Figure 5.10: Autoencoder-based unmixing enables more accurate morphochemical Raman imaging of neural organoids as validated against fluorescence microscopy. Analysis of a day-30 neural organoid cryosection (*Section 2*). Side-by-side comparison of fluorescence microscopy images of DAPI (nuclei) and DiO (membranes) (*left*), and RS-based abundance reconstructions of nucleic acid and lipids derived with three methods for Raman imaging, including univariate analysis, unmixing analysis with VCA and NNLS, and the autoencoder pipeline. Similarity metrics, namely Pearson's correlation coefficient (Pearson) and structural similarity index measure (SSIM), were performed with respect to the respective downscaled fluorescence images. The highest achieved similarity scores are indicated by underlined bold text. Arrow indicates the lumen of the present neural rosette. Scale bar: 50 µm. Fluorescence microscopy images and raw RS data acquired by Dr Daniel Reumann. Created/assembled in BioRender.com.

ity index measure (SSIM)[549] were computed between RS-derived abundance maps of nuclei and membranes and fluorescence images of DAPI and DiO, respectively. The performance of the AE-based method was benchmarked against two conventional chemometric approaches: (i) univariate analysis, and (ii) multivariate hyperspectral unmixing using VCA and NNLS. Univariate analysis was performed using the intensity profiles at 794 cm^{-1} and 1454 cm^{-1}, which were found to be indicative of nuclear and lipid-rich membrane regions, respectively (Figure 5.9A).

Univariate analysis using the 794 cm^{-1} band yielded nuclear reconstructions which were generally aligned with the DAPI signal. However, visual inspection revealed irregular structures, which impeded the precise localisation of individual nuclei (Figure 5.10, right, first column). Major membrane structures, such as the apical domain of the imaged

rosette, were also recovered via the 1454 cm^{-1} band, yet finer membrane features were not visually discernible, obscured by high background and non-specific signal. This could be attributed to the overlap with vibrations of proteins in the same spectral range or non-specific intensity variations due to acquisition or preprocessing settings, underscoring the need for more robust multivariate chemometric approaches[550]. Multivariate unmixing analysis with VCA and NNLS showed low reconstruction accuracy, with further reduced Pearson and SSIM scores compared to univariate analysis for both nuclei and membranes (Figure 5.10, right, second column). Visually, nuclear reconstructions were similarly non-distinct, with poorly defined nuclei. Moreover, while major membrane structures like the apical domain of the present rosette were recovered with higher specificity, finer membrane features were consistently undetected. By contrast, as shown in Figure 5.10 (right, third column), the autoencoder-based pipeline achieved the highest reconstruction accuracy across all evaluations – both quantitatively, with superior Pearson and SSIM scores, and qualitatively, with more interpretable and biologically relevant output. Nuclear regions were recovered with higher accuracy, with visually better-defined nuclei and lower non-specific signal. Membrane-associated structures were also more clearly delineated, enabling the visualisation of both large-scale features such as apical domains and individual cellular membranes.

5.5.5 Robustness to experimental noise

Furthermore, the AE approach demonstrated improved robustness to experimental noise, maintaining consistent reconstruction quality under reduced denoising conditions. In comparison, the performance of conventional methods deteriorated markedly, yielding noisier, less interpretable spectral components and abundance profiles, or failing to converge to a solution during unmixing altogether (Figure 5.11).

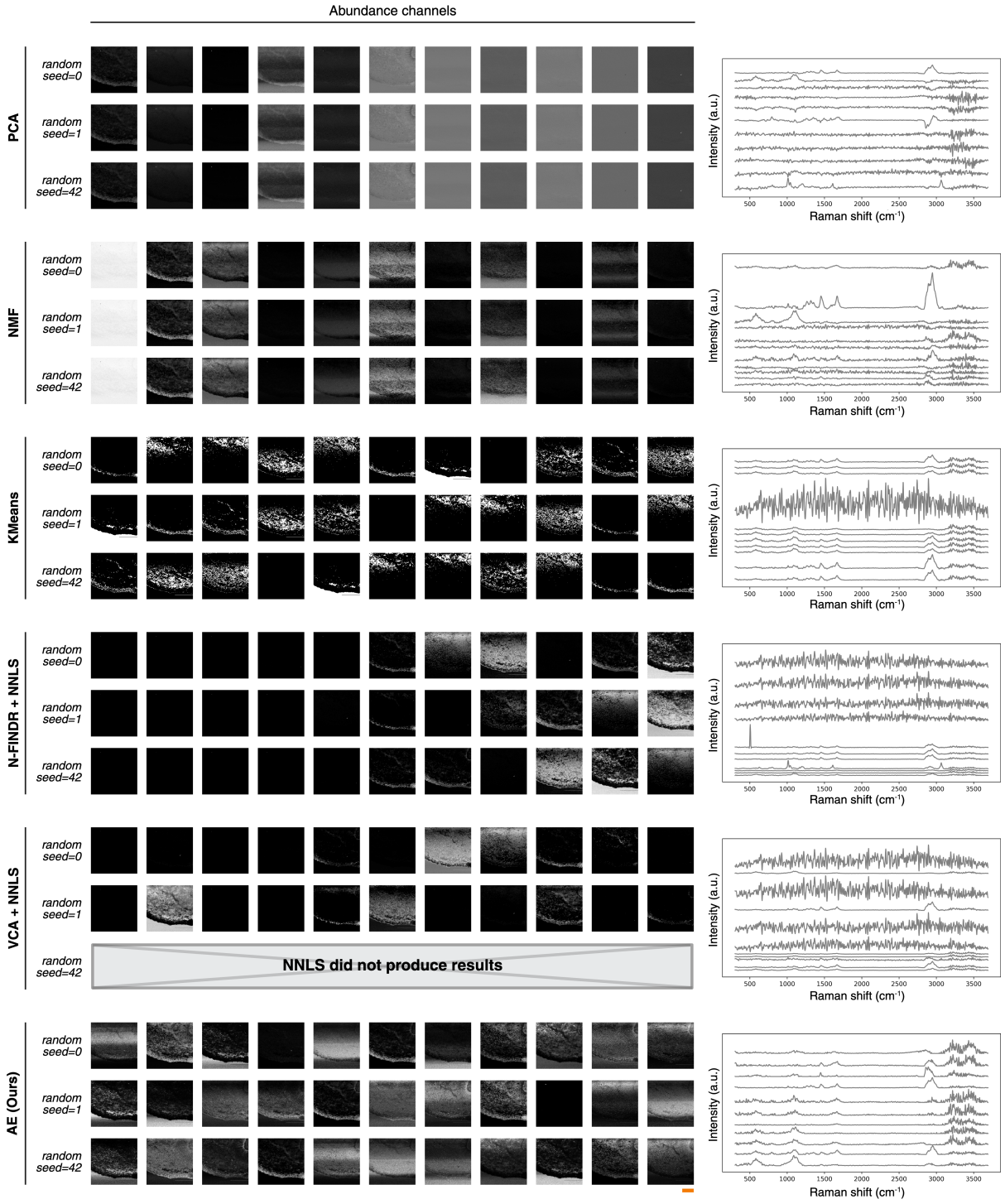


Figure 5.11: Comparative analysis of different chemometric methods in the presence of higher degrees of experimental noise. Higher noise levels were simulated by applying a more conservative denoising procedure with a second-order Savitzky-Golay filter with a window size of 7. Methods include: principal component analysis (PCA), non-negative matrix factorisation (NMF), k-means clustering, N-FINDR + non-negative least squares (NNLS), vertex component analysis (VCA) + NNLS, and the developed autoencoder (AE) pipeline. One of the analyses with VCA+NNLS failed to converge, with the NNLS implementation terminating after exceeding the maximum number of iterations. Abundance images were normalised for visualisation purposes. Endmember signatures displayed on the right correspond to the analyses performed with a random seed set to 42. Scale bar: 100 µm (bottom right, marked in orange). Raw RS data acquired by Dr Daniel Reumann. Assembled in BioRender.com.

5.6 Biochemical imaging in intact neural organoids at subcellular resolution

After demonstrating that the developed method achieves accurate subcellular imaging of 2D cryosectioned neural organoid tissue, higher-resolution Raman imaging was performed of a tissue region within an intact 3D neural organoid. The ability to resolve and visualise subcellular structures in intact organoid tissues in their native environment is critical for interrogating important neurodevelopmental processes, such as nuclear positioning, polarity establishment and lumen formation, and studying their effect on brain tissue organisation, architecture and function.

To this end, a larger 400 × 346 Raman imaging scan was acquired with 1 µm pixel size from an intact day-10 neural organoid sample (Figure 5.12A). Using the Raman imaging pipeline, the data were analysed and six tissue-specific components were identified. The spectral endmembers and corresponding abundance maps are presented in Figure 5.12B and Figure 5.12C, respectively (full unmixing data are available in Figure D.4 in Appendix D). Endmember 1 (red) exhibited a lipid-rich profile with prominent peaks at 1077 cm^{-1} (C–C and C–O stretching, phospholipids), 1305 cm^{-1}, 1448 cm^{-1} and 1657 cm^{-1}. Its abundance map outlined membrane-like structures, along with more confined bright spots resembling lipid droplets. Endmember 2 (grey) was characterised by vibrations associated with both protein and lipid species, with peaks at 1000 cm^{-1}, 1124 cm^{-1} (C–C and C–N stretching), 1249 cm^{-1}, 1311 cm^{-1}, 1454 cm^{-1} and 1654 cm^{-1}. Its abundance was observed in regions partially overlapping with membranes, as well as in isolated domains, possibly indicating cytoplasmic or extracellular matrix compartments. Endmember 3 (blue) was marked by strong nucleic acid bands at 790 cm^{-1} and 1095 cm^{-1}, alongside protein-associated bands at 1005 cm^{-1}, 1249 cm^{-1}, 1328 cm^{-1}, 1455 cm^{-1} and 1659 cm^{-1}. As in previous analyses, the abundance map revealed well-defined spherical compartments consistent with individual cell nuclei. Endmember 4 (green) presented a distinct protein-rich signature with peaks at 1005 cm^{-1}, 1251 cm^{-1}, 1454 cm^{-1} and 1662 cm^{-1}, along with vibrations associated with ring breathing modes in nucleic acid bases at 1573 cm^{-1}. The abundance map highlighted cellular-level compartments resembling cytoplasmic regions, as well as small punctate structures within nuclei with high signal intensity, potentially corresponding to biomolecular condensates. Endmember 5 (yellow) displayed a mixed profile comprising bands linked to proteins (1259 cm^{-1}, 1326 cm^{-1}, 1453 cm^{-1} and 1662 cm^{-1}) and nucleic acid (1095 cm^{-1}). This component appeared more diffusely distributed, with substantial spatial overlap with endmember 3. Endmember 6 (purple) presented a noisier signature, with carbohydrate-associated bands at 493 cm^{-1} (glycogen), 850 cm^{-1} (C–O–C skeletal mode, glycogen) and 938 cm^{-1} (C–C stretching), along with peaks linked to proteins and nucleic acids at 754 cm^{-1} and

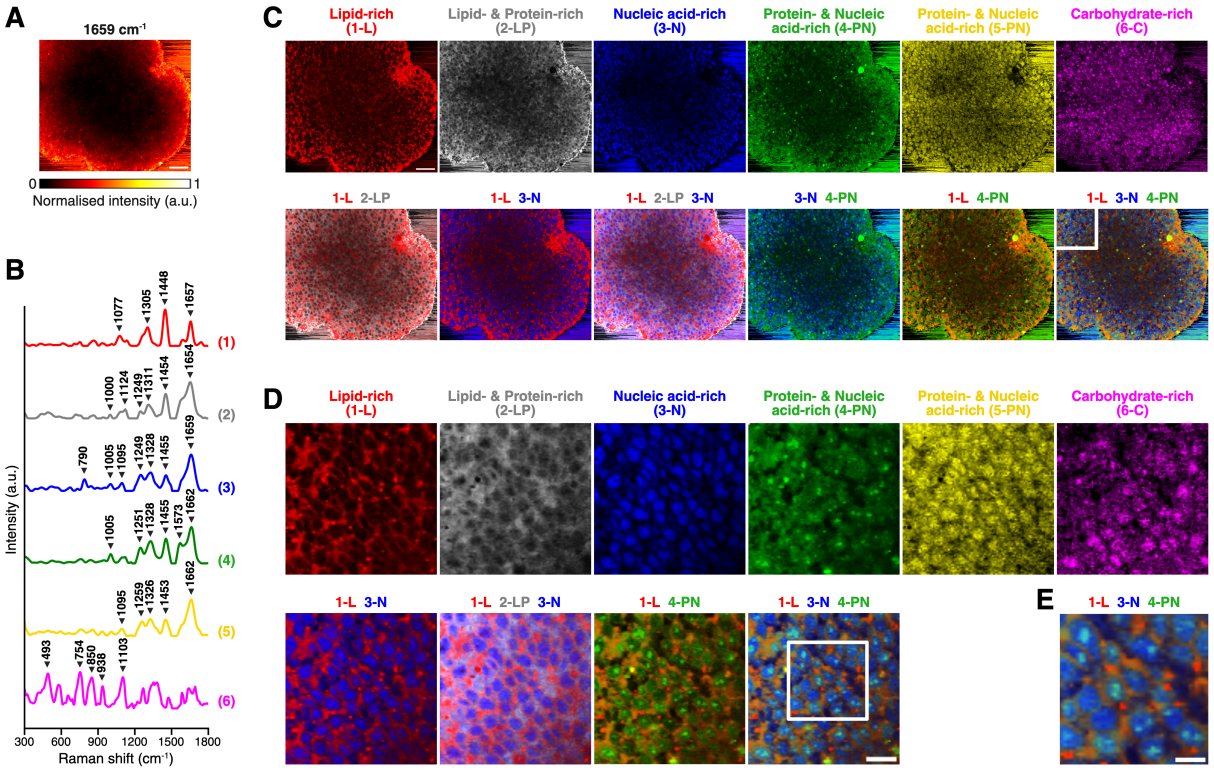


Figure 5.12: Biochemical imaging in intact neural organoids at subcellular resolution. A, Snapshot of collected Raman imaging data from an intact day-10 neural organoid sample, showing the intensity profile at the 1659 cm^{-1} band (proteins and lipids). Scale bar: 50 µm. Raw RS data acquired by Dr Ruoxiao Xie and Xiaoyu Zhao. **B–E,** Results of hyperspectral unmixing analysis. **B,** Derived tissue-specific endmember signatures. Endmembers scaled for visualisation purposes (maximum intensity set to 1). **C,** Corresponding abundance maps. Scale bar: 50 µm. **D,** Zoomed-in abundance images of a selected area (area marked in **C**, bottom right panel). Scale bar: 20 µm. **E,** Further zoomed-in abundance image (area marked in **D**). Scale bar: 10 µm. Raw RS data acquired by Dr Ruoxiao Xie and Xiaoyu Zhao. Assembled in BioRender.com.

1103 cm^{-1}. The abundance overlapped with endmember 4, particularly around nuclear condensates, potentially reflecting glycoproteins within cell nuclei.

To further inspect the spatial reconstructions produced by the developed imaging pipeline, magnified fields from selected regions within the imaged organoid tissue were examined (Figures 5.12D–E). The resulting reconstructions revealed consistent, well-defined cellular features, which delineated individual cells and subcellular structures resembling nuclei, cytoplasmic regions, cellular membranes and nuclear condensates.

These results demonstrate the effectiveness of the developed imaging platform across scales and resolutions, which resolved complex biochemical heterogeneity and interrogated key morphological features at cellular and subcellular scales in a single, label-free imaging modality.

5.7 Volumetric *in situ* imaging of a neural rosette within an intact neural organoid

Extending the subcellular imaging analysis to 3D, volumetric *in situ* imaging of a neural rosette structure within an intact neural organoid was performed using the developed pipeline.

Neural rosettes are important organisational units during neural organoid development, which recapitulate key aspects of neural tube formation in the embryonic central nervous system[551]. They typically form within the first two weeks of neural induction and appear as radially organised, layered structures composed primarily of neuroepithelial cells at early stages. As development progresses, these progenitor cells give rise to differentiated neural lineages, including neurons and glial cells. To illustrate this, fluorescence microscopy of a day-20 neural organoid tissue section was performed, with markers for nuclei (DAPI), neurons (MAP2), neural progenitor cells (SOX2) and apical domain (PKCζ). The resulting images, shown in Figure 5.13A, revealed multiple neural rosettes with characteristic radial structure. An abundance of neural progenitor cells was observed densely organised within the rosettes, with surrounding layers of differentiating neurons.

Following this, the Raman-based imaging platform was used to investigate whether it could visualise and analyse the structure of a neural rosette without the need for tissue sectioning and exogenous labelling. To this end, a volumetric RS scan of a neural rosette within an intact day-20 neural organoid was acquired, consisting of five imaging slices in z (organoid depth), with each slice measuring 95×95 pixels with a pixel size of 2 µm. Figure 5.13B displays the intensity profile at the 1659 cm^{-1} peak (proteins and lipids) of one of the collected layers, outlining a characteristic neural rosette structure. The autoencoder-based unmixing analysis yielded six tissue-specific components (Figure 5.13C–D; full unmixing data are available in Figure D.5 in Appendix D). Endmember 1 (red) presented a lipid-rich profile with prominent peaks at 874 cm^{-1} (phospholipids), 1295 cm^{-1} (CH_2 deformation), 1445 cm^{-1} and 1659 cm^{-1}. The corresponding abundance map highlighted structures resembling cellular membranes and lipid bodies. Endmember 2 (grey) contained features associated with proteins and lipids, with peaks at 1005 cm^{-1}, 1300 cm^{-1}, 1450 cm^{-1}, 1550 cm^{-1} (C=C stretching, tryptophan) and 1659 cm^{-1}. Its signal overlapped with that of endmember 1, appearing in membrane-like regions, but also extended into the cell interior, possibly indicating cytoplasmic and structural proteins. Endmember 3 (blue) showed a distinct nuclear signature, with nucleic acid-associated vibrations at 787 cm^{-1} and 1095 cm^{-1}, along protein bands at 1254 cm^{-1}, 1336 cm^{-1}, 1453 cm^{-1} and 1666 cm^{-1}. This component exhibited strong nuclear specificity, outlining well-defined spherical compartments consistent with

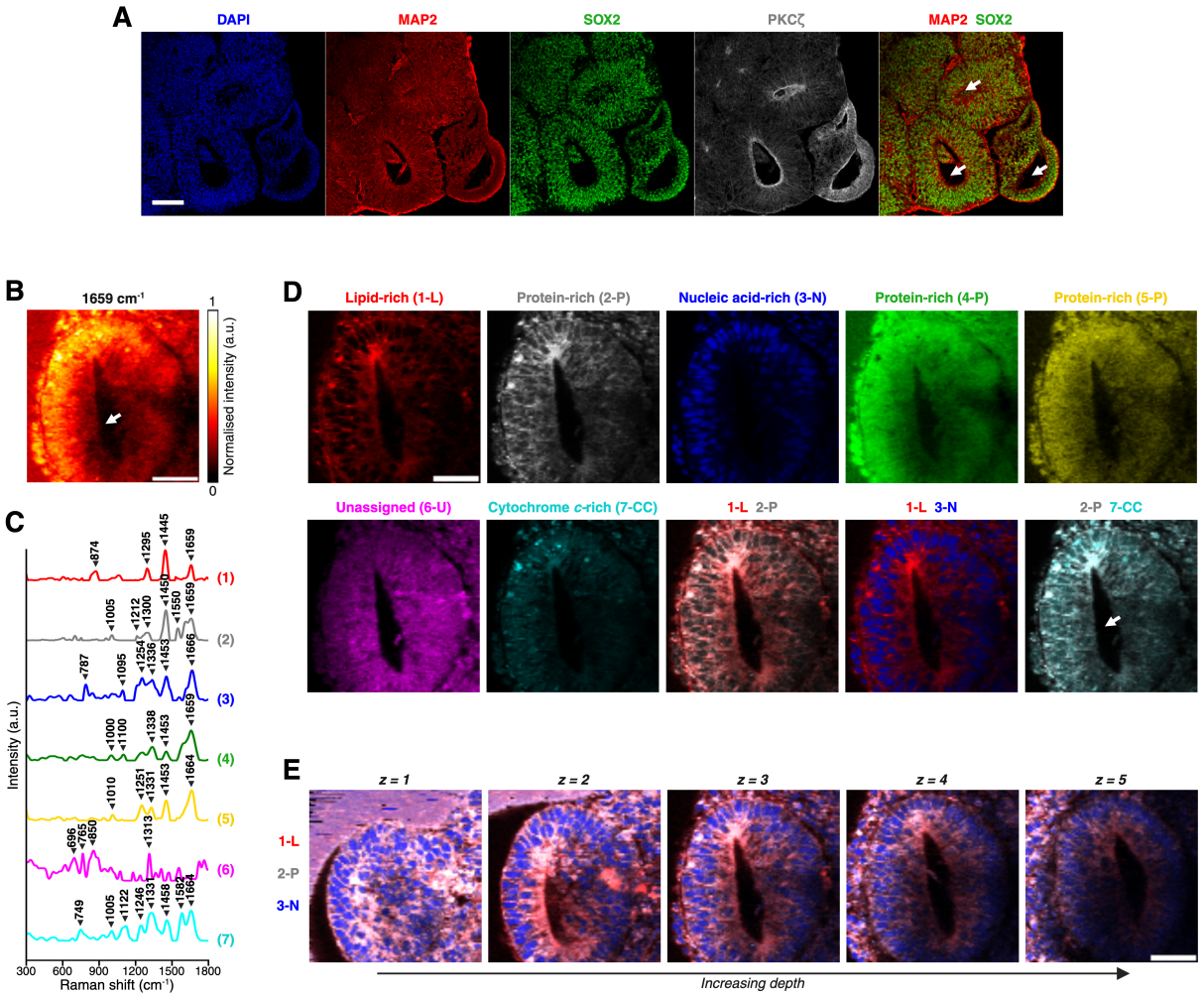


Figure 5.13: Volumetric *in situ* biochemical imaging of a neural rosette within an intact neural organoid. A, Fluorescence microscopy images of a day-20 neural organoid tissue section, showing characteristic neural rosette structures, with markers for nuclei (DAPI), neurons (MAP2), neural progenitor cells (SOX2) and apical domains (PKCζ). Arrows indicate the lumens of developed neural rosettes. Scale bar: 100 µm. Images acquired by Dr Ruoxiao Xie. **B,** Snapshot of collected volumetric Raman imaging data from an intact day-20 neural organoid sample (selected z-layer), showing the intensity profile at the 1659 cm^{-1} band (proteins and lipids). Arrow indicates the lumen of the present neural rosette. Scale bar: 50 µm. Raw RS data acquired by Dr Ruoxiao Xie and Xiaoyu Zhao. **C–E,** Results of hyperspectral unmixing analysis. **C,** Derived tissue-specific endmember signatures. Endmembers scaled for visualisation purposes (maximum intensity set to 1). **D,** Corresponding abundance maps for a representative z-layer. Arrow indicates the lumen of the present neural rosette, as also shown in **B**. Scale bar: 50 µm. **E,** Image reconstructions of selected endmember components at different depths in z. Scale bar: 50 µm. Assembled in BioRender.com.

individual cell nuclei. Endmembers 4 (green) and 5 (yellow) shared similar spectral features dominated by protein-related peaks around ∼1005 cm^{-1}, ∼1100 cm^{-1}, ∼1255 cm^{-1}, ∼1335 cm^{-1}, ∼1453 cm^{-1} and ∼1660 cm^{-1}, suggesting general protein composition. Endmember 6 (magenta) displayed a less conventional, noisier profile with peaks at 696 cm^{-1} (potentially related to cholesterol), 765 cm^{-1} (pyrimidine ring breathing in nucleic acids), 850 cm^{-1} (C–O–C skeletal mode in carbohydrates, glycogen) and 1313 cm^{-1} (CH_2/CH_3

twisting in lipids and collagens, or guanine in nucleic acid). This component was generally broadly distributed, with some confined regions with higher abundance intensity. Lastly, endmember 7 (cyan) showed a signature well-aligned with mitochondria, featuring peaks associated with cytochrome *c* at 749 cm^{-1}, 1122 cm^{-1} and 1582 cm^{-1}, alongside protein-related bands at 1005 cm^{-1}, 1246 cm^{-1}, 1331 cm^{-1}, 1458 cm^{-1} and 1664 cm^{-1}. This component was observed in cytoplasmic regions, in proximity to nuclei, consistent with intracellular localisation of mitochondria.

To examine the three-dimensional structure of the scanned rosette, the abundance maps of the derived lipid-rich (endmember 1), protein-rich (endmember 2) and nucleic acid-rich (endmember 3) components were overlaid across all five z-layers. The resulting images, presented in Figure 5.13E, demonstrated that the proposed imaging pipeline can effectively reconstruct key subcellular features – such as individual cellular membranes and nuclei – across all five imaging scans, enabling the interrogation of cellular architecture and organisation within the intact neural rosette in both 2D and 3D.

5.8 Tracing changes in biochemical composition during early neurodevelopment in intact neural organoids

During maturation, neural organoids recapitulate key stages of early human brain development, including the formation of distinct progenitor zones and the emergence of layered neural architectures[11,12,377,552]. To better understand the biochemical dynamics accompanying tissue differentiation and organisation, the developed imaging pipeline was used to trace changes in organoid composition during neural organoid development.

To this end, Raman imaging data were acquired from organoid samples at five developmental time points – days 5, 10, 20, 30 and 40 (Figure 5.14). In total, 72 Raman imaging scans were collected from organoids cultured in three independent culture batches using stem cells at different passage numbers. Scan numbers ranged between 11 and 17 per time point (Figure 5.15A). Each scan represented a 20×160 image across the outermost tissue layers, which normally contain the most metabolically active and viable cells[553,554]. Measurements were acquired with a lateral pixel size of 5 µm in the x-y plane and an axial step size of 10 µm along the z-axis, thereby imaging up to 800 µm across and 200 µm deep into the organoid samples.

Scans were preprocessed and hyperspectral unmixing analysis was performed. For quantitative analysis, manual background segmentation was performed using the derived abundance reconstructions as a reference. The obtained binary masks were used to (i) discard

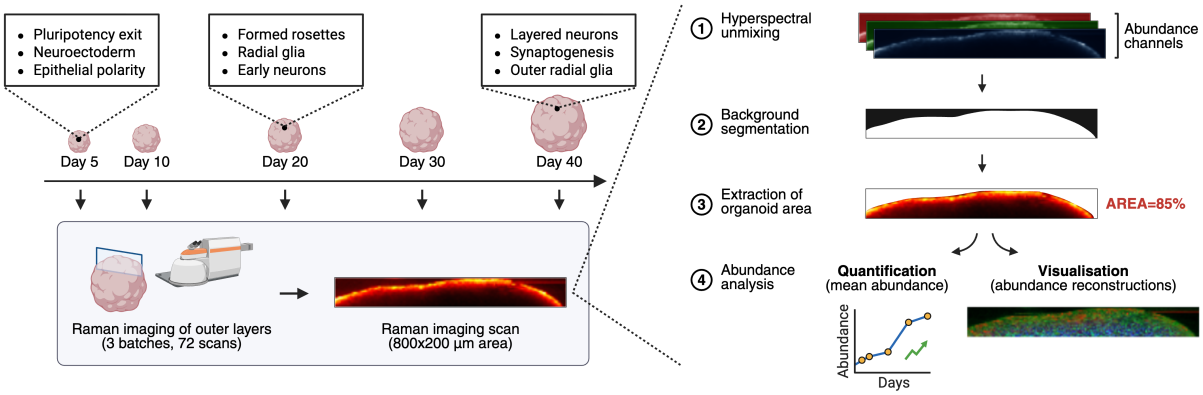


Figure 5.14: Timepoint analysis of intact neural organoids: Study design. The schematic illustrates the conducted timepoint study, depicting data acquisition (*left*) and analysis (*right*). Created in BioRender.com.

background pixels by setting abundance values in the background area to zero for all channels, and (ii) estimate the relative tissue portion in each scan. Scans with an organoid area below 30% were not used for the subsequent quantitative analysis to avoid potential biasing towards components in the outer tissue layers. In total, six scans out of the available 72 were discarded (Figure 5.15B). Finally, the mean abundance of endmembers was calculated for each scan and normalised by the organoid area.

Hyperspectral unmixing analysis yielded eight tissue-specific endmember species, shown in Figure 5.16A (full unmixing data are available in Figure D.6 in Appendix D). The corresponding abundance profiles were analysed to examine relative biochemical trends across developmental groups. Figure 5.16B displays the mean abundance of each component across time points, after background removal and area normalisation. Endmember 1 (red) exhibited a distinct lipid-rich spectral profile, characterised by prominent peaks at 1303 cm^{-1}, 1448 cm^{-1} and 1656 cm^{-1}, and appeared broadly distributed with high abundance across all organoids and time points (Figure 5.17A). Still, distinct spatial and temporal variations in its distribution across time points were observed. On day 5, the signal appeared relatively homogeneous, with slightly elevated intensity in small compartments near the outer layers. By day 10, its intensity increased markedly, particularly around the inner layers and surface. The high abundance in early time points may reflect shifts in lipid metabolism during early neurogenesis, in line with previous findings that lipid accumulation is essential for neural stem/progenitor cell activity and proliferation[555]. Following this, a sharp decline in signal was observed in day 20 organoids, where discrete, localised regions of elevated intensity emerged, likely corresponding to developed neural rosette structures. Between days 30 and 40, the signal progressively recovered, once again displaying a more uniform and widespread distribution. This gradual recovery may be linked to the expansion and functional maturation of neuronal populations,

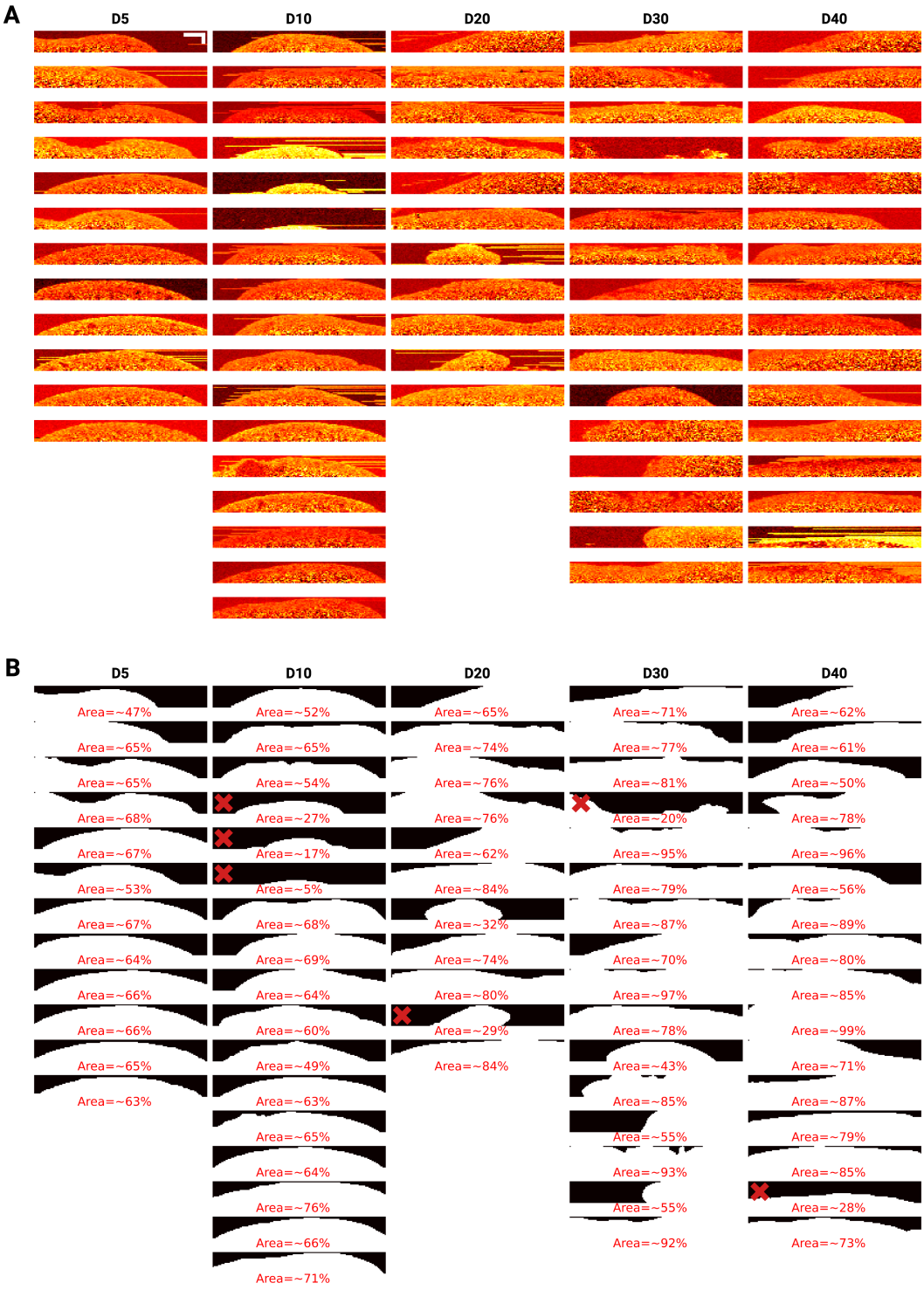


Figure 5.15: Preparation of timepoint Raman data for quantitative analysis. A, Intensity slices across the 1659 cm^{-1} band (linked to proteins and lipids) of the collected Raman imaging scans. Images were normalised for visualisation purposes. Scale bar: 100 µm (top left). **B,** Annotated organoid masks, along with estimated tissue portions. Scans with tissue areas smaller than 30% were not included in quantitative analysis (marked with a red cross). Raw RS data acquired by Dr Ruoxiao Xie and Xiaoyu Zhao. Assembled in BioRender.com.

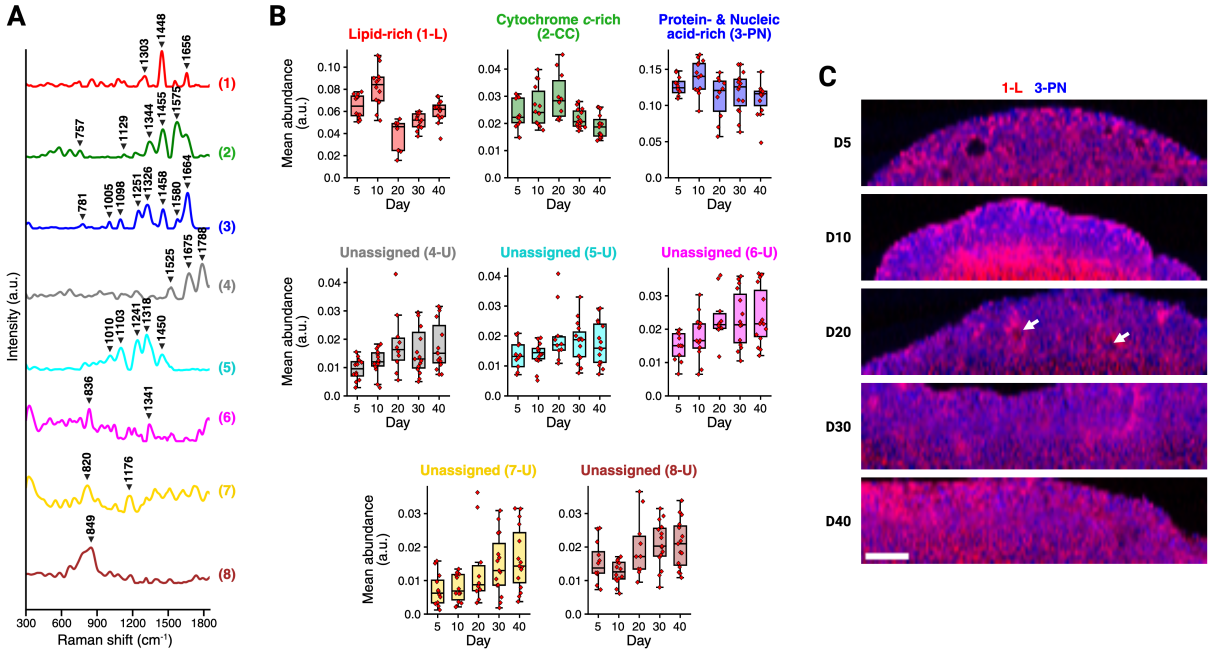


Figure 5.16: Tracing changes in biochemical composition during early neurodevelopment in intact neural organoids. Results of hyperspectral unmixing analysis. **A,** Derived tissue-specific endmember signatures. Endmembers scaled for visualisation purposes (maximum intensity set to 1). **B,** Quantitative analysis of the corresponding abundances over time. Sample sizes: day 5 ($n = 12$), day 10 ($n = 14$), day 20 ($n = 10$), day 30 ($n = 15$), day 40 ($n = 15$) from $N = 3$ independent organoid culture batches. **C,** Abundance reconstructions of selected organoid samples across the five timepoint groups. Arrows indicate structures resembling developed neural rosettes. Scale bars: 100 µm. Raw RS data acquired by Dr Ruoxiao Xie and Xiaoyu Zhao. Created/assembled in BioRender.com.

which rely on lipids not only as essential structural components of membranes but also as signalling molecules with key regulatory functions[556,557]. This is further supported by the observation that lipid-rich endmembers are well co-localised with TUJ1-positive regions – a marker of neurons (see Figures 5.4B and Figures 5.5A). In contrast, endmember 2 (green) presented a mixed spectral profile, with peaks linked to cytochrome *c* (757 cm^{-1}, 1129 cm^{-1}, 1575 cm^{-1}) and proteins (1344 cm^{-1}, 1455 cm^{-1}). This signature suggested regions enriched in mitochondria, and showed an abundance profile with an inverted U-shaped trajectory, gradually increasing from day 5 to day 20, followed by a decline at later time points. Despite these temporal changes, the spatial distribution remained generally uniform across the organoid tissue throughout the sampled time course. Endmember 3 (blue) comprised vibrational features associated with nucleic acids (781 cm^{-1}, 1098 cm^{-1}, 1580 cm^{-1}) and proteins (1005 cm^{-1}, 1251 cm^{-1}, 1326 cm^{-1}, 1458 cm^{-1}, 1664 cm^{-1}), likely corresponding to cytoplasmic and nuclear regions. Its abundance was generally stable across developmental stages, with a slight downward trend. A similar progression was reported by Bruno et al., who observed a gradual decrease in nucleic acid and protein features in maturing cortical organoids[508]. Visually, this component was highly abundant across all organoid samples, with a generally uniform distribution throughout organoids,

and slightly elevated intensity in the outer tissue layers (Figure 5.17B).

Endmembers 4–8 exhibited less conventional spectral profiles, with lower relative abundance predominantly present in deeper organoid regions (Figure D.6C). They shared similar developmental trajectories, each showing a gradual increase in abundance over time. While prominent peaks within these endmembers allow for tentative biochemical assignments, further validation is required to determine whether these components reflect representative biological features, or traces arising from higher noise levels in deeper tissue due to depth-dependent signal attenuation. As such, the interpretation of these endmembers warrants extra caution. To reflect this uncertainty, these five endmembers are left unassigned in the analysis. Nevertheless, they are retained to enable accurate calculation of the abundance of components 1–3, which represent the primary, well-characterised biological signals in this study. Endmember 4 (grey) featured two bigger peaks at 1675 cm^{-1} and 1788 cm^{-1}, likely attributable to protein species, and a smaller peak 1525 cm^{-1}, potentially indicating contributions from carotenoids (C–C and C=C stretching). Endmember 5 (cyan) was characterised by distinct protein-related peaks at 1010 cm^{-1}, 1103 cm^{-1}, 1241 cm^{-1}, 1318 cm^{-1} and 1450 cm^{-1}. Endmembers 6 (magenta) and 7 (yellow) exhibited noisier profiles, featuring multiple, less-defined peaks. Some of the more pronounced vibrations included peaks at 820 cm^{-1}, 836 cm^{-1}, 1176 cm^{-1} and 1341 cm^{-1}, all related to proteins. Lastly, endmember 8 (brown) showed a single dominant peak at 849 cm^{-1}, potentially corresponding to the C–O–C skeletal vibrations of saccharides. Bruno et al. reported a similar peak around 850 cm^{-1} as a prominent predictive marker of cortical organoid maturation, showing a significant, progressive increase between weeks 6 and 20[508]. The authors attributed this feature to glycogen and associated its rise with metabolic changes accompanying neuronal differentiation.

To illustrate the dynamic morphochemical composition of neural organoids during early neurodevelopment, spatial reconstructions of selected biochemical components were further generated across organoid samples at different developmental stages. Specifically, abundance images were produced without background removal, after scaling the data by a factor of 2 in z (organoid depth) to achieve an equal effective pixel size of 5 µm in x-y and z, and thus provide a more accurate depiction of the organoid morphology and dimensions. Figure 5.16C displays the created reconstructions of representative samples from each of the five time points, highlighting the two selected components associated with lipids, and proteins and nucleic acids, respectively (see Figure D.6C in Appendix D for other components). These visualisations enabled us to trace the evolving spatial organisation and architecture of neural organoids across their development, providing insights into neural organoid maturation and further supporting the quantitative trends observed in Figure 5.16B.

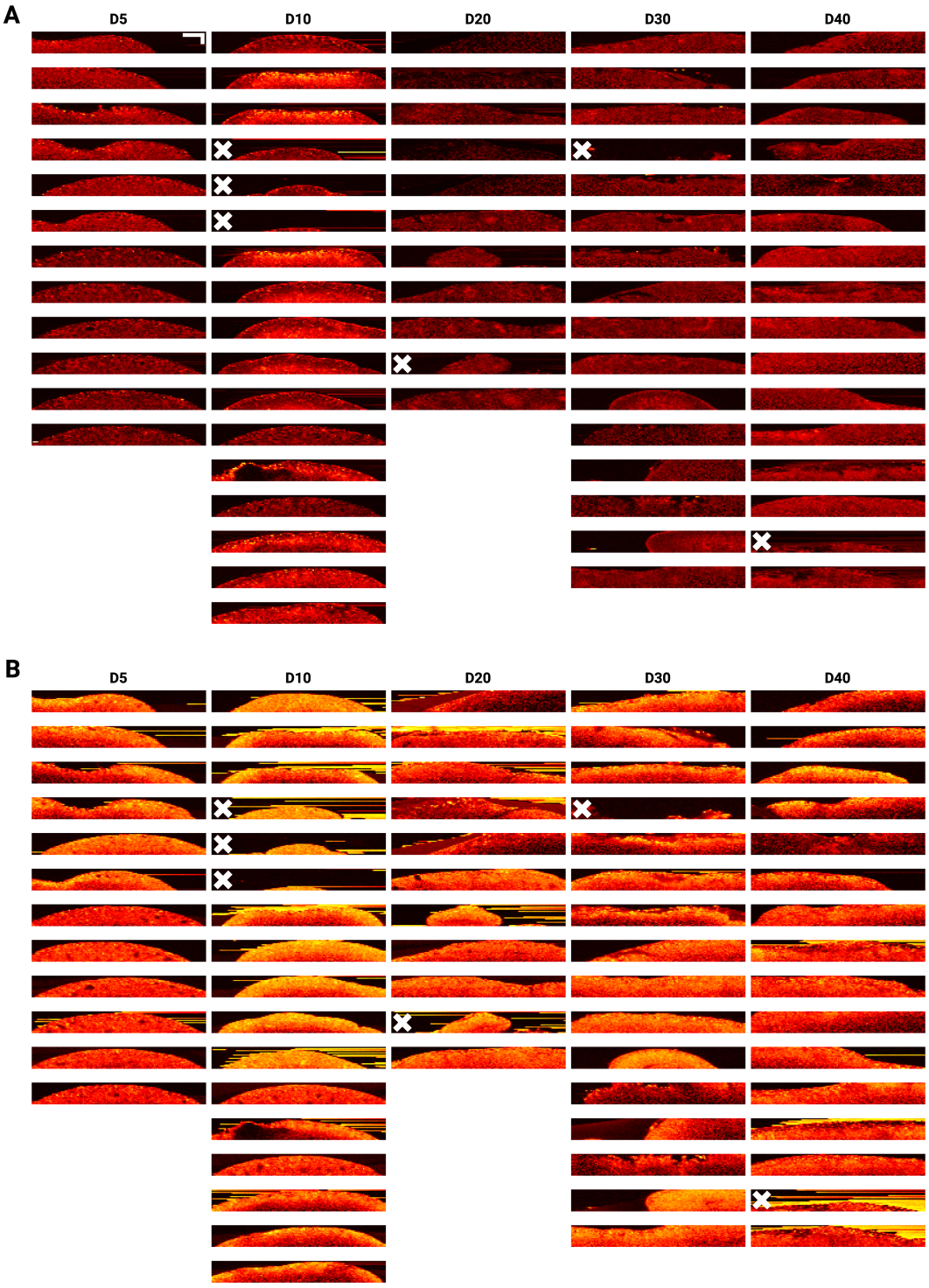


Figure 5.17: Abundance images of the two most abundant endmembers across timepoint Raman spectroscopy scans. A, Abundance images of endmember 1 (lipid-rich). Scale bar: 100 µm (top left). **B,** Abundance images of endmember 3 (protein- and nucleic acid-rich). Images were not normalised. Scans excluded from quantitative analysis are marked with a white cross. Raw RS data acquired by Dr Ruoxiao Xie and Xiaoyu Zhao. Assembled in BioRender.com.

These results serve as proof of concept of the feasibility of label-free, RS-based biomolec-

ular tracking of spatiotemporal dynamics in intact neural organoids during development.

5.9 Conclusion

In this chapter, a method for non-invasive, label-free organoid imaging was developed based on Raman microspectroscopy combined with unsupervised autoencoder-based hyperspectral analysis. Experiments showed that the proposed approach is effective at discerning key biochemical components in unlabelled Raman spectroscopy measurements, enabling 2D and 3D visualisation of structural and morphological features in both cryosectioned and intact neural organoids and at different scales and resolutions. Using the imaging platform, structural and biochemical features of early neurodevelopment in neural organoids were studied, and the ability to image neural rosettes within intact organoid samples in three dimensions was demonstrated. Furthermore, changes in the biomolecular composition of intact neural organoids during early developmental stages were interrogated, observing distinct spatiotemporal dynamics as organoids mature in major biomolecular classes, including lipids, proteins and nucleic acids. By eliminating the need for destructive preparation steps and external labels, the developed workflow offers simpler sample preparation that does not compromise organoid integrity, with potential for live-organoid analysis.

Data Availability

The research data supporting this chapter will be made available upon publication of the associated manuscript.

Code Availability

The codebase and scripts developed for the analyses presented in this chapter will be made available upon publication of the associated manuscript. To improve accessibility, a user-friendly, no-code platform will also be provided.

Chapter 6

Discussion

6.1 Thesis achievements

This thesis addresses key challenges in Raman spectroscopy and organoid analysis by advancing the methodological and computational toolkit of RS chemometrics and demonstrating the potential of RS combined with deep learning for non-invasive, label-free biochemical imaging in neural organoids. The main contributions are:

- **Open-source software platform for RS data analysis.** Released RamanSPy, an integrated, open-source Python framework that standardises RS workflows and improves accessibility, reproducibility and interoperability. It streamlines day-to-day analysis, supports the development of methods and protocols, and facilitates the integration of advanced data analytics.
- **Deep learning framework for RS chemometrics.** Developed autoencoder models for hyperspectral unmixing that achieved improved qualitative and quantitative analysis across synthetic and experimental data compared to standard baselines (e.g. VCA+NNLS). The models and curated benchmark datasets were openly released to facilitate fair validation, comparison and future innovation in RS chemometrics.
- **Platform for label-free organoid imaging via non-invasive RS.** Established an unsupervised RS imaging pipeline that integrates Raman microspectroscopy with deep learning-based hyperspectral unmixing for spatially resolved analysis in complex organoid samples. This enabled non-invasive, label-free biochemical mapping in intact neural organoids at subcellular resolution.
- **Volumetric imaging and spatiotemporal analysis in intact neural organoids.** Demonstrated *in situ* volumetric imaging of a neural rosette within a neural organoid, and investigated changes in biochemical composition during early developmental stages, revealing spatiotemporal variations in lipids, proteins and nucleic acids.

Overall, this thesis demonstrates the feasibility of RS combined with deep learning for label-free biochemical imaging and analysis in complex, three-dimensional neural organoid systems. It establishes a versatile methodological framework for high-content biochemical phenotyping that enables non-invasive interrogation of chemical composition, spatial organisation and temporal variations in intact organoids with minimal sample preparation and without prior knowledge of specific targets and the use of exogenous labels. In doing so, the work positions RS as a practical characterisation technique for organoid research and related areas, with potential for live-organoid analysis under physiologically relevant conditions.

6.2 Broader significance

By expanding the analytical toolkit for organoid analysis, this work opens new opportunities for studying brain development, neurological disorders and drug discovery. The application of RS imaging to neural organoid models could enable further research into self-organisation and tissue morphogenesis during early neurodevelopment, disease aetiology and pathogenesis in patient-derived models, and drug response and distribution analysis within intact tissues. The developed infrastructure and the downstream applications it supports represent an important step toward establishing organoids as experimental and preclinical platforms, with the potential to deepen our understanding of human brain development, neurological diseases, and accelerate translational discovery.

Beyond its biological applications, this thesis also contributes to more open and reproducible research practices in RS. The released software and datasets offer accessible resources that support standardisation, methodological innovation and validation. By reducing reliance on specialised, proprietary tools, the work broadens access to spectroscopy-based analysis for researchers both within and beyond traditional spectroscopy. In this way, it encourages interdisciplinary research at the interface of spectroscopy, chemical analysis and machine learning, fostering collaboration and community-driven innovation.

A testament to this impact is the rapid adoption of the developed tools across scientific research. Since its release, RamanSPy has gained substantial traction and supported research across diverse areas, including cell biology[558], cancer research[559], microbiology[560], materials science[561], environmental sciences[562,563], mineralogy[564], geology[565,566] and planetary science[567,568]. It has also facilitated methodological advances in both software[569,570] and hardware development[571]. As of January 2026, it has been downloaded close to

50,000 times* and has attracted strong community engagement on GitHub, including 160 stars and numerous contributions through raised issues and pull requests. Similarly, the unmixing autoencoders developed in Chapter 4 have already been applied in other domains, such as vaccine development[572]. The created benchmark datasets have also been adopted for algorithm development and validation[573–575]. To further extend its accessibility, the end-to-end RS imaging pipeline presented in Chapter 5 will also be released as a user-friendly, open-source tool, allowing researchers to run predictions on their own data with minimal coding requirements. Building on their demonstrated impact, these resources aim to continue driving innovation, fostering new research and broadening the applications of RS across science and technology.

6.3 Limitations

Despite the progress achieved in advancing RS and organoid imaging, several methodological and conceptual limitations persist.

Scope of RamanSPy. As previously discussed, RamanSPy provides a comprehensive suite of tools for spectral analysis, spanning data loading, handling, preprocessing, analysis and visualisation (Table B.1 in Appendix B.1). Nevertheless, given the breadth of methods and workflows in RS, several algorithms and procedures were not included in the initial release – for example, methods for calibration, outlier detection and peak fitting. Owing to its modular design, it is anticipated that community contributions will continue to extend its scope, incorporating new methods, algorithms, datasets and other resources as the field evolves. This can also include the integration of advanced AI techniques, tighter coupling with experimental setups and complementary software, and the development of graphical user interfaces.

Interpretability and failure modes of autoencoders. This thesis demonstrates that unmixing AEs can effectively capture (bio)chemically relevant spectral features, with validation across synthetic data, sugar mixtures and complex biological systems such as cells and organoids. Nevertheless, as with many unsupervised neural network methods, mechanistic interpretability remains limited. It is generally unclear how features are encoded, what biochemical or structural properties they relate to, or under what conditions reconstructions fail. Future work should focus on benchmarking AE performance against well-characterised datasets, systematically analysing failure modes, and elucidating the mechanisms by which latent representations emerge. Such efforts will be essential for en-

*According to statistics by *pepy.tech*: https://pepy.tech/projects/ramanspy.

suring trustworthiness, particularly in the context of quantitative interpretation. Other aspects to be explored include data requirements for stable AE performance and sensitivity to hyperparameter tuning.

Endmember characterisation. As with other unmixing approaches, endmember characterisation in this work remains largely manual. While derived endmembers were assigned using established Raman bands, some peak assignments remain ambiguous due to the inherent biochemical complexity of neural organoids and possible depth-dependent signal attenuation. Although partial automation is possible through library searches and related HQI methods, these remain limited in scope: they cannot enumerate all possible molecular species and are highly sensitive to variations in experimental conditions, acquisition parameters and instrumentation. Future work integrating Raman imaging with complementary modalities may help to resolve such ambiguities and strengthen both qualitative and quantitative interpretation.

6.4 Future research directions

The methods and tools developed in this thesis open up numerous avenues for future research. They represent flexible frameworks that can be extended and repurposed across a wide range of applications, which could further strengthen, broaden and translate this work. Several key directions for future work are outlined below

Expanding applications in neural organoid research. The natural extension of this work is to broaden its applications within neural organoid research. Future studies could extend this line of research and investigate cell type differences, self-organisation, tissue morphogenesis, patterning and region-specific organisation. In addition to neurodevelopment, the approach can also be applied to disease models of neurodegeneration, cancer and neurodevelopmental disorders, as well as to the analysis of drug distribution and response in intact organoids. To this end, further validation under varied imaging parameters, culture conditions and sample preparations will be essential to establish robustness and reproducibility.

Integration with advanced RS modalities toward real-time imaging and analysis. In this study, spontaneous RS was used to validate the developed pipeline under stable, reproducible conditions using fixed neural organoids. Although inherently time-consuming, spontaneous RS offers high spectral resolution and chemical specificity, and is more broadly accessible than faster, more advanced RS modalities – making it well-suited for method development and benchmarking. Importantly, the presented unmixing analysis framework is also compatible with emerging broadband coherent Raman scat-

tering modalities, which can achieve video-rate imaging over wide spectral ranges[576]. For example, Vernuccio et al. recently reported a microscopy system based on coherent anti-Stokes Raman scattering spectroscopy (CARS) capable of full-spectrum acquisitions over large fields of view at millisecond dwell times[577]. This corresponds to acquisition speeds 2–3 orders of magnitude faster than standard spontaneous Raman settings (e.g. approximately 500 times faster than the integration time of 0.5 seconds used for the Raman imaging data presented in Chapter 5), thereby reducing recording times for large imaging scans from hours/days to minutes. Vernuccio et al. demonstrated that, combined with standard methods for unmixing (N-FINDR+NNLS), their system enabled rapid biochemical imaging and analysis of cells and tissues[577]. In this context, spectral unmixing remains a key step in the analysis pipeline, and integrating the methods developed here with more advanced RS techniques could allow fast imaging and analysis with improved chemical accuracy. In particular, this could enable real-time, label-free phenotyping and monitoring of live organoids, further advancing their use as preclinical platforms for precision medicine and high-throughput drug screening. In the long term, these advances could feed into fully automated, closed-loop experimental platforms combining real-time RS acquisition, deep learning-based analysis and experimental control for self-driving laboratories.

To demonstrate the applicability of the approach to advanced RS modalities, the AE-based unmixing pipeline was further applied to full-spectrum CARS data from a HepG2 human liver cancer cell from the work of Vernuccio et al.[577]. Its performance was compared to that of the original approach based on N-FINDR combined with NNLS. Figure 6.1 shows that the AE method developed here yields visually improved abundance reconstruction of lipids, proteins and nucleic acids, offering clearer delineation of individual nuclei and other cellular organelles.

Modelling spatial correlations in Raman imaging data. In many imaging applications, and particularly in biological samples such as organoids and tissues, adjacent pixels share correlated biochemical features. However, most spectral unmixing methods, including the AEs presented in this thesis, treat each spectrum independently and thus disregard the underlying spatial organisation of chemical species. Explicitly modelling these dependencies could improve the accuracy and robustness of unmixing, especially under low-SNR conditions or when distinguishing subtle spectral differences. Traditional approaches often ignore spatial context for reasons of scalability and tractability, but neural network architectures offer more flexible alternatives. For instance, incorporating 2D or 3D convolutional layers into the encoder of an unmixing AEs could enable spatially aware representations that capture these correlations while maintaining computational efficiency. Recently, Luo et al. applied a similar strategy in the context of RS classifica-

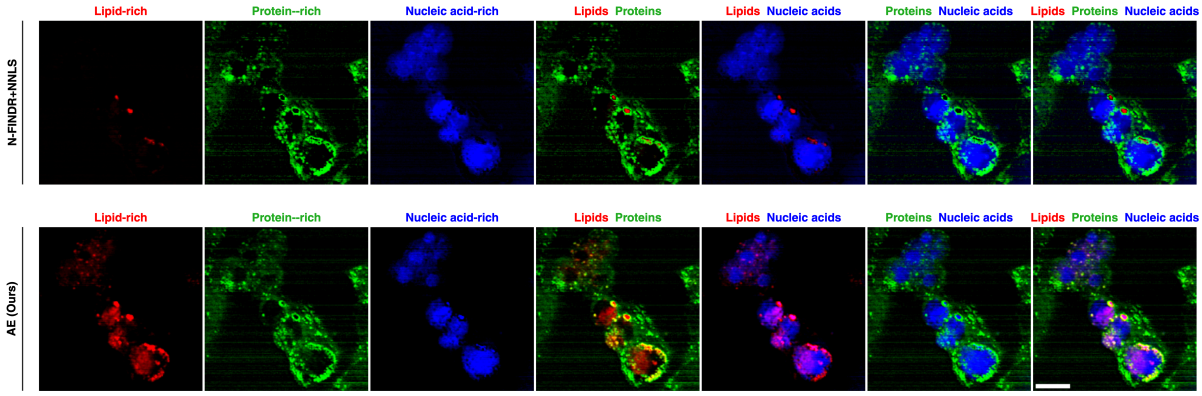


Figure 6.1: Hyperspectral unmixing analysis of broadband CARS imaging data from a cancer cell. Side-by-side comparison of N-FINDR+NNLS and the developed AE-based pipeline applied to broadband CARS data from a HepG2 human liver cancer cell (500 µm pixel size, 1 ms pixel dwell time). Scale bar: 20 µm. Preprocessed CARS data from Vernuccio et al.[577], provided by Dr Arianna Bresci. Assembled in BioRender.com.

tion[578]; however, its use in unsupervised unmixing remains unexplored and represents a promising avenue for future research.

More powerful unmixing AE models. Beyond spatial dependencies, unmixing AEs provide a versatile framework that can be extended in several other directions. Future research could investigate more advanced, non-linear decoder models[243], beyond the bilinear model tested in Figure 4.5C. Another opportunity includes further exploring training objectives more closely tailored to spectral reconstruction[548], expanding on the analysis presented in Figure 5.7, as well as encoder architectures for improved representation learning, such as variational models, deeper attention-based networks, and stacked or denoising autoencoders[533,535].

Automatic estimation of the number of endmembers. As with conventional methods (e.g. N-FINDR, VCA), the number of endmembers to extract remains a hyperparameter that must be carefully specified and optimised in the AE learning process. In practice, this is often inferred from prior knowledge or visual inspection *a posteriori*. Future work could address this limitation by incorporating sparsity regularisation or disentanglement strategies into the latent space of unmixing AEs, constraining the number of active dimensions and allowing the model to infer the effective number of components needed for reconstruction automatically. Such approaches could reduce bias, shorten optimisation times and enable more data-driven discovery of (bio)chemical features, particularly in exploratory analyses of unknown or heterogeneous samples.

Feature extraction for downstream applications. Another promising area is the use of AEs as a pre-training or feature extraction strategy, potentially in combination with

other AI-based approaches (e.g. deep learning models for denoising[209]). Beyond imaging, the unsupervised representations learned by unmixing AEs could serve as informative features for downstream tasks, such as organoid phenotyping, disease classification and drug response modelling. Building on this, the development of foundation models for RS based on unmixing AEs trained on diverse experimental datasets warrants further investigation[579]. Other research directions include transfer learning, and multimodal integration with complementary data types such as spatial transcriptomics or fluorescence microscopy.

Generalising to other samples and modalities. While this thesis focused on neural organoids, the imaging framework presented here can also be applied to other types of samples, including other types of organoids, tissues, and even non-biological specimens. Furthermore, while the focus of this thesis has been on RS and the unique requirements and challenges inherent to Raman data, many of the tools and methods developed here could be adapted to other spectroscopic modalities, such as IR and UV-vis spectroscopy.

6.5 Closing remarks

This thesis reflects the growing interest in both RS and organoid technology within the biomedical sciences and the need for more accessible, robust and well-integrated workflows for RS chemometrics in this context. It establishes a foundation for exciting future research at the intersection of organoids, label-free Raman imaging and deep learning. Looking ahead, the contributions of this thesis aim not only to deliver tools and methods, but also to promote a broader mindset of openness and interdisciplinarity. Such principles will be essential as ML and data-driven discovery become increasingly central to understanding complex biological systems, with usability, transparency and modularity key to maximising impact and ensuring that these advances benefit the wider scientific community. It is my hope that others will build upon these ideas, further advancing the tools introduced in this work and pushing the boundaries of what spectroscopy can reveal about living systems and beyond.

Bibliography

[1] John D Bancroft and Marilyn Gamble. *Theory and practice of histological techniques*. Elsevier health sciences, 2008.

[2] Paul W Harms, Timothy L Frankel, Myrto Moutafi, Arvind Rao, David L Rimm, Janis M Taube, Dafydd Thomas, May P Chan, and Liron Pantanowitz. Multiplex immunohistochemistry and immunofluorescence: a practical update for pathologists. *Modern Pathology*, 36(7):100197, 2023.

[3] Ben NG Giepmans, Stephen R Adams, Mark H Ellisman, and Roger Y Tsien. The fluorescent toolbox for assessing protein location and function. *Science*, 312(5771):217–224, 2006.

[4] Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353 (6294):78–82, 2016.

[5] Anjali Rao, Dalia Barkley, Gustavo S França, and Itai Yanai. Exploring tissue architecture using spatial transcriptomics. *Nature*, 596(7871):211–220, 2021.

[6] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):83, 2017.

[7] Madeline A Lancaster, Magdalena Renner, Carol-Anne Martin, Daniel Wenzel, Louise S Bicknell, Matthew E Hurles, Tessa Homfray, Josef M Penninger, Andrew P Jackson, and Juergen A Knoblich. Cerebral organoids model human brain development and microcephaly. *Nature*, 501 (7467):373–379, 2013.

[8] Madeline A Lancaster and Juergen A Knoblich. Organogenesis in a dish: modeling development and disease using organoid technologies. *Science*, 345(6194):1247125, 2014.

[9] Sergiu P Pașca. The rise of three-dimensional human brain cultures. *Nature*, 553(7689):437–445, 2018.

[10] Ilaria Chiaradia and Madeline A Lancaster. Brain organoids for the study of human neurobiology at the interface of in vitro and in vivo. *Nature Neuroscience*, 23(12):1496–1508, 2020.

[11] Giorgia Quadrato, Tuan Nguyen, Evan Z Macosko, John L Sherwood, Sung Min Yang, Daniel R Berger, Natalie Maria, Jorg Scholvin, Melissa Goldman, Justin P Kinney, et al. Cell diversity and network dynamics in photosensitive human brain organoids. *Nature*, 545(7652):48–53, 2017.

[12] Silvia Velasco, Amanda J Kedaigle, Sean K Simmons, Allison Nash, Marina Rocha, Giorgia Quadrato, Bruna Paulsen, Lan Nguyen, Xian Adiconis, Aviv Regev, et al. Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. *Nature*, 570(7762):523–527, 2019.

[13] Cleber A Trujillo, Richard Gao, Priscilla D Negraes, Jing Gu, Justin Buchanan, Sebastian Preissl, Allen Wang, Wei Wu, Gabriel G Haddad, Isaac A Chaim, et al. Complex oscillatory waves emerging from cortical organoids model early human brain network development. *Cell stem cell*, 25(4):558–569, 2019.

[14] Xuyu Qian, Hongjun Song, and Guo-li Ming. Brain organoids: advances, applications and challenges. *Development*, 146(8):dev166074, 2019.

[15] Elizabeth Di Lullo and Arnold R Kriegstein. The use of brain organoids to investigate neural development and disease. *Nature Reviews Neuroscience*, 18(10):573–584, 2017.

[16] Oliver L Eichmüller and Juergen A Knoblich. Human cerebral organoids—a new tool for clinical neurology research. *Nature Reviews Neurology*, 18(11):661–680, 2022.

[17] Zehra Yildirim, Kyle Swanson, Xuekun Wu, James Zou, and Joseph Wu. Next-Gen Therapeutics: Pioneering Drug Discovery with iPSCs, Genomics, AI, and Clinical Trials in a Dish. *Annual Review of Pharmacology and Toxicology*, 65, 2024.

[18] Norman Colthup. *Introduction to infrared and Raman spectroscopy*. Elsevier, 2012.

[19] Richard L McCreery. *Raman spectroscopy for chemical analysis*. John Wiley & Sons, 2005.

[20] Ewen Smith and Geoffrey Dent. *Modern Raman spectroscopy: a practical approach*. John Wiley & Sons, 2019.

[21] Holly J Butler, Lorna Ashton, Benjamin Bird, Gianfelice Cinque, Kelly Curtis, Jennifer Dorney, Karen Esmonde-White, Nigel J Fullwood, Benjamin Gardner, Pierre L Martin-Hirsch, et al. Using Raman spectroscopy to characterize biological materials. *Nature protocols*, 11(4):664–687, 2016.

[22] Kosuke Dodo, Katsumasa Fujita, and Mikiko Sodeoka. Raman spectroscopy for chemical biology research. *Journal of the American Chemical Society*, 144(43):19651–19667, 2022.

[23] Dustin W Shipp, Faris Sinjab, and Ioan Notingher. Raman spectroscopy: techniques and applications in the life sciences. *Advances in Optics and Photonics*, 9(2):315–428, 2017.

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

[25] Kenny Kong, Catherine Kendall, Nicholas Stone, and Ioan Notingher. Raman spectroscopy for medical diagnostics—From in-vitro biofluid assays to in-vivo cancer detection. *Advanced drug delivery reviews*, 89:121–134, 2015.

[26] Katherine JI Ember, Marieke A Hoeve, Sarah L McAughtrie, Mads S Bergholt, Benjamin J Dwyer, Molly M Stevens, Karen Faulds, Stuart J Forbes, and Colin J Campbell. Raman spectroscopy and regenerative medicine: a review. *NPJ Regenerative medicine*, 2(1):12, 2017.

[27] Isaac Pence and Anita Mahadevan-Jansen. Clinical instrumentation and applications of Raman spectroscopy. *Chemical Society Reviews*, 45(7):1958–1979, 2016.

[28] Álvaro Fernández-Galiana, Olga Bibikova, Simon Vilms Pedersen, and Molly M Stevens. Fundamentals and Applications of Raman-Based Techniques for the Design and Development of Active Biomedical Materials. *Advanced Materials*, page 2210807, 2023.

[29] Amrit Paudel, Dhara Raijada, and Jukka Rantanen. Raman spectroscopy in pharmaceutical product design. *Advanced drug delivery reviews*, 89:3–20, 2015.

[30] Lynn R Terry, Sage Sanders, Rebecca H Potoff, Jacob W Kruel, Manan Jain, and Huiyuan Guo. Applications of surface-enhanced Raman spectroscopy in environmental detection. *Analytical Science Advances*, 3(3-4):113–145, 2022.

[31] Shelby R Khandasammy, Marisia A Fikiet, Ewelina Mistek, Yasmine Ahmed, Lenka Halámková, Justin Bueno, and Igor K Lednev. Bloodstains, paintings, and drugs: Raman spectroscopy applications in forensic science. *Forensic Chemistry*, 8:111–133, 2018.

[32] Zanyar Movasaghi, Shazza Rehman, and Ihtesham U Rehman. Raman spectroscopy of biological tissues. *Applied Spectroscopy Reviews*, 42(5):493–541, 2007.

[33] Abdullah Chandra Sekhar Talari, Zanyar Movasaghi, Shazza Rehman, and Ihtesham Ur Rehman. Raman spectroscopy of biological tissues. *Applied spectroscopy reviews*, 50(1):46–111, 2015.

[34] Christoph Krafft, Benjamin Dietzek, and Jürgen Popp. Raman and CARS microspectroscopy of

cells and tissues. *Analyst*, 134(6):1046–1057, 2009.

[35] Thomas Huser and James Chan. Raman spectroscopy for physiological investigations of tissues and cells. *Advanced drug delivery reviews*, 89:57–70, 2015.

[36] Giuseppe Pezzotti. Raman spectroscopy in cell biology and microbiology. *Journal of Raman Spectroscopy*, 52(12):2348–2443, 2021.

[37] Rachael Smith, Karen L Wright, and Lorna Ashton. Raman spectroscopy: an evolving technique for live cell studies. *Analyst*, 141(12):3590–3600, 2016.

[38] Bastian Barton, James Thomson, Enrique Lozano Diz, and Raquel Portela. Chemometrics for Raman Spectroscopy Harmonization. *Applied Spectroscopy*, 76(9):1021–1041, 2022.

[39] Shuxia Guo, Jürgen Popp, and Thomas Bocklitz. Chemometric analysis in Raman spectroscopy from experimental design to machine learning–based modeling. *Nature Protocols*, 16(12):5426–5459, 2021.

[40] Hugh J Byrne, Peter Knief, Mark E Keating, and Franck Bonnier. Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells. *Chemical Society Reviews*, 45 (7):1865–1878, 2016.

[41] Rekha Gautam, Sandeep Vanga, Freek Ariese, and Siva Umapathy. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation*, 2:1–38, 2015.

[42] Thomas W Bocklitz, Shuxia Guo, Oleg Ryabchykov, Nadine Vogler, and Jürgen Popp. Raman based molecular imaging and analytics: a magic bullet for biomedical applications!? *Analytical chemistry*, 88(1):133–151, 2016.

[43] Liangrui Pan, Peng Zhang, Chalongrat Daengngam, Shaoliang Peng, and Mitchai Chongcheawchamnan. 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.

[44] Nicolas Coca-Lopez, Victor Alcolea-Rodriguez, Miguel A Bañares, Sandor Brockhauser, Julien Gorenflot, Alex Henderson, Ron Hildebrandt, Nina Jeliazkova, Nikolay Kochev, Enrique Lozano-Diz, et al. Open and FAIR Raman spectroscopy. Paving the way for artificial intelligence. *ChemRxiv preprint 10.26434/chemrxiv-2025-0q2mt*, 2025.

[45] Eugene Hecht. *Optics*. Pearson Education, fifth edition, 2017.

[46] Bahaa EA Saleh and Malvin Carl Teich. *Fundamentals of photonics*. John Wiley & Sons, third edition, 2019.

[47] Adolf Smekal. Zur quantentheorie der dispersion. *Naturwissenschaften*, 11(43):873–875, 1923.

[48] Chandrasekhara Venkata Raman and Kariamanikkam Srinivasa Krishnan. A new type of secondary radiation. *Nature*, 121(3048):501–502, 1928.

[49] Gr Landsberg and Leonid Mandelstam. Über die lichtzerstreuung in kristallen. *Zeitschrift für Physik*, 50(11):769–780, 1928.

[50] Bruce Chase. A new generation of Raman instrumentation. *Applied Spectroscopy*, 48(7):14A–19A, 1994.

[51] Johannes Arnoldus Koningstein. *Introduction to the Theory of the Raman Effect*. Springer Science & Business Media, 2012.

[52] Robin R Jones, David C Hooper, Liwu Zhang, Daniel Wolverson, and Ventsislav K Valev. Raman techniques: fundamentals and frontiers. *Nanoscale research letters*, 14:1–34, 2019.

[53] Herman A Szymanski. *Raman spectroscopy: theory and practice*. Springer Science & Business Media, 2012.

[54] Sebastian Schlücker, Michael D Schaeberle, Scott W Huffman, and Ira W Levin. Raman microspectroscopy: a comparison of point, line, and wide-field imaging methodologies. *Analytical Chemistry*,

75(16):4312–4318, 2003.

[55] Evtim V Efremov, Freek Ariese, and Cees Gooijer. Achievements in resonance Raman spectroscopy: Review of a technique with a distinct analytical chemistry potential. *Analytica chimica acta*, 606 (2):119–134, 2008.

[56] Bruno Robert. Resonance Raman spectroscopy. *Photosynthesis research*, 101(2):147–155, 2009.

[57] P Carey. *Biochemical applications of Raman and resonance Raman spectroscopes*. Elsevier, 2012.

[58] Eric Le Ru and Pablo Etchegoin. *Principles of Surface-Enhanced Raman Spectroscopy: and related plasmonic effects*. Elsevier, 2008.

[59] Song-Yuan Ding, En-Ming You, Zhong-Qun Tian, and Martin Moskovits. Electromagnetic theories of surface-enhanced Raman spectroscopy. *Chemical Society Reviews*, 46(13):4042–4076, 2017.

[60] Martin Moskovits. Surface-enhanced Raman spectroscopy: a brief retrospective. *Journal of Raman Spectroscopy: An International Journal for Original Work in all Aspects of Raman Spectroscopy, Including Higher Order Processes, and also Brillouin and Rayleigh Scattering*, 36(6-7):485–496, 2005.

[61] Paul L Stiles, Jon A Dieringer, Nilam C Shah, and Richard P Van Duyne. Surface-enhanced Raman spectroscopy. *Annu. Rev. Anal. Chem.*, 1:601–626, 2008.

[62] Sebastian Schlücker. Surface-Enhanced Raman Spectroscopy: Concepts and Chemical Applications. *Angewandte Chemie International Edition*, 53(19):4756–4795, 2014.

[63] Xiao Xia Han, Rebeca S Rodriguez, Christy L Haynes, Yukihiro Ozaki, and Bing Zhao. Surface-enhanced Raman spectroscopy. *Nature Reviews Methods Primers*, 1(1):87, 2021.

[64] William M Tolles, Joseph W Nibler, JR McDonald, and Albert B Harvey. A review of the theory and application of coherent anti-Stokes Raman spectroscopy (CARS). *Applied Spectroscopy*, 31(4): 253–271, 1977.

[65] Conor L Evans and X Sunney Xie. Coherent anti-Stokes Raman scattering microscopy: chemical imaging for biology and medicine. *Annu. Rev. Anal. Chem.*, 1(1):883–909, 2008.

[66] Ji-Xin Cheng and X Sunney Xie. Coherent anti-Stokes Raman scattering microscopy: instrumentation, theory, and applications, 2004.

[67] Sukesh Roy, James R Gord, and Anil K Patnaik. Recent advances in coherent anti-Stokes Raman scattering spectroscopy: Fundamental developments and applications in reacting flows. *Progress in Energy and Combustion Science*, 36(2):280–306, 2010.

[68] Conor L Evans, Eric O Potma, Mehron Puoris' haag, Daniel Côté, Charles P Lin, and X Sunney Xie. Chemical imaging of tissue in vivo with video-rate coherent anti-Stokes Raman scattering microscopy. *Proceedings of the national academy of sciences*, 102(46):16807–16812, 2005.

[69] Dana Cialla-May, Christoph Krafft, Petra Rösch, Tanja Deckert-Gaudig, Torsten Frosch, Izabella J Jahn, Susanne Pahlow, Clara Stiebing, Tobias Meyer-Zedler, Thomas Bocklitz, et al. Raman spectroscopy and imaging in bioanalytics. *Analytical chemistry*, 94(1):86–119, 2021.

[70] Gregory W Auner, S Kiran Koya, Changhe Huang, Brandy Broadbent, Micaela Trexler, Zachary Auner, Angela Elias, Katlyn Curtin Mehne, and Michelle A Brusatori. Applications of Raman spectroscopy in cancer diagnosis. *Cancer and Metastasis Reviews*, 37(4):691–717, 2018.

[71] Katharina Klein, Alexander M Gigler, Thomas Aschenbrenner, Roberto Monetti, Wolfram Bunk, Ferdinand Jamitzky, Gregor Morfill, Robert W Stark, and Jürgen Schlegel. Label-free live-cell imaging with confocal Raman microscopy. *Biophysical journal*, 102(2):360–368, 2012.

[72] Lu Wei, Fanghao Hu, Yihui Shen, Zhixing Chen, Yong Yu, Chih-Chun Lin, Meng C Wang, and Wei Min. Live-cell imaging of alkyne-tagged small biomolecules by stimulated Raman scattering. *Nature methods*, 11(4):410–412, 2014.

[73] Kenta Mizushima, Yasuaki Kumamoto, Shoko Tamura, Masahito Yamanaka, Kentaro Mochizuki, Menglu Li, Syusuke Egoshi, Kosuke Dodo, Yoshinori Harada, Nicholas I Smith, et al. Raman mi-

croscopy of cryofixed biological specimens for high-resolution and high-sensitivity chemical imaging. *Science Advances*, 10(50):eadn0110, 2024.

[74] Charalambos Kallepitis, Mads S Bergholt, Manuel M Mazo, Vincent Leonardo, Stacey C Skaalure, Stephanie A Maynard, and Molly M Stevens. Quantitative volumetric Raman imaging of three dimensional cell cultures. *Nature communications*, 8(1):14843, 2017.

[75] Tobias J Moritz, Douglas S Taylor, Denise M Krol, John Fritch, and James W Chan. Detection of doxorubicin-induced apoptosis of leukemic T-lymphocytes by laser tweezers Raman spectroscopy. *Biomedical optics express*, 1(4):1138–1147, 2010.

[76] Tatyana Chernenko, Christian Matthäus, Lara Milane, Luis Quintero, Mansoor Amiji, and Max Diem. Label-Free Raman Spectral Imaging of Intracellular Delivery and Degradation of Polymeric Nanoparticle Systems. *ACS Nano*, 3(11):3552–3559, 2009.

[77] Gabriela Romero, Irina Estrela-Lopis, Jie Zhou, Elena Rojas, Ana Franco, Christian Sanchez Espinel, Africa González Fernández, Changyou Gao, Edwin Donath, and Sergio E Moya. Surface Engineered Poly(lactide-co-glycolide) Nanoparticles for Intracellular Delivery: Uptake and Cytotoxicity–A Confocal Raman Microscopic Study. *Biomacromolecules*, 11(11):2993–2999, 2010.

[78] JL Deng, Q Wei, MH Zhang, YZ Wang, and YQ Li. Study of the effect of alcohol on single human red blood cells using near-infrared laser tweezers Raman spectroscopy. *Journal of Raman Spectroscopy: An International Journal for Original Work in all Aspects of Raman Spectroscopy, Including Higher Order Processes, and also Brillouin and Rayleigh Scattering*, 36(3):257–261, 2005.

[79] Christoph Krafft, Thomas Knetschke, Richard HW Funk, and Reiner Salzer. Studies on stress-induced changes at the subcellular level by Raman microspectroscopic mapping. *Analytical chemistry*, 78(13):4424–4429, 2006.

[80] Iwan W Schie and Thomas Huser. Methods and applications of Raman microspectroscopy to single-cell analysis. *Applied spectroscopy*, 67(8):813–828, 2013.

[81] Yirui Zhang, Kai Chang, Babatunde Ogunlade, Liam Herndon, Loza F Tadesse, Amanda R Kirane, and Jennifer A Dionne. From Genotype to Phenotype: Raman Spectroscopy and Machine Learning for Label-Free Single-Cell Analysis. *ACS nano*, 2024.

[82] Marcus T Cicerone and Charles H Camp. Histological coherent Raman imaging: a prognostic review. *Analyst*, 143(1):33–59, 2018.

[83] Mitsuru J Nakamura, Kohji Hotta, and Kotaro Oka. Raman spectroscopic imaging of the whole ciona intestinalis embryo during development. *PLoS One*, 8(8):e71739, 2013.

[84] Ralph Michael, Aufried Lenferink, Gijs FJM Vrensen, Ellen Gelpi, Rafael I Barraquer, and Cees Otto. Hyperspectral Raman imaging of neuritic plaques and neurofibrillary tangles in brain tissue from Alzheimer's disease patients. *Scientific reports*, 7(1):15603, 2017.

[85] Fiona Xi Xu, Rui Sun, Ryan Owens, Kailun Hu, and Dan Fu. Assessing drug uptake and response differences in 2D and 3D cellular environments using stimulated Raman scattering microscopy. *Analytical Chemistry*, 96(36):14480–14489, 2024.

[86] Erin L Dunnington, Brian S Wong, and Dan Fu. Innovative approaches for drug discovery: quantifying drug distribution and response with Raman imaging. *Analytical chemistry*, 96(20):7926–7944, 2024.

[87] Minbiao Ji, Daniel A Orringer, Christian W Freudiger, Shakti Ramkissoon, Xiaohui Liu, Darryl Lau, Alexandra J Golby, Isaiah Norton, Marika Hayashi, Nathalie YR Agar, et al. Rapid, label-free detection of brain tumors with stimulated Raman scattering microscopy. *Science translational medicine*, 5(201):201ra119–201ra119, 2013.

[88] Charles H Camp Jr and Marcus T Cicerone. Chemically sensitive bioimaging with coherent Raman scattering. *Nature photonics*, 9(5):295–305, 2015.

[89] Yasuyuki Ozeki, Wataru Umemura, Yoichi Otsuka, Shuya Satoh, Hiroyuki Hashimoto, Kazuhiko

Sumimura, Norihiko Nishizawa, Kiichi Fukui, and Kazuyoshi Itoh. High-speed molecular spectral imaging of tissue with stimulated Raman scattering. *Nature photonics*, 6(12):845–851, 2012.

[90] Vernon LaLone, Aleksandra Aizenshtadt, John Goertz, Frøydis Sved Skottvoll, Marco Barbero Mota, Junji You, Xiaoyu Zhao, Henriette Engen Berg, Justyna Stokowiec, Minzhi Yu, et al. Quantitative chemometric phenotyping of three-dimensional liver organoids by Raman spectral imaging. *Cell Reports Methods*, 3(4), 2023.

[91] JL Pichardo-Molina, C Frausto-Reyes, O Barbosa-García, R Huerta-Franco, JL González-Trujillo, CA Ramírez-Alvarado, G Gutiérrez-Juárez, and C Medina-Gutiérrez. Raman spectroscopy and multivariate analysis of serum samples from breast cancer patients. *Lasers in medical science*, 22 (4):229–236, 2007.

[92] Andrew T Harris, Anxhela Lungari, Christopher J Needham, Stephen L Smith, Michael A Lones, Sheila E Fisher, Xuebin B Yang, Nicola Cooper, Jennifer Kirkham, D Alastair Smith, et al. Potential for Raman spectroscopy to provide cancer screening using a peripheral blood sample. *Head & neck oncology*, 1(1):34, 2009.

[93] Guadalupe Donjuán-Loredo, Ricardo Espinosa-Tanguma, and Miguel G Ramírez-Elías. Raman spectroscopy in the diagnosis of metabolic syndrome. *Applied Spectroscopy Reviews*, 58(3):159–179, 2023.

[94] Alison J Hobro, Aki Konishi, Cevayir Coban, and Nicholas I Smith. Raman spectroscopic analysis of malaria disease progression via blood and plasma samples. *Analyst*, 138(14):3927–3933, 2013.

[95] Annika MK Enejder, Tae-Woong Koo, Jeankun Oh, Martin Hunter, Slobodan Sasic, Michael S Feld, and Gary L Horowitz. Blood analysis by Raman spectroscopy. *Optics letters*, 27(22):2004–2006, 2002.

[96] Lingwei Ma, Zhengjun Zhang, and Xiaogang Li. Non-invasive disease diagnosis using surface-enhanced Raman spectroscopy of urine and saliva. *Applied Spectroscopy Reviews*, 55(3):197–219, 2020.

[97] Pimporn Reokrungruang, Itthi Chatnuntawech, Tararaj Dharakul, and Suwussa Bamrungsap. A simple paper-based surface enhanced Raman scattering (SERS) platform and magnetic separation for cancer screening. *Sensors and Actuators B: Chemical*, 285:462–469, 2019.

[98] Yanlin Wang, Qiaoyu Li, Haimei Shi, Keqi Tang, Liang Qiao, Guopeng Yu, Chuanfan Ding, and Shaoning Yu. Microfluidic Raman biochip detection of exosomes: a promising tool for prostate cancer diagnosis. *Lab on a Chip*, 20(24):4632–4637, 2020.

[99] Shruti D Bindesri, Dalal S Alhatab, and Christa L Brosseau. Development of an electrochemical surface-enhanced Raman spectroscopy (EC-SERS) fabric-based plasmonic sensor for point-of-care diagnostics. *Analyst*, 143(17):4128–4135, 2018.

[100] Chi-Sing Ho, Neal Jean, Catherine A Hogan, Lena Blackmon, Stefanie S Jeffrey, Mark Holodniy, Niaz Banaei, Amr AE Saleh, Stefano Ermon, and Jennifer Dionne. Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nature communications*, 10(1): 1–8, 2019.

[101] Jijo Lukose, Ajaya Kumar Barik, N Mithun, Sanoop Pavithran, Sajan D George, VM Murukeshan, and Santhosh Chidangil. Raman spectroscopy for viral diagnostics. *Biophysical Reviews*, 15(2): 199–221, 2023.

[102] Damon DePaoli, Émile Lemoine, Katherine Ember, Martin Parent, Michel Prud'homme, Léo Cantin, Kevin Petrecca, Frédéric Leblond, and Daniel C Côté. Rise of Raman spectroscopy in neurosurgery: a review. *Journal of biomedical optics*, 25(5):050901–050901, 2020.

[103] Swati Tanwar, Santosh Kumar Paidi, Ram Prasad, Rishikesh Pandey, and Ishan Barman. Advancing Raman spectroscopy from research to clinic: Translational potential and challenges. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 260:119957, 2021.

[104] Inês P Santos, Elisa M Barroso, Tom C Bakker Schut, Peter J Caspers, Cornelia GF van Lanschot, Da-Hye Choi, Martine F Van Der Kamp, Roeland WH Smits, Remco Van Doorn, Rob M Verdijk, et al. Raman spectroscopy for cancer detection and cancer surgery guidance: translation to the clinics. *Analyst*, 142(17):3025–3047, 2017.

[105] Aaron M Mohs, Michael C Mancini, Sunil Singhal, James M Provenzale, Brian Leyland-Jones, May D Wang, and Shuming Nie. Hand-held spectroscopic device for in vivo and intraoperative tumor detection: contrast enhancement, detection sensitivity, and tissue penetration. *Analytical chemistry*, 82(21):9058–9065, 2010.

[106] Max Diem, Antonella Mazur, Kathleen Lenau, Jen Schubert, Ben Bird, Milo Miljković, Christoph Krafft, and Jürgen Popp. Molecular pathology via IR and Raman spectral imaging. *Journal of biophotonics*, 6(11-12):855–886, 2013.

[107] Damien Traynor, Isha Behl, Declan O'dea, Franck Bonnier, Siobhan Nicholson, Finbar O'connell, Aoife Maguire, Stephen Flint, Sheila Galvin, Claire M Healy, et al. Raman spectral cytopathology for cancer diagnostic applications. *Nature Protocols*, 16(7):3716–3735, 2021.

[108] Jin-Chao Dong, Xia-Guang Zhang, Valentín Briega-Martos, Xi Jin, Ji Yang, Shu Chen, Zhi-Lin Yang, De-Yin Wu, Juan Miguel Feliu, Christopher T Williams, et al. In situ Raman spectroscopic evidence for oxygen reduction reaction intermediates at platinum single-crystal surfaces. *Nature Energy*, 4(1):60–67, 2019.

[109] Khay M Tan, Ishan Barman, Narahara C Dingari, Gajendra P Singh, Tet F Chia, and Wee L Tok. Toward the development of Raman spectroscopy as a nonperturbative online monitoring tool for gasoline adulteration. *Analytical chemistry*, 85(3):1846–1851, 2013.

[110] Bernhard Zachhuber, Georg Ramer, Alison Hobro, Engelene t H Chrysostom, and Bernhard Lendl. Stand-off Raman spectroscopy: a powerful technique for qualitative and quantitative analysis of inorganic and organic compounds including explosives. *Analytical and bioanalytical chemistry*, 400 (8):2439–2447, 2011.

[111] Mingjin Wang, Nan Zheng, Defeng Zhao, Jing Shang, and Tong Zhu. Using Micro-Raman Spectroscopy to Investigate Chemical Composition, Mixing States, and Heterogeneous Reactions of Individual Atmospheric Particles. *Environmental Science & Technology*, 55(15):10243–10254, 2021.

[112] Olof Svensson, Mats Josefson, and Frans W Langkilde. Reaction monitoring using Raman spectroscopy and chemometrics. *Chemometrics and intelligent laboratory systems*, 49(1):49–66, 1999.

[113] Melissa Assirelli, Weiyin Xu, and Wee Chew. Reactor kinetics studies via process Raman spectroscopy, multivariate chemometrics, and kinetics modeling. *Organic Process Research & Development*, 15(3):610–621, 2011.

[114] Christian Hess. New advances in using Raman spectroscopy for the characterization of catalysts and catalytic reactions. *Chemical Society Reviews*, 50(5):3519–3564, 2021.

[115] Lotfi Merad, M Cochez, Samuel Margueron, F Jauchem, M Ferriol, Boumediene Benyoucef, and P Bourson. In-situ monitoring of the curing of epoxy resins by Raman spectroscopy. In *New Achievements in Materials and Environmental Sciences Names*, pages 155–158. EDP Sciences, 2007.

[116] El-Sayed M Sherif, RM Erasmus, and JD Comins. In situ Raman spectroscopy and electrochemical techniques for studying corrosion and corrosion inhibition of iron in sodium chloride solutions. *Electrochimica Acta*, 55(11):3657–3663, 2010.

[117] Lajos Rimai, David Gill, and John L Parsons. Raman spectra of dilute solutions of some steroisomers of vitamin A type molecules. *Journal of the American Chemical Society*, 93(6):1353–1357, 1971.

[118] Pinkie J Eravuchira, Rasha M El-Abassy, Sagar Deshpande, Marius F Matei, Soni Mishra, Poonam Tandon, Nikolai Kuhnert, and Arnulf Materny. Raman spectroscopic characterization of different

regioisomers of monoacyl and diacyl chlorogenic acid. *Vibrational Spectroscopy*, 61:10–16, 2012.

[119] Aurélie Dandeu, Bernard Humbert, Cédric Carteret, Hervé Muhr, Edouard Plasari, and Jean-Michel Bossoutrot. Raman spectroscopy–a powerful tool for the quantitative determination of the composition of polymorph mixtures: application to $caco_3$ polymorph mixtures. *Chemical Engineering & Technology: Industrial Chemistry-Plant Equipment-Process Engineering-Biotechnology*, 29(2):221–225, 2006.

[120] Karen A Esmonde-White, Maryann Cuellar, and Ian R Lewis. The role of Raman spectroscopy in biopharmaceuticals from development to manufacturing. *Analytical and Bioanalytical Chemistry*, 414(2):969–991, 2022.

[121] Stuart Farquharson, Chetan Shende, Frank E Inscore, Paul Maksymiuk, and Alan Gift. Analysis of 5-fluorouracil in saliva using surface-enhanced Raman spectroscopy. *Journal of Raman Spectroscopy: An International Journal for Original Work in all Aspects of Raman Spectroscopy, Including Higher Order Processes, and also Brillouin and Rayleigh Scattering*, 36(3):208–212, 2005.

[122] Marja Savolainen, Andrea Heinz, Clare Strachan, Keith C Gordon, Jouko Yliruusi, Thomas Rades, and Niklas Sandler. Screening for differences in the amorphous state of indomethacin using multivariate visualization. *European journal of pharmaceutical sciences*, 30(2):113–123, 2007.

[123] Sarah Al-Dulaimi, Adeyinka Aina, and Jonathan Burley. Rapid polymorph screening on milligram quantities of pharmaceutical material using phonon-mode Raman spectroscopy. *CrystEngComm*, 12(4):1038–1040, 2010.

[124] Karen A Esmonde-White, Maryann Cuellar, Carsten Uerpmann, Bruno Lenain, and Ian R Lewis. Raman spectroscopy as a process analytical technology for pharmaceutical manufacturing and bioprocessing. *Analytical and bioanalytical chemistry*, 409(3):637–649, 2017.

[125] Lynne S Taylor and Frans W Langkilde. Evaluation of solid-state forms present in tablets by Raman spectroscopy. *Journal of pharmaceutical sciences*, 89(10):1342–1353, 2000.

[126] Sabine Neuberger and Christian Neusüß. Determination of counterfeit medicines by Raman spectroscopy: systematic study based on a large set of model tablets. *Journal of pharmaceutical and biomedical analysis*, 112:70–78, 2015.

[127] Ahmad Aqel, Kholoud MM Abou El-Nour, Reda AA Ammar, and Abdulrahman Al-Warthan. Carbon nanotubes, science and technology part (I) structure, synthesis and characterisation. *Arabian Journal of Chemistry*, 5(1):1–23, 2012.

[128] Ying-Sing Li and Jeffrey S Church. Raman spectroscopy in the analysis of food and pharmaceutical nanomaterials. *Journal of food and drug analysis*, 22(1):29–48, 2014.

[129] Timothy TX Ong, Ewan W Blanch, and Oliver AH Jones. Surface Enhanced Raman Spectroscopy in environmental analysis, monitoring and assessment. *Science of The Total Environment*, 720: 137601, 2020.

[130] Zhiming Guo, Ping Chen, Nermeen Yosri, Quansheng Chen, Hesham R Elseedi, Xiaobo Zou, and Hongshun Yang. Detection of heavy metals in food and agricultural products by surface-enhanced Raman spectroscopy. *Food Reviews International*, 39(3):1440–1461, 2023.

[131] Yuxia Fan, Keqiang Lai, Barbara A Rasco, and Yiqun Huang. Analyses of phosmet residues in apples with surface-enhanced Raman spectroscopy. *Food Control*, 37:153–157, 2014.

[132] Catarina F Araujo, Mariela M Nolasco, Antonio MP Ribeiro, and Paulo JA Ribeiro-Claro. Identification of microplastics using Raman spectroscopy: Latest developments and future prospects. *Water research*, 142:426–440, 2018.

[133] Meng-Lei Xu, Yu Gao, Xiao Xia Han, and Bing Zhao. Detection of pesticide residues in food using surface-enhanced Raman spectroscopy: a review. *Journal of agricultural and food chemistry*, 65 (32):6719–6726, 2017.

[134] Shizhuang Weng, Wenxiu Zhu, Xueyan Zhang, Hecai Yuan, Ling Zheng, Jinling Zhao, Linsheng

Huang, and Ping Han. Recent advances in Raman technology with applications in agriculture, food and biosystems review. *Artificial Intelligence in Agriculture*, 3:1–10, 2019.

[135] Shintaro Pang, Tianxi Yang, and Lili He. Review of surface enhanced Raman spectroscopic (SERS) detection of synthetic chemical pesticides. *TrAC Trends in Analytical Chemistry*, 85:73–82, 2016.

[136] Shizhuang Weng, Xujin Hu, Jinghong Wang, Le Tang, Pan Li, Shouguo Zheng, Ling Zheng, Linsheng Huang, and Zhenghua Xin. Advanced application of Raman spectroscopy and surface-enhanced Raman spectroscopy in plant disease diagnostics: a review. *Journal of Agricultural and Food Chemistry*, 69(10):2950–2964, 2021.

[137] Narangerel Altangerel, Gombojav O Ariunbold, Connor Gorman, Masfer H Alkahtani, Eli J Borrego, Dwight Bohlmeyer, Philip Hemmer, Michael V Kolomiets, Joshua S Yuan, and Marlan O Scully. In vivo diagnostics of early abiotic plant stress response via Raman spectroscopy. *Proceedings of the National Academy of Sciences*, 114(13):3393–3396, 2017.

[138] Yu Fu, Christian Kuppe, Ventsislav K Valev, Hongbo Fu, Liwu Zhang, and Jianmin Chen. Surface-enhanced Raman spectroscopy: A facile and rapid method for the chemical component study of individual atmospheric aerosol. *Environmental Science & Technology*, 51(11):6260–6267, 2017.

[139] Vasanthi Sivaprakasam and Matthew B Hart. Surface-enhanced Raman spectroscopy for environmental monitoring of aerosols. *ACS omega*, 6(15):10150–10159, 2021.

[140] Marc Fries and Andrew Steele. Raman spectroscopy and confocal Raman imaging in mineralogy and petrography. In *Confocal Raman Microscopy*, pages 209–236. Springer, 2018.

[141] Soumya Das and M Jim Hendry. Application of Raman spectroscopy to identify iron minerals commonly found in mine wastes. *Chemical Geology*, 290(3-4):101–108, 2011.

[142] Danilo Bersani and Pier Paolo Lottici. Raman spectroscopy of minerals and mineral pigments in archaeometry. *Journal of Raman Spectroscopy*, 47(5):499–530, 2016.

[143] Shiv K Sharma, Paul G Lucey, Manash Ghosh, Hugh W Hubble, and Keith A Horton. Stand-off Raman spectroscopic detection of minerals on planetary surfaces. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 59(10):2391–2407, 2003.

[144] John M Chalmers, Howell G M Edwards, and Michael D Hargreaves, editors. *Infrared and Raman Spectroscopy in Forensic Science*. Wiley Online Library, 2012.

[145] Michael D. Hargreaves, Robert L. Green, Wayne Jalenak, Christopher D. Brown, and Craig Gardner. *Handheld Raman and FT-IR Spectrometers*, chapter 5.3, pages 275–287. John Wiley & Sons, Ltd, 2012. ISBN 9781119962328.

[146] Edward G. Bartick. *Criminal Forensic Analysis*, chapter 4, pages 87–109. John Wiley & Sons, Ltd, 2012. ISBN 9781119962328.

[147] Dong Yu, Wei Zhu, and Ai-Guo Shen. Raman encoding for security labels: a review. *Nanoscale Advances*, 5(23):6365–6381, 2023.

[148] Yuqing Gu, Chang He, Yuqing Zhang, Li Lin, Benjamin David Thackray, and Jian Ye. Gap-enhanced Raman tags for physically unclonable anticounterfeiting labels. *Nature communications*, 11(1):516, 2020.

[149] Yusai Zhou, Gui Zhao, Jingming Bian, Xiaoli Tian, Xuejun Cheng, Hong Wang, and Hongyu Chen. Multiplexed SERS barcodes for anti-counterfeiting. *ACS Applied Materials & Interfaces*, 12(25):28532–28538, 2020.

[150] Yan Cui, Ravi S Hegde, In Yee Phang, Hiang Kwee Lee, and Xing Yi Ling. Encoding molecular information in plasmonic nanostructures for anti-counterfeiting applications. *Nanoscale*, 6(1):282–288, 2014.

[151] Yan Cui, In Yee Phang, Yih Hong Lee, Mian Rong Lee, Qi Zhang, and Xing Yi Ling. Multiplex plasmonic anti-counterfeiting security labels based on surface-enhanced Raman scattering. *Chemical communications*, 51(25):5363–5366, 2015.

[152] Anil K Deisingh. Pharmaceutical counterfeiting. *Analyst*, 130(3):271–279, 2005.

[153] Mariana R de Almeida, Deleon N Correa, Werickson FC Rocha, Francisco JO Scafi, and Ronei J Poppi. Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation. *Microchemical Journal*, 109:170–177, 2013.

[154] Oleg Ryabchykov, Shuxia Guo, and Thomas Bocklitz. Analyzing Raman spectroscopic data. *Physical Sciences Reviews*, 4(2):20170043, 2018.

[155] Ravi teja Vulchi, Volodymyr Morgunov, Rajendhar Junjuri, and Thomas Bocklitz. Artifacts and anomalies in Raman spectroscopy review on origins and correction procedures. *Molecules*, 29(19): 4748, 2024.

[156] Shuxia Guo, Claudia Beleites, Ute Neugebauer, Sara Abalde-Cela, Nils Kristian Afseth, Fatima Alsamad, Suresh Anand, Cuauhtemoc Araujo-Andrade, Sonja Askrabic, Ertug Avci, et al. Comparability of Raman spectroscopic configurations: a large scale cross-laboratory study. *Analytical chemistry*, 92(24):15745–15756, 2020.

[157] Thomas Bocklitz, Angela Walter, Katharina Hartmann, Petra Rösch, and Jürgen Popp. How to pre-process Raman spectra for reliable and stable models? *Analytica chimica acta*, 704(1-2):47–56, 2011.

[158] Camilo LM Morais, Kássio MG Lima, Maneesh Singh, and Francis L Martin. Tutorial: multivariate classification for vibrational spectroscopy in biological samples. *Nature Protocols*, 15(7):2143–2162, 2020.

[159] Darren A Whitaker and Kevin Hayes. A simple algorithm for despiking Raman spectra. *Chemometrics and Intelligent Laboratory Systems*, 179:82–84, 2018.

[160] F Ehrentreich and L Sümmchen. Spike removal and denoising of Raman spectra by wavelet transform methods. *Analytical chemistry*, 73(17):4364–4373, 2001.

[161] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[162] Paul HC Eilers. A perfect smoother. *Analytical chemistry*, 75(14):3631–3636, 2003.

[163] Chad A Lieber and Anita Mahadevan-Jansen. Automated method for subtraction of fluorescence from biological Raman spectra. *Applied spectroscopy*, 57(11):1363–1367, 2003.

[164] Jianhua Zhao, Harvey Lui, David I McLean, and Haishan Zeng. Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy. *Applied spectroscopy*, 61 (11):1225–1232, 2007.

[165] Paul HC Eilers and Hans FM Boelens. Baseline correction with asymmetric least squares smoothing. *Leiden University Medical Centre Report*, 1(1):5, 2005.

[166] Shixuan He, Wei Zhang, Lijuan Liu, Yu Huang, Jiming He, Wanyi Xie, Peng Wu, and Chunlei Du. Baseline correction for Raman spectra using an improved asymmetric least squares method. *Analytical Methods*, 6(12):4402–4407, 2014.

[167] Sung-June Baek, Aaron Park, Young-Jin Ahn, and Jaebum Choo. Baseline correction using asymmetrically reweighted penalized least squares smoothing. *Analyst*, 140(1):250–257, 2015.

[168] J Carlos Cobas, Michael A Bernstein, Manuel Martín-Pastor, and Pablo García Tahoces. A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data. *Journal of Magnetic Resonance*, 183(1):145–151, 2006.

[169] Zhong Li, De-Jian Zhan, Jia-Jun Wang, Jing Huang, Qing-Song Xu, Zhi-Min Zhang, Yi-Bao Zheng, Yi-Zeng Liang, and Hong Wang. Morphological weighted penalized least squares for background correction. *Analyst*, 138(16):4483–4492, 2013.

[170] Mahalanobis Prasanta Chandra et al. On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India*, volume 2, pages 49–55, 1936.

[171] Daniël Lakens. Calculating and reporting effect sizes to facilitate cumulative science: a practical

primer for t-tests and ANOVAs. *Frontiers in psychology*, 4:863, 2013.

[172] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

[173] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[174] Ian Jolliffe. Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer, 2011.

[175] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[176] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[177] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 5, pages 281–298. University of California press, 1967.

[178] Stuart Lloyd. Least squares quantization in PCM. *IEEE transactions on information theory*, 28 (2):129–137, 1982.

[179] Robin Sibson. SLINK: an optimally efficient algorithm for the single-link cluster method. *The computer journal*, 16(1):30–34, 1973.

[180] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[181] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[182] Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis – a brief tutorial. *Institute for Signal and information Processing*, 18(1998):1–8, 1998.

[183] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[184] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[185] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[186] Paul Geladi and Bruce R Kowalski. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17, 1986.

[187] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[188] Steffen Möller, Stuart W Prescott, Lars Wirzenius, Petter Reinholdtsen, Brad Chapman, Pjotr Prins, Stian Soiland-Reyes, Fabian Klötzl, Andrea Bagnacani, Matúš Kalaš, 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.

[189] Afroditi Ntziouni, James Thomson, Ioannis Xiarchos, Xiang Li, Miguel A Bañares, Costas Charitidis, Raquel Portela, and Enrique Lozano Diz. Review of Existing Standards, Guides, and Practices for Raman Spectroscopy. *Applied Spectroscopy*, 76(7):747–772, 2022.

[190] Yaping Qi, Dan Hu, Yucheng Jiang, Zhenping Wu, Ming Zheng, Esther Xinyi Chen, Yong Liang, Mohammad A Sadi, Kang Zhang, and Yong P Chen. Recent Progresses in Machine Learning Assisted Raman Spectroscopy. *Advanced Optical Materials*, page 2203104, 2023.

[191] Xiaoqin Zhu, Tao Xu, Qingyu Lin, and Yixiang Duan. Technical development of Raman spectroscopy: from instrumental to advanced combined technologies. *Applied Spectroscopy Reviews*, 49

(1):64–82, 2014.

[192] Charles L Lawson and Richard J Hanson. *Solving least squares problems*. SIAM, 1995.

[193] Félix Lussier, Vincent Thibault, Benjamin Charron, Gregory Q Wallace, and Jean-Francois Masson. Deep learning and artificial intelligence methods for Raman and surface-enhanced Raman scattering. *TrAC Trends in Analytical Chemistry*, 124:115796, 2020.

[194] Jie Yang, Jinfan Xu, Xiaolei Zhang, Chiyu Wu, Tao Lin, and Yibin Ying. Deep learning for vibrational spectral analysis: Recent progress and a practical guide. *Analytica chimica acta*, 1081: 6–17, 2019.

[195] Ruihao Luo, Juergen Popp, and Thomas Bocklitz. Deep learning for Raman spectroscopy: a review. *Analytica*, 3(3):287–301, 2022.

[196] Carlos A Meza Ramirez, Michael Greenop, Lorna Ashton, and Ihtesham Ur Rehman. Applications of machine learning in spectroscopy. *Applied Spectroscopy Reviews*, 56(8-10):733–763, 2021.

[197] Xinyuan Bi, Li Lin, Zhou Chen, and Jian Ye. Artificial Intelligence for Surface-Enhanced Raman Spectroscopy. *Small methods*, 8(1):2301243, 2024.

[198] Weilai Lu, Xiuqiang Chen, Lu Wang, Hanfei Li, and Yu Vincent Fu. Combination of an artificial intelligence approach and laser tweezers Raman spectroscopy for microbial identification. *Analytical chemistry*, 92(9):6288–6296, 2020.

[199] Shuaishuai Yan, Shuying Wang, Jingxuan Qiu, Menghua Li, Dezhi Li, Dongpo Xu, Daixi Li, and Qing Liu. Raman spectroscopy combined with machine learning for rapid detection of food-borne pathogens at the single-cell level. *Talanta*, 226:122195, 2021.

[200] Liang Wang, Wei Liu, Jia-Wei Tang, Jun-Jiao Wang, Qing-Hua Liu, Peng-Bo Wen, Meng-Meng Wang, Ya-Cheng Pan, Bing Gu, and Xiao Zhang. Applications of Raman spectroscopy in bacterial infections: principles, advantages, and shortcomings. *Frontiers in Microbiology*, 12:683580, 2021.

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

[202] Jiaji Zhu, Arumugam Selva Sharma, Jing Xu, Yi Xu, Tianhui Jiao, Qin Ouyang, Huanhuan Li, and Quansheng Chen. 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.

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

[204] Hacer Akpolat, Mark Barineau, Keith A Jackson, Mehmet Z Akpolat, David M Francis, Yu-Ju Chen, and Luis E Rodriguez-Saona. High-Throughput Phenotyping Approach for Screening Major Carotenoids of Tomato by Handheld Raman Spectroscopy Using Chemometric Methods. *Sensors*, 20(13):3723, 2020.

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

[206] Abdullah CS Talari, Shazza Rehman, and Ihtesham U Rehman. 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.

[207] Howard Peng Sin Heng, Chi Shu, Wei Zheng, Kan Lin, and Zhiwei Huang. 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.

[208] Lihao Zhang, Chengjian Li, Di Peng, Xiaofei Yi, Shuai He, Fengxiang Liu, Xiangtai Zheng, Wei E

Huang, Liang Zhao, and Xia Huang. Raman spectroscopy and machine learning for the classification of breast cancers. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 264:120300, 2022.

[209] Conor C Horgan, Magnus Jensen, Anika Nagelkerke, Jean-Philippe St-Pierre, Tom Vercauteren, Molly M Stevens, and Mads S Bergholt. High-throughput molecular imaging via deep-learning-enabled Raman spectroscopy. *Analytical chemistry*, 93(48):15850–15860, 2021.

[210] Nabil Ibtehaz, Muhammad EH Chowdhury, Amith Khandakar, Serkan Kiranyaz, M Sohel Rahman, and Susu M Zughaier. RamanNet: a generalized neural network architecture for Raman spectrum analysis. *Neural Computing and Applications*, 35(25):18719–18735, 2023.

[211] Jacopo Acquarelli, Twan van Laarhoven, Jan Gerretzen, Thanh N Tran, Lutgarde MC Buydens, and Elena Marchiori. Convolutional neural networks for vibrational spectroscopic data analysis. *Analytica chimica acta*, 954:22–31, 2017.

[212] Jinchao Liu, Margarita Osadchy, Lorna Ashton, Michael Foster, Christopher J Solomon, and Stuart J Gibson. Deep convolutional neural networks for Raman spectrum recognition: a unified solution. *Analyst*, 142(21):4067–4074, 2017.

[213] Peng Wang, Liangsheng Guo, Yubing Tian, Jiansheng Chen, Shan Huang, Ce Wang, Pengli Bai, Daqing Chen, Weipei Zhu, Hongbo Yang, et al. Discrimination of blood species using Raman spectroscopy combined with a recurrent neural network. *OSA continuum*, 4(2):672–687, 2021.

[214] Onur Can Koyun, Reyhan Kevser Keser, Safa Onur Şahin, Damla Bulut, Mustafa Yorulmaz, Veysel Yücesoy, and Behçet Uğur Töreyin. RamanFormer: A Transformer-Based Quantification Approach for Raman Mixture Components. *ACS omega*, 2024.

[215] Ziyan Zhao, Zhenfang Liu, Mingqiang Ji, Xin Zhao, Qibing Zhu, and Min Huang. ConInceDeep: A novel deep learning method for component identification of mixture based on Raman spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 234:104757, 2023.

[216] Liangrui Pan, Pronthep Pipitsunthonsan, Chalongrat Daengngam, Sittiporn Channumsin, Suwat Sreesawet, and Mitchai Chongcheawchamnan. Identification of complex mixtures for Raman spectroscopy using a novel scheme based on a new multi-label deep neural network. *IEEE Sensors Journal*, 21(9):10834–10843, 2021.

[217] William John Thrift and Regina Ragan. Quantification of analyte concentration in the single molecule regime using convolutional neural networks. *Analytical chemistry*, 91(21):13337–13342, 2019.

[218] Zilong Wang, Yunfeng Li, Jinglei Zhai, Siwei Yang, Biao Sun, and Pei Liang. Deep learning-based Raman spectroscopy qualitative analysis algorithm: A convolutional neural network and transformer approach. *Talanta*, 275:126138, 2024.

[219] Bryce Manifold, Shuaiqian Men, Ruoqian Hu, and Dan Fu. A versatile deep learning architecture for classification and label-free prediction of hyperspectral images. *Nature machine intelligence*, 3 (4):306–315, 2021.

[220] Jing Zhang, Jian Zhao, Haonan Lin, Yuying Tan, and Ji-Xin Cheng. High-Speed Chemical Imaging by Dense-Net Learning of Femtosecond Stimulated Raman Scattering. *The journal of physical chemistry letters*, 11(20):8573–8578, 2020.

[221] Zhihao Wei, Xi Liu, Ruiqing Yan, Guocheng Sun, Weiyong Yu, Qiang Liu, and Qianjin Guo. Pixel-level multimodal fusion deep networks for predicting subcellular organelle localization from label-free live-cell imaging. *Frontiers in Genetics*, 13:1002327, 2022.

[222] Xiaoli Li, Ruiqing Zhou, Yifei Xu, Xuan Wei, and Yong He. Spectral unmixing combined with Raman imaging, a preferable analytic technique for molecule visualization. *Applied Spectroscopy Reviews*, 52(5):417–438, 2017.

[223] Victor Olmos, Laura Benitez, Monica Marro, Pablo Loza-Alvarez, Benjami Pina, Roma Tauler,

and Anna De Juan. Relevant aspects of unmixing/resolution analysis for the interpretation of biological vibrational hyperspectral images. *TrAC Trends in Analytical Chemistry*, 94:130–140, 2017.

[224] José M Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn Chanussot. Hyperspectral Unmixing Overview: Geometrical, Statistical, and Sparse Regression-Based Approaches. *IEEE journal of selected topics in applied earth observations and remote sensing*, 5(2):354–379, 2012.

[225] Nirmal Keshava and John F Mustard. Spectral unmixing. *IEEE signal processing magazine*, 19 (1):44–57, 2002.

[226] Nicolas Dobigeon, Jean-Yves Tourneret, Cédric Richard, José Carlos M Bermudez, Stephen McLaughlin, and Alfred O Hero. Nonlinear unmixing of hyperspectral images: Models and algorithms. *IEEE Signal processing magazine*, 31(1):82–94, 2013.

[227] Rob Heylen, Mario Parente, and Paul Gader. A review of nonlinear hyperspectral unmixing methods. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7 (6):1844–1868, 2014.

[228] Bruce Hapke. Bidirectional reflectance spectroscopy: 1. Theory. *Journal of Geophysical Research: Solid Earth*, 86(B4):3039–3054, 1981.

[229] Abderrahim Halimi, Yoann Altmann, Nicolas Dobigeon, and Jean-Yves Tourneret. Nonlinear Unmixing of Hyperspectral Images Using a Generalized Bilinear Model. *IEEE Transactions on Geoscience and Remote Sensing*, 49(11):4153–4162, 2011. doi: 10.1109/TGRS.2010.2098414.

[230] Wenyi Fan, Baoxin Hu, John Miller, and Mingze Li. Comparative study between a new nonlinear model and common linear model for analysing laboratory simulated-forest hyperspectral data. *International Journal of Remote Sensing*, 30(11):2951–2962, 2009.

[231] Abderrahim Halimi, Yoann Altmann, Nicolas Dobigeon, and Jean-Yves Tourneret. Unmixing hyperspectral images using the generalized bilinear model. In *2011 IEEE International Geoscience and Remote Sensing Symposium*, pages 1886–1889. IEEE, 2011.

[232] Rob Heylen and Paul Scheunders. A multilinear mixing model for nonlinear spectral unmixing. *IEEE transactions on geoscience and remote sensing*, 54(1):240–251, 2015.

[233] Yoann Altmann, Abderrahim Halimi, Nicolas Dobigeon, and Jean-Yves Tourneret. Supervised nonlinear spectral unmixing using a postnonlinear mixing model for hyperspectral imagery. *IEEE Transactions on Image Processing*, 21(6):3017–3025, 2012.

[234] Michael E. Winter. N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data. In Michael R. Descour and Sylvia S. Shen, editors, *Imaging Spectrometry V*, volume 3753, pages 266 – 275. International Society for Optics and Photonics, SPIE, 1999.

[235] José MP Nascimento and José MB Dias. Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE transactions on Geoscience and Remote Sensing*, 43(4):898–910, 2005.

[236] Daniel C Heinz et al. Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. *IEEE transactions on geoscience and remote sensing*, 39(3):529–545, 2001.

[237] Martin Hedegaard, Christian Matthäus, Søren Hassing, Christoph Krafft, Max Diem, and Jürgen Popp. Spectral unmixing and clustering algorithms for assessment of single cells by Raman microscopic imaging. *Theoretical Chemistry Accounts*, 130:1249–1260, 2011.

[238] A Thomas Harris. Spectral mapping tools from the earth sciences applied to spectral microscopy data. *Cytometry Part A: The Journal of the International Society for Analytical Cytology*, 69(8): 872–879, 2006.

[239] Burkni Palsson, Johannes R Sveinsson, and Magnus O Ulfarsson. Blind hyperspectral unmixing using autoencoders: A critical comparison. *IEEE Journal of Selected Topics in Applied Earth*

Observations and Remote Sensing, 15:1340–1372, 2022.

[240] Shaoquan Zhang, Yuanchao Su, Xiang Xu, Jun Li, Chengzhi Deng, and Antonio Plaza. Recent advances in hyperspectral unmixing using sparse techniques and deep learning. *Hyperspectral Image Analysis: Advances in Machine Learning and Signal Processing*, pages 377–405, 2020.

[241] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):213–247, 2022.

[242] Jignesh S Bhatt and Manjunath V Joshi. Deep learning in hyperspectral unmixing: A review. In *IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 2189–2192. IEEE, 2020.

[243] Jie Chen, Min Zhao, Xiuheng Wang, Cédric Richard, and Susanto Rahardja. Integration of physics-based and data-driven models for hyperspectral image unmixing: A summary of current methods. *IEEE Signal Processing Magazine*, 40(2):61–74, 2023.

[244] Nikolas Burzynski, Yuhao Yuan, Adiel Felsen, David Reitano, Zhibo Wang, Khalid A Sethi, Fake Lu, and Kenneth Chiu. Deep Learning Techniques for Unmixing of Hyperspectral Stimulated Raman Scattering Images. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5862–5864. IEEE, 2021.

[245] Damien Boildieu, David Helbert, Amandine Magnaudeix, Philippe Leproux, and Philippe Carré. Multivariate curve resolution with autoencoders for CARS microspectroscopy. In *Computational Imaging Conference, IS&T Electronic Imaging*, 2023.

[246] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Autoencoders. In *Deep learning*, chapter 14, pages 499–523. MIT Press, 2016.

[247] Kazi Tanzeem Shahid and Ioannis D Schizas. Unsupervised hyperspectral unmixing via nonlinear autoencoders. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021.

[248] Min Zhao, Mou Wang, Jie Chen, and Susanto Rahardja. Hyperspectral unmixing for additive nonlinear models with a 3-D-CNN autoencoder network. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2021.

[249] Olaf Sporns, Giulio Tononi, and Rolf Kötter. The human connectome: a structural description of the human brain. *PLoS computational biology*, 1(4):e42, 2005.

[250] KA Maldonado and K Alsayouri. *Physiology, Brain*. StatPearls Publishing, Treasure Island (FL), 2024.

[251] Danielle S Bassett and Michael S Gazzaniga. Understanding complexity in the human brain. *Trends in cognitive sciences*, 15(5):200–209, 2011.

[252] Stanley Finger. *Origins of neuroscience: a history of explorations into brain function*. Oxford University Press, 2001.

[253] Maxwell R Bennett and Peter Michael Stephan Hacker. *Philosophical foundations of neuroscience*. John Wiley & Sons, 2022.

[254] Russell A Poldrack and Martha J Farah. Progress and challenges in probing the human brain. *Nature*, 526(7573):371–379, 2015.

[255] Noelia Urbán and François Guillemot. Neurogenesis in the embryonic and adult brain: same regulators, different roles. *Frontiers in cellular neuroscience*, 8:396, 2014.

[256] P Mukherjee, S Roy, D Ghosh, and SK Nandi. Role of animal models in biomedical research: a review. *Laboratory Animal Research*, 38(1):18, 2022.

[257] L Syd M Johnson. The trouble with animal models in brain research. In *Neuroethics and nonhuman animals*, pages 271–286. Springer, 2020.

[258] Philipp J Keller. Imaging morphogenesis: technological advances and biological insights. *Science*, 340(6137):1234168, 2013.

[259] Françoise Barré-Sinoussi and Xavier Montagutelli. Animal models are essential to biological research: issues and perspectives. *Future science OA*, 1(4):FSO63–FSO63, 2015.

[260] N. Bryce Robinson, Katherine Krieger, Faiza M. Khan, William Huffman, Michelle Chang, Ajita Naik, Ruan Yongle, Irbaz Hameed, Karl Krieger, Leonard N. Girardi, and Mario Gaudino. The current state of animal models in research: A review. *International Journal of Surgery*, 72:9–13, 2019.

[261] Judith A Tello, Haley E Williams, Robert M Eppler, Michelle L Steinhilb, and May Khanna. Animal models of neurodegenerative disease: recent advances in fly highlight innovative approaches to drug discovery. *Frontiers in molecular neuroscience*, 15:883358, 2022.

[262] Debby Van Dam and Peter Paul De Deyn. Animal models for brain research. In *PET and SPECT of Neurobiological Systems*, pages 3–55. Springer, 2020.

[263] Urban Lendahl and Ronald DG McKay. The use of cell lines in neurobiology. *Trends in neurosciences*, 13(4):132–137, 1990.

[264] Stephen D Skaper, Laura Facci, Daria Milani, Alberta Leon, and Gino Toffano. Culture and use of primary and clonal neural cells. In *Methods in neurosciences*, volume 2, pages 17–33. Elsevier, 1990.

[265] Gary Banker and Kimberly Goslin. *Culturing nerve cells*. The MIT Press, 1998.

[266] Jane Bottenstein. *Cell culture in the neurosciences*. Springer Science & Business Media, 2012.

[267] Sergey Fedoroff. *Cell, tissue, and organ cultures in neurobiology*. Elsevier, 2012.

[268] Seongeun Cho, Andrew Wood, and Mark R Bowlby. Brain slices as models for neurodegenerative disease and screening platforms to identify novel therapeutics. *Current neuropharmacology*, 5(1): 19–33, 2007.

[269] C Humpel. Organotypic brain slice cultures review. *Neuroscience*, 305:86–98, 2015.

[270] Merve Alaylioğlu, Erdinç Dursun, Selma Yilmazer, and Duygu Gezen Ak. A bridge between in vitro and in vivo studies in neuroscience: Organotypic brain slice cultures. *Archives of Neuropsychiatry*, 57(4):333, 2020.

[271] Xin-Rui Qi, Ronald WH Verwer, Ai-Min Bao, Rawien A Balesar, Sabina Luchetti, Jiang-Ning Zhou, and Dick F Swaab. Human brain slice culture: a useful tool to study brain disorders and potential therapeutic compounds. *Neuroscience bulletin*, 35(2):244–252, 2019.

[272] Bart Ellenbroek and Jiun Youn. Rodent models in neuroscience research: is it a rat race? *Disease models & mechanisms*, 9(10):1079–1087, 2016.

[273] Bridgette D Semple, Klas Blomgren, Kayleen Gimlin, Donna M Ferriero, and Linda J Noble-Haeusslein. Brain development in rodents and humans: Identifying benchmarks of maturation and vulnerability to injury across species. *Progress in neurobiology*, 106:1–16, 2013.

[274] Heiko J Luhmann and Atsuo Fukuda. Can we understand human brain development from experimental studies in rodents? *Pediatrics International*, 62(10):1139–1144, 2020.

[275] M Angela Cenci, Ian Q Whishaw, and Timothy Schallert. Animal models of neurological deficits: how relevant is the rat? *Nature Reviews Neuroscience*, 3(7):574–579, 2002.

[276] Judith R Homberg. Measuring behaviour in rodents: towards translational neuropsychiatric research. *Behavioural brain research*, 236:295–306, 2013.

[277] Marija Markicevic, Iurii Savvateev, Christina Grimm, and Valerio Zerbi. Emerging imaging methods to study whole-brain function in rodent models. *Translational psychiatry*, 11(1):457, 2021.

[278] Nanna Marie Lind, Anette Moustgaard, Jacob Jelsing, Gabor Vajta, Paul Cumming, and Axel K Hansen. The use of pigs in neuroscience: modeling brain disorders. *Neuroscience & Biobehavioral Reviews*, 31(5):728–751, 2007.

[279] Nóra Bunford, Attila Andics, Anna Kis, Ádám Miklósi, and Márta Gácsi. Canis familiaris as a model for non-invasive comparative neuroscience. *Trends in neurosciences*, 40(7):438–452, 2017.

[280] Carlotta Gilardi and Nereo Kalebic. The ferret as a model system for neocortex development and evolution. *Frontiers in Cell and Developmental Biology*, 9:661759, 2021.

[281] David H Abbott, Deborah K Barnett, Ricki J Colman, M Emilia Yamamoto, and Nancy J Schultz-Darken. Aspects of common marmoset basic biology and life history important for biomedical research. *Comparative medicine*, 53(4):339–350, 2003.

[282] John P Capitanio and Marina E Emborg. Contributions of non-human primates to neuroscience research. *The Lancet*, 371(9618):1126–1135, 2008.

[283] Juan Carlos Izpisua Belmonte, Edward M Callaway, Sarah J Caddick, Patricia Churchland, Guoping Feng, Gregg E Homanics, Kuo-Fen Lee, David A Leopold, Cory T Miller, Jude F Mitchell, et al. Brains, genes, and primates. *Neuron*, 86(3):617–631, 2015.

[284] Charles G Jennings, Rogier Landman, Yang Zhou, Jitendra Sharma, Julia Hyman, J Anthony Movshon, Zilong Qiu, Angela C Roberts, Anna Wang Roe, Xiaoqin Wang, et al. Opportunities and challenges in modeling human brain disorders in transgenic primates. *Nature neuroscience*, 19 (9):1123–1130, 2016.

[285] US Food and Drug Administration. The Drug Development Process, 2018. https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process. Online; accessed 24 August 2025.

[286] Richard C Mohs and Nigel H Greig. Drug discovery and development: Role of basic biological research. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(4):651–657, 2017.

[287] Kayla Duval, Hannah Grover, Li-Hsin Han, Yongchao Mou, Adrian F Pegoraro, Jeffery Fredberg, and Zi Chen. Modeling physiological events in 2D vs. 3D cell culture. *Physiology*, 32(4):266–277, 2017.

[288] Caleb Jensen and Yong Teng. Is it time to start transitioning from 2D to 3D cell culture? *Frontiers in molecular biosciences*, 7:33, 2020.

[289] Eduarda GZ Centeno, Helena Cimarosti, and Angela Bithell. 2D versus 3D human induced pluripotent stem cell-derived cultures for neurodegenerative disease modelling. *Molecular neurodegeneration*, 13(1):27, 2018.

[290] Shubhada V Mangrulkar, Shalakha V Dabhekar, Pankaj Neje, Nivedita Parkarwar, Ankush Turankar, Brijesh G Taksande, Milind J Umekar, and Kartik T Nakhate. In vivo animal models development and their limitations for brain research. In *Application of Nanocarriers in Brain Delivery of Therapeutics*, pages 315–339. Springer, 2024.

[291] Jan H Lui, David V Hansen, and Arnold R Kriegstein. Development and evolution of the human neocortex. *Cell*, 146(1):18–36, 2011.

[292] Suzana Herculano-Houzel. The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in human neuroscience*, 3:857, 2009.

[293] John C Silbereis, Sirisha Pochareddy, Ying Zhu, Mingfeng Li, and Nenad Sestan. The cellular and molecular landscapes of the developing human central nervous system. *Neuron*, 89(2):248–268, 2016.

[294] Felipe Mora-Bermúdez, Farhath Badsha, Sabina Kanton, J Gray Camp, Benjamin Vernot, Kathrin Köhler, Birger Voigt, Keisuke Okita, Tomislav Maricic, Zhisong He, et al. Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. *Elife*, 5:e18683, 2016.

[295] Tomoki Otani, Maria C Marchetto, Fred H Gage, Benjamin D Simons, and Frederick J Livesey. 2D and 3D stem cell models of primate cortical development identify species-specific differences in progenitor behavior contributing to brain size. *Cell stem cell*, 18(4):467–480, 2016.

[296] David V Hansen, Jan H Lui, Philip RL Parker, and Arnold R Kriegstein. Neurogenic radial glia

in the outer subventricular zone of human neocortex. *Nature*, 464(7288):554–561, 2010.

[297] Simone A Fietz, Iva Kelava, Johannes Vogt, Michaela Wilsch-Bräuninger, Denise Stenzel, Jennifer L Fish, Denis Corbeil, Axel Riehn, Wolfgang Distler, Robert Nitsch, et al. OSVZ progenitors of human and ferret neocortex are epithelial-like and expand by integrin signaling. *Nature neuroscience*, 13(6):690–699, 2010.

[298] André MM Sousa, Kyle A Meyer, Gabriel Santpere, Forrest O Gulden, and Nenad Sestan. Evolution of the human nervous system function, structure, and development. *Cell*, 170(2):226–247, 2017.

[299] Xinyu Zhao and Anita Bhattacharyya. Human models are needed for studying human neurodevelopmental disorders. *The American Journal of Human Genetics*, 103(6):829–857, 2018.

[300] Eric J Nestler and Steven E Hyman. Animal models of neuropsychiatric disorders. *Nature neuroscience*, 13(10):1161–1169, 2010.

[301] Alexander Opitz, Arnaud Falchier, Gary S Linn, Michael P Milham, and Charles E Schroeder. Limitations of ex vivo measurements for in vivo neuroscience. *Proceedings of the National Academy of Sciences*, 114(20):5243–5246, 2017.

[302] David B Resnik. *The ethics of research with human subjects: Protecting people, advancing science, promoting trust*, volume 74. Springer, 2018.

[303] Gerard J Boer. Ethical guidelines for the use of human embryonic or fetal tissue for experimental and clinical neurotransplantation and research. *Journal of Neurology*, 242(1):1–13, 1994.

[304] Bonnie Steinbock. *Life before birth: the moral and legal status of embryos and fetuses*. Oxford University Press, 2011.

[305] Annelien L Bredenoord, Hans Clevers, and Juergen A Knoblich. Human tissues in a dish: the research and ethical implications of organoid technology. *Science*, 355(6322):eaaf9414, 2017.

[306] Aysha Karim Kiani, Derek Pheby, Gary Henehan, Richard Brown, Paul Sieving, Peter Sykora, Robert Marks, Benedetto Falsini, Natale Capodicasa, Stanislav Miertus, et al. Ethical considerations regarding animal experimentation. *Journal of preventive medicine and hygiene*, 63(2 Suppl 3):E255, 2022.

[307] National Research Council, Division on Earth, Institute for Laboratory Animal Research, Committee for the Update of the Guide for the Care, and Use of Laboratory Animals. *Guide for the care and use of laboratory animals*. National Academies Press, 2010.

[308] Asher Mullard. Parsing clinical success rates. *Nature Reviews Drug Discovery*, 15(7):447–448, 2016.

[309] Valentin K Gribkoff and Leonard K Kaczmarek. The need for new approaches in CNS drug discovery: Why drugs have failed, and what can be done to improve outcomes. *Neuropharmacology*, 120:11–19, 2017.

[310] Pandora Pound and Merel Ritskes-Hoitinga. Is it possible to overcome issues of external validity in preclinical animal research? why most animal models are bound to fail. *Journal of translational medicine*, 16(1):304, 2018.

[311] V. L. Feigin and T. Vos. Global Burden of Neurological Disorders: From Global Burden of Disease Estimates to Actions. *Neuroepidemiology*, 52(1-2):1–2, 2019.

[312] GBD 2016 Neurology Collaborators. Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet. Neurology*, 18(5):459–480, May 2019.

[313] World Health Organization. Neurological disorders : public health challenges, 2006.

[314] US Food, Drug Administration, et al. Roadmap to reducing animal testing in preclinical safety studies, 2024.

[315] HV Wilson. A new method by which sponges may be artificially reared. *Science*, 25(649):912–915,

1907.

[316] Ross G Harrison. Observations on the living developing nerve fiber. *Proceedings of the society for experimental biology and medicine*, 4(1):140–143, 1906.

[317] Alexis Carrel and Montrose T Burrows. Cultivation of adult tissues and organs outside of the body. *Journal of the American medical association*, 55(16):1379–1381, 1910.

[318] Umberto Galderisi, Francesco Paolo Jori, and Antonio Giordano. Cell cycle regulation and neural differentiation. *Oncogene*, 22(33):5208–5219, 2003.

[319] Jaap Van Pelt, Ildiko Vajda, Pieter S Wolters, Michael A Corner, and Ger JA Ramakers. Dynamics and plasticity in developing neuronal networks in vitro. *Progress in brain research*, 147:171–188, 2005.

[320] Laurence A Borden. GABA transporter heterogeneity: pharmacology and cellular localization. *Neurochemistry international*, 29(4):335–356, 1996.

[321] Latha Devi, Vijayendran Raghavendran, Badanavalu M Prabhu, Narayan G Avadhani, and Hindupur K Anandatheerthavarada. Mitochondrial import and accumulation of α-synuclein impair complex I in human dopaminergic neuronal cultures and Parkinson disease brain. *Journal of Biological Chemistry*, 283(14):9089–9100, 2008.

[322] Laura Gasparini, Laura Rusconi, Huaxi Xu, Piero Del Soldato, and Ennio Ongini. Modulation of β-amyloid metabolism by non-steroidal anti-inflammatory drugs in neuronal cell cultures. *Journal of neurochemistry*, 88(2):337–348, 2004.

[323] Jennifer Gordon and Shohreh Amini. General overview of neuronal cell culture. *Neuronal Cell Culture: Methods and Protocols*, pages 1–8, 2021.

[324] Gail R Martin. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences*, 78 (12):7634–7638, 1981.

[325] Martin J Evans and Matthew H Kaufman. Establishment in culture of pluripotential cells from mouse embryos. *nature*, 292(5819):154–156, 1981.

[326] James A Thomson, Joseph Itskovitz-Eldor, Sander S Shapiro, Michelle A Waknitz, Jennifer J Swiergiel, Vivienne S Marshall, and Jeffrey M Jones. Embryonic stem cell lines derived from human blastocysts. *Science*, 282(5391):1145–1147, 1998.

[327] John A Robertson. Embryo stem cell research: ten years of controversy. *Journal of Law, Medicine & Ethics*, 38(2):191–203, 2010.

[328] Roger A Barker and Inez de Beaufort. Scientific and ethical issues related to stem cell research and interventions in neurodegenerative disorders of the brain. *Progress in neurobiology*, 110:63–73, 2013.

[329] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–676, 2006.

[330] Kazutoshi Takahashi, Koji Tanabe, Mari Ohnuki, Megumi Narita, Tomoko Ichisaka, Kiichiro Tomoda, and Shinya Yamanaka. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*, 131(5):861–872, 2007.

[331] Jerome Mertens, Maria C Marchetto, Cedric Bardy, and Fred H Gage. Evaluating cell reprogramming, differentiation and conversion technologies in neuroscience. *Nature Reviews Neuroscience*, 17(7):424–437, 2016.

[332] Sujoy K Dhara and Steven L Stice. Neural differentiation of human embryonic stem cells. *Journal of cellular biochemistry*, 105(3):633–640, 2008.

[333] Leroy C Stevens. Embryonic potency of embryoid bodies derived from a transplantable testicular teratoma of the mouse. *Developmental biology*, 2(3):285–297, 1960.

[334] Gail R Martin and Martin J Evans. Differentiation of clonal lines of teratocarcinoma cells: forma-

tion of embryoid bodies in vitro. *Proceedings of the National Academy of Sciences*, 72(4):1441–1445, 1975.

[335] Joseph Itskovitz-Eldor, Maya Schuldiner, Dorit Karsenti, Amir Eden, Ofra Yanuka, Michal Amit, Hermona Soreq, and Nissim Benvenisty. Differentiation of human embryonic stem cells into embryoid bodies comprising the three embryonic germ layers. *Molecular medicine*, 6(2):88–95, 2000.

[336] Su-Chun Zhang, Marius Wernig, Ian D Duncan, Oliver Brüstle, and James A Thomson. In vitro differentiation of transplantable neural precursors from human embryonic stem cells. *Nature Biotechnology*, 19(12):1129–1133, 2001.

[337] Maya Schuldiner, Ofra Yanuka, Joseph Itskovitz-Eldor, Douglas A Melton, and Nissim Benvenisty. Effects of eight growth factors on the differentiation of cells derived from human embryonic stem cells. *Proceedings of the National Academy of Sciences*, 97(21):11307–11312, 2000.

[338] Kiichi Watanabe, Daisuke Kamiya, Ayaka Nishiyama, Tomoko Katayama, Satoshi Nozaki, Hiroshi Kawasaki, Yasuyoshi Watanabe, Kenji Mizuseki, and Yoshiki Sasai. Directed differentiation of telencephalic precursors from embryonic stem cells. *Nature neuroscience*, 8(3):288–296, 2005.

[339] Stuart M Chambers, Christopher A Fasano, Eirini P Papapetrou, Mark Tomishima, Michel Sadelain, and Lorenz Studer. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nature Biotechnology*, 27(3):275–280, 2009.

[340] David B Edelman and Edward W Keefer. A cultural renaissance: in vitro cell biology embraces three-dimensional context. *Experimental neurology*, 192(1):1–6, 2005.

[341] Brent A Reynolds and Samuel Weiss. Generation of neurons and astrocytes from isolated cells of the adult mammalian central nervous system. *Science*, 255(5052):1707–1710, 1992.

[342] Benjamin E Reubinoff, Pavel Itsykson, Tikva Turetsky, Martin F Pera, Etti Reinhartz, Anna Itzik, and Tamir Ben-Hur. Neural progenitors from human embryonic stem cells. *Nature Biotechnology*, 19(12):1134–1140, 2001.

[343] Mototsugu Eiraku, Kiichi Watanabe, Mami Matsuo-Takasaki, Masako Kawada, Shigenobu Yonemura, Michiru Matsumura, Takafumi Wataya, Ayaka Nishiyama, Keiko Muguruma, and Yoshiki Sasai. Self-organized formation of polarized cortical tissues from ESCs and its active manipulation by extrinsic signals. *Cell stem cell*, 3(5):519–532, 2008.

[344] Teruko Danjo, Mototsugu Eiraku, Keiko Muguruma, Kiichi Watanabe, Masako Kawada, Yuchio Yanagawa, John LR Rubenstein, and Yoshiki Sasai. Subregional specification of embryonic stem cell-derived ventral telencephalic tissues by timed and combinatory treatment with extrinsic signals. *Journal of Neuroscience*, 31(5):1919–1933, 2011.

[345] Jessica Mariani, Maria Vittoria Simonini, Dean Palejev, Livia Tomasini, Gianfilippo Coppola, Anna M Szekely, Tamas L Horvath, and Flora M Vaccarino. Modeling human cortical development in vitro using induced pluripotent stem cells. *Proceedings of the National Academy of Sciences*, 109(31):12770–12775, 2012.

[346] Toshiro Sato, Robert G Vries, Hugo J Snippert, Marc Van De Wetering, Nick Barker, Daniel E Stange, Johan H Van Es, Arie Abo, Pekka Kujala, Peter J Peters, et al. Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature*, 459(7244):262–265, 2009.

[347] Hidetaka Suga, Taisuke Kadoshima, Maki Minaguchi, Masatoshi Ohgushi, Mika Soen, Tokushige Nakano, Nozomu Takata, Takafumi Wataya, Keiko Muguruma, Hiroyuki Miyoshi, et al. Self-formation of functional adenohypophysis in three-dimensional culture. *Nature*, 480(7375):57–62, 2011.

[348] Mototsugu Eiraku, Nozomu Takata, Hiroki Ishibashi, Masako Kawada, Eriko Sakakura, Satoru Okuda, Kiyotoshi Sekiguchi, Taiji Adachi, and Yoshiki Sasai. Self-organizing optic-cup morphogenesis in three-dimensional culture. *Nature*, 472(7341):51–56, 2011.

[349] Francesco Antonica, Dominika Figini Kasprzyk, Robert Opitz, Michelina Iacovino, Xiao-Hui Liao, Alexandra Mihaela Dumitrescu, Samuel Refetoff, Kathelijne Peremans, Mario Manto, Michael Kyba, et al. Generation of functional thyroid from embryonic stem cells. *Nature*, 491(7422):66–71, 2012.

[350] Sergiu P Pașca, Paola Arlotta, Helen S Bateup, J Gray Camp, Silvia Cappello, Fred H Gage, Jürgen A Knoblich, Arnold R Kriegstein, Madeline A Lancaster, Guo-Li Ming, et al. A nomenclature consensus for nervous system organoids and assembloids. *Nature*, 609(7929):907–910, 2022.

[351] Taisuke Kadoshima, Hideya Sakaguchi, Tokushige Nakano, Mika Soen, Satoshi Ando, Mototsugu Eiraku, and Yoshiki Sasai. Self-organization of axial polarity, inside-out layer pattern, and species-specific progenitor dynamics in human ES cell–derived neocortex. *Proceedings of the National Academy of Sciences*, 110(50):20284–20289, 2013.

[352] Junghyun Jo, Yixin Xiao, Alfred Xuyang Sun, Engin Cukuroglu, Hoang-Dai Tran, Jonathan Göke, Zi Ying Tan, Tzuen Yih Saw, Cheng-Peow Tan, Hidayat Lokman, et al. Midbrain-like organoids from human pluripotent stem cells contain functional dopaminergic and neuromelanin-producing neurons. *Cell stem cell*, 19(2):248–257, 2016.

[353] Keiko Muguruma, Ayaka Nishiyama, Hideshi Kawakami, Kouichi Hashimoto, and Yoshiki Sasai. Self-organization of polarized cerebellar tissue in 3D culture of human pluripotent stem cells. *Cell reports*, 10(4):537–550, 2015.

[354] Yuki Miura, Min-Yin Li, Fikri Birey, Kazuya Ikeda, Omer Revah, Mayuri Vijay Thete, Jin-Young Park, Alyssa Puno, Samuel H Lee, Matthew H Porteus, et al. Generation of human striatal organoids and cortico-striatal assembloids from human pluripotent stem cells. *Nature Biotechnology*, 38(12):1421–1430, 2020.

[355] Joshua A Bagley, Daniel Reumann, Shan Bian, Julie Lévi-Strauss, and Juergen A Knoblich. Fused cerebral organoids model interactions between brain regions. *Nature methods*, 14(7):743–751, 2017.

[356] Yangfei Xiang, Yoshiaki Tanaka, Benjamin Patterson, Young-Jin Kang, Gubbi Govindaiah, Naomi Roselaar, Bilal Cakir, Kun-Yong Kim, Adam P Lombroso, Sung-Min Hwang, et al. Fusion of regionally specified hpsc-derived organoids models human brain development and interneuron migration. *Cell stem cell*, 21(3):383–398, 2017.

[357] Fikri Birey, Jimena Andersen, Christopher D Makinson, Saiful Islam, Wu Wei, Nina Huber, H Christina Fan, Kimberly R Cordes Metzler, Georgia Panagiotakos, Nicholas Thom, et al. Assembly of functionally integrated human forebrain spheroids. *Nature*, 545(7652):54–59, 2017.

[358] Xuyu Qian, Ha Nam Nguyen, Mingxi M Song, Christopher Hadiono, Sarah C Ogden, Christy Hammack, Bing Yao, Gregory R Hamersky, Fadi Jacob, Chun Zhong, et al. Brain-region-specific organoids using mini-bioreactors for modeling ZIKV exposure. *Cell*, 165(5):1238–1254, 2016.

[359] Steven A Sloan, Spyros Darmanis, Nina Huber, Themasap A Khan, Fikri Birey, Christine Caneda, Richard Reimer, Stephen R Quake, Ben A Barres, and Sergiu P Paşca. Human astrocyte maturation captured in 3D cerebral cortical spheroids derived from pluripotent stem cells. *Neuron*, 95 (4):779–790, 2017.

[360] Sunghee Estelle Park, Andrei Georgescu, and Dongeun Huh. Organoids-on-a-chip. *Science*, 364 (6444):960–965, 2019.

[361] Kaja I Ritzau-Reid, Sebastien JP Callens, Ruoxiao Xie, Martina Cihova, Daniel Reumann, Christopher L Grigsby, Lino Prados-Martin, Richard Wang, Axel C Moore, James PK Armstrong, et al. Microfibrous scaffolds guide stem cell lumenogenesis and brain organoid engineering. *Advanced Materials*, 35(41):2300305, 2023.

[362] Pelin Saglam-Metiner, Utku Devamoglu, Yagmur Filiz, Soheil Akbari, Goze Beceren, Bakiye Goker, Burcu Yaldiz, Sena Yanasik, Cigir Biray Avci, Esra Erdal, et al. Spatio-temporal dynamics enhance cellular diversity, neuronal function and further maturation of human cerebral organoids.

Communications Biology, 6(1):173, 2023.

[363] Minoru Takasato, Pei X Er, Han S Chiu, Barbara Maier, Gregory J Baillie, Charles Ferguson, Robert G Parton, Ernst J Wolvetang, Matthias S Roost, Susana M Chuva de Sousa Lopes, et al. Kidney organoids from human iPS cells contain multiple lineages and model human nephrogenesis. *Nature*, 526(7574):564–568, 2015.

[364] Kyle W McCracken, Emily M Catá, Calyn M Crawford, Katie L Sinagoga, Michael Schumacher, Briana E Rockich, Yu-Hwai Tsai, Christopher N Mayhew, Jason R Spence, Yana Zavros, et al. Modelling human development and disease in pluripotent stem-cell-derived gastric organoids. *Nature*, 516(7531):400–404, 2014.

[365] Briana R Dye, David R Hill, Michael AH Ferguson, Yu-Hwai Tsai, Melinda S Nagy, Rachel Dyal, James M Wells, Christopher N Mayhew, Roy Nattiv, Ophir D Klein, et al. In vitro generation of human pluripotent stem cell derived lung organoids. *elife*, 4:e05098, 2015.

[366] Pablo Hofbauer, Stefan M Jahnel, Nora Papai, Magdalena Giesshammer, Alison Deyett, Clara Schmidt, Mirjam Penc, Katherina Tavernini, Nastasja Grdseloff, Christy Meledeth, et al. Cardioids reveal self-organizing principles of human cardiogenesis. *Cell*, 184(12):3299–3317, 2021.

[367] Susanne C Van den Brink, Peter Baillie-Johnson, Tina Balayo, Anna-Katerina Hadjantonakis, Sonja Nowotschin, David A Turner, and Alfonso Martinez Arias. Symmetry breaking, germ layer specification and axial organisation in aggregates of mouse embryonic stem cells. *Development*, 141(22):4231–4242, 2014.

[368] Naomi Moris, Kerim Anlas, Susanne C van den Brink, Anna Alemany, Julia Schröder, Sabitri Ghimire, Tina Balayo, Alexander van Oudenaarden, and Alfonso Martinez Arias. An in vitro model of early anteroposterior organization during human development. *Nature*, 582(7812):410–415, 2020.

[369] Jesse V Veenvliet, Adriano Bolondi, Helene Kretzmer, Leah Haut, Manuela Scholze-Wittler, Dennis Schifferl, Frederic Koch, Léo Guignard, Abhishek Sampath Kumar, Milena Pustet, et al. Mouse embryonic stem cells self-organize into trunk-like structures with neural tube and somites. *Science*, 370(6522):eaba4937, 2020.

[370] Marina Simian and Mina J Bissell. Organoids: a historical perspective of thinking in three dimensions. *Journal of Cell Biology*, 216(1):31–40, 2017.

[371] Bonsang Koo, Baekgyu Choi, Hoewon Park, and Ki-Jun Yoon. Past, present, and future of brain organoid technology. *Molecules and cells*, 42(9):617–627, 2019.

[372] Claudia Corrò, Laura Novellasdemunt, and Vivian SW Li. A brief history of organoids. *American Journal of Physiology-Cell Physiology*, 2020.

[373] Madeline A Lancaster. Pluripotent stem cell-derived organoids brief history of curiosity-led discoveries. *BioEssays*, 46(12):2400105, 2024.

[374] Aliya Fatehullah, Si Hui Tan, and Nick Barker. Organoids as an in vitro model of human development and disease. *Nature Cell Biology*, 18(3):246–254, 2016.

[375] Chongyuan Luo, Madeline A Lancaster, Rosa Castanon, Joseph R Nery, Juergen A Knoblich, and Joseph R Ecker. Cerebral organoids recapitulate epigenomic signatures of the human fetal brain. *Cell reports*, 17(12):3369–3384, 2016.

[376] J Gray Camp, Farhath Badsha, Marta Florio, Sabina Kanton, Tobias Gerber, Michaela Wilsch-Bräuninger, Eric Lewitus, Alex Sykes, Wulf Hevers, Madeline Lancaster, et al. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proceedings of the National Academy of Sciences*, 112(51):15672–15677, 2015.

[377] Alex A Pollen, Aparna Bhaduri, Madeline G Andrews, Tomasz J Nowakowski, Olivia S Meyerson, Mohammed A Mostajo-Radji, Elizabeth Di Lullo, Beatriz Alvarado, Melanie Bedolli, Max L Dougherty, et al. Establishing cerebral organoids as models of human-specific brain evolution.

Cell, 176(4):743–756, 2019.

[378] Sabina Kanton, Michael James Boyle, Zhisong He, Malgorzata Santel, Anne Weigert, Fátima Sanchís-Calleja, Patricia Guijarro, Leila Sidow, Jonas Simon Fleck, Dingding Han, et al. Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature*, 574(7778): 418–422, 2019.

[379] Jessica Mariani, Gianfilippo Coppola, Ping Zhang, Alexej Abyzov, Lauren Provini, Livia Tomasini, Mariangela Amenduni, Anna Szekely, Dean Palejev, Michael Wilson, et al. FOXG1-dependent dysregulation of GABA/glutamate neuron differentiation in autism spectrum disorders. *Cell*, 162 (2):375–390, 2015.

[380] Hongwon Kim, Hyeok Ju Park, Hwan Choi, Yujung Chang, Hanseul Park, Jaein Shin, Junyeop Kim, Christopher J Lengner, Yong Kyu Lee, and Jongpil Kim. Modeling G2019S-LRRK2 sporadic Parkinson's disease in 3D midbrain organoids. *Stem cell reports*, 12(3):518–531, 2019.

[381] J Jeya Vandana, Cassandra Manrique, Lauretta A Lacko, and Shuibing Chen. Human pluripotent-stem-cell-derived organoids for drug discovery and evaluation. *Cell Stem Cell*, 30(5):571–591, 2023.

[382] Cleber A Trujillo, Jason W Adams, Priscilla D Negraes, Cassiano Carromeu, Leon Tejwani, Allan Acab, Ben Tsuda, Charles A Thomas, Neha Sodhi, Katherine M Fichter, et al. Pharmacological reversal of synaptic and network pathology in human MECP2-KO neurons and cortical organoids. *EMBO Molecular Medicine*, 13(1):e12523, 2021.

[383] Anand Ramani, Giovanni Pasquini, Niklas J Gerkau, Vaibhav Jadhav, Omkar Suhas Vinchure, Nazlican Altinisik, Hannes Windoffer, Sarah Muller, Ina Rothenaigner, Sean Lin, et al. Reliability of high-quantity human brain organoids for modeling microcephaly, glioma invasion and drug screening. *Nature Communications*, 15(1):10703, 2024.

[384] Se Hoon Choi, Young Hye Kim, Matthias Hebisch, Christopher Sliwinski, Seungkyu Lee, Carla D'Avanzo, Hechao Chen, Basavaraj Hooli, Caroline Asselin, Julien Muffat, et al. A three-dimensional human neural cell culture model of Alzheimer's disease. *Nature*, 515(7526):274–278, 2014.

[385] Wei Zhang, Si-Lu Yang, Mei Yang, Stephanie Herrlinger, Qiang Shao, John L Collar, Edgar Fierro, Yanhong Shi, Aimin Liu, Hui Lu, et al. Modeling microcephaly with cerebral organoids reveals a WDR62–CEP170–KIF2A pathway promoting cilium disassembly in neural progenitors. *Nature communications*, 10(1):2612, 2019.

[386] Xiangling Meng, David Yao, Kent Imaizumi, Xiaoyu Chen, Kevin W Kelley, Noah Reis, Mayuri Vijay Thete, Arpana Arjun McKinney, Shravanti Kulkarni, Georgia Panagiotakos, et al. Assembloid CRISPR screens reveal impact of disease genes in human neurodevelopment. *Nature*, 622(7982): 359–366, 2023.

[387] Nita A Farahany, Henry T Greely, Steven Hyman, Christof Koch, Christine Grady, Sergiu P Pașca, Nenad Sestan, Paola Arlotta, James L Bernat, Jonathan Ting, et al. The ethics of experimenting with human brain tissue. *Nature*, 556(7702):429–432, 2018.

[388] Henry T Greely. Human brain surrogates research: the onrushing ethical dilemma. *The American Journal of Bioethics*, 21(1):34–45, 2021.

[389] Dide De Jongh, Emma K Massey, and Eline M Bunnik. Organoids: a systematic review of ethical issues. *Stem Cell Research & Therapy*, 13(1):337, 2022.

[390] Anahita Amiri, Gianfilippo Coppola, Soraya Scuderi, Feinan Wu, Tanmoy Roychowdhury, Fuchen Liu, Sirisha Pochareddy, Yurae Shin, Alexias Safi, Lingyun Song, et al. Transcriptome and epigenome landscape of human cortical development modeled in organoids. *Science*, 362(6420): eaat6720, 2018.

[391] Zhijian Zhang, Richard O'Laughlin, Hongjun Song, and Guo-Li Ming. Patterning of brain organoids derived from human pluripotent stem cells. *Current opinion in neurobiology*, 74:102536,

2022.

[392] Ilaria Chiaradia, Ivan Imaz-Rosshandler, Benedikt S Nilges, Jerome Boulanger, Laura Pellegrini, Richa Das, Nachiket D Kashikar, and Madeline A Lancaster. Tissue morphology influences the temporal program of human brain organoid development. *Cell Stem Cell*, 30(10):1351–1367, 2023.

[393] Ugomma C Eze, Aparna Bhaduri, Maximilian Haeussler, Tomasz J Nowakowski, and Arnold R Kriegstein. Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nature neuroscience*, 24(4):584–594, 2021.

[394] Zhisong He, Ashley Maynard, Akanksha Jain, Tobias Gerber, Rebecca Petri, Hsiu-Chuan Lin, Malgorzata Santel, Kevin Ly, Jean-Samuel Dupré, Leila Sidow, et al. Lineage recording in human cerebral organoids. *Nature methods*, 19(1):90–99, 2022.

[395] Jonas Simon Fleck, Sophie Martina Johanna Jansen, Damian Wollny, Fides Zenk, Makiko Seimiya, Akanksha Jain, Ryoko Okamoto, Malgorzata Santel, Zhisong He, J Gray Camp, et al. Inferring and perturbing cell fate regulomes in human brain organoids. *Nature*, 621(7978):365–372, 2023.

[396] Rebecca M Marton and Sergiu P Pașca. Organoid and assembloid technologies for investigating cellular crosstalk in human brain development and disease. *Trends in cell biology*, 30(2):133–143, 2020.

[397] Carla G Silva, Elise Peyre, and Laurent Nguyen. Cell migration promotes dynamic cellular interactions to control cerebral cortex morphogenesis. *Nature Reviews Neuroscience*, 20(6):318–329, 2019.

[398] Aparna Bhaduri, Madeline G Andrews, Walter Mancia Leon, Diane Jung, David Shin, Denise Allen, Dana Jung, Galina Schmunk, Maximilian Haeussler, Jahan Salma, et al. Cell stress in cortical organoids impairs molecular subtype specification. *Nature*, 578(7793):142–148, 2020.

[399] Manuel Schröter, Congwei Wang, Marco Terrigno, Philipp Hornauer, Ziqiang Huang, Ravi Jagasia, and Andreas Hierlemann. Functional imaging of brain organoids using high-density microelectrode arrays. *MRS bulletin*, 47(6):530–544, 2022.

[400] Hideya Sakaguchi, Yuki Ozaki, Tomoka Ashida, Takayoshi Matsubara, Naotaka Oishi, Shunsuke Kihara, and Jun Takahashi. Self-organized synchronous calcium transients in a cultured human neural network derived from cerebral organoids. *Stem Cell Reports*, 13(3):458–473, 2019.

[401] Ranmal A Samarasinghe, Osvaldo A Miranda, Jessie E Buth, Simon Mitchell, Isabella Ferando, Momoko Watanabe, Thomas F Allison, Arinnae Kurdian, Namie N Fotion, Michael J Gandal, et al. Identification of neural oscillations and epileptiform changes in human brain organoids. *Nature neuroscience*, 24(10):1488–1500, 2021.

[402] Daniel Reumann, Christian Krauditsch, Maria Novatchkova, Edoardo Sozzi, Sakurako Nagumo Wong, Michael Zabolocki, Marthe Priouret, Balint Doleschall, Kaja I Ritzau-Reid, Marielle Piber, et al. In vitro modeling of the human dopaminergic system using spatially arranged ventral midbrain–striatum–cortex assembloids. *Nature Methods*, 20(12):2034–2047, 2023.

[403] Hans Clevers. Modeling Development and Disease with Organoids. *Cell*, 165(7):1586–1597, 2016.

[404] Neal D. Amin and Sergiu P. Paşca. Building Models of Brain Disorders with Three-Dimensional Organoids. *Neuron*, 100(2):389–405, 2018.

[405] Hansen Wang. Modeling Neurological Diseases With Human Brain Organoids. *Frontiers in synaptic neuroscience*, 10:15–15, 2018.

[406] Jaydeep Sidhaye and Jürgen A Knoblich. Brain organoids: an ensemble of bioassays to investigate human neurodevelopment and disease. *Cell Death & Differentiation*, 28(1):52–67, 2021.

[407] Xiaoxiang Lu, Jiajie Yang, and Yangfei Xiang. Modeling human neurodevelopmental diseases with brain organoids. *Cell Regeneration*, 11(1):1, 2022.

[408] Chong Li, Jonas Simon Fleck, Catarina Martins-Costa, Thomas R Burkard, Jan Themann, Marlene Stuempflen, Angela Maria Peer, Ábel Vertesy, Jamie B Littleboy, Christopher Esk, et al. Single-cell

brain organoid screening identifies developmental defects in autism. *Nature*, 621(7978):373–380, 2023.

[409] Marina Bershteyn, Tomasz J Nowakowski, Alex A Pollen, Elizabeth Di Lullo, Aishwarya Nene, Anthony Wynshaw-Boris, and Arnold R Kriegstein. Human iPSC-derived cerebral organoids model cellular features of lissencephaly and reveal prolonged mitosis of outer radial glia. *Cell stem cell*, 20(4):435–449, 2017.

[410] Ana Rita Gomes, Tiago G Fernandes, Sandra H Vaz, Teresa P Silva, Evguenia P Bekman, Sara Xapelli, Sofia Duarte, Mehrnaz Ghazvini, Joost Gribnau, Alysson R Muotri, et al. Modeling Rett syndrome with human patient-specific forebrain organoids. *Frontiers in cell and developmental biology*, 8:610427, 2020.

[411] Vanesa Nieto-Estévez and Jenny Hsieh. Human brain organoid models of developmental epilepsies. *Epilepsy Currents*, 20(5):282–290, 2020.

[412] Joseph Park, Isaac Wetzel, Ian Marriott, Didier Dréau, Carla D'Avanzo, Doo Yeon Kim, Rudolph E Tanzi, and Hansang Cho. A 3D human triculture system modeling neurodegeneration and neuroinflammation in Alzheimer's disease. *Nature neuroscience*, 21(7):941–951, 2018.

[413] P Conforti, D Besusso, VD Bocchi, A Faedo, E Cesana, G Rossetti, V Ranzani, CN Svendsen, LM Thompson, M Toselli, et al. Faulty neuronal determination and cell polarization are reverted by modulating HD early phenotypes. *Proceedings of the National Academy of Sciences*, 115(4): E762–E771, 2018.

[414] Kornélia Szebényi, Léa MD Wenger, Yu Sun, Alexander WE Dunn, Colleen A Limegrover, George M Gibbons, Elena Conci, Ole Paulsen, Susanna B Mierau, Gabriel Balmus, et al. Human ALS/FTD brain organoid slice cultures display distinct early astrocyte and targetable neuronal pathology. *Nature neuroscience*, 24(11):1542–1554, 2021.

[415] Chun-Ting Lee, Raphael M Bendriem, Wells W Wu, and Rong-Fong Shen. 3D brain organoids derived from pluripotent stem cells: promising experimental models for brain development and neurodegenerative disorders. *Journal of biomedical science*, 24(1):59, 2017.

[416] Wenqiang Fan, Kimberly M Christian, Hongjun Song, and Guo-li Ming. Applications of brain organoids for infectious diseases. *Journal of molecular biology*, 434(3):167243, 2022.

[417] Jason Dang, Shashi Kant Tiwari, Gianluigi Lichinchi, Yue Qin, Veena S Patil, Alexey M Eroshkin, and Tariq M Rana. Zika virus depletes neural progenitors in human cerebral organoids through activation of the innate immune receptor TLR3. *Cell stem cell*, 19(2):258–265, 2016.

[418] Fernanda R Cugola, Isabella R Fernandes, Fabiele B Russo, Beatriz C Freitas, João LM Dias, Katia P Guimarães, Cecília Benazzato, Nathalia Almeida, Graciela C Pignatari, Sarah Romero, et al. The Brazilian Zika virus strain causes birth defects in experimental models. *Nature*, 534 (7606):267–271, 2016.

[419] Patricia P Garcez, Erick Correia Loiola, Rodrigo Madeiro da Costa, Luiza M Higa, Pablo Trindade, Rodrigo Delvecchio, Juliana Minardi Nascimento, Rodrigo Brindeiro, Amilcar Tanuri, and Stevens K Rehen. Zika virus impairs growth in human neurospheres and brain organoids. *Science*, 352(6287):816–818, 2016.

[420] Fadi Jacob, Sarshan R Pather, Wei-Kai Huang, Feng Zhang, Samuel Zheng Hao Wong, Haowen Zhou, Beatrice Cubitt, Wenqiang Fan, Catherine Z Chen, Miao Xu, et al. Human pluripotent stem cell-derived neural cells and brain organoids reveal SARS-CoV-2 neurotropism predominates in choroid plexus epithelium. *Cell stem cell*, 27(6):937–950, 2020.

[421] Laura Pellegrini, Anna Albecka, Donna L Mallery, Max J Kellner, David Paul, Andrew P Carter, Leo C James, and Madeline A Lancaster. SARS-CoV-2 infects the brain choroid plexus and disrupts the blood-CSF barrier in human brain organoids. *Cell stem cell*, 27(6):951–961, 2020.

[422] Guoqiang Sun, Flavia Chiuppesi, Xianwei Chen, Cheng Wang, E Tian, Jenny Nguyen, Mindy Kha,

Daniel Trinh, Hannah Zhang, Maria C Marchetto, et al. Modeling human cytomegalovirus-induced microcephaly in human iPSC-derived brain organoids. *Cell Reports Medicine*, 1(1), 2020.

[423] Veronica Krenn, Camilla Bosone, Thomas R Burkard, Julia Spanier, Ulrich Kalinke, Arianna Calistri, Cristiano Salata, Raissa Rilo Christoff, Patricia Pestana Garcez, Ali Mirazimi, et al. Organoid modeling of Zika and herpes simplex virus 1 infections reveals virus-specific responses leading to microcephaly. *Cell stem cell*, 28(8):1362–1379, 2021.

[424] Jie Wen, Fangkun Liu, Quan Cheng, Nathaniel Weygant, Xisong Liang, Fan Fan, Chuntao Li, Liyang Zhang, and Zhixiong Liu. Applications of organoid technology to brain tumors. *CNS Neuroscience & Therapeutics*, 29(10):2725–2743, 2023.

[425] Shan Bian, Marko Repic, Zhenming Guo, Anoop Kavirayani, Thomas Burkard, Joshua A Bagley, Christian Krauditsch, and Jürgen A Knoblich. Genetically engineered cerebral organoids model brain tumor formation. *Nature methods*, 15(8):631–639, 2018.

[426] Teresa G Krieger, Stephan M Tirier, Jeongbin Park, Katharina Jechow, Tanja Eisemann, Heike Peterziel, Peter Angel, Roland Eils, and Christian Conrad. Modeling glioblastoma invasion using human brain organoids and single-cell transcriptomics. *Neuro-oncology*, 22(8):1138–1149, 2020.

[427] Fadi Jacob, Ryan D Salinas, Daniel Y Zhang, Phuong TT Nguyen, Jordan G Schnoll, Samuel Zheng Hao Wong, Radhika Thokala, Saad Sheikh, Deeksha Saxena, Stefan Prokop, et al. A patient-derived glioblastoma organoid model and biobank recapitulates inter-and intra-tumoral heterogeneity. *Cell*, 180(1):188–204, 2020.

[428] Junko Ogawa, Gerald M Pao, Maxim N Shokhirev, and Inder M Verma. Glioblastoma model using human cerebral organoids. *Cell reports*, 23(4):1220–1229, 2018.

[429] Amanda Linkous, Demosthenes Balamatsias, Matija Snuderl, Lincoln Edwards, Ken Miyaguchi, Teresa Milner, Batsheva Reich, Leona Cohen-Gould, Andrew Storaska, Yasumi Nakayama, et al. Modeling patient-derived glioblastoma with cerebral organoids. *Cell reports*, 26(12):3203–3211, 2019.

[430] Toshikatsu Matsui and Tadahiro Shinozawa. Human organoids for predictive toxicology research and drug development. *Frontiers in Genetics*, 12:767621, 2021.

[431] Alexander P Haring, Harald Sontheimer, and Blake N Johnson. Microphysiological human brain and neural systems-on-a-chip: potential alternatives to small animal models and emerging platforms for drug discovery and personalized medicine. *Stem cell reviews and reports*, 13(3):381–406, 2017.

[432] Tanya Chhibber, Sounak Bagchi, Behnaz Lahooti, Angela Verma, Abraham Al-Ahmad, Manash K Paul, Gurudutt Pendyala, and Rahul Dev Jayant. CNS organoids: an innovative tool for neurological disease modeling and drug neurotoxicity screening. *Drug discovery today*, 25(2):456–465, 2020.

[433] Gianluca Costamagna, Giacomo Pietro Comi, and Stefania Corti. Advancing drug discovery for neurological disorders using iPSC-derived neural organoids. *International Journal of Molecular Sciences*, 22(5):2659, 2021.

[434] Zhi Wang, Shu-Na Wang, Tian-Ying Xu, Zhu-Wei Miao, Ding-Feng Su, and Chao-Yu Miao. Organoid technology for brain and therapeutics research. *CNS neuroscience & therapeutics*, 23 (10):771–778, 2017.

[435] Momoko Watanabe, Jessie E Buth, Neda Vishlaghi, Luis De La Torre-Ubieta, Jiannis Taxidis, Baljit S Khakh, Giovanni Coppola, Caroline A Pearson, Ken Yamauchi, Danyang Gong, et al. Self-organized cerebral organoids with human-specific features predict effective drugs to combat Zika virus infection. *Cell reports*, 21(2):517–532, 2017.

[436] Mei Tang, Caili Chen, Guoqing Wang, Yuelong Wang, Zongliang Zhang, Hexian Li, Qizhong Lu, Zeng Wang, Shasha Zhao, Chen Yang, et al. Evaluation of B7-H3 targeted immunotherapy in a

3D organoid model of craniopharyngioma. *Biomolecules*, 12(12):1744, 2022.

[437] Yvonne L. Latour, Robin Yoon, Sarah E. Thomas, Christina Grant, Cuiling Li, Miguel Sena-Esteves, Maria L. Allende, Richard L. Proia, and Cynthia J. Tifft. Human GLB1 knockout cerebral organoids model system for testing AAV9-mediated GLB1 gene therapy for reducing GM1 ganglioside storage in gm1 gangliosidosis. *Molecular Genetics and Metabolism Reports*, 21:100513, 2019. ISSN 2214-4269.

[438] Mikaela Grönholm, Michaela Feodoroff, Gabriella Antignani, Beatriz Martins, Firas Hamdan, and Vincenzo Cerullo. Patient-derived organoids for precision cancer immunotherapy. *Cancer research*, 81(12):3149–3155, 2021.

[439] Meghan Logun, Xin Wang, Yusha Sun, Stephen J Bagley, Nannan Li, Arati Desai, Daniel Y Zhang, MacLean P Nasrallah, Emily Ling-Lin Pai, Bike Su Oner, et al. Patient-derived glioblastoma organoids as real-time avatars for assessing responses to clinical CAR-T cell therapy. *Cell Stem Cell*, 32(2):181–190, 2025.

[440] Shree Bose, Hans Clevers, and Xiling Shen. Promises and challenges of organoid-guided precision medicine. *Med*, 2(9):1011–1026, 2021.

[441] Zilong Zhou, Lele Cong, and Xianling Cong. Patient-derived organoids in precision medicine: drug screening, organoid-on-a-chip and living organoid biobank. *Frontiers in oncology*, 11:762184, 2021.

[442] Agnieska Brazovskaja, Barbara Treutlein, and J Gray Camp. High-throughput single-cell transcriptomics on organoids. *Current opinion in biotechnology*, 55:167–171, 2019.

[443] Henrik Renner, Martha Grabos, Katharina J Becker, Theresa E Kagermeier, Jie Wu, Mandy Otto, Stefan Peischard, Dagmar Zeuschner, Yaroslav TsyTsyura, Paul Disse, et al. A fully automated high-throughput workflow for 3D-based chemical screening in human midbrain organoids. *elife*, 9: e52904, 2020.

[444] Franziska L Lampart, Dagmar Iber, and Nikolaos Doumpas. Organoids in high-throughput and high-content screenings. *Frontiers in Chemical Engineering*, 5:1120348, 2023.

[445] Anna S Monzel, Kathrin Hemmer, Tony Kaoma, Lisa M Smits, Silvia Bolognin, Philippe Lucarelli, Isabel Rosety, Alise Zagare, Paul Antony, Sarah L Nickels, et al. Machine learning-assisted neurotoxicity prediction in human midbrain organoids. *Parkinsonism & Related Disorders*, 75:105–109, 2020.

[446] George A Truskey. Human microphysiological systems and organoids as in vitro models for toxicological studies. *Frontiers in public health*, 6:185, 2018.

[447] Yi Cao. The uses of 3D human brain organoids for neurotoxicity evaluations: A review. *Neurotoxicology*, 91:84–93, 2022.

[448] Yan Huang, Yanping Dai, Min Li, Lulu Guo, Chulin Cao, Yuting Huang, Rui Ma, Shengyue Qiu, Xiaoyi Su, Kai Zhong, et al. Exposure to cadmium induces neuroinflammation and impairs ciliogenesis in hESC-derived 3D cerebral organoids. *Science of The Total Environment*, 797:149043, 2021.

[449] Alessandro Mariani, Davide Comolli, Roberto Fanelli, Gianluigi Forloni, and Massimiliano De Paola. Neonicotinoid pesticides affect developing neurons in experimental mouse models and in human induced pluripotent stem cell (iPSC)-derived neural cultures and organoids. *Cells*, 13 (15):1295, 2024.

[450] Thiago Arzua, Yasheng Yan, Congshan Jiang, Sarah Logan, Reilly L Allison, Clive Wells, Suresh N Kumar, Richard Schäfer, and Xiaowen Bai. Modeling alcohol-induced neurotoxicity using human induced pluripotent stem cell-derived three-dimensional cerebral organoids. *Translational psychiatry*, 10(1):347, 2020.

[451] Shiqun Chen, Yue Chen, Yifei Gao, Bin Han, Tao Wang, Huajiang Dong, and Liqun Chen. Toxic effects and mechanisms of nanoplastics on embryonic brain development using brain organoids

model. *Science of the Total Environment*, 904:166913, 2023.

[452] Luciana Simões Rafagnin Marinho, Gabrielly Maria Denadai Chiarantin, Juliane Midori Ikebara, Débora Sterzeck Cardoso, Théo Henrique de Lima-Vasconcellos, Guilherme Shigueto Vilar Higa, Mariana Sacrini Ayres Ferraz, Roberto De Pasquale, Silvia Honda Takada, Fabio Papes, et al. The impact of antidepressants on human neurodevelopment: Brain organoids as experimental tools. In *Seminars in Cell & Developmental Biology*, volume 144, pages 67–76. Elsevier, 2023.

[453] Ming-Fen Ho, Cheng Zhang, Irene Moon, Xiujuan Zhu, Brandon J Coombes, Joanna Biernacka, Michelle Skime, Tyler S Oesterle, Victor M Karpyak, Kristen Schmidt, et al. Single cell transcriptomics reveals distinct transcriptional responses to oxycodone and buprenorphine by iPSC-derived brain organoids from patients with opioid use disorder. *Molecular psychiatry*, 29(6):1636–1646, 2024.

[454] Braulio Gutiérrez-Medina. Making sense of blobs, whorls, and shades: methods for label-free, inverse imaging in bright-field optical microscopy. *Biophysical Reviews*, pages 1–11, 2025.

[455] Francesca Ciarpella, Raluca Georgiana Zamfir, Alessandra Campanelli, Giulia Pedrotti, Marzia Di Chio, Emanuela Bottani, and Ilaria Decimo. Generation of mouse hippocampal brain organoids from primary embryonic neural stem cells. *STAR protocols*, 4(3):102413, 2023.

[456] Héloïse Castiglione, Lucie Madrange, Camille Baquerre, Benoît Guy Christian Maisonneuve, Thomas Lemonnier, Jean-Philippe Deslys, Frank Yates, Thibault Honegger, Jessica Rontard, and Pierre-Antoine Vigneron. Towards a quality control framework for cerebral cortical organoids. *Scientific Reports*, 15(1):29431, 2025.

[457] Job O de Jong, Ceyda Llapashtica, Matthieu Genestine, Kevin Strauss, Frank Provenzano, Yan Sun, Huixiang Zhu, Giuseppe P Cortese, Francesco Brundu, Karlla W Brigatti, et al. Cortical overgrowth in a preclinical forebrain organoid model of CNTNAP2-associated autism spectrum disorder. *Nature Communications*, 12(1):4087, 2021.

[458] Maria Vinci, Sharon Gowan, Frances Boxall, Lisa Patterson, Miriam Zimmermann, William Court, Cara Lomas, Marta Mendiola, David Hardisson, and Suzanne A Eccles. Advances in establishment and analysis of three-dimensional tumor spheroid-based functional assays for target validation and drug evaluation. *BMC biology*, 10(1):29, 2012.

[459] Sarah Kessel, Scott Cribbes, Olivier Déry, Dmitry Kuksin, Eric Sincoff, Jean Qiu, and Leo Li-Ying Chan. High-throughput 3D tumor spheroid screening method for cancer drug discovery using celigo image cytometry. *SLAS TECHNOLOGY: Translating Life Sciences Innovation*, 22 (4):454–465, 2017.

[460] Michael A Borten, Sameer S Bajikar, Nobuo Sasaki, Hans Clevers, and Kevin A Janes. Automated brightfield morphometry of 3D organoid populations by OrganoSeg. *Scientific reports*, 8(1):5319, 2018.

[461] Nicola Gritti, Jia Le Lim, Kerim Anlaş, Mallica Pandya, Germaine Aalderink, Guillermo Martínez-Ara, and Vikas Trivedi. MOrgAna: accessible quantitative analysis of organoids with machine learning. *Development*, 148(18):dev199611, 2021.

[462] Reid T Powell, Micheline J Moussalli, Lei Guo, Goeun Bae, Pankaj Singh, Clifford Stephan, Imad Shureiqi, and Peter J Davies. deepOrganoid brightfield cell viability model for screening matrix-embedded organoids. *SLAS Discovery*, 27(3):175–184, 2022.

[463] Cameron J Wells, Najwa Labban, Shayna L Showalter, Roza K Przanowska, and Kevin Janes. Fast, flexible, learning-free organoid quantification and tracking with OrganoSeg2. *bioRxiv*, pages 2025–08, 2025.

[464] Julian Schröter, Luca Deininger, Blaz Lupse, Petra Richter, Steffen Syrbe, Ralf Mikut, and Sabine Jung-Klawitter. A large and diverse brain organoid dataset of 1,400 cross-laboratory images of 64 trackable brain organoids. *Scientific Data*, 11(1):514, 2024.

[465] Jeff W Lichtman and José-Angel Conchello. Fluorescence microscopy. *Nature methods*, 2(12): 910–919, 2005.

[466] Malte Renz. Fluorescence microscopy—a historical and technical perspective. *Cytometry Part A*, 83(9):767–779, 2013.

[467] Douglas B Murphy and Michael W Davidson. *Fundamentals of light microscopy and electronic imaging*. John Wiley & Sons, 2012.

[468] Ulrich Kubitscheck. *Fluorescence microscopy: from principles to biological applications*. John Wiley & Sons, 2017.

[469] Richard N Day and Michael W Davidson. The fluorescent protein palette: tools for cellular imaging. *Chemical Society Reviews*, 38(10):2887–2921, 2009.

[470] Gregor PC Drummen. Fluorescent probes and fluorescence (microscopy) techniques—illuminating biological and biomedical research. *Molecules*, 17(12):14067–14090, 2012.

[471] Mark Terasaki and Laurinda A Jaffe. Labeling of cell membranes and compartments for live cell fluorescence microscopy. *Methods in cell biology*, 74:469–489, 2004.

[472] Kyuseok Im, Sergey Mareninov, M Fernando Palma Diaz, and William H Yong. *An introduction to performing immunofluorescence staining*. Springer, 2019.

[473] Andreas Ettinger and Torsten Wittmann. Fluorescence live cell imaging. *Methods in cell biology*, 123:77–94, 2014.

[474] Clara Brémond Martin, Camille Simon Chane, Cédric Clouchoux, and Aymeric Histace. Recent trends and perspectives in cerebral organoids imaging and analysis. *Frontiers in Neuroscience*, 15: 629067, 2021.

[475] Rashmiparvathi Keshara, Yung Hae Kim, and Anne Grapin-Botton. Organoid imaging: seeing development and function. *Annual review of cell and developmental biology*, 38(1):447–466, 2022.

[476] Daniel Reumann, Martin Colombini, Paul Möseneder, Agnieszka Piszczek, and Jürgen A Knoblich. A cost-and time-efficient method for high-throughput cryoprocessing and tissue analysis using multiplexed tissue molds. *Cell Reports Methods*, 5(4), 2025.

[477] Johanna F Dekkers, Maria Alieva, Lianne M Wellens, Hendrikus CR Ariese, Paul R Jamieson, Annelotte M Vonk, Gimano D Amatngalim, Huili Hu, Koen C Oost, Hugo JG Snippert, et al. High-resolution 3D imaging of fixed and cleared organoids. *Nature protocols*, 14(6):1756–1771, 2019.

[478] Juan Eduardo Rodriguez-Gatica, Vira Iefremova, Liubov Sokhranyaeva, Si Wah Christina Au Yeung, Yannik Breitkreuz, Oliver Brüstle, Martin Karl Schwarz, and Ulrich Kubitscheck. Imaging three-dimensional brain organoid architecture from meso-to nanoscale across development. *Development*, 149(20):dev200439, 2022.

[479] Anne Beghin, Gianluca Grenci, Geetika Sahni, Su Guo, Harini Rajendiran, Tom Delaire, Saburnisha Binte Mohamad Raffi, Damien Blanc, Richard de Mets, Hui Ting Ong, et al. Automated high-speed 3D imaging of organoid cultures with multi-scale phenotypic quantification. *Nature Methods*, 19(7):881–892, 2022.

[480] Mariana Oksdath Mansilla, Camilo Salazar-Hernandez, Sally L Perrin, Kaitlin G Scheer, Gökhan Cildir, John Toubia, Kristyna Sedivakova, Melinda N Tea, Sakthi Lenin, Elise Ponthier, et al. 3D-printed microplate inserts for long term high-resolution imaging of live brain organoids. *BMC Biomedical Engineering*, 3(1):6, 2021.

[481] Akanksha Jain, Gilles Gut, Fátima Sanchis-Calleja, Reto Tschannen, Zhisong He, Nicolas Luginbühl, Fides Zenk, Antonius Chrisnandy, Simon Streib, Christoph Harmel, et al. Morphodynamics of human early brain organoid development. *Nature*, pages 1–10, 2025.

[482] Gustavo de Medeiros, Raphael Ortiz, Petr Strnad, Andrea Boni, Franziska Moos, Nicole Repina, Ludivine Challet Meylan, Francisca Maurer, and Prisca Liberali. Multiscale light-sheet organoid

imaging framework. *Nature Communications*, 13(1):4864, 2022.

[483] Sangjun Park, Cheol Hong Min, Eunjin Choi, Jeong-Sun Choi, Kyungjin Park, Seokyoung Han, Wonjun Choi, Hyun-Jong Jang, Kyung-Ok Cho, and Moonseok Kim. Long-term tracking of neural and oligodendroglial development in large-scale human cerebral organoids by noninvasive volumetric imaging. *Scientific Reports*, 15(1):2536, 2025.

[484] Ulrike Schnell, Freark Dijk, Klaas A Sjollema, and Ben NG Giepmans. Immunolabeling artifacts and the need for live-cell imaging. *Nature methods*, 9(2):152–158, 2012.

[485] Lixue Shi, Mian Wei, Yupeng Miao, Naixin Qian, Lingyan Shi, Ruth A Singer, Richard KP Benninger, and Wei Min. Highly-multiplexed volumetric mapping with Raman dye imaging and tissue clearing. *Nature Biotechnology*, 40(3):364–373, 2022.

[486] Hiromichi Tsurui, Hiroyuki Nishimura, Susumu Hattori, Sachiko Hirose, Ko Okumura, and Toshikazu Shirai. Seven-color fluorescence imaging of tissue samples based on Fourier spectroscopy and singular value decomposition. *Journal of Histochemistry & Cytochemistry*, 48(5):653–662, 2000.

[487] Francesco Cutrale, Vikas Trivedi, Le A Trinh, Chi-Li Chiu, John M Choi, Marcela S Artiga, and Scott E Fraser. Hyperspectral phasor analysis enables multiplexed 5D in vivo imaging. *Nature methods*, 14(2):149–152, 2017.

[488] Daniel Axelrod. Total internal reflection fluorescence microscopy in cell biology. *Traffic*, 2(11): 764–774, 2001.

[489] Ernst HK Stelzer, Frederic Strobl, Bo-Jui Chang, Friedrich Preusser, Stephan Preibisch, Katie McDole, and Reto Fiolka. Light sheet fluorescence microscopy. *Nature Reviews Methods Primers*, 1(1):73, 2021.

[490] Peter TC So, Chen Y Dong, Barry R Masters, and Keith M Berland. Two-photon excitation fluorescence microscopy. *Annual review of biomedical engineering*, 2(1):399–429, 2000.

[491] Bo Huang, Mark Bates, and Xiaowei Zhuang. Super-resolution fluorescence microscopy. *Annual review of biochemistry*, 78(1):993–1016, 2009.

[492] Andrew H Fischer, Kenneth A Jacobson, Jack Rose, and Rolf Zeller. Hematoxylin and eosin staining of tissue and cell sections. *Cold spring harbor protocols*, 2008(5):pdb–prot4986, 2008.

[493] Madeline A Lancaster, Nina S Corsini, Simone Wolfinger, E Hilary Gustafson, Alex W Phillips, Thomas R Burkard, Tomoki Otani, Frederick J Livesey, and Juergen A Knoblich. Guided self-organization and cortical plate formation in human brain organoids. *Nature Biotechnology*, 35(7): 659–666, 2017.

[494] Cesar Gonzalez, Enrique Armijo, Javiera Bravo-Alegria, Andrea Becerra-Calixto, Charles E Mays, and Claudio Soto. Modeling amyloid beta and tau pathology in human cerebral organoids. *Molecular psychiatry*, 23(12):2363–2374, 2018.

[495] Rebecca M Brown, Pranav SJB Rana, Hannah K Jaeger, John M O'Dowd, Onesmo B Balemba, and Elizabeth A Fortunato. Human cytomegalovirus compromises development of cerebral organoids. *Journal of virology*, 93(17):10–1128, 2019.

[496] Daojing Wang and Steven Bodovitz. Single cell analysis: the new frontier in 'omics'. *Trends in biotechnology*, 28(6):281–290, 2010.

[497] Zhisong He, Leander Dony, Jonas Simon Fleck, Artur Szałata, Katelyn X Li, Irena Slišković, Hsiu-Chuan Lin, Malgorzata Santel, Alexander Atamian, Giorgia Quadrato, et al. An integrated transcriptomic cell atlas of human neural organoids. *Nature*, 635(8039):690–698, 2024.

[498] Ivano Legnini, Lisa Emmenegger, Alessandra Zappulo, Agnieszka Rybak-Wolf, Ricardo Wurmus, Anna Oliveras Martinez, Cledi Cerda Jara, Anastasiya Boltengagen, Talé Hessler, Guido Mastrobuoni, et al. Spatiotemporal, optogenetic control of gene expression in organoids. *Nature Methods*, 20(10):1544–1552, 2023.

[499] Ray F Egerton et al. *Physical principles of electron microscopy*. Springer, 2016.

[500] Richard M Caprioli, Terry B Farmer, and Jocelyn Gile. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. *Analytical chemistry*, 69(23):4751–4760, 1997.

[501] Seulgi Noh, Yurim Park, Beomsue Kim, and Ji Young Mun. Structural analysis of cerebral organoids using confocal microscopy and transmission/scanning electron microscopy. *Microscopy and Microanalysis*, 31(1):ozae119, 2025.

[502] Gerarda Cappuccio, Saleh M Khalil, Sivan Osenberg, Feng Li, and Mirjana Maletic-Savatic. Mass spectrometry imaging as an emerging tool for studying metabolism in human brain organoids. *Frontiers in Molecular Biosciences*, 10:1181965, 2023.

[503] Chia-Chen Hsu, Jiabao Xu, Bas Brinkhof, Hui Wang, Zhanfeng Cui, Wei E Huang, and Hua Ye. A single-cell Raman-based platform to identify developmental stages of human pluripotent stem cell-derived neurons. *Proceedings of the National Academy of Sciences*, 117(31):18412–18423, 2020.

[504] Adrian Ghita, Flavius C Pascut, Melissa Mather, Virginie Sottile, and Ioan Notingher. Cytoplasmic RNA in undifferentiated neural stem cells: a potential label-free Raman spectral marker for assessing the undifferentiated status. *Analytical chemistry*, 84(7):3155–3162, 2012.

[505] Adrian Ghita, Flavius C Pascut, Virginie Sottile, Chris Denning, and Ioan Notingher. Applications of Raman micro-spectroscopy to stem cell technology: label-free molecular discrimination and monitoring cell differentiation. *EPJ Techniques and Instrumentation*, 2:1–14, 2015.

[506] Junnan Geng, Wei Zhang, Cheng Chen, Han Zhang, Anhong Zhou, and Yu Huang. Tracking the differentiation status of human neural stem cells through label-free Raman spectroscopy and machine learning-based analysis. *Analytical Chemistry*, 93(30):10453–10461, 2021.

[507] Arno Germond, Yulia Panina, Mikio Shiga, Hirohiko Niioka, and Tomonobu M Watanabe. Following embryonic stem cells, their differentiated progeny, and cell-state changes during iPS reprogramming by Raman spectroscopy. *Analytical Chemistry*, 92(22):14915–14923, 2020.

[508] Giulia Bruno, Michal Lipinski, Koseki J Kobayashi-Kirschvink, Christian Tentellino, Peter TC So, Jeon Woong Kang, and Francesco De Angelis. Label-free detection of biochemical changes during cortical organoid maturation via Raman spectroscopy and machine learning. *Analytical Chemistry*, 97(9):5029–5037, 2025.

[509] Dimitar Georgiev, Simon Vilms Pedersen, Ruoxiao Xie, Alvaro Fernández-Galiana, Molly M Stevens, and Mauricio Barahona. RamanSPy: An open-source Python package for integrative Raman spectroscopy data analysis. *Analytical Chemistry*, 96(21):8492–8500, 2024.

[510] Offer Rozenstein, Tarin Paz-Kagan, Christoph Salbach, and Arnon Karnieli. 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.

[511] Esra'a Alshdaifat, Doa'a Alshdaifat, Ayoub Alsarhan, Fairouz Hussein, and Subhieh Moh'd Faraj S El-Salhi. The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance. *Data*, 6(2):11, 2021.

[512] Feng Zhang, Xiaojun Tang, Angxin Tong, Bin Wang, Jingwei Wang, Yangyu Lv, Chunrui Tang, and Jie Wang. Baseline correction for infrared spectra using adaptive smoothness parameter penalized least squares method. *Spectroscopy Letters*, 53(3):222–233, 2020.

[513] Mads S Bergholt, Jean-Philippe St-Pierre, Giovanni S Offeddu, Paresh A Parmar, Michael B Albro, Jennifer L Puetzer, Michelle L Oyen, and Molly M Stevens. Raman spectroscopy reveals new insights into the zonal organization of native and tissue-engineered articular cartilage. *ACS central science*, 2(12):885–895, 2016.

[514] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion,

Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[515] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[516] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. TensorFlow: A system for large-scale machine learning. In *Osdi*, volume 16, pages 265–283. Savannah, GA, USA, 2016.

[517] Fred A Kruse, AB Lefkoff, JW Boardman, KB Heidebrecht, AT Shapiro, PJ Barloon, and AFH Goetz. The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. *Remote sensing of environment*, 44(2-3):145–163, 1993.

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

[519] Yoav Benjamini and Yosef Hochberg. 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.

[520] Shankar Rao Pandala. Lazy Predict help build a lot of basic models without much code and helps understand which models works better without any parameter tuning. https://github.com/shankarpandala/lazypredict.

[521] Gang Yin, Lintao Li, Shun Lu, Yu Yin, Yuanzhang Su, Yilan Zeng, Mei Luo, Maohua Ma, Hongyan Zhou, Lucia Orlandini, et al. An efficient primary screening of COVID-19 by serum Raman spectroscopy. *Journal of Raman Spectroscopy*, 52(5):949–958, 2021.

[522] Gang Yin, Lintao Li, Shun Lu, Yu Yin, Yuanzhang Su, Yilan Zeng, Mei Luo, Maohua Ma, Hongyan Zhou, Dezhong Yao, Gang Liu, and Jinyi Lang. Data and code on serum Raman spectroscopy as an efficient primary screening of coronavirus disease in 2019 (COVID-19). *Figshare*, 2020.

[523] Stefano Fornasaro, Fatima Alsamad, Monica Baia, Luis AE Batista de Carvalho, Claudia Beleites, Hugh J Byrne, Alessandro Chiadò, Mihaela Chis, Malama Chisanga, Amuthachelvi Daniel, 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.

[524] Ayşe ŞEN, Ibrahim Kecoglu, Muhammad Ahmed, Uğur Parlatan, and Mehmet Unlu. Differentiation of advanced generation mutant wheat lines: Conventional techniques versus Raman spectroscopy. *Frontiers in Plant Science*, 14, 2023.

[525] Barbara Lafuente, Robert T Downs, Hexiong Yang, and Nate Stone. The power of databases: The RRUFF project. In *Highlights in mineralogical crystallography*, pages 1–30. De Gruyter, 2015.

[526] Dimitar Georgiev, Álvaro Fernández-Galiana, Simon Vilms Pedersen, Georgios Papadopoulos, Ruoxiao Xie, Molly M Stevens, and Mauricio Barahona. Hyperspectral unmixing for Raman spectroscopy via physics-constrained autoencoders. *Proceedings of the National Academy of Sciences*, 121(45):e2407439121, 2024.

[527] Xiangrong Zhang, Yujia Sun, Jingyan Zhang, Peng Wu, and Licheng Jiao. Hyperspectral unmixing via deep convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 15(11): 1755–1759, 2018.

[528] Burkni Palsson, Magnus O Ulfarsson, and Johannes R Sveinsson. Convolutional autoencoder for spectral–spatial hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 59(1):535–549, 2020.

[529] Menna M Elkholy, Marwa Mostafa, Hala M Ebied, and Mohamed F Tolba. Hyperspectral unmixing

using deep convolutional autoencoder. *International Journal of Remote Sensing*, 41(12):4799–4819, 2020.

[530] Preetam Ghosh, Swalpa Kumar Roy, Bikram Koirala, Behnood Rasti, and Paul Scheunders. Hyperspectral unmixing using transformer network. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

[531] Savas Ozkan, Berk Kaya, and Gozde Bozdagi Akar. Endnet: Sparse autoencoder network for endmember extraction and hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):482–496, 2018.

[532] Ying Qu and Hairong Qi. uDAS: An untied denoising autoencoder with sparsity for spectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 57(3):1698–1712, 2018.

[533] Yuanchao Su, Andrea Marinoni, Jun Li, Javier Plaza, and Paolo Gamba. Stacked nonnegative sparse autoencoders for robust hyperspectral unmixing. *IEEE Geoscience and Remote Sensing Letters*, 15(9):1427–1431, 2018.

[534] Yuanchao Su, Andrea Marinoni, Jun Li, Antonio Plaza, and Paolo Gamba. Nonnegative sparse autoencoder for robust endmember extraction from remotely sensed hyperspectral images. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 205–208. IEEE, 2017.

[535] Ying Qu, Rui Guo, and Hairong Qi. Spectral unmixing through part-based non-negative constraint denoising autoencoder. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 209–212. IEEE, 2017.

[536] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[537] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[538] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[539] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, 2017.

[540] Xu Zhang, Maarten BJ Roeffaers, Srinjan Basu, Joseph R Daniele, Dan Fu, Christian W Freudiger, Gary R Holtom, and X Sunney Xie. Label-free live-cell imaging of nucleic acids using stimulated Raman scattering microscopy. *ChemPhysChem*, 13(4):1054–1059, 2012.

[541] Leonard Kritharides, Aimee Christian, Genevieve Stoudt, Dianne Morel, and George H Rothblat. Cholesterol metabolism and efflux in human THP-1 macrophages. *Arteriosclerosis, thrombosis, and vascular biology*, 18(10):1589–1599, 1998.

[542] Alan R Tall and Laurent Yvan-Charvet. Cholesterol, inflammation and innate immunity. *Nature Reviews Immunology*, 15(2):104–116, 2015.

[543] Amit K Saha, Shatha F Dallo, Ariana L Detmar, Pawel Osmulski, Maria Gaczynska, Tim Hui-Ming Huang, and Anand K Ramasubramanian. Cellular cholesterol regulates monocyte deformation. *Journal of biomechanics*, 52:83–88, 2017.

[544] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[545] Sergey Ioffe and Christian Szegedy. Batch normalizationccelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[546] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural

networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[547] Shehroz S Khan and Michael G Madden. New similarity metrics for Raman spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, 114:99–108, 2012.

[548] Ashok Zachariah Samuel, Ryo Mukojima, Shumpei Horii, Masahiro Ando, Soshi Egashira, Takuji Nakashima, Masato Iwatsuki, and Haruko Takeyama. On selecting a suitable spectral matching method for automated analytical applications of Raman spectroscopy. *ACS omega*, 6(3):2060–2065, 2021.

[549] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[550] Tarek Eissa, Liudmila Voronina, Marinus Huber, Frank Fleischmann, and Mihaela Žigman. The perils of molecular interpretations from vibrational spectra of complex samples. *Angewandte Chemie International Edition*, 63(50):e202411596, 2024.

[551] Mattia Miotto, Maria Rosito, Matteo Paoluzzi, Valeria de Turris, Viola Folli, Marco Leonetti, Giancarlo Ruocco, Alessandro Rosa, and Giorgio Gosti. Collective behavior and self-organization in neural rosette morphogenesis. *Frontiers in Cell and Developmental Biology*, 11:1134091, 2023.

[552] Sarah Fernandes, Davis Klein, and Maria C Marchetto. Unraveling human brain development and evolution using organoid models. *Frontiers in Cell and Developmental Biology*, 9:737429, 2021.

[553] Ana Uzquiano, Amanda J Kedaigle, Martina Pigoni, Bruna Paulsen, Xian Adiconis, Kwanho Kim, Tyler Faits, Surya Nagaraja, Noelia Antón-Bolaños, Chiara Gerhardinger, et al. Proper acquisition of cell class identity in organoids allows definition of fate specification programs of the human cerebral cortex. *Cell*, 185(20):3770–3788, 2022.

[554] Ábel Vértesy, Oliver L Eichmüller, Julia Naas, Maria Novatchkova, Christopher Esk, Meritxell Balmaña, Sabrina Ladstaetter, Christoph Bock, Arndt von Haeseler, and Juergen A Knoblich. Gruffi: an algorithm for computational removal of stressed cells from brain organoid transcriptomic datasets. *The EMBO Journal*, 41(17):e111118, 2022.

[555] Mergim Ramosaj, Sofia Madsen, Vanille Maillard, Valentina Scandella, Daniel Sudria-Lopez, Naoya Yuizumi, Ludovic Telley, and Marlen Knobloch. Lipid droplet availability affects neural stem/progenitor cell metabolism and proliferation. *Nature communications*, 12(1):7362, 2021.

[556] Gaia Cermenati, Nico Mitro, Matteo Audano, Roberto C Melcangi, Maurizio Crestani, Emma De Fabiani, and Donatella Caruso. Lipids in the nervous system: from biochemistry and molecular biology to patho-physiology. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1851(1):51–60, 2015.

[557] TJ Tracey, SE Kirk, FJ Steyn, and ST Ngo. The role of lipids in the central nervous system and their pathological implications in amyotrophic lateral sclerosis. In *Seminars in Cell & Developmental Biology*, volume 112, pages 69–81. Elsevier, 2021.

[558] Swarang Sachin Pundlik, Ashwin Venkateshvaran, Hitesh Mamgain, and Arvind Ramanathan. Analysis of Intracellular Fatty Acid Metabolism in Senescent Cells using Raman Microscopy. *bioRxiv*, pages 2025–09, 2025.

[559] Hao Peng, Yu Wang, Xusheng Tang, Lindong Shang, Fuyuan Chen, Peng Liang, Yuntong Wang, and Bei Li. Adjustable spot wide-field Raman spectroscopy combined with machine learning for accurate classification of breast cancer cells. *Talanta*, page 128862, 2025.

[560] Callum Highmore, Niall Hanrahan, Yoshiki Cook, Ysanne Pritchard, Adam Lister, Kirsty Cooper, George Devitt, Alasdair PS Munro, Saul N Faust, Sumeet Mahajan, et al. Identification and antimicrobial resistance profiling of Pseudomonas aeruginosa using multi-excitation Raman spectroscopy and computational analytics. *npj Antimicrobials and Resistance*, 3(1):74, 2025.

[561] Omar E Uribe-Juárez, Luis A Silva Valdéz, Flor Ivon Vivar Velázquez, Fidel Montoya-Molina, José A Moreno-Razo, María G Flores-Sánchez, Juan Morales-Corona, and Roberto Olayo-González. Analysis of Chemical Heterogeneity in Electrospun Fibers Through Hyperspectral Raman Imaging Using Open-Source Software. *Polymers*, 17(13):1883, 2025.

[562] Meral Yüce, Nazlı Öncer, Ceren Duru Çınar, Beyza Nur Günaydın, Zeynep İdil Akçora, and Hasan Kurt. Comprehensive Raman Fingerprinting and Machine Learning-Based Classification of 14 Pesticides Using a 785 nm Custom Raman Instrument. *Biosensors*, 15(3):168, 2025.

[563] Szymon Wójcik, Magdalena Król, and Paweł Stoch. Machine learning-assisted Raman spectroscopy for enhanced plastic identification. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, page 126973, 2025.

[564] Carlos VA Ribeiro, Elton L Dantas, and Reinhardt A Fuck. Effects of high-temperature annealing and low-temperature metamictization on Archean zircon: Constraints from U-Pb isotopes, trace elements, and Raman dating. *American Mineralogist*, 110(8):1291–1302, 2025.

[565] K Stephan, K Rammelkamp, M Baqué, S Schröder, A Pisello, K Gwinner, G Ortenzi, P Irmisch, F Sohl, and V Unnithan. Multi-spectral field study of planetary analog material in extreme environments—alteration products of volcanic deposits of Vulcano/Italy. *Earth and Space Science*, 12(5):e2024EA004036, 2025.

[566] Penny E Wieser, Charlotte L DeVitre, and Isabelle Susman. Best practices for the analyses of co_2 fluids by Raman Spectroscopy, 2025.

[567] Giorgia Di Stefano, Mickael Baqué, Jean-Pierre Paul de Vera, Micol Bellucci, Manuele Ettore Michel Gangi, and Daniela Billi. Biosignature preservation in FaRLiP cyanobacteria after prolonged desiccation and its relevance to space missions searching for life. *International Journal of Astrobiology*, 24:e14, 2025.

[568] Alessia Cassaro, Claudia Pacelli, Giuseppina Fanelli, Mickael Baqué, Alessandro Maturilli, Patrick Leo, Veronica Lelli, Jean-Pierre Paul de Vera, Silvano Onofri, and Annamaria Timperio. Biomarker Preservation in Antarctic Sandstones after Prolonged Space Exposure Outside the International Space Station During the ESA EXPOSE-E Lichens and Fungi Experiment. *Astrobiology*, 25(5): 331–345, 2025.

[569] Simone Innocente, Siddra Maryam, Stefan Andersson-Engels, Katarzyna Komolibus, Rekha Gautam, and Andrea Visentin. A comprehensive pipeline to integrate preprocessing and machine learning techniques for accurate classification in Raman spectroscopy. In *Data Science for Photonics and Biophotonics*, volume 13011, pages 51–57. SPIE, 2024.

[570] Nicola Rossberg, Rekha Gautam, Katarzyna Komolibus, Barry O'Sullivan, and Andrea Visentin. Explainable AI-Based Feature Selection Approaches for Raman Spectroscopy. *Diagnostics*, 15(16): 2063, 2025.

[571] Audrey Bell, Sunanda Sharma, Megan L Hochstrasser, Keith Cheveralls, David G Mets, Ilya Kolb, and Ben Braverman. AutoOpenRaman: Low-cost, automated Raman spectroscopy, 2025.

[572] Jana Hahn, Pooja Gune, Sascha Hein, Wolf Holtkamp, Marcel Holger Schulz, Walter Matheis, Volker Oeppling, and Christel Kamp. Quantifying Components in a Model Vaccine with Machine-Learning-Augmented Raman Spectroscopy, 2025. ChemRxiv preprint 10.26434/chemrxiv-2025-zxszr.

[573] Yue Wang, Anqi Liu, Lin Tan, Jinyu Sun, Hailiang Zhang, Ting Xie, Zhimin Zhang, and Hongmei Lu. A Pre-training Enhanced Deep Learning Framework for Robust Sparse Unmixing in Chemical Imaging. *Analytica Chimica Acta*, page 344524, 2025.

[574] Tongshu Wu, Fanqiang Kong, Dan Li, Yunsong Li, and Fei Ye. DEAE: Diffusion-Enhanced Autoencoder Network for Unsupervised Nonlinear Hyperspectral Unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[575] Marcelo Terán, José Javier Ruiz, Pablo Loza-Álvarez, David Masip, and David Merino. Open Raman spectral library for biomolecule identification. *Chemometrics and Intelligent Laboratory Systems*, page 105476, 2025.

[576] Dario Polli, Vikas Kumar, Carlo M Valensise, Marco Marangoni, and Giulio Cerullo. Broadband coherent Raman scattering microscopy. *Laser & Photonics Reviews*, 12(9):1800020, 2018.

[577] Federico Vernuccio, Renzo Vanna, Chiara Ceconello, Arianna Bresci, Francesco Manetti, Salvatore Sorrentino, Silvia Ghislanzoni, Flavia Lambertucci, Omar Motiño, Isabelle Martins, et al. Full-spectrum CARS microscopy of cells and tissues with ultrashort white-light continuum pulses. *The Journal of Physical Chemistry B*, 127(21):4733–4745, 2023.

[578] Ruihao Luo, Shuxia Guo, Julian Hniopek, and Thomas Bocklitz. 3D Hyperspectral Data Analysis with Spatially Aware Deep Learning for Diagnostic Applications. *Analytical Chemistry*, 97(14): 7729–7737, 2025.

[579] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022.

[580] John R. Ferraro, Kazuo Nakamoto, and Chris W. Brown. Chapter 1 - Basic Theory. In John R. Ferraro, Kazuo Nakamoto, and Chris W. Brown, editors, *Introductory Raman Spectroscopy*, pages 1–94. Academic Press, San Diego, second edition edition, 2003.

[581] Zhi-Min Zhang, Shan Chen, and Yi-Zeng Liang. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, 135(5):1138–1146, 2010.

[582] Degang Xu, Song Liu, Yaoyi Cai, and Chunhua Yang. Baseline correction method based on doubly reweighted penalized least squares. *Applied optics*, 58(14):3913–3920, 2019.

[583] Jianfeng Ye, Ziyang Tian, Haoyun Wei, and Yan Li. Baseline correction method based on improved asymmetrically reweighted penalized least squares for the Raman spectrum. *Applied Optics*, 59 (34):10933–10943, 2020.

[584] Juntao Liu, Jianyang Sun, Xiuzhen Huang, Guojun Li, and Binqiang Liu. Goldindec: a novel algorithm for Raman spectrum baseline correction. *Applied spectroscopy*, 69(7):834–842, 2015.

[585] Quanjie Han, Silong Peng, Qiong Xie, Yifan Wu, and Genwei Zhang. Iterative reweighted quantile regression using augmented lagrangian optimization for baseline correction. In *2018 5th International Conference on Information Science and Control Engineering (ICISCE)*, pages 280–284.

IEEE, 2018.

[586] Yuanjie Liu, Xiaoguang Zhou, and Yude Yu. A concise iterative method using the Bezier technique for baseline construction. *Analyst*, 140(23):7984–7996, 2015.

[587] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[588] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2.

[589] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

Appendix A

Classical derivation of Raman scattering

Classically, Raman scattering is explained through the interaction between the oscillating electric field of incident light and the polarisability of a molecule[580]. When a molecule is irradiated by an electromagnetic wave, the electric field $E(t)$ induces a time-dependent dipole moment $\mu(t)$ given by:

$$\mu(t) = \alpha(t)E(t), \tag{A.1}$$

where $E(t) = E_0 \cos(\omega_0 t)$ is the electric field of the incident light at frequency ω_0, and $\alpha(t)$ is the molecular polarisability varying with the vibrational motion of the molecule.

As the molecule vibrates, its polarisability $\alpha(t)$ varies with time. Assuming small vibrational displacements, this time dependence can be approximated by a first-order Taylor expansion:

$$\alpha(t) = \alpha_0 + \left(\frac{d\alpha}{dQ}\right) Q(t), \tag{A.2}$$

where α_0 is the equilibrium polarisability, and $Q(t) = Q_0 \cos(\omega_v t)$ represents the vibrational displacement oscillating at frequency ω_v.

Substituting into the expression for the dipole $\mu(t)$ yields:

$$\mu(t) = \underbrace{\alpha_0 E_0 \cos(\omega_0 t)}_{\text{Elastic (Rayleigh)}} + \underbrace{\left(\frac{d\alpha}{dQ}\right) Q_0 E_0 \cos(\omega_v t)\cos(\omega_0 t)}_{\text{Inelastic (Raman)}}. \tag{A.3}$$

The first term, $\alpha_0 E_0 \cos(\omega_0 t)$, describes a dipole oscillating at the same frequency as the incident light and corresponds to elastic Rayleigh scattering.

The second term involves the product of two cosine functions, which can be expanded using the identity:

$$\cos(\omega_v t)\cos(\omega_0 t) = \frac{1}{2}\left[\cos(\omega_0 + \omega_v)t + \cos(\omega_0 - \omega_v)t\right]. \tag{A.4}$$

These two frequency components represent inelastic Raman scattering: the $\omega_0 + \omega_v$ term corresponds to anti-Stokes scattering, while $\omega_0 - \omega_v$ corresponds to Stokes scattering.

Note that the second term only contributes if the derivative of the polarisability with respect to the vibrational coordinate is non-zero, i.e.:

$$\frac{d\alpha}{dQ} \neq 0 \quad \Longleftrightarrow \quad \text{Raman-active vibrational mode.} \tag{A.5}$$

This provides a condition for Raman activity: a vibrational mode is Raman-active and produces scattered light at shifted frequencies only if it induces a change in the molecular polarisability.

Appendix B

RamanSPy: Supplementary information

B.1 List of features in RamanSPy

Table B.1: Main features of RamanSPy (October 2025). Reproduced from Georgiev et al.[509], published by American Chemical Society, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

DATA LOADING AND MANAGEMENT

Common data format

- Single spectra
- Raman line scans
- Raman imaging data
- Volumetric Raman imaging data
- Collections of the above

Data loaders

- Built-in data loaders:
  - WITec Suite (WITec)
  - WiRE (Renishaw)
  - OceanView (Ocean Insight)
  - LabSpec (HORIBA)
- Interface for integration of new data loaders

Datasets

- Bacteria data[100]
- Volumetric THP-1 cell data[74]
- MDA-MB-231 cells data[209]
- COVID-19 data[521,522]
- Adenine data[523]
- Wheat lines data[524]
- RRUFF data[525]
- Synthetic data generator (Chapter 4)[526]

Continued on next page ...

Built-in preprocessing methods

- Miscellaneous
  - Spectral cropping
  - Subtracting a fixed background signal
- Cosmic rays removal
  - Whitaker-Hayes despiking[159]
- Denoising
  - Savitzky-Golay filtering[161]
  - Discrete Penalised Least Squares (Whittaker-Henderson smoothing)[162]
  - Kernel/window smoothers
  - Gaussian denoising
- Baseline correction
  - Asymmetric Least Squares (AsLS)[165]
  - Improved Asymmetric Least Squares (IAsLS)[166]
  - Adaptive Iteratively Reweighted Penalised Least Squares (airPLS)[581]
  - Asymmetrically Reweighted Penalised Least Squares (arPLS)[167]
  - Doubly Reweighted Penalised Least Squares (drPLS)[582]
  - Improved Asymmetrically Reweighted Penalised Least Squares (IarPLS)[583]
  - Adaptive Smoothness Parameter Penalised Least Squares (asPLS)[512]
  - Polynomial fitting
  - Modified polynomial fitting[163]
  - Improved modified polynomial fitting[164]
  - Goldindec method[584]
  - Iterative Reweighted Spline Quantile Regression (IRSQR)[585]
  - Corner Cutting[586]
  - Fully automatic baseline correction (FABC)[168]

Continued on next page ...

Table B.1 (continued)

SPECTRAL PREPROCESSING

Built-in preprocessing methods (continued)

- Normalisation/Scaling
  - Vector normalisation
  - Min-Max normalisation
  - Max intensity normalisation
  - Area-under-the-curve normalisation
- Interface for integration of custom algorithms

Preprocessing pipelines

- Interface for development of custom pipelines
- Built-in pipelines
  - 3 preprocessing protocols used in Figure 3.3
  - Protocol based on previous work by Bergholt et al.[513]

Continued on next page ...

DATA ANALYSIS

Built-in analysis methods

- Decomposition:
  - Principal component analysis (PCA)[172–174]
  - Non-negative matrix factorisation (NMF)[187]
  - Independent component analysis (ICA)[587]
- Clustering:
  - k-means[177,178]
- Spectral unmixing:
  - N-FINDR[234]
  - Vertex component analysis (VCA)[235]
  - Non-negative least squares (NNLS)[192]
  - Fully constrained least squares (FCLS)[236]

Metrics

- Mean absolute error (MAE)
- Mean squared error (MSE)
- Root mean square deviation (RMSE)
- Spectral angle divergence (SAD)[517]
- Spectral Information Divergence (SID)[518]

AI/ML integration

- Interface with common AI & ML frameworks

Continued on next page ...

Table B.1 (continued)

VISUALISATION

Spectra

- Collections of spectra
- Distributions of spectra
- Peak finding

Raman imaging data

- Image slices across specific spectral bands
- Spectra at specific locations

Volumetric Raman imaging data

- Volumetric slices across specific spectral bands
- Image slices of specific z-slices
- Spectra at specific locations

B.2 RamanSPy manual

*Reproduced from Georgiev et al.[509], published by American Chemical Society, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

Installation

RamanSPy has been deposited in the Python Package Index (https://pypi.org/project/ramanspy) and can thus be directly installed via the common package installer *pip* for Python:

```
pip install ramanspy
```

To access the functionalities of the package after installation, users only need to import RamanSPy in their Python scripts. One can import the whole package:

```
import ramanspy
# or import ramanspy as rp
```

or individual modules or methods:

```
# individual modules
from ramanspy import load, preprocessing

# individual methods
from ramanspy.analysis.unmix import NFINDR
```

Core infrastructure

Data management

Data in RamanSPy is represented by a set of custom data container classes based on scalable, computationally efficient array programming based on *numpy*[588], which correspond to different spectroscopic modalities. This includes the generic *SpectralContainer* class, as well as the more specialised *Spectrum*, *SpectralImage* and *SpectralVolume* classes representing single-point spectra (1D), imaging data (3D), and volumetric data (4D), respectively. These classes define data-specific information and behaviour in the background to allow a smooth, user-friendly experience, regardless of the data of interest.

The containers can be initialised by providing the corresponding intensity data, the spectral axis (in cm^{-1}) and other relevant (meta) data. For instance:

```
raman_spectrum = ramanspy.Spectrum(intensity_data, spectral_axis, *args, **kwargs)
raman_image = ramanspy.SpectralImage(intensity_data, spectral_axis, *args, **kwargs)
```

Once created, data containers can be manipulated, visualised, saved and loaded as needed using the built-in tools in RamanSPy.

Note that for the most part, users would not need to manually populate these containers. Instead, they can take advantage of the data loading functionalities that RamanSPy provides.

Data loading

To support data loading, RamanSPy offers easy-to-use data loaders compatible with experimental Raman spectroscopic data from a range of instrumental vendors in the area. These loaders - available within *ramanspy.load* - automatically parse relevant data files and return the appropriate spectral container. As an example, users can load MATLAB files exported from WITec's ProjectFOUR/FIVE software using the following command:

```
raman_object = ramanspy.load.witec(<PATH>)
```

A full list of the data loaders built into RamanSPy is available as part of the documentation of the package at https://ramanspy.readthedocs.io/en/latest/loading.html.

Raman data can also be loaded via established data-loading tools in Python. For instance, one can use *pandas*[589] to load a spectrum from a *.csv* file with two columns storing the intensity data and the spectral axis by using:

```
import pandas as pd

data = pd.read_csv(csv_filename)
raman_spectrum = ramanspy.Spectrum(data["<intensity_column>"], data["<axis_column>"])
```

Spectral preprocessing

Preprocessing logic in RamanSPy is defined by the *PreprocessingStep* class, which defines most of the necessary preprocessing infrastructure in the background to ensure a smooth, data-agnostic experience via a single point of contact specified through their *apply()* method.

Yet, as with data loading, for the most part, users are not expected to use this class to manually implement and optimise such preprocessing methods themselves. Instead, the RamanSPy package provides a comprehensive toolbox of ready-to-use preprocessing methods, which users can access, customise and employ to compile a wide variety of

preprocessing procedures. These preprocessing procedures are given as predefined classes within *ramanspy.preprocessing* which extend the *PreprocessingStep* class. To use these built-in methods, users need to create an instance of the selected technique. For instance:

```
denoiser = ramanspy.preprocessing.denoise.SavGol(*args, **kwargs)
baseline_corrector = ramanspy.preprocessing.baseline.ASLS(*args, **kwargs)
normaliser = ramanspy.preprocessing.normalise.MaxIntensity(*args, **kwargs)
```

Note that RamanSPy offers full control over relevant parameters, which can be supplied during initialisation via the **args* and ***kwargs* arguments.

As the methods inherit all operational logic defined within the parent *PreprocessingStep* class, they can be directly accessed and used on any data loaded in the framework through their *apply()* method:

```
preprocessesd_objects = denoiser.apply(<spectral object or collection of spectral
    objects>)
preprocessesd_objects = baseline_corrector.apply(<spectral object or collection of
    spectral objects>)
```

A full list of the methods for spectral preprocessing built into RamanSPy is available as part of the documentation of the package at https://ramanspy.readthedocs.io/en/latest/preprocessing.html.

If needed, users can also incorporate in-house methods into RamanSPy by wrapping the corresponding preprocessing logic into instances of the *PreprocessingStep* class. This can be done as follows:

```
def preprocessing_func(intensity_data, spectral_axis, *args, **kwargs):
    # Preprocess intensity_data and spectral_axis
    ...

    return updated_intensity_data, updated_spectral_axis

# wrapping the function together with the relevant *args and **kwargs
custom_preprocessing_method = ramanspy.preprocessing.PreprocessingStep(
    preprocessing_func, *args, **kwargs)
```

Then, the custom preprocessing method is fully compatible with the rest of RamanSPy's functionalities and out-of-the-box applicable to any data integrated within the package via its *apply()* method:

```
custom_preprocessing_method.apply(<spectral object or collection of spectral objects>)
```

Note that this class structure implies that these instances can then be saved (e.g. as *pickle* files) and, therefore, reused and shared as required afterwards.

Spectral analysis

As with preprocessing classes, users can access any built-in analysis method (available within the *ramanspy.analysis* sub-module) by creating an object instance of the corresponding class as follows:

```
nmf = ramanspy.analysis.decompose.NMF(*args, **kwargs)
kmeans = ramanspy.analysis.cluster.KMeans(*args, **kwargs)
unmixer = ramanspy.analysis.unmix.NFINDR(*args, **kwargs)
```

Once created, instances can be similarly accessed via their *apply()* method on any data loaded in RamanSPy.

```
cluster_maps, cluster_centres = kmeans.apply(<spectral object or collection of spectral
    objects>)
abundance_fractions, endmemebrs = unmixer.apply(<spectral object or collection of
    spectral objects>)
```

A full list of the methods for spectral analysis built into RamanSPy is available as part of the documentation of the package at https://ramanspy.readthedocs.io/en/latest/analysis.html.

Visualisation

The RamanSPy package also provides various visualisation tools available within the *ramanspy.plot* sub-module. As an example, one can plot spectra using the *spectra* function:

```
ramanspy.plot.spectra(<spectra or collection of spectra>)
ramanspy.plot.show()  # or plt.show() after import matplotlib.pyplot as plt
```

Note that these functions are highly customisable. This can be done by providing relevant parameters to control the plot generation, as well as through *matplotlib*'s customisation workflow.

```
import matplotlib.pyplot as plt

plt.figure(figsize = (5, 5))
ax = ramanspy.plot.spectra(<spectra or collection of spectra>, title="<str>", label="<
    str or list[str]>")
ax.set_ylabel("<str>")  # adding a label to the y-axis
plt.show()  # or ramanspy.plot.show()
```

A full list of the methods for data visualisation built into RamanSPy is available as part of the documentation of the package at https://ramanspy.readthedocs.io/en/latest/plot.html.

Preprocessing pipelines

Pipelining behaviour is defined by the *Pipeline* class in RamanSPy, which ensures that pipelines are accessible, simple-to-use and fully compatible with the rest of RamanSPy.

Creating a custom preprocessing pipeline

To assemble a preprocessing pipeline, one simply needs to stack relevant methods (built-in or custom) into the intended order of execution. For instance:

```
preprocessing_pipeline = ramanspy.preprocessing.Pipeline([
    ramanspy.preprocessing.denoise.SavGol(*args, **kwargs),
    ramanspy.preprocessing.baseline.ASLS(*args, **kwargs),
    ramanspy.preprocessing.normalise.MaxIntensity(*args, **kwargs),
    custom_preprocessing_method(*args, **kwargs)  # custom in-house method
])
```

Constructed pipelines can then be applied exactly as single methods via their *apply()* method to any data loaded within RamanSPy.

```
preprocessesd_objects = preprocessing_pipeline.apply(<spectral object or collection of
    spectral objects>)
```

As pipelines in RamanSPy are objects, they can also be directly saved in a convenient file format, such as *pickle* files. As such, they can then be reloaded, reused and shared as needed.

Access a predefined preprocessing pipeline

RamanSPy also provides a collection of built-in preprocessing pipelines. To access them, one can select the desired protocol from *ramanspy.preprocessing.protocols* as follows:

```
preprocessing_pipeline = ramanspy.preprocessing.protocols.PROTOCOL_X
```

A pre-defined *Pipeline* instance will be returned, which can similarly be employed directly through its *apply()* method.

A full list of the protocols for spectral preprocessing built into RamanSPy is available as part of the documentation of the package at https://ramanspy.readthedocs.io/en/latest/preprocessing.html#established-protocols.

AI integration

Integrate AI methods into RamanSPy

To integrate new techniques for spectral preprocessing and analysis, users can take advantage of the extensible architecture of RamanSPy and wrap models and algorithms into custom classes. For instance, one can create a new denoiser method based on a *PyTorch* model for denoising by simply creating a function, which defines how the model can be used to preprocess a generic intensity data array, and then wrapping the method within a *PreprocessingStep* instance.

```
def nn_preprocesing(intensity_data, wavenumber_axis):
    intensity_data = v.reshape(-1, intensity_data.shape[-1])
    output = model(torch.Tensor(intensity_data).unsqueeze(1)).cpu().detach().numpy()
    output = np.squeeze(output).reshape(intensity_data.shape)

    return output, wavenumber_axis

nn_denoiser = ramanspy.preprocessing.PreprocessingStep(nn_preprocesing)
```

Integrated methods are automatically rendered fully compatible with the rest of RamanSPy's functionalities in the background, so one can simply use the *apply()* method of the constructed denoiser to preprocess any data loaded within RamanSPy as any built-in preprocessing class.

Export data from RamanSPy to AI frameworks

The data management core of RamanSPy allows a direct interface with the entire Python ecosystem, including frameworks for statistical modelling, machine learning and deep learning. To do that, users can simply feed relevant data from RamanSPy to functions and tools they want to use elsewhere. For instance, one can pass the intensity data stored in a spectral container to a specific model from the *scikit-learn*[514] framework for statistical and ML modelling directly via their *fit()* method:

```
model.fit(spectral_container.spectral_data)
```

Datasets

To access the Raman spectroscopic datasets available in RamanSPy, users can employ custom data-loading methods built into RamanSPy under *ramanspy.datasets*. These would automatically parse the relevant data into the corresponding spectral container. For instance, one can load the bacteria data from Ho et al.[100] using the following function:

```
data_container, labels = ramanspy.datasets.bacteria(dataset="train", <PATH>)
```

Note that, depending on where each dataset was deposited and the license it was deposited under, some of these methods will automatically download the given dataset, whereas others may require the manual download of the data. Users are pointed to the documentation of each method for instructions on how to properly load each dataset.

A full list of the datasets built into RamanSPy is available as part of the documentation of the package at https://ramanspy.readthedocs.io/en/latest/datasets.html.

Metrics

Users can likewise readily access relevant spectroscopic metrics, such as MSE, SAD and SID, from *ramanspy.metrics*. These can be used to measure the similarity between spectral data by using the respective methods, e.g.:

```
ramanspy.metrics.SID(spectrum_I, spectrum_II)
```

A full list of the metrics built into RamanSPy is available as part of the documentation of the package at https://ramanspy.readthedocs.io/en/latest/metrics.html.

Appendix C

Unmixing Raman autoencoders: Supplementary information

Table C.1: Benchmark results on synthetic Raman mixture datasets. Each value represents the average result of 25 replicates, including 5 dataset and 5 model initialisations. Confidence intervals are calculated as one standard deviation around the sample mean. Values rounded to 3 decimal places. Best and second best results (mean value) are given in **underlined bold** and **bold**, respectively. Reproduced from Georgiev et al. [526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

		Chessboard		*Gaussian*		*Dirichlet*	
	Method	Endmembers (SAD) ↓	Abundances (MSE) ↓	Endmembers (SAD) ↓	Abundances (MSE) ↓	Endmembers (SAD) ↓	Abundances (MSE) ↓
ideal	PCA	0.874 ± 0.139	22.592 ± 8.969	1.051 ± 0.081	5.387 ± 1.876	0.847 ± 0.092	3.811 ± 1.291
	N-FINDR + FCLS	**0.000** ± **0.000**	**0.000** ± **0.000**	0.473 ± 0.042	0.033 ± 0.001	0.101 ± 0.040	**0.001** ± **0.000**
	VCA + FCLS	**0.000** ± **0.000**	**0.000** ± **0.000**	0.482 ± 0.047	0.038 ± 0.009	0.105 ± 0.031	**0.003** ± **0.003**
	Dense AE	0.015 ± 0.065	0.002 ± 0.008	0.255 ± 0.037	**0.010** ± **0.005**	**0.003** ± **0.001**	0.012 ± 0.007
	Convolutional AE	0.042 ± 0.111	0.002 ± 0.007	**0.168** ± **0.033**	**0.009** ± **0.006**	**0.003** ± **0.001**	0.012 ± 0.006
	Transformer AE	**0.002** ± **0.000**	**0.000** ± **0.000**	**0.212** ± **0.043**	**0.009** ± **0.005**	**0.007** ± **0.002**	0.012 ± 0.006
	Convolutional Transformer AE	**0.002** ± **0.000**	**0.000** ± **0.000**	0.215 ± 0.042	**0.009** ± **0.005**	0.017 ± 0.062	0.013 ± 0.007
+ artifacts	PCA	0.819 ± 0.072	79.38 ± 49.727	1.009 ± 0.082	61.488 ± 48.586	0.869 ± 0.069	60.154 ± 48.330
	N-FINDR + FCLS	0.629 ± 0.111	0.073 ± 0.036	0.828 ± 0.075	0.041 ± 0.015	0.599 ± 0.116	0.033 ± 0.015
	VCA + FCLS	0.353 ± 0.159	0.071 ± 0.042	0.609 ± 0.093	0.059 ± 0.015	0.392 ± 0.175	0.036 ± 0.019
	Dense AE	0.072 ± 0.081	0.042 ± 0.032	**0.347** ± **0.137**	0.029 ± 0.021	**0.068** ± **0.070**	0.021 ± 0.011
	Convolutional AE	**0.033** ± **0.003**	0.023 ± 0.013	**0.385** ± **0.166**	**0.027** ± **0.014**	**0.068** ± **0.072**	**0.017** ± **0.008**
	Transformer AE	**0.039** ± **0.031**	**0.022** ± **0.011**	0.393 ± 0.129	**0.026** ± **0.014**	**0.073** ± **0.105**	**0.017** ± **0.009**
	Convolutional Transformer AE	**0.033** ± **0.004**	**0.021** ± **0.010**	0.399 ± 0.164	**0.027** ± **0.015**	0.112 ± 0.140	**0.019** ± **0.010**
+ realistic	PCA	0.967 ± 0.074	88.132 ± 50.053	1.079 ± 0.045	63.735 ± 48.513	0.993 ± 0.089	61.244 ± 48.733
	N-FINDR + FCLS	0.361 ± 0.062	0.072 ± 0.040	0.478 ± 0.047	0.043 ± 0.013	0.299 ± 0.087	0.030 ± 0.019
	VCA + FCLS	0.173 ± 0.079	0.061 ± 0.038	0.400 ± 0.075	0.063 ± 0.016	0.229 ± 0.088	0.030 ± 0.017
	Dense AE	0.045 ± 0.024	0.030 ± 0.023	0.177 ± 0.050	0.015 ± 0.011	**0.073** ± **0.012**	**0.011** ± **0.006**
	Convolutional AE	0.054 ± 0.013	**0.019** ± **0.008**	0.168 ± 0.030	**0.011** ± **0.004**	**0.071** ± **0.022**	**0.008** ± **0.003**
	Transformer AE	**0.039** ± **0.005**	**0.018** ± **0.009**	**0.156** ± **0.044**	**0.011** ± **0.005**	0.085 ± 0.014	**0.008** ± **0.003**
	Convolutional Transformer AE	**0.040** ± **0.005**	**0.018** ± **0.008**	**0.151** ± **0.033**	**0.010** ± **0.004**	0.081 ± 0.012	**0.008** ± **0.003**
+ bilinear	PCA	—"—	—"—	1.117 ± 0.071	67.189 ± 48.404	1.038 ± 0.112	63.915 ± 48.423
	N-FINDR + FCLS	—"—	—"—	0.456 ± 0.049	0.039 ± 0.016	0.287 ± 0.074	0.025 ± 0.015
	VCA + FCLS	—"—	—"—	0.391 ± 0.058	0.057 ± 0.018	0.277 ± 0.082	0.030 ± 0.015
	Dense AE (bilinear)	—"—	—"—	0.247 ± 0.070	0.017 ± 0.011	**0.094** ± **0.011**	**0.010** ± **0.005**
	Convolutional AE (bilinear)*	—"—	—"—	***0.194*** ± **0.030**	**0.010** ± **0.004**	**0.087** ± **0.010**	**0.008** ± **0.003**
	Transformer AE (bilinear)	—"—	—"—	0.222 ± 0.084	0.012 ± 0.008	0.105 ± 0.012	**0.008** ± **0.003**
	Convolutional Transformer AE (bilinear)	—"—	—"—	**0.208** ± **0.051**	**0.011** ± **0.006**	0.100 ± 0.010	**0.008** ± **0.003**

Table C.2: Benchmark results on Raman spectroscopy data from sugar solutions. Each value represents the average result of 5 replicates. Confidence intervals are calculated as one standard deviation around the sample mean. Values rounded to 3 decimal places. Best and second best results (mean value) are given in **underlined bold** and **bold**, respectively. Reproduced from Georgiev et al.[526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

		High SNR				*Low SNR*			
		with endmembers		*without endmembers*		*with endmembers*		*without endmembers*	
	Method	Endmembers (SAD) ↓	Abundances (MSE) ↓	Endmembers (SAD) ↓	Abundances (MSE) ↓	Endmembers (SAD) ↓	Abundances (MSE) ↓	Endmembers (SAD) ↓	Abundances (MSE) ↓
	PCA	1.018 ± 0.000	0.073 ± 0.000	1.026 ± 0.000	0.072 ± 0.000	1.173 ± 0.000	0.075 ± 0.000	1.210 ± 0.000	0.073 ± 0.000
	N-FINDR+FCLS	**0.202 ± 0.000**	0.005 ± 0.000	0.515 ± 0.035	0.037 ± 0.029	0.900 ± 0.000	**0.026 ± 0.000**	1.077 ± 0.000	0.154 ± 0.000
	VCA+FCLS	0.322 ± 0.044	**0.004 ± 0.000**	0.525 ± 0.026	0.042 ± 0.029	0.677 ± 0.022	0.077 ± 0.035	0.901 ± 0.017	0.108 ± 0.006
SAD	Dense AE	0.212 ± 0.004	0.057 ± 0.001	0.219 ± 0.001	0.058 ± 0.001	**0.462 ± 0.038**	**0.058 ± 0.005**	**0.503 ± 0.046**	**0.054 ± 0.007**
SAD	Convolutional AE	**0.203 ± 0.001**	0.058 ± 0.001	**0.215 ± 0.002**	0.058 ± 0.001	0.503 ± 0.055	**0.058 ± 0.002**	0.529 ± 0.041	**0.057 ± 0.003**
SAD	Transformer AE	0.206 ± 0.001	0.058 ± 0.001	**0.218 ± 0.002**	0.028 ± 0.002	**0.496 ± 0.050**	0.061 ± 0.002	0.545 ± 0.037	0.069 ± 0.013
SAD	Conv. Trans. AE	0.208 ± 0.002	0.057 ± 0.001	**0.218 ± 0.001**	0.057 ± 0.001	0.521 ± 0.059	0.068 ± 0.013	0.533 ± 0.035	0.062 ± 0.011
MSE+SAD	Dense AE	0.242 ± 0.008	**0.003 ± 0.001**	0.265 ± 0.040	**0.003 ± 0.002**	0.498 ± 0.038	0.070 ± 0.006	**0.509 ± 0.045**	0.069 ± 0.008
MSE+SAD	Convolutional AE	0.264 ± 0.068	**0.004 ± 0.004**	0.240 ± 0.007	**0.002 ± 0.000**	0.568 ± 0.057	0.083 ± 0.017	0.593 ± 0.052	0.080 ± 0.013
MSE+SAD	Transformer AE	0.234 ± 0.006	**0.004 ± 0.001**	0.240 ± 0.009	0.004 ± 0.001	0.606 ± 0.034	0.101 ± 0.026	0.632 ± 0.014	0.097 ± 0.028
MSE+SAD	Conv. Trans. AE	0.321 ± 0.121	0.008 ± 0.005	0.333 ± 0.123	0.007 ± 0.006	0.570 ± 0.068	0.113 ± 0.012	0.665 ± 0.179	0.164 ± 0.037

Table C.3: Benchmark results on Raman spectroscopy data from sugar solutions – endmember similarity based on the Pearson correlation coefficient. Each value represents the average result of 5 replicates. Confidence intervals are calculated as one standard deviation around the sample mean. Values rounded to 3 decimal places. Best and second best results (mean value) are given in **underlined bold** and **bold**, respectively. The PCC between two endmember signatures x and y is defined as $\mathrm{PCC}(x,y) = 1 - \frac{\sum (x_i - \bar{x}) \sum (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$. Reproduced from Georgiev et al.[526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

		High SNR					*Low SNR*			
		with endmembers		*without endmembers*			*with endmembers*		*without endmembers*	
	Method	Endmembers (SAD) ↓	Abundances (MSE) ↓	Endmembers (SAD) ↓	Abundances (MSE) ↓		Endmembers (SAD) ↓	Abundances (MSE) ↓	Endmembers (SAD) ↓	Abundances (MSE) ↓
	PCA	0.547 ± 0.000	0.073 ± 0.000	0.541 ± 0.000	0.072 ± 0.000		0.641 ± 0.000	0.075 ± 0.000	0.691 ± 0.000	0.073 ± 0.000
	N-FINDR+FCLS	0.039 ± 0.000	0.005 ± 0.000	0.203 ± 0.055	0.037 ± 0.029	0.520 ± 0.000		**<u>0.026</u> ± <u>0.000</u>**	0.641 ± 0.000	0.154 ± 0.000
	VCA+FCLS	0.099 ± 0.047	**0.004 ± 0.000**	0.221 ± 0.056	0.042 ± 0.029		0.325 ± 0.019	0.077 ± 0.035	0.482 ± 0.005	0.108 ± 0.006
SAD	Dense AE	0.029 ± 0.001	0.057 ± 0.001	**0.031 ± 0.000**	0.058 ± 0.001		**<u>0.164</u> ± <u>0.042</u>**	**0.058 ± 0.005**	**0.198 ± 0.046**	**<u>0.054</u> ± <u>0.007</u>**
SAD	Convolutional AE	**<u>0.027</u> ± <u>0.000</u>**	0.058 ± 0.001	**<u>0.030</u> ± <u>0.002</u>**	0.058 ± 0.001		0.207 ± 0.052	**0.058 ± 0.002**	0.228 ± 0.030	**0.057 ± 0.003**
SAD	Transformer AE	**0.028 ± 0.000**	0.058 ± 0.001	**0.031 ± 0.000**	0.028 ± 0.002		0.200 ± 0.048	0.061 ± 0.002	0.243 ± 0.032	0.069 ± 0.013
SAD	Conv. Trans. AE	**0.028 ± 0.000**	0.057 ± 0.001	**0.031 ± 0.000**	0.057 ± 0.001		0.226 ± 0.059	0.068 ± 0.013	0.236 ± 0.034	0.062 ± 0.011
MSE+SAD	Dense AE	0.037 ± 0.002	**<u>0.003</u> ± <u>0.001</u>**	0.047 ± 0.018	**0.003 ± 0.002**		**0.174 ± 0.032**	0.070 ± 0.006	**<u>0.185</u> ± <u>0.042</u>**	0.069 ± 0.008
MSE+SAD	Convolutional AE	0.056 ± 0.044	**0.004 ± 0.004**	0.037 ± 0.002	**<u>0.002</u> ± <u>0.000</u>**		0.243 ± 0.044	0.083 ± 0.017	0.272 ± 0.040	0.080 ± 0.013
MSE+SAD	Transformer AE	0.035 ± 0.002	**0.004 ± 0.001**	0.037 ± 0.002	0.004 ± 0.001		0.296 ± 0.041	0.101 ± 0.026	0.311 ± 0.020	0.097 ± 0.028
MSE+SAD	Conv. Trans. AE	0.104 ± 0.096	0.008 ± 0.005	0.108 ± 0.099	0.007 ± 0.006		0.250 ± 0.072	0.113 ± 0.012	0.348 ± 0.176	0.164 ± 0.037

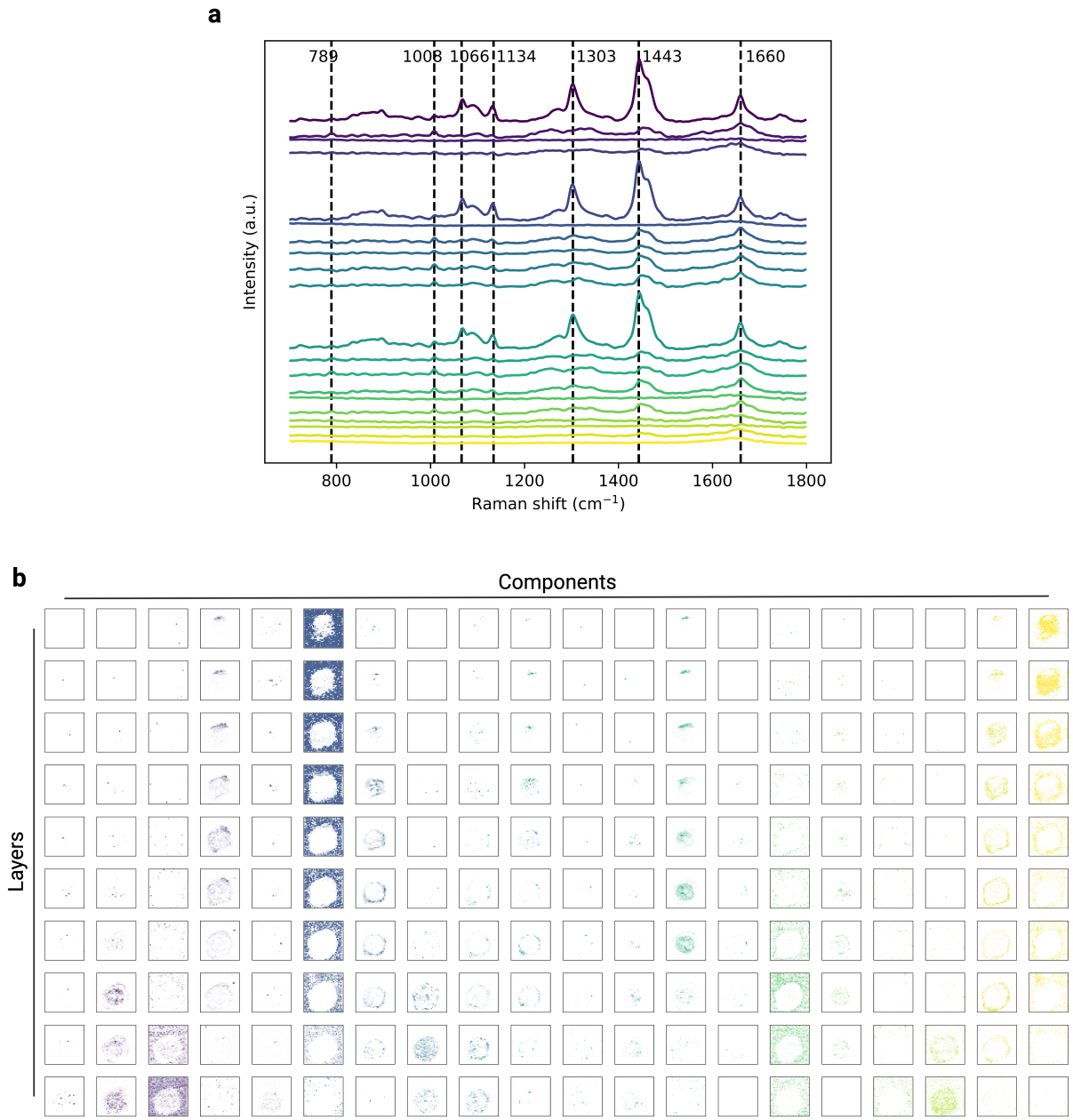


Figure C.1: Full unmixing results obtained with VCA + NNLS on THP-1 cell data. **a,** Derived endmembers. **b,** Derived fractional abundances. Assembled in BioRender.com. Reproduced from Georgiev et al.[526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

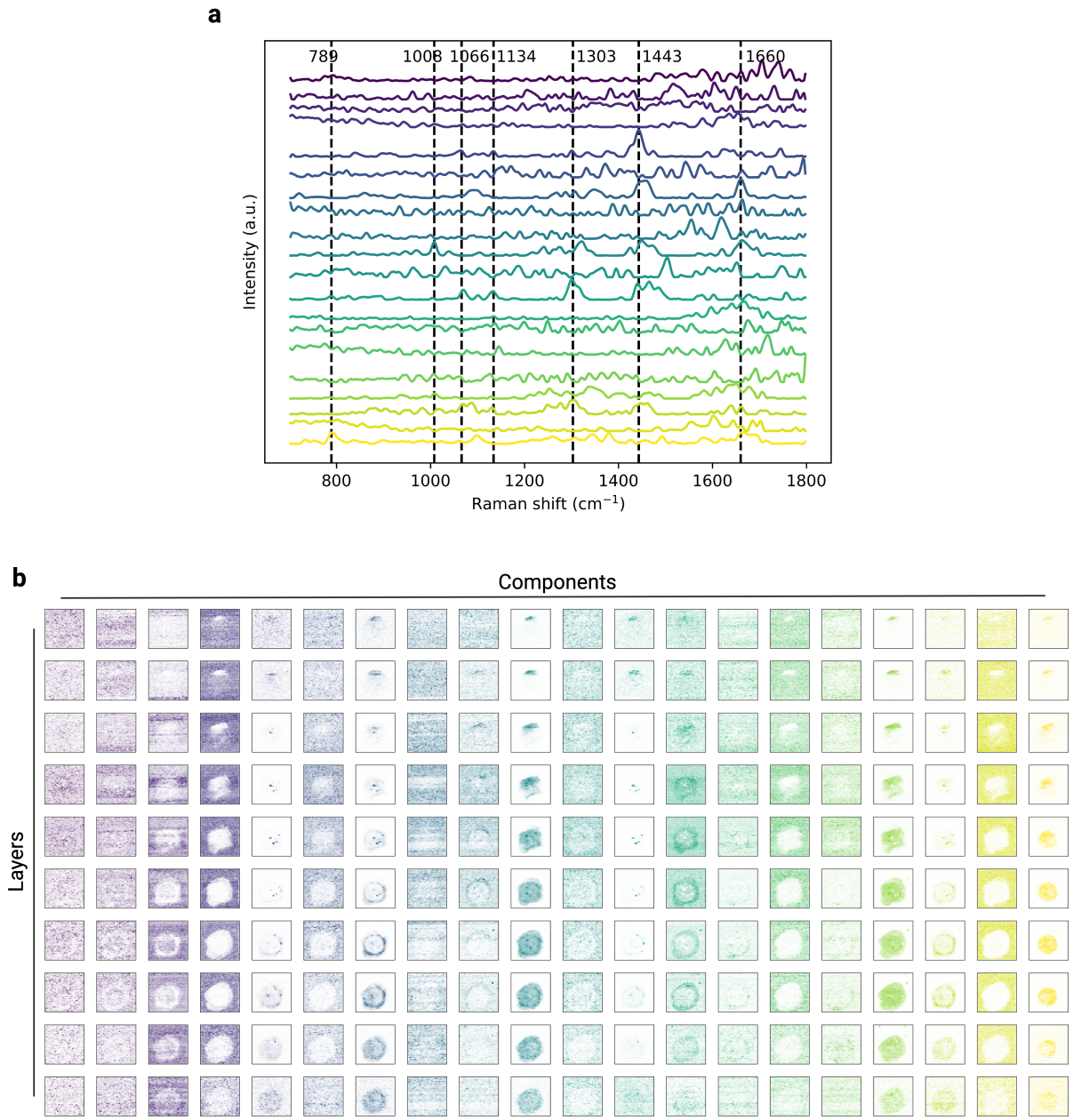


Figure C.2: Full unmixing results obtained with *Dense AE* model on THP-1 cell data. **a,** Derived endmembers. **b,** Derived fractional abundances. Assembled in BioRender.com. Reproduced from Georgiev et al.[526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

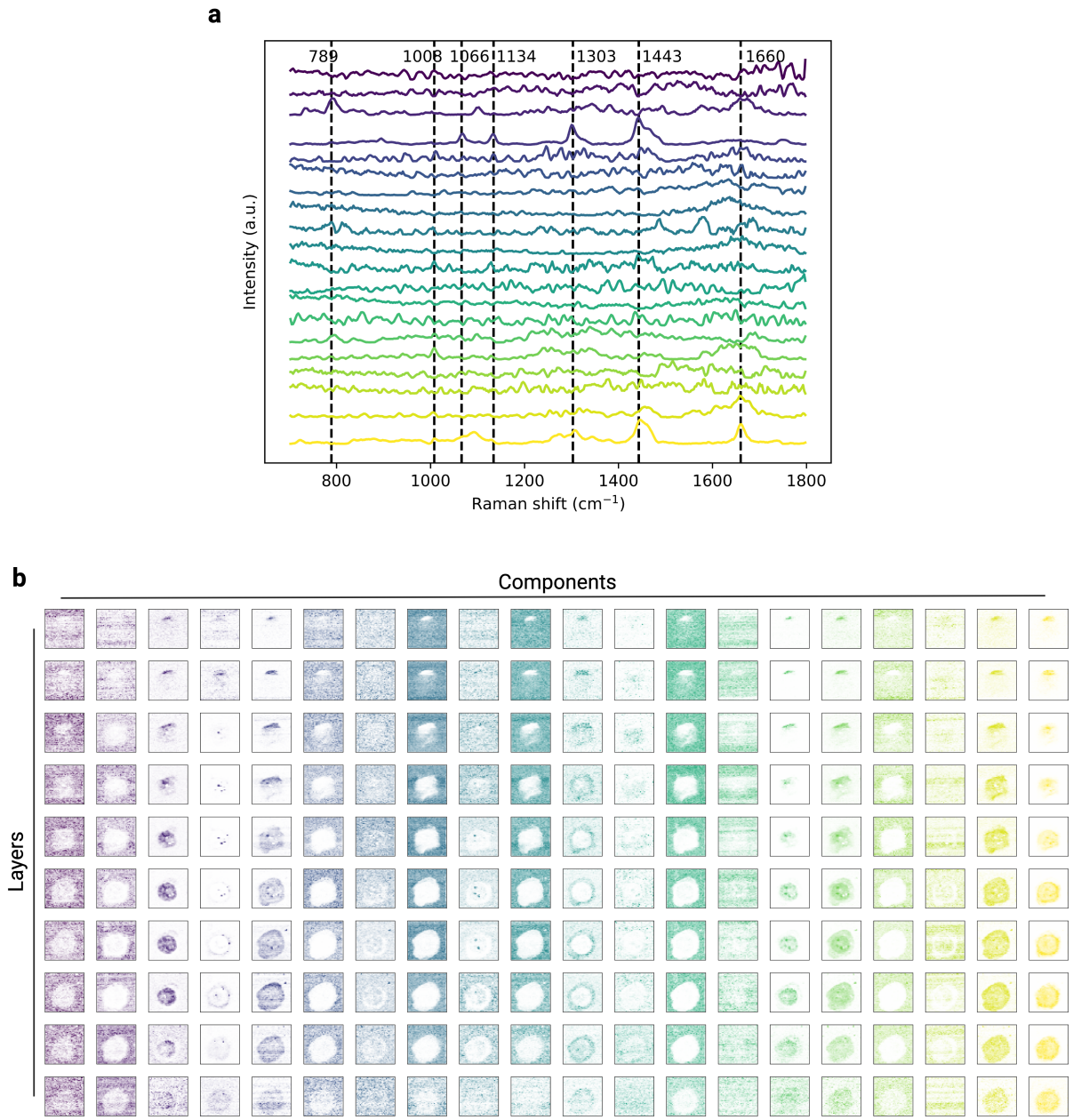


Figure C.3: Full unmixing results obtained with *Deep Dense AE* model on THP-1 cell data. **a,** Derived endmembers. **b,** Derived fractional abundances. Assembled in BioRender.com. Reproduced from Georgiev et al.[526], published by PNAS, under the CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/). © 2024 The Authors.

Appendix D

Organoid imaging: Supplementary information

Table D.1: Primary antibodies used for immunofluorescence staining. Extended information about the primary antibodies used for immunofluorescence staining. 'Cat. no.' stands for 'Catalogue number'.

Figure panel(s)	Antigen	Species	Producer	Cat. no.	Dilution
Figure 1C	SOX2	Rabbit	Abcam	ab97959	1:600
	PKCζ	Mouse	Santa Cruz Biotechnology	sc-17781	1:200
Figure 2C, Figure S3B	SOX2	Goat	R&D Systems	AF2018	1:100
	TUJ1	Mouse	BioLegend	801201	1:100
Figure 2D, Figure S5B	SOX2	Goat	R&D Systems	AF2018	1:100
	TUJ1	Mouse	BioLegend	801201	1:100
Figure 4A	SOX2	Rabbit	Abcam	ab97959	1:600
	PKCζ	Mouse	Santa Cruz Biotechnology	sc-17781	1:200
	MAP2	Chicken	Abcam	ab5392	1:500

Table D.2: Secondary antibodies used for immunofluorescence staining. Extended information about the secondary antibodies used for immunofluorescence staining. 'Cat. no.' stands for 'Catalogue number'.

Figure panel(s)	Species	Anti-	Producer	Cat. no.	Dilution
Figure 1C	Goat	Rabbit	Invitrogen	A-11008	1:500
	Donkey	Mouse	Invitrogen	A-31571	1:500
Figure 2C, Figure S3B	Donkey	Goat	Invitrogen	A-21432	1:500
	Donkey	Mouse	Invitrogen	A-31571	1:500
Figure 2D, Figure S5B	Donkey	Goat	Invitrogen	A-21432	1:500
	Donkey	Mouse	Invitrogen	A-31571	1:500
Figure 4A	Goat	Rabbit	Invitrogen	A-11008	1:500
	Donkey	Mouse	Invitrogen	A-31571	1:500
	Goat	Chicken	Invitrogen	A-21437	1:500

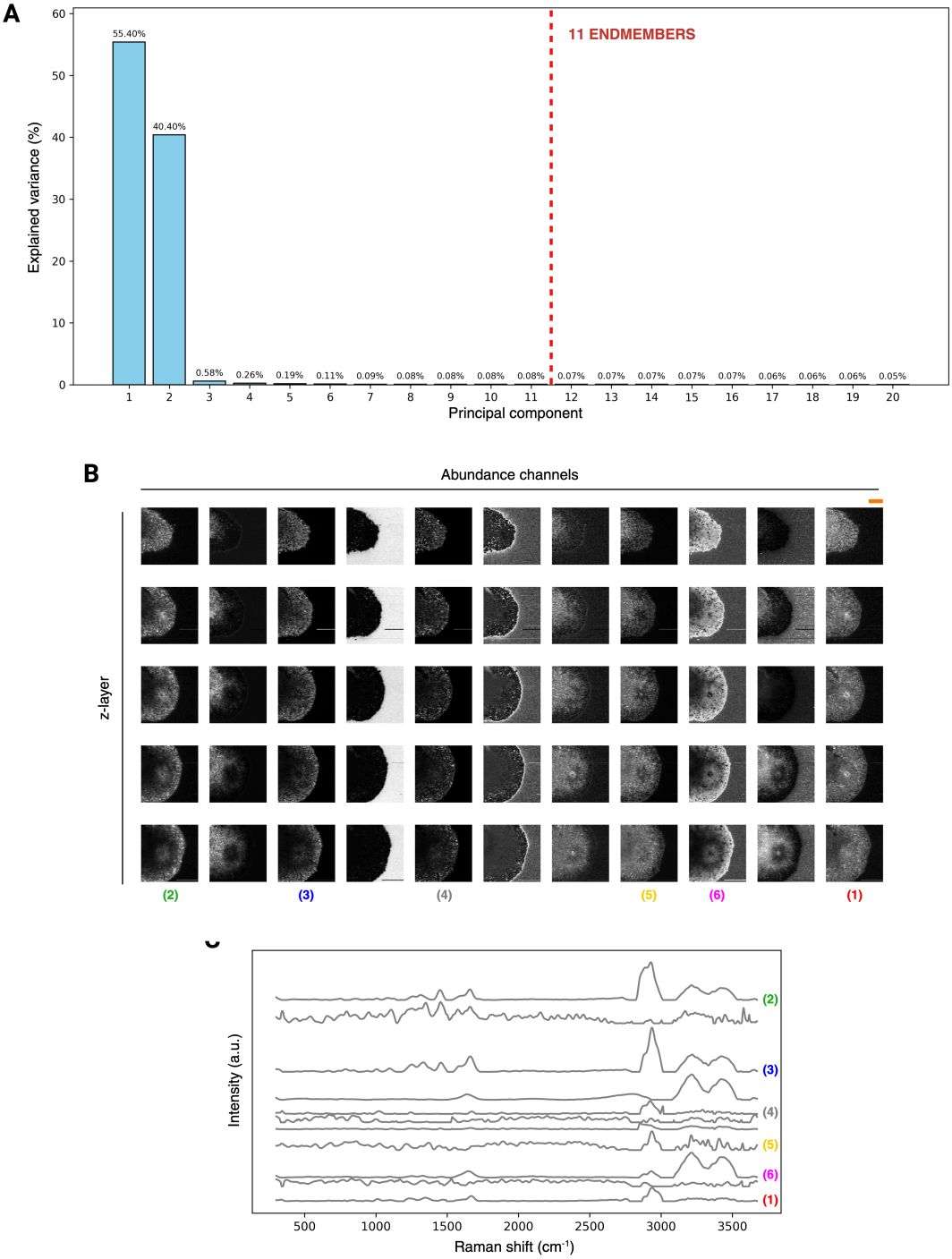


Figure D.1: Full unmixing results for low-resolution volumetric Raman imaging scan of an intact day-20 neural organoid (as shown in Figures 5.3D–H in Chapter 5). A, Principal component analysis of preprocessed RS data. The number of endmember components to extract was set to 11. **B,** Derived abundance images (normalised for visualisation purposes). Scale bar: 100 µm (top right, marked in orange). **C,** Corresponding endmembers. Endmember and abundance numbering indicate the components presented in Figures 5.3D–H. Raw RS data acquired by Dr Ruoxiao Xie and Xiaoyu Zhao. Assembled in BioRender.com.

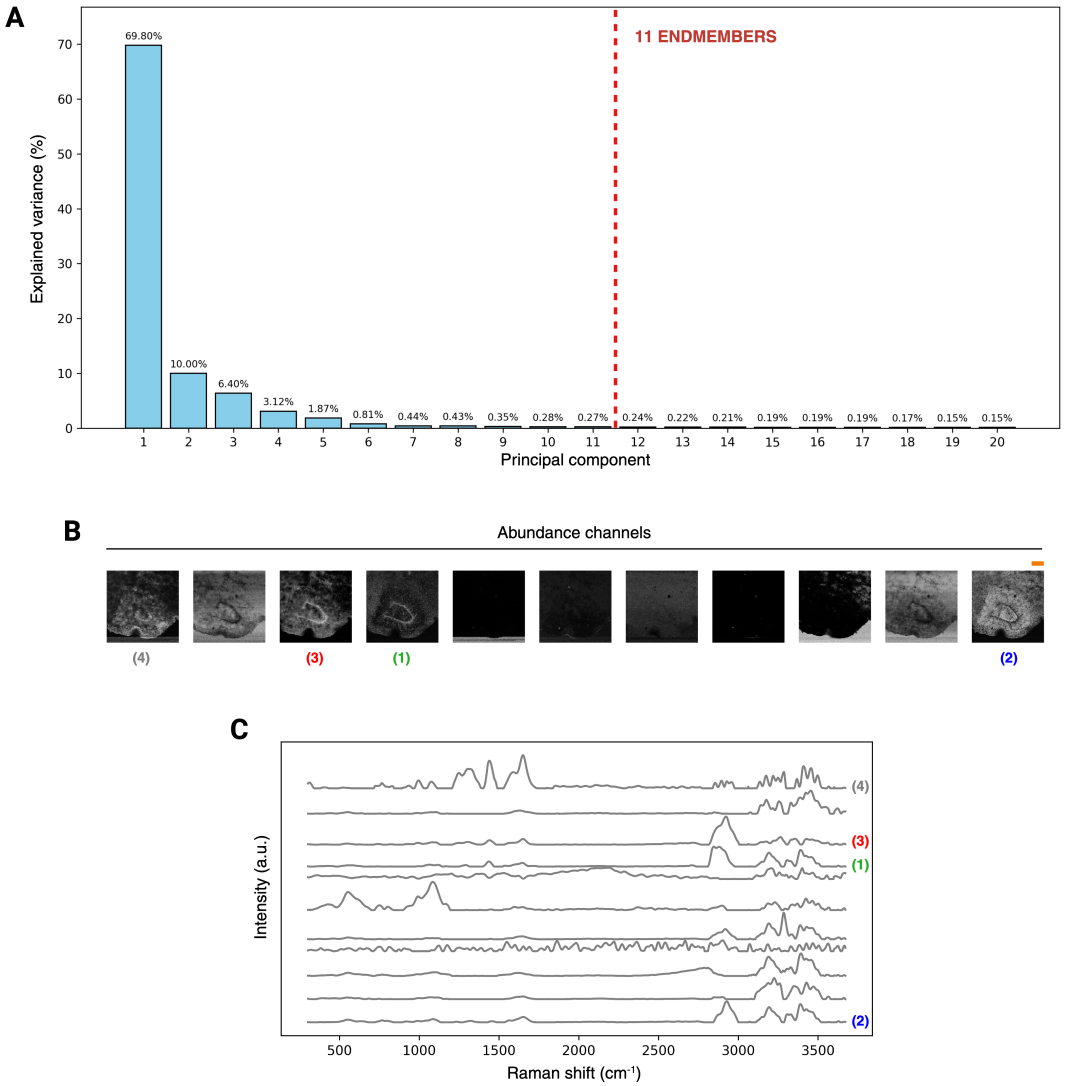


Figure D.2: Full unmixing results for Raman imaging scan of a day-30 neural organoid cryosection (as shown in Figure 5.5 in Chapter 5). A, Principal component analysis of preprocessed RS data. The number of endmember components to extract was set to 11. **B,** Derived abundance images (normalised for visualisation purposes). Scale bar: 100 µm (top right, marked in orange). **C,** Corresponding endmembers. Endmember and abundance numbering indicate the components presented in Figure 5.5. Raw RS data acquired by Dr Daniel Reumann. Assembled in BioRender.com.

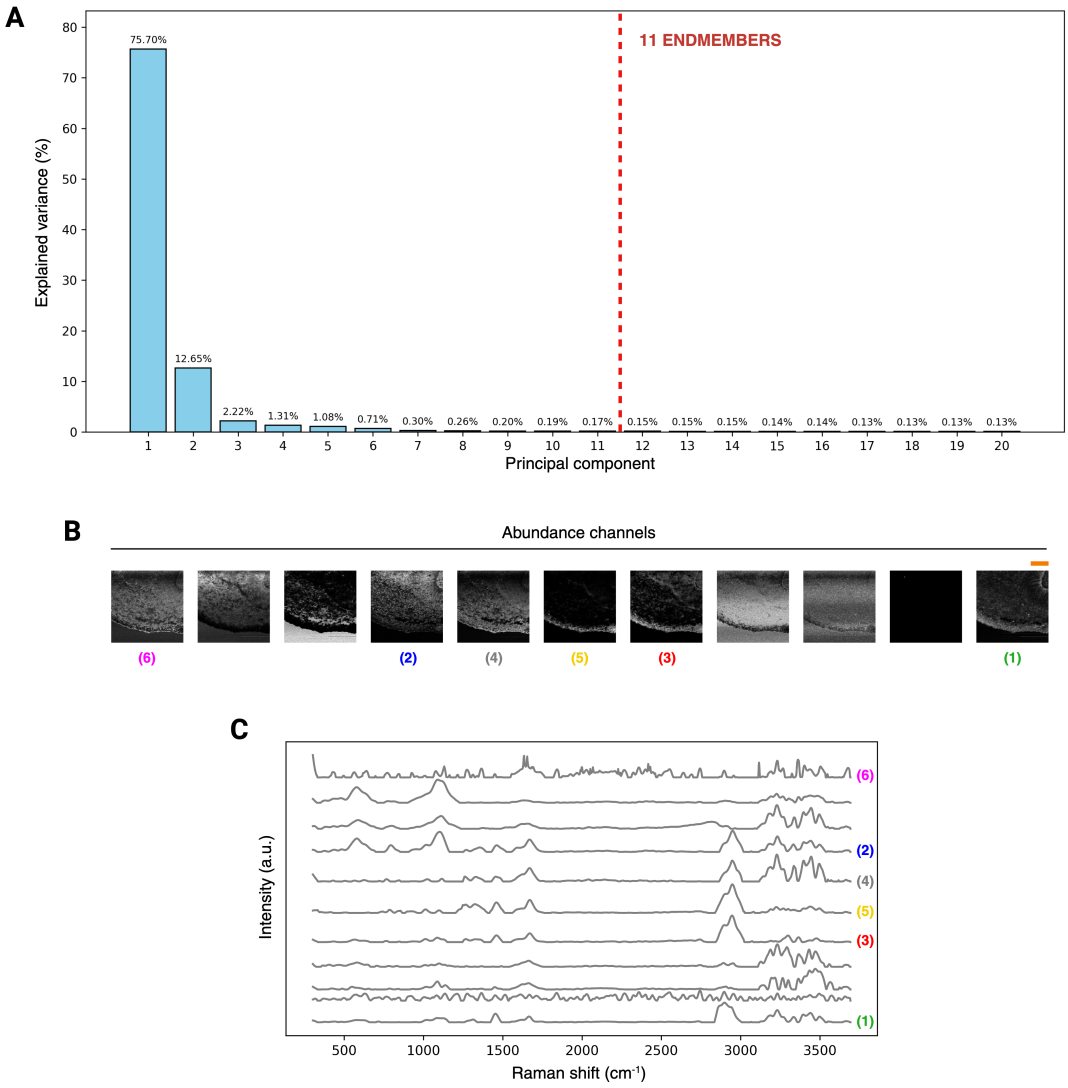


Figure D.3: Full unmixing results for Raman imaging scan of a second day-30 neural organoid cryosection (as shown in Figure 5.9 in Chapter 5). A, Principal component analysis of preprocessed RS data. The number of endmember components to extract was set to 11. **B,** Derived abundance images (normalised for visualisation purposes). Scale bar: 100 µm (top right, marked in orange). **C,** Corresponding endmembers. Endmember and abundance numbering indicate the components presented in Figure 5.9. Raw RS data acquired by Dr Daniel Reumann. Assembled in BioRender.com.

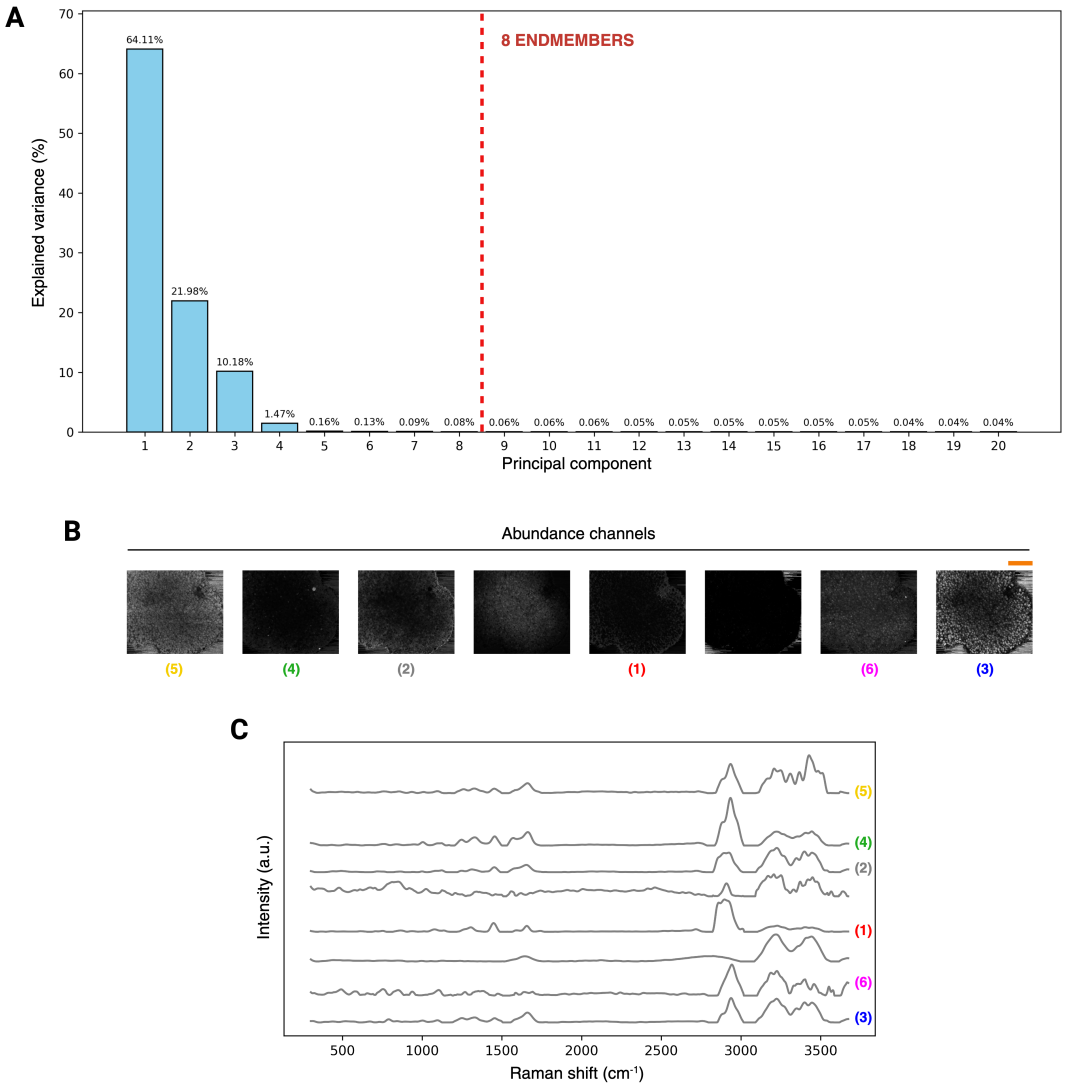


Figure D.4: Full unmixing results for high-resolution Raman imaging scan of an intact day-10 neural organoid (as shown in Figures 5.12B–E in Chapter 5). A, Principal component analysis of preprocessed RS data. The number of endmember components to extract was set to 8. **B,** Derived abundance images (normalised for visualisation purposes). Scale bar: 100 µm (top right, marked in orange). **C,** Corresponding endmembers. Endmember and abundance numbering indicate the components presented in Figures 5.12B–E. Raw RS data acquired by Dr Ruoxiao Xie and Xiaoyu Zhao. Assembled in BioRender.com.

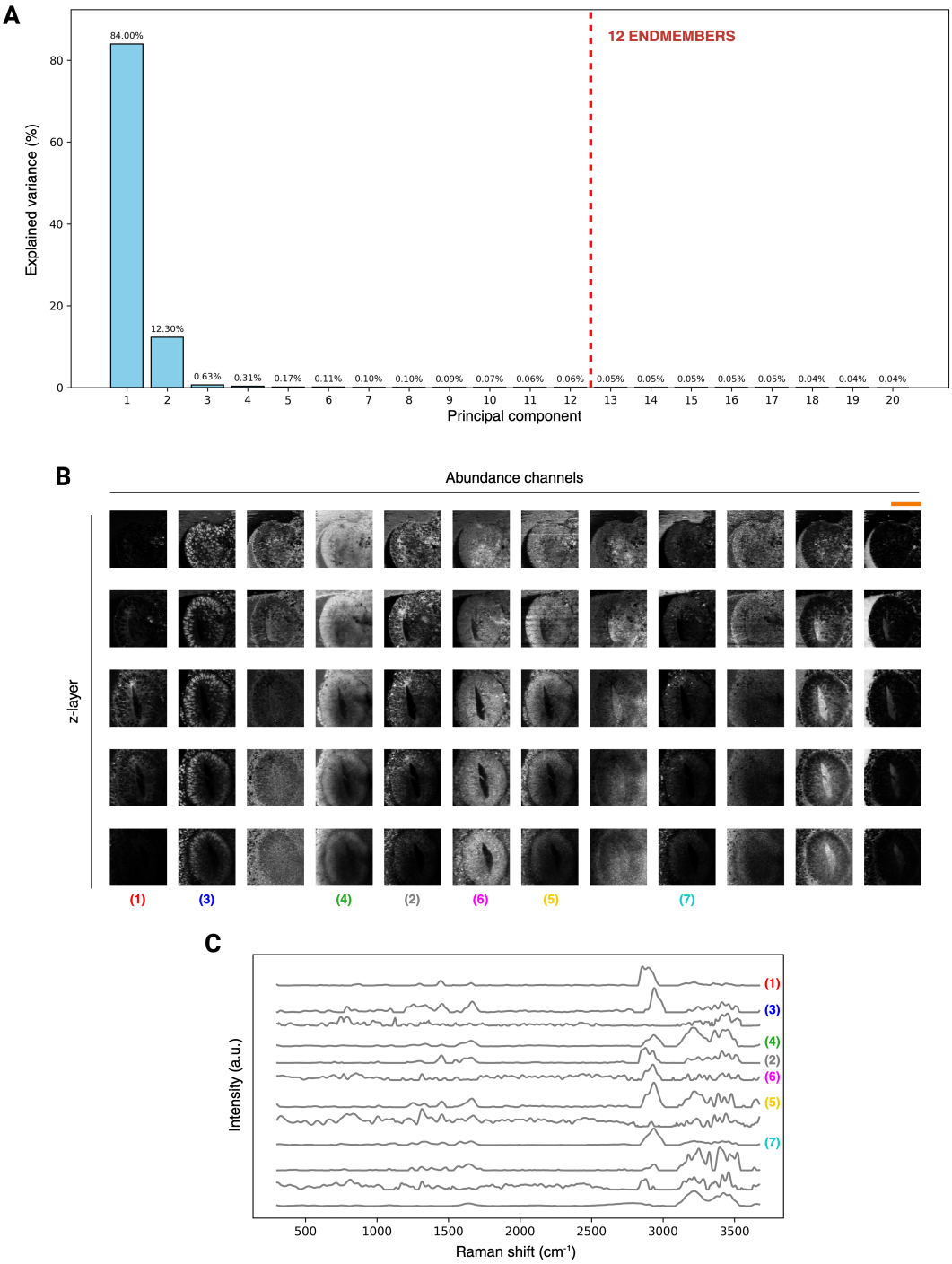


Figure D.5: Full unmixing results for volumetric Raman imaging scan of a neural rosette within an intact day-20 neural organoid (as shown in Figures 5.13B–E in Chapter 5). A, Principal component analysis of preprocessed RS data. The number of endmember components to extract was set to 12. **B,** Derived abundance images (normalised for visualisation purposes). Scale bar: 100 µm (top right, marked in orange). **C,** Corresponding endmembers. Endmember and abundance numbering indicate the components presented in Figures 5.13B–E. Raw RS data acquired by Dr Ruoxiao Xie and Xiaoyu Zhao. Assembled in BioRender.com.

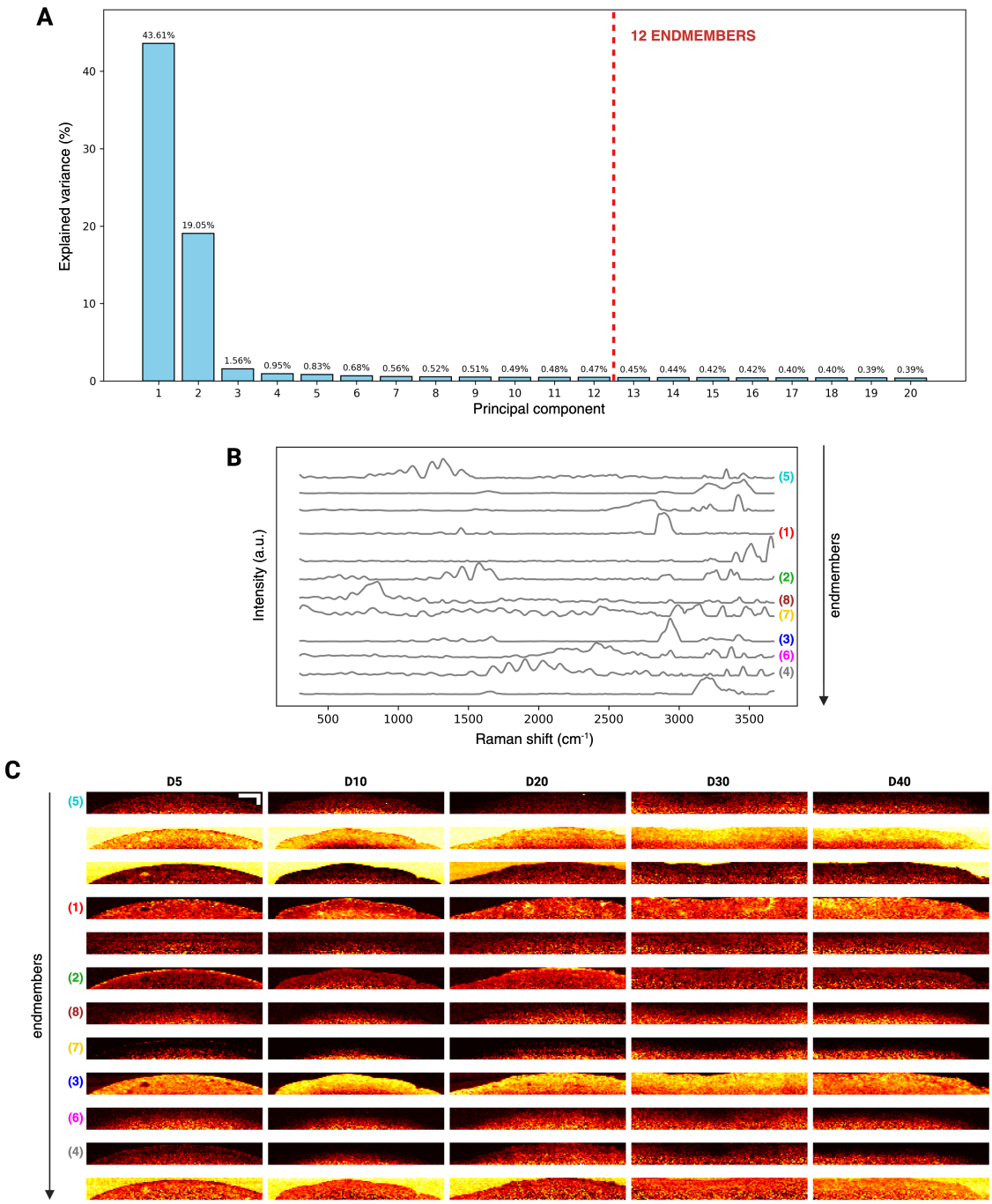


Figure D.6: Full unmixing results for timepoint Raman imaging data (as shown in Figure 5.16 in Chapter 5). A, Principal component analysis of preprocessed RS data. The number of endmember components to extract was set to 12. **B,** Derived endmembers. **C,** Corresponding abundance images (normalised for visualisation purposes) for five representative organoid samples. Endmember and abundance numbering indicate the components presented in Figure 5.16. Scale bar: 100 µm (top left). Raw RS data acquired by Dr Ruoxiao Xie and Xiaoyu Zhao. Assembled in BioRender.com.

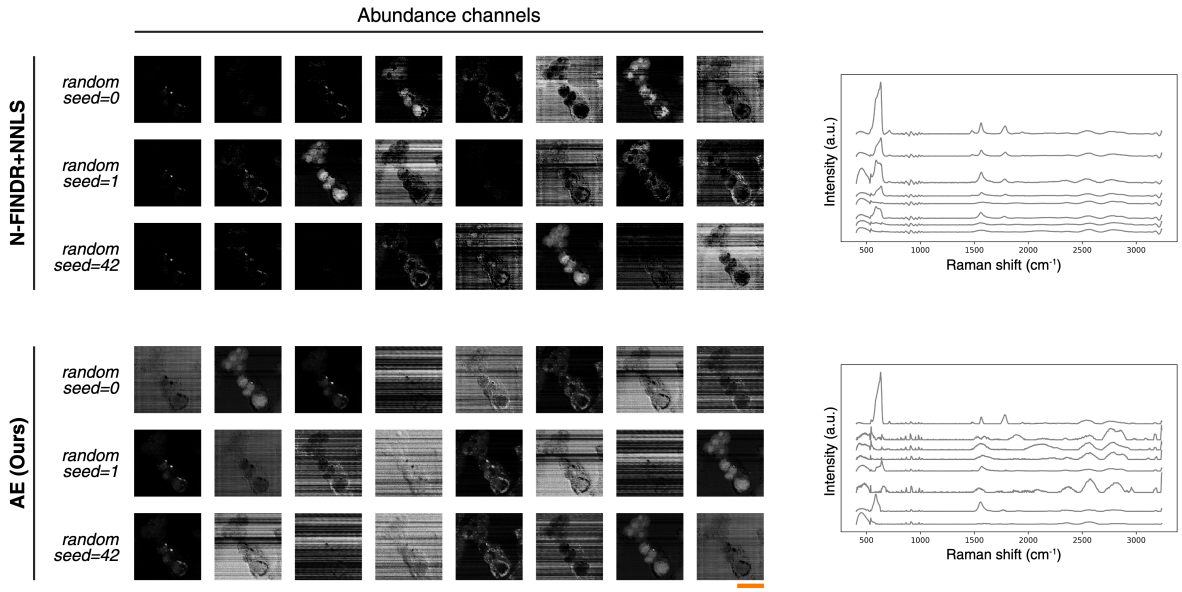


Figure D.7: Comparative analysis on broadband CARS imaging data from a cancer cell (as shown in Figures 6.1 in Chapter 6). Unmixing analyses with N-FINDR+NNLS and the autoencoder (AE) pipeline presented in Chapter 5 on broadband CARS data from a HepG2 human liver cancer cell. Abundance images were normalised for visualisation purposes. Endmember signatures displayed on the right correspond to the analyses performed with a random seed set to 42. Scale bar: 20 µm (bottom right, marked in orange). Preprocessed CARS data from Vernuccio et al.[577], provided by Dr Arianna Bresci. Assembled in BioRender.com.