

Applications of Single-Cell Raman Microspectroscopy in Bacteria Metabolism and Physiology



Jiabao Xu
St Edmund Hall
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Michaelmas 2020

Acknowledgements

Over the last couple of years, I was fortunate to be supported by a number of outstanding people, both academically and socially.

First and foremost, I would like to express my gratitude to my supervisor, Prof Wei Huang, who led me into this amazing field and has given me not only sound guidance but also encouragement and confidence throughout my DPhil study. I am incredibly grateful for his eternal stream of new ideas and all of the time he put into discussions about my work, which have been fundamental to the progress of my DPhil.

I would also like to thank every member of our simcell group and Begbroke Farmhouse for their insight and support. I would like to thank Dr Yizhi Song in particular for his training, ideas and his inspiring scientific attitude. I would like to express my special thanks to Dr Lorena Preciado-Llanes and Dr Anna Aulicino, who have selflessly spent so much time and been providing tremendous help on the *Salmonella* project. I also want to thank my other collaborators and coauthors for their valuable input, especially Dr Isabel Webb and Prof Philip Poole on the *Rhizobium* projec, Mr Christoph Martin Decker, Prof Frank Schmidt and Prof Alison Simmons on the *Salmonella* project, Dr Karl Morten and Dr Tiffany Lodge on the ME/CFS project, and Dr Chiachen Hsu and Prof Cathy Ye on the stem-cell project.

I am grateful for the financial support from the ME Association as well as the Engineering Department and St Edmund Hall.

I am very thankful for the friendship with a lot of wonderful people in Oxford like Dr Xuejie Ding, Dr Matthias Qian, Ms Yi Wang, Dr Yixing Wu, Mr Guotai Shi, Mr Yifan Sun, Dr Diming Xu, Ms Yanyao Zhou and many more. They have always been able to cheer me up when I needed it. I am also incredibly grateful to my friends away from Oxford, particularly Ms Zongqi Li, Dr Dong Ding, Dr Tszkan Ma, Dr Xiaoyan Lin and Dr Yonghua Yin. They have been a great support to me from the beginning.

I would like to thank my husband-to-be Dr Zhaoxu Hou, my biggest supporter and encourager. I can't thank him enough for his endless love and inspiration.

Finally, it would be difficult to overstate my appreciation and gratitude to my parents, who encouraged and made it possible for me to pursue this degree and have always done everything they could for me.

Declaration

I declare that the work described within this thesis is my own, except where explicitly stated below, and that this thesis is my own composition.

Chapter 3

The strains and genetic mutants of *Rhizobium* were provided by Isabel Webb, as indicated in Table 3.1. The bacterioid preparation was performed by Isabel Webb. The transmission electron microscopic image was provided by Isabel Webb.

Chapter 4

The measurement of part of the Raman spectra was helped by Di Zhu. The theoretical calculation of Raman frequencies was performed by Christopher C. R. Allen, Christopher M. Gibson and Patrick W. Fowler.

Chapter 5

The *Salmonella* infection experiment was helped by Lorena Preciado-Llanes and Anna Aulicino. The proteomic measurement and protein identification were performed by Christoph Martin Decker and Frank Schmidt.

List of Publications

Partial work of this thesis is based on the following publications:

1. **Jiabao Xu**, Isabel Webb, Phillip Poole, Wei E. Huang. "Label-Free Discrimination of Rhizobial Bacteroids and Mutants by Single-Cell Raman Microspectroscopy". *Analytical Chemistry*, 89, 6336-6340 (2017).
2. **Jiabao Xu**, Di Zhu, Aliyu D. Ibrahim, Christopher C. R. Allen, Christopher M. Gibson, Patrick W. Fowler, Yizhi Song, Wei E. Huang. "Raman Deuterium Isotope Probing Reveals Microbial Metabolism at the Single-Cell Level". *Analytical Chemistry*, 89, 13305-13312 (2017).
3. **Jiabao Xu**, Lorena Preciado-Llanes, Anna Aulicino, Christoph Martin Decker, Maren Depke, Manuela Gesell Salazar, Frank Schmidt, Alison Simmons, Wei E. Huang. "Single-Cell and Time-Resolved Profiling of Intracellular Salmonella Metabolism in Primary Human Cells". *Analytical Chemistry*, 91, 7729-7737 (2019).

Other Raman-centred publications:

4. Yizhi Song, Li Cui, Jose Angel Siles Lopez, **Jiabao Xu**, Yong-guan Zhu, Ian P. Thompson, Wei E. Huang. "Raman-Deuterium Isotope Probing for in-situ identification of antimicrobial resistant bacteria in Thames River". *Scientific Reports*, 7, 16648 (2017).
5. Xiaofei Yuan, Yanqing Song, Yizhi Song, **Jiabao Xu**, Yinhu Wu, Andrew Glidle, Maggie Cusack, Umer Z. Ijaz, Jonathan M. Cooper, Wei E. Huang, Huabing Yin. "Effect of Laser Irradiation on Cell Function and Its Implications in Raman Spectroscopy". *Applied and Environmental Microbiology*, 84, e02508-17 (2018).
6. **Jiabao Xu**, Michelle Potter, Cara Tomas, Joanna L. Elson, Karl J. Morten, Joanna Poulton, Ning Wang, Hanqing Jin, Zhaoxu Hou, Wei E. Huang. "A new approach to find biomarkers in chronic fatigue syndrome/myalgic encephalomyelitis (CFS/ME) by single-cell Raman micro-spectroscopy". *Analyst*, 144, 913-920 (2019).

7. Bo Liang, Yuan Gao, **Jiabao Xu**, Yizhi Song, Liming Xuan, Ting Shi, Ning Wang, Zhaoxu Hou, Yi-Lei Zhao, Wei E. Huang, Zi-Jiang Chen. "Raman profiling of embryo culture medium to identify aneuploid and euploid embryos". *Fertility and Sterility*, 111, 753-762 (2019).
8. Yi Wang, **Jiabao Xu**, Lingchao Kong, Tang Liu, Lingbo Yi, Hongjuan Wang, Wei E. Huang, Chunmiao Zheng. "Raman–deuterium isotope probing to study metabolic activities of single bacterial cells in human intestinal microbiota". *Microbial Biotechnology* (2019).

Rest of the publications:

9. Jack Xiaoyu Chen, Harrison Steel, Yin-Hu Wu, Yun Wang, **Jiabao Xu**, Cordelia P. N. Rampley, Ian P. Thompson, Antonis Papachristodoulou, Wei E. Huang. "Development of Aspirin-Inducible Biosensors in Escherichia coli and SimCells". *Applied and Environmental Microbiology*, 85, e02959-18 (2019).
10. Yi Wang, Zi-Feng Zhang, Xi-Jun Xu, Chuan Chen, **Jia-Bao Xu**, Ling-Chao Kong, Peng Xie, Chun-Miao Zheng, Nan-Qi Ren, Duu-jong Ren. "Effective removal of methyl siloxane from water by sewage activated sludge microbes: biodegradation behavior and characteristics of microbial community". *Bioresource Technology Reports*, 7, 100209 (2019).

Manuscripts under preparation:

11. Yi Wang, **Jiabao Xu**, Lingchao Kong, Wei E. Huang, Chunmiao Zheng. "Raman-activated sorting of antibiotic-resistant bacteria in human gut microbiota". *Environmental Microbiology* (under review).
12. Chia-Chen Hsu, **Jiabao Xu**, Bas Brinkhof, Hui Wang, Zhanfeng Cui, Wei E. Huang, Hua Ye. "A Single-Cell Raman-based Platform to Identify Developmental Stages of Human Pluripotent Stem Cell-derived Neurons".
13. Chun Kiat Ng, **Jiabao Xu**, Zhao Cai, Liang Yang, Ian Thompson, Wei E. Huang, Bin Cao. "Elevated intracellular cyclic-di-GMP level in *Shewanella oneidensis* increases expression of c-type cytochromes".
14. Isabel Webb, Jiabao Xu, Ramakrishnan Karunakaran, Elaine Barclay, Wei E. Huang, N. J. Watmough, Philip S. Poole. "Regulation of the fixABCXnifAB operon in *Rhizobium leguminosarum* in symbiosis with *Pisum sativum*".

Abstract

Bacteria play important roles in almost all ecosystems, including the natural environment and the human body, and they are extremely abundant and versatile. New generations of sequencing technologies can provide phylogenetic diversity and metagenomic profiling of microbial communities, but metabolic functions of individual microbes and their molecular mechanisms are often elusive, owing to the limited culturability and the complexity of microbial communities.

Single-cell Raman microspectroscopy (SCRM) is label-free vibrational spectroscopy that measures a collection of molecular vibrations coming from all biomolecules in a single bacterial cell. This thesis emphasises the capabilities of SCRM in a wide spectrum of applications for investigating bacterial metabolism and physiology. SCRM is applied to study bacterial phenotypes in a symbiotic setting with a plant host. SCRM is combined with deuterium-isotope-probing (DIP) and applied to study carbon metabolism in pure cultures, as well as *in vivo* metabolic adaptations by human-invasive pathogens. Especially, the power of single-cell analysis in studies of heterogeneous bacterial populations is explored and highlighted.

This thesis also outlines an analytical pipeline to standardise and automate the process of spectral analysis in SCRM studies, from data tidying and data preprocessing to a variety of analytical approaches for either exploratory or diagnostic purposes. By following the established pipeline and executing the developed programming script, single-cell Raman analysis can be performed in a simple and rapid manner.

To conclude, by applying SCRM to provide intrinsic biochemical fingerprints of individual bacterial cells and connecting SCRM with the power of computational analysis, this thesis provides the possibility of probing bacterial metabolism and physiological states in ever-changing and multifaceted ecological settings.

Contents

List of Figures	x
List of Tables	xviii
List of Abbreviations	xx
1 Introduction	1
1.1 Aim and objectives	2
1.2 Structure of the thesis and key findings	2
2 Background	4
2.1 Principles of Raman spectroscopy	5
2.1.1 Classical description of Raman scattering	6
2.1.2 Quantum mechanical description of Raman scattering	8
2.2 Raman practices in biological studies	10
2.2.1 Advances in spatial resolution of a Raman microscope	10
2.2.2 Advantages of Raman microspectroscopy in biological studies . .	12
2.2.3 Enhancement of Raman cross-section	14
2.3 Phenotyping bacteria by Raman microspectroscopy	18
2.3.1 A cultivation-free approach to study bacterial single cells	19
2.3.2 Identification of cellular components	20
2.3.3 Identification of bacteria	24
2.3.4 Computational strategies for the analysis of Raman spectra	27
2.4 You are what you eat—stable isotope probing of bacteria	31
2.4.1 Principles of Raman-SIP	32
2.4.2 ¹³ C-labelling in Raman-SIP	33
2.4.3 ¹⁵ N-labelling in Raman-SIP	34
2.4.4 D-labelling in Raman-SIP: Raman-DIP	36
2.5 Outlook	40

3	Application of SCRM in phenotyping symbiotic plant bacteria	41
3.1	Introduction	41
3.2	Materials and Methods	45
3.3	Results and Discussion	47
3.3.1	SCRS distinguish cells in different physiological states	47
3.3.2	SCRS link phenotypic changes with gene knockouts	52
3.3.3	Single-cell resolution allows heterogeneity analysis within a single bacterial population	58
3.4	Conclusions	63
4	Application of SCRM-DIP in revealing microbial carbon metabolism	66
4.1	Introduction	66
4.2	Materials and Methods	67
4.3	Results and Discussion	70
4.3.1	SCRM-DIP detects carbon metabolism by using deuterated carbon sources	70
4.3.2	SCRM-DIP reveals distinct biosynthesis pathways of fatty acids starting from different deuterated carbon sources	77
4.3.3	SCRM-DIP reveals distinct biosynthesis pathways of phenylalanine starting from different deuterated carbon sources	79
4.4	Conclusions	87
5	Application of SCRM-DIP in dynamic mapping of intracellular <i>Salmonella</i> metabolism	88
5.1	Introduction	89
5.2	Materials and Methods	90
5.3	Results and Discussion	97
5.3.1	Phenotypic profiles based on SCRS successfully distinguish closely related <i>Salmonella</i> strains	97
5.3.2	Intracellular <i>Salmonella</i> shows distinctive metabolic activities revealed by SCRM-DIP	102
5.3.3	SCRS map dynamic phenotypic profiles of intracellular <i>Salmonella</i> within primary cells	108
5.3.4	Single-cell Raman analysis resolves phenotypic heterogeneity in isogenic populations	113
5.3.5	Intracellular Raman profiles are consistent with the proteomic analysis	116
5.4	Conclusions	122

6 Pipeline of SCRS Analysis	125
6.1 Introduction	125
6.2 Data tidying and data preprocessing	126
6.3 Data analysis	132
6.3.1 Data visualisation	132
6.3.2 Classification	140
6.3.3 Feature Extraction	151
6.4 Conclusion	156
7 Summary and Outlook	162
Appendices	
A R Code for Raman Analysis Pipeline	168
A.1 Data tidying	168
A.2 Data Pre-processing	169
A.2.1 Baseline correction	169
A.2.2 Normalisation	169
A.3 Data Analysis	170
A.3.1 Visualisation	170
A.3.2 Classification	172
A.3.3 Feature extraction	174
References	176

List of Figures

2.1	(A) Electromagnetic radiation interacting with a vibrating molecule. (B) The appearance of the Stokes ν_0 and the anti-Stokes lines.	5
2.2	Schematic energy diagrams of various processes of photon–molecule interaction. The energy spacing between two vibrational levels equals to ν_m	8
2.3	Selective enhancement of chromophores in RR spectroscopy (picture taken from reference [3])	15
2.4	Schematic energy diagrams of CRS. The energy spacing between two vibrational levels equals to ν_m	17
2.5	SCRS of (A) <i>E. coli</i> , (B) <i>Rhizobium</i> and (C) <i>Rhodobacter</i> . The assignments of the Raman band are labelled in superscripts at the wavenumbers. The abbreviations are listed as follows: ‘NA’ stands for nucleic acids; ‘CH’ stands for carbohydrates; ‘AA’ stands for aromatic amino acids; ‘Lip’ stands for lipids, ‘PHB’ stands for polyhydroxybutyrate; ‘cyc’ stands for cytochrome <i>c</i> ; ‘carot’ stands for carotenoid.	21
2.6	Raman spectra of <i>E. coli</i> with highlighted bands caused by incorporation of (A) ^{13}C or ^{12}C and (B) D or H into the biomass, modified from reference [126].	35
2.7	D ₂ O as a unique and universal tracer for different biomolecules. ‘rds’ stands for rate-determining step.	38
2.8	D incorporation during the reduction reactions to synthesise biomolecules via the action of NADP ⁺ /NADPH.	38
3.1	<i>Rhizobium</i> -legume symbiosis to form root nodules: (A) Bacteria sense flavonoids secreted from the roots of host leguminous plants. (B) Bacteria produce Nod factors which in turn causes root hair growth and deformation and formation of infection threads. (C) Bacteria are encapsulated in infection threads. (D) Infection threads, after reaching the root cortex, release bacteria into plant cells and initiate nodule development. Inside root nodules, bacteria turn into bacteroids. The figure is taken and modified from reference [154].	43

3.2	(A) SCRS of the wild-type Rlv3841 as either symbiotic bacteroids in root nodule (red) or free-living bacteria in liquid culture (blue), averaged from 30-50 single cells in each condition. Standard deviation is plotted as shaded area. (B) PCA plot along PC1 and PC2 based on the SCRS of Rlv3841 as either bacteroids (red) or bacteria (blue).	48
3.3	PCA contribution plots of squared PCA loading scores of Raman wavenumber to PC1 and PC2	48
3.4	(A) SCRS with four cytochrome <i>c</i> bands labelled. (B) Spectral quantification of cytochrome <i>c</i> bands. (C) SCRS with two DNA/RNA bands labelled. (D) Spectral semi-quantification of DNA/RNA bands.	50
3.5	Chemical structure of cytochrome <i>c</i>	51
3.6	(A) SCRS of the wild-type Rlv3814 and three derived mutants. (B) SCRS of the wild-type RIVA34 and two derived mutants.	53
3.7	(A) Quantification of PHB. (B) Quantification of glycogen. (C) Quantification of polyphosphate.	56
3.8	Bright-field microscopic images of the wild-type Rlv3841 mature bacteroids (A & B), and the $\Delta fixAB$ mutant LMB777 with two distinct morphologies (C & D). Scale bar 2 μm	58
3.9	(A) SCRS of the wild-type Rlv3841 and the mutant LMB777, showing two distinct Raman phenotypes. (B) The probability density function for a normal distribution of cell sizes quantified for each individual cell of Rlv3814 and LMB777. Cells of LMB777 were grouped by two Raman phenotypes of either PHB+ or PHB-. The average cell size in μm^2 for each group is annotated in bold font.	60
3.10	Individual cell size versus Raman quantification of (A) PHB and (B) polyphosphate. A t-test performed on the slope of the linear regression is indicated in red.	61
3.11	Electron microscopic images of sections of nodules taken from <i>P. sativum</i> inoculated with LMB777.	62
4.1	SCRS of (A) <i>E. coli</i> DH5 α and (B) <i>P. putida</i> UWC1 grown in glucose with different ratios of deuterated glucose- d_{12} . (C) <i>P. putida</i> UWC1 (pWH2-Nah ⁺) and (D) <i>P. putida</i> G7 grown in naphthalene with different ratios of deuterated naphthalene- d_8 . The characteristic C-D Raman band centred at 2170 cm^{-1} is highlighted.	71
4.2	(A) SCRS of <i>P. putida</i> UWC1 (pWH2-Nah ⁺) grown in DMSO without any naphthalene and <i>P. putida</i> UWC1 grown in 100% naphthalene- d_8 , both of which have a silent Raman window between 2070 - 2300 cm^{-1} . (B) The growth curves of <i>P. putida</i> UWC1 (pWH2-Nah ⁺) grown in DMSO without any naphthalene and <i>P. putida</i> UWC1 grown in 100% naphthalene- d_8 , measured by OD ₆₀₀	72

4.3	Growth curves created by OD ₆₀₀ measurements of cultures of (A) <i>E. coli</i> DH5 α and (B) <i>P. putida</i> UWC1 grown with different ratios of deuterated glucose- d_{12} , (C) <i>P. putida</i> UWC1 (pWH2-Nah ⁺) and (D) <i>P. putida</i> G7 grown with different ratios of deuterated naphthalene- d_8	73
4.4	Biological mechanism of the incorporation of deuterium from deuterated carbon source, using glucose- d_{12} as an example. Deuterium in the deuterated carbon source contributes to a new C–D bond during the reduction reactions to synthesise biomolecules such as lipids via the action of NADP ⁺ /NADPH.	74
4.5	Comparison of the percentage of D content in carbon source with the band intensity ratio of C–D / (C–D + C–H) in SCRS of <i>E. coli</i> DH5 α and <i>P. putida</i> UWC1 grown with glucose, and of <i>P. putida</i> UWC1 (pWH2-Nah ⁺) and <i>P. putida</i> G7 grown with naphthalene, with different percentages of glucose- d_{12} or naphthalene- d_8	74
4.6	(A) Molecular structure of palmitic acid. The optimisation yielded a fully extended conformer, with mirror symmetry. (B)-(E). Simulated Raman spectra for palmitic acid as a simplified model for aliphatic CH environments in lipids. The spectra from top to bottom are calculated for (B) the all-H form of the molecule, (C) the all-D form, (D) a form in which all H positions to one side of the mirror plane are replaced by D and (E) a sample form of single-D substitution on the methyl group at the end.	76
4.7	Fatty acids metabolic pathways starting from (A) glucose- d_{12} and (B) naphthalene- d_8	77
4.8	Enlarged view of phenylalanine shifts from 1004 cm ⁻¹ to 988, 976 and 962 cm ⁻¹ in Figure 4.1 with respect to different ratios of deuterated growth glucose- d_{12} or naphthalene- d_8	79
4.9	Chemical structure of phenylalanine with 5 C–H positions on the phenyl ring. Only the red-labelled 1, 3 and 5 sites of the symmetric ring-breathing mode are responsible for the observed phenylalanine Raman shifts. . . .	80
4.10	Four distinct Raman-band positions are found at 1004, 989–990, 976 and 962–963 cm ⁻¹ , for the 32 possible isotopomers of phenylalanine	83
4.11	Comparison of D content in the phenylalanine band in SCRS of <i>E. coli</i> DH5 α and <i>P. putida</i> UWC1 grown in glucose, and of <i>P. putida</i> UWC1 (pWH2-Nah ⁺) and <i>P. putida</i> G7 grown in naphthalene, with different percentages of glucose- d_{12} or naphthalene- d_8	85
4.12	Phenylalanine metabolic pathways of (A) <i>E. coli</i> grown in glucose- d_{12} , (B) <i>P. putida</i> UWC1 grown in glucose- d_{12} , and (C) <i>P. putida</i> UWC1 (pWH2-Nah ⁺) grown in naphthalene- d_8	86

- 5.1 (A) Average SCRS of overnight cultures of STM-LT2, STM-D23580 and ST-Ty2 with biological assignments of important biomolecules. (B) PCA plot demonstrating distinctive metabolic properties of STM-LT2, STM-D23580 and ST-Ty2 in the Raman fingerprint region. (C) Boxplots show semi-quantification of selected biomolecules by integrating relevant Raman bands. Abbreviations of the biomolecules are listed as follows: “CH” stands for “Carbohydrates”; “DNA” stands for “DNA/RNA”; “Tyr” stands for “Tyrosine”; “Phe” stands for “Phenylalanine”. 98
- 5.2 Loading plot of SCRS of STM-LT2, STM-D23580 and ST-Ty2, with Raman wavenumbers against their contributions to PC1 of the PCA. 100
- 5.3 SCRS of STM-LT2 and *E. coli* DH5 α . Bacterial cultures were grown overnight in LB broth in the presence of 35% D₂O. A broad C–D band indicates an incorporation of D during active metabolism. The shaded area represents standard deviation from 30 single-cell measurements. . . 102
- 5.4 SCRS and PCA plots of D-labelled STM-LT2 under different conditions: (A) and (B) growth at post-freezing or fresh state; (C) and (D) growth with or without sucrose, 103
- 5.5 (A) SCRS of *Salmonella* serovars under four experimental conditions: extracellular growth in liquid culture, and intracellular metabolism during infection of a differentiated THP-1 cell line, primary monocyte-derived dendritic cells (MoDCs) and primary macrophages (Mf). (B) Boxplots quantifying D content at different time points. Linear regression shows the D decreasing speed for each strain at each experimental condition. Notation above the box plots indicates statistical significance of the three strains at each time point by one-way ANOVA test (ns: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; ****: $p < 0.0001$). Linear regression was fitted to compare the D decreasing speed of each strain in each infection condition and the slope was labelled as β . The absolute value of β represents the metabolism speed. 105
- 5.6 (A) Extracellular growth of D-labelled bacteria in RPMI. Intracellular growth of D-labelled bacteria in (B) THP-1, (C) MoDCs and (D) Mf. An average of two experiments is shown. 106
- 5.7 Raman profiling of **carbohydrates** in intracellular bacteria. Raman bands of Carbohydrates I, II and III were quantified in fingerprint region of SCRS of STM-LT2, STM-D23580 and ST-Ty2 at different time point of infection in (A) THP-1 cells, (B) dendritic cells and (C) macrophages. 109
- 5.8 Raman profiling of **lipids** in intracellular bacteria. Raman bands of Lipid I and TUFA/TFA were quantified in fingerprint region of SCRS of STM-LT2, STM-D23580 and ST-Ty2 at different time point of infection in (A) THP-1 cells, (B) dendritic cells and (C) macrophages. 110

- 5.9 Raman profiling of **nucleic acids** in intracellular bacteria. Raman bands of Adenine, DNA/RNA, and the ratio of DNA/RNA to Amide I were quantified in fingerprint region of SCRS of STM-LT2, STM-D23580 and ST-Ty2 at different time point of infection in (A) THP-1 cells, (B) dendritic cells and (C) macrophages. 112
- 5.10 Raman profiling of **amino acids and proteins** in intracellular bacteria. Raman bands of Tyrosine, Phenylalanine, Amide III and Amide I were quantified in fingerprint region of SCRS of STM-LT2, STM-D23580 and ST-Ty2 at different time point of infection in (A) THP-1 cells, (B) dendritic cells and (C) macrophages. 113
- 5.11 (A) A t-SNE plot of single cells (perplexity = 10) computed based on SCRS of STM-LT2, STM-D23580 and ST-Ty2 at 2 hr p.i. of MoDCs and Mf. Three bacterial subpopulations were clustered in MoDCs (MoDC1, MoDC2 and MoDC3) or in Mf (Mf1, Mf2 and Mf3). (B) t-SNE plots coloured based on quantification of different biomolecules. The intensity gradient is mapped from green (low) to red (high). 114
- 5.12 Single-cell heatmap from Raman integrations of important biomolecules demonstrates serovar-specific heterogeneity during the course of infection of (A) THP-1, (B) MoDCs and (C) Mf. The heatmap is column-scaled and each column of the heatmap represents one-time point and within each time point, Raman intensities were ordered from the highest to the lowest. Each row of the heatmap represents a group of biomolecule and rows were ordered based on hierarchical clustering dendrogram to view proximity and relationships of different biomolecules. 116
- 5.13 Venn diagram of proteins detected in STM-LT2, STM-D23580 and ST-Ty2, in at least 3 time points. 117

- 5.14 Dotplots of the abundance of DEPs (Log2 of the LFQ intensity) related to (A) energy and (B) biosynthesis pathways in STM-LT2 (green), STM-D23580 (blue) and ST-Ty2 (red). The large dot represents the average of all Log2 protein expression values belonging to a specific biosynthesis pathway at one time-point. Each of the small dots represents a data value for an individual protein within the pathway. Time-wise changes were calculated as linear regression slopes and represented as “++” and “+” being upregulated ($>10^{-1}$ and $>10^{-2}$), “-” and “-” being downregulated ($<-10^{-1}$ and $<-10^{-2}$), and “ns” being non-significant ($<10^{-3}$) over time. Abbreviations of the pathways are listed as follows: “AA” stands for “Amino Acid Biosynthesis”; “Amine” stands for “Amines and Polyamines Biosynthesis”; “Carbo” stands for “Carbohydrates Biosynthesis”; “CellStr” stands for “Cell Structures Biosynthesis”; “Cofactor” stands for “Co-factors, Prosthetic Groups, Electron Carriers Biosynthesis”; “FA/Lip” stands for “Fatty Acids and Lipids Biosynthesis”; “MetReg” stands for “Metabolic Regulators Biosynthesis”; “Nucleo” stands for “Nucleosides and Nucleotides Biosynthesis”; “Other” stands for “Other Biosynthesis”; “SecMet” stands “Secondary Metbolites Biosynthesis”. 118
- 5.15 Heatmap of the row-normalised abundance of differentially expressed proteins (DEPs) related to energy pathways in (A) STM-LT2, (B) STM-D23580 and (C) ST-Ty2 or biosynthesis pathways in (D) STM-LT2, (E) STM-D23580 and (F) ST-Ty2. Colour bars on the left present assignment of protein to one specific energy pathway and individual protein names are listed on the right. Abbreviations of the pathways are listed as follows: “Aero” stands for “Aerobic Respiration”; “Anaero” stands for “Anaerobic Respiration”; “Chemo” stands for “Chemoautotrophic Energy Metabolism”; “Ferment” stands for “Fermentation”; “Glycol” stands for “Glycolysis”; “Other” stands for “Other Energy Pathways”; “PPP” stands for “Pentose Phosphate Pathway”; “TCA” stands for “The Citric Acid Cycle”. 119

5.16 Heatmap of the row-normalised abundance of differentially expressed proteins (DEPs) related to biosynthesis pathways in (A) STM-LT2, (B) STM-D23580 and (C) ST-Ty2. Colour bars on the left present assignment of protein to one specific energy pathway and individual protein names are listed on the right. Abbreviations of the pathways are listed as follows: “AA” stands for “Amino Acid Biosynthesis”; “Amine” stands for “Amines and Polyamines Biosynthesis”; “Carbo” stands for “Carbohydrates Biosynthesis”; “CellStr” stands for “Cell Structures Biosynthesis”; “Cofactor” stands for “Cofactors, Prosthetic Groups, Electron Carriers Biosynthesis”; “FA/Lip” stands for “Fatty Acids and Lipids Biosynthesis”; “MetReg” stands for “Metabolic Regulators Biosynthesis”; “Nucleo” stands for “Nucleosides and Nucleotides Biosynthesis”; “Other” stands for “Other Biosynthesis”; “SecMet” stands for “Secondary Metabolites Biosynthesis”.	121
6.1 Raman analysis pipeline.	126
6.2 Data tidying process immediately after acquisition of Raman spectra. . .	127
6.3 Comparison of an original spectrum and the baseline-corrected spectrum with the use of (A) modified polynomial fitting (degree = 4), (B) asymmetric least squares ($\lambda = 7$) and (C) rolling ball (wm = 200, ws = 100). Two badly corrected examples are given in (D) using modified polynomial fitting (degree = 20) and (E) using asymmetric least squares ($\lambda = 1$).	129
6.4 The original spectra (A) are baseline-corrected (B).	130
6.5 The original spectra (A) are baseline-corrected (B) and then normalised (C).	131
6.6 Data visualisation by (A) averaged spectra, (B) averaged spectra with standard deviation, (C) PCA, (D) LDA and (E) t-SNE.	134
6.7 Examples of PCA plots from Chapter 3 and Chapter 5.	135
6.8 (A) Scree plot showing the percentage of the variances explained by each PC. (B) The contribution of each variable to PC1 and PC2.	136
6.9 Stacked histograms of the values of (A) the first discriminant function and (B) the second discriminant function from the LDA of the Raman spectra in different groups.	137
6.10 Visualisation results of PCA and t-SNE with a perplexity value of 5, 10, 30, 50, 100 or 500.	139
6.11 The pipeline of machine-learning classification using Raman spectra. . .	141
6.12 An example of model building and parameter tuning process using kNN. .	142
6.13 An example of the results of the model building and parameter tuning process using GBM.	143

6.14 Performance of 8 models on stem-cell dataset.	146
6.15 Accuracy scatter plots of 50 resampled predictions between pairs of the eight primary models.	149
6.16 Performance of eight models and the stacked model on stem-cell dataset.	150
6.17 Feature extraction during the Raman analysis pipeline.	151
6.18 LDA loading plots showing the contribution of Raman wavenumber to (A) the first discriminant function LD1 and (B) the second discriminant function LD2. The wavenumbers with high contribution are labelled. . .	152
6.19 Variable importance extracted from the SVM-RBF model, showing top 50 wavenumbers responsible for classification of iPSC.	153
6.20 LDA loading plots showing the contribution of Raman wavenumber to (A) the first discriminant function LD1 and (B) the second discriminant function LD2. The wavenumbers with high contribution are identified. .	155
6.21 Quantification of (A) glycogen, (B) phenylalanine and (C) cytochrome c by integrating relevant Raman bands.	156
7.1 Potential of SCRM-DIP to study cell activity and metabolism.	165

List of Tables

3.1	Strains used in Chapter 3	45
3.2	Assignment of Raman bands in section 2.3.1	49
3.3	Assignment of Raman bands in section 2.3.2 and 2.3.3	54
4.1	Strains and plasmid used in Chapter 2	68
4.2	The equilibrium geometry of 32 possible isotopomers of the phenylalanine molecule calculated at the B3LYP/6 31G** level of theory. ν_1 and ν_2 are two Raman wavenumbers generated with an intensity of I1 and I2, respectively. Equiv. represents any free rotation equivalents of a specific isotopmer. Averaged ν is an average from two wavenumbers, ruling out equivalents from free rotations along symmetric axis.	81
4.3	The collected peaks from theoretical modelling in Table 4.2 and wavenumbers scaled by a factor of 1004/1017 for comparison with the experimental values, represented as scaled peaks.	82
4.4	Calculated band positions correlate with the experimental observations of phenylalanine Raman shifts. Theoretical Raman band wavenumbers have been scaled by a factor of 1004/1017. Black crosses represent theoretical calculation and red crosses represent experimental shifts observed in SCRS of <i>E. coli</i> as an example.	83
5.1	Assignment of Raman bands in section 4.3.1	99
5.2	Confusion matrix and performance evaluation of the kNN classification model classifying 116 STM-LT2 spectra, 152 STM-D23580 spectra and 125 ST-Ty2 spectra.	101
6.1	Models trained in this Chapter	145
6.2	Resampling correlations of the models trained in this Chapter	148
6.3	Biological and vibrational assignments of Raman bands of Proteins.	157
6.4	Biological and vibrational assignments of Raman bands of lipids.	158
6.5	Biological and vibrational assignments of Raman bands of nucleic acids.	159
6.6	Biological and vibrational assignments of Raman bands of carbohydrates.	160
6.7	Biological and vibrational assignments of Raman bands of special molecules with resonance effect under a 532-nm laser.	161

6.8 Biological and vibrational assignments of Raman band shifts related to Raman-DIP.	161
--	-----

List of Abbreviations

16S-rRNA		16S ribosomal ribonucleic acid
6PG		6-phosphogluconate
ABC		Ammonium bicarbonate
AFM		Atomic Force Microscopy
CARS		Coherent anti-Stokes Raman scattering
CCD		Charge-coupled device
CFU		Colony-forming units
CRS		Coherent Raman scattering
D₂O		Deuterium oxide
DC		Dendritic cell
DEP		Differentially expressed protein
DIP		Deuterium isotope probing
DNA		Deoxyribonucleic acid
ED		Entner-Doudoroff
EMP		Embden-Meyerhof-Parnas
<i>E. coli</i>/EC		<i>Escherichia coli</i>
FBS		Fetal bovine serum
FDR		False discovery rate
FISH		Fluorescence <i>in situ</i> hybridisation
G6P		Glucose 6-phosphate
GC-MS		Gas chromatography-mass spectrometry
GMP		Guanosine-5'-monophosphate
GM-CSF		Granulocyte macrophage-colony-stimulating factor
HCA		Hierarchical cluster analysis
HPLC		High performance liquid chromatography

iNTS		Invasive non-typhoidal salmonellosis
iPSC		Human induced pluripotent stem cell
IR		Infrared
KDPG		2-Keto-3-deoxy-6-phosphogluconate
kNN		k-nearest neighbours
LC-MS/MS	...	Liquid chromatography-tandem mass spectrometry
LDA		Linear discriminant analysis
LDC		Linear discriminant classification
LFQ		Label-free quantification
LOOCV		Leave-one-out- cross-validation
LSPR		Localised surface plasmon resonance
Mf		Macrophages
MoDC		Monocyte-derived dendritic cell
MOI		Multiplicity of infection
MS		Mass spectrometry
M-CSF		Macrophage-colony-stimulating factor
NADP⁺/NADPH		Nicotinamide adenine dinucleotide phosphate
NMR		Nuclear magnetic resonance
NO		Nitric oxide
NSC		Neural stem cell
OD		Optical density
PBMC		Peripheral blood mononuclear cell
PBS		Phosphate-buffered saline
PCA		Principal component analysis
PEP		Phosphoenolpyruvate
PFA		Paraformaldehyde
PHB		Polyhydroxybutyrate
PLS		Partial least squares
PTS		Phosphotransferase
<i>P. fluorescens</i>	..	<i>Pseudomonas fluorescens</i>
p.i.		Post infection

<i>P. putida</i>	<i>Pseudomonas putida</i>
<i>P. sativum</i>	<i>Pisum sativum</i>
qPCR	Quantitative polymerase chain reaction
RNA	Ribonucleic acid
RR	Raman resonant
<i>R. leguminosarum</i>	<i>Rhizobium leguminosarum</i>
SCRM	Single-cell Raman microspectroscopy
SCRS	Single-cell Raman spectra/spectrum
SCV	<i>Salmonella</i> -containing vacuole
SERS	Surface-enhanced Raman spectroscopy
SIMS	Secondary ion mass spectrometry
SIP	Stable isotope labelling
SNR	Signal-to-noise ratio
SPA	Successive projections algorithm
SRS	Stimulated Raman scattering
ST	Sequence type
STM-D23580 . .	<i>Salmonella enterica</i> serovar Typhimurium D23580
STM-LT2	<i>Salmonella enterica</i> serovar Typhimurium LT2
ST-Ty2	<i>Salmonella enterica</i> serovar Typhi Ty2
SVM	Support vector machine
<i>S. enterica</i>	<i>Salmonella enterica</i>
<i>S. Paratyphi</i> . . .	<i>Salmonella enterica</i> serovar Paratyphi
<i>S. Typhi</i>	<i>Salmonella. enterica</i> serovar Typhi
<i>S. Typhimurium</i>	<i>Salmonella enterica</i> serovar Typhimurium
T2SS	Type II secretion system
TEM	Transmission electron microscopy
TERS	Tip-Enhanced Raman Spectroscopy
TFA	Trifluoroacetic acid
t-SNE	t-Distributed Stochastic Neighbour Embedding
TUFA/TFA	Total unsaturated fatty acids/total fatty acids
UV	Ultraviolet
VBNC	Viable but non-culturable

Chapter 1

Introduction

Bacteria express a wide spectrum of phenotypic traits with high relevance for research and development. Bacteria inhabit a wide variety of ecological niches, including colonisation in extreme environments and symbiosis with plant or animal hosts, being mutualistic, commensal or parasitic. This is made possible by a plethora of metabolic and physiological adaptations observed in bacteria, such as multiple abilities for oxygen and nitrate respiration, switches between nutrient sources, resistance against stressors and molecular interactions with host cells. These traits of metabolism and physiology can all be broadly construed as bacterial phenotypes.

While whole genome sequences are rapidly growing in numbers (206,448 bacterial genomes are found in the NCBI genome database, accessed in May 2019), advances and annotations of phenotypic attributes of bacteria are not keeping pace. This also means that the potential for comprehensive genotype–phenotype association studies cannot be realised.

Single-cell Raman microspectroscopy (SCRM) is vibrational spectroscopy that measures the inelastic light scattering of molecular vibrations. SCRM has recently been applied in various biological fields to study phenotypes of single cells. It surpasses other analytical techniques in terms of its single-cell resolution. While bulk-scale cellular analysis averages signals from an entire population and loses track of rare and low-level events, single-cell approaches are able to draw out the least prevalent but most important phenotypes. SCRM is also label-free and non-destructive. By applying SCRM to provide intrinsic biochemical fingerprints of individual bacteria cells, we can study bacterial metabolism and physiology in multifaceted ecological settings.

1.1 Aim and objectives

The main aim of this thesis is to explore some of the uncharted areas in the biological applications of SCRM in profiling bacterial phenotypes, as well as to develop and validate a robust pipeline of single-cell Raman analysis.

The detailed objectives include:

- To study bacteria in a symbiotic setting with a plant host and link phenotypic alterations with genetic mutations by application of SCRM
- To consolidate a technique combining SCRM and deuterium-isotope-probing (DIP)
- To apply SCRM-DIP to study bacterial carbon synthetic and metabolic pathways
- To apply SCRM-DIP in a symbiotic setting with a human host and study metabolic adaptations of a pathogen
- To explore the power of single-cell analysis in heterogeneous bacterial populations
- To develop and standardise an analytical approach of SCRM suitable for a wide range of biological applications

1.2 Structure of the thesis and key findings

The thesis is structured and key findings are summarised as follows:

Chapter 2 briefly introduces the theoretical and technical background of Raman spectroscopy. In addition, applications using Raman spectroscopy for bacterial investigations are reviewed with an emphasis on the advantages of single-cell measurements. The technique of Raman-DIP is also discussed, followed by its applications in exploring bacterial metabolism.

Chapter 3 investigates the physiology and metabolism of a symbiotic plant bacterium *Rhizobium leguminosarum*. Single-cell Raman spectra (SCRS) are used as

biochemical fingerprints of *Rhizobium* cells to differentiate different physiological states, link phenotypic changes and specific genetic mutations, and assess single-cell heterogeneity within a single bacterial population. By analysing SCRS of different strains and mutants, it provides insights into metabolite production and carbon regulatory network of rhizobia.

Chapter 4 elaborates on the realisation of SCRM-DIP to reveal microbial carbon metabolism. By incubating bacteria with deuterated carbon sources, a distinguishable C–D vibration band in SCRS can be used to semi-quantitatively and sensitively indicate the degree of carbon metabolism and lipid conversion. Different pathways of phenylalanine adopted by different microorganisms are also revealed by analysing corresponding Raman shifts. By harnessing low-cost and versatile deuterated substrates, this work demonstrates that SCRM-DIP has the potential to probe a wide range of metabolic pathways and functions at the single-cell level.

Chapter 5 continues to explore the symbiotic lifestyle of a pathogen *Salmonella* inside human immune cells. SCRM-DIP reveals metabolic changes of three *Salmonella* strains. Reverse labelling of deuterium tracks their metabolic changes in the time-course infection of immune cells. In comparison with a non-invasive serovar, the invasive *Salmonella* strains are found to down-regulate metabolic activity inside human macrophages and dendritic cells and use lipids as an alternative carbon source, perhaps as a strategy to escape from the host immune response. Proteomic analysis using high sensitivity mass spectrometry validates the findings of SCRM-DIP analysis.

In Chapter 6, a pipeline of SCRS analysis in biological studies is established and validated, devised from a collection of avenues in Chapter 3–5. This pipeline involves data tidying, data preprocessing and differing chemometric and machine learning techniques to fulfil a variety of research purposes. By implementing the established protocol, Raman analysis can be performed by a non-specialist in a simple and rapid manner.

The final Chapter 7 provides a short summary of the contents of this thesis and concludes with an outlook.

Chapter 2

Background

Contents

2.1 Principles of Raman spectroscopy	5
2.1.1 Classical description of Raman scattering	6
2.1.2 Quantum mechanical description of Raman scattering	8
2.2 Raman practices in biological studies	10
2.2.1 Advances in spatial resolution of a Raman microscope	10
2.2.2 Advantages of Raman microspectroscopy in biological studies	12
2.2.3 Enhancement of Raman cross-section	14
2.3 Phenotyping bacteria by Raman microspectroscopy	18
2.3.1 A cultivation-free approach to study bacterial single cells . . .	19
2.3.2 Identification of cellular components	20
2.3.3 Identification of bacteria	24
2.3.4 Computational strategies for the analysis of Raman spectra . .	27
2.4 You are what you eat—stable isotope probing of bacteria	31
2.4.1 Principles of Raman-SIP	32
2.4.2 ¹³ C-labelling in Raman-SIP	33
2.4.3 ¹⁵ N-labelling in Raman-SIP	34
2.4.4 D-labelling in Raman-SIP: Raman-DIP	36
2.5 Outlook	40

This chapter introduces the background of Raman spectroscopy, starting with two descriptions of the underlying mechanism of Raman scattering and working through to the instrumentation and advances of modern Raman microscopes that benefit biological studies. Some of the many applications using Raman spectroscopy for bacteria investigations will be explored, with a specific focus on the advantages of single-cell measurements. Finally, the technique of the combination of stable isotope labelling and Raman spectroscopy will be discussed, followed by its applications in studying bacterial metabolism.

2.1 Principles of Raman spectroscopy

Spectroscopy is a class of analytical tools based on the interaction between electromagnetic radiation and matter which ranges from absorption, emission, fluorescence to scattering¹. Raman spectroscopy is a spectroscopic technique based on Raman scattering. It was first experimentally observed by C. V. Raman in 1928 for which he was awarded the Nobel Prize in 1930².

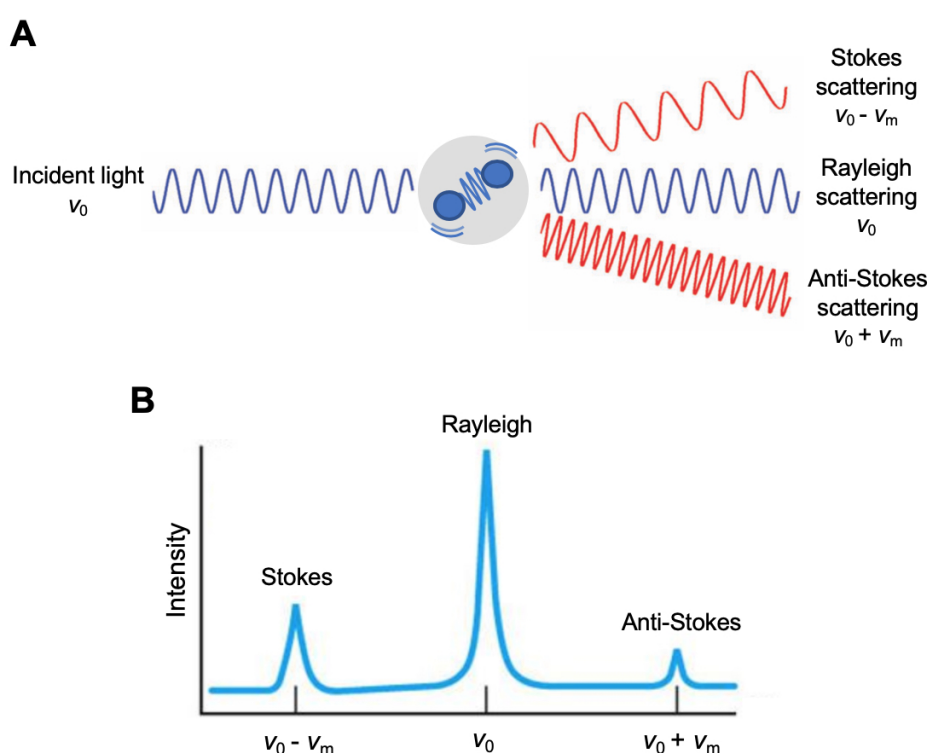


Figure 2.1: (A) Electromagnetic radiation interacting with a vibrating molecule. (B) The appearance of the Stokes ν_0 and the anti-Stokes lines.

In Raman spectroscopy, a sample is illuminated with a monochromatic laser beam (Figure 2.1A). The incident light with a frequency of ν_0 interacts with a molecule in the sample that has a molecular vibration at ν_m and originates scattered light. Most scattering is elastic and has a frequency equal to the frequency of the incident radiation ν_0 , so-called Rayleigh scattering. A small fraction of the scattered radiation is inelastic and has a frequency different from the frequency of the incident light, collectively called Raman scattering. Raman scattering consists of both Stokes scattering, which

has a lower frequency at $\nu_0 - \nu_m$, and anti-Stokes scattering, which has a higher frequency at $\nu_0 + \nu_m$. The appearance of the anti-Stokes lines is identical to the Stokes lines, only with strongly reduced intensities (Figure 2.1B). Thus, the Stokes spectrum is often presented as the Raman spectrum.

2.1.1 Classical description of Raman scattering

The classical picture of light scattering is the interaction between an oscillating electromagnetic field and a molecule. An incident light wave serves as the source of a travelling electric field, which will exert force on all electrons in the molecule and tends to displace them from their average positions around the positively charged nuclei. This change in the electronic charge distribution in the molecule thereby results in an induced dipole moment, which also emits radiation with characteristic radiation. The induced dipole moment, μ , is proportional to the electric field E :

$$\mu = \alpha E \quad (2.1)$$

where α is the polarisability of the molecule, describing the susceptibility of the charge distribution of a molecule to have its charges displaced by the electric field.

The electric field E of the light wave varies with time and can be described by its vector amplitude E_0 and its oscillation/excitation frequency ν_0 :

$$E = E_0 \cos(\nu_0 t) \quad (2.2)$$

Thus,

$$\mu = \alpha E_0 \cos(\nu_0 t) \quad (2.3)$$

The polarisability of a molecule is not a simple scalar quantity and can be expanded in a Taylor series in the normal co-ordinates q of the nuclear displacements:

$$\alpha(q) = \alpha_0 + \sum_{m=1}^Q \left[\left(\frac{\partial \alpha}{\partial q} \right)_0 q_n + \dots \right] \quad (2.4)$$

where there are $Q = 3M - 6$ normal modes and α_0 is the polarisability at the equilibrium position when $q = 0$.

In turn, the normal co-ordinates can be approximated by

$$q(t) = q_0 \cos(\nu_m t) \quad (2.5)$$

Inserting Equation 2.4 and Equation 2.5 into Equation 2.3 yields:

$$\mu(t) = \left[\alpha_0 + \left(\frac{\partial \alpha}{\partial q} \right)_0 q_0 \cos(\nu_m t) \right] E_0 \cos(\nu_0 t) \quad (2.6)$$

By using the formula for the product of two cosine functions Equation 2.6 can be written as:

$$\mu(t) = \alpha_0 E_0 \cos(\nu_0 t) + \frac{1}{2} \left(\frac{\partial \alpha}{\partial q} \right)_0 q_0 E_0 \cos(\nu_0 - \nu_m) t + \frac{1}{2} \left(\frac{\partial \alpha}{\partial q} \right)_0 q_0 E_0 \cos(\nu_0 + \nu_m) t \quad (2.7)$$

Equation 2.7 represents the conceptual core of Raman scattering. The classical electromagnetic theory states that an oscillating dipole radiates energy in the form of scattered light. Thus, Equation 2.7 demonstrates that the time-dependent induced dipole moment μ that acts as the secondary source of scattered excitation has three components:

1. The first term on the right-hand side of Equation, 2.7 $\alpha_0 E_0 \cos(\nu_0 t)$, represents the scattered light at the same frequency as the incident radiation ν_0 , i.e. elastic Rayleigh scattering.
2. The second term represents the component vibrating at a frequency of $\nu_0 - \nu_m$, i.e. the frequency of the incident light minus that of the molecular vibration. The scattered light resulting from this term is thus the redshifted Stokes Raman scattering.
3. The third term is characterised by $\nu_0 + \nu_m$, i.e. the sum of the frequencies of incident light and molecular vibrations, thereby the anti-Stokes scattering.

2.1.2 Quantum mechanical description of Raman scattering

The classical approach offers a useful conceptual model for understanding Raman scattering, however, it also has some shortfalls and fails to explain some of the Raman phenomena. To give a simple example, the appearance of discrete lines in a rotational Raman spectrum of a gaseous molecule cannot be explained as the classical theory does not assign discrete frequencies to rotational transitions³.

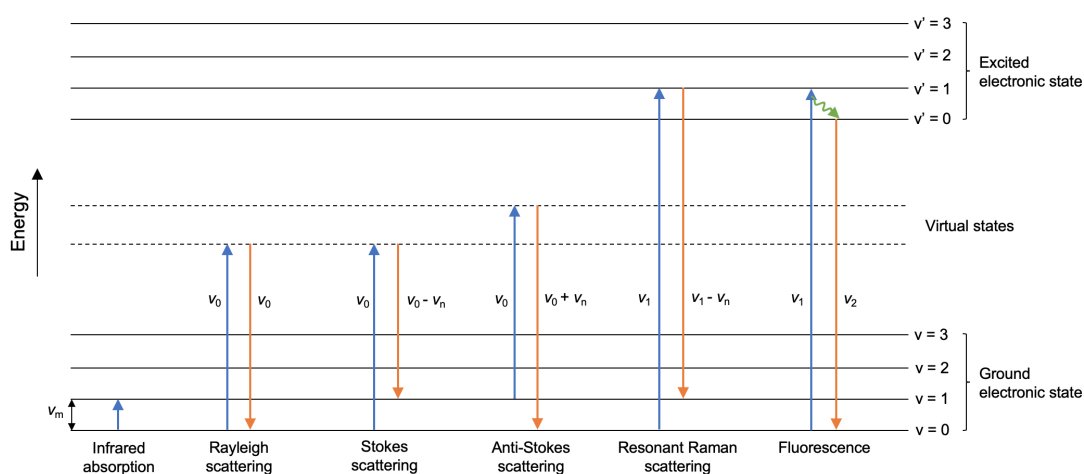


Figure 2.2: Schematic energy diagrams of various processes of photon–molecule interaction. The energy spacing between two vibrational levels equals to ν_m .

The quantum mechanical description of Raman scattering is quite different. The concept of photons that are quanta of light particles, is used to depict the wave–particle duality of light. Moreover, the quantisation of vibrational and electronic energies is taken into consideration. In the quantum mechanical model, scattering is described as a two-photon process, shown as the blue and red arrows in Figure 2.2. The blue arrow represents the first step of the process where a photon interacts with a molecule and raise it to a higher energy state. The red arrow represents the second step of the release of a photon after a very short period of time, achieving overall conservation of energy.

The various processes of photon–molecule interaction can now be outlined as in Figure 2.2. Rayleigh scattering involves transitions of photons with the same energy (frequency). In Stokes scattering, the photon returns to a vibrational level that is higher than the starting vibrational level, thus the second photon has an energy of

$\nu_0 - \nu_n$ where ν_n equals to the energy difference between two vibrational energy states. The anti-Stokes scattering involves exciting the molecule in a higher vibrational state and returning to a lower vibrational state, thus the photon released has a quantised energy of $\nu_0 + \nu_n$.

In the classical model, Equation 2.7 depicts the three scattering processes without noting any differences between the Stokes and anti-Stokes scattering as they share the same coefficient. However, this quantum mechanical picture shows that the molecule must be in a higher vibrational state for the anti-Stokes scattering to take place. According to the Boltzmann distribution, the population of higher states is much smaller than that of lower states. The relative ratio of populations of the $\nu = 1$ to the $\nu = 0$ vibrational states can be denoted as N_1 / N_0 and calculated as:

$$\frac{N_1}{N_0} = \exp\left(-\frac{h\nu}{k_B T}\right) \quad (2.8)$$

where h is Planck's constant, ν is the vibrational frequency, k_B is the Boltzmann constant and T is the temperature.

This means that at a temperature of 300 K, the ratio of N_1 / N_0 is 0.1 with a vibrational wavenumber of 480 cm^{-1} and 0.001 with a wavenumber of 1440 cm^{-1} . This results in much reduced intensities of anti-Stokes lines compared with Stokes lines in Raman practices. Therefore, a Stokes spectrum is often chosen to represent the Raman spectrum of a sample.

Having explored the transitions between vibrational levels, we now also consider the transitions between two electronic states. This process is associated with a very important phenomenon called resonant Raman (RR) scattering, which significantly improves the intrinsically weak Raman intensities. When a photon has energy enough to excite the molecule to the upper electronic state, the direct absorption of the photon can occur and the emission of the second photon gives rise to the RR process (Figure 2.2).

The band intensities in an RR spectrum can be orders of magnitude higher than a Stokes Raman spectrum⁴. In a biochemical context, this not only means easier detection but also introduces the possibility of selectively probing some key molecules

within a complex biological system by choosing an excitation wavelength affiliated with chromophores that are of interest.

However, RR is not the only consequence of photon absorption as fluorescence is also one of the pathways electrons use to return to their ground states (Figure 2.2). As the fluorescence cross-section is about six orders of magnitude larger than the RR cross-section, strong fluorescence background can in many cases hamper the observation of Stokes scattering and RR scattering. An instrumental approach to suppress fluorescence is based on the different time dependences of Raman scattering and fluorescence. Because Raman scattering has a much shorter lifetime (less than a picosecond) than fluorescence (nanoseconds), ultra-fast picosecond pulsed excitation can be used to achieve a significant reduction in fluorescence^{5,6}.

2.2 Raman practices in biological studies

Owing to the recent advances made in the instrumentation of Raman spectroscopy, it has become an essential tool for chemists, physicists, biologists and materials scientists. A Raman spectrum can be seen as a collection of molecular vibrations in a sample, therefore, providing compositional and chemical information in a biological context. The high molecular specificity, spatial resolution and non-destructive nature make Raman microspectroscopy an excellent technique for biological analysis, from probing the secondary structural changes of proteins to investigating more complex biological systems like bacteria. In this section, we shall start with the instrumental advances in modern Raman techniques, followed by discussing the advantages and challenges of using Raman microspectroscopy in biological studies.

2.2.1 Advances in spatial resolution of a Raman microscope

Having discussed two models describing Raman scattering in the last section, we now consider the technical and instrumental aspects of modern Raman techniques. Although first being discovered in the 1920s, the number of publications using Raman techniques didn't soar until the 1960s when the laser (monochromatic and coherent sources of light) was first invented and used as an excitation source in

Raman spectroscopy. The 1980s and 1990s ushered in the confocal systems with better spatial resolution, charge-coupled device (CCD) detectors with a faster acquisition and effective Rayleigh filters to remove the Rayleigh scattering⁷. This is the modern Raman microspectroscopy that we know of today.

In the 90s, the Raman spectrometer was first coupled with an optical microscope and so-called micro-Raman spectroscopy was established. A microscope was used to focus the excitation light to a small spot instead of a simple lens. For the first time, Raman spectra could be obtained for samples of only a few micrometres in diameter.

Nowadays, the spatial resolution of a Raman microscope is further improved by using more advanced microscope objectives with high collection efficiency. The minimal size of a laser spot is defined by the diffraction limit:

$$D = \frac{1.22\lambda}{NA} \quad (2.9)$$

where λ is the wavelength of the incident light and NA is the numerical aperture of the objective. The diffraction limit equation is derived from:

$$d = \frac{\lambda}{2n\sin\theta} = \frac{\lambda}{2NA} \quad (2.10)$$

where n is the refractive index of the medium between the lens and the sample and θ is the included angle of the lens which depends on the diameter of the lens and its focal length. The factor 1.22 in the angular resolution is decided by the angle spanned by the central bright spot of the diffraction pattern produced by light, therefore the focus of a laser beam. It is derived from a calculation of the position of the first dark circular ring surrounding the central Airy disc of the diffraction pattern.

With an objective of 100 $\times$ with an NA of 0.8 and an incident wavelength of 532 nm (used throughout this thesis as an excitation source), the spot diameter is $\sim 0.8 \mu\text{m}$. The spatial resolution is half of the laser spot diameter thus $\sim 0.4 \mu\text{m}$. Such spatial resolution allows investigation of a single bacterial cell with a typical size larger than $1 \mu\text{m}^2$.

The spatial resolution can be further improved by using shorter wavelength and higher NA objectives. High NA objectives are associated with shorter working distances

thus suitable for thin samples. The ultraviolet (UV) excitation can reduce the laser spot size but UV objectives with high NA are extremely expensive and hardly accessible⁷.

Apart from using advanced microscopes and changing wavelengths of light, the spatial resolution of a Raman microscope can be more significantly improved by combining it with other scanning probe modalities. For example, a Raman microscope integrating Atomic Force Microscopy (AFM) can provide physical characterisations on the nanometer scale, such as topography, conductivity and thermal measurements, all together with the chemical information obtained from Raman spectroscopy. Tip-Enhanced Raman Spectroscopy (TERS) is another specialist approach in which the enhancement of Raman scattering occurs at the point of a near atomically sharp pin, typically coated with gold, therefore achieving spatial resolution that is otherwise unattainable by any other analytical method⁸.

2.2.2 Advantages of Raman microspectroscopy in biological studies

Raman microspectroscopy has become increasingly popular among biologists. A Raman spectrum can be regarded as a phenotype of a biological system because it provides an overall molecular vibrational profile, containing ~1000 Raman bands for major cellular building blocks such as proteins, nucleic acids, lipids and carbohydrates. The advantages of using Raman microspectroscopy in biological studies include:

1. *High spatial resolution.* As we discussed in the last section, by combining the power of optical magnification and direct visualisation, Raman microspectroscopy can probe biological systems at a subcellular resolution. It can provide a single-cell Raman spectrum for a single bacterial cell with all macromolecules studied at a single-cell level. For larger biological systems like mammalian cells and tissues, it can not only derive single point spectra but can collectively produce Raman images with subcellular structural and chemical information.
2. *Ability to analyse aqueous samples.* Biological samples are usually prepared in an aqueous environment. Water has very strong absorption in infrared (IR) spectroscopy, which is also a vibrational spectroscopy that is often compared in

parallel with Raman spectroscopy. A water molecule has very low polarizability, thus does not hinder sample signals in Raman spectroscopy. This is particularly advantageous in biological studies because it avoids laborious sample preparation which may also alter the biochemistry of the samples.

3. *Intrinsic and label-free.* Molecular labelling using genetically encoded reporters (e.g. green fluorescent protein) has been used extensively for monitoring cellular events in biological systems⁹. They introduce unnatural functional groups and risk interfering with the native biological processes. Raman spectroscopy, on the other hand, is label-free and able to probe biological samples in their natural setting. It also does not require any prior knowledge of the particular substrate for selective labelling. It unbiasedly probes all macromolecules and collectively displays them in one spectrum.
4. *Non-contacting and non-destructive.* Raman spectroscopy is a non-contacting and non-destructive technique, as opposed to other analytical tools such as gas chromatography and mass spectrometry that destroy the sample to achieve results. This non-invasive nature renders Raman microspectroscopy suitable for *in situ* and *in vivo* biological investigations.
5. *Easy preparation and small sample volume.* Various kinds of substrates like liquid, solution, gases, films, surfaces and single crystals are easily amenable to examination, therefore minimising the process of sample preparation, extraction and purification. The requirement of only a very small amount of sample in Raman microspectroscopy is also very beneficial to applications from small-volume blood samples to unculturable bacterial communities.
6. *Time-resolved snapshots.* The extremely fast time scales of Raman scattering make a Raman spectrum an instant snapshot of a biological system where rapid chemical exchanges are always occurring. Relaxation effects on line shapes in nuclear magnetic resonance (NMR) are not readily observed in a Raman spectrum.¹⁰ Furthermore, coupling with ultra-fast instrumental techniques

allows studies of the dynamics of short-lived intermediates and reactions in the microsecond to femtosecond time domain¹¹.

Collectively, the ability easily to acquire label-free and comprehensive biochemical information at a subcellular and time-resolved resolution non-destructively and without the interference of water makes Raman microspectroscopy a unique and invaluable tool for *in vivo* biological investigations.

2.2.3 Enhancement of Raman cross-section

Having discussed the advantages of Raman microspectroscopy, the largest challenge remains to be the intrinsically weak Raman scattering as approximately only 1 in 10^7 photons experiences inelastic scattering¹². The improvement of the small Raman cross-section can be understood by looking at Equation 2.1 at the beginning of Section 2.1.1. In Equation 2.1 we see that the induced dipole can be enhanced with larger (i) polarisability α or/and (ii) electric field E . Consequently, prominent enhancement mechanisms address these two entities, including the two routinely used techniques, RR scattering and surface-enhanced Raman scattering (SERS).

Resonant Raman (RR) scattering

As mentioned in Section 2.1.2, RR scattering greatly improves the electric field E as the excitation wavelength falls within an electronic excitation of the molecule. The enhancement factor¹³ compared with the non-resonant Raman scattering can achieve up to 10^4 to 10^6 . In biological systems, selective enhancement of different chromophores can be achieved by varying the excitation source to be resonant with different electronic transitions.

Figure 2.3 shows an example of selective enhancement in RR spectroscopy. Guanosine-5'-monophosphate (GMP) is a component of nucleic acid. Exciting GMP at 532 nm generates a non-resonant Raman spectrum with a comprehensive picture of all molecular vibrations in the molecule. Excitation wavelengths at 244 nm and 218 nm, on the other hand, generate much simpler and signal-enhanced RR spectra, corresponding to two electronic transitions at two molecular sites of GMP. The 244 nm excitation

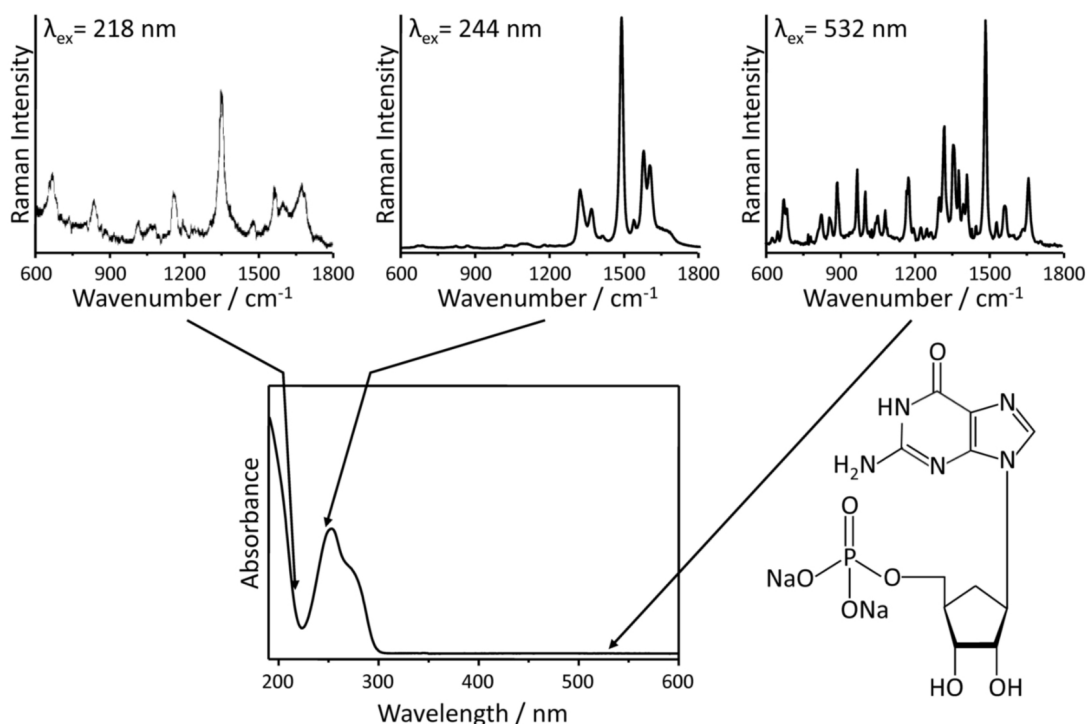


Figure 2.3: Selective enhancement of chromophores in RR spectroscopy (picture taken from reference [3])

corresponds to the π - π^* transition in the aromatic guanine ring and is applicable and beneficial to the study of all DNA/RNA bases and aromatic amino acids¹⁴.

The high specificity and selectivity in RR spectroscopy allow the study of subsystems in larger entities, from exploring an active group in one protein molecule to the selective observation of macromolecules such as proteins or DNA in a cellular system. One can also combine a RR spectrum with a spontaneous Raman spectrum by simply changing the excitation laser, thus obtaining rich vibrational information and selective probing substrates of interest at the same time.

Surface-enhanced Raman scattering (SERS)

The other mechanism, SERS, offers improvements in both E and α (Equation 2.2) and can achieve an enhancement factor¹⁵ on the order of $10^6 \sim 10^{14}$. The effect of SERS from pyridine adsorbed on electrochemically roughened silver was first observed by Fleischmann, Hendra and McQuillan at the University of Southampton in the UK in

1973¹⁶. In 1977, two independent groups each proposed a mechanism for the observed enhancement of signal. Jeanmaire and Richard Van Duyne proposed an electromagnetic effect¹⁷, while Albrecht and Creighton proposed a charge-transfer effect¹⁸. The electromagnetic theory proposes the excitation of localised surface plasmons, which was further supported by a research in 2015 on a more powerful extension of the SERS technique called SLIPSERS (Slippery Liquid-Infused Porous SERS)¹⁹.

The so-called localised surface plasmon resonance (LSPR) effect happens when strong localised electromagnetic fields are generated by the nanostructured metallic surfaces, e.g. silver, gold, and copper (enhanced E)²⁰. Enhancement in α can be understood as the chemisorption of the molecule on the metal substrate disturbing the electronic structure of the substrate²¹.

Applications of SERS have progressed from bulk samples to single molecules^{22,23} and single molecular events inside living cells²⁴. Despite its advantages of signal enhancement and fluorescence quenching²⁵, practical biomedical applications of SERS would have to address a few challenges because i) SERS Raman bands vary depending on wavelengths of incident lasers and SERS particle properties²⁶, ii) SERS involves invasive labelling with metallic nanoparticles which may change the intrinsic molecular behaviours, iii) SERS has a distance-dependent effect in the range of 0–4 nm so only chemical bonds close to the metallic surface can be detected²⁷ and iv) the interpretation and the reproducibility of SERS spectra from biological samples are often challenging.

Coherent Raman scattering (CRS)

Another approach is emerging in recent decades to obtain much stronger vibrational signals by employing coherent Raman scattering (CRS), which employs multiple light sources to produce coherent Raman signals²⁸. In contrast to the conventional linear and non-coherent spontaneous Raman scattering, CRS is non-linear and incoherent. The linearity of an optic signal is defined through its dependence on the intensity of the incident radiation, i.e. the linear optical signals scale linearly with the average power of the incident light while the non-linear signals exhibit a quadratic

or higher-order dependence on the intensity of the input radiation. Compared with incoherent spontaneous Raman where the scattered waves have random relative phase, in CRS the oscillators in the sample are correlated by the light fields, producing radiation from different locations in the sample with a well-defined phase relationship. The non-linearity and coherence of CRS make the produced signal greatly intensified and highly directional.

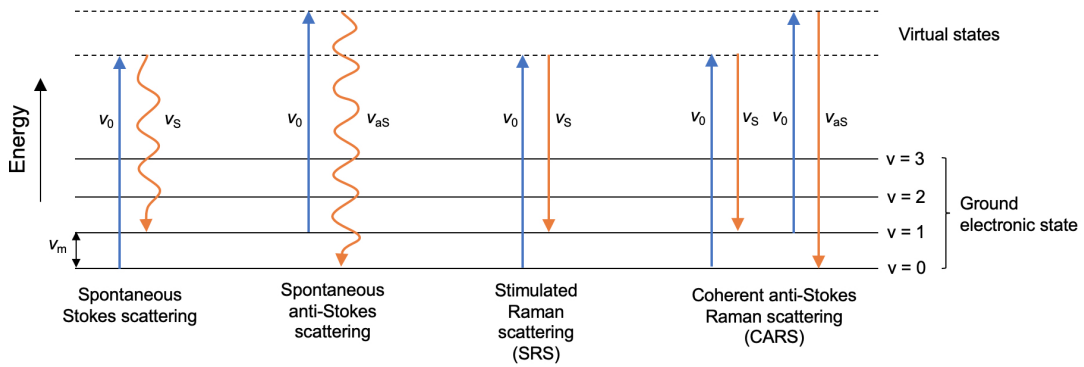


Figure 2.4: Schematic energy diagrams of CRS. The energy spacing between two vibrational levels equals to ν_m .

Coherent anti-Stokes Raman scattering (CARS) and stimulated Raman scattering (SRS) are two CRS processes²⁹. Figure 2.4 illustrates the energy-level diagrams of the CARS and SRS processes as compared with the spontaneous Raman scattering. During spontaneous Raman scattering, the pump beam with a frequency of ν_0 is incident upon the sample generating either a Stokes signal ν_S or an anti-Stokes signal ν_{aS} . In SRS, two laser beams at frequencies ν_0 and ν_S are directed onto the sample, such that the frequency difference $\nu_0 - \nu_S$ matches the frequency of a molecular vibrational mode ν_m . This process causes stimulated excitation of a chemically specific signature and brings the advantage of suppressing non-resonant background. Experimentally, either stimulated-Raman gain (SRG) of the Stokes or stimulated-Raman loss (SRL) of the pump is measured and optical modulation and demodulation are required to extract the proper SRS signal³⁰.

CARS involves a complex four-beam mixing process probing as shown in Figure 2.4. As in SRS, the resonance occurs when the difference between the pump and Stokes

beam $\nu_0 - \nu_S$ matches a molecular vibration in the sample. This resonance is then probed by the third field at ν_0 and generates an anti-Stokes signal at $2\nu_0 - \nu_S$. With respect to SRS, no complex demodulation schemes are required in CARS. However, the presence of a strong non-resonant background due to the four-wave mixing that does not carry any chemical information can distort and overwhelm the resonant signal³¹.

With the advantages of dramatically enhanced Raman signal and little autofluorescence interference, CARS and SRS microscopy are emerging techniques for real-time analysis and video-rate imaging of single cells in a living system with high imaging speed and 3D spatial resolution³². Nonetheless, both techniques are limited to the complex set-up of two synchronised trains of laser pulses. CARS and SRS are also limited by the small spectral range due to the narrow bandwidths between lasers, therefore so far in most cases, they have been demonstrated for chemical signatures with high concentrations and strong Raman signals, for example, the CH_2 stretching vibrations which are typically found in lipids in cells³³. In addition, CARS suffers from the complicated interpretation of Raman bands due to changed band frequencies³⁴. SRS, on the other hand, has comparable spectra as spontaneous Raman. Its signal also has an intensity linearly proportional to the concentration of the targeted molecules and is free of non-resonant background. SRS would be a complementary tool to spontaneous Raman scattering³⁵. The investigation of a biological system can begin with spontaneous Raman scattering (acquisition time from seconds to minutes) covering the entire spectral window of molecular vibrations and looking for a defined marker of interest. It then can be followed with SRS focusing on a predefined spectral window and offering high-speed benefit (acquisition time microseconds) for imaging and collection of a huge amount of spectral data.

2.3 Phenotyping bacteria by Raman microspectroscopy

In the past few decades, Raman microspectroscopy has been shown to be a powerful analytical tool in biological studies, including in the field of molecular biology, microbiology, ecology, biomedicine, live imaging of biological materials and disease

diagnosis. Having discussed the advantages and various techniques of Raman microspectroscopy, in this section we will focus specifically on the applications of Raman microspectroscopy in investigating bacteria.

2.3.1 A cultivation-free approach to study bacterial single cells

Bacteria are extremely versatile and abundant, and they play important roles in almost all aspects of modern life. Rapid and robust detection, characterisation and identification of microorganisms are unarguably of vital importance. To name a few examples, the industrial production of drinking water, food and pharmaceutical substances requires tight controls for bacterial contamination to prevent any health risks. Reliable characterisation and detection methods are also of great interest in environmental and agricultural science, either to utilise microorganisms in bioremediation to break down environmental pollutants or to detect plant pathogens that can cause ecological and economic damage. More examples and increasing attention can be found in the practice of antimicrobial stewardship in clinical fields as timely identification of a pathogenic infection can be lifesaving for patients³⁶.

Culture-based methods remain the mainstream standard for current bacterial identification. Bacteria are isolated from samples under investigation, grown on agar plates and analysed with chemical, optical and genotypic techniques. However, these methods introduce four major disadvantages in practice:

1. Biomass enrichment is often the most time-demanding step during sample preparation. In a clinical setting, it takes at least 24 hours for incubation of bacterial cultures from clinical specimens, which dependent on the bacterial species, can be extended to up to 7 days. An additional 24 hours is required for downstream characterisation for identification and susceptibility testing³⁷.
2. The majority of microorganisms remain unculturable using standard cultivation methods³⁸. In addition, many bacteria species are found to enter into a viable but non-culturable (VBNC) state under stress conditions³⁹.

3. Cultivation-based methods obtaining averages from bulk measurements can overlook substantial cell-to-cell heterogeneity. Bacterial subpopulations can exist due to stochastic processes and various physiological states⁴⁰.
4. The bacteria isolated from the sample matrices are not characterised *in situ*. This might produce a disturbed metabolism in bacteria away from their natural setting such as symbiotic bacteria in their host plants or pathogenic bacteria in blood⁴¹.

Here is what sets SCRM apart from other techniques. As discussed in the last section, Raman microspectroscopy makes it possible to measure single bacterial cells due to its high spatial resolution, thus providing a cultivation-free and label-free single-cell approach for phenotyping and profiling bacteria.

2.3.2 Identification of cellular components

Figure 2.5 shows the SCRS of *Escherichia coli* (*E. coli*) (Figure 2.5A), *Rhizobium* (Figure 2.5B) and *Rhodobacter* (Figure 2.5C). They are used to exemplify the bacterial Raman spectra which contain all kinds of biomolecules that serve as essential building blocks of life, such as proteins, nucleic acids, lipids, carbohydrates and species-specific compounds. Luckily, each of these biomolecules has distinct vibrational properties thus generating different Raman bands in a bacterial Raman spectrum.

A bacterial SCRS can be divided into three regions: the ‘fingerprint’ region which contains essential bio-information and can be seen as a fingerprint of a cell (400–1800 cm^{-1}); the ‘silent’ region which usually does not involve vibrational modes contributed by biomolecules formed of naturally occurring isotopes (1800–2700 cm^{-1}); the high-wavenumber region specifically contributed by stretching vibrations of CH, NH and OH groups (2700–3500 cm^{-1}).

The most prominent Raman band is usually found in the high-wavenumber region at 2800–3040 cm^{-1} with a centre around 2930 cm^{-1} . This broad band represents all C–H stretching vibrations that are present in all biomolecules, superimposing with each other (Figure 2.5A). The decomposition of the C-H band has been demonstrated

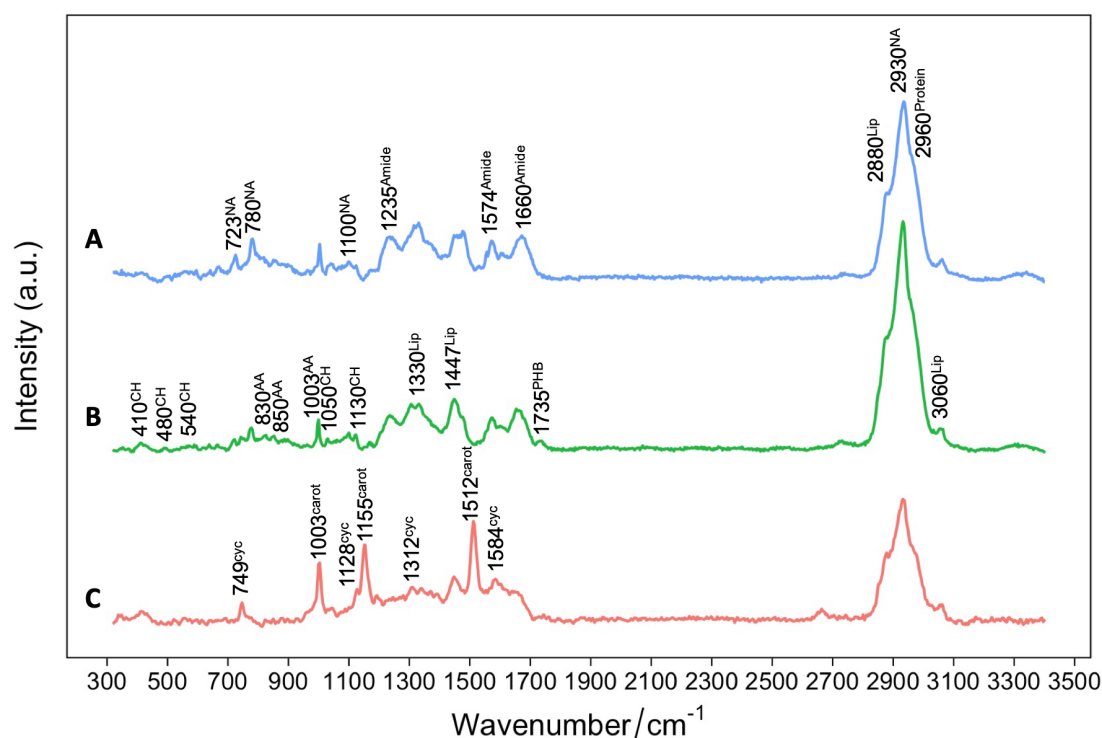


Figure 2.5: SCRS of (A) *E. coli*, (B) *Rhizobium* and (C) *Rhodobacter*. The assignments of the Raman band are labelled in superscripts at the wavenumbers. The abbreviations are listed as follows: ‘NA’ stands for nucleic acids; ‘CH’ stands for carbohydrates; ‘AA’ stands for aromatic amino acids; ‘Lip’ stands for lipids; ‘PHB’ stands for polyhydroxybutyrate; ‘cyc’ stands for cytochrome c; ‘carot’ stands for carotenoid.

to map the distributions of cellular lipids ($\sim 2880 \text{ cm}^{-1}$), proteins ($\sim 2930 \text{ cm}^{-1}$) and DNA ($\sim 2960 \text{ cm}^{-1}$), with a particular interest of using SRS microscopy for label-free imaging of mammalian cells^{42–45}.

Proteins

The most characteristic features for proteins can be represented by three Amide bands that are present in peptide bonds (CONH) linking consecutive amino acids in all proteins⁴⁶. In Figure 2.5A, Amide I band centres at 1660 cm^{-1} and contains $\sim 100\%$ C=O stretching. Amide II appears near 1574 cm^{-1} and is composed of 60% N–H and 40% C–N stretching. The Raman band of Amide III that centres around 1235 cm^{-1} contains 40% C–N stretching and 30% N–H bending. The band intensities and positions of the three Amide bands can provide valuable information on secondary structures of proteins, including the α -helix and the β -sheet structures⁴⁷.

In addition to the characterisation of the amide bonding in proteins, additional spectral features also appear in Raman spectra which allows descriptions of individual amino acids⁴⁶. Special interests have been directed to aromatic amino acids (Figure 2.5B), including phenylalanine at 1003 cm^{-1} (phenyl ring vibration), tyrosine doublet at 830 and 850 cm^{-1} and tryptophan doublet at 1340 and 1360 cm^{-1} . The narrowness of the bands of these individual amino acids illustrates that the vibrations, such as the phenyl ring vibration, are relatively stable in response to surrounding chemical environments.

Lipids

Lipids are the most abundant and key components of bacterial cell membranes. The most evident features of lipids in a Raman spectrum are related to the presence of the hydrocarbon chain in all lipids. In Figure 2.5B, two bands at 1447 and 1330 cm^{-1} can be assigned to the scissoring and twisting vibrations of the CH_2 and CH_3 groups, respectively⁴⁸. In the high-wavenumber region, the band centred at 3060 cm^{-1} is representative of unsaturated $=\text{CH}$ stretching which is present in unsaturated fatty acids. A ratio of total unsaturated fatty acids/total fatty acids (TUFA/TFA) is commonly calculated by the ratio of the integrated area of the band at 3060 cm^{-1} to the integrated area of the lipid CH band at 2880 cm^{-1} . The TUFA/TFA ratio is often used in Raman studies to characterise the profile of the lipidic profile of a cell⁴⁹.

Nucleic acids

Five canonical nucleobases: adenine, thymine, cytosine, guanine and uracil, can form base pairs and stack upon one another, which leads to long-chain helical structures in nucleic acids of either deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). Very strong Raman signals of the five bases can be observed by the ring-breathing modes in the region of $600\text{--}800\text{ cm}^{-1}$ ⁵⁰. For example, cytosine and uracil have vibrational modes of ring breathing at 780 cm^{-1} (Figure 2.5A). Adenine, thymine and guanine give rise to the bands at 723 (Figure 2.5A), 740 and 650 cm^{-1} , respectively. Therefore, the region of $600\text{--}800\text{ cm}^{-1}$ can be used to differentiate different nucleobases and characterise the number of nucleic acids in a cell. In addition to the

ring vibrations of the bases, the DNA and RNA backbone has a symmetric PO_2^- bond stretching and displays a signal at 1100 cm^{-1} (Figure 2.5A).

Carbohydrates

Carbohydrates function as the main source of energy, the major constituent of the cell envelope and molecular agents in cellular recognition in microorganisms⁵¹. They can be divided into three main subgroups: monosaccharides (often referred to as 'sugar'), oligosaccharides and polysaccharides. Vibrational spectra of carbohydrates are considerably less specific than spectra of proteins, lipids or nucleic acids because their complex functional groups contribute to multiple bands spread in the whole fingerprint and high-wavenumber regions⁵².

The O–H stretching is often found at the far end in the high-wavenumber region ($3300\text{--}3500\text{ cm}^{-1}$), however, it can overlap with the O–H stretching present in water. Other fingerprint regions that can be attributed to carbohydrates include $300\text{--}800\text{ cm}^{-1}$ (deformational modes of CCC/CCO/OCO), $800\text{--}1200\text{ cm}^{-1}$ (stretching vibrations of C–O/C–C) and $1200\text{--}1500\text{ cm}^{-1}$ (deformational modes of CH/CH₂).

For example, for a monosaccharide, glucose, some characteristic bands appear around 540 and 410 cm^{-1} (exocyclic and endocyclic ring deformation, respectively) in Figure 2.5B. Glycogen is a polysaccharide which is a readily available glucose reserve accumulated by microorganisms to supply carbon and energy under starvation conditions⁵³. A band is specifically observed at $\sim 480\text{ cm}^{-1}$ (Figure 2.5B) in the spectra of glycogen which can be assigned to C1–O–C4 glycosidic bond (straight chains). Other COC vibrations of a glycosidic bond can be found at 850 and 940 cm^{-1} . Other dominant features include bands at 1050 cm^{-1} (CO and CC stretching) and 1130 cm^{-1} (CC, CO and COC stretching in the glycosidic link and ring-breathing vibration), as marked in Figure 2.5B.

Specific biomarkers

In addition to cell contents that are common for all bacteria such as proteins, lipids, DNA/RNA and carbohydrates, bacteria also exhibit many biocomponents that are

more or less specific to certain genera, species, or even strains. An additional ester-orientated band centred around 1735 cm^{-1} appears in some bacterial spectra, for example, in SCRS of *Rhizobium* in Figure 2.5B. This band is representative of C=O stretching of the ester functional group in storage molecules like polyhydroxybutyrate (PHB) and can appear in bacteria like rhizobia who synthesise and store PHB under temporary nutrient-limited conditions⁵⁴.

Another class of biomolecules that has been studied heavily is heme proteins, which have a diverse range of structures and biological functions in many life processes. Heme proteins can easily be studied with the aid of resonance Raman (RR) scattering using a laser which has an excitation wavelength coinciding with the porphyrin ring's absorption band. Cytochrome *c* is a heme protein involved in electron transfer chain located in the membrane of the mitochondrion. Using a 532-nm laser, cytochrome *c* can be observed at 749 (porphyrin ring breathing), 1128 (C–N stretching), 1312 (C–H deformation) and 1584 cm^{-1} (C–C stretching), shown in the SCRS of *Rhodobacter* (Figure 2.5C).

Production of carotenoids in cyanobacteria also gives strong RR signal with an excitation wavelength around 532 nm. Carotenoid generates three sharp bands at around 1003, 1155 and 1512 cm^{-1} (Figure 2.5C). They can be assigned to the C=C stretching, C–C stretching and C–CH₃ deformation in carotenoids, respectively. Minor shifts occur at these three positions depending on the structure of the carotenoid, therefore, structural information of different carotenoids can also be obtained^{55,56}.

2.3.3 Identification of bacteria

Having understood how to interpret the bacterial spectrum by identifying different cellular contents and biomarkers, we now look at the complex spectrum as a whole. An SCRS can be seen as a fingerprint of a whole cell. While the identification of biomarkers can be used to obtain biochemical information and semi-quantify the intracellular concentrations of important cellular components, the analysis of SCRS to discriminate bacteria at a species or subspecies level is of increasing interest⁵⁷.

Identification of clinical pathogens

Fast and reliable detection and identification of bacteria are unquestionably crucial in the field of medicine to bring benefits for a fast and precise diagnosis and treatment. Raman spectral fingerprints combined with statistical analysis have been demonstrated to identify clinically relevant bacteria under defined laboratory conditions⁵⁸, including in the study of *E. coli*, *Klebsiella pneumoniae*, *Proteus mirabilis*, *Pseudomonas aeruginosa*, *Enterobacter cloacae*, *Staphylococcus aureus*, *Staphylococcus epidermidis*, *Salmonella*, *Chlamydiae*, *Enterococcus*, mycobacteria and many more^{59–69}.

Raman-based evaluation can also be directly applied to the analysis of body fluids for the identification of pathogenic bacteria present in patients⁵⁸. Harz et al. have applied SCRM to detect bacteria in the cerebrospinal fluid of patients with bacterial meningitis via hierarchical cluster analysis (HCA)⁷⁰. Kloss et al. have built a Raman database containing more than 10000 SCRS for the identification of 34 bacterial strains out of 13 species in ascitic fluids⁷¹. The same group also built up a database of 11 important pathogens in urinary tract infections⁷². An accuracy of 92% on a species level was achieved via a support vector machine (SVM). Identification of bacteria in the sputum sample using SCRS was also achieved on the basis of a principal component analysis (PCA) followed by linear discriminant analysis (LDA)⁷³. SERS also enabled rapid identification of pathogens in human blood which usually has a lower bacterial count compared with other body fluids^{61,74}.

The effect of antibiotics–bacteria interactions is also an important point in a clinical approach. Munchbert et al. investigated how antibiotic treatment of bacteria influenced the bacterial SCRS⁶⁶. Before the treatment, an LDA model was built to identify two strains of *E. coli* and two species of *Pseudomonas* spp. Upon treatment of four antibiotics, namely, ampicillin, ciprofloxacin, gentamicin and sulfamethoxazole, the model succeeded at bacterial identification with accuracies of 85.4%, 95.3%, 89.9%, and 97.3%, respectively. In addition to the successful identification of bacteria treated with antibiotics, a study by Moritz et al. used laser tweezers Raman spectroscopy to characterise the metabolic states and the spectral changes associated with the cellular response of *E. coli* cells to cefazolin⁷⁵. A PCA demonstrated an accurate identification

of different metabolic states and suggested a great potential of Raman biomarkers towards the identification of cell responses to antibiotic drug treatment.

Identification of food- and water-borne pathogens

The presence of pathogens in food can also be examined with SCRM, which has the potential for routine uses in a food safety test. Meisel et al. have constructed a database of 4622 SCRS and a three-level classification tree based on SVM successfully discriminated 19 meat-associated bacterial species⁷⁶. The same group also illustrated the applicability of SCRM to identify food-borne pathogens in spiked milk^{77,78}. Assaf et al. have applied Raman spectroscopy to identify *Salmonella* spp. from rice, oat, wheat, maize, chicken and pork⁷⁹. This study also suggested a standardised protocol of using Raman-based control systems in the food-production field.

In the field of guarding water quality, several proof-of-concept studies using Raman spectroscopy can be found. Raman identification of *Pseudomonas* strains, *E. coli* and coliform strains was successfully performed in drinking water^{80,81}. In a study by Kusic et al. cultivation-free identification of *Legionella* species in the VBNC state was achieved by Raman microspectroscopy combined with multiclass SVMs⁸². Raman spectra of 22 strains of the water pathogen *Legionella* were distinguished from those of other common aquatic bacteria *E. coli*, *Klebsiella pneumoniae* and *Pseudomonas aeruginosa*.

Identification of environmental bacteria

While detection and identification of microorganisms are key in medicine and the food industry, one should also realise the amazingly diverse distribution and importance of bacteria in environments of soil, ocean, air, as well as the plants and animals that inhabit these domains and artificial environments such as bioreactors or microbial fuel cells.

Productions of chlorophyll, carotenoids and other pigments serve as survival strategies in microorganisms such as halophilic heterotrophic bacteria, cyanobacteria and algae to cope with hazardous environments with extreme temperature, pressure,

pH, salinity, radiation insolation and restricted nutrient and oxygen availability⁸³. Detection of those photosynthetic or protective pigments is key in understanding terrestrial extremophilic strategies and searching for extinct or extra-terrestrial life. Because these pigments give strong resonant RR signals under certain excitations, they are recognised as specific biomarkers in Raman spectra and give unique advantages of applying Raman spectroscopy in *in situ* geological and astrobiological studies of bacteria.

For example, Raman spectra of different carotenoids from representative species of archaea, bacteria, cyanobacteria and algae were obtained with excitation by a 514.5-nm laser⁸⁴, by targeting the three carotenoid bands at 1003, 1155 and 1512 cm^{-1} . A review of Raman spectroscopic investigations in halophiles also summarised the studying of microbial carotenoids and carotenoid-like pigments in evaporites, salt crystals and evaporitic crusts⁸⁵.

Raman detection of chlorophylls including bacteriochlorophylls has also been applied in the study of photosystems in photosynthetic bacteria⁸⁶. An 1064-nm laser excitation suppresses the accompanying fluorescence⁸⁷ and enables the observation of key signatures of the aromatic pyrrole rings coordinated with the central magnesium(II) ion in the region of 400 to 1,600 cm^{-1} ⁸⁸. These characteristic bands have been adopted to identify the presence of microbial chlorophylls in biogeological specimens^{89,90}.

2.3.4 Computational strategies for the analysis of Raman spectra

Spectroscopic studies rapidly generate large datasets that require computational processing to derive useful biochemical information. Because of the high number of features (i.e. Raman wavenumbers) present in a Raman spectrum, the analysis of a spectrum is often challenging and requires careful selection and performance of a computational analysis. Depending on either exploratory purposes or diagnostic/-classification purposes, applying different techniques in Raman analysis is possible to extract informative images, patterns and biomarkers, as well as to classify the samples based on their entire spectra.

Feature extraction techniques

Feature extraction techniques are responsible for extracting a smaller number of variables that are more informative out of a pool of original wavenumber variables. They can either reconstruct the data into smaller dataset thus aiding the visualisation or help to select a few important biomarkers.

Dimension-reducing techniques such as PCA and partial least squares (PLS) are often desirable in the study of multivariate datasets like Raman spectra. PCA is often the first analysis performed by researchers on a new Raman dataset⁹¹. It is a powerful tool to reduce the high dimensionality and uncover hidden relationships⁹². PCA reconstructs a new co-ordinate system with the new variables as the principal components (PCs) and each PC is a linear combination of the original variables, i.e. the wavenumbers and intensities measured in a Raman spectrum. This new co-ordinate system is better at conveying the information present in the data than the original measurement variables because it filters background noise and retains only the most important features. In the case of a Raman spectrum that contains a lot of correlated biological information, only a few PCs are enough to explain most of the spectral variance.

LDA is another linear transformation technique that is often used by spectroscopists⁹³. Unlike unsupervised PCA, LDA is a supervised approach (i.e. classes are taken into account prior to the analysis). As in any supervised method, LDA can overfit small datasets and having the number of spectra at least five times bigger than the number of variables can potentially avoid this problem⁹⁴. Using a PCA at a prior stage is one way to achieve segregation with a smaller dataset, i.e. PCA-LDA, but this does not guarantee that the scores are unbiased⁹¹. Techniques like PCA, LDA and PCA-LDA allow one to visualise the dataset in a reduced-dimensional space and extract the features that convey most of the spectral information.

Except for PCA and LDA as feature reconstruction techniques, some more specific feature selection techniques are useful to extrapolate existing features from the dataset, such as specific wavenumbers that can be identified as biomarkers. Approaches based on successive projections algorithm and genetic algorithm have proven to be

useful to remove redundant variables and only include informative variables in the resultant model⁹⁵.

Classification techniques

In Raman approaches, besides feature extraction to reduce and visualise the original complicated dataset and to find informative biomarkers, classification based on samples' spectra is often desirable for diagnostic and practical purposes.

Unsupervised classification usually relies on clustering methods that are designed to cluster the spectra into groups, based on the similarity between spectra. Fuzzy *c*-means clustering, *k*-means clustering and HCA are the three most common options⁹⁶. Although initially time-consuming, once computed, an HCA results in a hierarchy of clusters in which the number of clusters can be selected without recalculation and often gives the best correlations among the three clustering techniques⁹¹.

For diagnostic purposes, supervised classification models are more desirable, and they are used in most of the studies mentioned in Section 2.3.3. In a supervised classification model, spectra are often divided into two sets: the training set and the test set. In the training set, each spectrum has a user-defined class label and it will be used to train one or a combination of several classifiers. The trained classifier is then able to categorise each individual spectrum in the unlabelled train set. It is also a common approach to use an independent dataset to validate the model's performance. The sample size is highly influential to the performance and validity of a model⁹⁷. It has been shown that 75–100 samples will usually be needed to train a classifier in spectroscopic studies⁹⁸. In a smaller dataset, cross-validation approaches to resample and iterate many training and test sets may be suitable for effectively using as much of the dataset⁹⁹.

Next, I will introduce some of the most common machine learning algorithms for supervised classification. Many of the techniques and the power of them will be further exploited and compared on a real-world Raman dataset in Chapter 6.

Linear classifiers in for example linear discriminant classification (LDC), which is a classifier version of LDA, are quick and easy to fit but may not be sophisticated enough to accurately induce the boundaries between non-linear data classes.

A common and conceptually simple approach is k-nearest neighbours (kNN)¹⁰⁰. Because of its simplicity and flexibility, it is widely applied in genetic and spectroscopic studies^{101–105}. In a kNN model, an undefined sample is assigned by the label that is most frequent among the k (a user-defined number) samples nearest to that query sample, based on Euclidean distance. It suffers significantly from the high dimensionality of data and often requires variable reduction prior to classification⁹¹.

Random Forest (RF)¹⁰⁶ is a popular choice in the area of bioinformatics such as genetic analysis, metabolomics, medical image analysis as well as Raman spectroscopic analysis^{107–111} owing to its unique advantages in handling small sample size, high-dimensional feature space and noisy dataset¹¹². RF is an ensemble learning method that operates by constructing multiple Decision Trees. One decision tree is a tree-like model in which each decision node branches and splits the data according to certain parameters and each leaf presents a decision. An RF constructs hundreds and thousands of Decision Trees and then predicts the outcome of the classification by majority votes. Like Decision Trees, RF naturally incorporates variable screening and feature selection during the process. While retaining many benefits of Decision Trees, RF improves in terms of stability and accuracy by random sampling and bagging. The main disadvantage of RFs is their complexity, therefore more time and computational resources are required to train RF models.

SVM is another class of classifiers that is increasingly popular and shown to be powerful in bioinformatics. It can be applied to DNA and protein sequences^{113–116} as well as spectroscopic data^{117–119}. The SVM technique was first introduced by Vapnik and his group in 1995¹²⁰. The original idea is to use a linear hyperplane to separate the data in a high-dimensional space. For problems which cannot be linearly separated, SVMs with non-linear transformations (kernels) are introduced. In the transformed feature space it is more possible to find an optimal separating hyperplane. Building an

SVM requires careful tuning of parameters deciding the hyperplane, such as soft/hard margin, the choice of the kernel and misclassification tolerance.

Artificial neural networks (ANNs)¹²¹ are now commonly applied in pattern recognition and their effectiveness in bioinformatics has been explored. This is due to their ability to cope with the highly complex dataset by simulating layers of biological neurons¹²². However, due to the complexity of their designing, ANNs are easy to overfit.

The choice of a machine learning algorithm and the training of the model are vital in the task of classification based on Raman spectra. By employing many powerful machine learning techniques and constantly building and updating the databases, it is possible to classify bacteria with different genera, species or even strains based on their Raman spectra with very complex information and often minute differences.

2.4 You are what you eat—stable isotope probing of bacteria

Bacteria have evolved a remarkably versatile spectrum of metabolic capabilities. They are constantly changing and adapting according to different environmental conditions. Interactions between microbes and their environment have been increasingly studied using stable isotope probing (SIP)¹²³. SIP involves exposing microbes to isotopically labelled substrates that are consequently assimilated into microbial biomass by the microorganisms that are actively involved in specific metabolic processes. Molecular analysis of the labelled biomarkers thereby reveals the functional information about the bacteria responsible for the metabolism of a particular substrate. Using isotopes that mimic their naturally abundant counterparts provides a direct link between microbial identities and their functions without altering the natural substrate pool.

While modalities like isotope ratio mass spectrometry and nuclear magnetic resonance (NMR) spectroscopy require at least 300,000 bacterial cells, Raman microspectroscopy or secondary ion mass spectrometry (SIMS) coupled with SIP enables an analysis of microbial functions at a single-cell level^{124,125}. However, SIMS is limited by its destructive and expensive features. Raman microspectroscopy exceeds other

techniques in its non-destructive, sensitive and relatively cheap nature. While a Raman spectrum already conveys intrinsic, phenotypic information, more insight into bacteria functions can be obtained by combining Raman microspectroscopy with SIP using ^2H , ^{13}C and ^{15}N -containing substrates¹²⁶.

2.4.1 Principles of Raman-SIP

Because Raman spectroscopy considers changes in vibrational modes of molecules, a mechanical model can be used to consider a molecular vibration as a classical spring model. A classical ‘two-balls-on-a-spring’ approach provides an equivalence of a diatomic molecule with two atoms with masses m_1 and m_2 connected by a chemical bond. The vibrational frequency ν of such a molecule is described as:

$$\nu = \frac{1}{2\pi c} \sqrt{\frac{k}{\mu}} \quad (2.11)$$

where c is the speed of light, k is the force constant of the spring (diatomic bond) and μ is the reduced mass given by:

$$\mu = \frac{m_1 m_2}{m_1 + m_2} \quad (2.12)$$

Therefore, the vibrational frequency ν is inversely proportional to the square root of the reduced mass μ . When an atom is replaced by its heavier isotope, μ increases, therefore, ν decreases. Taking C–D replacement of C–H as an example, if the ^1H atom is replaced by the ^2D atom, μ almost doubles and the ratio of the new ν to the old ν will be 1.36. Therefore, C–D vibrations usually appear in the silent zone of a Raman spectrum at around $2100\text{--}2200\text{ cm}^{-1}$, red-shifted from the C–H vibrations at $2800\text{--}3000\text{ cm}^{-1}$. Thus, a strategy to quantify D incorporation is to calculate the percentage of the integrated C–D band area to the sum of the integrated areas of the C–D and C–H bands:

$$D_{\text{incorp}} = \frac{A_{\text{C-D}}}{A_{\text{C-D+C-H}}} \quad (2.13)$$

Redshifts of Raman bands can also be observed when replacing ^{12}C , ^{14}N , ^{18}O or ^{32}S with their heavier analogues, although ^2D , ^{13}C and ^{15}N are mostly used in Raman-SIP studies for functional studies of microbes¹²⁶. Compared with the C–D shifts, the

extent of the red shift is considerably less dramatic in ^{13}C and ^{15}N substitutions due to the comparably low changes in μ .

By feeding microbes with isotope-labelled substrates, active microbes metabolise these substrates and the isotopic-dependent shifts in Raman spectra can be used as indicators of isotopic incorporation, thereby ‘you are what you eat’. In addition, Raman-SIP can be directly combined with fluorescence *in situ* hybridisation (FISH) to connect the identities and functions of metabolically active members in a microbial community at the single-cell level¹²⁶.

2.4.2 ^{13}C -labelling in Raman-SIP

^{13}C - and ^{15}N -labelled substrates are most widely applied in Raman-SIP to show that bacteria specifically harness the labelled substrates as a carbon/nitrogen source³⁶.

The first Raman-SIP shifts were observed by Huang et al. in 2004 when feeding *Pseudomonas fluorescens* with ^{13}C -glucose as the sole carbon source¹²⁷. The most pronounced band shift occurred from 1001 to 966 cm^{-1} , which can be assigned to the symmetric benzene ring breathing of phenylalanine. This study was followed by many other Raman-SIP studies using ^{13}C -labelled carbon sources. For example, by using a mixture of ^{12}C - and ^{13}C -glucose, Li et al. observed the Raman shifts of phenylalanine and thymine in the carbon flow from bacterial prey to its predators¹²⁸. In a more recent study, the SCRS of *Deltaproteobacterium* grown with ^{13}C -naphthalene exhibited four peaks at 1001, 909, 978 and 968 cm^{-1} , originating from isotopologues of phenylalanine with 0, 2, 4 or 6 ^{13}C atoms in the aromatic ring of phenylalanine, respectively¹²⁹.

Phenylalanine serves as a good indicator for observing Raman shifts, not only because of the prominent shifts as a result of D substitutions on the aromatic ring but also as the phenylalanine shifts indicate a *de novo* synthesis of aromatic amino acids from the labelled substrates. This was illustrated in a study by Eichorst et al., in which no shifts were detected by Raman microspectroscopy in soil bacteria fed with ^{13}C cellulose¹³⁰. Although the soil bacteria actively metabolised cellulose, they fulfilled their need for amino acids by directly taking up from the soil environment instead of synthesising them from the carbon source. This study also suggests the

importance to keep a clean sample background in the applicability of Raman-SIP in the studies of environmental samples.

In the case of unobservable phenylalanine shifts, other ^{13}C induced shifts should be used (Figure 2.6A) such as in Raman bands of Amide I (~ 1662 to $\sim 1621\text{ cm}^{-1}$), Amide II (~ 1577 to $\sim 1567\text{ cm}^{-1}$), CH_3 stretching (~ 1340 to $\sim 1313\text{ cm}^{-1}$), Amide III (~ 1247 to $\sim 1232\text{ cm}^{-1}$) and ring-breathing mode of cytosine and uracil (~ 780 to $\sim 765\text{ cm}^{-1}$)^{126,127,131,132}.

Raman-SIP studies with red-shifted Raman bands of bacterial chromophores have emerged recently, utilising the RR effect that has been discussed in Section 2.2.3 and 2.3.3. Li et al. showed that fixation of $^{13}\text{CO}_2$ into carotenoids by cyanobacteria produced redshifts at all three carotenoid positions¹³³. In this single-cell Raman-SIP approach, single cells capable of carbon fixation were identified from complex samples such as mixed cultures and natural seawater. *Geobacter metallireducens* is an iron-reducer capable of degrading aromatic pollutants and accumulating a large amount of *c* type cytochromes. In another study, SCRS of *Geobacter metallireducens* fed with ^{13}C -acetate were obtained to study the assimilation of ^{13}C from carbon sources to cytochrome *c*¹²⁹. No redshift was observed at 748 cm^{-1} , indicating that no or only a few ^{13}C atoms were involved in the porphyrin breathing mode. However, the peaks at 1128 , 1311 and 1583 cm^{-1} all showed clear redshifts of about 5 cm^{-1} to 1123 , 1307 and 1577 cm^{-1} , respectively.

2.4.3 ^{15}N -labelling in Raman-SIP

^{15}N -labelled substrates are rarely used in Raman-SIP studies. Because of a larger atomic mass, the ^{15}N -induced Raman shifts are smaller than those caused by ^{13}C -substitutions (Equations 2.11 and 2.12) and can be masked by other Raman bands.

Most frequently observed redshifts are caused by an incorporation of ^{15}N into nucleic acids^{134,135}. In a study by Muhamadali et al., SCRS of *E. coli* cells grown with ^{15}N ammonium chloride exhibited a shift at the cytosine and uracil band, moving from 781 to 778 cm^{-1} . It should be noted that a combination of ^{13}C - and ^{15}N -labelling was used in this research. Quantitative differentiation of bacterial cells incubated with

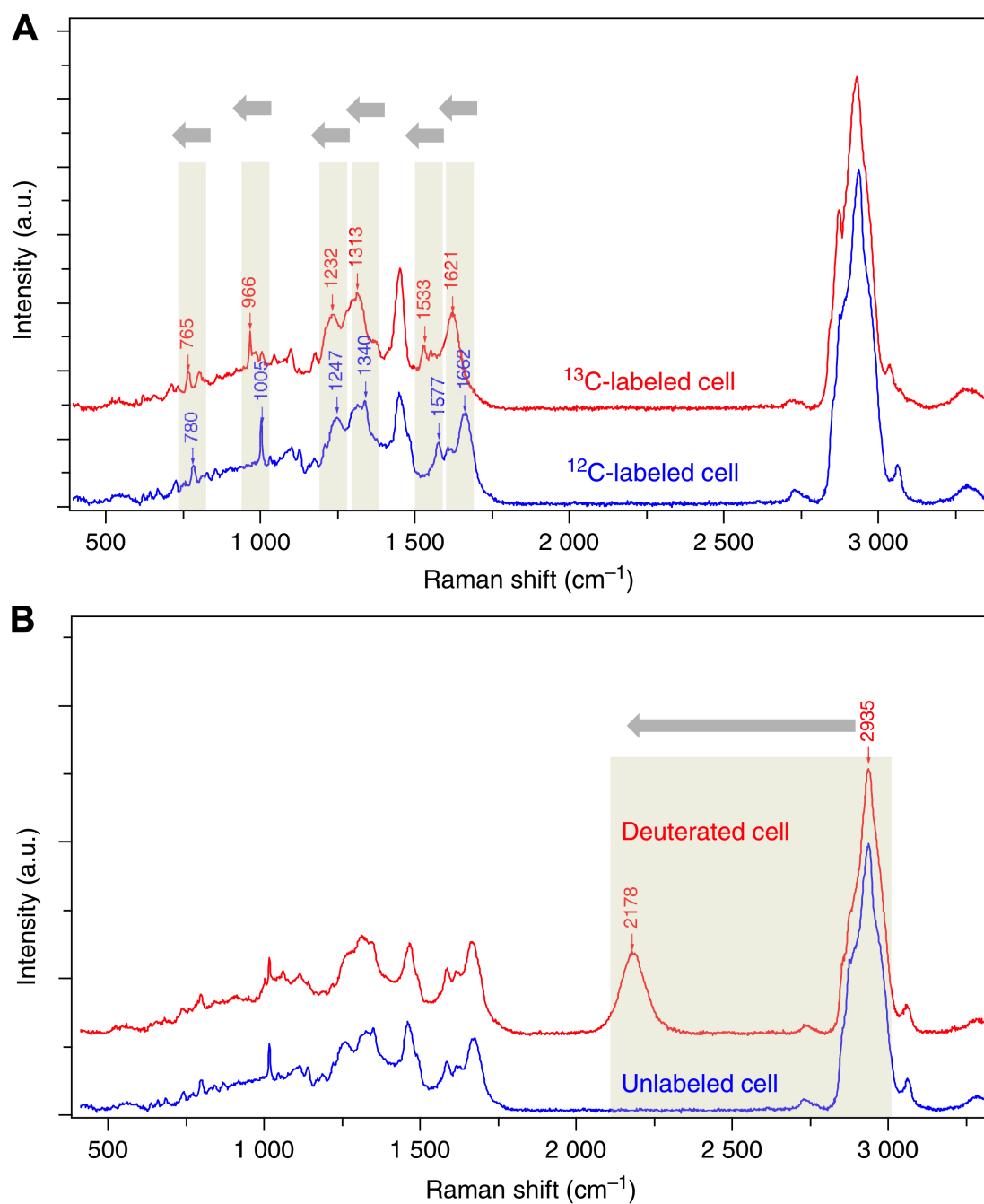


Figure 2.6: Raman spectra of *E. coli* with highlighted bands caused by incorporation of (A) ^{13}C or ^{12}C and (B) D or H into the biomass, modified from reference [126].

different ratios of $^{13}\text{C}/^{12}\text{C}$ and $^{15}\text{N}/^{14}\text{N}$ was also enabled at the single-cell level by employing PCA- and PLS-based clustering analysis¹³⁶. Two more recent ^{15}N -Raman-SIP studies utilised the SERS effect and observed different SERS spectra of bacteria with ^{15}N assimilation^{137,138}. Cui et al. discovered that both ^{15}N -labelled single cells and colonies displayed strong shifts at various positions (653, 730, 793, 957, 1216 and 1325 cm^{-1}), mostly associated with ring-breathing modes of nucleic acids and C–N stretching modes in proteins¹³⁷.

2.4.4 D-labelling in Raman-SIP: Raman-DIP

Raman-DIP with D-labelled substrates to monitor specific metabolic features

Similar to the ^{13}C and ^{15}N approaches, D-labelled substrates can be given to microbes and the *de novo* synthesis of particular biomolecules can be tracked by observing corresponding shifts in the Raman spectra. It should be noted that D-labelled substrates are often used simultaneously and complementarily with ^{13}C -labelling in Raman-SIP experiments. *Geobacter metallireducens* grown with D-acetate displayed a Raman spectrum with isotopic phenylalanine shifts¹²⁹. Similar shifts in phenylalanine were also observed in another study using D-labelled glucose as sole carbon source¹³². Compared with the diverse Raman-SIP studies utilising ^{13}C -labelling, Raman-DIP with D-labelled substrates to study metabolic pathways of bacteria is still new and under-explored. In Chapter 4, one of the aspects in Raman-DIP using deuterated carbon source to link bacterial single cells to their carbon metabolism will be discussed and applied.

Different deuterated substrates have also been used to monitor the metabolic features in eukaryotic cells. Manen et al. reported the detection of isotope-labelled proteins using D-labelled phenylalanine, tyrosine and methionine in single HeLa cells¹³⁹. A similar approach with SRS microscopy was used by Wei et al.¹⁴⁰. Through the incorporation of D-labelled amino acids, they imaged the newly synthesised proteins in live HeLa and HEK293T cells. Moreover, SRS microscopy with higher spatial and temporal resolution could readily generate spatial maps of the quantitative ratio between newly synthesised and total proteomes. Stiebing et al. visualised lipid

uptake and storage within single macrophages using deuterated fatty acids¹⁴¹. Li et al. used glucose-d₇ as a tracer to visualise the *de novo* lipogenesis in cancer cells by employing the SRS effect¹⁴².

Raman-DIP with D₂O to probe general metabolic activity

Heavy water (D₂O) has been suggested as a unique and universal tool to monitor the synthesis of biomolecules on a global scale^{143,144}. As an isotopologue of water, D₂O can rapidly and freely equilibrate with the total body water inside a cell and D atoms can exchange with the H atoms to form a variety of X–D bonds. Different from the often non-enzymatic reactions to form heteroatomic X–D bonds such as O–D, N–D and S–D, C–D bond formation depends solely on enzyme-catalysed chemical reactions due to the stronger C–H bond strength. In C–H/C–D exchanges, D atoms from D₂O are rapidly incorporated into metabolic precursors of different classes of biomolecules such as deoxyriboses, acetyl-CoA/NADPH, amino acids and phosphoenolpyruvate (PEP) (Figure 2.7). These precursors with D labelling are then slowly incorporated into their final products, which are nucleic acids, lipids, proteins and carbohydrates (Figure 2.7). For example, the reduction reactions to synthesise lipids are through the action of NADP⁺/NADPH, and D is incorporated due to free exchange with the hydrogen ion (Figure 2.8). The rate-determining step of the biomolecules' synthesis is governed by cellular metabolic activities. Therefore, D₂O serves as an ideal agent to probe general metabolic activities through the emergence of a variety of D-labelled macromolecules.

Raman-DIP using D₂O has been demonstrated to measure general metabolic activities of microbial single cells, first reported by Berry et al¹⁴⁵. By incubating microbes with D₂O, physiologically active cells can be rapidly and sensitively identified with Raman by measuring the C–D peak at 2170–2300 cm^{−1} which is the silent region with no presence of other biological bands (Figure 2.6B). The metabolic activity can be semi-quantified and compared by measuring the degree of D incorporation (Equation 2.13). In fast-growing *E. coli* cells, D detection was already possible after 20 min¹⁴⁵.

Furthermore, Raman-DIP via D₂O incubation can be combined with FISH for the identification of active microbes¹⁴⁶. By labelling the mouse caecum microbiota with

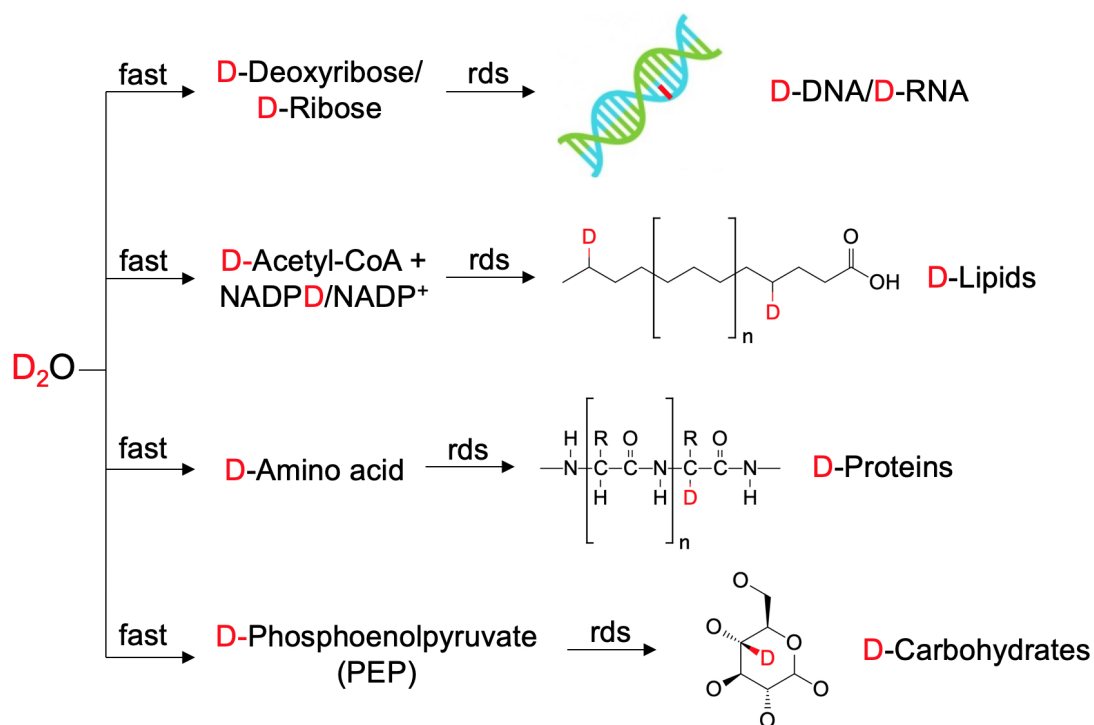


Figure 2.7: D_2O as a unique and universal tracer for different biomolecules. 'rds' stands for rate-determining step.

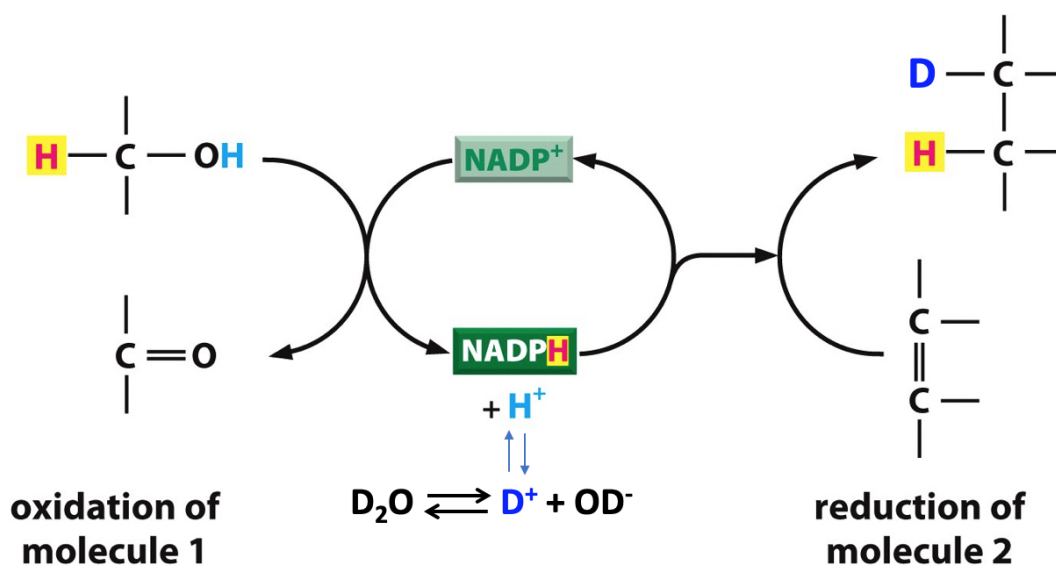


Figure 2.8: D incorporation during the reduction reactions to synthesise biomolecules via the action of $NADP^+$ / $NADPH$.

FISH probes and monitoring the activity of single cells after incubation with various substrates, it was revealed that *Akkermansia muciniphila* and *Bacteroides acidifaciens* exhibited distinguishable response patterns to the addition of mucin and sugars¹⁴⁵. This study also introduced the possibility to pinpoint microorganisms metabolising any substrate of interest: by adding D₂O and any unlabelled substrate together, only microbes that are active and stimulated by the added substrate will show elevated C–D signals compared with the non-active members in the community. In a study by Eichorst et al., Raman-DIP using D₂O was first applied to investigate the ratio of active microbial cells in the soil as well as in soil with supplemented carbon¹³⁰. In a recent study, Tao et al. applied single-cell Raman-DIP in antimicrobial stewardship to determine the minimum inhibitory concentration based on metabolic activity of bacteria under different antibiotic treatments¹⁴⁷. The same approach was applied to *in situ* identify the antimicrobial-resistant bacteria in the Thames River within 24 hours after dosing carbenicillin or/and kanamycin¹⁴⁸. In Chapter 5, the application of Raman-DIP via D₂O incubation will be further investigated in the intracellular pathogen *Salmonella*.

Advantages of Raman-DIP in metabolic studies

SCRM-DIP has advantages over other analytical technologies of SIP in terms of its non-destructive nature and detection at the single-cell level. Its non-destructive nature offers possibilities for downstream single-cell sorting and omics (e.g. single-cell genomics) analysis^{126,149}. When combined with metabolic studies, it has the potential to provide a culture-independent approach to probe metabolic activity at the single-cell level in complex microbial communities.

In SIP practice, ¹³C-labelled or ¹⁵N-labelled derivatives of substrates are generally expensive and many of them are not commercially available, which limits their use in Raman-SIP studies^{126,143}. This problem can be overcome by using much less expensive D-labelled substrates of either cheap D₂O or deuterated compounds synthesised more efficiently and more cost-effectively by H/D exchange reactions in D₂O environment¹⁵⁰.

The relatively low cost and high versatility of D-labelled substrates allow wider applicability to make a broad range of complex deuterated substances, including natural products, medicines and deuterated mixtures containing multiple chemicals. Raman-DIP with these D-labelled substrates is therefore invaluable in the functional studies of microbial cells in their natural habitats at the single-cell level.

2.5 Outlook

This chapter has provided a basic outline of the principles of Raman microspectroscopy and Raman-DIP, with an emphasis on their applications in phenotyping bacteria. SCRM will be put to a significant application in the next chapter to explore the phenotypes of a plant bacterium in symbiosis with its host plant. SCRM combined with Raman-DIP, thereby SCRM-DIP will be explored further and put to specific applications in Chapters 4 and 5.

Chapter 3

Application of SCRM in phenotyping symbiotic plant bacteria

Contents

3.1 Introduction	41
3.2 Materials and Methods	45
3.3 Results and Discussion	47
3.3.1 SCRS distinguish cells in different physiological states	47
3.3.2 SCRS link phenotypic changes with gene knockouts	52
3.3.3 Single-cell resolution allows heterogeneity analysis within a single bacterial population	58
3.4 Conclusions	63

In the last chapter of background, I have repeated that Raman microspectroscopy has unique advantages in phenotyping bacteria and Raman spectra can be used to identify a variety of cellular components. In this chapter, SCRM will be employed to an uncharted area of investigating genetic mutants of a symbiotic plant bacterium *Rhizobium*. I will demonstrate that SCRS can be regarded as biochemical fingerprints of *Rhizobium* cells, consisting of a collection of biomolecules that can be used to identify and characterise bacteria at a single-cell level. SCRS will be used to differentiate different physiological states of bacteria, link phenotypic changes with specific genetic mutations, and assess heterogeneity within a single bacterial population.

3.1 Introduction

Nitrogen fixation by *Rhizobium*-legume symbiosis provides a significant fraction of the available nitrogen in the biosphere, making it agronomically and ecologically

important¹⁵¹. This symbiosis starts with the soil bacteria *Rhizobium* stimulating leguminous plants to develop root nodules, which the bacteria infect and inhabit (Figure 3.1). Upon ‘infection’ of leguminous plants such as *Pisum sativum* (pea), cells of *Rhizobium leguminosarum* (*R. leguminosarum*) are carried into the host plant cells via an infection thread and released into the plant nodules where they terminally differentiate into bacteroids. The root nodule is an optimal low-oxygen environment required for nitrogen fixation as the nitrogenase enzyme is inactivated by oxygen¹⁵². Ultimately, the two organisms dwell together in unity and establish metabolic cooperation: the plant reduces carbon dioxide into organic carbon in the form of dicarboxylic acids during photosynthesis and translocates these to the root to fuel the bacteria; the bacteria reduce molecular nitrogen into ammonia and export to the plant for assimilation¹⁵³.

As bacteria differentiate into nitrogen-fixing bacteroids, they undergo morphological and transcriptional changes. The free-living rhizobia appear as straight and regular rods; in root nodules, the nitrogen-fixing bacteroids are irregular in structure and are often Y-shaped¹⁵⁵. In bacteroids, genes involved in motility and reproduction are suppressed, and those required for nitrogen fixation are upregulated¹⁵⁶. Genes required for this nitrogen-fixing machinery are found on a *sym* plasmid, the genetic organisation of which varies between strains and hosts, allowing for tight control over host specificity. The *sym* plasmid encodes for nitrogenase subunits *nifHDK* as well as an operon encoding the *fixABCX* genes. The operon *fixABCX* encodes an electron transfer flavoprotein, FixAB, and its electron acceptor electron transfer flavoprotein quinone oxidase, FixCX. These proteins carry electrons to the nitrogenase enzyme (which requires 8 electrons per molecule of nitrogen fixed). Previous studies have shown the *fixABCX* gene products to be essential for a functional symbiosis in several species^{157–159}.

The high electron requirement of nitrogen fixation requires careful coordination with other cellular processes in order to maintain the redox balance within nitrogen-fixing bacteroids. Carbon storage sinks play an important role in maintaining this balance. *R. leguminosarum* oxidises a wide range of carbon sources and stores excess

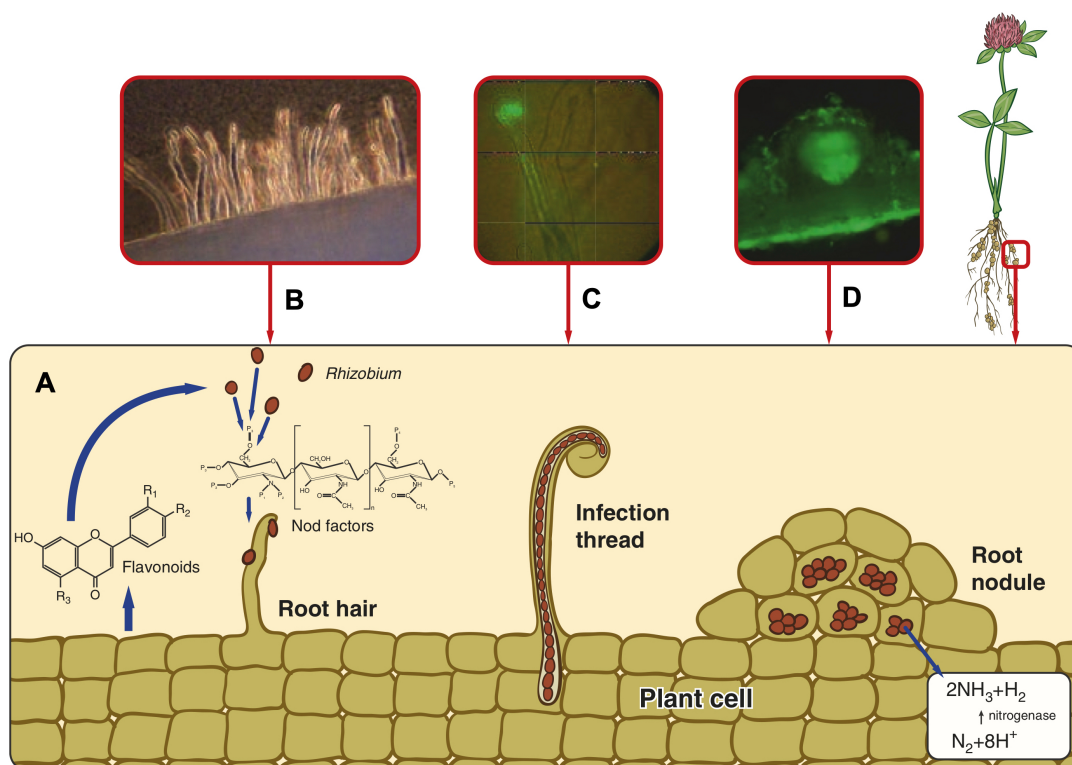


Figure 3.1: *Rhizobium*-legume symbiosis to form root nodules: (A) Bacteria sense flavonoids secreted from the roots of host leguminous plants. (B) Bacteria produce Nod factors which in turn causes root hair growth and deformation and formation of infection threads. (C) Bacteria are encapsulated in infection threads. (D) Infection threads, after reaching the root cortex, release bacteria into plant cells and initiate nodule development. Inside root nodules, bacteria turn into bacteroids. The figure is taken and modified from reference [154].

carbon and reductant in the forms of polyhydroxybutyrate (PHB) and glycogen. These storage polymers can be detected in rhizobia, including in their differentiated bacteroid state^{160,161}. PHB is produced by many bacteria under nitrogen-limiting but carbon-rich conditions¹⁶², and can be degraded to produce acetyl-CoA when cells switch to carbon-starvation states¹⁶³. In *R. leguminosarum* bv. *viciae*, PHB is found inside the infection threads carrying undifferentiated rhizobia into their host plant cells¹⁶⁴. This PHB is thought to be degraded to fuel bacteroid differentiation process^{160,164}. In developed bacteroids within pea nodules, glycogen is considered to be the main carbon storage compound used to relieve redox stress under conditions where carbon is in excess¹⁶⁰. Transmission electron microscopy (TEM) of infection threads¹⁶⁰ shows granules of a third, electron-dense storage compound, which is thought to be polyphosphate¹⁶⁵.

In a similar manner to PHB, polyphosphate is synthesised under conditions of stress and builds up as granules known as volutin¹⁶⁶.

To examine these biologically significant molecules in the process of bacteroid development as well as in symbiotic efficiency, *R. leguminosarum* bv. *viciae* strains were obtained, including from both liquid culture and from root nodules. Genetic mutants of PHB synthase, glycogen synthase and/or polyphosphate synthase, and a knockout mutant of a nitrogenase gene were also obtained.

Traditional chemical analysis is often used to study bacterial metabolism and measure the bulk abundance level of the molecules. In the case of rhizobia, bacteroids are terminally differentiated and cannot be cultured outside of the nodule. This requests large volumes of plant material to carry out biochemical assays. Furthermore, a bulk measurement approach could overlook important information on cellular heterogeneity which results from stochastic gene expression and various physiological states, such as between infection-thread bacteria and differentiated nitrogen-reducing bacteroids. Therefore, evaluation and quantification at the intracellular level has been difficult and dependent on molecular tools such as fluorescence *in situ* hybridisation (FISH) and quantitative polymerase chain reaction (qPCR)^{167,168}. These techniques, however, can only be applied to *in vitro* and pre-known genetic conditions, and often require external labelling.

Here I present SCRM, a label-free technique which measures molecular vibrational fingerprints of single cells. In this study, SCRS are used to discriminate different physiological states of *R. leguminosarum* and, for the first time, differentiate genetically defined mutants from their wild types. This study will demonstrate that SCRS are able to detect the relative abundance of intracellular biomolecules in single cells, such as cytochrome *c*, glycogen, PHB and polyphosphate. Notably, biochemical analysis at the the single-cell level allows discovery of heterogeneity within a single bacterial population.

3.2 Materials and Methods

Strains and culture conditions. Strains used in this study are detailed in Table 3.1. Free-living *R. leguminosarum* bv. *viciae* strains were grown at 28 °C in TY media¹⁶⁹ and the cultures were centrifuged at 4000 g for 20 minutes. The wild-type strain Rlv3841 and RlvA34 were grown in free-living laboratory culture and isolated as bacteroids from *P. sativum* nodules. Glycogen synthase (*glgA*) and PHB synthase (*phaC*) mutants were generated according to Lodwig et al.¹⁶⁰ and Terpollili et al.¹⁶². Mutants of polyphosphate synthase (*ppk*) and nitrogenase gene (*fixAB*) were constructed as part of this work by I. Webb. The mutants were inoculated on *P. sativum* and bacteroids were harvested.

Table 3.1: Strains used in Chapter 3

Strain	Description/mutation	Reference
Rlv3841	Wild type; <i>Rhizobium leguminosarum</i> bv. <i>viciae</i> str ^r derivative of strain 300	[170, 171]
RlvA34	Wild type; Sym plasmid pRL1JI (<i>viciae</i>) derivative of 8401 bv. <i>viciae</i> , str ^r	[172, 173]
LMB777	Rlv3841 $\Delta fixAB$	I. Webb (unpublished)
LMB816	Rlv3841 <i>phaC1::Tn5</i> , <i>phaC2::Ω</i> , str ^r , kan ^r , sp ^r	[174]
OPS706	Rlv3841 <i>ppk::Ω</i> , str ^r , tc ^r	I. Webb (unpublished)
OPS744	LMB816 <i>ppk::Ω</i> , <i>phaC1::Tn5</i> , <i>phaC2::Ω</i> , kan ^r , sp ^r , str ^r , tc ^r	I. Webb (unpublished)
RU1448	RlvA34 <i>glgA::TnB60</i> , str ^r , kan ^r	[160]
RU1478	RlvA34 <i>phaC::Ω</i> , <i>glgA::TnB60</i> , str ^r , kan ^r , spc ^r	[160]

Bacteroid preparation. *P. sativum* cv. Avola seeds were surface sterilised using 95% ethanol and 2% sodium hypochlorite. Seeds were sown in 1 L pots containing sterilised vermiculite and 400 mL of nitrogen-free rooting solution¹⁷⁵. Seeds were inoculated with 1×10^7 cells of the appropriate rhizobial strain. After 28 days, the plants were harvested and the nodules were collected. The collected nodules were ground in an isolation buffer (8 mM K₂HPO₄, 2 μM KH₂PO₄, 30 mM sucrose, 2 mM

MgCl₂). The resultant brei was strained through muslin and centrifuged at 1000 g for 5 minutes to remove plant material. The resulting supernatant was centrifuged at 6000 g for 5 minutes to give a bacteroid pellet.

SCRS measurements. Samples were washed with Milli-Q water three times to remove traces of media. They were diluted to make a concentration of $\sim 1,000$ cells mL⁻¹ to observe individual cells under microscope. Cells were spread onto an aluminium-coated slide and observed under a 100 $\times$ microscope objective (Olympus). Individual SCRS were acquired using an HR Evolution confocal Raman microscope (Horiba Jobin-Yvon, France) equipped with a 532 nm neodymium-yttrium aluminium garnet laser and 600 grooves/mm diffraction grating. The laser power on single cells was attenuated to 4.7 mW by neutral density (ND) filters. An objective with a magnification of 100 $\times$ 0.8 was used to focus the laser beam on single cells with a laser spot size of ~ 1 μ m². Raman scattering was detected by a charge coupled device (CCD) cooled at -68 °C. The acquisition time was 20 s per spectrum and the spectra were obtained in the wavenumber range of 600 to 1900 cm⁻¹. For each sample, 30-75 single cells were measured and from them, average spectra were plotted with the shaded area as standard deviation. The spectra were occasionally artificially offset to allow better visualisation.

Data processing and analysis of Raman spectra. All spectra were pre-processed by cosmic ray correction and polynomial baseline fitting using LabSpec 6 (Horiba). Spectral normalisation was done by vector normalisation of the entire spectral region. Data analysis, statistics and visualisation were done under an R environment using in-house scripts. Quantification of intracellular biomolecule concentration was done by integrating individual Raman bands and presented as box plots. The rectangle in the box plots represents the second and third quartiles with a line inside representing the median. The lower and upper quartiles are drawn as lines outside the box. Stars were used to indicate statistical significance where**** represents $p < 0.0001$, *** represents $p < 0.001$, ** represents $p < 0.01$, * represents $p < 0.05$, and ns represents not significant. Bacteroid sizes were determined by counting pixels in the microscopic images using ImageJ (National Institutes of Health, US).

Transmission electron microscopic analysis. Nodules were fixed in a solution of 2.5% glutaraldehyde in 0.05 M sodium cacodylate (v/v) at pH 7.3. Samples were washed in 0.05 M sodium cacodylate and post-fixed with 1% (w/v) OsO₄ in 0.05 M sodium cacodylate for 60 minutes at room temperature¹⁶⁰. Samples were then washed and dehydrated with ethanol before gradual infiltration with LR white resin (London Resin Company, Reading, UK) according to the manufacturer's protocol. Nodules were then sectioned using a Leica UC6 ultramicrotome (Leica). Ultrathin sections of 90 nm were picked up on 200 mesh hold grids coated in pyroxylin and carbon. Grids were viewed on an FEI Tecnai 20 transmission electron microscope at 2000 kV. Digital images were taken using an AMT XR60B digital camera (Deben).

3.3 Results and Discussion

3.3.1 SCRS distinguish cells in different physiological states

The transformation of *Rhizobium* cells from vegetative bacteria into nitrogen-fixing bacteroids involves a dramatic alteration of cell physiology. Understanding the differences between two cell types can help to reveal the underlying developmental pathway of the bacteroid differentiation which is extremely versatile and important. As a tool independent of intrinsic genetics and capable of single-cell resolution, SCRM was applied to characterise these isogenic but distinct physiological forms of *Rhizobium* and distinguish them at a single-cell level.

Figure 3.2A shows the averaged SCRS of *R. leguminosarum* bv. *viciae* wild-type Rlv3841, either as symbiotic bacteroids in mature root nodules (28 days post inoculation) or as free-living bacteria grown in liquid culture. Distinct spectral patterns can be observed. As around 800 Raman peaks were present in one spectrum in the range of 600 to 1800 cm⁻¹, unsupervised principal component analysis (PCA) was used to resolve the high-dimensional Raman dataset. PC1 and PC2 of the PCA explain 60.5% and 10.4% of the variances among single SCRS, respectively. A PCA plot along PC1 and PC2 reveals two distinct clusters, indicating clear differences between two phenotypic profiles of bacteria in two different physiological states (Figure 3.2B).

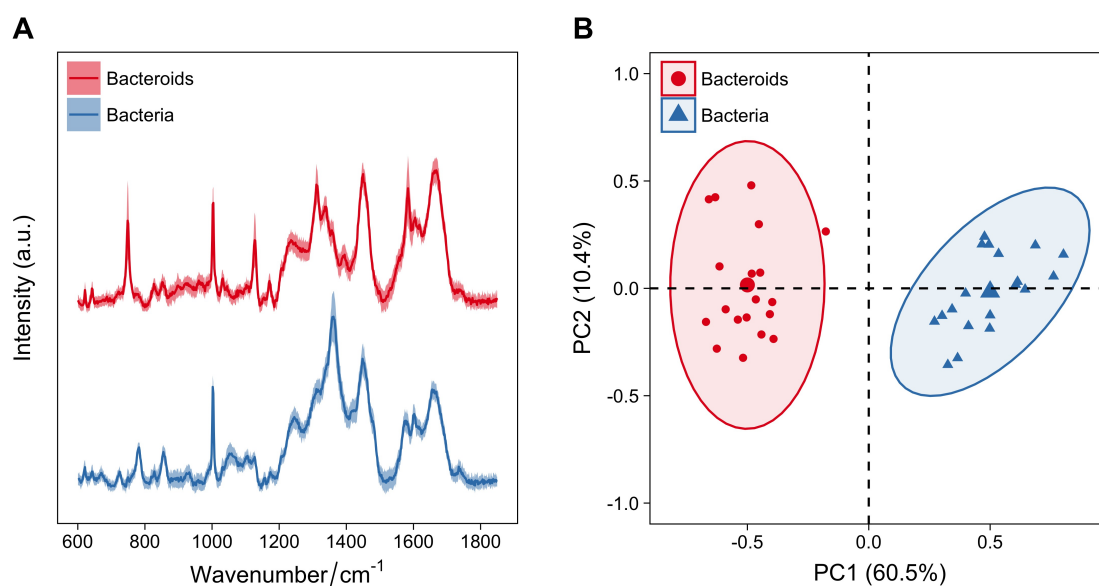


Figure 3.2: (A) SCRS of the wild-type Rlv3841 as either symbiotic bacteroids in root nodule (red) or free-living bacteria in liquid culture (blue), averaged from 30-50 single cells in each condition. Standard deviation is plotted as shaded area. (B) PCA plot along PC1 and PC2 based on the SCRS of Rlv3841 as either bacteroids (red) or bacteria (blue).

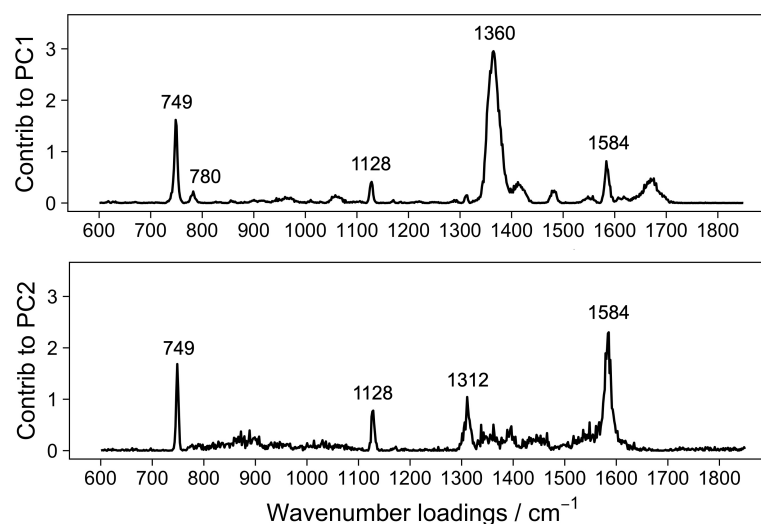


Figure 3.3: PCA contribution plots of squared PCA loading scores of Raman wavenumber to PC1 and PC2

Next, PCA contribution plots of squared PCA loading scores were obtained to map Raman wavenumbers against their contributions to a specific PC and to identify the most significant Raman bands contributing to the observed clustering. In loading plots of both PC1 and PC2 (Figure 3.3), four common Raman wavenumbers can be found at 749, 1128, 1312 and 1584 cm^{-1} . Table 3.2 summarises biological assignments and vibrational modes of Raman bands appeared in this section. All four bands can be attributed to the heme molecule cytochrome *c*, characteristic of vibrational modes at 749 (porphyrin ring breathing), 1128 (C–N stretching), 1312 (C–H deformation) and 1584 cm^{-1} (C–C stretching).

In Figure 3.4A with labelled cytochrome *c* positions, these cytochrome *c* bands were found high in the bacteroid cells isolated from the nodules and were much lower in the free-living cells incubated in rich media. This spectral observation was further confirmed by integrating individual Raman bands, thus semiquantifying concentrations of intracellular cytochrome *c*. Figure 3.4B shows significant decrease at all four positions of cytochrome *c* in bacteria grown in culture compared with symbiotic bacteroids ($p < 0.0001$).

Table 3.2: Assignment of Raman bands in section 2.3.1

Wavenumber/ cm^{-1}	Assignment	Vibrational mode	Reference
749	Cytochrome <i>c</i>	ν (porphyrin)	[176, 177]
1128	Cytochrome <i>c</i>	ν (C $_{\alpha}$ N)	[176, 177]
1312	Cytochrome <i>c</i>	δ (C $_m$ H)	[176, 177]
1584	Cytochrome <i>c</i>	ν (C $_{\alpha}$ C $_m$)	[176, 177]
780	DNA/RNA	ν (cytosine, uracil)	[50, 178]
1360	DNA/RNA	ν (guanine, cytosine)	[50, 179]

As introduced in the last chapter, cytochrome *c* is a heme protein absorbing visible light at a wavelength of around 530 nm, therefore, using a 532 nm excitation laser light generates a strong resonant Raman effect of the heme group¹⁷⁷ (Figure 3.5). This significantly enhances the Raman scattering cross section compared to the detection of other non-resonant intracellular constituents. These cytochrome subunits correspond to high-affinity cytochrome *cbb*₃ oxidase FixNOQP, which is used

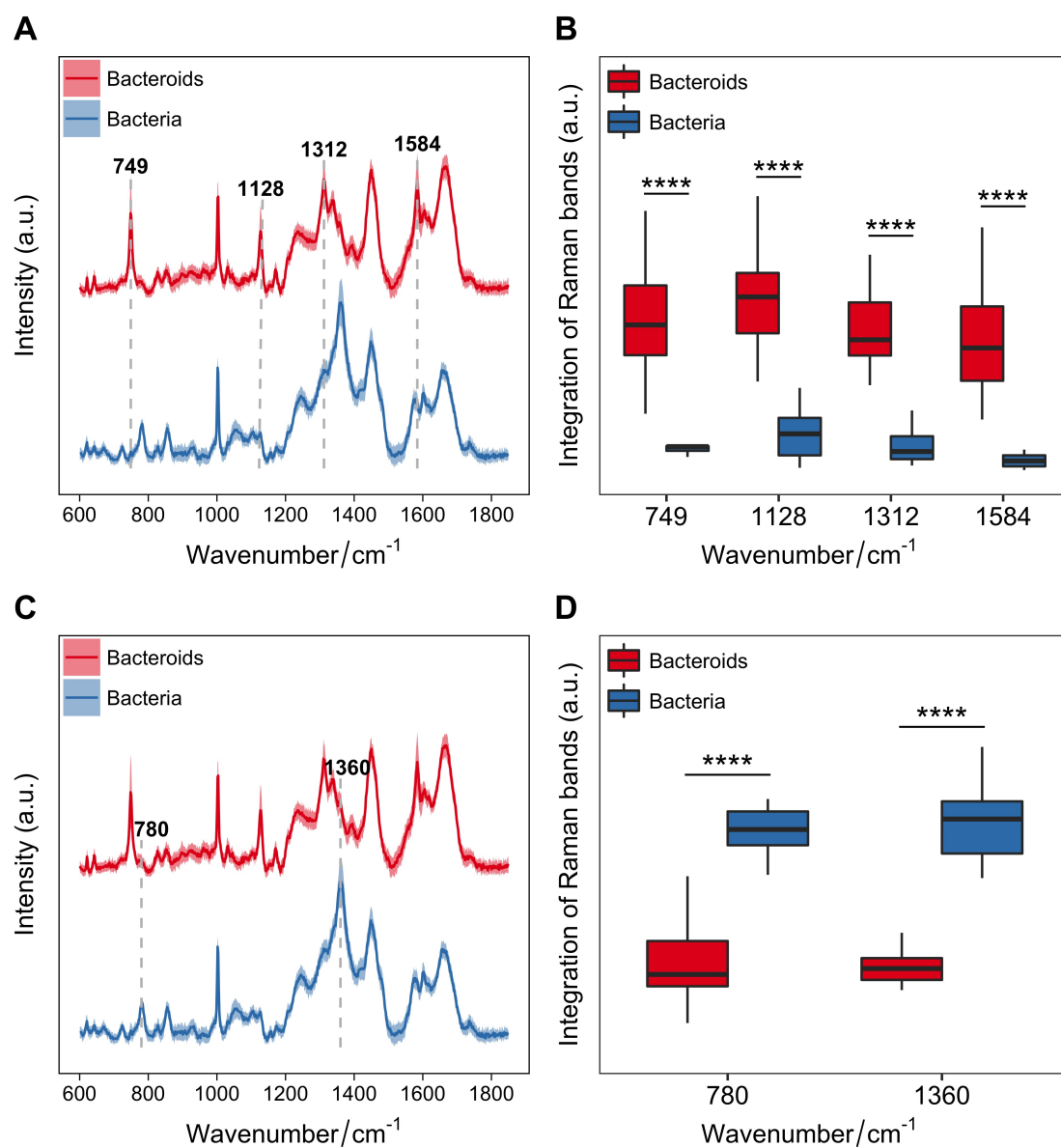


Figure 3.4: (A) SCRS with four cytochrome *c* bands labelled. (B) Spectral quantification of cytochrome *c* bands. (C) SCRS with two DNA/RNA bands labelled. (D) Spectral semi-quantification of DNA/RNA bands.

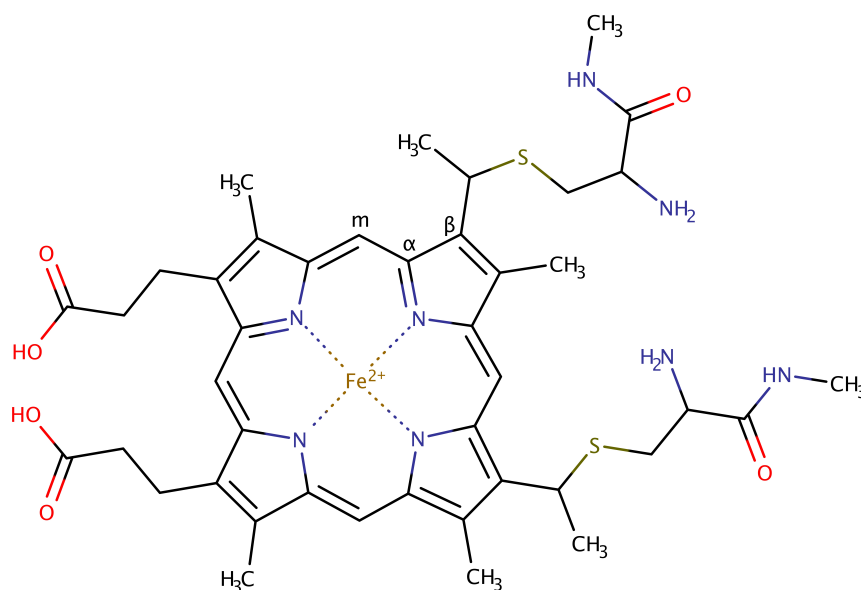


Figure 3.5: Chemical structure of cytochrome *c*

to maintain respiration within the microaerobic environment required for nitrogenase activity¹⁸⁰. This result is in good agreement with the physiological observation of *R. leguminosarum* bv. *viciae*, as cytochrome *c* is only expressed under low-oxygen conditions within the nodule, and it should be weak or absent in the free-living, aerobic culture¹⁸¹.

In addition to the difference in intracellular cytochromes, PCA loading plots also reveal variances in nucleotide contents at 780 and 1360 cm^{-1} (Figure 3.3), which can be assigned to cytosine and uracil ring breathing, and guanine and cytosine ring breathing, respectively (Table 3.2). Both of the Raman bands were observed to be lower in cells within nodules compared to cultured bacteria (Figure 3.4C), which suggests differences in nucleic acid contents between two physiological states of *R. leguminosarum* bv. *viciae*. This result was confirmed by semi-quantification/comparison of the two bands, showing significant differences at both positions ($p < 0.0001$) (Figure 3.4D).

Conflicting results on DNA content of rhizobia have been reported. While some publications reported a higher DNA content of *Rhizobium* bacteroids than that of broth-cultured *Rhizobium* bacteria^{182,183}, others showed DNA content of bacteroids either equal to or less than that of cultured bacteria^{184,185}. The non-dividing feature

of bacteroids might be attributed to the Raman observation of lower nucleic acid contents¹⁸⁶. Nevertheless, nucleic acid content can vary largely from measurement to measurement depending on different strains, host types, nodule ages, cell sizes, growth conditions, methods for preparing bacteroids and methods of detection^{187,188}. Traditional nucleic acid quantification methods like quantitative polymerase chain reaction (qPCR), fluorescence and spectrophotometry, require labouring extraction process, pre-known genetic knowledge, external labelling, and/or offer only bulk information^{189–191}. This study points SCRM as a possible alternative for measuring nucleic acid content which is quick, label-free and offers a single-cell resolution.

Together, these results suggest that SCRS are able to reflect phenotypic profiles of *R. leguminosarum* in free-living or bacteroid states. It provides an insight into bacterial cells in different physiological states without external labelling or pre-known knowledge of specific expression. At the same time, it also provides semi-quantification of biomolecules that of interest at the single-cell level.

3.3.2 SCRS link phenotypic changes with gene knockouts

R. leguminosarum synthesises a wide range of molecules as its carbon storage compounds during its symbiosis with its host plant. To examine the roles of these compounds and the underlying carbon managing network, single and double mutants of *R. leguminosarum* by *viciae* were made which lack PHB synthase, glycogen synthase, polyphosphate synthase or any two of them. SCRS were obtained for the wild-type Rlv3814 and its derived mutants LMB816 (*phaC1*::Tn5/*phaC2*:: Ω), OPS706 (*ppk*:: Ω) and OPS744 (*ppk*:: Ω /*phaC1*/*phaC2*), and the wild-type RlvA34 and its derived mutants RU1448 (*glgA*::TnB60) and RU1478 (*glgA*::TnB60/*phaC*:: Ω) (See Table 3.1 for details).

The ability to synthesise glycogen, PHB or/and polyphosphate is expected to be compromised in these mutants (*glgA*, glycogen synthase mutation; *phaC*/*phaC1*/*phaC2*, PHB synthase mutation; *ppk*, polyphosphate kinase mutation). Strain RlvA34 has a single gene encoding PHB synthase, *phaC*, whilst Rlv3841 has two PHB synthases, encoded by *phaC1* (on the chromosome) and a type II PHB synthase encoded by *phaC2-phaE* (on the *sym* plasmid and expressed only in bacteroids). All PHB mutants

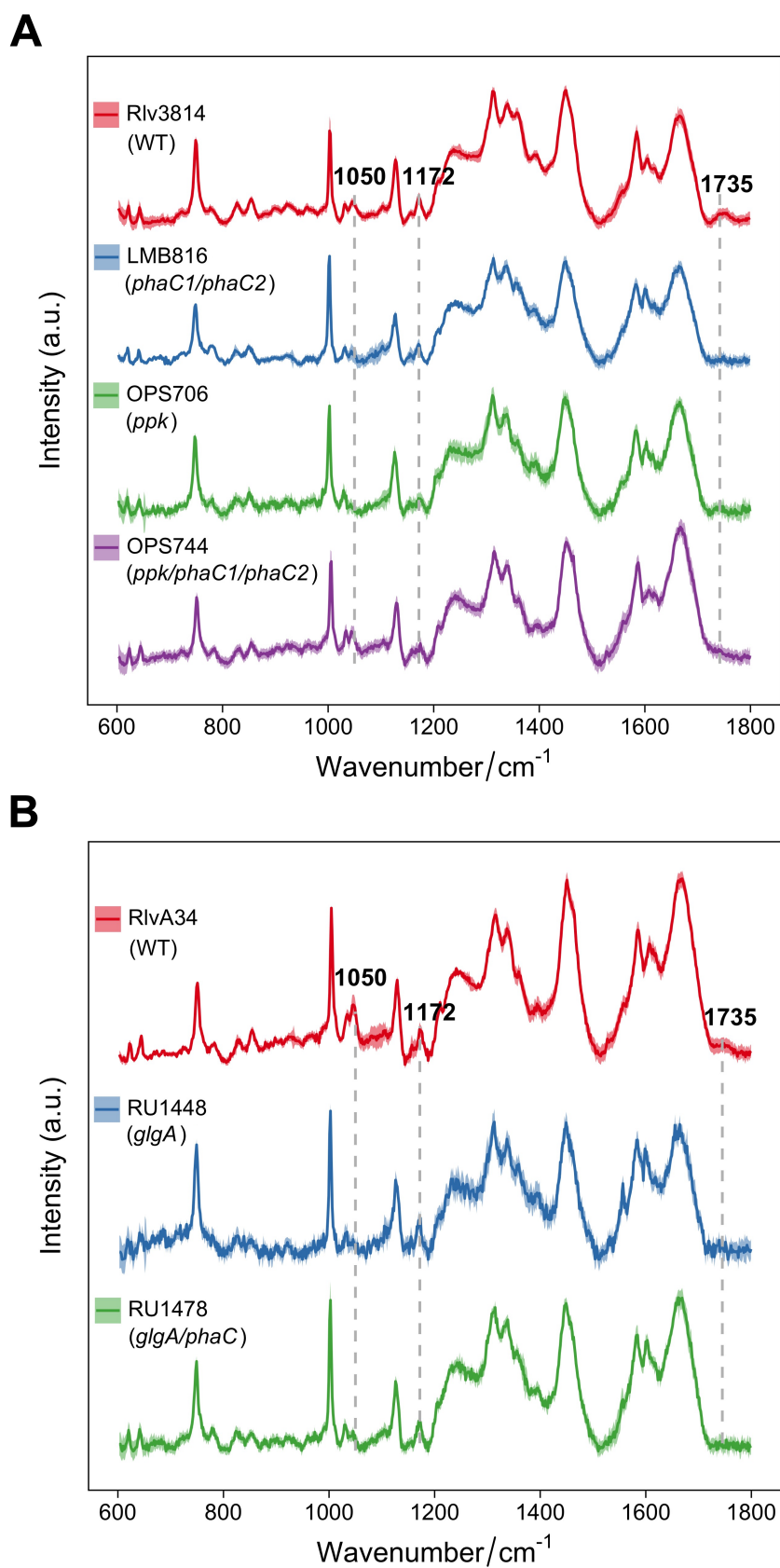


Figure 3.6: (A) SCRS of the wild-type Rlv3814 and three derived mutants. (B) SCRS of the wild-type RlvA34 and two derived mutants.

in the Rlv3841 background had both *phaC* genes mutated. The chemical phenotypes of PHB and glycogen mutants were confirmed with lower or non-detectable PHB or glycogen accumulation¹⁶⁰. The phenotype of the mutant OPS706 with *ppk* mutation was confirmed by a bulk DAPI-based assay which shows a significantly lower polyphosphate accumulation (I. Webb and J. Xu et al., unpublished).

Similar to the results of bulk chemical analysis, SCRS revealed clear phenotypic changes in the mutants due to the loss of the ability to synthesise different carbon storage molecules. Figure 3.6A shows the Raman spectra of the wild-type Rlv3814 and three derived mutants. Figure 3.6B shows the wild-type RlvA34 and its two derived mutants. Three highlighted bands at 1050, 1172 and 1735 cm^{-1} can be assigned to C–O stretching vibrations of glycogen, PO_2^{-1} stretching vibrations of polyphosphate and C=O stretching vibrations of the ester groups in PHB, respectively (Table 3.3).

Table 3.3: Assignment of Raman bands in section 2.3.2 and 2.3.3

Wavenumber/ cm^{-1}	Assignment	Vibrational mode	Reference
1050	Glycogen	$\nu(\text{CO})$	[192]
1172	Polyphosphate	$\nu(\text{PO}_2^{-1})$	[168]
1735	Polyhydroxybutyrate (PHB)	$\nu(\text{C=O})$	[168, 193]

Comparing to their corresponding wild types, a decrease in the characteristic PHB Raman band at 1735 cm^{-1} indicates that PHB synthesis in the mutants LMB816 (*phaC1/phaC2*), OPS744 (*ppk/phaC1/phaC2*) and RU1478 (*glgA/phaC*) have been disrupted, corresponding to the different mutations of PHB synthases (Figure 3.6).

Comparing SCRS of RlvA34 with RU1448 and RU1478 that contain defective *glgA* (glycogen synthase), a clear reduction at 1050 cm^{-1} (glycogen) in RU1448 and RU1478 indicates that the glycogen synthesis has been compromised (Figure 3.6B). A reduction in the PHB band is also seen in the single mutant of glycogen mutant RU1448 (*glgA*), suggesting a link between biosynthesis of these two carbon storage molecules.

SCRS of the PHB mutant LMB816 (*phaC1/phaC2*) also show a reduction in the glycogen band (Figure 3.6A), indicating that the disruption of PHB synthesis has an

impact on glycogen synthesis. This supports the hypothesis that the biosynthesis of these storage molecules is linked.

A reduction at 1172 cm^{-1} (polyphosphate) of the mutants OPS706 (*ppk*) and OPS744 (*ppk/phaC1/phaC2*) compared to the wild-type Rlv3814 suggests that the mutants have compromised the ability to make polyphosphate, which is consistent with a mutation in a *ppk* gene encoding polyphosphate kinase (Figure 3.6A).

Without the need of biochemical assays, SCRS of both single and double mutants of one or two carbon storage molecules reveal a blockage of specific biosynthesis by identifying relevant Raman bands.

Statistical analysis of SCRS from 30 single cells has been applied to quantify the relative abundance of the intracellular polymer *in situ* and confirm the link between gene knockouts and phenotypic changes in mutants. Figure 3.7 shows the abundance distributions of different levels of intracellular carbon storage polymers by individually examining Raman bands at 1735 cm^{-1} (PHB), 1050 cm^{-1} (glycogen) and 1172 cm^{-1} (polyphosphate).

Figure 3.7A shows a statistical distribution of intracellular PHB levels in bacteroid single cells. It reveals a remarkably lower PHB accumulation ($p < 0.0001$) in mutants LMB816 (*phaC1/phaC2*), OPS744 (*ppk/phaC1/phaC2*) and RU1478 (*glgA/phaC*), all of which completely prevent PHB synthase (either *phaC* in the A34 background, or *phaC1/phaC2* in the Rlv3841 background). A correlation between PHB and glycogen, which both serve as carbon storage compounds, is also found in both SCRS (Figure 3.6) and single-cell statistics (Figure 3.7A). A single gene mutation of glycogen synthesis in the mutant RU1448 (*glgA*) leads to a decrease in PHB level compared to in the wild-type RlvA34 ($p < 0.01$). This suggests a link between biosynthesis of these two carbon storage molecules; in the absence of glycogen stores, PHB was more rapidly broken down in order to provide carbon or reductants. On the other hand, a single mutation in a polyphosphate kinase *ppk* in OPS706 shows no effect on the intracellular PHB levels ($p > 0.05$).

The glycogen quantification was compared by integrating the Raman band at 1050 cm^{-1} of each strain (Figure 3.7B). It is observed that RU1448 (*glgA*) and RU1478

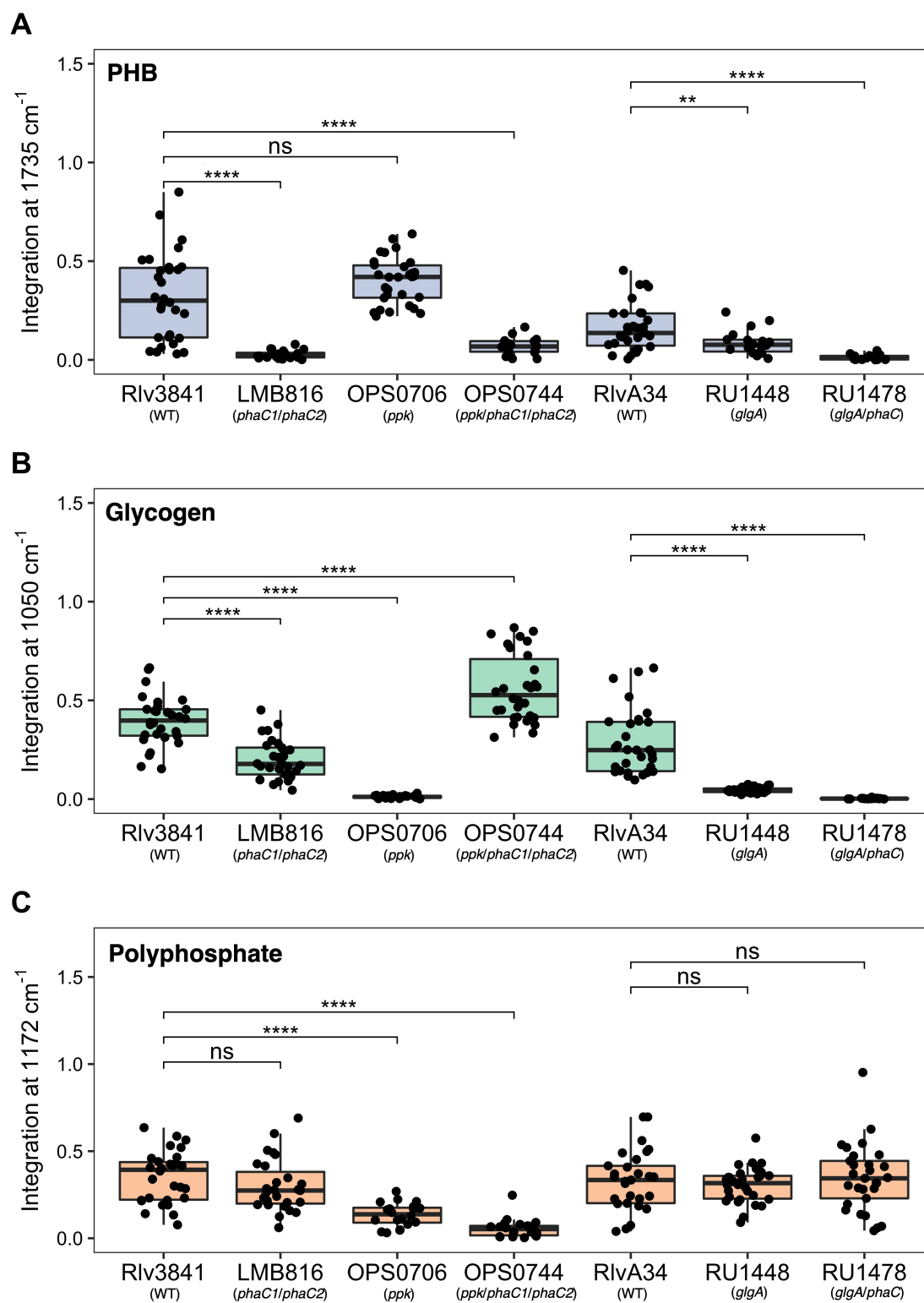


Figure 3.7: (A) Quantification of PHB. (B) Quantification of glycogen. (C) Quantification of polyphosphate.

(*glgA/phaC*) have a significant decrease in the glycogen level, compared to their wild-type strain RlvA34 in bacteroid state ($p < 0.0001$). Both single (RU1448) and double (RU1478) mutants lose their ability to store glycogen in the carbon-excessive bacteroid state, which is in a good agreement to the previous report¹⁶⁰. Notably, a single mutation in the polyphosphate kinase (*ppk*) leads to a significantly compromised glycogen accumulation in OPS706 ($p < 0.0001$). Polyphosphate plays a role in redox stress protection and metabolism, therefore polyphosphate disruption may result in higher redox stress and a requirement for glycogen as a reductant sink^{173,194}. SCRS and statistical quantification of the PHB mutant LMB816 (*phaC1/phaC2*) display a glycogen level lower than the wild-type Rlv3841 ($p < 0.0001$), supporting the hypothesis that biosynthesis of glycogen and PHB is linked. Surprisingly, the double gene mutant OPS744 (*ppk/phaC1/phaC2*) isolated from the root nodules shows a strong glycogen accumulation, significantly higher than the wild type ($p < 0.0001$). The double PHB synthase mutant LMB816 (*phaC1/phaC2*) does not show this overproduction of glycogen, suggesting that the removal of a polyphosphate synthesis gene could have an effect on glycogen synthesis in *R. leguminosarum* bv. *viciae*.

In Figure 3.7C comparing intracellular levels of polyphosphate, Raman analysis shows a significant decrease in both OPS706 (*ppk*) and OPS744 (*ppk/phaC1/phaC2*) compared to their wild-type Rlv3814 ($p < 0.0001$), which confirms the compromised capability of polyphosphate synthesis in these novel mutants. On the other hand, removing PHB synthases and glycogen synthase seems to have no effect on polyphosphate production.

The observation of a coupled relationship between PHB and glycogen disagrees with previous experiments of *Rhizobium tropici* and *Rhizobium etli*, reporting that a mutation in PHB synthesis led to increased glycogen accumulation¹⁹⁵. Our results suggest a clear linkage between these two storage molecules in which a knockout in glycogen synthesis compromises the ability to synthesise PHB and a knockout in PHB synthesis reduces glycogen accumulation. However, glycogen is specifically increased when PHB and polyphosphate are knocked out at the same time. The production of these carbon storage molecules may be tightly coupled in *R. leguminosarum* bv.

viciae, under the control of a so-far unknown global carbon regulatory network acting to maintain redox balance. Further experiments on double mutation in *ppk* and *glgA* could be done to investigate the reciprocal relationship on PHB accumulation when polyphosphate is involved.

3.3.3 Single-cell resolution allows heterogeneity analysis within a single bacterial population

The ability to link phenotypic biopolymers (e.g. PHB, glycogen and polyphosphate) and genetic mutations by SCRS was further investigated in LMB777 ($\Delta fixAB$), a derived mutant of the wild-type Rlv3841. The *fixABCX* mutations in *R. leguminosarum* has an essential role for a functional nitrogen-fixing symbiosis with the host. The *fix* gene mutants can cause nodule development, but the nodules are not capable of fixing nitrogen, thus demonstrating a Nod⁺ Fix⁻ phenotype¹⁵³. Here, SCRM was applied to carry out biological characterisation of the mature LMB777 bacteroids at 28 days post inoculation.

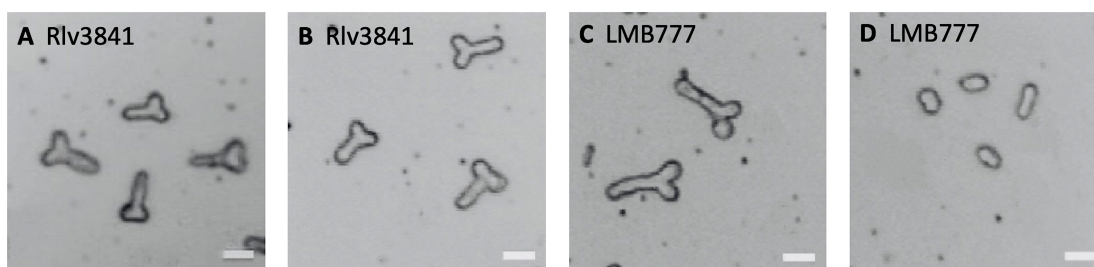


Figure 3.8: Bright-field microscopic images of the wild-type Rlv3841 mature bacteroids (A & B), and the $\Delta fixAB$ mutant LMB777 with two distinct morphologies (C & D). Scale bar 2 μ m.

Under microscope, drastic morphological differences were observed between the wild-type bacteroids (Rlv3841) and the $\Delta fixAB$ mutant cells (LMB777) (Figure 3.8). While the wild type showed a homogeneous Y-shaped morphology (Figure 3.8A and B), two distinct morphologies were found in the mutant (Figure 3.8C and D). The first morphology is of swollen and larger Y-shaped bacteroids; the second morphology is rod shape and of smaller cells. This indicates two separate fates for mutated cells that cannot fix nitrogen and might account for distinct biological phenotypes. The heterogeneity in phenotypes within a single bacteroid population makes it difficult

to carry out conventional biochemical assays, as the phenotypes mask one another. Here, the use of SCRM provides a solution to unveiling biochemical phenotypes of the mutant at the single-cell level.

Figure 3.9A shows the SCRS of the wild-type Rlv3841 ($n = 57$) and the mutant LMB777 ($n = 75$). Intriguingly, two distinct Raman phenotypes were observed; either with a PHB band larger than that in the wild type (PHB+) or with a substantially smaller PHB band at 1735 cm^{-1} (PHB-).

The sizes of individual cells were calculated for the wild-type Rlv3841, LMB777 (PHB+) and LMB777 (PHB-) by ImageJ. It was found that these two phenotypes corresponded to differing cell sizes compared to the wild type (Figure 3.9B). A fitted normal distribution of the single-cell measurements shows an average cell size of $3.312\text{ }\mu\text{m}^2$ for the PHB+ phenotype and $1.386\text{ }\mu\text{m}^2$ for the PHB- phenotype, whilst the wild-type sits in the middle with a size of $2.284\text{ }\mu\text{m}^2$. The PHB- phenotype features a smaller standard deviation in cell size, matching the microscopic observation that 95% of this phenotype (18 out of 19) retained rod shape with a smaller size (Figure 3.8D). The PHB+ phenotype, on the other hand, shows a wider bell-shaped distribution, and 75% of the population (42 out of 56) being Y-shaped (Figure 3.8C).

This result combined with the bright-field microscopic images confirms that there is a correlation between individual cell size and biological phenotype: PHB tends to accumulate in larger and Y-shaped cells and disappears in smaller and rod-shaped bacteroids. The glycogen levels characterised by the 1050 cm^{-1} band appeared to follow the same trend as glycogen being higher in the PHB+ phenotype, indicating larger bacteroids of *R. leguminosarum* that are unable to fix N_2 accumulate both glycogen and PHB. This result also confirms the linkage between the two storage molecules, PHB and glycogen, as described in the last section. The polyphosphate levels at 1172 cm^{-1} were also differing in two phenotypes (Figure 3.9A); the PHB+ subgroups showed a substantial decrease in polyphosphate level compared to the wild type while the PHB- cells accumulate more intracellular polyphosphate.

Our results indicate a bimodal heterogeneity existed in a single population of bacteroids due to a mutation in FixABCX. The first phenotype is that of large and Y-shaped

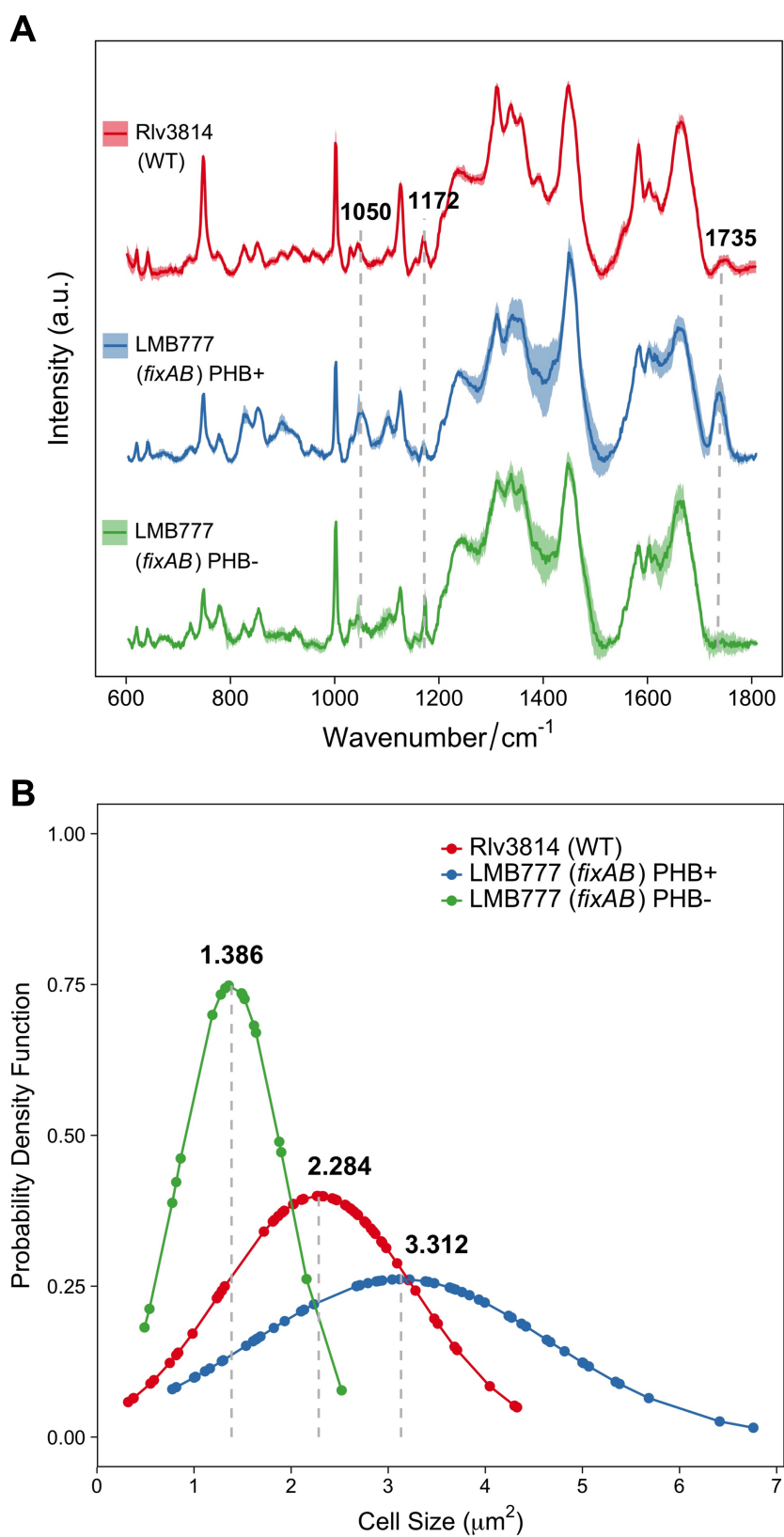


Figure 3.9: (A) SCRS of the wild-type Rlv3814 and the mutant LMB777, showing two distinct Raman phenotypes. (B) The probability density function for a normal distribution of cell sizes quantified for each individual cell of Rlv3814 and LMB777. Cells of LMB777 were grouped by two Raman phenotypes of either PHB+ or PHB-. The average cell size in μm^2 for each group is annotated in bold font.

bacteroids containing granules of PHB and glycogen; the second cell type resembles in both size and appearance undifferentiated bacteria from infection threads, producing and accumulating more polyphosphate.

R. leguminosarum bv. *viciae* 3841 has two genes encoding PHB synthase. The first, *phaC1*, is found on the chromosome, whilst the second, *phaC2*, is bacteroid-specific, encoded on the *sym* plasmid. In *Azotobacter* spp. PHB has been proposed to play a role in protecting cells from excess reductant under microaerobic conditions¹⁹⁶. Since FixABCX donates electrons to nitrogenase, the absence of FixAB will prevent electrons from being accepted by nitrogenase, leading to an excess of reduced nucleotides in bacteroids. If FixAB does pass electrons to FixCX and onwards to the respiratory chain, this would add to the build-up of reductant and production of PHB. In terms of polyphosphate, bacteria such as *E. coli* and *Lactobacillus* have been shown to accumulate polyphosphate under conditions of stress including osmotic stress, oxidative stress and nutrient starvation^{196,197}.

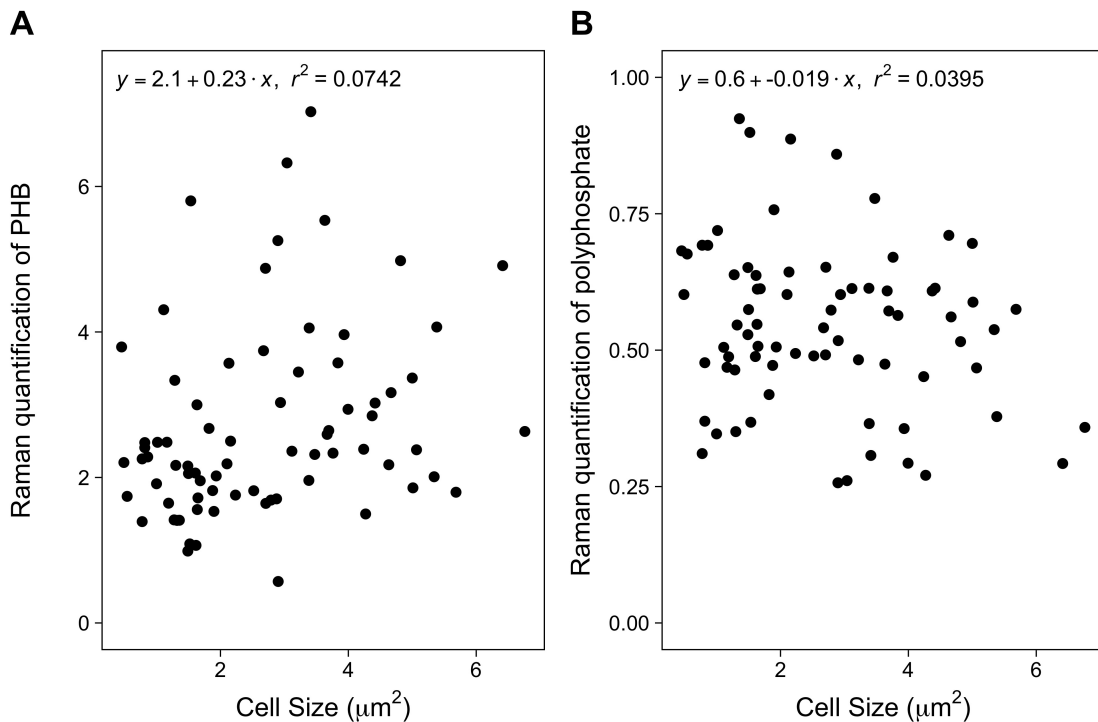


Figure 3.10: Individual cell size versus Raman quantification of (A) PHB and (B) polyphosphate. A t-test performed on the slope of the linear regression is indicated in red.

To further examine the role of the carbon storage polymers in FixABCX mutation, intracellular quantification of PHB and polyphosphate was done by integrating the Raman bands at 1735 and 1172 cm^{-1} , respectively, and mapped with the individual cell size (Figure 3.10). Although a linear fitting was calculated as at each upper-right corner of the scattered plots, the linear relationship was not significant enough to draw conclusions on any proportional relationships between the PHB or polyphosphate level with the corresponding cell size.

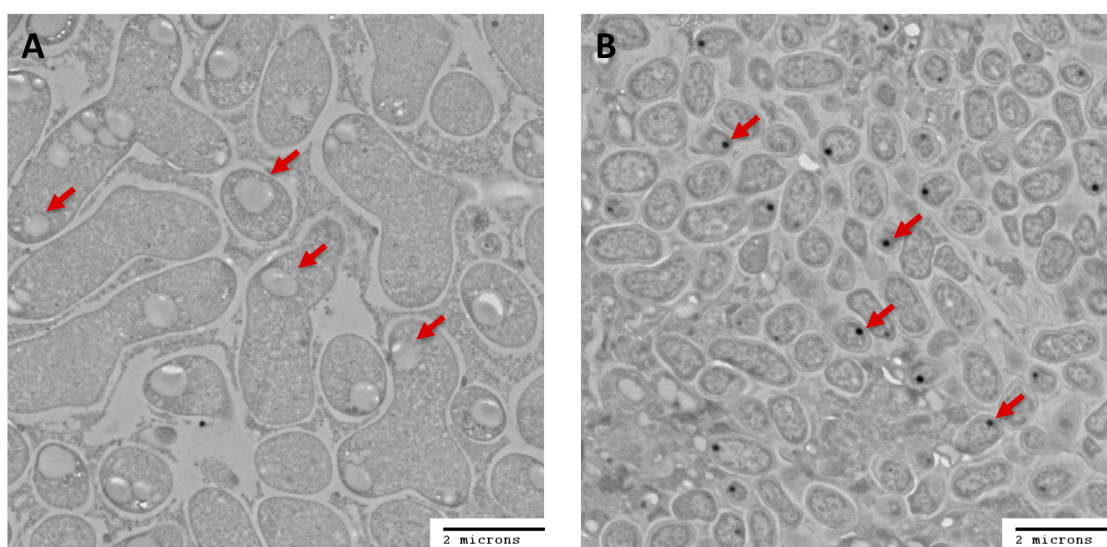


Figure 3.11: Electron microscopic images of sections of nodules taken from *P. sativum* inoculated with LMB777.

Transmission electron microscopic (TEM) images of LMB777 confirm the Raman findings by revealing two different morphological phenotypes (Figure 3.11). Similar as in the Raman results, the first phenotype is that of large bacteroids containing large granules of PHB (indicated by the red arrows in Figure 3.11A). The second cell type is of smaller cells and contains dark polar inclusions that thought to be electron-dense polyphosphate granules (indicated by the red arrows in Figure 3.11B). These phenotypes are found within the same nodule, though homogeneity was seen within individual plant cells.

While TEM images reveal morphological details of heterogeneous bacteroids, it would be difficult to reveal the identity of these granules and discover details of such previously unknown phenomenon in non-fixing rhizobial species without the use of

SCRM. The heterogeneity seen between isogenic bacteroids within different plant cells reveals developmental feedback from the plant when there is a lack of provision of fixed nitrogen. This heterogeneity only emerges in mature nodules, demonstrating a delay in the response from the plant host. The appearance of both these cell types in a single nodule means that traditional bulk biochemical analysis of *fixABCX* mutant nodules would average out the dual phenotype. Alternative strategies for cell sorting such as fluorescence labelling in this system would require large volumes of bulk material, and terminal bacteroid differentiation means that a large number of plants would be required. Using SCRM requires a small volume of bacteroids and allows analysis without prior labelling. This is particularly beneficial for unculturable samples such as bacteroids.

3.4 Conclusions

Individual biochemical assays require a large volume of starting materials, which can be difficult in the case of terminally differentiated bacteria that have lost their ability to reproduce, such as bacteroids. Bacteroids within a single nodule have extremely heterogeneous populations of which important information can be averaged out in bulk chemical analysis. Single-cell assays could also be beneficial in the case of organisms that are difficult to culture, such as unknown soil bacteria or marine microorganisms. In this study, SCRM was employed to study the symbiotic diazotroph *R. leguminosarum* bv. *viciae*.

Firstly, it has been shown that SCRS can distinguish rhizobial cells in different physiological states. SCRS of free-living culture and terminally differentiated bacteroids in pea nodules primarily differ, as the free-living cells lack Raman bands on cytochrome *c* which are present in the *cbb3* complex. These cytochromes correspond to the symbiotic terminal oxidase required to maintain respiration under the low-oxygen conditions found within root nodules. The Raman band at 780 cm^{-1} also indicates a difference at the nucleic acid level. The terminally differentiated bacteroids undergo a suite of metabolic and genetic changes in order to carry out symbiotic nitrogen fixation. PCA analysis shows a clear classification between bacteroids and free-living bacteria.

In together, SCRS and PCA analysis reflect their phenotypic changes due to different physiological states despite being genetically identical.

Secondly, this study demonstrates that, for the first time, Raman microscopic measurements are able to discriminate genetic mutants, including both single and double mutants of carbon storage polymers. Genetic mutations can be confirmed and semiquantified at the intracellular level by analysing characteristic Raman biomarkers (e.g. PHB, glycogen and polyphosphate). By creating metabolic profiles of single mutant cells and quantifying intracellular levels of macromolecules, SCRM is applicable to the study of complex carbon regulatory networks. Many biochemical assays only focus on one specific molecule analysis, so collective protocols would be required to create a basic metabolic profile. Other technologies such as fluorescent labelling require information on DNA/RNA or protein in order to design fluorescent tags. SCRS analysis allows a semi-quantitative analysis of several macromolecules within a single experiment and reveals a comprehensive intrinsic chemical profile in single cells without prior knowledge. This can provide novel insights into productions of different metabolites and has the potential of making new scientific discoveries of cell phenotypes. More accurate quantification can be achieved if used in conjunction with other analytical techniques such as gas chromatography/mass spectrometry which provide a standardised reference.

Last but not least, SCRM offers single-cell resolution which allows heterogeneity analysis within an isogenic bacterial population. The SCRS of a *fixAB* mutant revealed a previously unknown heterogeneity in bacteroid morphology and phenotype in non-fixing rhizobia. Two distinct phenotypes were identified; the first phenotype is that of large bacteroids containing a large amount of PHB and glycogen, whilst the second phenotype is of smaller cells and accumulates polyphosphate. A proportional relationship between cell size and intracellular concentration of PHB was also identified. The appearance of two cell types in a single nodule means that traditional bulk biochemical analysis of nodules would average out the dual phenotype. SCRM provides a strategy for biochemical characterisation at the single-cell level, which is particularly beneficial for unculturable and terminally differentiated samples such

as bacteroids. Furthermore, where genomics cannot reach, SCRM can be potentially applied to a wide range of isogenic cells or mutants with heterogeneous behaviours.

Chapter 4

Application of SCRM-DIP in revealing microbial carbon metabolism

Contents

4.1 Introduction	66
4.2 Materials and Methods	67
4.3 Results and Discussion	70
4.3.1 SCRM-DIP detects carbon metabolism by using deuterated carbon sources	70
4.3.2 SCRM-DIP reveals distinct biosynthesis pathways of fatty acids starting from different deuterated carbon sources	77
4.3.3 SCRM-DIP reveals distinct biosynthesis pathways of phenylalanine starting from different deuterated carbon sources	79
4.4 Conclusions	87

In the last chapter, we have shown SCRS as a snapshot of bacterial phenotype, consisting of a collection of biomolecules that reflects bacterial physiological states, links with genetic mutations and reveals bacterial subpopulations. In this chapter, I will present SCRM-DIP, the tool that combines SCRM and deuterium isotope probing (DIP), to probe microbial metabolism. Specifically, SCRM-DIP can not only indicate the metabolic activity of bacteria using deuterated carbon sources but also reveal different metabolic pathways adopted by different microorganisms.

4.1 Introduction

Among all isotopes in Raman stable-isotope-probing (SIP) studies, it has been found that Raman spectroscopy gives sensitive detection of the incorporation of deuterium in single cells, as the carbon-deuterium (C–D) vibrations shifted from

the carbon-hydrogen (C–H) vibrations exhibit a clear and strong band in the silent region of SCRS which has no detectable Raman bands in most, if not all, cells growing with naturally abundant elements (e.g, ^1H , ^{14}N , ^{16}O and ^{12}C). Deuterated substrates can be synthesised by hydrogen-deuterium exchange reactions in the presence of D_2O , which is not only much more cost-effective than other stable isotope substrates but also able to produce more complex substrate or mixed compounds^{150,198,199}. Hence, providing a high variety of possible isotopic substrates, SCRM-DIP is invaluable in the functional study of cells in their natural conditions at the single-cell level.

In this study, we exploited one aspect of SCRM-DIP using deuterated carbon source to link cells to their carbon metabolism. We applied SCRM-DIP to study the assimilation of specific deuterated carbon substrates such as glucose and naphthalene into lipids and proteins in single microbial cells. We used a mixture of glucose/glucose- d_{12} or naphthalene/naphthalene- d_8 to probe various anabolic pathways of carbon sources and semi-quantify the biosynthesis of fatty acids and phenylalanine in single cells of *Escherichia coli* (*E. coli*) and *Pseudomonas putida* (*P. putida*). Given its ability to probe metabolic reactions, we believe SCRM-DIP can provide culture-independent insights in understanding the complex physiological systems of microbial communities.

4.2 Materials and Methods

Strains, culture conditions and plasmid transfer. Strains and plasmid used in this study are detailed in Table 4.1. For all strains tested, media were supplemented with carbon sources containing different ratios of D-glucose (Sigma, UK, D9434) to D-glucose- d_{12} (Sigma, UK, 616338) in water, or naphthalene (Sigma, UK, 84679) to naphthalene- d_8 (Sigma, UK, 176044) dissolved in dimethyl sulfoxide (DMSO).

Pure cultures of *Escherichia coli* (*E. coli*) DH5 α were grown aerobically overnight with 150 rpm shaking at 37 °C in M9 minimal medium with 10 mM glucose as the sole carbon source. Cultures of *Pseudomonas putida* (*P. putida*) strains UWC1 and G7 were grown aerobically for 72 hours with 150 rpm shaking at 30 °C in M9 medium with 10 mM naphthalene or glucose as the sole carbon source. Optical density at 600 nm (OD_{600}) was measured during cultivation with deuterated carbon sources.

Table 4.1: Strains and plasmid used in Chapter 2

Strain/plasmid	Description	Reference
<i>Escherichia coli</i> DH5 α	a non-pathogenic strain developed for laboratory cloning use	Invitrogen U.K.
<i>Pseudomonas putida</i> UWC1	Mutant of <i>P. putida</i> KT2440; Rip ^r ; able to use glucose as sole carbon source	[200, 201]
<i>Pseudomonas putida</i> G7	Naphthalene-degrading bacterium with plasmid NAH7; cannot use glucose as sole carbon source	[202]
<i>Pseudomonas fluorescens</i> WH2	Naphthalene-degrading bacterium with plasmid pWH2-Nah	[203]
<i>Pseudomonas putida</i> UWC1 (pWH2-Nah ⁺)	Transconjugant of <i>P. putida</i> UWC1 bearing pWH2-Nah plasmid from <i>P. fluorescens</i> WH2	This study
pWH2-Nah	Naphthalene-degrading plasmid of <i>P. fluorescens</i> WH2	[203]

Pseudomonas fluorescens (*P. fluorescens*) WH2 contains the pWH2-Nah plasmid carrying the naphthalene-degradation operon and the plasmid was transferred via conjugation from *P. fluorescens* WH2 to *P. putida* UWC1 (Table 4.1). Briefly, *P. fluorescens* WH2 and *P. putida* UWC1 were grown separately in LB medium overnight at 30 °C. One hundred μ L of each were mixed and added on a 0.22 μ m membrane filter (47 mm diameter, Millipore UK) placing on an LB agar plate, followed by incubation at 30 °C for 16 hours. The bacterial biofilm grown on the filter was taken out and resuspended in 5 mL of PBS by vortex. Transconjugants bearing the naphthalene-degradation plasmid pWH2-Nah were selected in M9 minimal medium with 100 μ g/mL rifampicin and naphthalene as the sole carbon source. As *P. putida* UWC1 cannot assimilate naphthalene but is rifampicin-resistant, and *P. fluorescens* WH2 does not resist rifampicin, only the transconjugants can grow on the selective plates. Colony PCR was performed to amplify 16S ribosomal RNA (16S-rRNA) of the transconjugants using primers 63f and 1387r²⁰⁴. The identity of the transconjugants *P. putida* UWC1 (pWH2-Nah⁺) were confirmed by DNA sequencing of the PCR product.

SCRS measurements. Samples were washed with Milli-Q water three times to remove traces of media. They were then diluted to a concentration of $\sim 1,000$ cells mL⁻¹ and spotted onto an aluminium-coated slide. Individual SCRS were acquired using an

HR Evolution confocal Raman microscope (Horiba Jobin-Yvon, France) equipped with a 532 nm neodymium-yttrium aluminium garnet laser and 300 grooves/mm diffraction grating. The laser power on single cells was attenuated to 4.7 mW by neutral density (ND) filters. An objective with a magnification of 100×0.8 (Olympus) was used to focus the laser beam on single cells with a laser spot size of $\sim 1 \mu\text{m}^2$. Raman scattering was detected by a charge coupled device (CCD) cooled at -68°C . The acquisition time was 10 s per spectrum and the spectra were obtained in the wavenumber range of 340 to 3500 cm^{-1} . For each sample, 30 single cells were obtained and from them, averaged spectra were plotted with the shaded area as standard deviation. The spectra were occasionally manually offset to allow better visualisation.

Data processing and analysis of Raman spectra. All spectra were pre-processed by cosmic ray correction and polyline baseline fitting using LabSpec 6 (Horiba). Spectral normalisation was done by vector normalisation of the entire spectral region. Data analysis, statistics and visualisation were done under an R environment using in-house scripts. Quantification of intracellular biomolecule concentration was done by integrating individual Raman bands. The ratio of C–D / (C–D + C–H) was used to indicate the extent of deuterium incorporation from deuterated carbon source^{128,205}.

Theoretical calculation of Raman frequencies The equilibrium geometry of the palmitic acid molecule was calculated at the B3LYP/6 31G** level of theory using GAUSSIAN09. The conformer obtained had C_s point group symmetry with the aliphatic backbone lying in the single mirror plane. Wavenumbers of Raman-active vibrations were then calculated for various substitution patterns on the molecular backbone, including all-H and all-D forms, the distinct single-D substitutions in the all-H form, the distinct single-H substitutions in the all-D form, and the form with full deuteration along half the length of the backbone, starting from either methyl or carboxylic terminal.

Raman wavenumbers of phenylalanine were then calculated at the same level for 32 possible isotopomers obtained by H/D substitution on the phenyl ring. The ring breathing vibration at 1017 cm^{-1} in the all-H structure was identified as the match for the experimental peak at 1004 cm^{-1} , as it was the only intense peak in that region of the spectrum. Data from each of the 32 structures were then grouped by %D in the

ring (i.e., 0%, 20%, 40%, 60%, 80% and 100%, depending on whether there were 0, 1, 2, 3, 4 or 5 D on the ring), and they were further grouped by mirror symmetry to account for the approximately free rotation about the ring-backbone bond.

4.3 Results and Discussion

4.3.1 SCRM-DIP detects carbon metabolism by using deuterated carbon sources

Both *E. coli* DH5 α and *P. putida* UWC1 are able to use glucose as the sole carbon source^{200,206} (Table 4.1). Figure 4.1A and 4.1B show SCRS of *E. coli* DH5 α and *P. putida* UWC1 grown overnight in glucose as a sole carbon source with different percentages of fully deuterated glucose- d_{12} . With the presence of deuterated glucose, a distinguishable Raman band appeared in the region between 2070 and 2300 cm^{-1} in SCRS, and this broad band was attributed by the C–D stretching vibrations^{126,145}. The intensity of this C–D band enhanced while that of the C–H band at 2800 – 3100 cm^{-1} reduced with an increasing percentage of deuterated glucose- d_{12} in the growth media. This spectral pattern demonstrates exchanges of intracellular H by D due to the catabolism of deuterated carbon sources.

P. fluorescens WH2 and *P. putida* G7 are two bacteria capable of utilising naphthalene as the sole carbon and energy source (Table 4.1). The naphthalene-degradation trait is conferred by plasmid pWH2-Nah in *P. fluorescens* WH2²⁰³ and NAH7 in *P. putida* G7²⁰². *P. putida* UWC1 received the pWH2-Nah plasmid by conjugation with *P. fluorescens* WH2, obtaining the ability to metabolise naphthalene. Figure 4.1C and 4.1D show SCRS of *P. putida* UWC1 (pWH2-Nah⁺) and *P. putida* G7 (NAH7) grown in naphthalene as the sole carbon source with different percentage of fully deuterated naphthalene- d_8 . Similar to deuterated glucose- d_{12} metabolism shown above, the intensity of the C–D band proportionally increased with an increasing percentage of deuterated naphthalene- d_8 in the growth medium. The control experiments of *P. putida* UWC1 (pWH2-Nah⁺) without naphthalene and *P. putida* UWC1 (pWH2-Nah⁻) with 100% naphthalene- d_8 displayed a flat and ‘silent’ C–D region between 2070 and

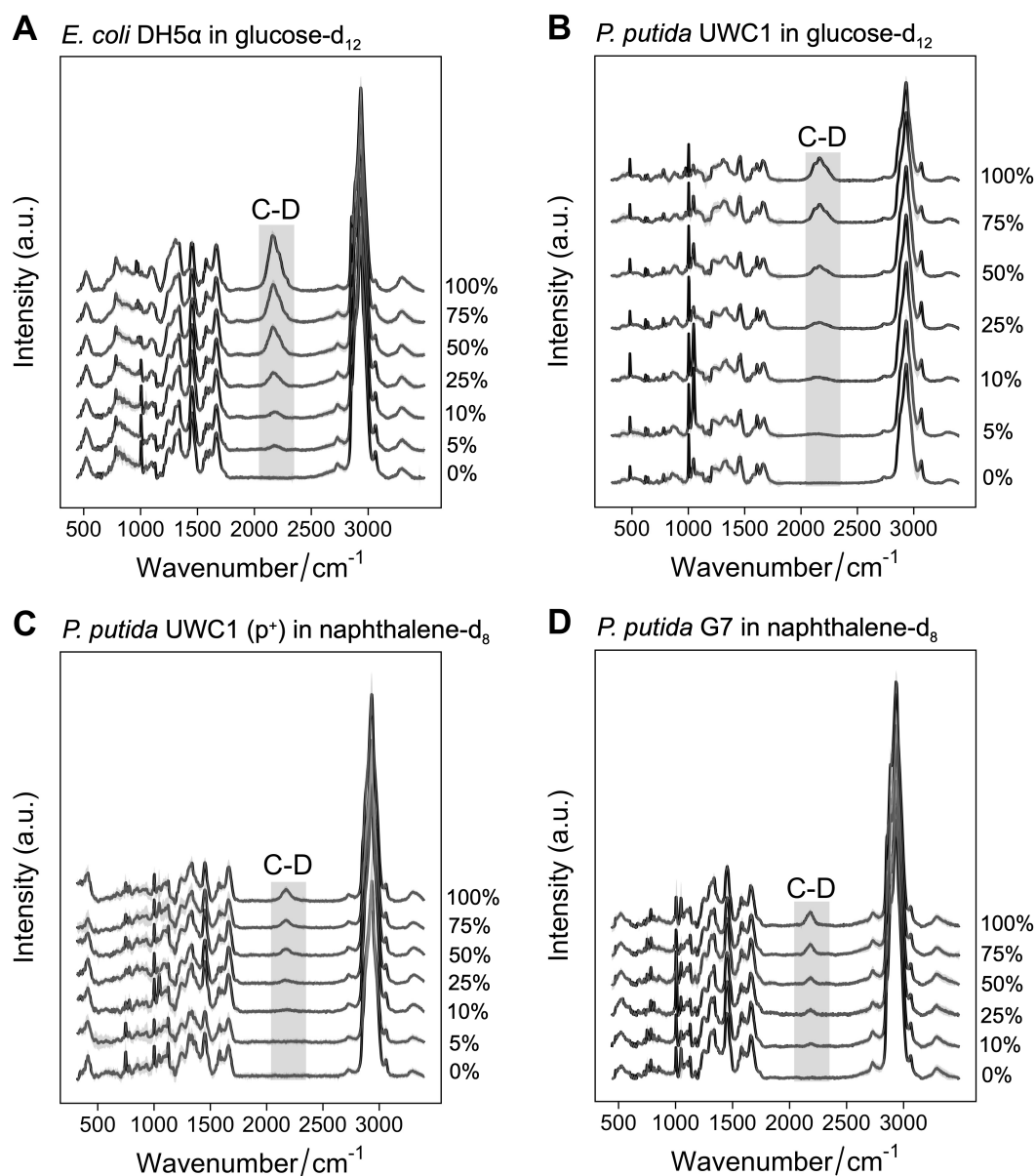


Figure 4.1: SCRS of (A) *E. coli* DH5α and (B) *P. putida* UWC1 grown in glucose with different ratios of deuterated glucose- d_{12} . (C) *P. putida* UWC1 ($p^{WH2-Nah^+}$) and (D) *P. putida* G7 grown in naphthalene with different ratios of deuterated naphthalene- d_8 . The characteristic C-D Raman band centred at 2170 cm^{-1} is highlighted.

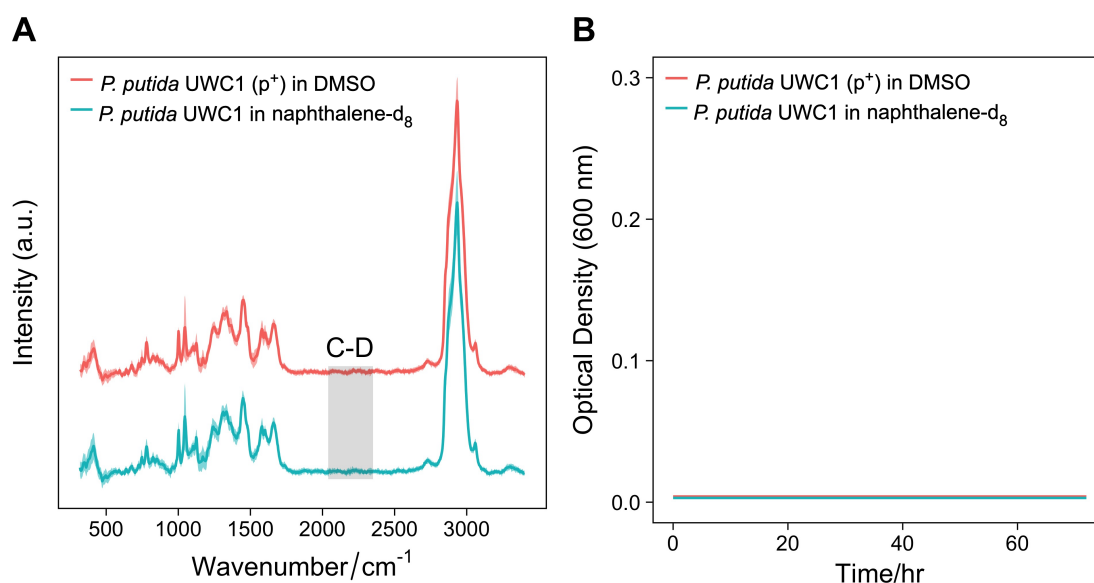


Figure 4.2: (A) SCRS of *P. putida* UWC1 (pWH2-Nah⁺) grown in DMSO without any naphthalene and *P. putida* UWC1 grown in 100% naphthalene-*d*₈, both of which have a silent Raman window between 2070 - 2300 cm⁻¹. (B) The growth curves of *P. putida* UWC1 (pWH2-Nah⁺) grown in DMSO without any naphthalene and *P. putida* UWC1 grown in 100% naphthalene-*d*₈, measured by OD₆₀₀.

2300 cm⁻¹ in their SCRS (Figure 4.2A) and no increase in OD₆₀₀ after incubating for 72 hours (Figure 4.2B), ruling out any solvent effect or abiotic H-D exchanges in the cells.

Deuterium and hydrogen are nearly chemically identical as deuteration has a negligible effect on atom size, molecular shape, equilibrium bond length or stiffness, apart from small vibrational effects²⁰⁷. Deuteration does, however, influence Raman positions for spectral modes in which the H/D atom is moving, and kinetics of reactions in which the H/D atom is involved. Therefore, we assume that, to some extent, bacterial cells are unable to distinguish between undeuterated carbon sources (glucose and naphthalene) and fully deuterated carbon sources (glucose-*d*₁₂ and naphthalene-*d*₈) in carbon metabolism.

Figure 4.3 shows the growth curves of microbes in the presence of different percentages of deuterated carbon sources measured by OD₆₀₀. The three *P. putida* strains UWC1, UWC1 (pWH2-Nah⁺) and G7 were unable to distinguish deuterated and undeuterated glucose or naphthalene and exhibited no growth differences among conditions from 0% to 100% deuterated glucose or naphthalene ($p > 0.01$) (Figure 4.3B, C and D). In Figure 4.3A, there was no significant observable effect on the growth kinetics

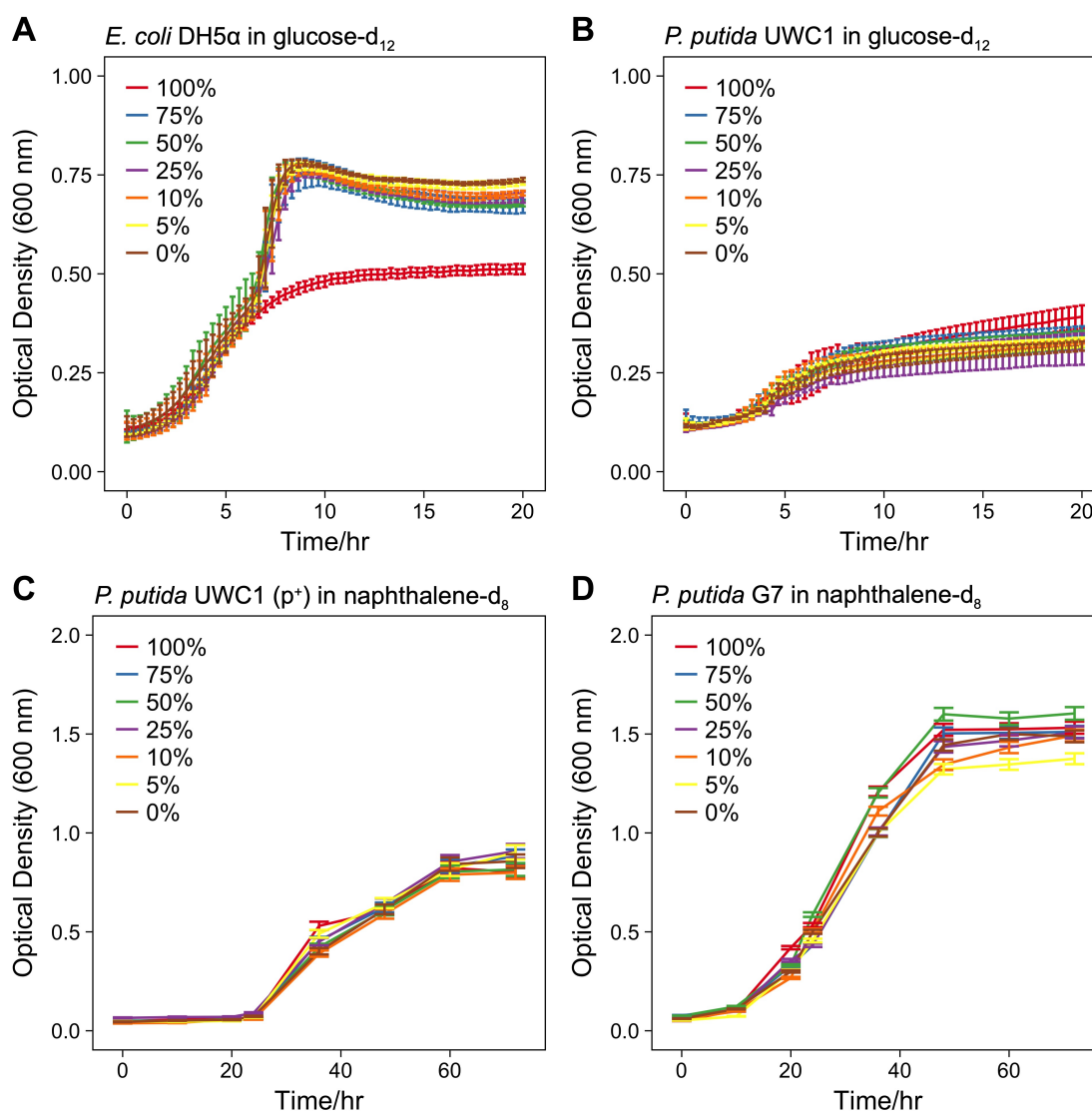


Figure 4.3: Growth curves created by OD₆₀₀ measurements of cultures of (A) *E. coli* DH5α and (B) *P. putida* UWC1 grown with different ratios of deuterated glucose-*d*₁₂, (C) *P. putida* UWC1 (pWH2-Nah⁺) and (D) *P. putida* G7 grown with different ratios of deuterated naphthalene-*d*₈.

of *E. coli* grown in 0–100% deuterated glucose in the initial 6 hours ($p = 0.058$). Although 0–75% deuterated glucose conditions featured indistinguishable growth curves, cells with 100% deuterated glucose exhibited a similar log phase but a subsequent repressed stationary phase, compared to in other less deuterated conditions.

Formation of C–D bonds in cells with a deuterated carbon source can be explained by the action of nicotinamide adenine dinucleotide phosphate (NADP⁺/NADPH). For example, the D in glucose-*d*₁₂ is passed to NADP⁺ to form NADPD instead of NADPH, which is then used to facilitate a new C–D bond during reduction reactions

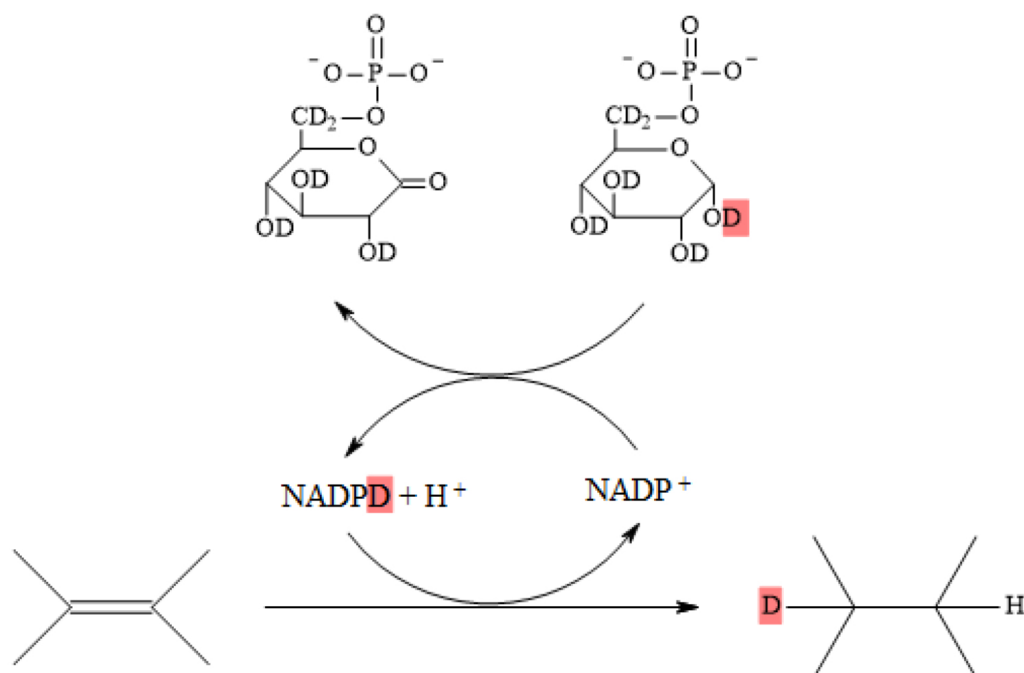


Figure 4.4: Biological mechanism of the incorporation of deuterium from deuterated carbon source, using glucose- d_{12} as an example. Deuterium in the deuterated carbon source contributes to a new C-D bond during the reduction reactions to synthesise biomolecules such as lipids via the action of $\text{NADP}^+/\text{NADPH}$.

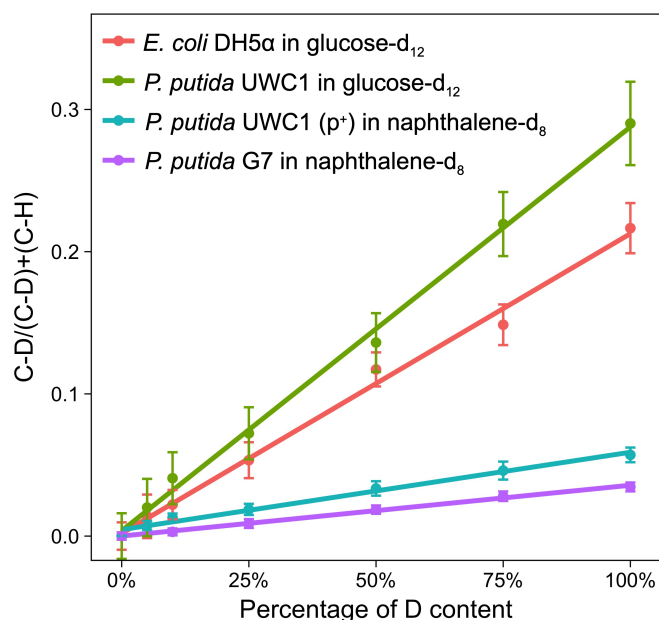


Figure 4.5: Comparison of the percentage of D content in carbon source with the band intensity ratio of C-D / (C-D + C-H) in SCRS of *E. coli* DH5α and *P. putida* UWC1 grown with glucose, and of *P. putida* UWC1 (p^+) and *P. putida* G7 grown with naphthalene, with different percentages of glucose- d_{12} or naphthalene- d_8 .

to synthesise cellular building blocks such as lipids (Figure 4.4). This C–D bond leads to a clearly distinguishable broad band at 2070 - 2300 cm^{-1} in SCRS.

Figure 4.5 shows that the intensity ratio of C–D / (C–D + C–H) in SCRS to the percentage of deuterated carbon source followed a linear relationship in all four cases, regardless of carbon source and bacterial species. The linearity suggests that bacteria incorporate D from deuterated carbon source into the cells, with an amount proportional to the percentage of the deuterated carbon source in the medium. This result is in good agreement with our hypothesis that cells cannot distinguish deuterated carbon sources from their unlabelled counterparts. The SCRS demonstrate a sensitive D detection limit at as low as 5% deuterated glucose ($p < 0.001$) and 5% deuterated naphthalene ($p < 0.01$) (Figure 4.5).

Theoretical calculations on the palmitic acid molecule as a typical aliphatic fatty acid were used to shed further light on the above spectral observations (Figure 4.6). With a reference to the molecular structure of palmitic acid (Figure 4.6A), Raman spectra were computationally generated for the molecular form with an all-H backbone (Figure 4.6B), an all-D backbone (Figure 4.6C), an half-H/half-D backbone (Figure 4.6D), as well as a backbone with a single-D substitution (Figure 4.6E).

In summary: the stretches of C–H/C–D bonds on the aliphatic backbone are found to be effectively decoupled from other molecular vibrations. Each D substitution in a CH_2 or CH_3 group of the backbone produces an extra contribution to the band in the C–D stretch region (2150–2350 cm^{-1} in the calculated spectrum, corresponding to an overestimation by a few percent of the vibrational force constant) and removes intensity from the C–H stretch region (2950–3150 cm^{-1} in the calculated spectrum), as predicted by the $\sqrt{2}$ rule arising from reduced mass effects (Equation 2.10 in Chapter 2), with the contribution of each C–H and C–D to band area being in approximately 2:1 ratio. The effects of D substitution are more or less independent of the number and position of substitutions on carbon atoms of the backbone. Thus, spectral intensities could be used to estimate C–D/C–H ratios here and in similar aliphatic environments.

Although lipids synthesis plays the most important part in the D incorporation at 2070–2300 cm^{-1} as it contributes roughly 20–35% to the total cellular C–H content¹⁴⁵,

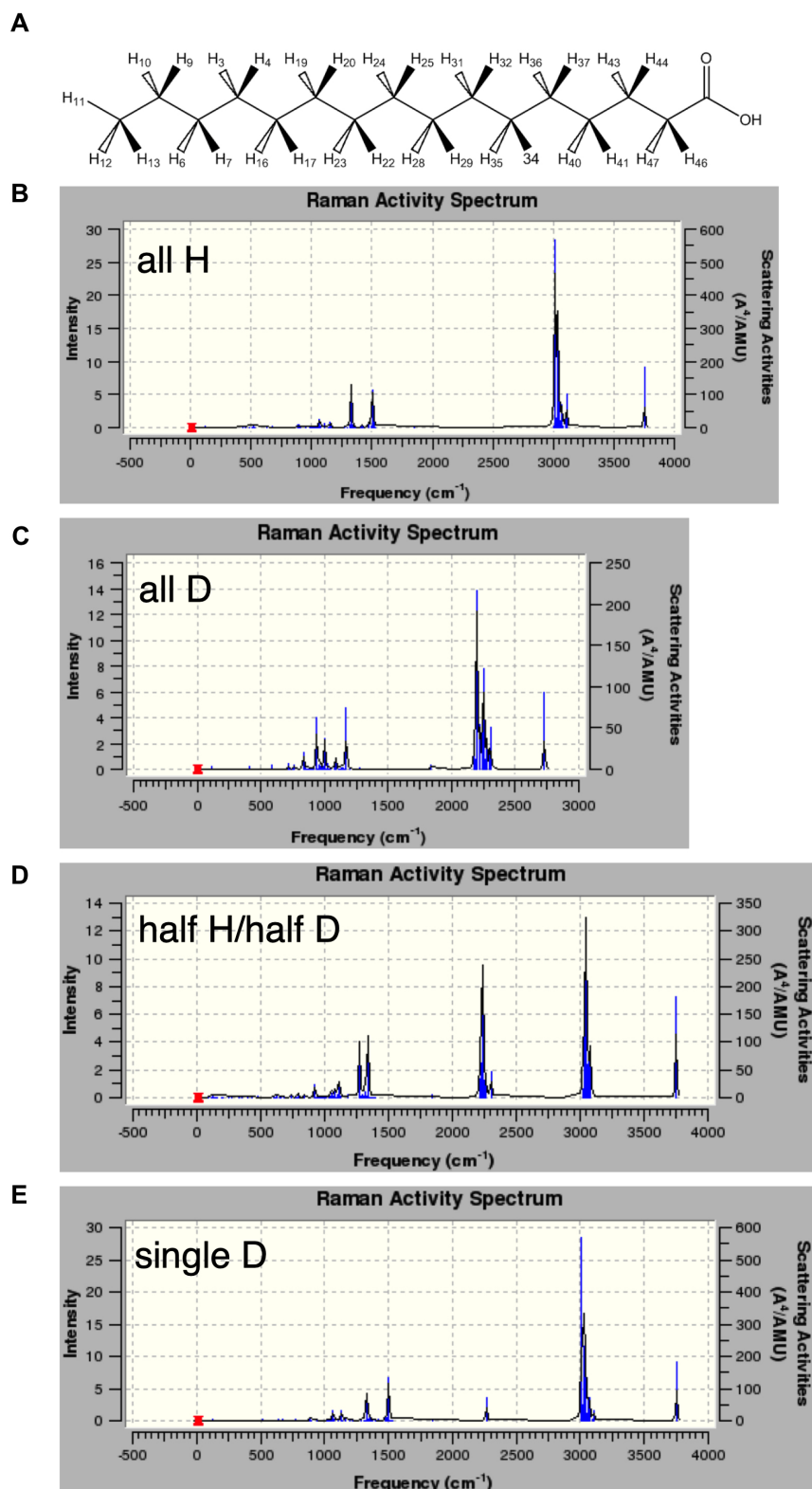


Figure 4.6: (A) Molecular structure of palmitic acid. The optimisation yielded a fully extended conformer, with mirror symmetry. (B)-(E). Simulated Raman spectra for palmitic acid as a simplified model for aliphatic CH environments in lipids. The spectra from top to bottom are calculated for (B) the all-H form of the molecule, (C) the all-D form, (D) a form in which all H positions to one side of the mirror plane are replaced by D and (E) a sample form of single-D substitution on the methyl group at the end.

theoretically the C–D Raman band can include any possible C–D vibrations in lipids, proteins, nucleic acids and carbohydrates (Figure 2.7 in Chapter 2). Our results suggest that this process is a ubiquitous metabolic process that can be sensitively detected by SCRS. Taken together, a deuterated carbon source can be used as a universal indicator to probe carbon utilisation.

4.3.2 SCRM-DIP reveals distinct biosynthesis pathways of fatty acids starting from different deuterated carbon sources

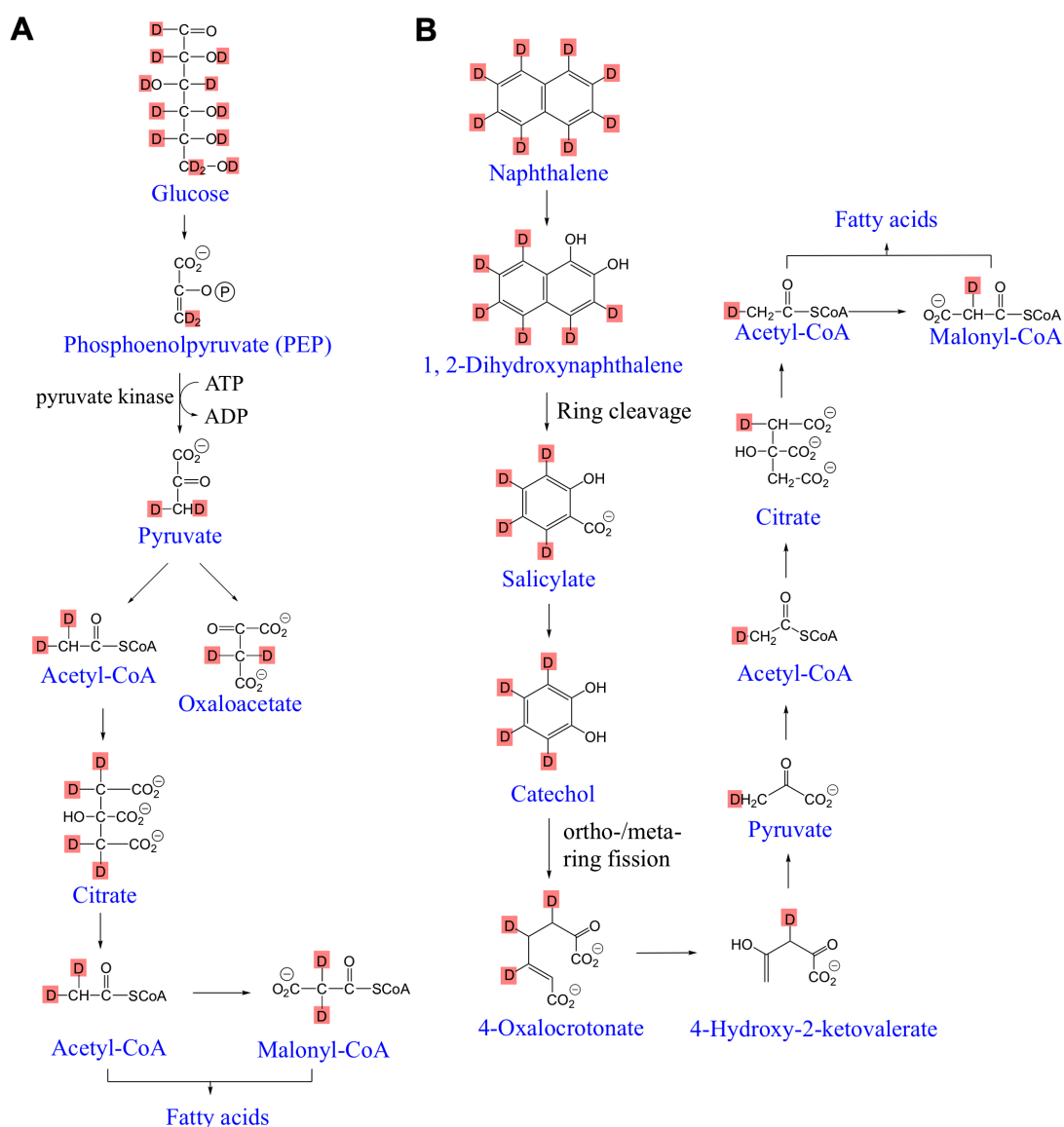


Figure 4.7: Fatty acids metabolic pathways starting from (A) glucose- d_{12} and (B) naphthalene- d_8 .

In Figure 4.5, despite showing a linear relationship in both glucose and naphthalene conditions, the slopes of the lines are significantly different ($p < 0.001$). When the cells were grown in 100% glucose- d_{12} as the sole carbon source, the ratio of C–D / (C–D + C–H) was 21.7% in *E. coli* and 29.0% in *P. putida* UWC1. However, when 100% naphthalene- d_8 was used as the sole carbon source, the ratio decreased greatly to 5.7% in *P. putida* UWC1 (pWH2-Nah⁺) and 3.5% in *P. putida* G7.

As lipid biosynthesis contributes roughly 20-35% to the total cellular C–H content¹⁴⁵, it plays a crucial part in the appearance of the broad C–D band through different lipid synthesis pathways degrading either glucose or naphthalene (Figure 4.7).

The glycolytic conversion of fatty acids from carbohydrates generates a common intermediate phosphoenolpyruvate (PEP) through both Embden-Meyerhof-Parnas (EMP) and Entner-Doudoroff (ED) pathways in *E. coli*²⁰⁸ and almost exclusively through the ED route in *P. putida*²⁰⁹ (Figure 4.7A).

The aerobic degradation of naphthalene, however, proceeds in more steps involving two-stage cleavages of aromatic rings in naphthalene²¹⁰ (Figure 4.7B). The first cleavage stage produces a central intermediate, salicylate, which is further converted to catechol. This is followed by the second stage which includes ring fission by either ortho-cleavage between the hydroxyl groups or meta-cleavage adjacent to one of the hydroxyls. The resulting intermediates then get into central metabolism to produce TCA cycle intermediates such as acetyl-CoA, succinyl-CoA, oxaloacetate and pyruvate. Hence, the pathway to make key metabolites for fatty acids synthesis via naphthalene metabolism in microbes involves more anabolic steps than the pathway via glucose metabolism (Figure 4.7A). A significant amount of D in the initial naphthalene- d_8 substrate was lost during the ring-cleavage reactions, which contributed to the lower intensity ratio of C–D / (C–D + C–H) observed in cells grown with deuterated naphthalene than in *P. putida* UWC1 grown with deuterated glucose (Figure 4.5). This pattern was not limited to *P. putida* UWC1, it was also true in other strains such as *E. coli* DH5 α grown in deuterated glucose and *P. putida* G7 grown in deuterated naphthalene (Figure 4.5), which collectively validates the mechanistic explanation in Figure 4.7.

In summary, the intensity ratio of C–D / (C–D + C–H) that is dependent on the degree of cellular H/D exchanges stems from different deuterated carbon sources, proceeds through various numbers of metabolic steps, and finally yields essential metabolites for lipid biosynthesis.

4.3.3 SCRM-DIP reveals distinct biosynthesis pathways of phenylalanine starting from different deuterated carbon sources

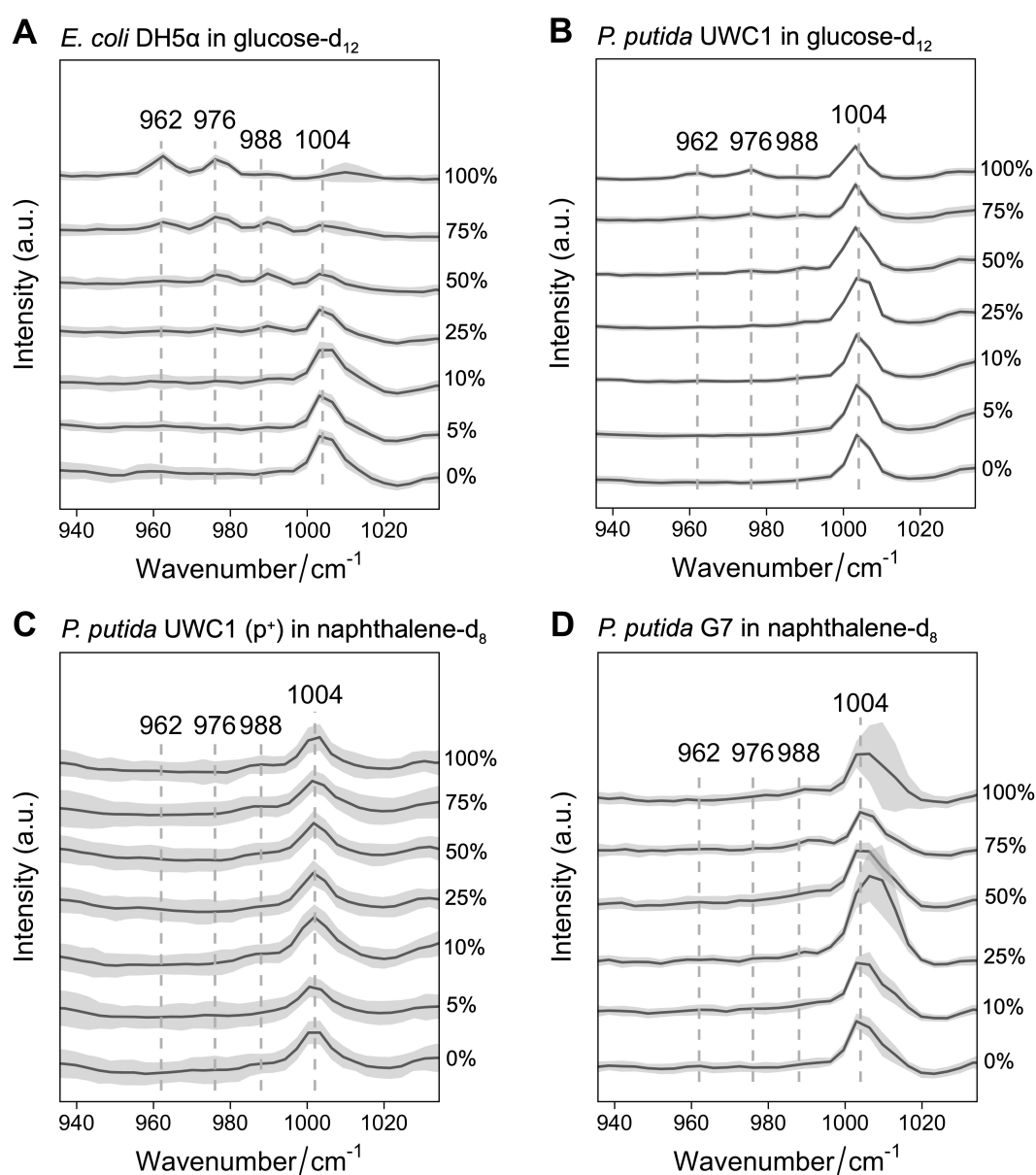


Figure 4.8: Enlarged view of phenylalanine shifts from 1004 cm^{-1} to 988 , 976 and 962 cm^{-1} in Figure 4.1 with respect to different ratios of deuterated growth glucose- d_{12} or naphthalene- d_8 .

In addition to the C–D spectral feature observed in SCRS at 2070 - 2300 cm^{-1} , isotopic shifts were also observed at 1004 cm^{-1} which can be assigned to the aromatic amino acid phenylalanine (Figure 4.1). Figure 4.8 highlights the Raman shift around 1004 cm^{-1} in a magnified view between 940 – 1020 cm^{-1} . Here we observe red shifts of band from 1004 cm^{-1} to 988, 976 and 962 cm^{-1} when cells were grown with deuterated glucose. This is in agreement with a previous study of *Geobacter metallireducens* grown with deuterated acetate, in which a splitting pattern of phenylalanine Raman band was observed¹²⁹.

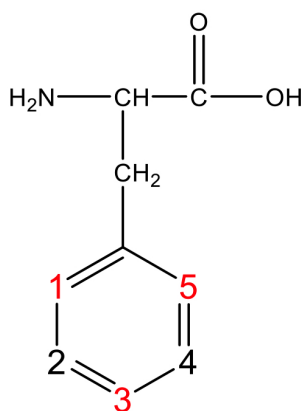


Figure 4.9: Chemical structure of phenylalanine with 5 C–H positions on the phenyl ring. Only the red-labelled 1, 3 and 5 sites of the symmetric ring-breathing mode are responsible for the observed phenylalanine Raman shifts.

Depending on the number of D substitutions on the five C–H sites on the phenyl ring, there are 32 possible isotopomers of phenylalanine (Figure 4.9 and Table 4.2). To understand the observed shifts of the phenylalanine band, we calculated the equilibrium geometry and Raman frequencies for the 32 isotopomers of phenylalanine (Table 4.2). One (ν_1) or two (ν_1 and ν_2) Raman wavenumbers were generated with intensities I1 and I2, respectively. The average of the wavenumbers was calculated as the theoretical Raman wavenumber. The data were grouped by mirror symmetry to account for the free rotation of the ring in real samples. Substitutions at only three sites on the phenyl ring (sites 1, 3 and 5) are found to affect the position of the vibration (Figure 4.9). Ruling out equivalents from free rotations along the symmetric axis

Table 4.2: The equilibrium geometry of 32 possible isotopomers of the phenylalanine molecule calculated at the B3LYP/6 31G** level of theory. $\nu 1$ and $\nu 2$ are two Raman wavenumbers generated with an intensity of I1 and I2, respectively. Equiv. represents any free rotation equivalents of a specific isotopmer. Averaged ν is an average from two wavenumbers, ruling out equivalents from free rotations along symmetric axis.

	12345	$\nu 1/\text{cm}^{-1}$	$\nu 2/\text{cm}^{-1}$	I1	I2	No. of D	Equiv.	Averaged ν/cm^{-1}
1	HHHHH	1017		24.33		0	-	1017
2	DHHHH	1002		24.4		1	6	1002
3	HDHHH	1017		24.68		1	5	1017
4	HHDHH	1003	1001	21.3	12.17	1	-	1002
5	HHHDH	1017		25.9		1		
6	HHHHD	1001		23.97		1		
7	DDHHH	1002		25.76		2	16	1002
8	DHDHH	988		29.92		2	15	989
9	DHHDH	1002		23.74		2	13	1002
10	DHHHD	988	988	17.65	8.87	2	-	988
11	HDDHH	1003	1001	21.07	10.15	2	14	1002
12	HDHDH	1017		26.26		2	-	1017
13	HDHHD	1001		23.67		2		
14	HHDDH	1002		31.12		2		
15	HHDDH	989		28.36		2		
16	HHHDD	1002		26.28		2		
17	DDDHH	988		29.44		3	26	989
18	DDHDH	1002		26.21		3	25	1002
19	DDHHD	989		27.51		3	22	989
20	DHDDH	989		28.55		3	24	989
21	DHDHD	974		28.93		3	-	974
22	DHHDD	988		26.61		3		
23	HDDDH	1003		29.77		3	-	1003
24	HDDHD	989		28.06		3		
25	HDHDD	1001		26.19		3		
26	HHDDD	989		28.29		3		
27	DDDDH	988	992	24.37	8.49	4	31	990
28	DDDHD	974		27.78		4	30	974
29	DDHDD	988		26.81		4	-	988
30	DHDDD	974		29.49		4		
31	HDDDD	989		28.12		4		
32	DDDDD	974		28.8		5	-	974

Table 4.3: The collected peaks from theoretical modelling in Table 4.2 and wavenumbers scaled by a factor of 1004/1017 for comparison with the experimental values, represented as scaled peaks.

No. of D	Collected peaks /cm ⁻¹	Scaled peaks /cm ⁻¹
0	1017	1004
1	1017	1004
	1002	989
2	1017	1004
	1002	989
	989	976
3	1003	990
	989	976
	974	962
4	989	976
	974	962
5	974	962

and equivalent wavenumbers, only 4 possible Raman positions of all H/D mixtures were collectively found at 1017, 1002–1003, 989 and 976 cm⁻¹ (Table 4.3). The ring breathing vibration at 1017 cm⁻¹ in the all-H structure was identified as the match for the experimental peak at 1004 cm⁻¹. All theoretical wavenumbers were then scaled by a factor of 1004/1017 for comparison with the experimental peaks and listed as scaled peaks in Table 4.3. This resulted in a total of 4 possible Raman positions at 1004, 989–990, 976 and 962 cm⁻¹ (Figure 4.10), for the 32 possible isotopomers of phenylalanine with different H/D ratios and distributions around the aromatic ring, which correlate well with the experimental band positions in Figure 4.8.

The 4 theoretical calculations of Raman wavenumbers were then mapped with the experimental observation of phenylalanine Raman shifts in SCRS of *E. coli* in Figure 4.8A, with regard to %D in the ring (i.e., 0, 20, 40, 60, 80 and 100%, depending on whether there was 0, 1, 2, 3, 4 or 5 D on the ring). A good correlation was found (Table 4.4). This result not only demonstrates the consistency between theoretical calculations and experimental observations, but also offers a good explanation for the differences in isotopic shifts concerning different D percentages: the form of the vibrational mode, in which the motion is a three-fold asymmetrical breathing of three C–H (or C–D) units against the rest of the molecular framework, with little participation

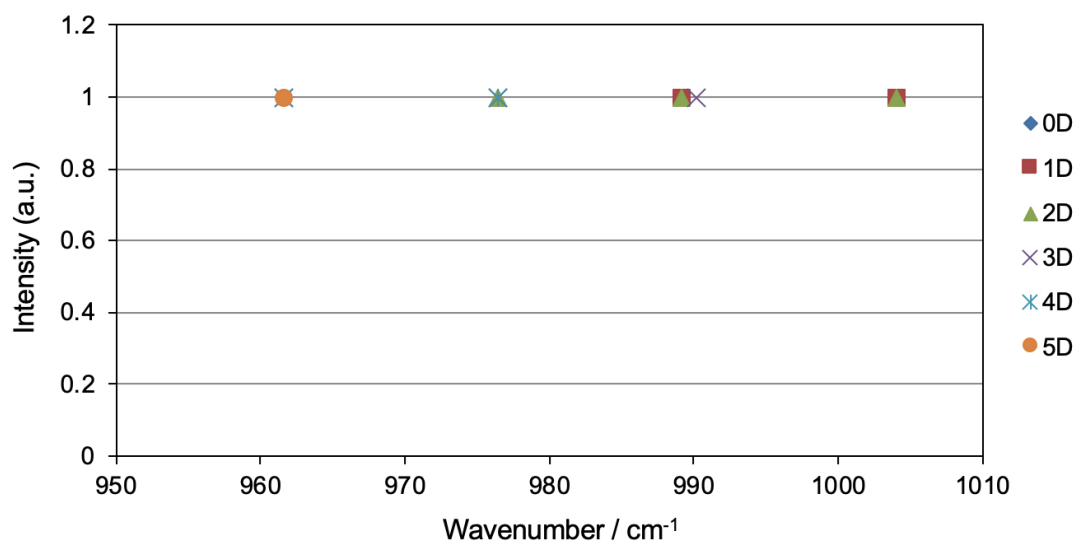


Figure 4.10: Four distinct Raman-band positions are found at 1004, 989–990, 976 and 962–963 cm^{-1} , for the 32 possible isotopomers of phenylalanine

Table 4.4: Calculated band positions correlate with the experimental observations of phenylalanine Raman shifts. Theoretical Raman band wavenumbers have been scaled by a factor of 1004/1017. Black crosses represent theoretical calculation and red crosses represent experimental shifts observed in SCRS of *E. coli* as an example.

% of D	1004 cm^{-1}	989–990 cm^{-1}	976 cm^{-1}	962 cm^{-1}
0	xx			
5	x			
10	x	x		
20	x	x		
25	x	x		
40	x	x	x	
50	x	x	x	
60	x	x	x	x
75	x	x	x	x
80		x	x	x
100		x	xx	xx

of the remaining ring C–H units. The band at 1004 cm^{-1} corresponds to structures where sites 1, 3 and 5 are H, and so can occur when some D is present in the sample. The band at $989\text{--}990\text{ cm}^{-1}$ occurs when any of the sites 1, 3 and 5 are D. The band at 976 cm^{-1} occurs when any two of the sites 1, 3 and 5 are D, and therefore only presents at higher D/H ratios. The bands at 962 cm^{-1} corresponds to structures where all sites 1, 3 and 5 are D, and hence can only appear at D/H ratios above 60%, which correlates nicely with the observation that the bands at 962 cm^{-1} were only present in cells with 75% or 100% deuterated glucose (Figure 4.8A and B).

The wavenumber shifts of roughly 3%, 6% and 9% from the all-H vibrations are consistent with the effect of changing the mass of the small units vibrating harmonically against a rigid framework. Each active D contributes additively to the higher effective reduced mass for the vibration and hence gives rise to a square root dependence in mode wavenumbers.

The degree of D incorporation into phenylalanine in Figure 4.8 was then quantified by calculating the ratio of total D to total H and D as follows, in which A represents the area under the curve centred at a defined wavenumber:

$$\frac{D}{D + H} = \frac{A_{962\text{ cm}^{-1}} + \frac{2}{3}A_{976\text{ cm}^{-1}} + \frac{1}{3}A_{989\text{ cm}^{-1}}}{A_{976\text{ cm}^{-1}} + A_{989\text{ cm}^{-1}} + A_{1004\text{ cm}^{-1}}}$$

High variability of the cellular D content at the same level of deuterated carbon source was observed regardless of carbon sources and or species. In Figure 4.11, linearity is found between D percentage and cellular D uptake in two glucose conditions. However, *E. coli* grown with 100% deuterated glucose- d_{12} displays a much higher D incorporation (58.1%) than *P. putida* UWC1 with the same carbon source (28.3%).

This difference can be explained by the less efficient utilisation of glucose in *P. putida* compared to *E. coli* (Figure 4.12A and B). The catabolism of glucose in *E. coli* DH5 α has been proven mainly via the EMT pathway and the phosphotransferase (PTS) system, in which glucose is phosphorylated to glucose 6-phosphate (G6P) and

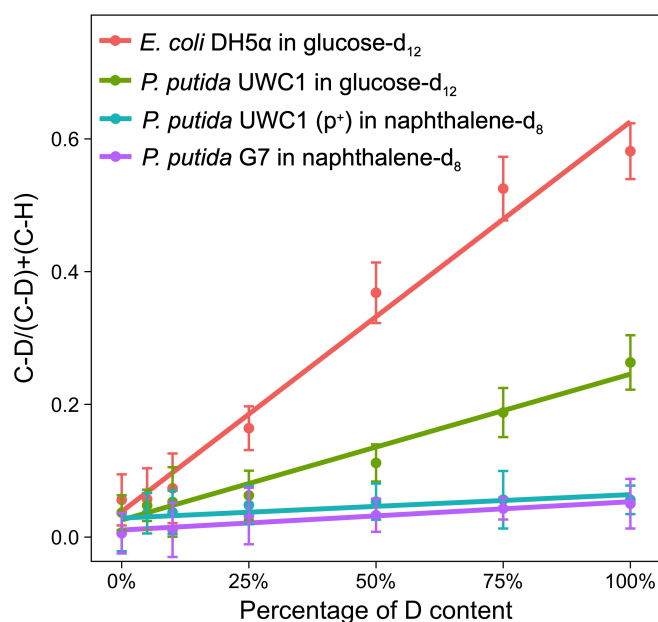


Figure 4.11: Comparison of D content in the phenylalanine band in SCRS of *E. coli* DH5α and *P. putida* UWC1 grown in glucose, and of *P. putida* UWC1 (pWH2-Nah⁺) and *P. putida* G7 grown in naphthalene, with different percentages of glucose- d_{12} or naphthalene- d_8 .

eventually converted to PEP for the synthesis of aromatic amino acids via the shikimate pathway²¹¹ (Figure 4.12A). In contrast, it is reported that bacteria of *Pseudomonas* spp. catabolise glucose almost exclusively through the ED pathway^{212,213} (Figure 4.12B). This pathway in *P. putida* involves three simultaneous sub-pathways starting with glucose, gluconate or 2-ketogluconate that subsequently converge at the level of 6-phosphogluconate (6PG)²¹⁴. 6PG is further converted to a key intermediate, 2-keto-3-deoxy-6-phosphogluconate (KDPG), which continues to yield the production of PEP. The interconversion and cellular transportation of glucose, gluconate and 2-ketogluconate in three convergent pathways might largely contribute to the D loss observed in the SCRS of *Pseudomonas* bacteria grown with deuterated glucose.

On the other hand, when deuterated naphthalene was used as the sole carbon source, no shifts of the phenylalanine band at 1004 cm^{-1} , presumably caused by D incorporation, were detected in *P. putida* UWC1 (pWH2-Nah⁺) and *P. putida* G7 cells (Figure 4.8C and D). These cells were unable to directly convert naphthalene into the shikimate pathway intermediates before entering central metabolism cycle (Figure

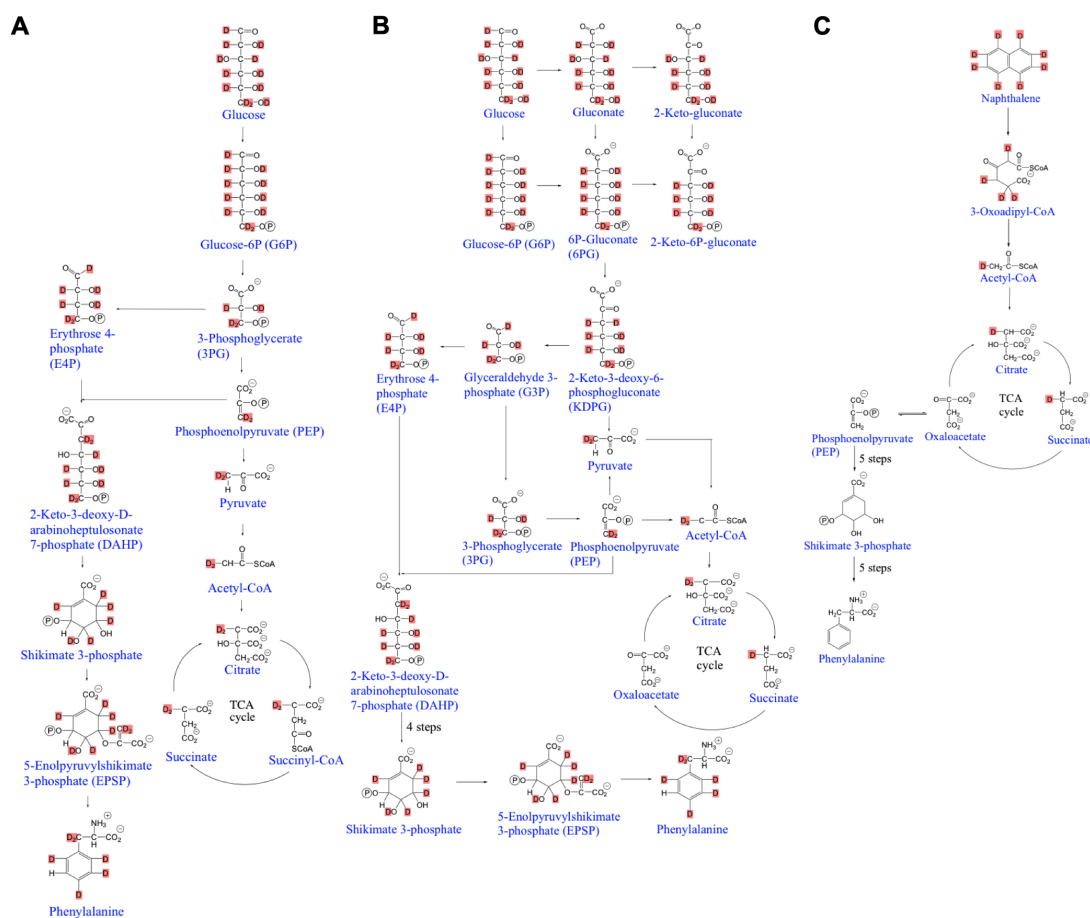


Figure 4.12: Phenylalanine metabolic pathways of (A) *E. coli* grown in glucose- d_{12} , (B) *P. putida* UWC1 grown in glucose- d_{12} , and (C) *P. putida* UWC1 (pWH2-Nah⁺) grown in naphthalene- d_8 .

4.11C). D in naphthalene was largely lost during the two-phase ring cleavage processes and further diluted in the TCA cycle before D was incorporated into phenylalanine. Therefore, D in naphthalene- d_8 was largely lost due to many dilution steps in the naphthalene-degradation pathway, while D was less diluted and remained in the intermediate metabolites to make phenylalanine in the glucose metabolic pathway (Figure 4.11). Together, these results suggest that SCRM-DIP could indicate relative numbers of dilution steps in metabolic pathways by analysing isotopic shifts of the phenylalanine band in SCRS, and therefore predicting metabolic pathways adopted by different microorganisms.

4.4 Conclusions

This study illustrates that SCRM coupled with DIP (SCRM-DIP) provides a culture-independent and non-destructive approach to probe metabolic pathways of carbon substrates at the single-cell level.

A clearly distinguishable C–D vibration band at 2070–2300 cm^{-1} was found in SCRS of *Escherichia coli* grown with deuterated glucose or *Pseudomonas* spp. grown with deuterated naphthalene as the sole carbon source. The intensity of the C–D band is proportional to the extent of deuteration in carbon source, and as little as 5% deuteration can be distinguished by an analysis of the SCRS. It suggests that SCRM-DIP using deuterated carbon sources could be used to semi-quantitatively and sensitively indicate carbon metabolism in microbes.

A lower lipid conversion rate from deuterated naphthalene compared to that from deuterated glucose was observed, presumably owing to different metabolic pathways to synthesise fatty acids from two two carbon sources.

Apart from the C–D band shift from C–H, SCRS also reveal three isotopic shifts of the phenylalanine band at 1004 cm^{-1} , of which the positions correlate well with a computational model. A reduction in phenylalanine deuteration in *Pseudomonas* compared to that in *E. coli* is found due to the dilution effect during phenylalanine biosynthesis in *Pseudomonas*.

Collectively, we demonstrate that SCRM-DIP can not only indicate metabolic activity using deuterated carbon sources but also reveal different metabolic pathways by analysing SCRS. Providing a high variety of possible isotopic substrates, SCRM-DIP has the potential to probe cell activity and cellular metabolism of a broad range of important biomolecules including proteins, lipids, carbohydrates and nucleic acids.

Chapter 5

Application of SCRM-DIP in dynamic mapping of intracellular *Salmonella* metabolism

Contents

5.1 Introduction	89
5.2 Materials and Methods	90
5.3 Results and Discussion	97
5.3.1 Phenotypic profiles based on SCRS successfully distinguish closely related <i>Salmonella</i> strains	97
5.3.2 Intracellular <i>Salmonella</i> shows distinctive metabolic activities revealed by SCRM-DIP	102
5.3.3 SCRS map dynamic phenotypic profiles of intracellular <i>Salmonella</i> within primary cells	108
5.3.4 Single-cell Raman analysis resolves phenotypic heterogeneity in isogenic populations	113
5.3.5 Intracellular Raman profiles are consistent with the proteomic analysis	116
5.4 Conclusions	122

In the last two chapters, I have shown that SCRS can be used as phenotypic profiles of single bacterial cells and have applied the methodology of SCRM-DIP to probe bacterial general activity and carbon metabolism. In this chapter, I will apply SCRM-DIP in a more complex example in which intracellular pathogen meets host human cells. SCRM-DIP will be applied to measure the metabolic activity of single *Salmonella* cells during infection in human primary cells. For the first time, reverse labelling will be applied to intracellular infectious bacteria. By comparing and mapping the activities and metabolic profiles of 3 *Salmonella* strains in a dynamic and time-resolved manner, I will highlight SCRM-DIP as a novel tool to detect the mechanisms adopted

by intracellular pathogens to regulate its metabolic pathways in response to host cell during the early intracellular stage.

5.1 Introduction

Salmonella enterica (*S. enterica*) is a Gram-negative, facultative anaerobic, intracellular pathogen, responsible for a wide range of diseases, such as gastroenteritis, bacteremia and typhoid fever. *Salmonella* accounts for 16 million cases of typhoid fever, 1.3 billion cases of gastroenteritis and 3 million deaths per year²¹⁵. In humans, the outcome of *S. enterica* infection mostly depends on the infecting serovar. Infection with human-restricted typhoidal serovars, such as *S. enterica* serovar Typhi (*S. Typhi*) and Paratyphi (*S. Paratyphi*), results in typhoid fever; whilst non-typhoidal broad host range serovars, such as *S. enterica* serovar Typhimurium (*S. Typhimurium*), are usually responsible of localised self-limiting gastroenteritis²¹⁶.

Nevertheless, highly invasive multi-drug resistant *S. Typhimurium* serovars have emerged in sub-Saharan Africa, as a major cause of lethal bacteremia in malnourished anaemic children and individuals affected by underlying diseases such as malaria or HIV^{217–222}. *S. Typhimurium* isolates from Kenya and Malawi are predominantly of a novel multi-locus sequence type (ST) named ST313²²¹. The *S. Typhimurium* ST313 pathovar D23580 (STM-D23580) is a representative blood-stream clinical isolate which demonstrates genome degradation, resembling that of the human restricted pathogen *S. Typhi*. STM-D23580 has been largely used in studies of host-pathogen interactions to gain deeper insights into its increased pathogenicity, named as invasive non-typhoidal salmonellosis (iNTS).

In order to survive and replicate within host cells, *Salmonella* undergoes dramatic metabolic changes to utilise the available intracellular nutrients and to overcome adverse endosomal conditions (e.g. high osmotic pressure, acidic pH and poor oxygen availability)^{223,224}. These adaptations, alongside with genetic variances, might contribute to the different disease outcomes upon *Salmonella* infections.

Technological perspectives have emerged to obtain a comprehensive picture of the intracellular pathogen lifestyle. Time-course transcriptomic²²⁵ and proteomic

studies^{226,227} give insights into the mechanisms adopted by *Salmonella* to regulate its metabolic pathways in response to host cell environment. However, these technologies rely on highly expressed genes and proteins, thus overseeing subtle metabolic changes and a crucial fraction of non-enzymatic metabolic reactions²²⁸. Finally, while most of these investigations provide a bulk average measurement of the total bacterial population, they neglect the astonishing customised single-bacteria heterogeneity related to the surrounding environment^{229–231}.

In this study, we applied SCRM-DIP and reverse labelling to obtain a comprehensive and time-resolved picture of the intracellular metabolism of three *Salmonella* strains responsible for different clinical outcomes: the gastroenteritis-associated *S. Typhimurium* LT2 (STM-LT2), the non-typhoidal invasive *S. Typhimurium* D23580 (STM-D23580) and the typhoidal *S. Typhi* Ty2 (ST-Ty2). The results from single-cell Raman analysis were further validated by proteomics using high sensitivity mass spectrometry (MS). To our knowledge, this is the first research where SCRM-DIP has been successfully applied to disclose metabolism characteristic of invasive human-restricted *Salmonella* serovars.

5.2 Materials and Methods

Bacterial strains and culture conditions. *Salmonella enterica* serovar Typhimurium (STM) strain LT2 (ATCC 700220) and the clinical isolate STM-D23580, were used as representative non-typhoidal *Salmonella* sequence type 19 (ST19) and type 313 (ST313), respectively. Strain D23580 was isolated from the blood of an HIV-negative Malawian child with malaria and anemia. *Salmonella enterica* serovar Typhi (ST) strain Ty2 (ATCC 700931) was used as representative typhoidal serovar. *Escherichia coli* strain DH5 α (EC-DH5 α) (ATCC 67878) was used as an external reference for comparison with *Salmonella* strains.

The ability of *Salmonella* to incorporate D₂O was tested by growing bacteria overnight in LB Lennox broth (Sigma, UK) supplemented with 35% D₂O (v/v), at 37 °C under continuous shaking at 200 rpm. The degree of D₂O incorporation by STM-LT2 was compared with EC-DH5 α , used as a reference. Assimilation of intracellular

deuterium (D) was tested under different experimental conditions (before and after cryopreservation at -80°C ; with or without the presence of 10% sucrose) to test the robustness of the SCRM-DIP method.

For the infection experiments, STM-LT2, STM-D23580 and ST-Ty2 were grown to logarithmic growth phase in LB Lennox broth (Sigma) supplemented with 10% sucrose (Sigma) and 50% D_2O . Aliquots were kept frozen at -80°C until use, whilst bacterial viability was monitored periodically. For each experiment, an aliquot of bacteria was thawed, washed and resuspended in RPMI 1640 (Sigma) to obtain a multiplicity of infection (MOI) of 30:1 (bacteria:THP-1), 15:1 (bacteria:macrophage) or 10:1 (bacteria:dendritic cell). As a control for bacterial extracellular growth, RPMI was inoculated with each strain at 37°C under continuous shaking at 200 rpm. At specific time points, corresponding to the kinetic of the infection, bacteria were collected, washed and processed for the SCRS measurements and analysis, as described below. The number of microorganisms was assessed at each time point post infection (p.i.) by plating 10-fold dilutions of the bacterial suspension, in quadruplicate, on LB Lennox agar (Sigma). The number of bacteria was determined as colony-forming units (CFU).

Propagation of THP-1 cells and infection. THP-1 cells were obtained from the American Type Culture Collection (ATCC, TIB-202). The cells were maintained in vented T75 flasks (Cellstar) at 37°C , 5% CO_2 and split every 2-3 days when confluent (2×10^6 cells/mL). One day before the infection, THP-1 cells were harvested, washed, resuspended at a density of 1×10^7 cells/mL/well in 24-well plates (Costar) in complete RPMI medium containing 10% fetal bovine serum (FBS, Sigma), 2 mM L-Glutamine, 100 $\mu\text{g}/\text{mL}$ of penicillin and 100 $\mu\text{g}/\text{mL}$ streptomycin (Sigma), and supplemented with 25 ng/mL of phorbol 12-myristate 13-acetate (Sigma) to induce differentiation into macrophage-like cells, while incubating at 37°C , 5% CO_2 for 24 hours. Two hours before infection, the medium was replaced with fresh complete medium without antibiotics and cells were infected with D-labelled STM-LT2, STM-D23580 or ST-Ty2 at an MOI of 30 bacteria per cell. After 30 min incubation at 37°C , extracellular bacteria were removed by washing cells twice with phosphate-buffered saline (PBS) and further incubated for 30 min in RPMI supplemented with 10 % FBS, 2mM L-glutamine and

100 µg/mL of gentamycin (MP Biomedicals) to kill extracellular bacteria. After that, the gentamycin concentration was reduced to 30 µg/mL for the rest of the experiment.

Generation of monocyte-derived dendritic cells and macrophages and infection.

Leukocyte Reduction System cones were obtained from the UK National Blood Centre with informed consent following local ethical guidelines. Blood from healthy donors was diluted in PBS and layered on a standard density gradient (Lymphoprep™). Peripheral blood mononuclear cells (PBMCs) were collected from the interface and washed three times in PBS at 4 °C. Monocytes were obtained by magnetic isolation using human CD14+ MicroBeads (Miltenyi Biotec, Germany) according to manufacturer's protocol.

Monocyte-derived dendritic cells (MoDCs) and macrophages (Mf) were obtained by incubation of freshly isolated monocytes in complete medium in the presence of 40 ng/mL of recombinant human granulocyte macrophage-colony-stimulating factor (GM-CSF) plus 40 ng/mL rh Interleukin-4 (IL-4) or 100 ng/mL of recombinant human macrophage-colony-stimulating factor (M-CSF) (Peprotech), respectively, at 37 °C in humidified atmosphere at 5% CO₂.

Cells differentiating into MoDCs were plated in tissue culture dishes (Falcon) at a density of 1×10^6 cells/mL, while cells differentiating into Mf were plated in 24-well plates (Costar) at a density of 5×10^5 cells/mL.

After 5 days of culture, MoDCs were harvested, washed in PBS, resuspended in complete medium without antibiotics and seeded in polypropylene tubes (Falcon) at a density of 1×10^6 cells/tube. MoDCs were infected with D-labelled STM-LT2, STM-D23580 or ST-Ty2 at MOI of 10 bacteria per cell. Bacterial contact was maximised by 5 min centrifugation at 500 g at room temperature. After incubation for 45 min at 37 °C, extracellular bacteria were removed by washing cells with PBS. MoDCs were then incubated for 30 min in RPMI supplemented with 10% FBS, 2 mM L-glutamine and 100 µg/mL of gentamycin to kill extracellular bacteria. After that, the gentamycin concentration was reduced to 30 µg/mL for the rest of the experiment.

After 6 days of culture, Mf were washed with PBS and 500 µL of complete medium without antibiotics was added to each well. Mf were infected with D-labelled STM-LT2,

STM-D23580 or ST-Ty2 at an MOI of 15 bacteria per cell. After incubation for 30 min at 37 °C, extracellular bacteria were removed by washing cells twice with PBS. Mf were then incubated for 30 min in RPMI supplemented with 10% FBS, 2mM L-glutamine and 100 µg/mL of gentamycin to kill extracellular bacteria. After that, the gentamycin concentration was reduced to 30 µg/mL for the rest of the experiment.

At each time point, eukaryotic cells were lysed by an addition of 500 µL of saponin 2% (w/v) (Sigma) in PBS, followed by 5 min incubation at 37 °C. Cell lysates were used for SCRS (see below) or were serially diluted 10-fold in PBS and plated onto LB agar (LB Lennox Broth with agar; Sigma) to count the number of intracellular bacteria as CFU.

SCRS measurements and analysis. At specific time point post infection (p.i.), bacteria were grown in RPMI and cell lysates of THP-1, MoDCs and Mf were washed with PBS, fixed in 4% paraformaldehyde (PFA, EMS) in PBS (v/v) at room temperature for 15 min and washed twice with Milli-Q water. Samples were then diluted to a degree where single bacteria could be observed under the microscope and spotted onto a microscopic slide with aluminium coating to minimise background signal. Individual cells were observed under a 100×/0.8 microscope objective. SCRS were acquired using an HR Evolution confocal Raman microscope (Horiba Jobin-Yvon, France) equipped with a 532 nm neodymium-yttrium aluminium garnet laser. The laser power was attenuated to 4.7 mW by neutral density (ND) filters. An objective with a magnification of 100× was used to focus the laser beam on single cells with a laser spot size of $\sim 1 \mu\text{m}^2$ and Raman scattering was detected by a charge coupled device (CCD) cooled at -68°C . The spectra were acquired in the range of 500 to 3500 cm^{-1} with a 300 grooves/mm diffraction grating. The acquisition time was 10 s per spectrum and at least 30 single bacterial cells were acquired per each condition.

Raman data pre-processing and analysis. All spectra were recorded and pre-processed by cosmic ray correction and polyline baseline fitting using LabSpec 6 (Horiba Scientific, France). Spectral normalisation was done by vector normalisation of the entire spectral region. Data analysis, statistics and visualisation were done under an R environment using in-house scripts. Each averaged spectrum is an average from 30-50 single-cell measurements and the shaded area in SCRS represents the

standard deviation. Quantification of intracellular biomolecule and D content was done by integrating corresponding Raman bands in SCRS. Bands associated with carbohydrates I (382 - 470 cm^{-1}), carbohydrates II (510 - 600 cm^{-1}), adenine (705 - 736 cm^{-1}), DNA/RNA (765 - 795 cm^{-1}), tyrosine (843 - 869 cm^{-1}), phenylalanine (985 - 1015 cm^{-1}), carbohydrates III (1020 - 1070 cm^{-1}), amide III (1208 - 1275 cm^{-1}), lipid I (1275 - 1357 cm^{-1}), lipid II (1425 - 1470 cm^{-1}), amide I (1630 - 1710 cm^{-1}), saturated C-H (2880 - 2950 cm^{-1}) and unsaturated C-H (3040 - 3070 cm^{-1}) were integrated to quantify intracellular biomolecule concentration. The C-D (2070 - 2300 cm^{-1}) and C-H (2800-3030 cm^{-1}) band areas were calculated to determine deuterium content originating from incubation with D_2O . The ratio of C-D / (C-D + C-H) was used to indicate the extent of deuterium incorporation and this ratio of three *Salmonella* strains was mean-normalised at $t = 0$ hr. Quantification results were represented as box plots or single-cell heatmaps. Donor-wise centring has been done in all analysis which rendered any phenotypic heterogeneity due to metabolically different bacteria and not due to differences among donors.

Machine learning classification model. A machine learning classification model was built based on the fingerprint region (400 - 1800 cm^{-1}) of the SCRS of inoculum STM-LT2, STM-D23580 and ST-Ty2 (116, 152, 125 SCRS respectively). The model was based on k-Nearest-Neighbor (kNN) algorithm which is a simple machine learning classifier based on Euclidean distance between individuals. Leave-one-out-cross-validation (LOOCV) was used to evaluate the performance of the model and performance measures were computed as sensitivity for each strain as well as an overall accuracy rate.

Sample preparation for proteomics and liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis. For proteomic experiments, human MoDCs were infected with *Salmonella* strains as described above. Two biological and three technical replicates were performed in the proteomics experiments. At specific time points p.i., MoDCs were lysed by addition of 500 μL of saponin 2% (w/v) in PBS followed by 5 min incubation at 37 $^{\circ}\text{C}$, and bacterial pellets were resuspended in 20 mM ammonium bicarbonate (ABC; Sigma-Aldrich, St. Louis, MO, USA). The cell

solution was then supplemented with 50% trifluoroacetic acid (TFA; Uvasol, Merck, Darmstadt, Germany) to reach a final concentration of 1%. After 15 min of incubation at 37 °C pH was adjusted to pH 7.0 using 0.5 M ABC. Sequencing grade modified trypsin (0.3 µg; Promega, Madison, WI, USA) was added and incubated for 16 h at 37 °C. Tryptic digestion was stopped by acetic acid (RotiPuran; Carl Roth GmbH + Co. KG, Karlsruhe, Germany) at a final concentration of 1%. The suspension was centrifuged at 13,000 g for 10 min at 4 °C. Supernatant was collected and purified using ZipTip C18 (Millipore, Schwalbach, Germany). Purified tryptic peptides were dried by lyophilisation, dissolved in Buffer A consisting of 2% acetonitrile (J.T. Baker; part of Fisher Scientific, Waltham, MA, USA), 0.1% acetic acid (Roth) in water (J.T. Baker), and subjected to LC-MS/MS analysis.

The procedure for the lysis of bacteria cells is different to common preparation methods for *in vitro* experiments using urea or detergents. Due to the high pathogenicity of *S. Typhi* we optimised the way to safely kill all bacteria in BLS-3 area in order to continue with the peptide preparation outside. Previous studies have shown a strong lytic effect of low pH conditions on gram negative bacteria which in turn kills them rigorously^{232,233}. Tests with *S. Typhi* incubated with 1% TFA for 15 min did not show any further bacterial growth and after neutralisation, trypsin was efficiently working. *E. coli* was also treated under identical conditions using a low number of bacteria and was analysed in MS. Using this method, we could identify a sufficient number of proteins and we did not find any significant peptide modifications caused by the use of TFA. Due to the neutralisation with ABC after 1% TFA treatment, the liquid volume increased and a further addition of additional compounds would lead to very high volumes. Additional precipitation steps would be necessary and would yield to a loss of proteins/peptides. In addition, any use of e.g. urea leads to unspecific artificial modification on a various number of peptides^{232,233}. This leads to a split of the primary peptide intensity into more variances with lower intensities and further to less identifications, because only a limited number of bacteria and proteins/peptides were available. We therefore tried to limit the chemical treatments to a minimum.

MS data analysis, protein identification, quantitation and analysis. LC-MS/MS analyses of tryptic peptides were performed using an UltiMate 3000 RSLC (Thermo-Fisher Scientific, Idstein, Germany) coupled to a Q Exactive plus mass spectrometer (Thermo-Fisher Scientific) via a Nanospray Flex Ion Source (Thermo-Fisher Scientific). Due to the very low number of bacteria isolated, a classical pre-normalisation by total protein concentration determination could not be applied. Similar to previous *in vivo* studies on *Staphylococcus aureus*²³⁴, all available bacteria cells were used to get a maximum of proteins per sample. In order to get equal amounts in MS, we used a 2-mg ZipTip bed to roughly equalise the peptide amounts before loading them on high performance liquid chromatography (HPLC) column.

Raw data files from the LC-MS/MS runs were analysed using MaxQuant²³⁵ version 1.5.3.8. The analysis was performed in strain-specific batches (10 files each for STM-D23580 and ST-Ty2, 9 files for STM-LT2). Protein identification for STM-LT2 and STM-D23580 was based upon a *Salmonella* FASTA database file which contained 4,547 target entries (*S. Typhimurium* strain LT2/SGSC1412/ATCC 700720, UniProt; proteome ID UP000001014; November 28, 2017). A FASTA database with 4,295 *S. Typhi* entries (*S. Typhi* strain Ty2/ATCC 700931, UniProt, proteome ID UP000002622; February 27, 2018) was applied for the ST-Ty2 MS data.

In all MaxQuant analyses, parameters were set to specific Trypsin/P digestion with maximal 2 missed cleavages. Oxidation of methionine, oxidation of cysteine, deamidation of asparagine, and N-terminal acetylation was chosen as variable modifications with at most 5 modifications allowed per peptide. Fixed modifications were not set. The peptide mass was limited to 9,900 Da. Label-free quantification (LFQ) with an LFQ minimal ratio count of 2 was selected. Protein identification was supported by the alignment between the samples of each analysis batch using the “match between runs” feature with a match time window of 0.7 min and an alignment time window of 20 min. The false discovery rate (FDR) was controlled using a target-decoy approach and limited to 1%. A post-measurement statistical normalisation was done by default in LFQ calculation in MaxQuant. It is based on normalization introduced by Cox et al. in 2014²³⁶. This normalisation approach considers that most proteins do not

change between different samples of the same organism, so an average behaviour is used as a standard for normalisation. Supporting in this thesis 602 proteins were found in all three analysed conditions.

Quantification was performed on normalised data and a minimum ratio count of 2, with unique and razor peptides, including unmodified as well as modified peptides. Other parameters were set to default values. Resulting data of protein identifications were filtered prior to analysis of biological content. Contaminants, false-positive (reverse) hits and protein groups containing more than single proteins were removed.

The following analysis was done in an R environment. Log2 transformed LFQ intensities were calculated and pairwise comparisons of post-infection time points 0.75 hr, 2 hr, 4 hr and 6 hr (6 pairs in total) were done using empirical Bayes approach in LIMMA package²³⁷. Proteins with $p < 0.05$ (moderated t-test) at any pairwise comparison were determined as differentially expressed proteins (DEPs). DEPs were then functionally annotated using Biocyc database²³⁸. The proteins related to energy and biosynthesis pathways were visualised by dot plots and heatmaps to observe changes in the expression level of proteins during courses of infection.

5.3 Results and Discussion

5.3.1 Phenotypic profiles based on SCRS successfully distinguish closely related *Salmonella* strains

The fingerprint region ($400\text{--}1800\text{ cm}^{-1}$) can be used as a phenotypic profile to discriminate closely related bacteria based on their metabolic states. In order to understand whether the three genetically close *Salmonella* strains (STM-LT2, STM-D23580 and ST-Ty2), yet responsible of distinct clinical manifestations in humans, can be discriminated by their metabolic profiles, SCRS of bacterial cultures were obtained and analysed.

Figure 5.1A shows the SCRS of STM-LT2, STM-D23580 and ST-Ty2, averaged from 30–50 single-cell measurements of each strain, where the shaded area represents the standard deviation from single-cell measurements. Wavenumbers and corresponding

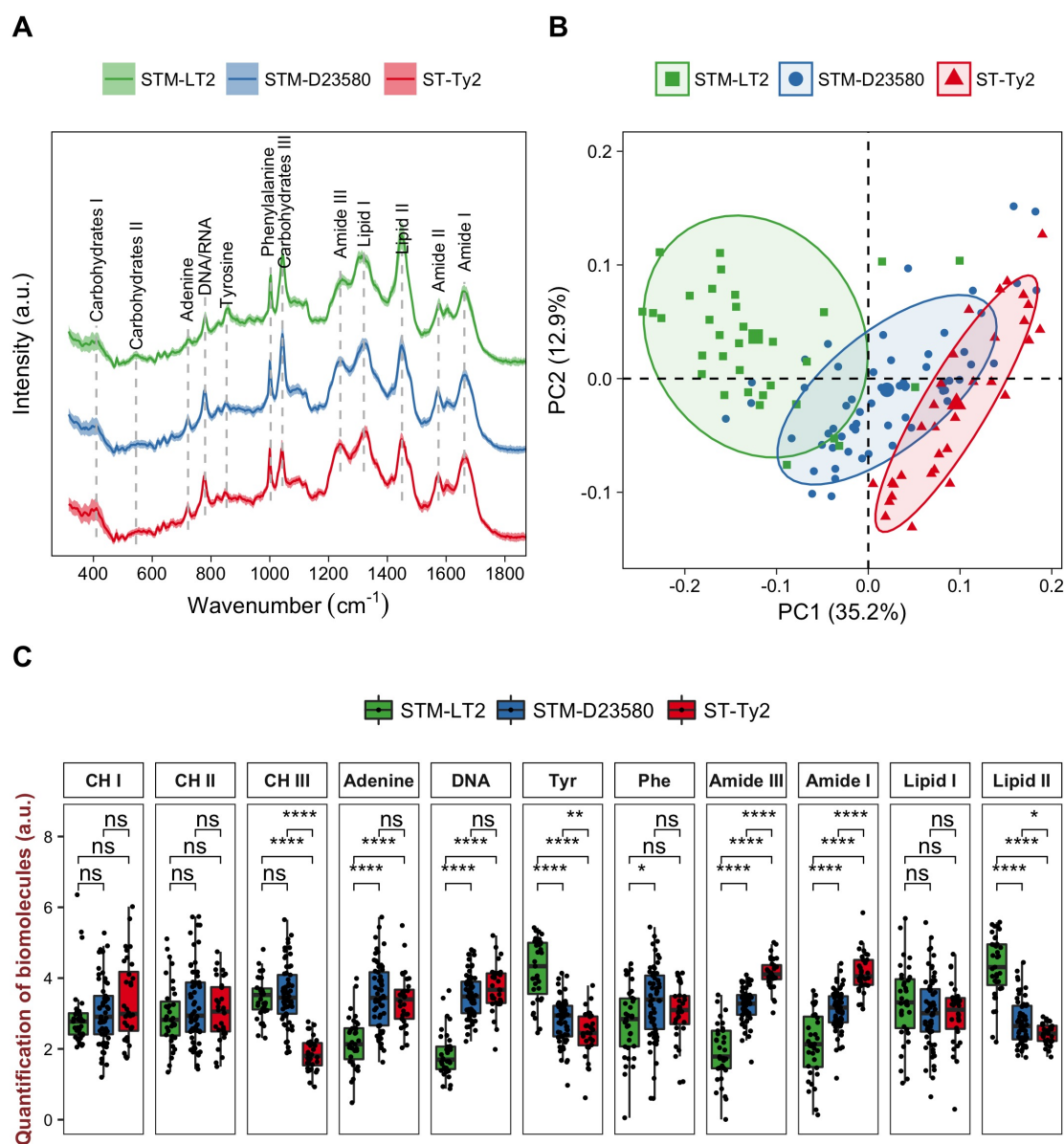


Figure 5.1: (A) Average SCRS of overnight cultures of STM-LT2, STM-D23580 and ST-Ty2 with biological assignments of important biomolecules. (B) PCA plot demonstrating distinctive metabolic properties of STM-LT2, STM-D23580 and ST-Ty2 in the Raman fingerprint region. (C) Boxplots show semi-quantification of selected biomolecules by integrating relevant Raman bands. Abbreviations of the biomolecules are listed as follows: “CH” stands for “Carbohydrates”; “DNA” stands for “DNA/RNA”; “Tyr” stands for “Tyrosine”; “Phe” stands for “Phenylalanine”.

Table 5.1: Assignment of Raman bands in section 4.3.1

Wavenumber/cm ⁻¹	Assignment	Vibrational mode	Reference
410	Carbohydrates I	skeletal mode of carbohydrates	[239]
545	Carbohydrates II	C–O–C glycosidic ring deformation	[178]
722	Adenine	Ring breathing	[50]
780	DNA/RNA	cytosine, uracil ring breathing	[178, 179]
853	Tyrosine	Ring breathing	[50]
1003	Phenylalanine	Ring breathing	[240]
1043	Carbohydrates III	C–O stretching	[241]
1240	Amide III of proteins	40% C–N stretching, 30% N–H bending	[46]
1320	Lipid I	C–H deformation	[48]
1450	Lipid II	CH ₂ deformation	[48]
1574	Amide II of proteins	60% N–H bending, 40% C–H stretching	[46]
1662	Amide I of proteins	80% C=O stretching	[46]

vibrational modes of spectral features assigned to nucleic acids, proteins, lipids and carbohydrates are summarised in Table 5.1.

Since one SCRS contains hundreds of Raman bands, unsupervised principal component analysis (PCA) was used to reduce the high dimensional Raman dataset and reveal key information responsible for single-cell variances. PCA plot along the first two principal components (PC1 and PC2) separates the three bacterial strains, in which the invasive, non-typhoidal STM-D23580 sits in between the non-invasive, non-typhoidal STM-LT2 and the invasive, typhoidal ST-Ty2 (Figure 5.1B). It indicates that the three *Salmonella* strains (STM-LT2, STM-D23580 and ST-Ty2) have distinctive phenotypic profiles.

Semi-quantification of intracellular biomolecules was then performed by integrating relevant Raman bands to highlight the phenotypic differences among the three strains (Figure 5.1C and Table 5.1). The most significant differences were observed in carbohydrates (Carbohydrates III), nucleic acids (Adenine and DNA/RNA), lipids (Lipid II), amino acids and proteins (Tyrosine, Amide III and Amide I). These differences

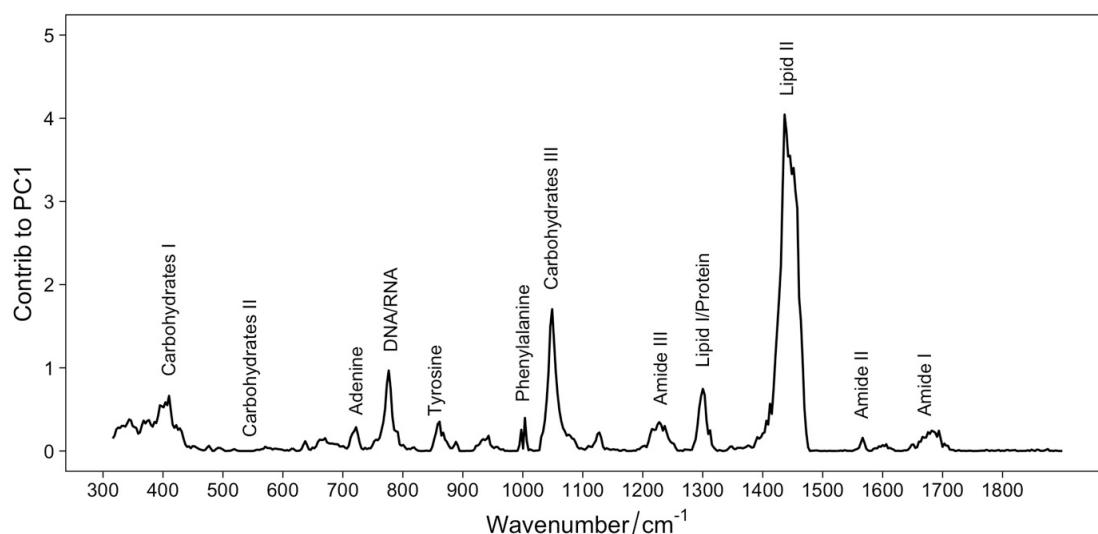


Figure 5.2: Loading plot of SCRS of STM-LT2, STM-D23580 and ST-Ty2, with Raman wavenumbers against their contributions to PC1 of the PCA.

correlate with the Raman contributions along PC1 of the PCA, which accounts for the highest variances among single bacterial cells (Figure 5.2).

A number of traits in the Raman metabolic profiles indicate a closer congruity between the invasive STM-D23580 and ST-Ty2 than to the non-invasive STM-LT2 (Figure 5.1C). The most significant changes in peak intensity between the two invasive strains and STM-LT2 were observed in Raman wavenumbers at 722 and 780 cm^{-1} , corresponding to nucleic acids (Adenine and DNA/RNA), and at 1574 and 1662 cm^{-1} , corresponding to proteins (Amide III and Amide I). An increased DNA/RNA intensity could be due to different DNA states, transitioning from a condensate form in STM-LT2 to an uncoiled form in STM-D23580 and ST-Ty2, ready for more transcription activities in these invasive strains. The relaxed states of DNA facilitate efficient transcription, which may also account for the high intensity of the Amide peaks, and thus more protein synthesis in STM-D23580 and ST-Ty2.

It was previously reported that *S. Typhi* no longer retains the pathway for tyrosine utilisation²⁴² because this serovar does not remain long enough in the gut to utilise the aromatic compounds produced by the human intestinal flora²⁴³. In agreement, the Tyrosine peak (853 cm^{-1}) is significantly lower in ST-Ty2 and STM-D23580 as

Table 5.2: Confusion matrix and performance evaluation of the kNN classification model classifying 116 STM-LT2 spectra, 152 STM-D23580 spectra and 125 ST-Ty2 spectra.

		Actual class		
		STM-LT2	STM-D23580	ST-Ty2
Predicted class	STM-LT2	106	10	0
	STM-D23580	10	135	3
	ST-Ty2	0	7	122
Performance evaluation	Sensitivity	91.4%	88.8%	97.6%
	Accuracy		92.4%	

compared to STM-LT2 (Figure 5.1C).

Surface lipids and proteins are often antigenic and represent the first line of contact with the host immune system²⁴⁴. Lipid II at 1450 cm^{-1} refers to saturated lipids that are abundant in the bacterial cell wall, in particular, myristate and palmitate⁴⁸. Lipid II peak is significantly higher in STM-LT2 as compared to STM-D23580 and ST-Ty2 (Figure 5.1C). The development of a robust outer membrane with an increased level of saturated lipids confers resistance to antimicrobial peptides²⁴⁴. This characteristic might be of relevance to STM-LT2, which primarily resides in the gut, but can be lost in those pathovars that do not remain long enough in the gut, such as ST-Ty2 and STM-D23580.

The Carbohydrates III (1043 cm^{-1}) quantification (Figure 5.1C), on the other hand, indicates dissimilarities between the typhoidal and the typhimurium strains, with ST-Ty2 exhibiting a significantly lower content. The genome of the extra-intestinal *Salmonella* pathovars, such as *S. Typhi*, *S. Paratyphi*, *S. Gallinarum* and *S. Choleraesuis* exhibits degradation of regulators for a variety of processes, including carbohydrates metabolism²⁴⁵. For instance, the human-restricted serovars *S. Typhi* and *S. Paratyphi* harbour pseudogenes related to the transport of D-glucarate into the cell, suggesting that the utilization of this carbon source may not be advantageous in causing invasive infection²⁴².

Whilst feature extraction provides invaluable biochemical information on the phenotypic differences, machine learning algorithms can be used to resolve complex Raman information, classify bacterial species and benefit diagnostic purposes. A classification model based on k-Nearest-Neighbour (kNN) algorithm was applied

to analyse the SCRS of STM-LT2, STM-D23580 and ST-Ty2. A leave-one-out-cross-validation evaluated the kNN model, achieving an overall accuracy rate of 92.4% (Table 5.2). This classification model correctly classifies 106 out of 116 STM-LT2 spectra (sensitivity 91.4%), 135 out of 152 STM-D23580 spectra (sensitivity 88.8%) and 122 out of 125 ST-Ty2 spectra (97.6%). The highest sensitivity rate is achieved in classifying ST-Ty2 indicating that substantial differences emerge between the typhoidal and the non-typhoidal strains. Notably, all incorrect assignments of STM-LT2 and ST-Ty2 solely belong to the class of STM-D23580, whilst STM-D23580 is incorrectly classified as both STM-LT2 and ST-Ty2. This is consistent with both results from the PCA and quantification of biomolecules (Figure 5.1B and C). While STM-LT2 and ST-Ty2 bear distant resemblance, STM-D23580 is closely related to STM-LT2 at a taxonomic level but features many of the ST-Ty2's traits.

5.3.2 Intracellular *Salmonella* shows distinctive metabolic activities revealed by SCRM-DIP

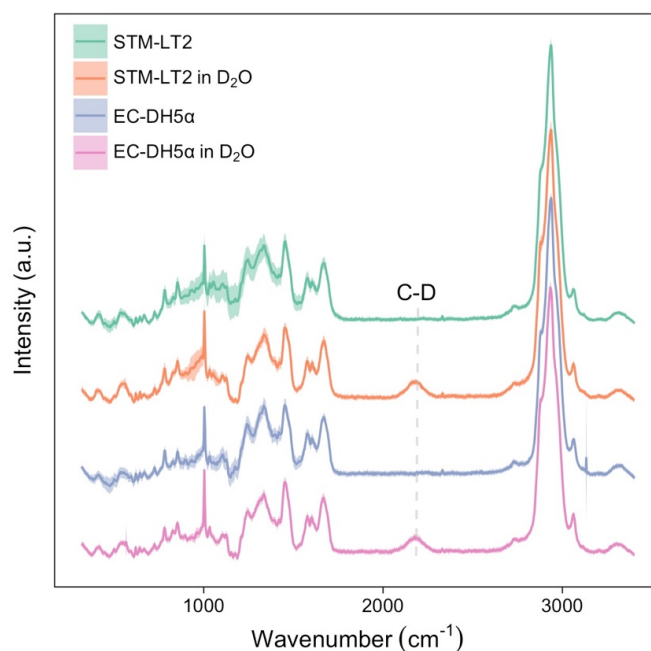


Figure 5.3: SCRS of STM-LT2 and *E. coli* DH5 α . Bacterial cultures were grown overnight in LB broth in the presence of 35% D₂O. A broad C–D band indicates an incorporation of D during active metabolism. The shaded area represents standard deviation from 30 single-cell measurements.

We then sought to implement SCRM-DIP and reverse labeling²⁰⁵ to detect the metabolic rates of *Salmonella* serovars. Due to the incorporation of D from D₂O during active metabolism, a significant C–D Raman band centred at 2170 cm^{−1} appears in bacterial SCRS, shifted from the C–H vibrations at 2800–3100 cm^{−1}. Hence, the intensity of the C–D band can be used to indicate the metabolism rate of the bacteria. In the presence of D₂O, the SCRS of *Salmonella* spp. showed a distinguishable C–D band, similar to *Escherichia coli* DH5 α (EC-DH5 α) (Figure 5.3), which has been used extensively in SCRM-DIP as a reference strain^{205,246}. Hence, the D incorporation in cells can be used as a universal indicator of metabolic activity applicable to *Salmonella* spp..

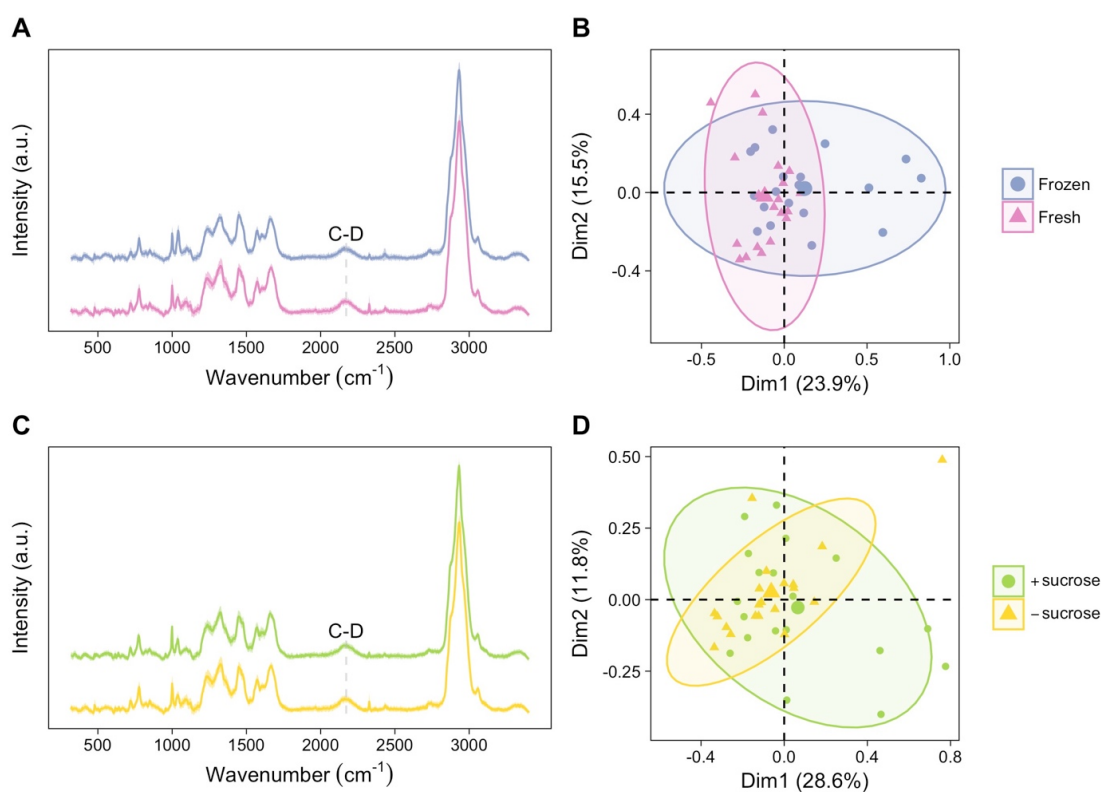


Figure 5.4: SCRS and PCA plots of D-labelled STM-LT2 under different conditions: (A) and (B) growth at post-freezing or fresh state; (C) and (D) growth with or without sucrose,

The degree of D₂O incorporation by STM-LT2 was compared with EC-DH5 α , used as a reference. Assimilation of intracellular D was tested under different experimental conditions (before and after cryopreservation at −80 °C; with or without the presence of 10 mM sucrose) to prove the robustness of the SCRM-DIP method. Bacteria were

able to retain the same level of D and metabolic fingerprints after cryopreservation at -80°C (Figure 5.4A and B). Moreover, the ability to incorporate D and the metabolic states of *Salmonella* were not affected by the presence of sucrose in liquid culture (Figure 5.4C and D).

To better understand the metabolic features that underlie *Salmonella*'s survival and replication within the host cells, the three *Salmonella* strains were initially labelled with deuterium and then reverse labelling was employed to trace the metabolic changes over time during the interaction of *Salmonella* with human host cells. Figure 5.5A shows SCRS of STM-LT2, STM-D23580 and ST-Ty2 recovered at different time points post infection (p.i.) of THP-1 cell line, human monocyte-derived dendritic cells (MoDCs) and human macrophages (Mf). The three strains growing in liquid culture were used as extracellular controls.

In THP-1 and Mf, time points of 0 hr, 0.5 hr, 1 hr, 2 hr, 4 hr and 6 hr p.i were selected for analysis while cells were lysed at 0 hr, 0.75 hr, 2 hr, 4 hr and 6 hr p.i. for MoDCs. The difference in time points reflects the effector functions of each of the cell types herein included. Mf and differentiated THP-1 cells are specialised phagocytic cells that quickly uptake and clear pathogens by engulfing them. Across *ex vivo* settings, incubation time of 30 minutes or even less guarantees successful internalisation of *Salmonella* by human Mf^{247,248}. Compared to Mf, the phagocytic activity of dendritic cells (DCs) is less pronounced²⁴⁹. After sensing pathogens via receptors or via inflammatory stimuli, DCs switch on expression of chemokine receptors, migrate to draining lymphoid organs and undergo a process of maturation. Unlike Mf, the major task of DC is not to clear pathogens but to present antigens to T lymphocytes. Thus, pathogen uptake, phagosomal degradation and acidification are much lower in DCs than in Mf²⁵⁰. Taking into account the above and in order to maximise the phagocytic activity of our MoDCs, we prolonged the incubation time from 30 to 45 minutes, becoming the first time point p.i.. Indeed, the selection of time points is the actual result of experimental optimisation. Our model is in line with other publications²⁵¹⁻²⁵³.

Figure 5.5B illustrates that, at each time point, cellular D content was quantified by calculating the ratio of C-D / (C-D + C-H) in SCRS of single bacterial cells. The overall

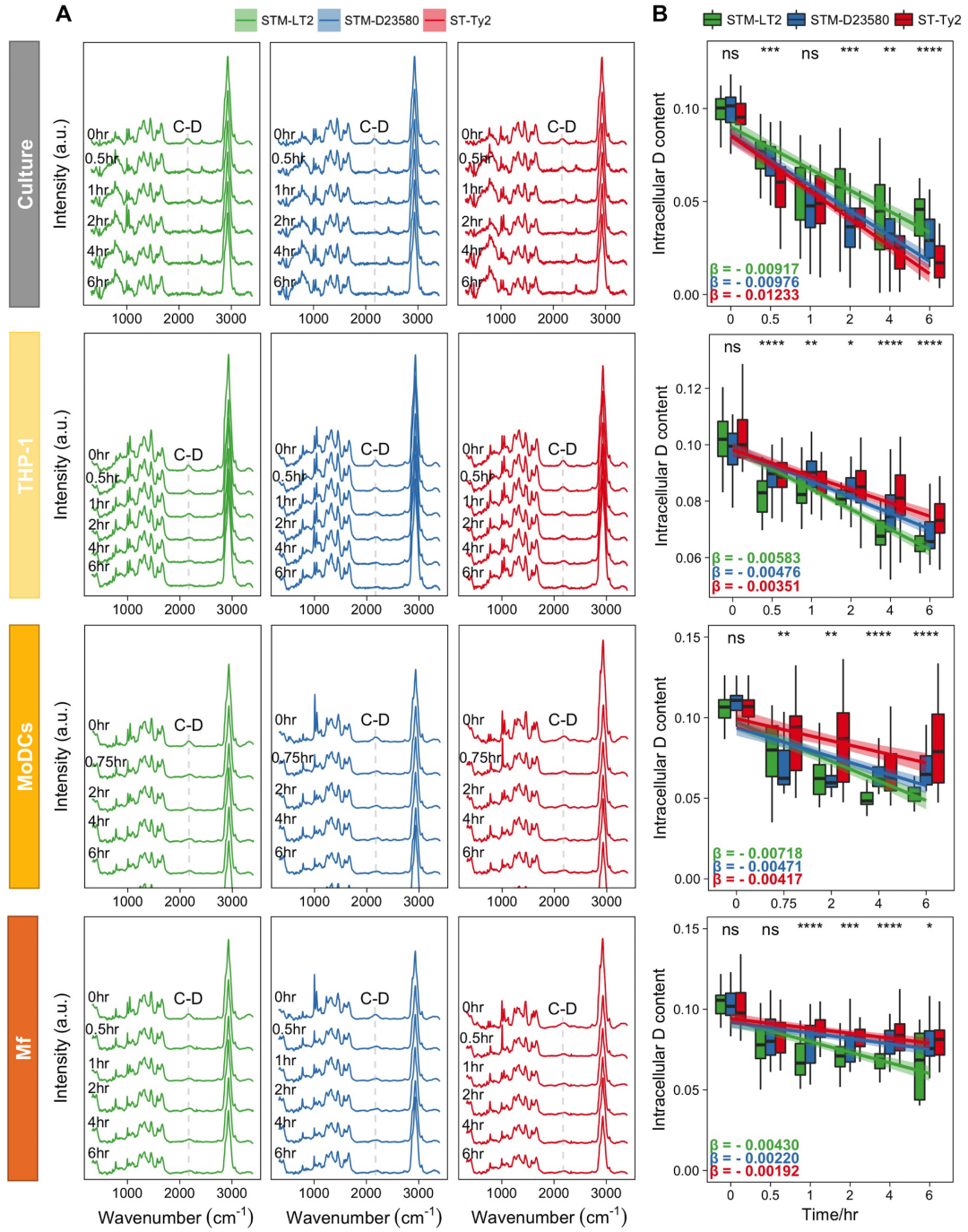


Figure 5.5: (A) SCRS of *Salmonella* serovars under four experimental conditions: extracellular growth in liquid culture, and intracellular metabolism during infection of a differentiated THP-1 cell line, primary monocyte-derived dendritic cells (MoDCs) and primary macrophages (Mf). (B) Boxplots quantifying D content at different time points. Linear regression shows the D decreasing speed for each strain at each experimental condition. Notation above the box plots indicates statistical significance of the three strains at each time point by one-way ANOVA test (ns: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; ****: $p < 0.0001$). Linear regression was fitted to compare the D decreasing speed of each strain in each infection condition and the slope was labelled as β . The absolute value of β represents the metabolism speed.

metabolic rate was determined as the gradients (β) of D loss during growth in liquid culture and replication in either THP-1 cells, MoDCs or Mf. In this reverse labelling process, the initial labelled C–D band gradually decreased due to a replacement of C–H band during active metabolism. Therefore, a greater C–D loss speed β represents a higher bacterial metabolic rate.

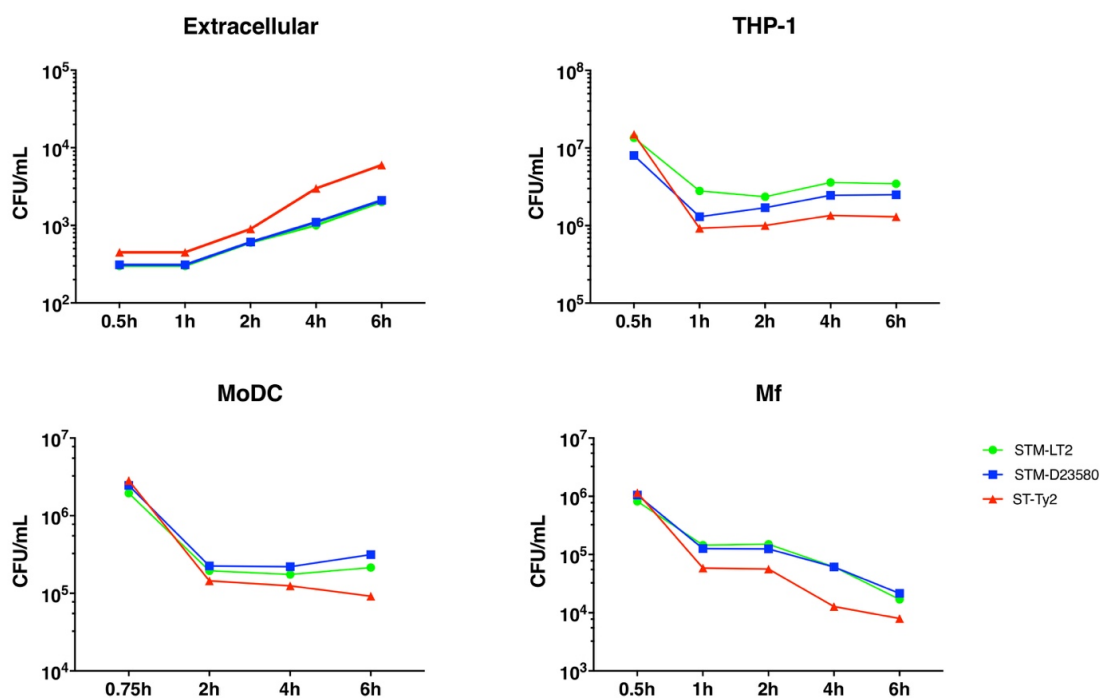


Figure 5.6: (A) Extracellular growth of D-labelled bacteria in RPMI. Intracellular growth of D-labelled bacteria in (B) THP-1, (C) MoDCs and (D) Mf. An average of two experiments is shown.

Under a free-living condition in the growth medium, SCRS of extracellular *Salmonella* illustrate a gradually decreasing C–D band over time in all three *Salmonella* strains. ST-Ty2 displays the fastest D loss (-0.01233), compared to STM-LT2 (-0.00917) and STM-D23580 (-0.00976), indicating that ST-Ty2 is the most active and fastest duplicative strain in liquid culture. This is consistent with the higher colony forming unit (CFU) counts in Figure 5.6A in comparison to two Typhimurium strains. In support of these results, the Raman phenotypic quantification in Figure 5.1C also shows that in a nutrient-rich environment, the higher energy requirements of ST-Ty2, necessary for its replication, is fulfilled by higher consumption of carbohydrate stocks (Carbohydrates III). The extensive growth is accompanied by a higher content of nucleic

acids (DNA/RNA) ready for transcription and increased protein production (Amide III and Amide I) (Figure 5.1C).

A decrease of the C–D band in the recovered intracellular *Salmonella* was also observed in all strains, during time-course infection of THP-1 cells, MoDCs and Mf (5.5A). In contrast to the observation in the liquid growth medium, all infection models show that ST-Ty2 was the strain with the slowest metabolic rate (Figure 5.5B). In THP-1 cells, at 6 hr p.i., a visible amount of the C–D band is still present in the spectra of ST-Ty2, as opposed to the complete absence of the C–D band in STM-LT2 and STM-D23580. Quantification of D content confirms that ST-Ty2 has the slowest metabolic rate (-0.00351), comparing to STM-LT2 (-0.00583) and STM-D23580 (-0.00476). A similar spectral pattern is also observed in human MoDCs and Mf, where ST-Ty2 has the slowest metabolism (-0.00417 in MoDCs, -0.00192 in Mf), followed by STM-D23580 (-0.00471 in MoDCs, -0.00220 in Mf) and ST-LT2 (0.00718 in MoDCs, -0.00430 in Mf). The slowest metabolic rate of ST-Ty2 during infection in THP-1, MoDCs and Mf also correlates with the CFU counting of the three *Salmonella* strains (Figure 5.6B, C and D). Therefore, in terms of *Salmonella* metabolic activity in all infection models, the invasive ST-Ty2 has the slowest metabolic rate, and the non-invasive STM-LT2 has the fastest metabolic rate, whilst the activity of the non-typhoidal but invasive clinical strain STM-D23580 is in between.

It is known that the invasive host-adapted *Salmonella* replicates slowly and elicits weaker pro-inflammatory responses²¹⁶ in order to avoid immune-surveillance and to disseminate inside host cells. Our results highlight a plausible strategy adopted also by STM-D23580 to slow down its central metabolism in an effort to survive intracellularly while avoiding immune recognition.

Overall, all three *Salmonella* strains exhibited a slower metabolism in intracellular conditions than in free-living, liquid growth medium. The metabolic activity was particularly reduced in Mf as compared to THP-1 and MoDCs, most likely as a result of the hostile acidic phagosomal environment²⁵⁴. Likewise, Mf exhibited increased killing ability compared to THP-1 and MoDCs, as shown in the lower CFU counts within Mf

compared with other cell models in Figure 5.6. This result suggests that inside the fatal compartment of macrophages, *Samonella* remains silent and not very active.

5.3.3 SCRS map dynamic phenotypic profiles of intracellular *Salmonella* within primary cells

We hypothesise that *Salmonella* could transit from aerobic to anaerobic lifestyle and switch at the level of nutrient uptake and biomass synthesis to adapt in the hostile environmental niches inside host cells. As SCRS provide a snapshot of bacterial phenotypic profiles, SCRS of intracellular cells were analysed for each *Salmonella* strain by quantifying relevant Raman bands at different time points p.i., thus providing a phenotypic comparison across strains. Statistics was done by using one-way ANOVA with a global p-value to indicate general differences among time points (ns: $p > 0.05$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; ****: $p < 0.0001$). Linear regression was fitted to compare the time-wise changes of one specific quantification in each strain and the slope was labelled at the top right corner, where a negative value indicates a decrease over time and a positive value indicates an increase over time. Time-resolved Raman profiles are summarised and discussed in terms of carbohydrates, lipids, nucleic acids and proteins as follows:

Carbohydrates

Variations in bacterial content of carbohydrates (Carbohydrates I, II and III) were particularly evident in THP-1 and in Mf, as compared to MoDCs. Previous studies have indicated that the intracellular lifestyle of *Salmonella* in MoDCs differs from that in Mf^{254–256}, suggesting that, in MoDCs, the bacteria do not require the gene expression program necessary for survival in Mf. For instance, a high C-O stretching at 1050 cm^{-1} (Carbohydrates III), the typical vibrational mode of glycogen²⁴¹, was observed for all the strains at 1 hr p.i. in Mf (t-test, **** against other time points in all strains). Since the role of glycogen depots in the virulence of *Salmonella* has been shown²⁵⁷, this result may suggest that more intense stress conditions might be encountered by the bacteria in Mf as compared to MoDC and 1 hr p.i. might be an important stage of host-exerted stress and bacterial tolerance. Distant adjustments between MoDCs and Mf could

suggest differences in host environments and accessibility of different metabolites in the *Salmonella*-containing vacuole (SCV) during the course of infection.

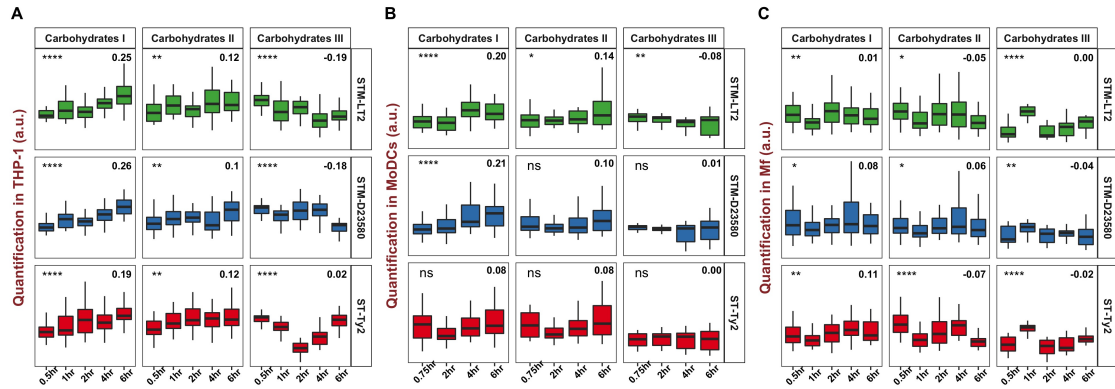


Figure 5.7: Raman profiling of **carbohydrates** in intracellular bacteria. Raman bands of Carbohydrates I, II and III were quantified in fingerprint region of SCRS of STM-LT2, STM-D23580 and ST-Ty2 at different time point of infection in (A) THP-1 cells, (B) dendritic cells and (C) macrophages.

Across three strains, an increased consumption of Carbohydrates III in STM-LT2 was observed in THP-1 (-0.19) and MoDCs (-0.08) as compared to STM-D23580 (-0.18 in THP-1; 0.01 in MoDCs) and ST-Ty2 (0.00 in THP-1; -0.02 in MoDCs). This indicates a greater utilisation of carbohydrate stocks by the non-invasive STM-LT2 during infection.

Lipids

On the contrary to the carbohydrates quantification, a decrease of Lipid I (saturated long chain fatty acids) at 1320 cm^{-1} is observed in STM-D23580 (-0.2 in THP-1; -0.11 in MoDC; -0.12 in Mf) and ST-Ty2 (-0.16 in THP-1; -0.23 in MoDCs; -0.15 in Mf), compared with in STM-LT2 (-0.12 in THP-1; -0.05 in MoDCs, -0.11 in Mf). The results of favouring lipids over carbohydrates during the course of infection suggest that the invasive *Salmonella* ST-Ty2 and STM-D23580 should adopt a switch from carbohydrate consumption to lipid consumption. In line with a previous study²⁵⁸, our results suggest that invasive *Salmonella* appears to switch off the glycolytic pathways, using fatty acids as carbon sources instead.

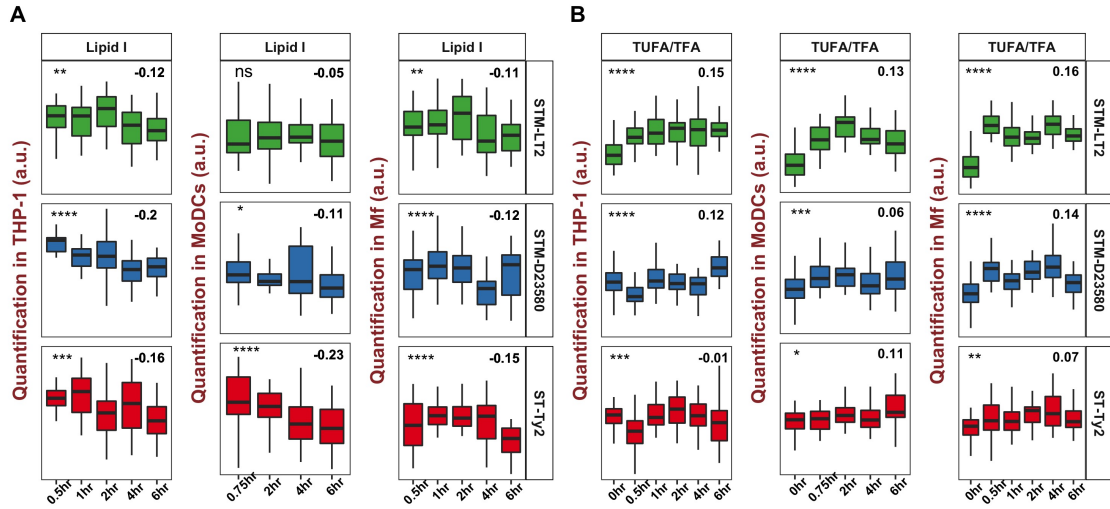


Figure 5.8: Raman profiling of **lipids** in intracellular bacteria. Raman bands of Lipid I and TUFA/TFA were quantified in fingerprint region of SCRS of STM-LT2, STM-D23580 and ST-Ty2 at different time point of infection in (A) THP-1 cells, (B) dendritic cells and (C) macrophages.

The contribution of lipid metabolism to the virulence and survival of *Salmonella* in the intravacuolar environment has been highlighted in a number of studies^{258,259}. For instance, *fadE*, which encodes for an acyl-CoA dehydrogenase specific for medium-/long-chain fatty acids, was significantly increased in *S. Typhimurium* during carbon starvation²⁶⁰. The development of a robust outer membrane, due to the increased level of saturated lipids confers certain resistance to antimicrobial peptides²⁶¹, in particular during the stages of gut infection.

To further assess the modulation of *Salmonella* lipidic phenotype during its intracellular life, we analysed the higher wavenumber region of the SCRS (2700 – 3300 cm^{-1}). The band centered at 3060 cm^{-1} in SCRS is representative of unsaturated =CH stretching and the one at 2930 cm^{-1} represents all CH, CH₂ and CH₃ stretching in the cell²⁶². The ratio of total unsaturated fatty acids/total fatty acids (TUFA/TFA) is commonly used in Raman studies to characterize lipid membrane profile⁴⁹. The TUFA/TFA ratio is also used to study changes in lipid content of bacterial cell wall, as an index of bacterial adaptation and response to oxidative stress²⁶³. Increased unsaturation of lipids causes fluidising of the membrane and allows membrane modifications that enhance the tolerance of bacteria under stress conditions. As bacterial membrane modification

is particularly upon internalisation, we compared the TUFA/TFA ratio between the intracellular and the extracellular conditions. The TUFA/TFA ratio in all conditions show a marked upregulation in STM-LT2 (0.15 in THP-1; 0.13 in MoDCs; 0.16 in Mf) compared to STM-D23580 (0.12 in THP-1; 0.06 in MoDCs; 0.14 in Mf) and ST-Ty2 (0.16 in THP-1; 0.11 in MoDCs; 0.07 in Mf), highlighting the differences in membrane alteration between the invasive and non-invasive strains. Notably, the TUFA/TFA ratio in all strains increased dramatically from 0 hr to 0.5 hr in Mf (t-test, **** in STM-LT2, *** in STM-D23580 and * in ST-Ty2), suggesting a rearrangement of the surface moieties upon phagocytosis necessary to resist to the hostile intra-macrophage environment.

Nucleic acids

Transient accumulation of adenine was observed for all the bacterial strains during an early stage of MoDC infection. This result can be explained by the fact that bacteria might have sufficient purine reserves to sustain only the initial host cell attack. It is also possible to suggest that the bacteria utilise more purine reserves at a later stage of infection for intracellular survival and replication. However, in Mf, significant alterations of the Adenine peak were only detected for the Typhimurium strains. Purine nucleotides are critical for intra-macrophage survival of *Salmonella*. Under purine-limiting environments like the macrophage phagosome, *Salmonella* is unable to repair the DNA damage, which ultimately results in bacterial death²⁶⁴.

In Figure 5.9B, a clear variation in DNA/RNA macromolecules is visible during MoDC infection for all bacterial strains, with a marked upregulation at 2 hr p.i. (t-test, * in STM-LT2, *** in STM-D23580 and ** in ST-Ty2), likely due to increased DNA synthesis and transcriptions. A decrease was observed in the invasive STM-D23580 (-0.23) and ST-Ty2 (-0.20), as compared to STM-LT2 (0.12). Amide I (Figure 5.10), a representative peak of C=O stretching in proteins (Table 5.1), showed different time-wise modulation as compared to DNA/RNA. We calculated the ratio of nucleic acids/protein for all three *Salmonella* strains during infection as an indicator of intracellular bacterial replication (Figure 5.9). Following a short increase of the ratio, a significant decrease is observed in the invasive STM-D23580 (-0.22) and ST-Ty2 (-0.16) compared to STM-LT2 (-0.04). This

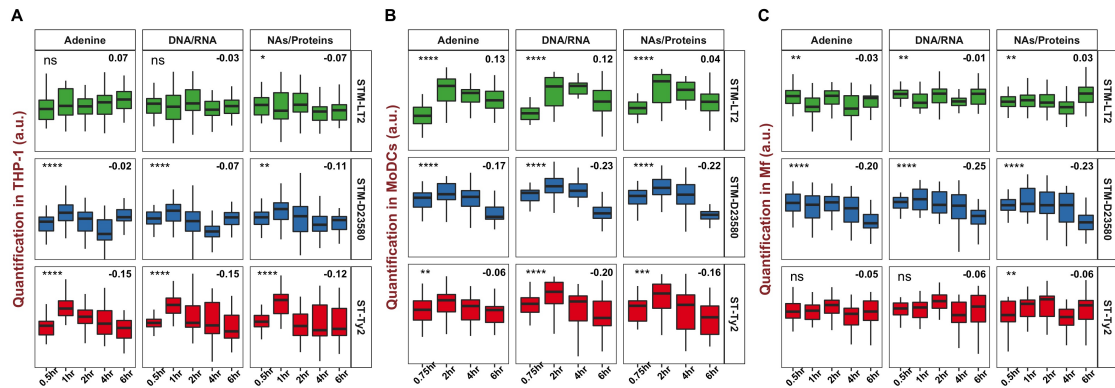


Figure 5.9: Raman profiling of **nucleic acids** in intracellular bacteria. Raman bands of Adenine, DNA/RNA, and the ratio of DNA/RNA to Amide I were quantified in fingerprint region of SCRS of STM-LT2, STM-D23580 and ST-Ty2 at different time point of infection in (A) THP-1 cells, (B) dendritic cells and (C) macrophages.

indicates an active intracellular bacterial replication at earlier time points, followed by a decrease later in the infection, probably due to a reduction of DNA synthesis and arrested growth of the invasive strains as a consequence of the adverse intracellular environment. Nevertheless, a great heterogeneity can be observed in the nucleic acids/protein ratios, implying bacterial sub-populations in different growth phases.

Aromatic amino acids and proteins

Significant fluctuation in the levels of phenylalanine is observed in STM-D23580 and ST-Ty2 as compared to STM-LT2 in both MoDCs and Mf. The contribution of aromatic amino acid metabolism to intracellular *Salmonella* proliferation has been shown to be important and distinct in different cell types²⁶⁵. Notably, in MoDCs, all bacteria have a bimodal expression of phenylalanine, suggesting that bacteria in different metabolic stages may be simultaneously present inside the host.

Significant variations in the Amide III peak were also detected except for STM-LT2 in MoDCs, possibly due to either intracellular stress that changes the properties of bacterial proteins and enzymes, or protein denaturation as a result of bacteria being killed. Indeed, significant changes in the Amide III peak are noticeable during Mf infection, in agreement with the decreased bacterial survival (Figure 5.6) and metabolic activity (Figure 5.5) previously observed. Additionally, a decrease can be

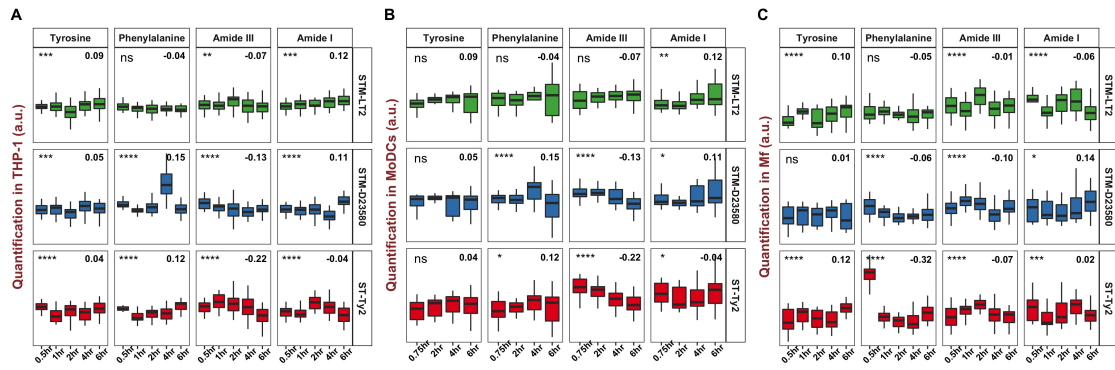


Figure 5.10: Raman profiling of **amino acids and proteins** in intracellular bacteria. Raman bands of Tyrosine, Phenylalanine, Amide III and Amide I were quantified in fingerprint region of SCRS of STM-LT2, STM-D23580 and ST-Ty2 at different time point of infection in (A) THP-1 cells, (B) dendritic cells and (C) macrophages.

observed in STM-D23580 (-0.15) and ST-Ty2 (-0.22) during infection of MoDCs as compared to STM-LT2 (-0.07), which might contribute to invasive characteristics of these strains in MoDCs.

5.3.4 Single-cell Raman analysis resolves phenotypic heterogeneity in isogenic populations

Salmonella displays significant cell-to-cell variation resulting in distinct subpopulations with different metabolic phenotypes. During infection, multiple outcomes are observed when *Salmonella* encounters the host cell: bacteria may reside in the SCV or escape in the cytosol, they can replicate and spread to neighbour cells or die due to antimicrobial defence mechanisms. Single-cell measurements help to reveal different sub-groups of bacteria in otherwise seemingly homogeneous populations. We, thus, generated a t-Distributed Stochastic Neighbour Embedding (t-SNE) of all bacterial cells in MoDCs and Mf at 2 hr p.i. to exemplify and visualise the multimodal behaviours of intracellular bacteria. In Figure 5.11, three subpopulations were clustered for bacteria in MoDCs (MoDC1, MoDC2 and MoDC3) and Mf (Mf1, Mf2 and Mf3).

Bacteria in the MoDC3 cluster were enriched in lipid content while displaying low levels of carbohydrates. This metabolic phenotype is compatible with that of bacteria restricted to phagosomal compartments, in which glucose is less available and alternative carbon sources are required²⁶⁶. The nitric oxide (NO) produced in the

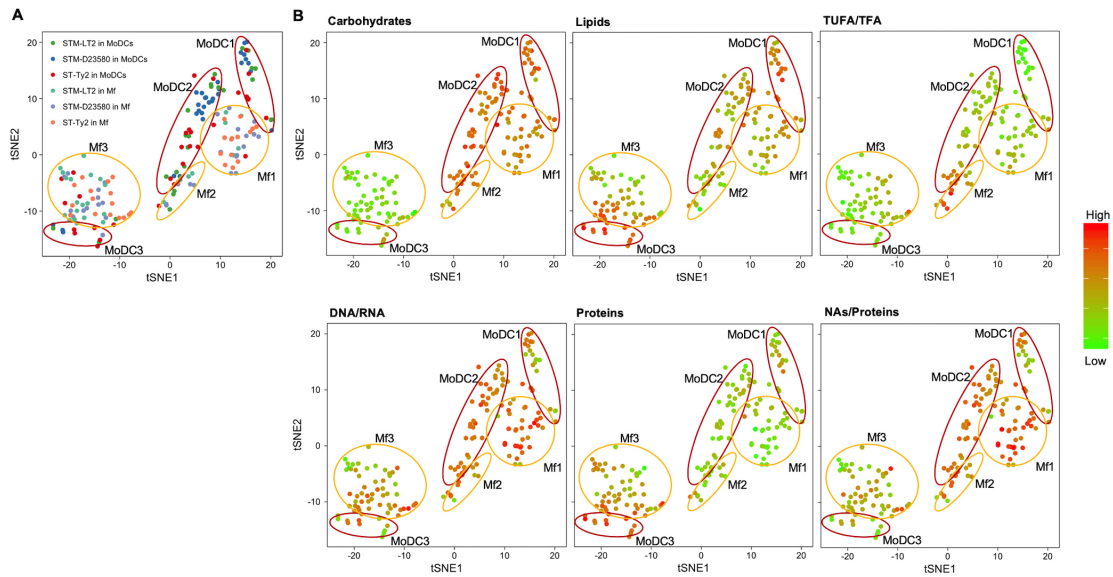


Figure 5.11: (A) A t-SNE plot of single cells (perplexity = 10) computed based on SCRS of STM-LT2, STM-D23580 and ST-Ty2 at 2 hr p.i. of MoDCs and Mf. Three bacterial subpopulations were clustered in MoDCs (MoDC1, MoDC2 and MoDC3) or in Mf (Mf1, Mf2 and Mf3). (B) t-SNE plots coloured based on quantification of different biomolecules. The intensity gradient is mapped from green (low) to red (high).

SCV targets the pyruvate and α -ketoglutarate dehydrogenase complexes, which would strongly affect glycolysis, glycogenesis and the TCA cycle²⁶⁷. The higher lipid content, on the other hand, suggests the employment of alternative energy pathways.

MoDC3 and Mf3 clusters also had a low TUFAs/TFA ratio. When exposed to acidic pHs, such as in the SCV, *Salmonella* induces a profound rearrangement of lipid membranes by increasing the unsaturated fatty acids content²⁶⁸. On the contrary, the higher TUFAs/TFA ratio, observed for the bacteria in the clusters MoDC2 and Mf2, could suggest continuous internal turnover, damage, and/or loss to the environment, which has previously been proposed as a potential strategy adopted by the persisters²⁶⁹. Diversity in proliferation is also observed in different bacterial groups. A low DNA/RNA contents and NAs/protein ratio was observed for the bacteria of the “intraphagosomal” clusters MoDC3 and Mf3. In fact, under particular stress conditions, such as the presence of NO, *Salmonella* cell division is reduced^{270,271}. On the contrary, high protein-specific peak intensities indicate an increase in protein synthesis, which may be required for the correct assembly of the type II secretion system (T2SS) machinery²⁷². On the other hand, bacteria in the clusters of MoDC1, MoDC2, Mf1 and Mf2 exhibit

a different metabolic profile characterised by a high content of carbohydrates and nucleic acids concomitant with a reduced TUFA/TFA ratio.

The Mf3 population is bigger than MoDC3, suggesting that a more efficient compartmentalisation of *Salmonella* in the SCV could be achieved by Mf. Previous studies have indicated that the intracellular lifestyle of *Salmonella* in MoDCs differs from that in Mf^{254–256}, implying that *Salmonella*'s gene expression required for survival inside MoDCs is not the same as the one needed for survival in Mf.

A heatmap of single-cell Raman band intensities was then generated in Figure 5.12. It indicates the proximity and relationships of different biomolecule modules over the course of infection Raman profiles of Carbohydrates I, II, Amide I and Lipid II clustered closely in all strains in human MoDCs and Mf. These spectral features are associated with macromolecules that generally contribute to cell surface properties of bacteria (e.g. glycoprotein, glycolipid and lipopolysaccharide). Simultaneous regulation of these molecules might indicate a rearrangement of the surface antigen moieties, which in turn, could allow *Salmonella* to resist the hostile intracellular environment.

Figure 5.12 also indicates a certain level of heterogeneity in biomolecule synthesis and utilisation. At a later stage of the infection, such heterogeneity was particularly accentuated in primary cells as compared to THP-1 cells. For example, compared to a rather homogeneous distribution of Carbohydrate I across all time points in THP-1, subpopulations of very high or low Carbohydrate I content started to develop in MoDCs and Mf.

Accurate results on metabolic responses are often hampered by the use of cancer cell lines that already exhibit various de-regulated reactions in the primary carbon metabolism. Cell lines have already experienced massive metabolic changes which may be similar to those triggered by the bacteria and hence may mask pathogen heterogeneity²⁷³. Importantly, donor-wise centring proved that this phenotypic heterogeneity was due to metabolically different bacteria and not due to differences among donors.

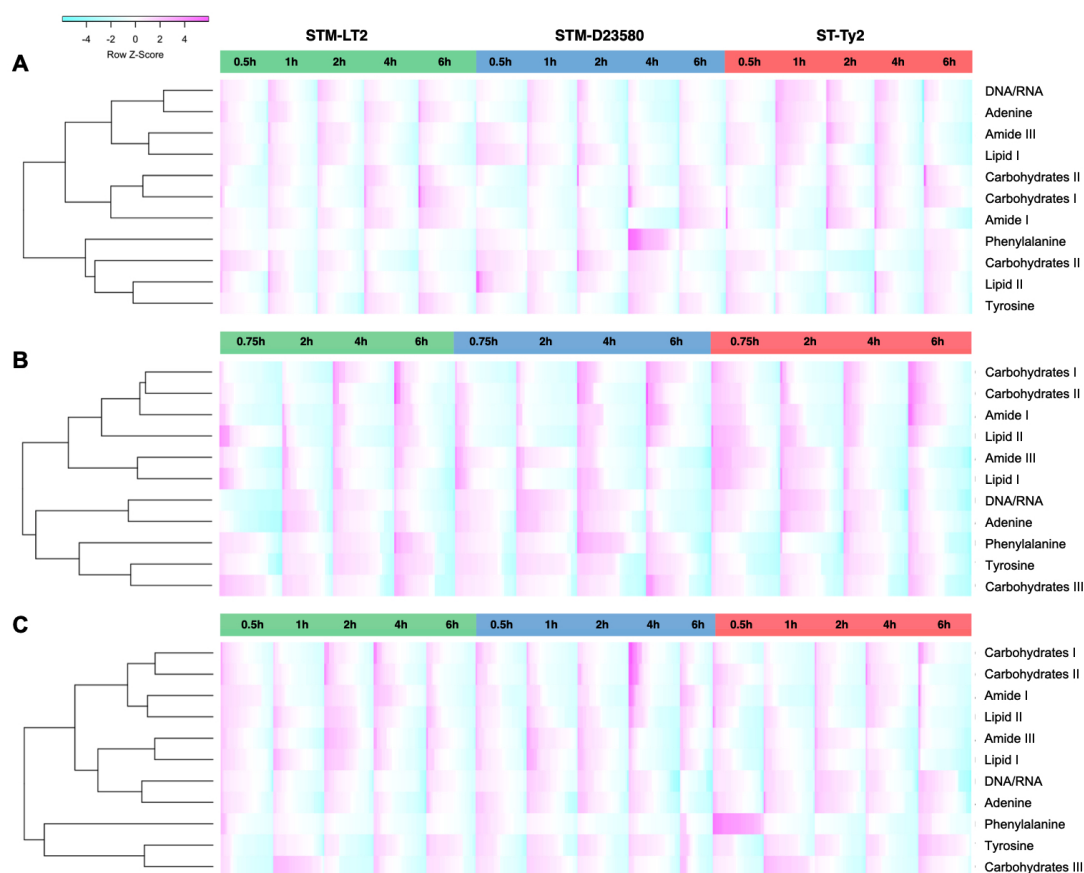


Figure 5.12: Single-cell heatmap from Raman integrations of important biomolecules demonstrates serovar-specific heterogeneity during the course of infection of (A) THP-1, (B) MoDCs and (C) Mf. The heatmap is column-scaled and each column of the heatmap represents one-time point and within each time point, Raman intensities were ordered from the highest to the lowest. Each row of the heatmap represents a group of biomolecule and rows were ordered based on hierarchical clustering dendrogram to view proximity and relationships of different biomolecules.

5.3.5 Intracellular Raman profiles are consistent with the proteomic analysis

MoDCs are the most important antigen-presenting cells. The ability of *Salmonella* to subvert MoDC functions may contribute to their survival and dissemination within the host^{251,274,275}. In order to validate the Raman observations, we performed time-course proteomics experiments of intracellular *Salmonella* recovered from MoDCs. We identified 687 proteins in STM-LT2, 1,277 proteins in STM-D23580 and 1408 proteins in ST-Ty2 (Figure 5.13). For each bacterial strain, differentially-expressed proteins (DEPs; $p < 0.05$) were determined across the time points. Those DEPs were then functionally

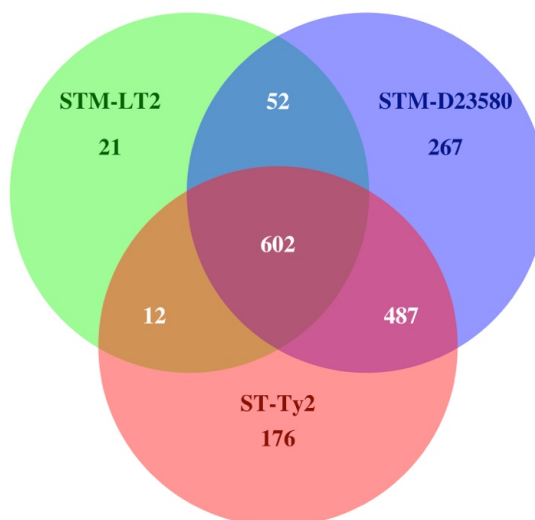


Figure 5.13: Venn diagram of proteins detected in STM-LT2, STM-D23580 and ST-Ty2, in at least 3 time points.

annotated to bacterial pathways related to either energy metabolism (Figure 5.14A and Figure 5.15) or biosynthesis (Figure 5.14B and Figure 5.16).

Overall, proteins at 2 hr p.i. were less abundant compared to other time points (Figure 5.15 and Figure 5.16). It has been shown that early SCV undergoes extensive remodeling²⁷⁶ and 2 hr p.i. might represent the stage at which *Salmonella* adapts to the intra-SCV lifestyle and switches its metabolic requirements. This transient downregulation of protein expression can be a consequence of an early host response. Once the bacteria adapt to the hostile intracellular environment, they restore their metabolic activity.

DEPs related to energy pathways

Carbon metabolism and energy utilisation are key to all bacterial metabolic events. We, therefore, hypothesise that regulation in energy pathways could support the observation that the three *Salmonella* strains demonstrate distinctive metabolism rates by SCRM-DIP. Figure 5.14 shows expression of DEPs related to energy pathways. Most of the proteins related to bacterial energy pathways were upregulated during the intracellular life of STM-LT2 except for those responsible for glycolysis and fermentation. STM-D23580 exhibited a mixture of upregulated and downregulated pathways. All pathways related to energy metabolism had a decreasing trend in ST-Ty2. This is in

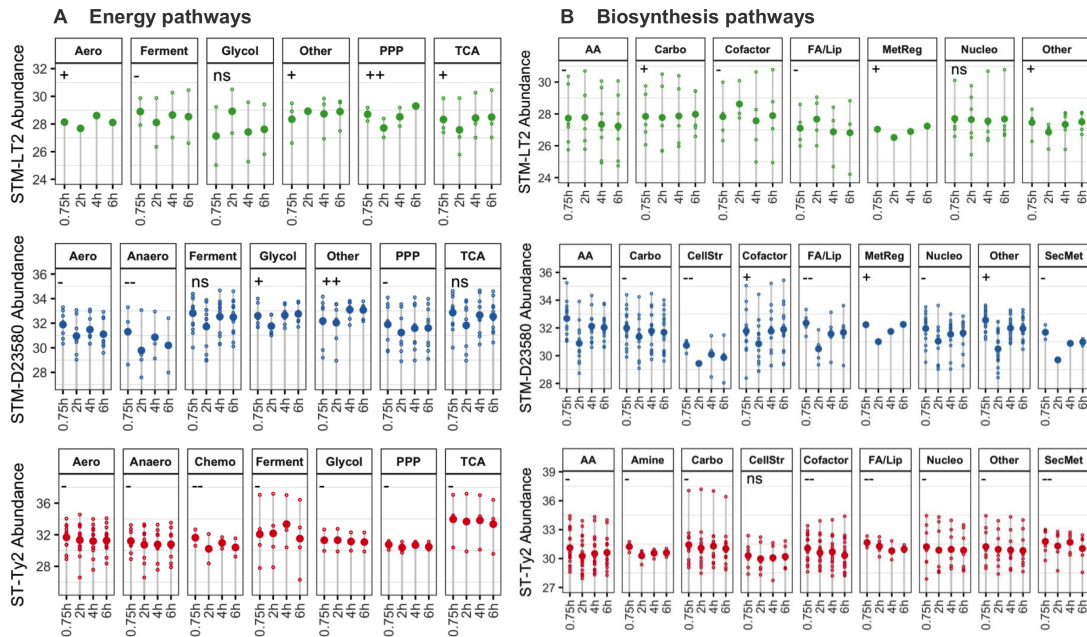


Figure 5.14: Dotplots of the abundance of DEPs (Log2 of the LFQ intensity) related to (A) energy and (B) biosynthesis pathways in STM-LT2 (green), STM-D23580 (blue) and ST-Ty2 (red). The large dot represents the average of all Log2 protein expression values belonging to a specific biosynthesis pathway at one time-point. Each of the small dots represents a data value for an individual protein within the pathway. Time-wise changes were calculated as linear regression slopes and represented as “++” and “+” being upregulated ($>10^{-1}$ and $>10^{-2}$), “-” and “-” being downregulated ($<-10^{-1}$ and $<-10^{-2}$), and “ns” being non-significant ($<10^{-3}$) over time. Abbreviations of the pathways are listed as follows: “AA” stands for “Amino Acid Biosynthesis”; “Amine” stands for “Amines and Polyamines Biosynthesis”; “Carbo” stands for “Carbohydrates Biosynthesis”; “CellStr” stands for “Cell Structures Biosynthesis”; “Cofactor” stands for “Cofactors, Prosthetic Groups, Electron Carriers Biosynthesis”; “FA/Lip” stands for “Fatty Acids and Lipids Biosynthesis”; “MetReg” stands for “Metabolic Regulators Biosynthesis”; “Nucleo” stands for “Nucleosides and Nucleotides Biosynthesis”; “Other” stands for “Other Biosynthesis”; “SecMet” stands “Secondary Metabolites Biosynthesis”.

a good agreement with the SCRM-DIP results, indicating that the invasive typhoidal ST-Ty2 had the lowest metabolic activity among three *Salmonella* pathovars (Figure 5.5), by slowing down most metabolic reactions to obtain energy and nutrients.

Figure 5.15 listed the expression and names of the DEPs related to energy pathways. Overall, a downregulation of many metabolic enzymes in the TCA cycle is observed in all *Salmonella* strains at 2 hr p.i., in good agreement with the previous report²⁷⁷. For instance, the aspartate aminotransferase AspC, an important aminotransferase enzyme catalyzing the transfer of amino acids in the TCA cycle²⁷⁸, was downregulated. Other TCA cycle related enzymes, the NADP-specific glutamate dehydrogenase

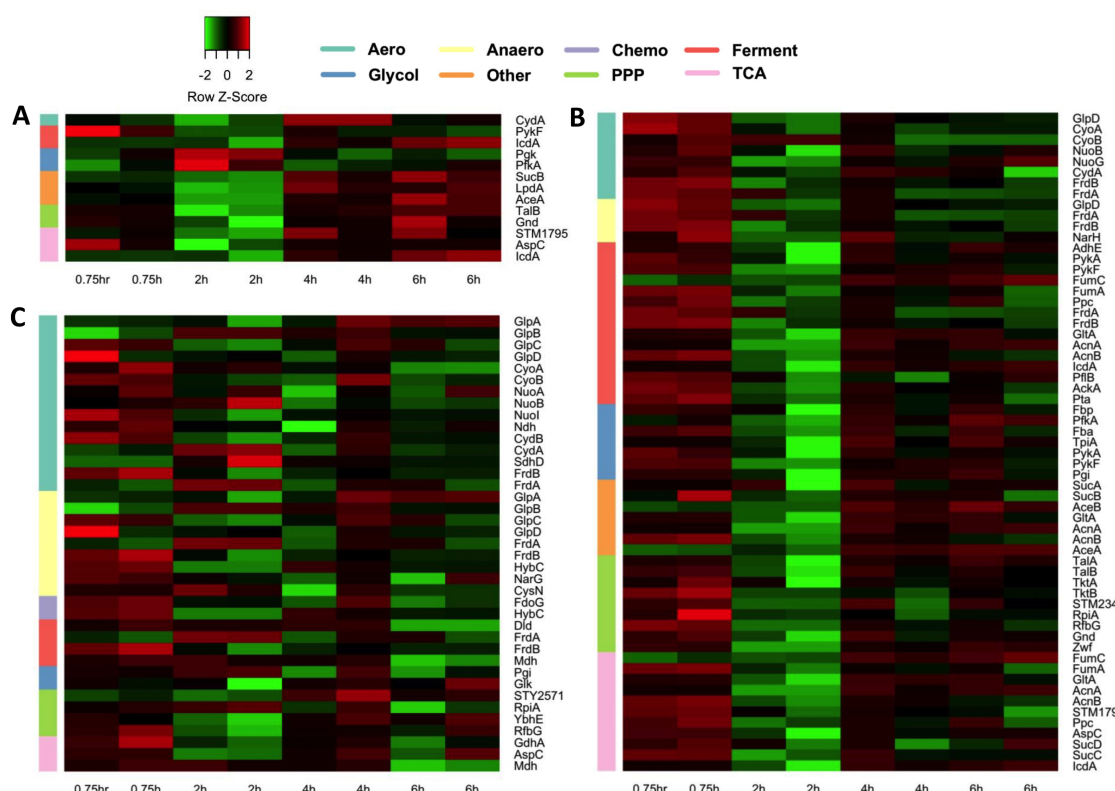


Figure 5.15: Heatmap of the row-normalised abundance of differentially expressed proteins (DEPs) related to energy pathways in (A) STM-LT2, (B) STM-D23580 and (C) ST-Ty2 or biosynthesis pathways in (D) STM-LT2, (E) STM-D23580 and (F) ST-Ty2. Colour bars on the left present assignment of protein to one specific energy pathway and individual protein names are listed on the right. Abbreviations of the pathways are listed as follows: “Aero” stands for “Aerobic Respiration”; “Anaero” stands for “Anaerobic Respiration”; “Chemo” stands for “Chemoautotrophic Energy Metabolism”; “Ferment” stands for “Fermentation”; “Glycol” stands for “Glycolysis”; “Other” stands for “Other Energy Pathways”; “PPP” stands for “Pentose Phosphate Pathway”; “TCA” stands for “The Citric Acid Cycle”.

GdhA, the succinyl-CoA synthetases SucC and SucD, and AcnA and AcnB were also downregulated at a later stage of the infection. The inactivation of the TCA cycle increases the ability of *Salmonella* to survive within resting and activated murine macrophages²⁷⁹, which might explain a more prominent disruption of TCA cycle and aerobic respiration in STM-D23580 and ST-Ty2 (Figure 5.14A). This result supports the dynamic Raman profiling in MoDCs, where a significant consumption of carbon stock (Carbohydrates III) was observed in STM-LT2, whilst higher consumption of lipids occurred in STM-D23580 and ST-Ty2 (Figure 5.7). These results highlight a possible strategy adopted by the invasive strains to survive intracellularly: switching energy

sources from aerobic respiration to lipid-driven metabolisms. This conclusion is also supported by a decrease in lipid biosynthesis pathway in STM-LT2, STM-D23580 and ST-Ty2 (Figure 5.14B).

DEPs related to biosynthesis pathways

Next, we sought to correlate the DEPs in biosynthesis pathways with the dynamic Raman profile of intracellular bacteria in MoDCs. Biosynthesis pathways related to amino acids (AA), amine and polyamine were found to be differentially regulated during the infection and differences among strains were observed (Figure 5.14B). While proteins associated with glutamine and lysine synthesis were upregulated at final time points in STM-LT2 (STM1795 and DapD), proteins associated with these pathways (e.g. STM1795, GlnA, ThrA, IlvC, IlvD, DapA, DapD) were detected at significantly lower level at 6 hr, as compared to the initial stage of the infection, in STM-D23580 and ST-Ty2 (Figure 5.16). Overall, AA biosynthesis was more repressed in STM-D23580 and ST-Ty2 during the infection, as a significant decrease of Amide III was observed in the invasive strains while non-significant changes were detected in STM-LT2 (Figure 5.10).

Phenylalanine biosynthesis proteins were significantly downregulated during infection in STM-D23580 (PheT and PheS) and ST-Ty2 (PheA), but not in STM-LT2 (Figure 5.16), as significantly modulating phenylalanine bands observed only in STM-D23580 and ST-Ty2 (Figure 5.10).

More evidently reduced abundance of nucleosides and nucleotides (NN) synthesis proteins was observed in STM-D23580 and ST-Ty2 (Figure 5.14B), similar to marked decreasing DNA/RNA bands in the Raman profile of the invasive strains (Figure 5.9). On the contrary, ATP binding proteins AtpA, NrdA and Ndk in STM-LT2 increased over time (Figure 5.16), in agreement with a higher DNA/RNA band in the Raman profile of STM-LT2 at the later time point of the infection (Figure 5.9).

Purine metabolism proteins (e.g. PurA, PurB, PurG, PurM) were differentially abundant in all the strains with a transient decrease and an upregulation at later time points (Figure 5.16). Purine nucleotides are thought to be critical for intracellular survival and bacterial virulence and induction of purine biosynthesis pathways was undetected for *Salmonella* cultured *in vitro*^{280,281}. The results from both the Raman

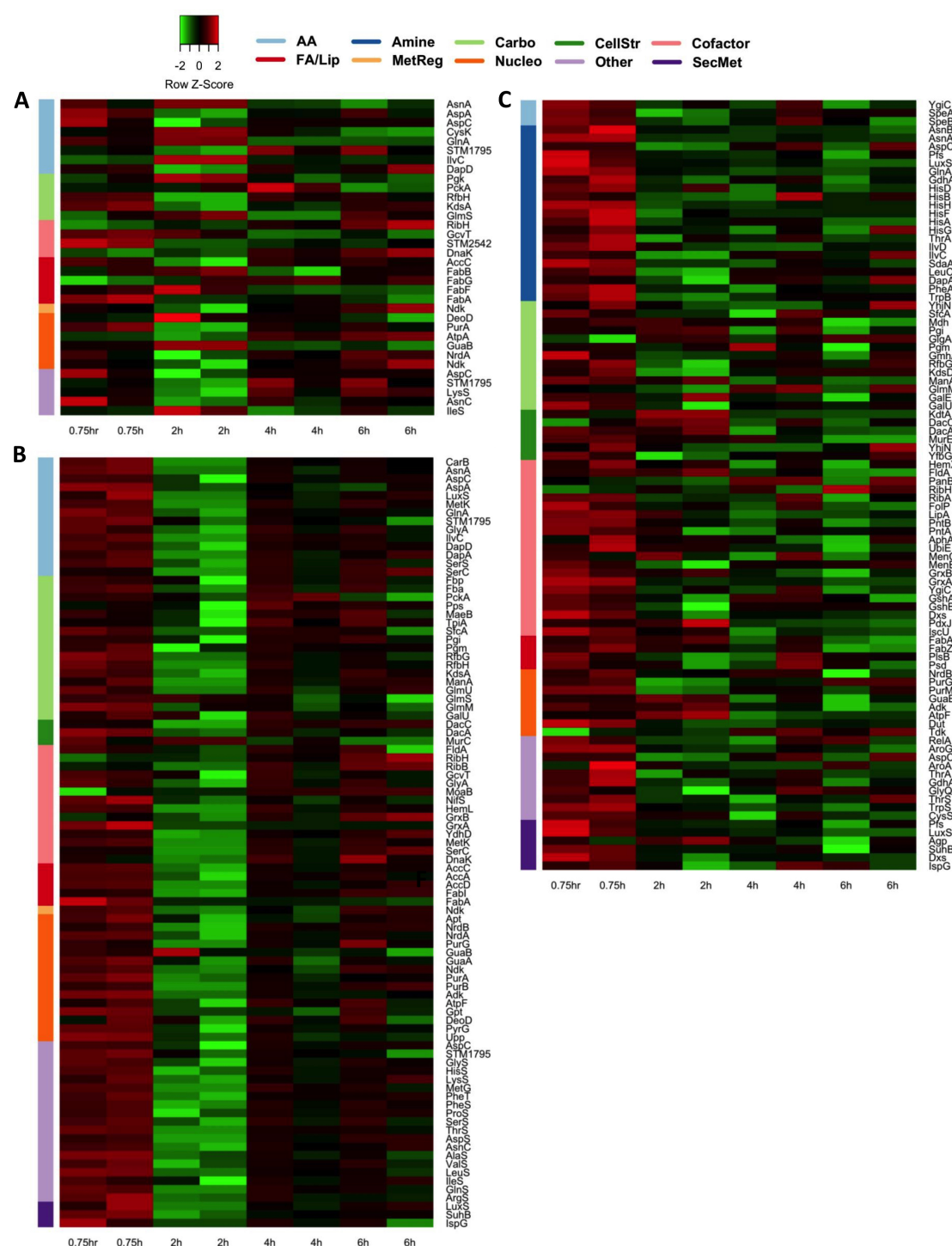


Figure 5.16: Heatmap of the row-normalised abundance of differentially expressed proteins (DEPs) related to biosynthesis pathways in (A) STM-LT2, (B) STM-D23580 and (C) ST-Ty2. Colour bars on the left present assignment of protein to one specific energy pathway and individual protein names are listed on the right. Abbreviations of the pathways are listed as follows: “AA” stands for “Amino Acid Biosynthesis”; “Amine” stands for “Amines and Polyamines Biosynthesis”; “Carbo” stands for “Carbohydrates Biosynthesis”; “CellStr” stands for “Cell Structures Biosynthesis”; “Cofactor” stands for “Cofactors, Prosthetic Groups, Electron Carriers Biosynthesis”; “FA/Lip” stands for “Fatty Acids and Lipids Biosynthesis”; “MetReg” stands for “Metabolic Regulators Biosynthesis”; “Nucleo” stands for “Nucleosides and Nucleotides Biosynthesis”; “Other” stands for “Other Biosynthesis”; “SecMet” stands for “Secondary Metabolites Biosynthesis”.

profile and the proteomics suggest an initial disruption of purine metabolism, possibly due to a host response, followed by active synthesis and consumption of purine sources at later stages.

The Fad-family proteins are known to be the key players in the metabolism of saturated and unsaturated fatty acids and are indicative of modulation of bacterial membrane fluidity in response to environmental stress²⁸². While all Fab proteins were downregulated in STM-D23580 and ST-Ty2, STM-LT2 exhibited an increased abundance of FabB and FabG, responsible for the *de novo* synthesis of unsaturated fatty acids via the *fabA-fabB* pathway^{282,283} (Figure 5.16). This suggests variability in the membrane composition between the invasive strains and the non-invasive strain, verified by the markedly increased Raman TUFA/TFA ratio in STM-LT2 (Figure 5.8).

Several proteins that regulate properties of cell membrane differed during infection in all *Salmonella* strains. For instance, RfpH and RfpG, involved in the biosynthesis of the O-Antigen, were downregulated at later time points in all strains (Figure 5.16). ManA, which is involved in the production of surface polysaccharide, was downregulated in both STM-D23580 and ST-Ty2. Some proteins related to cell structure biosynthesis including MurC, MurE, DacA and DacC, responsible for peptidoglycan biosynthesis, were also differentially expressed during infection with the invasive *Salmonella*. These findings suggest a dynamic rearrangement of surface antigen moieties directed by invasive *Salmonella* to resist the intracellular environment and avoid recognition.

5.4 Conclusions

Typhoid fever and invasive non-typhoidal salmonellosis (iNTS) are caused by *S. Typhi* and *S. Typhimurium*, respectively, but they share similar disease manifestations in humans like high fever, hepatosplenomegaly and infrequent diarrhoea. Despite these clinical similarities each serovar interacts differently with the human immune system. In this work, we used SCRM to describe the different metabolic profiles of the typhoidal *S. typhi* Ty2 and the non-typhoidal *S. Typhimurium* LT2 and D23580. The fingerprint region of SCRS shows that the three strains are metabolically different.

STM-D23580 has metabolic traits lying in between the non-typhoidal STM-LT2 and the virulent ST-Ty2 and it shares features in common with the typhoidal strain relative to the ability to achieve successful extraintestinal infection. A kNN machine learning classification achieved an accuracy of 92.4% correctly identifying these closely related strains based on their SCRS.

In order to survive and proliferate inside eukaryotic host cells, *Salmonella* and other intracellular pathogens must adapt their metabolism to available nutrients and to physical conditions. We used SCRM-DIP to characterise the metabolic activity of different *Salmonella* during infection of human cell lines and primary monocyte-derived dendritic cells and macrophages. The results unveil different metabolic activities of bacteria in liquid culture and in an intracellular environment, due to the adverse host environment. The slowest metabolism rate was found in the most invasive ST-Ty2, as opposed to its fastest extracellular growth. With the non-invasive STM-LT2 being the most active metabolizer, our results highlighted a possible strategy adopted by the invasive salmonellae to slow down central metabolism, hide from host surveillance and allow for further dissemination. A quantitative proteomic analysis of proteins related to bioenergetic functions confirmed the SCRM-DIP results, with most of the energy pathways being upregulated in STM-LT2 and downregulated in ST-Ty2.

Salmonella in vivo survival requires *de novo* synthesis of biocomponents such as proteins, carbohydrates, lipids and nucleic acids. A dynamic Raman profile of *Salmonella* inside host cells unveils the metabolic adaptations of these biocomponents during their intracellular life as well as the differences among three strains. The results suggest that the invasive *Salmonella* might switch off the glycolytic pathways and uses fatty acids as carbon sources instead. We also note a possible role of amino acids and purine nucleotides in contributing to different virulence phenotypes of *Salmonella*, which was further confirmed in the proteomic analysis. In addition, variations in subpopulations and diverse adaptations between in cell lines and in primary human cells were observed, suggesting the importance of choosing single-cell tools and correct cell culture models in the study of intracellular *Salmonella*.

In summary, we have demonstrated that single-cell SCRM-DIP technologies are capable of revealing metabolic changes, highlighting molecular heterogeneity and cell-to-cell variations among invasive *Salmonella* serovars. The dynamic Raman profiles offered sufficient information at the single-cell level, to unveil metabolic adaptations undertaken by *Salmonella* when transitioning from extracellular space to an intracellular compartment. Through comprehensive profiling and comparison of three different *Salmonella* pathovars, this study sheds light into bacterial metabolic traits and adaptations inside human host cells, which might contribute to the distinct pathogenicity of *Salmonella*. As several aspects of metabolism can influence antimicrobial susceptibility and tolerance, a better understanding of the *Salmonella*-host metabolic landscape might open new perspectives for antimicrobial development.

Chapter 6

Pipeline of SCRS Analysis

Contents

6.1 Introduction	125
6.2 Data tidying and data preprocessing	126
6.3 Data analysis	132
6.3.1 Data visualisation	132
6.3.2 Classification	140
6.3.3 Feature Extraction	151
6.4 Conclusion	156

In Chapters 3, 4 and 5, three applications of SCRM in investigating bacterial phenotypes are presented. In this Chapter, I aim to summarise and standardise the analytical pipeline of SCRS developed over the course of my DPhil study. By following this pipeline, I intend to enable a typical Raman analysis to be performed by a non-specialist user to achieve robust and high-quality visualisation, feature reduction and classification in a range of biological applications.

6.1 Introduction

Although the Raman scattering effect was first discovered in the 1920s, it has been traditionally applied in analytical chemistry. SCRM is still a novel tool yet to be fully validated in biological research laboratories. As a consequence, most of the data handling is done semi-manually, using programmes that come with the spectrometer manufacturer and with transfer of data between different software. This process is further complicated by the larger number of possible avenues for data analysis.

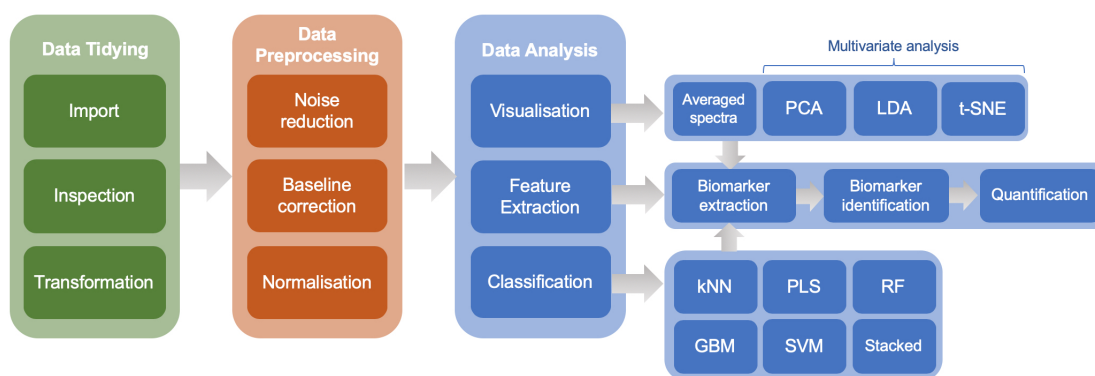


Figure 6.1: Raman analysis pipeline.

In this Chapter, I have rationalised a pipeline of Raman data analysis for biological subjects and developed a programming code in R language to automate execution of the pipeline. Immediately after spectra acquisition, this pipeline can be performed by running a single R script, attached in the Appendix B. The Raman analysis pipeline follows the flowchart in Figure 6.1, starting from raw-data tidying through to data preprocessing, and continuing on to various techniques of data analysis. It aims to fulfil both descriptive and predictive purposes in various biological research contexts. This pipeline will serve as a useful guide for Raman spectroscopic analysis in biological sciences. By implementing the established protocols, Raman analysis can be performed in a simple and rapid manner.

6.2 Data tidying and data preprocessing

For the purpose of illustrating the Raman analysis pipeline in this Chapter, a large number of spectra are obtained by measuring three developmental stages of stem cells, namely iPSCs (human induced pluripotent stem cells), NSCs (neural stem cells), and neurons. The stem-cell dataset that consists of 8774 SCRS (3316 iPSC, 2342 NSC and 3116 neuron spectra) will be used as an example to implement the R code and validate the Raman pipeline.

Immediately after acquiring the spectra, data tidying should be conducted to allow further processing and analysis as in the following steps:

1. *Import* the data into R

2. *Inspect* and screen the data of spectra for anomalies or poor qualities
3. *Transform* the dataset into appropriate forms for further processing

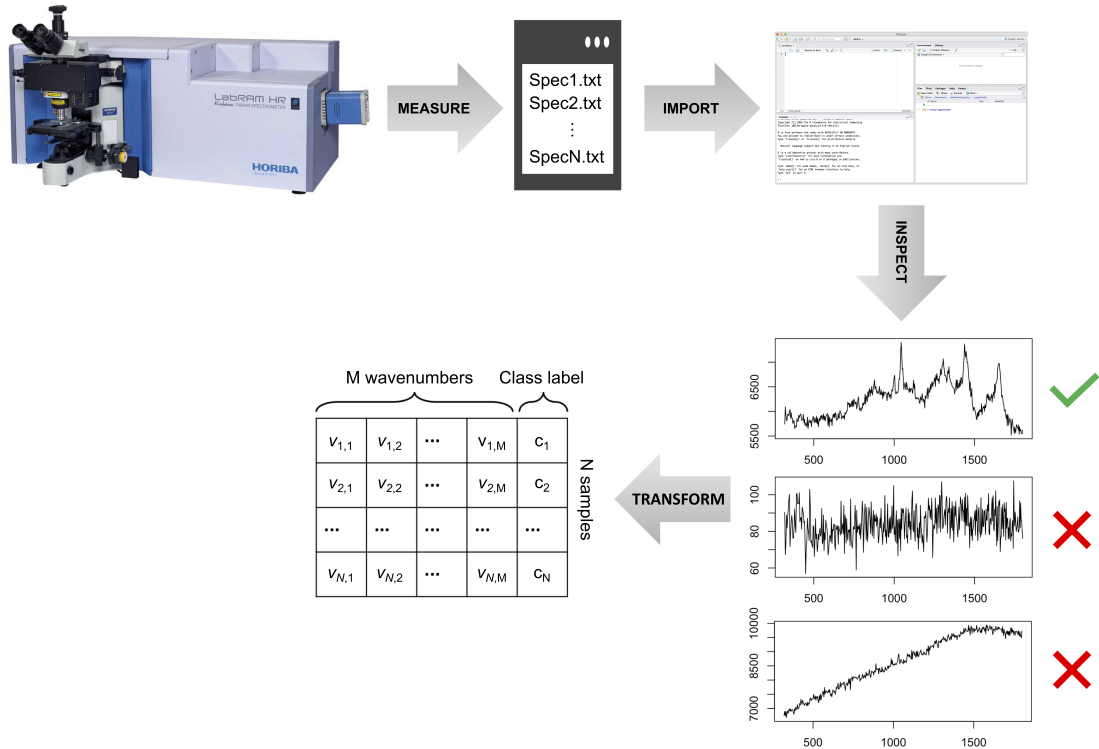


Figure 6.2: Data tidying process immediately after acquisition of Raman spectra.

The above steps are described in Figure 6.2 and exercised on the stem-cell dataset by using the code in Appendix B.1. After measurement with a Raman spectrometer, a list of files containing individual spectra can be exported into the computer. They are then imported into an R environment (such as RStudio). The data are visually inspected for quality, and spectra with high noise or fluorescence are screened out. Finally, the inspected dataset is transformed into an R-operable numeric matrix containing all spectral information. It should be noted that in biological contexts, one may sometimes desire to select a specific range of wavenumbers such as the fingerprint region from $300 - 1800 \text{ cm}^{-1}$ that confers most of the biological information. This step is carried out in the stem-cell dataset to focus on the fingerprint region of the stem-cell spectra as well as to reduce dimensionality and complexity of the data for further analysis.

After tidying the raw data, data preprocessing steps should be followed. The selection of prep-processing steps and the order in which they are performed has been shown to have a profound impact on the results of the spectral analysis⁹⁷. Therefore, I recommend that a uniform approach should be applied to all data for reproducibility and robustness. The recommended preprocessing steps are as follows and are executed by the R code in Appendix B.2:

4. Conduct a *noise-reduction* technique to account for poor SNR
5. Choose and apply a *baseline-correction* algorithm to account for fluorescence interference
6. Perform a spectral *normalisation* to account for sample variations

Noise reduction

Raman spectra are sometimes prone to noise and may require noise-reducing techniques to improve the signal-to-noise ratio (SNR). Mathematical methods include Savitzky–Golay smoothing²⁸⁴, wavelet de-noising⁹¹ and principal component analysis (PCA)⁹⁷. However, poorly practised smoothing can significantly reduce spectral resolution as well as biochemical features. The preferred approach to adjusting spectral quality is always to alter instrumental settings, such as by using higher laser power and longer acquisition time. Therefore, this step should only be performed if absolutely necessary. In the stem-cell dataset, noise reduction is not applied.

Baseline correction

Raman signals can be easily corrupted by a fluorescent sample background, as well as thermal fluctuations of the CCD, both of which markedly affect the spectral baseline. Therefore, the first essential step of data prep-processing is always to perform a baseline-correction algorithm to suppress the spectral background. Various mathematical methods are available for correcting spectral baselines, including least squares²⁸⁵, wavelet transform²⁸⁶, median filter²⁸⁷, polynomial fitting (one of the most common forms of baseline correction due to its simplicity and convenience of its many different variants²⁸⁸) and rolling ball²⁸⁹.

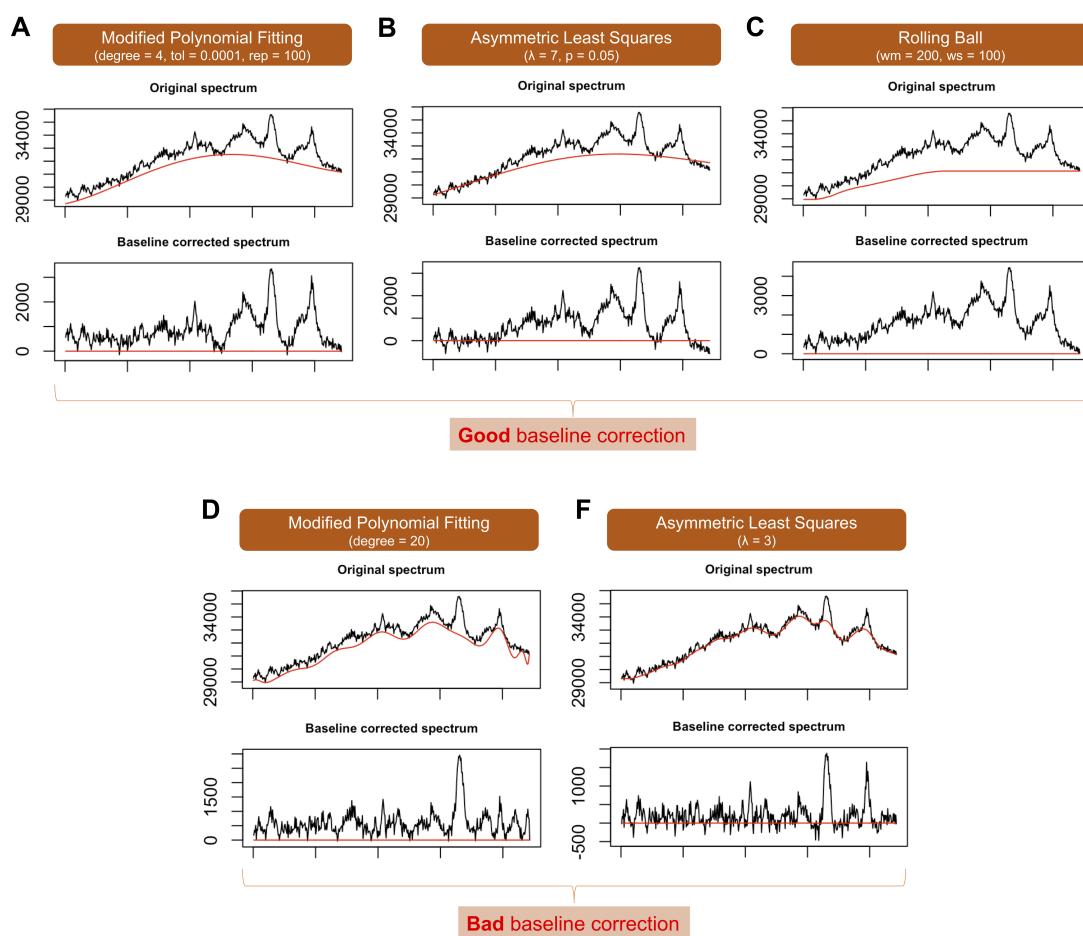


Figure 6.3: Comparison of an original spectrum and the baseline-corrected spectrum with the use of (A) modified polynomial fitting (degree = 4), (B) asymmetric least squares ($\lambda = 7$) and (C) rolling ball (wm = 200, ws = 100). Two badly corrected examples are given in (D) using modified polynomial fitting (degree = 20) and (E) using asymmetric least squares ($\lambda = 1$).

It is down to the preferences of individual researcher to choose baseline-correcting algorithms. Each algorithm will require careful tuning of parameters, which consequently decides the outcomes of the corrected spectra. In the stem-cell dataset, I have applied and tuned three baseline-correcting algorithms: modified polynomial fitting, asymmetric least squares, and rolling ball using the code described in Appendix B.2.1. In Figure 6.3A-C, one spectrum with a high fluorescence background is compared with its baseline-corrected version using each of the three algorithms. With careful tuning, all algorithms achieve good baseline-correction results.

Modified polynomial fitting was introduced by Lieber and Mahadevan-Jansen²⁹⁰ to make an iterative polynomial curve-fitting that assigns a n th degree polynomial to

the spectra. Three parameters can be adjusted and they are tuned in the example as the degree of the polynomial function (degree = 4), tolerance of difference between iterations (tol = 0.0001), and maximum number of iterations (rep = 100).

Asymmetric least squares is described by Eilers and Boelens²⁸⁵. It uses an iterative least-squares algorithm applying smoothing by adding a 2nd-derivative constraint and asymmetrically weighting residuals from previous iterations. Two parameters are tuned: 2nd-derivative constraint as the amount of smoothing ($\lambda = 7$) and weighting of positive residuals from previous iterations ($p = 0.05$).

The rolling ball algorithm is introduced by Kneen and Annegarn in 1996²⁸⁹. Originally proposed for fitting X-ray spectra backgrounds, this algorithm imagines a large ball rolling underneath a spectrum where the trace of the topmost point of the ball fits the spectral baseline. Two parameters are adjusted for fitting this algorithm to the dataset: width of local window for minimisation/maximisation (wm = 200) and width of local window for smoothing (ws = 100).

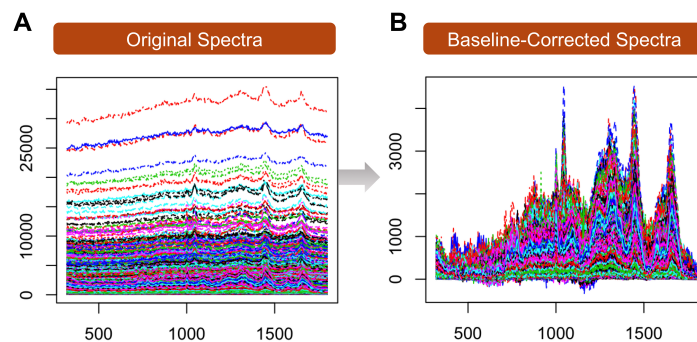


Figure 6.4: The original spectra (A) are baseline-corrected (B).

Care should always be taken with all baseline-subtraction routines as they can introduce unintended artefacts. Figures 6.3D and F present two examples of bad practices of modified polynomial fitting (degree = 20) and the asymmetric least squares algorithm ($\lambda = 3$). The applied algorithms with unsuitable parameters distort the spectrum and degrade the spectral features. Therefore, cautious tuning and application of any baseline-correction algorithm are suggested at all times.

After tuning and deciding on one baseline-correction method, the chosen method can be applied to the whole Raman dataset. Figure 6.4 shows the transformation of the original spectra (A) in the stem-cell dataset into the baseline-corrected spectra (B).

Normalisation

After baseline correction, the spectra will require normalisation to account for any sample and experimental variances, such as varying concentrations of analyte across the sample, and the thickness of the sample. In biological studies, leaving spectra non-normalised for further analysis will highlight spectral differences based primarily on Raman intensity, whereas normalisation will reveal biochemical differences from cell to cell and group to group. To minimise experimental variances, all spectra should be scaled according to a specific criterion.

Normalisation to a particular peak (e.g. Amide I or Amide II peak) is applied in some Raman and IR studies^{97,291}. However, I do not recommend using this method in biological studies as the content of one specific group of biomolecules (e.g. proteins that are representative of the Amide peaks) can vary significantly in cells with different physiologies and metabolic states.

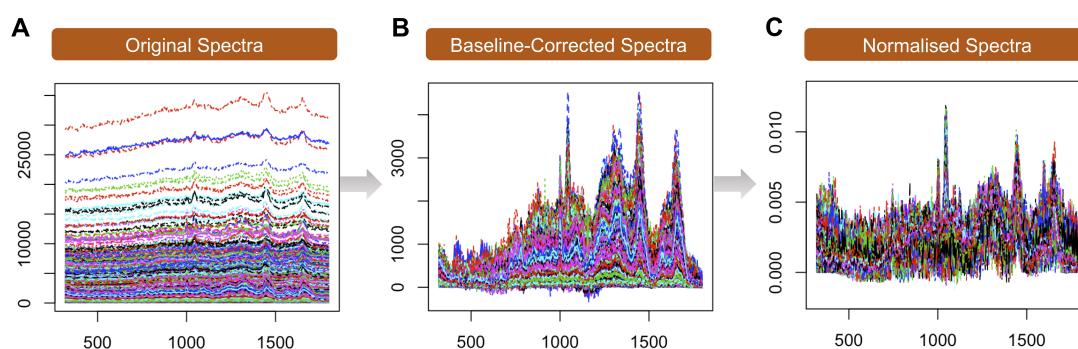


Figure 6.5: The original spectra (A) are baseline-corrected (B) and then normalised (C).

Area normalisation (L_1 -norm) and vector normalisation (L_2 -norm) are two popular methods that can be applied to spectra²⁹². In the stem-cell study in this Chapter and all studies in other Chapters, area normalisation is carried out as part of the data prep-processing using the R code described in Appendix B.2.2. Figure 6.5C shows the normalised spectra after baseline correction and normalisation. With area

normalisation, the sum of all intensity values is uniform in each spectrum. It is useful in single-cell studies as it assumes equal cellular content in all spectra, therefore, allowing comparison of intracellular biomolecules across different cells or groups.

6.3 Data analysis

Data analysis with various mathematical and statistical methods is necessary to efficiently access spectral data, describing the underlying biological properties and relationships, and predicting the target features and behaviours. The purpose of analysing a complicated Raman dataset varies greatly from study to study. This section divides the analytical techniques into three purposes that fit into most biological investigations: visualisation, classification and feature extraction. For biological analysis, the combination of different techniques is advisable in order to scrutinise a complex biological system phenotyped by a Raman spectrum.

6.3.1 Data visualisation

"The simple graph has brought more information to the data analyst's mind than any other device." —John Tukey

Data visualisation is important. It is an important aspect of visual communication in many disciplines, including biosciences. Due to the way the human brain processes information, graphical forms convey messages more clearly and effectively than spreadsheets of words and values. In the numeric matrix that we created in the last section, patterns, trends, correlations and differences are implicit and undetectable by a human. Data visualisation can reveal those features by placing it in a visual context.

Data visualisation is especially important in Raman analysis due to the high dimensionality (hundreds and thousands of wavenumbers) of a Raman dataset. A human's ability to visualise is limited to two or three dimensions. Therefore, different visualisation approaches and/or dimension-reducing techniques are explored in this section and performed by executing the code in Appendix B.3.1.

Averaged-spectra visualisation

The most intuitive way to visualise one spectrum is to treat it as a dataset with two variables: wavelength and intensity. In such a way, one can visualise a spectrum in a two-dimensional plot. With an increasing number of spectra, it becomes impossible to plot all of them in one graph. One common approach to visualising a large number of spectra is averaging them according to different subgroups. In this stem-cell dataset, all spectra in the group "iPSC", "NSC" or "Neuron" are averaged and plotted (Figure 6.6A). Often, spectra from different groups are manually shifted along the intensity axis to aid visualisation.

One drawback of averaging the spectra is the loss of single-cell information. To compensate for this drawback, I suggest a small modification to the spectra-plotting routine. The standard deviation from the single-cell population can be plotted as shades around the averaged spectrum (Figure 6.6B). Variations at one single Raman band among the population can now be visualised and spotted. The code for plotting the averaged spectra and the averaged spectra with standard deviation is included in the Appendix.

Multivariate visualisation

Another less intuitive but more informative way to look at a spectral dataset is to treat it as a multi-dimensional dataset, in which each wavenumber represents a discrete variable. By analysing spectral data in this way, we are able to extract a smaller number of variables that are more important out of a pool of original wavenumber-variables.

Depending on the instrumental parameters, the number of wavenumber-variable can range from hundreds to thousands. In this dataset, 1019 wavenumber-variables are present in one complete spectrum and 443 wavenumber-variables are present in the fingerprint region of one spectrum. Our ability to visualise is limited to two or three dimensions. Multivariate, dimension-reducing techniques are essential when reconstructing data into a dataset with a smaller number of variables and visualising data in a two-dimensional graph.

Multivariate visualisation - PCA

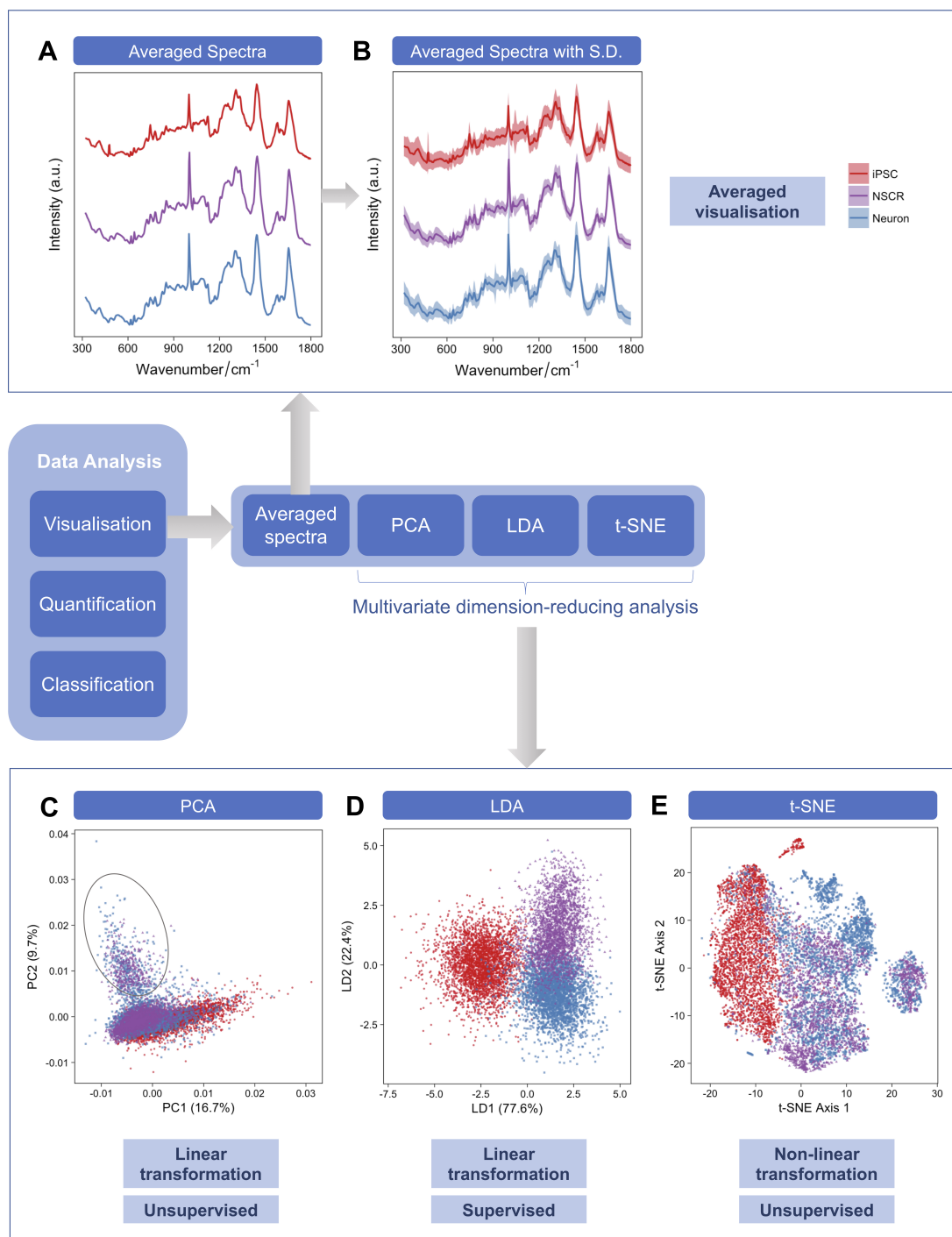


Figure 6.6: Data visualisation by (A) averaged spectra, (B) averaged spectra with standard deviation, (C) PCA, (D) LDA and (E) t-SNE.

Principal component analysis (PCA) is often the first analysis performed by Raman researchers on a new spectral dataset⁹¹. It is an unsupervised dimension-reducing technique that linearly transforms the original variables to new variables as principal components (PCs) and each PC is a linear combination of the original wavenumbers. This new coordinate system is better at conveying information present in the data compared with the original measurement variables, as it filters background noise and retains only the most important features. In addition, Raman statistical models suffer from high correlation between covariates (wavenumbers that are correlated with each other), and techniques like PCA can be used to produce linear combinations of variables that are uncorrelated with each other.

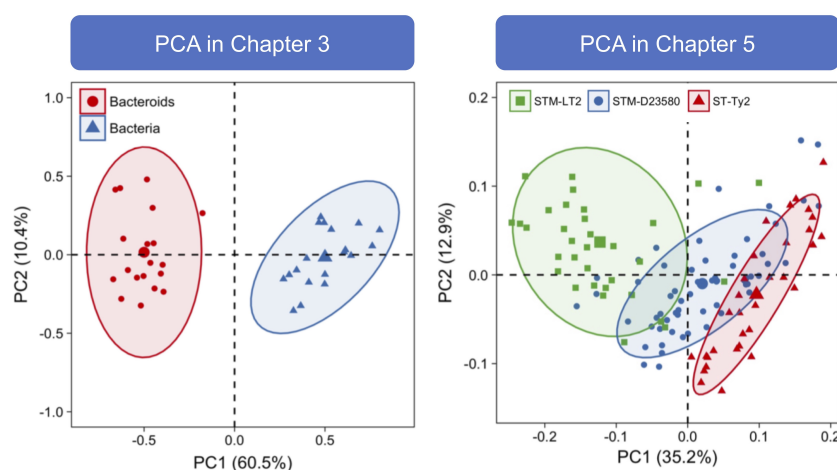


Figure 6.7: Examples of PCA plots from Chapter 3 and Chapter 5.

Often, in relatively homogeneous biological populations with little difference within the same group and large differences among different groups, PCA of the Raman dataset is sufficient to create a visualisation with well-separated clusters; for example, the PCA plots illustrated in Chapter 3 and Chapter 5 (Figure 6.7).

However, PCA-aided visualisation can sometimes achieve poor results. In Figure 6.6C, the PCA plot based on PC1 and PC2 of the stem-cell dataset reveals overlapping populations rather than distinct clusters. The limitation of PCA on this particular dataset can be explained by two key factors (Figure 6.8):

possible separation between three groups (iPSC, NSC and Neuron). The resulting LDA plot is shown in Figure 6.6D. Compared to the PCA plot in Figure 6.6C, it displays more distinctly isolated clusters of different groups. In addition, the first two LDs succeed in explaining 100% of the variances in the data (77.6% by LD1 and 22.4% by LD2).

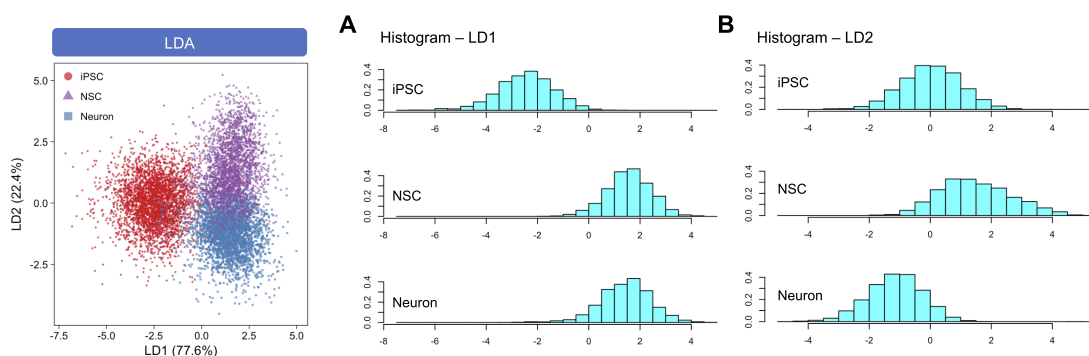


Figure 6.9: Stacked histograms of the values of (A) the first discriminant function and (B) the second discriminant function from the LDA of the Raman spectra in different groups.

Besides the scatter plot of individual data points, another effective way of visualising the results of an LDA is to make a stacked histogram of the values of the discriminant function for the individual spectrum from different groups. In Figure 6.9A, the histogram of LD1 shows that the first discriminant function nicely separates the iPSC group from the others. The histogram of LD2 in Figure 6.9B, on the other hand, reveals differences between the spectra of iPSCs and NSCs and the spectra of neurons.

Despite a more satisfying visualisation result of LDA over PCA on the stem-cell Raman dataset, one must bear in mind that PCA is unsupervised while LDA is a supervised approach. As in any supervised method, LDA can overfit small datasets and having the number of spectra at least five times bigger than the number of variables can potentially avoid this problem⁹⁴ as in this example where we have 443 variables and 8774 Raman spectra.

Multivariate visualisation - t-SNE

t-distributed Stochastic Neighbor Embedding (t-SNE) is a more recent algorithm for dimension reduction and visualisation developed by Laurens van der Maaten and Geoffrey Hinton²⁹³. t-SNE is a non-linear technique. It is often compared with PCA for embedding high-dimensional data in a low-dimensional space of two or three

dimensions for visualisation and it has been shown to outperform PCA in many fields. Nowadays, t-SNE is extensively applied in many fields such as image processing, natural language processing, speech processing and bioinformatics such as genomics, transcriptomics and mass spectrometry imaging^{294–299}. However, its applications to Raman studies are still limited and under-explored. Here, I introduce the potential for applying the powerful t-SNE on a biological Raman dataset for visualisation.

Unlike PCA or LDA, the preservation in t-SNE is non-linear: it minimises the Kullback-Leibler divergence between two distributions. The first distribution measures pairwise attraction of individual data points in the high-dimensional space, and the second heavy-tailed Student's t-distribution that measures pairwise attraction of data points in the low-dimensional embedding space. The most important parameter to tune in t-SNE is perplexity. Perplexity governs how many of its nearest neighbours each point is effectively attracted to. Here in this example, the perplexity is set equal to 100 and the resulting visualisation is shown in Figure 6.6E.

Visualisation based on t-SNE with multiple perplexity values is also performed, from perplexity = 5 to perplexity = 500, and compared with visualisation based on PCA (Figure 6.10). Overall, all t-SNE visualisation demonstrates much improvement with clearly separable clusters over PCA visualisation. In this dataset, it is observed that higher perplexity value leads to a more preserved local structure where compact clusters with a small number of data points exist; and with lower perplexity, a more global structure is preserved where there is more sparseness between data points and between clusters. The visualisation with a perplexity value of 50 or 100 is slightly better in terms of the separation between each cluster as well as the preservation of the global structure.

In all t-SNE figures, the iPSC cluster is well separated while individuals from the other two groups entangle and overlap with each other. Biologically, iPSCs are the most homogeneous population before any induced proliferation or development; therefore, Raman spectra of iPSCs are of least variances and form distinct clusters. NSCs and neurons, on the other hand, are two developmental stages of the stem cells and are often mixed because each individual cell proliferates and develops into a neuron at a

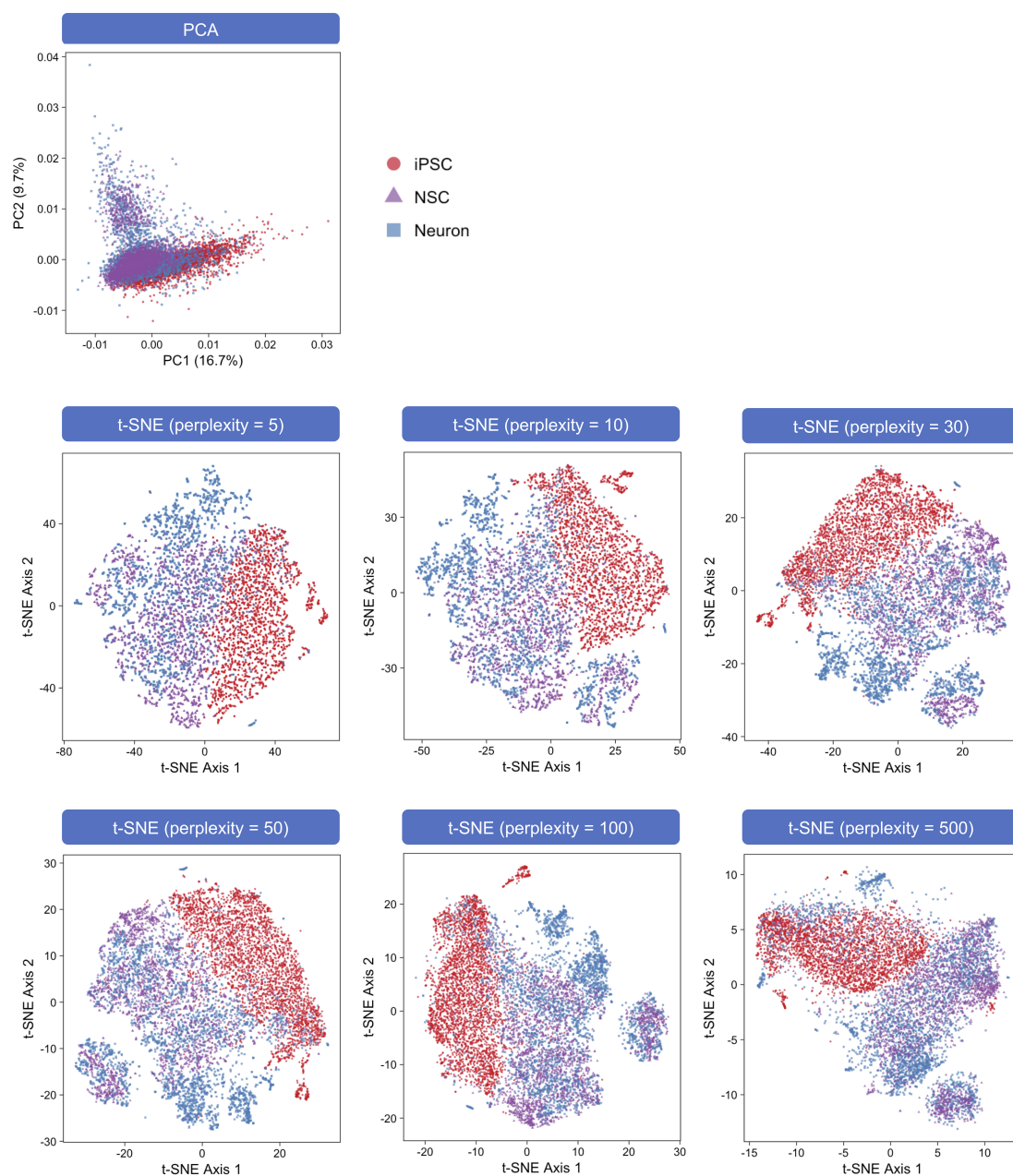


Figure 6.10: Visualisation results of PCA and t-SNE with a perplexity value of 5, 10, 30, 50, 100 or 500.

different speed. Therefore, sub-clusters exist in each of the two groups and overlapping of the two clusters is observed. Compared to t-SNE, neither PCA nor LDA shows visual results that preserve these biological features of the cells.

In contrast to supervised learning methods, PCA and t-SNE are unbiased, unsupervised approaches. Compared with a standard PCA that is extensively used in current Raman studies, t-SNE produces a two-dimensional visualisation with more distinctly isolated clusters, shows high tolerance towards outliers that are common in real-world samples, and preserves a better global structure of the data. Therefore, I recommend the use of t-SNE in a standard data analysis pipeline to aid visualisation of a multi-dimensional dataset of Raman spectra.

Because t-SNE maps the multi-dimensional data to a lower dimensional space non-linearly, the limitation of t-SNE is that the input features are no longer identifiable. Thus one cannot make any inference based only on the output of t-SNE, meaning that it is essentially mainly a data exploration and visualisation technique. However, t-SNE can be used prior to the process of classification and clustering by using its output as the input feature for other classification algorithms.

6.3.2 Classification

After data visualisation that shows a promising distinction between different groups, classification based on samples' spectra is often desirable for diagnostic and practical purposes. For example, the use of stem cells in clinical therapy has been hindered partly because there is not yet a standardised technique to characterise and classify the developmental stages of cells. If a Raman spectrum of a cell can point to a specific proliferation stage (iPSC, NSC, or neuron), it will be of remarkable benefit to the quality control process of stem-cell transplants.

Supervised classification depends on machine-learning algorithms, known as classifiers, that are able to understand and learn patterns from complex data. When encountering new data, the classifier will associate the patterns with the data and try to assign a diagnosis. This approach can be applied to Raman spectroscopic studies; the pipeline of using machine-learning models to classify samples based on

their Raman spectra is shown in Figure 6.11, and is performed by executing the R code in Appendix B.3.2.

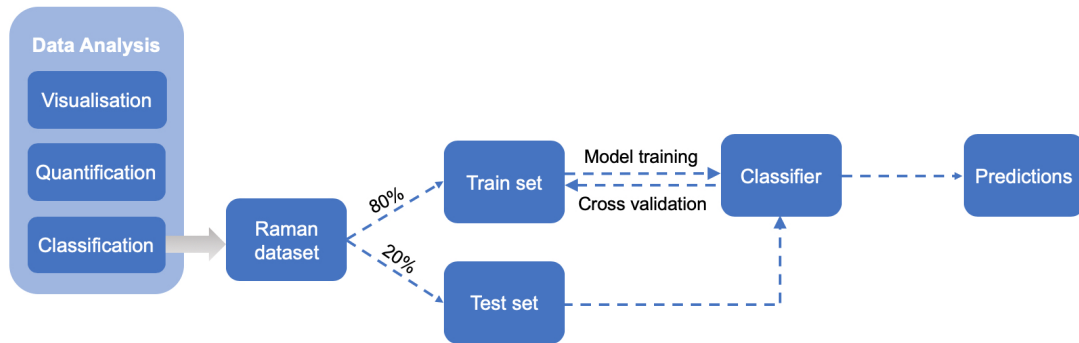


Figure 6.11: The pipeline of machine-learning classification using Raman spectra.

Data splitting

As in any machine-learning classification practice, it is important to avoid bias. A model will most certainly perform better when it is tested by spectra that were used when creating the model. The true performance of the model should be evaluated with spectra from an independent batch of samples. Therefore, data splitting should be the first step in the process of classification to divide the data into the "training set" and the "test set". Both the training set and the test set should be created by random sampling and should preserve the overall class distribution of the data. An advisable split would be 80/20% so that 80% of the spectra constitute the training set (7020 spectra in this example) and 20% of the spectra go to the test set (1754 spectra in this example).

Model training with the training set

In the training set, each spectrum has a user-defined class label (in this example it will be "iPSC", "NSC" or "Neuron"). The training set is then used for model training and parameter tuning. The process involves:

1. Choose a specific classifier/algorithm that is to be used to create the model.
2. Define a set of parameters to evaluate in the model. For example, if fitting a k-Nearest Neighbour (kNN) model, the range of the value of k (number of nearest neighbour) should be defined.

3. Specify the type of resampling. During model training, the training set is divided into subsets of data to cross-validate the model for best parameters and best performance. A k-fold cross-validation is a common approach. For a smaller dataset, leave-one-out cross-validation can also be implemented.
4. For each parameter set and each resampling iteration, the model holds out specific samples controlled by k-fold cross-validation, fits the algorithm on the remainder and finally predicts the hold-out samples.
5. Calculate the average performance across hold-out predictions.
6. Determine the optimal parameter set.
7. Fit the final model to all data in the training set using the optimal parameter set.

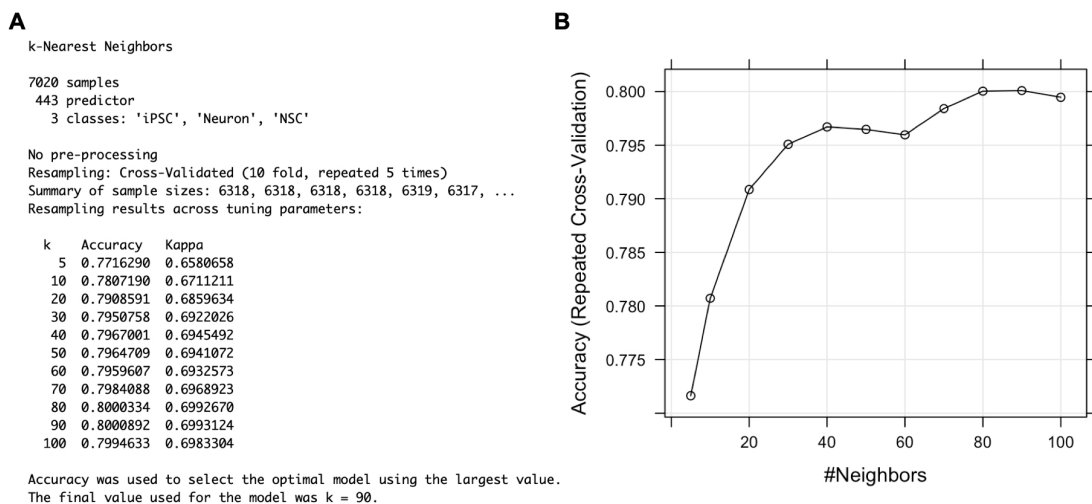


Figure 6.12: An example of model building and parameter tuning process using kNN.

An example of the results of the model-building and parameter-tuning process is shown in Figure 6.12, following the steps:

1. The classifier chosen to create the model is k-Nearest Neighbour (kNN), in which the class of one sample is decided by its k number of nearest neighbours.
2. The parameter to tune in kNN is the value of k which is set to be 5, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100.

3. The training set is resampled into 10 subsets for 10-fold cross-validation and this process is repeated 5 times.
4. For each k and each resampling iteration, the model holds out 1 subset and fits the algorithm on the remaining 9 subsets and finally predicts the hold-out samples.
5. The average accuracy across hold-out predictions is calculated for each k .
6. The best accuracy is found when $k = 90$.
7. The final model achieves an accuracy of 80.0% when it is fit to all data in the training set.

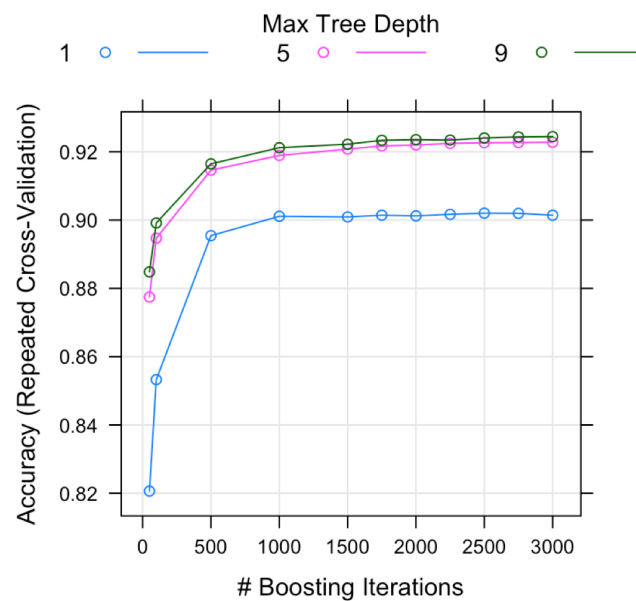


Figure 6.13: An example of the results of the model building and parameter tuning process using GBM.

Some aspects of the parameter-tuning process are worthy of note. Usually, finding an optimal set of parameter values for a learning algorithm is a time-consuming task. Grid search is the parameter-tuning method of choice in this thesis. In grid search, the user defines a set of candidate values to explore (in the same way as in the kNN model training above) and the algorithm will exhaustively try every possible combination

of the values of individual parameters. Grid search is usually a good choice when there are a small number of parameters to optimise. The main problem is that grid search is an exponential time algorithm - the time cost grows exponentially with the number of parameters. In the kNN model above with 1 parameter and 11 grid values, the algorithm runs in $O(11)$ time considering $O(1)$ time for each of the values. Turning to another model that has been trained using a stochastic gradient boosting machine (GBM) where 2 parameters are tuned; 3 max tree depths and 11 number of boosting iterations (Figure 6.13). This model runs in $O(33)$ time. The complexity increases exponentially with more parameter choices. When there are p parameters and each one takes at most v values, the model runs in $O(v^p)$ time. In this situation, one can try to run the model multiple times with different grids. For example, in the GBM model, if we set the first grid of boosting iterations as [1000, 2000, 3000] and the best models return with the value of 3000, we should try to refine our grid values by zooming in around 3000. To do this, we set up a new grid search and redefine the number of boosting iterations as [2500, 2750, 3000, 3250, 3500]. In this case, a good combination of parameters is found in $O(24)$ time instead of $O(33)$ time in the one grid-search.

Using the above model training, parameter tuning and cross-validation approach, various models are trained based on the training set of the Raman spectral data, all of which have been introduced in Chapter 2, Section 2.3.4. Eight models are created using k-Nearest Neighbour (kNN), Linear Discriminant Classifier (LDC), Partial Least Squares (PLS), Random Forest (RF), stochastic Gradient Boosting Machine (GBM), eXtreme Gradient Boosting (XGB), Support Vector Machine with linear kernel (SVM-linear) and Support Vector Machine with Radial Basis Function kernel (SVM-RBF). In Table 6.1, 8 models are described and the optimal parameters after the tuning process are listed. The training speed of each model is also compared.

The simple kNN model behaves poorest among all models with an accuracy of 80.0%. Two linear classifiers, LDC and PLS, have good accuracy of 90.1% and 90.6% respectively, while outperforming all others in terms of training speed. RF, GBM and XGB are three ensemble techniques that, although shown to have better predictive power than a single Decision Tree classifier, suffer from the lengthy training process.

Table 6.1: Models trained in this Chapter

	Classifier	Optimal Parameters	Accuracy on training set	Training speed
kNN	k-Nearest Neighbour	k = 80	80.0%	Fast
LDC	Linear Discriminant Classifier	Dimension = 2	90.1%	Very fast
PLS	Partial Least Squares	ncomp = 11	90.6%	Very fast
RF	Random Forest	mtry = 443 split.rule = extratrees min.node.size = 1	87.4%	Slow
GBM	Stochastic Gradient Boosting Machine	n.trees = 3000 interaction.depth = 9 shrinkage = 0.1 n.minobsinnode = 10	92.4%	Very slow
XGB	eXtreme Gradient Boosting	nrounds = 150 max.depth = 3 eta = 0.4 gamma = 0 subsample = 1 colsample.bytree = 0.8 rate.drop = 0.01 skip.drop = 0.95 min.child.weight = 1	90.6%	Very slow
SVM-linear	Support Vector Machine with linear kernel	cost = 0.25	88.4%	Moderate
SVM-RBF	Support Vector Machine with Radial Basis Function kernel	sigma = 0.001603621 C = 1	92.6%	Moderate

These models achieve an accuracy of 87.4%, 92.4% and 90.6%, respectively, on the training set of the stem-cell dataset. In cases of the SVMs, although a linear SVM (88.4%) does not perform as well as the linear classifiers and two of the ensemble models, an SVM with an RBF kernel exhibits the best predictive performance with an accuracy of 92.6% on the training set. The SVM-RBF model also has the additional merit of moderate time cost, compared with the extensive time spent for training a GBM that has a similar performance score.

This result agrees with a lot of Raman studies that utilise SVMs to build classification models in biological studies including identification of bacteria, classification of

proteins and nucleic acids and disease diagnosis^{72,82,104,113,116,300–303}. Due to their good performance and the relatively easy training process, I suggest that SVMs can serve as good candidates when building a machine learning model based on samples' Raman spectra.

Model predictions with the test set

After training a model with optimal parameter sets, the true performance of the model should be tested on an independent dataset - the test set. The test set will be input into the model and the model will assign a class estimation to each spectrum in the test set. The model performance will then be evaluated by comparing the estimated class to the actual class of one spectrum.

The performance of the eight models created above is evaluated individually and summarised in Figure 6.14. In each model, the predictions of the classes of 1754 Raman spectra are compared with their true classes. The overall accuracy and the accuracy per class (iPSC, NSC or Neuron) are calculated for each model.

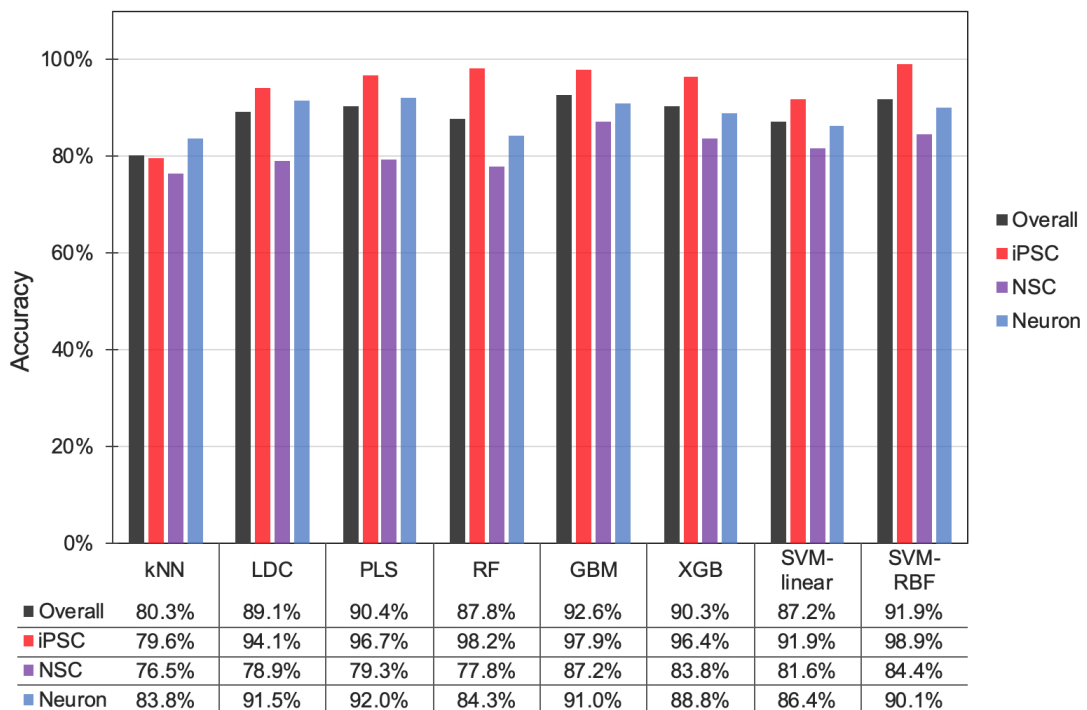


Figure 6.14: Performance of 8 models on stem-cell dataset.

Let us turn first of all to the overall accuracy obtained by each model. Upon predict-

ing the classes of the test set, GBM and SVM-RBF remain the best performers, achieving an accuracy of 92.6% and 91.9%, respectively. All models achieve a similar accuracy score on the test set compared to that on the training set. This illustrates that all models adequately capture the underlying structure of the data without corresponding too exactly to the training set, in another word, overfitting. Overfitting can often be problematic, especially when dealing with a Raman dataset with a large number of parameters and a small number of samples. It can be checked by comparing the performance of the model on the training set against the performance on the test set. An overfitted model has optimal performance on a previously seen training set and its suitability drops considerably when dealing with an independent dataset. The challenge of overfitting a machine model is overcome in this case by using a Raman dataset with a much larger number of observations (8774) compared to the number of wavenumber-variables (449).

All models achieve the best performance when predicting spectra from the iPSC group (except for kNN) and the worst when dealing with spectra in the Neuron group. The greatest difference is seen in RF where an accuracy of 98.2% is achieved for the iPSCs but only 84.3% of the Neuron spectra are correctly predicted. This result is compatible with the visualisation results using t-SNE (Figure 6.10) and the biological stem that the NSC are the developmental stage between iPSC and neuron.

Model stacking

After understanding and building single classification models based on Raman spectra, I would like to introduce a new concept in Raman machine learning: stacking. Stacking is an ensemble technique to build a model that combines multiple models (typically of differing types) and a supervisor model that learns how to best combine the predictions of the primary models. Rather than using a single classifier, which is common in Raman practices, the stacked model will usually perform better than the primary models due to its ability to highlight each base model where it performs best and discredit each base model where it performs poorly. For this reason, stacking is most effective when the primary models are significantly different. We have recently published a paper which implements stacking classification to successfully identify

aneuploid or euploid embryos by Raman spectra of embryo culture medium with an accuracy of 95.9%³⁰⁴.

Here, we try to build a stacked model based on the eight models created in the last section. Before stacking, the correlations between models should be examined. It is desirable that the predictions made by the eight primary models have low correlations (< 0.75). This would suggest that the models are effective but in different ways, allowing the stacking classifier to get the best from each model for an improved score. Otherwise, with highly correlated predictions (> 0.75) among models, primary models would be making the same or very similar predictions most of the time, thus reducing the benefit of stacking multiple models.

The correlation between models is investigated by resampling the predictions by 50 subsets, and correlation scatter plots are made according to the resampling accuracies in Figure 6.15. As we can see from the plots, each model has a tight resampling profile (similar accuracies) and each two of the models are relatively uncorrelated (a high correlation would be indicated by a linear regression line through the dashed line). The models' correlations are then confirmed in Table 6.2. We can see that all pairs of predictions have generally low correlation. The two algorithms with the highest correlations between their predictions are kNN and SVM-linear at a correlation coefficient of 0.180 which is still considerably low, given a high correlation is considered to be > 0.75 .

Table 6.2: Resampling correlations of the models trained in this Chapter

	KNN	LDC	PLS	RF	GBM	XGB	SVMlin	SVMRBF
KNN	1	-0.053	-0.030	-0.008	0.167	0.138	0.180	0.005
LDC	-0.053	1	0.029	-0.060	-0.232	0.163	-0.066	0.026
PLS	-0.030	0.029	1	-0.235	-0.040	0.053	-0.109	-0.115
RF	-0.008	-0.060	-0.235	1	0.274	-0.127	-0.118	0.133
GBM	0.167	-0.232	-0.040	0.274	1	0.015	-0.125	-0.088
XGB	0.138	0.164	0.053	-0.127	0.015	1	0.142	-0.210
SVMlin	0.180	-0.066	-0.109	-0.118	-0.125	0.142	1	-0.172
SVMRBF	0.005	0.026	-0.115	0.133	-0.088	-0.210	-0.172	1

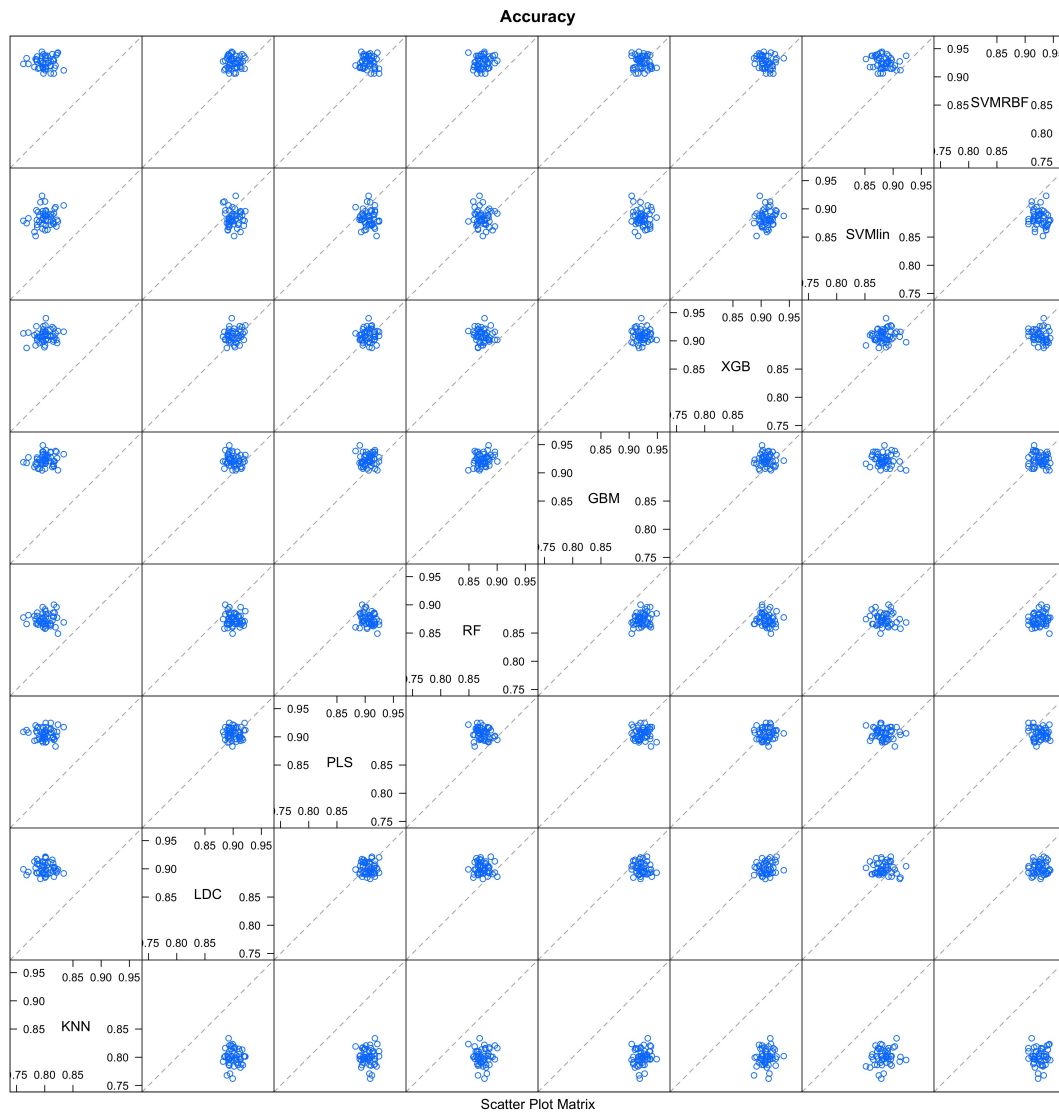


Figure 6.15: Accuracy scatter plots of 50 resampled predictions between pairs of the eight primary models.

After confirming the low correlations between the pairs of the primary models, we initiate the stacking progress. The main strategy is to use the predictions of the base models as input features for the stacked model. Another algorithm should be chosen to discern where each model performs well and where each model performs poorly. Here, GBM is used to train the stacked model. The same resampling and cross-validation profile are performed and the final parameters selected are $n.trees = 50$, $interaction.depth = 3$, $shrinkage = 0.1$, and $n.minobsinnode = 10$.

As we can see in Figure 6.16, the stacked model achieves an accuracy of 95.3%,

exceeding all primary classifiers which have an averaged accuracy of 88.7% and a best accuracy (SVM-RBF) of 91.9%. A significant improvement is also seen in terms of class differentiation. In the class of NSC that has been poorly predicted by the primary models, the percentage of correctly predicted NSC spectra increases from an average of 81.2% to 92.1% in the stacked model.

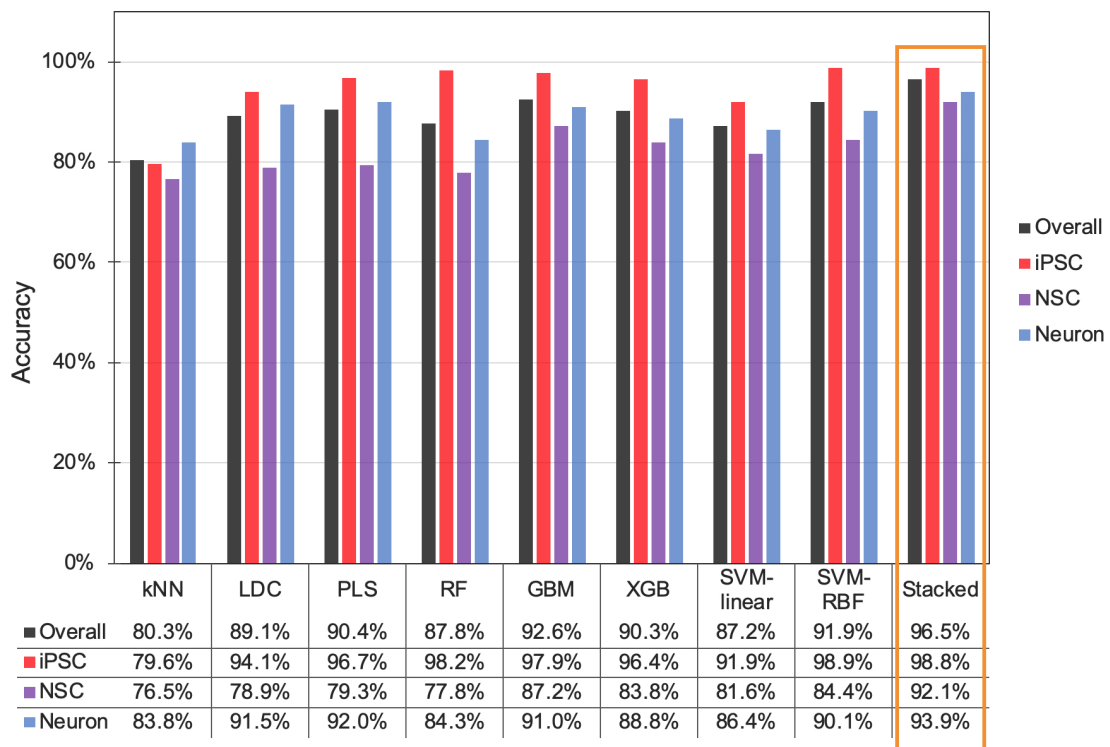


Figure 6.16: Performance of eight models and the stacked model on stem-cell dataset.

Most of the errors from a model's learning are from three main factors: variance, noise, and bias. By using stacking methods, the stability of the model can be significantly improved by reducing errors. By combining many models, stacking is able to group weak learners together to form one strong learner and, as shown in this stem-cell case, improve model performance significantly. By employing many powerful machine-learning techniques and constantly building and combining the databases, it is possible to further improve the predictive power of classification models on Raman spectra that have very complex information and often minute differences.

6.3.3 Feature Extraction

Multivariate analysis allows visualisation of the high-dimensional Raman data and classification algorithms to aid diagnostic practice in real-world approaches. However, in order to understand the underlying biochemistry of a biological system, feature extraction is essential to extrapolate the most important wavenumbers from a larger number of features and render them with biological meanings. The procedures for feature extraction in Raman analysis are summarised in Figure 6.17 and implemented by the code in Appendix B.3.3.

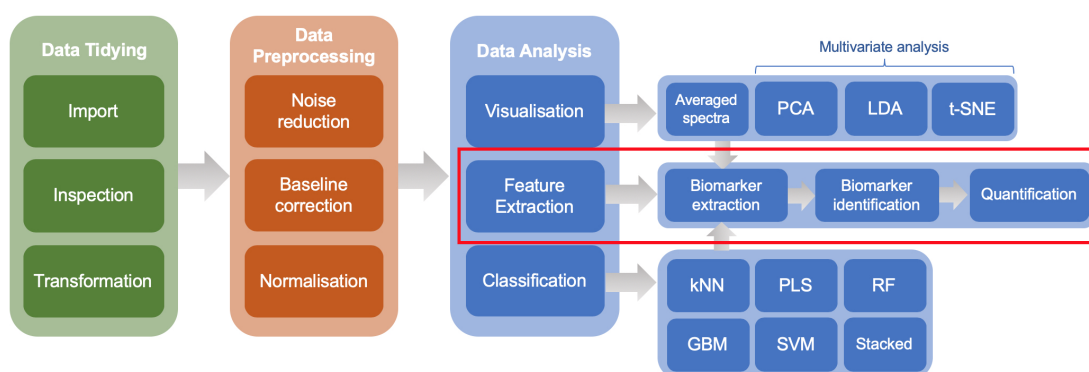


Figure 6.17: Feature extraction during the Raman analysis pipeline.

Biomarker extraction

The biomarker extraction stage is responsible for producing a smaller number of wavenumber-variables that are more informative than the original whole set of variables. Luckily, it can be simply done after the visualisation and classification, as described in the last two sections. The algorithms used in visualisation and classification, such as PCA, LDA, kNN and SVM, provide a means to identify the most important wavenumbers identified in the modelling process.

During the visualisation stage, both PCA and LDA reconstruct the original variables in a low-dimensional feature space where the new coordinates (PCs and LDs) are linear combinations of the original wavenumbers that are called "factors". The "factor loadings" are correlations between the original variables and the factors, and are the key to understanding the underlying nature of a particular factor. For both PCA and

LDA, a "loading plot" can be plotted to connect the original wavenumbers to their contributions to the PCs in PCA or LDs in LDA.

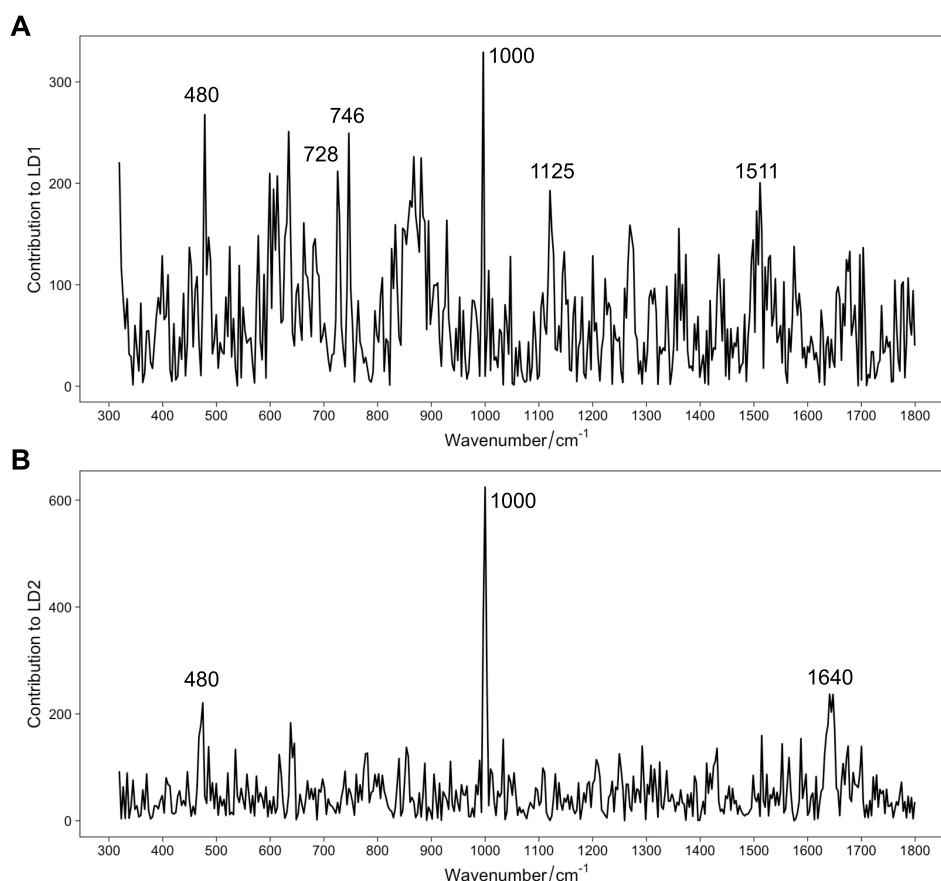


Figure 6.18: LDA loading plots showing the contribution of Raman wavenumber to (A) the first discriminant function LD1 and (B) the second discriminant function LD2. The wavenumbers with high contribution are labelled.

In some cases where PCA visualisation gives well separated clusters, PCA loading plots can be generated, such as shown in Figure 3.3 in Chapter 3 and Figure 5.2 in Chapter 5. In this stem-cell dataset, LDA produces better visualisation results compared to PCA (Figure 6.10), therefore LDA loading plots are obtained for the first and the second discriminant functions (LD1 and LD2). In Figure 6.18, the wavenumbers that have a relatively high contribution in LD1 and LD2 are labelled.

It should be noted that the t-SNE technique in the visualisation stage is not a feature-extraction technique as are PCA and LDA. Due to its non-linear transformation of variables, biological meanings cannot be extrapolated from the t-SNE results.

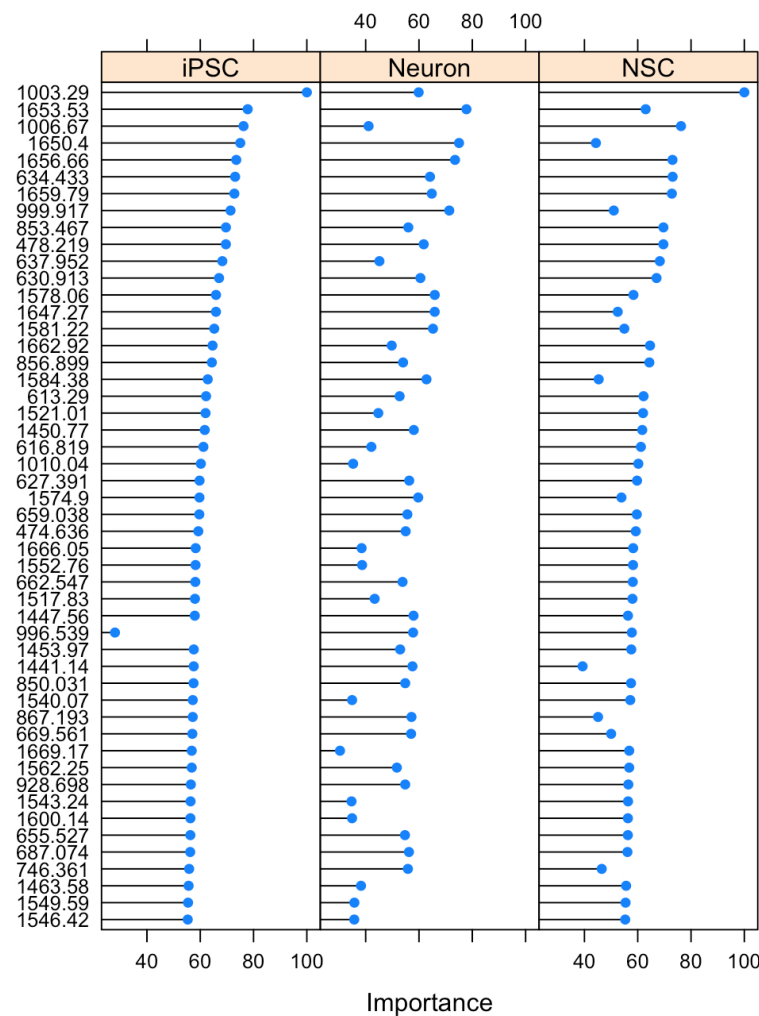


Figure 6.19: Variable importance extracted from the SVM-RBF model, showing top 50 wavenumbers responsible for classification of iPSC.

Biomarker extraction is also possible from the classification models in a similar fashion. Figure 6.19 is an example of displaying variable importance that is calculated by the SVM-RBF model, which achieves the best performance of classification among all models, especially when assigning the class of the iPSCs. The top 50 wavenumbers with the highest importance in classifying iPSCs are presented, along with their importance in classifying neurons and NSCs. As in the LDA loading plots, this allows researchers to identify the biomarkers in a specific group or the differences between different groups. Moreover, one can choose a different extraction route according to the performance of different models on a specific group.

Biomarker identification

After the extraction of specific wavenumbers, it is important to assign the identified wavenumbers to biomolecules and relate back to their biological mechanisms of action. This will require linking specific wavenumbers to their vibrational modes and then to the biological molecules that are responsible for the vibrations. During my DPhil, I have collected and summarised tables of some most important Raman assignments in biological studies, gathering from various reviews and research papers^{45,46,48–52,126,140,176–179,240,262,305}. The tables are shown below with wavenumbers, vibrational assignments, biological assignments and some useful comments. They are categorised into assignments of proteins (Table 6.3), lipids (Table 6.4), nucleic acids (Table 6.5), carbohydrates (Table 6.6) and resonant molecules (Table 6.7). Major Raman bands shifted in cells due to deuterium isotope substitution in Raman-DIP are also summarised (Table 6.8).

With biological connections, we are now able to identify the individual Raman bands as well as the biomarkers selected by multivariate analysis and machine learning models. The wavenumbers identified in Figure 6.18 can now be assigned to their biological identification (Figure 6.20).

Quantification

The number of molecules generating a band can be estimated by calculating the area of that band above the baseline using integration. After identifying the individual biomarkers, we are able to semi-quantify and compare the concentrations of specific biomolecules by calculating the integrated area of the corresponding Raman band, where absolute quantification may be achieved through a calibration curve using other analytical techniques. It should be noted that the pre-processing step contributes greatly to the quantification results, therefore a consistent and robust approach of applying baseline correction and normalisation should be identified.

Instead of looking at an averaged value from the calculation as in other bulk techniques, it is important to consider the whole single-cell population. A boxplot is a standard way of displaying the distribution of data. The rectangle in a box plot represents the second and third quartiles with a line inside representing the median. The lower and upper quartiles are drawn as lines outside the box. Boxplots are used

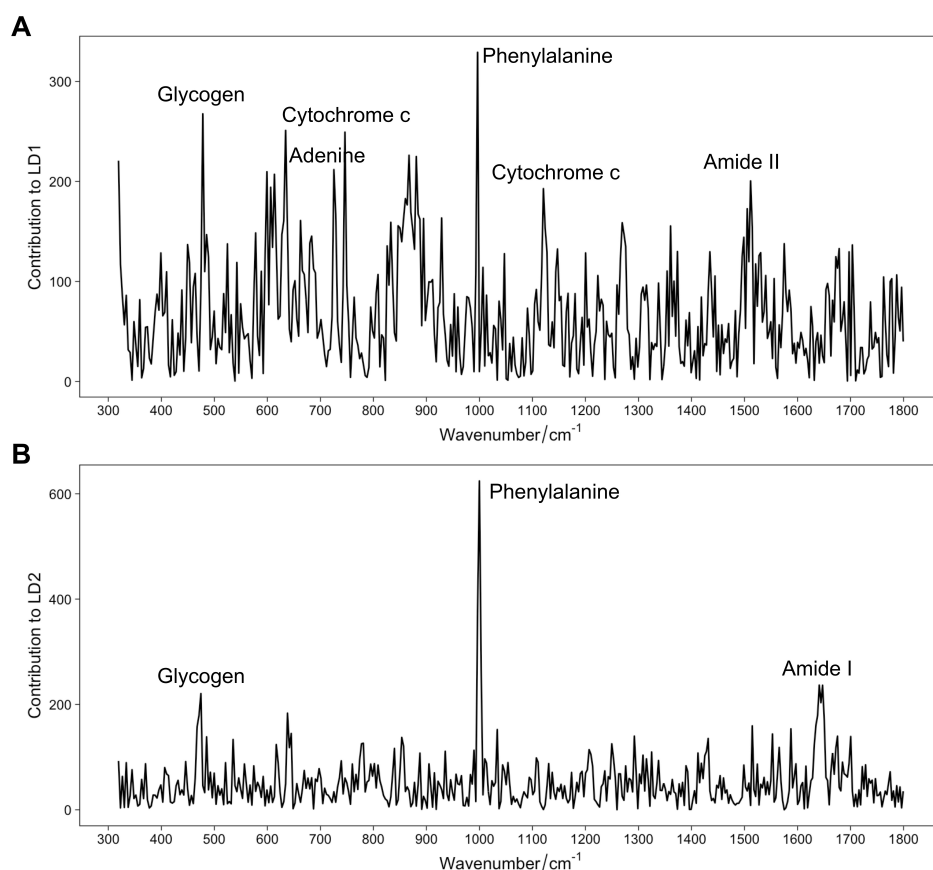


Figure 6.20: LDA loading plots showing the contribution of Raman wavenumber to (A) the first discriminant function LD1 and (B) the second discriminant function LD2. The wavenumbers with high contribution are identified.

to describe, quantify and compare one Raman band from all single-cell spectra in Chapter 3 (Figure 3.4B, Figure 3.4D and Figure 3.7) and Chapter 5 (Figure 5.1C, Figure 5.5B and Figure 5.7-5.10).

For the stem-cell dataset, similar quantification is done on three of the biomarkers identified in the previous section. The intracellular concentration of glycogen, phenylalanine and cytochrome can be semi-quantified by integrating the Raman bands at 480, 1003 and 1125 cm⁻¹, respectively. The quantification by Raman bands does not reveal the absolute concentration of the analytes without corroborating by calibration curves. Therefore, the results should be employed for comparison between groups and the comparison should always be validated by a statistical test. As in Figure 6.10A, the boxplot and a Welch's t-test reveal a significantly higher glycogen accumulation in cells of iPSC. Figure 6.10B shows significantly lower phenylalanine

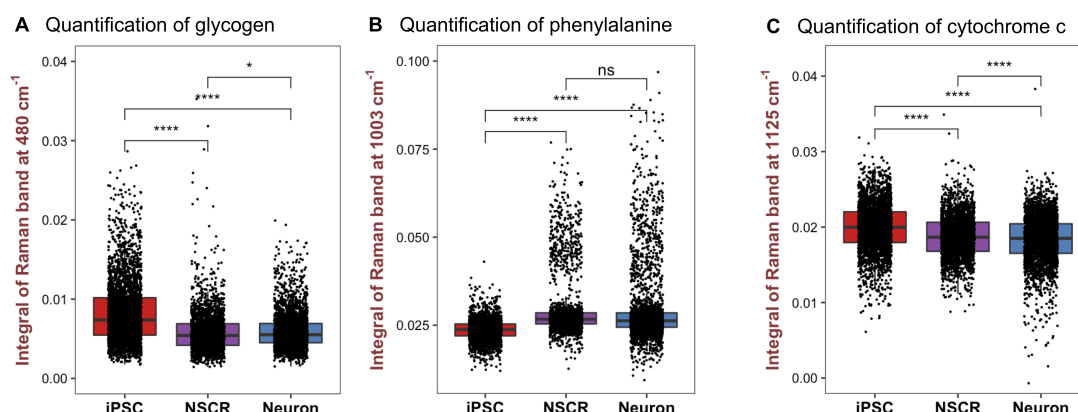


Figure 6.21: Quantification of (A) glycogen, (B) phenylalanine and (C) cytochrome c by integrating relevant Raman bands.

content in the initial stage of stem-cell differentiation. Quantification of cytochrome c in Figure 6.10C suggests a decreasing trend of the electron-transport molecule during the differentiation of stem cells. By quantifying the Raman biomarkers and comparing them between different groups, we are able to connect the spectroscopic studies with their underlying biological stems.

6.4 Conclusion

A pipeline of Raman analysis has been established and validated. Immediately after the spectral acquisition, data tidying and preprocessing can be performed, followed by a variety of analytical tools for either exploratory or diagnostic, qualitative or quantitative purposes. By following the established pipeline and executing the programming script, Raman analysis can be performed in a simple and rapid manner. In the future, more user-friendly software with the developed code can be used by bio-researchers to automatically perform a number of analytical and statistical tasks. This will help harmonise the Raman spectroscopic analysis and shift the work of researchers from complicated and labour-intensive data analysis to scientific exploration.

Table 6.3: Biological and vibrational assignments of Raman bands of Proteins.

Wavenumber/cm ⁻¹	Biological Assignment	Vibrational assignment	Comments
510	S-S bridge	S-S stretch	Proteins with sulphide bridges
620	Phenylalanine	C-C twist	
643	Tyrosine	C-C twist	
667	Cystine	C-S stretch	
855	Tyrosine (w)	Ring breath	
924-943	Proline, valine	Skeletal C-C stretch & protein backbone	α -helix
1000	Phenylalanine (s)	Phe symmetric ring breath	Aromatic AA
1030	Phenylalanine (w)	C-H in plane bend	
1230-1300	Amide III	40% C-H stretch 30% N-H bend	
1235-1240	Amide III		Exist in β -sheet Absence in α -helix
1340	Tryptophan	C-H deformation	Aromatic AA Fermi doublet
1360	Tryptophan	C-H deformation	
1480-1580	Amide II	60% N-H bend 40% C-N stretch	
1600-1690	Amide I	80% C=O stretch	
1660-1670	Amide I		β -sheet
2890	Aliphatic AA	C-H stretch	Proteins main C-H
2930	Aliphatic AA	C-H stretch	2850: lipids C-H 2930: proteins C-H 2970: DNA C-H
3060	Aromatic AA	C-H stretch	
3100	Amide B	N-H stretch	
3500	Amide A	N-H stretch	

Table 6.4: Biological and vibrational assignments of Raman bands of lipids.

Wavenumber/cm ⁻¹	Biological Assignment	Vibrational assignment	Comments
1060 1098 1120	Sat. FA	$\nu(\text{C-C})$	Saturated FA
1265	Unsat. FA	$\delta(\text{=C-H})$	Unsaturated FA; Intensity increases with double bonds
1295	FA	$\delta(\text{CH}_2)$	Sharp in sat. FA
1440	FA	$\delta(\text{CH}_2, \text{CH}_3)$	
1655	Unsat. FA	$\nu(\text{C=C})$	Sharp in UFA and fats; Intensity increases with double bonds
1735	Fats	$\nu(\text{C=O})$	Ester group e.g. cholesteryl esters e.g. polymer like PHB e.g. phospholipids
2850	Lipids	$\nu(\text{C-H})$	2850: lipids C-H 2930: protein C-H 2970: DNA C-H
3010	Unsat. Lipids	$\nu(\text{=C-H})$	

Table 6.5: Biological and vibrational assignments of Raman bands of nucleic acids.

Wavenumber/cm ⁻¹	Biological Assignment	Vibrational assignment	Comments
617	Thymine	Ring breath	Pyrimidine bases
650	Guanine	Ring breath	Purine bases
670	DNA	Thymine Guanine	
723	Adenine	Ring breath	Purine bases
740	Thymine	Ring breath	
785*	DNA/RNA	Ring breath of C, T and U	
790	Uracil	Ring breath	Pyrimidine bases
790	Cytosine	Ring breath	Pyrimidine bases
828	DNA/RNA	O-P-O stretch	Nucleic acid backbone
1094*	DNA/RNA	Symmetric stretch of PO ₂ ⁻	Nucleic acid backbone
1235	Uracil		
1275	DNA	Cytosine Guanine Adenine	
1376	DNA	Thymine Guanine Adenine	
1490	DNA	Guanine Adenine	
1578	DNA	Guanine Adenine	
1671	DNA	(C=O); Thymine Guanine Cytosine	

Table 6.6: Biological and vibrational assignments of Raman bands of carbohydrates.

Wavenumber/cm ⁻¹	Biological Assignment	Vibrational assignment	Comments
484*	Glycogen; polysaccharide	$\nu(\text{C1-O-C4})$	glycosidic bond; straight chain
856	Glycogen; polysaccharide	$\nu(\text{C1-O-C4})$; ring breath	glycosidic bond; straight chain
940	Glycogen; polysaccharide	$\nu(\text{C1-O-C6})$	glycosidic bond; bridge chain
1050	Glycogen; polysaccharide	C-O stretch	
1084	Glycogen; polysaccharide	C-O stretch	
1130	Glycogen; polysaccharide	C-O ring breath	
407	Glucose; oligosaccharide	Skeletal mode	
424	Glucose; oligosaccharide	Skeletal mode	
518	Glucose; oligosaccharide	Skeletal mode	
542	Glucose; oligosaccharide	Ring deformation	
843	Glucose; oligosaccharide	$\nu(\text{C-O-C})$	
914	Glucose; oligosaccharide	Ring vibration	

Table 6.7: Biological and vibrational assignments of Raman bands of special molecules with resonance effect under a 532-nm laser.

Wavenumber/cm ⁻¹	Biological Assignment	Vibrational assignment	Comments
750	Cytochrome C	$\nu(\text{pyr breathing});$ $\nu(\text{C}_\alpha\text{C}_\beta)$	Heme protein
1128	Cytochrome C	$\nu(\text{C}_\alpha\text{N})$	
1312	Cytochrome C	$\nu(\text{C}_\alpha\text{C}_\beta)$	
1585	Cytochrome C	$\nu(\text{C}_\alpha\text{C}_m)_{\text{asym}}$	
1003	β carotene; carotenoid	Ring breathing	Orange pigments
1156	β carotene; carotenoid	$\nu(\text{C-C})$	
1515	β carotene, carotenoid	$\nu(\text{C=C})$	

Table 6.8: Biological and vibrational assignments of Raman band shifts related to Raman-DIP.

Wavenumber of unlabelled molecule/cm ⁻¹	Wavenumber of labelled molecule/cm ⁻¹	Biological and Vibrational Assignment
1003	989	Phe ring vibration; one of sites 1, 3, 5 labelled
	976	Phe ring vibration; two of sites 1, 3, 5 labelled
	962	Phe ring vibration; all of sites 1, 3, 5 labelled
2850	2140	C-H stretch for lipids
2930	2180	C-H stretch for proteins
2800-3030	2210	C-H stretch for DNA
2070-2300	2070-2300	C-H stretch for all molecules

Chapter 7

Summary and Outlook

The main aspect of this work is to apply SCRM and DIP to various biological scenarios of bacterial metabolism and physiology. Some of the previously uncharted areas have been explored and, during the process, a robust pipeline of single-cell Raman analysis has been developed and validated. The power of single-cell analysis in a heterogeneous bacterial population in real-world scenarios has been specifically addressed.

In the first application, SCRM is applied to study bacteria in a symbiotic setting with a plant host. SCRS of the symbiotic diazotroph *R. leguminosarum* bv. *viciae* successfully differentiate different physiological states of bacteria despite being genetically identical. This study also demonstrates that, for the first time, Raman microscopic measurements are able to discriminate genetic mutants, including both single and double mutants of carbon storage polymers. Furthermore, by analysing SCRS of a nitrogenase mutant, it offers single-cell resolution which allows heterogeneity analysis of dual phenotypes within an isogenic bacterial population. By creating metabolic profiles of single mutant cells and quantifying intracellular levels of macromolecules, SCRM is applicable to the study of complex carbon regulatory networks in the rhizosphere.

The ability of SCRM to perform a semi-quantitative analysis of several macromolecules within a single experiment is revealed in this study. It reveals a comprehensive intrinsic chemical profile in single cells without prior knowledge. This can provide novel insights into production of different metabolites and has the potential for making new scientific discoveries of cell phenotypes. Compared with traditional analytical techniques like chromatography, mass spectrometry and quantitative spectroscopy, SCRM is still new in biological sciences and requires validation and standardisation through the efforts of all Raman scientists. For example, more accurate quantification of biomolecules through SCRS can be achieved if used in conjunction with other

analytical techniques such as gas chromatography–mass spectrometry (GC–MS) which provides a standardised reference.

The second application implements the technique of SCRM-DIP in the study of bacterial carbon metabolism. It provides a culture-independent and non-destructive approach to probe metabolic pathways of carbon substrates at the single-cell level. By incubating microorganisms with deuterated carbon sources, a clearly distinguishable C–D vibration band at 2070–2300 cm^{-1} can be found, the intensity of which is proportional to the extent of deuteration in the carbon source. A lower lipid conversion rate from deuterated naphthalene compared with that from deuterated glucose is revealed owing to different metabolic pathways to synthesise fatty acids. Three isotopic shifts of the phenylalanine band at 1004 cm^{-1} correlate well with a computational model. A reduction in phenylalanine deuteration in *Pseudomonas* compared with that in *E. coli* is found due to the dilution effect during phenylalanine biosynthesis. Collectively, it shows that SCRM-DIP not only indicates general metabolic activity of bacteria but can also reveal different metabolic pathways of specific carbon sources adopted by different microorganisms.

In the third application, SCRM-DIP is further employed from pure bacterial cultures to another symbiotic setting with a human host and, for the first time, to study metabolic adaptations of a pathogen inside the host cells. Three *Salmonella* strains responsible for different clinical outcomes are analysed using SCRM-DIP. The fingerprint region of SCRS shows that the three strains are metabolically different. STM-D23580 has metabolic traits lying in between the non-typhoidal STM-LT2 and the virulent ST-Ty2, and it shares features in common with the typhoidal strain relative to the ability to achieve successful extraintestinal infection. A kNN machine learning classification achieves an accuracy of 92.4% correctly identifying these closely related strains based on their SCRS. SCRM-DIP and reverse labelling reveal the slowest metabolism rate in the most invasive ST-Ty2, as opposed to its fastest extracellular growth. With the non-invasive STM-LT2 being the most active metaboliser, this study highlights a possible strategy adopted by the invasive salmonellae to slow down central metabolism, hide from host surveillance and allow for further dissemination. A quantitative proteomic

analysis of proteins related to bioenergetic functions confirms the SCRM-DIP results, with most of the energy pathways being upregulated in STM-LT2 and downregulated in ST-Ty2. A dynamic Raman profile of *Salmonella* inside host cells reveals the metabolic adaptations of biocomponents during their intracellular life as well as the differences among three strains. The results suggest that the invasive *Salmonella* might switch off the glycolytic pathways and use fatty acids as carbon sources instead. By comparing and mapping the activities and metabolic profiles of three *Salmonella* strains in a dynamic and time-resolved manner, the mechanisms adopted by intracellular pathogens to regulate their metabolic pathways in response to the host cell are revealed.

Our understanding of intracellular bacteria metabolism has been largely built upon genetic studies, in which individual genes or a small set of genes are chosen to be inactivated and studied. These studies are often insufficient to reveal a comprehensive picture of bacterial metabolism network. Recent advances in omics technologies enable a more systemic view of *Salmonella* metabolism. However, these techniques have disadvantages such as insufficient resolution, low turnover rates, incomplete coverage and more importantly, loss of heterogeneity due to the averaged read-outs. During the intracellular life of *Salmonella*, while most of the bacteria remain initially in host phagosomes, some salmonellae escape to the cell cytosol and feature vigorously different proliferation and metabolism. Consequently, bulk measurements can generate misleading conclusions originated from the diverse bacterial subpopulations. SCRM and SCRM-DIP are capable of highlighting such molecular heterogeneity and cell-to-cell variation among *Salmonella* subsets. Single-cell Raman approaches thereby offer new perspectives on understanding metabolic adaptations and single-cell heterogeneity of host-infected *Salmonella*, and can be potentially applied to a wide range of intracellular conditions with heterogeneous behaviours.

In this thesis, SCRM-DIP as a tool to probe cell activity and cellular metabolism is introduced by exposing cells to heavy water or deuterium-labelled carbon substrates. It shines a light on various possible routes for SCRM-DIP to investigate cell activity and cellular metabolism. Providing a high variety of possible isotopic substrates, SCRM-DIP has the potential to probe the metabolism of most of the important biomolecules,

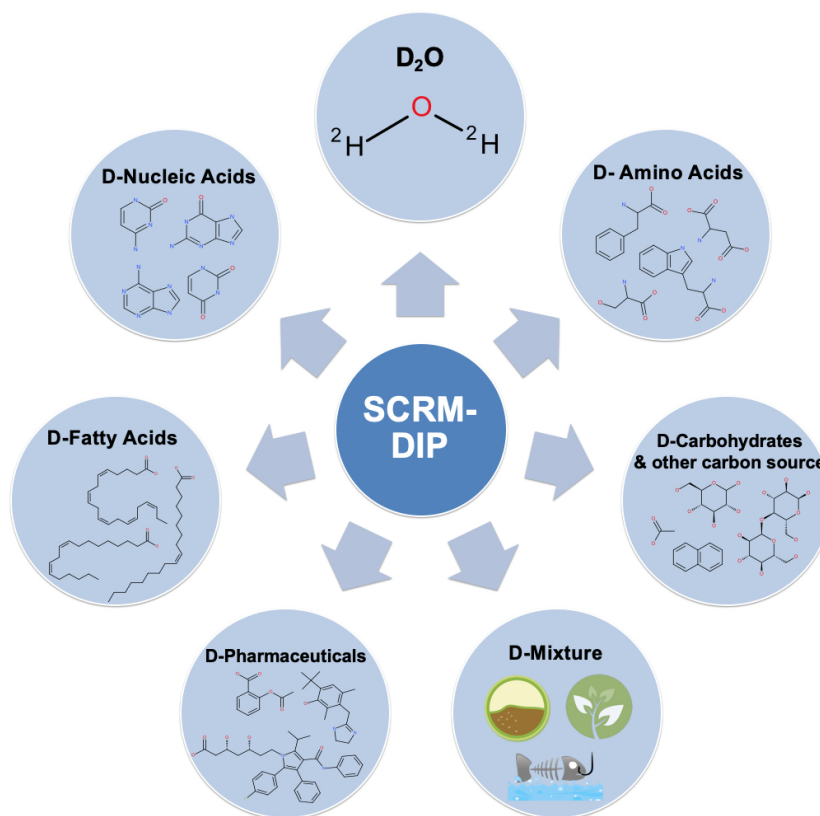


Figure 7.1: Potential of SCRM-DIP to study cell activity and metabolism.

including proteins, lipids, carbohydrates and nucleic acids (Figure 7.1). For example, D-labelled nucleotides can detect DNA synthesis in proliferating cells, D-amino acids can selectively and quantitatively trace synthesis of proteins from specific precursors, D-fatty acids allows tracking of lipid uptake and storage. Furthermore, SCRM-DIP using more complex substances including natural products, medicines and mixtures containing multiple chemicals can be invaluable in the study of biomedicine and environmental sciences (Figure 7.1). By exposing the cells to deuterium-labelled substrates designed to mimic as closely as possible their naturally more abundant counterpart, it provides an explanation of cellular metabolism without altering the natural substrate pool. SCRM-DIP studies using these D-labelled substrates are therefore invaluable and requires further exploration in the functional study of cells in their natural conditions at the single-cell level.

Finally, to wrap up what has been applied above, a pipeline of Raman data analysis for biological subjects is rationalised, together with a programming script for

automatically executing the pipeline. The pipeline starts from raw-data tidying to data preprocessing and continues on to various techniques of data analysis to fulfil both descriptive and predictive purposes in differing biological research. Data visualisation techniques, namely, PCA, LDA and t-SNE are employed and compared. Various approaches to machine learning classification based on samples' SCRS are also discussed. Performances of kNN, LDC, PLS, RF, GBM, XGB and two SVMs are compared. The power of stacking different models is emphasised and enabled in this protocol. Biomarker identification and quantification are made possible by introducing feature extraction techniques, summarising Raman assignment tables and providing quantification methods. Following the established pipeline, a rapid, systematic and automatic Raman spectroscopic analysis in biological sciences can be performed by non-specialists.

Overall, this thesis demonstrates SCRM as a label-free and non-destructive technique to study bacterial phenotypes by producing biochemical profiles of single cells, consisting of a collection of biomolecules that can identify and characterise single cells. To where genomics cannot reach, the combination of SCRM and DIP can answer a plethora of questions in terms of cellular physiology and metabolic adaptations.

The work outlined here is just a first step towards developing SCRM and SCRM-DIP into a universal tool capable of answering a wide range of questions which remain open in the vast world of cell biology, microbiology and beyond. The development of computational power and machine learning techniques can also push the advance of the field by employing algorithms or models to understand the biological origins buried underneath the overlapping and complex Raman spectra.

Appendices

Appendix A

R Code for Raman Analysis Pipeline

Contents

A.1 Data tidying	168
A.2 Data Pre-processing	169
A.2.1 Baseline correction	169
A.2.2 Normalisation	169
A.3 Data Analysis	170
A.3.1 Visualisation	170
A.3.2 Classification	172
A.3.3 Feature extraction	174

A.1 Data tidying

```
1 # Set working directly to the folder containing spectra in .txt
2 setwd("/Users/Jiabao/Desktop/IBME collaboration/MapTxt_FilteredDespiked")
3 files <- list.files(path = "/Users/Jiabao/Desktop/IBME collaboration/MapTxt_
   FilteredDespiked")
4
5 # Create a dataframe containing all spectra
6 combinedata <- read.table(files[1], header = FALSE, sep = "\t")[1,]
7 combinedata$Filename <- "Filename"
8 for (filename in files)
9 {
10   data <- read.table(filename, header = FALSE, sep="\t")[-1,]
11   data$Filename <- filename
12   combinedata <- rbind(combinedata, data)
13 }
14 row.names(combinedata)[1] <- "Wavenumber"
15 combinedata <- combinedata[,-c(1:2)]
16
17 # Remove observations rows with 0s
18 combinedata[combinedata==0] <- NA
19 combinedata <- combinedata[complete.cases(combinedata),]
20
21 # Create a numeric matrix
22 spec <- as.matrix(combinedata[-1,-1020]) # Numeric matrix for intensities
23 wn <- as.numeric(combinedata[1,-1020]) # Numeric vector for wavenumber
24 t.spec <- t(spec) # Transpose
25
26 # Select FP region
27 spec <- spec[,c(1:443)]
28 wn <- wn[c(1:443)]
29 t.spec <- t(spec)
30
31 # Plot and inspect spectra
```

```

32 matplot(wn, t.spec, type="l") # See all
33 plot(wn, spec[1,], type="l") # See any one
34 which.max(spec[,400]) # Useful to find the highest fluorescence
35 which.min(spec[,400]) # Useful to find the highest noise

```

A.2 Data Pre-processing

A.2.1 Baseline correction

```

1 library(baseline)
2 sapply(baselineAlgorithms, description) # See all correction methods
3 # Choose one method, recommend rollingBall, als, modpolyfit
4 spec.bl <- baseline(spec, method="modpolyfit", degree = 4, tol = 0.0001, rep =
  100)
5 spec.bl <- baseline(spec, method='als', lambda=7, p = 0.05)
6 spec.bl <- baseline(spec, method='irls', lambda1 = 3, lambda2 = 10)
7 spec.bl <- baseline(spec, method='rfbaseline', span = 2/3)
8 spec.bl <- baseline(spec, method='rollingBall', wm=200, ws=100)
9
10 # Plot any one original vs corrected spectrum
11 plot(spec.bl, specNo = 3554)
12
13 # Output
14 corr <- getCorrected(spec.bl)
15 t.corr <- t(corr)
16
17 # Plot
18 matplot(wn, t.spec, type="l") # original spectra
19 matplot(wn, t.corr, type="l") # baseline-corrected spectra

```

A.2.2 Normalisation

```

1 # normalize a row by the sum of its entries
2 for (n in 1:nrow(corr)) {
3   S <- sum(corr[n,]) # Add range here if specific wn desired
4   corr[n,] <- corr[n,]/S
5 }
6
7 t.corr <- t(corr)
8
9 # Plot normalised spectra
10 matplot(wn, t.corr, type="l")
11
12 # Transform matrix back to data frame and change names of the group
13 data <- data.frame(corr, row.names = NULL)
14 colnames(data) <- wn
15
16 # Add file name
17 library(stringr)
18 str(data)
19 data$File <- combineddata[-1,-(444:1019)]$Filename
20 data$Group <- word(data$File,1,sep = fixed("_"))
21 data$CellLine <- word(data$File,2,sep = fixed("_"))
22 data$Replicate <- word(data$File,3,sep = fixed("_"))
23 data$Time <- word(data$File,4,sep = fixed("_"))
24 data$Cell <- word(data$File,5,sep = fixed("_"))

```

A.3 Data Analysis

A.3.1 Visualisation

To plot averaged spectra from single-cell spectra:

```

1 ##### Plot spectra
2 library(reshape2)
3 library(plyr)
4 # Wide to long
5 data.long <- melt(data, variable.name = "Wavenumber",
6                   idvar = c("File","Group","CellLine", "Replicate", "Time", "Cell"
7                             ))
8
9 # Summary stats
10 stat.1 <- ddply(data.long, c("Group", "Wavenumber"), summarise,
11                 average = mean(value), stdev = sd(value),
12                 error = sd(value)/sqrt(length(value)))
13 # Factor group information
14 stat.1$Group <- factor(stat.1$Group, levels = c("iPSC","NSCR","Neuron"))
15 # Shift spectral visualisation
16 stat.1$modified_average = stat.1$average - 0.005 * as.numeric(as.factor(stat.1$
17   Group))
18
19 # Plot averages and ribbons
20 library(ggplot2)
21 library(RColorBrewer)
22 # Create colours desired for plotting
23 mypalette = brewer.pal(8, "Set1")[c(1,4,2)]
24
25 # Plot according to different cell groups
26 ggplot(stat.1, aes(x = Wavenumber, y = modified_average, group = Group)) +
27   geom_line(aes(color = Group),size=0.6) +
28   geom_ribbon(aes(ymin = modified_average - stdev,
29                 ymax = modified_average + stdev,
30                 fill = Group),
31             alpha = 0.5) +
32   labs(title="",x=expression(Wavenumber/cm^{-1}), y = "Intensity (a.u.)") +
33   scale_fill_manual(values = mypalette) +
34   scale_color_manual(values = mypalette) +
35   scale_x_continuous(breaks=seq(300,1800,300)) +
36   theme_linedraw() +
37   theme(panel.grid.major = element_blank(),
38         panel.grid.minor = element_blank(),
39         legend.position = "right",
40         legend.background = element_blank(),
41         legend.title = element_blank(),
42         axis.text.y = element_blank(),
43         axis.ticks.y = element_blank()) +
44   ggtitle("")

```

To perform PCA and visualise the results:

```

1 library(FactoMineR)
2 library(factoextra)
3 set.seed(329)
4 # Perform PCA
5 pca <- PCA(data[, -(444:449)], scale.unit = FALSE, ncp = 10,
6            graph = FALSE)
7
8 PCA <- fviz_pca_ind(pca, axes = c(2,1),
9                   #select.ind = list(cos2 = 0.2),
10                  label="none",
11                  habillage = as.factor(data$Group),
12                  repel = TRUE,

```

```

13         palette = mypalette,
14         addEllipses=TRUE, ellipse.level=0.9) +
15 theme_linedraw() +
16 # coord_cartesian(xlim=c(-0.02,0.025),ylim = c(-0.02,0.02))+
17 theme(panel.grid.major = element_blank(),
18       panel.grid.minor = element_blank(),
19       legend.position = "right",
20       legend.title = element_blank()) +
21 ggtitle("")
22 # Plot PCA
23 plot(PCA)

```

To perform LDA and visualise the results:

```

1 ##### LDA
2 library(MASS)
3 library(caret)
4 # Remove class label and other labels
5 dataLDA <- data[,-c(444,446:449)]
6 dataLDA$Group <- factor(dataLDA$Group, levels = c("iPSC","NSC","Neuron"))
7 # Build LDA model
8 lda <- lda(Group~., data = dataLDA)
9
10 # Proportion of trace explains the percentage of variances explained by each
    discriminant function
11 lda # LD1 - 77.6%, LD2 - 22.4%
12
13 # Use LDA to predict class in probability of assigning correctly
14 predictions <- predict(lda)
15 predictions
16 round(predictions$posterior, 2)
17
18 # A Stacked Histogram of the LDA Values
19 # Make a stacked histogram of the values of the discriminant function for the
    samples from different groups
20 ldahist(data = predictions$x[,1], g=dataLDA$Group) # 1st DF
21 ldahist(data = predictions$x[,2], g=dataLDA$Group) # 2nd DF
22
23 # Plot and visualise - Scatterplots of the Discriminant Functions
24 lda.data <- cbind(data, predict(lda)$x)
25 lda.data$Group <- factor(lda.data$Group, levels = c("iPSC","NSC","Neuron"))
26 ggplot(lda.data, aes(LD1, LD2)) +
27   geom_point(aes(color = Group, shape = Group),size=0.5,alpha=0.7)+
28   scale_fill_manual(values = mypalette) +
29   scale_color_manual(values = mypalette) +
30   # coord_cartesian(ylim = c(0, 5)) +
31   theme_linedraw() +
32   labs(x = "LD1 (77.6%)", y = "LD2 (22.4%)" )+
33   theme(panel.grid.major = element_blank(),
34         panel.grid.minor = element_blank(),
35         legend.position = "right",
36         legend.background = element_blank(),
37         legend.title = element_blank()
38   ) + ggtitle("")

```

To perform t-SNE and visualise the results:

```

1 # Perform 2-D t-SNE
2 tsne <- Rtsne(data[,-c(444:449)], perplexity = 100,
3             pca=T, pca_scale = T)
4
5 # Plot 2-D t-SNE
6 d_tsne <- data.frame(tsne$Y)
7 d_tsne$Group <- as.character(data$Group)
8 d_tsne$Group <- factor(d_tsne$Group, levels = c("iPSC","NSC","Neuron"))
9 ggplot(d_tsne, aes(X1, X2, color = d_tsne$Group)) +
10   geom_point(size=0.3,alpha=0.5)+

```



```

11 labs(x="t-SNE Axis 1", y="t-SNE Axis 2") +
12 scale_fill_manual(values = mypalette) +
13 scale_color_manual(values = mypalette) +
14 theme_linedraw() +
15 theme(panel.grid.major = element_blank(),
16       panel.grid.minor = element_blank(),
17       legend.position = "right",
18       # legend.background = element_blank(),
19       legend.title = element_blank()
20 ) + ggtitle("")

```

A.3.2 Classification

To build machine learning models by individual classifiers:

```

1 library(caret)
2 # Grep the indexes of labels
3 grep("File", colnames(data)) #444
4 grep("Group", colnames(data)) #445
5 grep("CellLine", colnames(data)) #446
6 grep("Replicate", colnames(data)) #447
7 grep("Time", colnames(data)) #448
8 grep("Cell", colnames(data)) #449
9
10 data$Group <- as.factor(data$Group)
11 data$CellLine <- as.factor(data$CellLine)
12 data$Replicate <- as.factor(data$Replicate)
13 data$Time <- as.factor(data$Time)
14 data$Cell <- as.factor(data$Cell)
15
16 # Create balanced splits of the data into training set and test set
17 set.seed(329)
18 trainIndex <- createDataPartition(data$Group, p = .8,
19                                  list = FALSE,
20                                  times = 1)
21 head(trainIndex)
22 dataTrain <- data[ trainIndex,]
23 dataTest <- data[-trainIndex,]
24
25 # Parameter tuning
26 fitControl <- trainControl(## 10-fold CV
27   method = "repeatedcv",
28   number = 10,
29   ## repeated 5 times
30   repeats = 5,
31   allowParallel=TRUE)
32
33 ##### Diff models
34
35 # Note: parallel computing can be done using the code below
36 library(doParallel) # for parallel computing
37 library(tictoc) # for time counting
38 tic()
39 cl <- makeCluster(detectCores())
40 registerDoParallel(cl)
41 YourModelHere
42 stopCluster(cl)
43 registerDoSEQ()
44 toc()
45
46 # kNN - k best at 90
47 knnGrid <- expand.grid(k = c(5,10,20,30,40,50,60,70,80,90,100))
48 knnFit <- train(Group ~ ., data = dataTrain[, -c(444,446:449)],
49               method = "knn",
50               trControl = fitControl,
51               tuneGrid = knnGrid)

```

```

52
53 # LDA - no need for grid expansion, dim = 2
54 ldaFit <- train(Group ~ ., data = dataTrain[, -c(444,446:449)],
55               method = "lda2",
56               trControl = fitControl)
57
58 # PLS
59 plsGrid <- expand.grid(ncomp = c(1,3,5,7,9,11,13,15,17,19))
60 plsFit <- train(Group ~ ., data = dataTrain[, -c(444,446:449)],
61               method = "pls",
62               trControl = fitControl,
63               tuneGrid = plsGrid)
64
65 # RF
66 rffit <- train(Group ~ ., data = dataTrain[, -c(444,446:449)],
67               method = "ranger",
68               trControl = fitControl)
69
70 # GBM
71 gbmGrid <- expand.grid(interaction.depth = c(1,5,9,11,15),
72                       n.trees = c
73                           (500,1000,1500,2000,2500,3000,3500,4000,4500,5000,5500)
74                           ,
75                       shrinkage = 0.1,
76                       n.minobsinnode = 10)
77 gbmFit <- train(Group ~ ., data = dataTrain[, -c(444,446:449)],
78               method = "gbm",
79               trControl = fitControl,
80               tuneGrid = gbmGrid,
81               verbose = FALSE)
82
83 # XGB
84 xgbFit <- train(Group ~ ., data = dataTrain[, -c(444,446:449)],
85               method = "xgbDART",
86               trControl = fitControl)
87
88 # SVM-Linear
89 svmFitLin <- train(Group ~ ., data = dataTrain[, -c(444,446:449)],
90               method = "svmLinear2", # or svmLinear
91               trControl = fitControl)
92
93 # SVM-RBF
94 svmFitRBF <- train(Group ~ ., data = dataTrain[, -c(444,446:449)],
95               method = "svmRadialSigma", # or svmRadialWeights
96               trControl = fitControl)

```

To test model performance with the test set:

```

1 # Predict
2 knnPred <- predict(knnFit, dataTest[, -c(444:449)])
3 ldaPred <- predict(ldaFit, dataTest[, -c(444:449)])
4 plsPred <- predict(plsFit, dataTest[, -c(444:449)])
5 rffPred <- predict(rffit, dataTest[, -(444:449)])
6 gbmPred <- predict(gbmFit, dataTest[, -(444:449)])
7 xgbPred <- predict(xgbFit, dataTest[, -(444:449)])
8 svmPred <- predict(svmFitLin, dataTest[, -(444:449)])
9 svmPred2 <- predict(svmFitRBF, dataTest[, -(444:449)])
10
11
12 confusionMatrix(data = knnPred, reference = dataTest$Group)
13 confusionMatrix(data = ldaPred, reference = dataTest$Group)
14 confusionMatrix(data = plsPred, reference = dataTest$Group)
15 confusionMatrix(data = rffPred, reference = dataTest$Group)
16 confusionMatrix(data = gbmPred, reference = dataTest$Group)
17 confusionMatrix(data = xgbPred, reference = dataTest$Group)
18 confusionMatrix(data = svmPred, reference = dataTest$Group)
19 confusionMatrix(data = svmPred2, reference = dataTest$Group)

```

To build a stacked model and test its performance:

```

1 # Compare models by resampling
2 resamps <- resamples(list(KNN = knnFit,
3                           LDC = ldaFit,
4                           PLS = plsFit,
5                           RF = rfFit,
6                           GBM = gbmFit,
7                           XGB = xgbFit,
8                           SVMlin = svmFitLin,
9                           SVMRBF = svmFitRBF))
10 resamps
11 summary(resamps)
12 modelCor(resamps) # Models being stacked should have low correlations (<0.75)
13 splom(resamps)
14
15 # Stack the predictions from multiple models
16 stackPred <- data.frame(knnPred, ldaPred, plsPred,
17                          gbmPred, rfPred, xgbPred,
18                          svmPred, svmPred2, svmPred2,
19                          class = dataTest$Group)
20
21 # Apply GBM to the stacked predictions
22 tic()
23 cl <- makeCluster(detectCores())
24 registerDoParallel(cl)
25 stackgbmFit <- caret::train(class ~ ., data = stackPred,
26                             method = "gbm",
27                             trControl = fitControl)
28 stopCluster(cl)
29 toc()
30
31 # Test the performance of the stacked model
32 stackgbmPred <- predict(stackgbmFit, stackPred[, -grep("class", colnames(stackPred))
33   ])
34 confusionMatrix(data = stackgbmPred, reference = stackPred[, grep("class", colnames(
35   stackPred))])

```

A.3.3 Feature extraction

To extract features from LDA:

```

1 # View loadings/slopes/coefficients/weights of each variable on each discriminant
  function
2 round(lda$scaling, 2)
3 loadings = data.frame(lda$scaling)
4 loadings$AbsVal_LD1 = (loadings$LD1^2)^0.5
5 loadings$AbsVal_LD2 = (loadings$LD2^2)^0.5
6 loadings$wn <- wn
7
8 # Plot LDA loadings
9 ggplot(loadings, aes(x=wn, y=AbsVal_LD2)) +
10   geom_line() +
11   labs(title="", x=expression(Wavenumber/cm^{-1}), y = "Contribution to LD2") +
12   scale_fill_manual(values = mypalette) +
13   scale_color_manual(values = mypalette) +
14   scale_x_continuous(breaks=seq(200,1800,100)) +
15   # coord_cartesian(ylim = c(0, 5)) +
16   theme_linedraw() +
17   theme(panel.grid.major = element_blank(),
18         panel.grid.minor = element_blank(),
19         legend.position = "right",
20         legend.background = element_blank(),
21         legend.title = element_blank())
22 ) + ggtitle("")

```

To extract features from SVM:

```
1 # Extract variable importance from the model
2 svm_varimp <- data.frame(varImp(svmFitLin, scale = FALSE)$importance)
3
4 # Plot variable importance
5 plot(varImp(svmFitRBF), top = 50)
```

To quantify the biomolecules by calculating the Raman band area:

```
1 library(plyr)
2 library(easyGgplot2)
3 library(ggpubr)
4 my_comparisons <- list( c("iPSC", "NSC"), c("iPSC", "Neuron"), c("NSC", "Neuron")
5 )
6
7 # Integration of cytochrome c at 1127
8 data$'1114-1134' <- rowSums(data[,232:238])
9
10 # Plot quantification results with stats
11 ggplot2::boxplot(data,
12   xName = "Group",
13   yName = data$'1114-1134',
14   groupName = "Group",
15   backgroundColor = "white",
16   ytitle = expression(bold(paste(Integral~of~Raman~band~at~"1125"~cm
17     ^"-1"))),
18   xtitle = "",
19   addDot = T, dotSize = 0.5, dotPosition = "jitter",
20   removePanelGrid = TRUE, PanelBorder = c(1, "solid", "black"),
21   axisLine=c(0, "solid", "black"),
22   legendPosition = "none",
23   # xShowTickLabel = FALSE,
24   outlier.shape=NA
25 )+
26 coord_cartesian(ylim=c(0.00,0.042))+
27 scale_fill_manual(values = mypalette) +
28 scale_color_manual(values = mypalette) +
29 stat_compare_means(label = "p.signif", method = "t.test",
30   comparisons = my_comparisons,
31   label.y = c(0.033,0.036,0.04))+
32 theme(axis.text.x = element_text(size=11, face="bold"),
33   axis.text.y=element_text(size=9, face="plain"),
34   axis.title.y = element_text(size=12, face="bold", colour="#993333"))
```

References

1. Bumbrah, G. S. & Sharma, R. M. Raman spectroscopy – Basic principle, instrumentation and selected applications for the characterization of drugs of abuse. *Egyptian Journal of Forensic Sciences* **6**, 209–215. ISSN: 2090-536X (Sept. 1, 2016).
2. Raman, C. V. & Krishnan, K. S. A New Type of Secondary Radiation. *Nature* **121**, 501. ISSN: 1476-4687 (Mar. 1928).
3. Toporski, J., Thomas Dieing & Hollricher, O. *Confocal Raman Microscopy* 596 pp. ISBN: 978-3-319-75378-2 (Springer International Publishing, 2018).
4. Sharma, D. B. K. *Instrumental Methods of Chemical Analysis* Google-Books-ID: igR_jsqfcowC. 1726 pp. ISBN: 978-81-8283-019-6 (Krishna Prakashan Media, 1981).
5. Kostamovaara, J. *et al.* Fluorescence suppression in Raman spectroscopy using a time-gated CMOS SPAD. *Optics Express* **21**, 31632–31645. ISSN: 1094-4087 (Dec. 16, 2013).
6. Wei, D., Chen, S. & Liu, Q. Review of Fluorescence Suppression Techniques in Raman Spectroscopy. *Applied Spectroscopy Reviews* **50**, 387–406. ISSN: 0570-4928 (May 28, 2015).
7. Lewandowska, R. in *Encyclopedia of Materials: Science and Technology* (eds Buschow, K. H. J. *et al.*) 1–6 (Elsevier, Oxford, Jan. 1, 2010). ISBN: 978-0-08-043152-9. doi:10.1016/B978-008043152-9.02247-8. <http://www.sciencedirect.com/science/article/pii/B9780080431529022478> (2019).
8. Sonntag, M. D., Pozzi, E. A., Jiang, N., Hersam, M. C. & Van Duyne, R. P. Recent Advances in Tip-Enhanced Raman Spectroscopy. *The Journal of Physical Chemistry Letters* **5**, 3125–3130 (Sept. 18, 2014).
9. Thorn, K. Genetically encoded fluorescent tags. *Molecular Biology of the Cell* **28**, 848–857. ISSN: 1059-1524 (Apr. 1, 2017).
10. Carey, P. *Biochemical Applications of Raman and Resonance Raman Spectroscopies* 276 pp. ISBN: 978-0-12-159650-7 (Academic Press, 1982).
11. Sahoo, S. K., Umapathy, S. & Parker, A. W. Time-resolved resonance Raman spectroscopy: exploring reactive intermediates. *Applied Spectroscopy* **65**, 1087–1115. ISSN: 1943-3530 (Oct. 2011).
12. Zhang, Y., Hong, H. & Cai, W. Imaging with Raman Spectroscopy. *Current pharmaceutical biotechnology* **11**, 654–661. ISSN: 1389-2010 (Sept. 1, 2010).
13. Jensen, L., Zhao, L. L., Autschbach, J. & Schatz, G. C. Theory and method for calculating resonance Raman scattering from resonance polarizability derivatives. *The Journal of Chemical Physics* **123**, 174110. ISSN: 0021-9606 (Nov. 1, 2005).
14. Schimelman, J. B. *et al.* Optical properties and electronic transitions of DNA oligonucleotides as a function of composition and stacking sequence. *Physical chemistry chemical physics: PCCP* **17**, 4589–4599. ISSN: 1463-9084 (Feb. 14, 2015).
15. Le Ru, E. C., Blackie, E., Meyer, M. & Etchegoin, P. G. Surface Enhanced Raman Scattering Enhancement Factors: A Comprehensive Study. *The Journal of Physical Chemistry C* **111**, 13794–13803. ISSN: 1932-7447, 1932-7455 (Sept. 2007).
16. Fleischmann, M., Hendra, P. J. & McQuillan, A. J. Raman spectra of pyridine adsorbed at a silver electrode. *Chemical Physics Letters* **26**, 163–166. ISSN: 0009-2614 (May 15, 1974).

17. Jeanmaire, D. L. & Duyne, R. P. V. Surface raman spectroelectrochemistry. Part I. Heterocyclic, aromatic, and aliphatic amines adsorbed on the anodized silver electrode. *Journal of Electroanalytical Chemistry* **84**, 1–20. ISSN: 0022-0728 (Nov. 10, 1977).
18. Albrecht, M. G. & Creighton, J. A. Anomalous intense Raman spectra of pyridine at a silver electrode. *Journal of the American Chemical Society* **99**, 5215–5217. ISSN: 0002-7863 (June 1, 1977).
19. Yang, S., Dai, X., Stogin, B. B. & Wong, T.-S. Ultrasensitive surface-enhanced Raman scattering detection in common fluids. *Proceedings of the National Academy of Sciences* **113**, 268–273. ISSN: 0027-8424, 1091-6490 (Jan. 12, 2016).
20. Haynes, C. L., McFarland, A. D. & Van Duyne, R. P. Surface-Enhanced Raman Spectroscopy. *Analytical Chemistry* **77**, 338 A–346 A. ISSN: 0003-2700 (Sept. 1, 2005).
21. Otto, A. The ‘chemical’ (electronic) contribution to surface-enhanced Raman scattering. *Journal of Raman Spectroscopy* **36**, 497–509. ISSN: 1097-4555 (2005).
22. Kneipp, K. *et al.* Single Molecule Detection Using Surface-Enhanced Raman Scattering (SERS). *Physical Review Letters* **78**, 1667–1670 (Mar. 3, 1997).
23. Nie, S. & Emory, S. R. Probing Single Molecules and Single Nanoparticles by Surface-Enhanced Raman Scattering. *Science* **275**, 1102–1106. ISSN: 0036-8075, 1095-9203 (Feb. 21, 1997).
24. Vo-Dinh, T., Yan, F. & Wabuyele, M. B. Surface-enhanced Raman scattering for medical diagnostics and biological imaging. *Journal of Raman Spectroscopy* **36**, 640–647. ISSN: 1097-4555 (2005).
25. McNay, G., Eustace, D., Smith, W. E., Faulds, K. & Graham, D. Surface-enhanced Raman scattering (SERS) and surface-enhanced resonance Raman scattering (SERRS): a review of applications. *Applied Spectroscopy* **65**, 825–837. ISSN: 1943-3530 (Aug. 2011).
26. *Nanomaterials and Nanoarchitectures: A Complex Review of Current Hot Topics and their Applications* (eds Bardosova, M. & Wagner, T.) (Springer Netherlands, 2015). ISBN: 978-94-017-9920-1. <https://www.springer.com/gb/book/9789401799201> (2019).
27. Camden, J. P., Dieringer, J. A., Zhao, J. & Van Duyne, R. P. Controlled Plasmonic Nanostructures for Surface-Enhanced Spectroscopy and Sensing. *Accounts of Chemical Research* **41**, 1653–1661. ISSN: 0001-4842 (Dec. 16, 2008).
28. *Coherent Raman scattering microscopy* (eds Cheng, J.-X. & Xie, X. S.) (CRC Press, Taylor & Francis Group, Boca Raton, 2013). 590 pp. ISBN: 978-1-4398-6765-5.
29. J. Tipping, W., Lee, M., Serrels, A., G. Brunton, V. & N. Hulme, A. Stimulated Raman scattering microscopy: an emerging tool for drug discovery. *Chemical Society Reviews* **45**, 2075–2089 (2016).
30. Liao, C.-S. & Cheng, J.-X. In Situ and In Vivo Molecular Analysis by Coherent Raman Scattering Microscopy. *Annual review of analytical chemistry (Palo Alto, Calif.)* **9**, 69–93. ISSN: 1936-1327 (June 12, 2016).
31. Marangoni, M. *et al.* Fiber-format CARS spectroscopy by spectral compression of femtosecond pulses from a single laser oscillator. *Optics Letters* **34**, 3262. ISSN: 0146-9592, 1539-4794 (Nov. 1, 2009).
32. Zhang, C., Zhang, D. & Cheng, J.-X. Coherent Raman Scattering Microscopy in Biology and Medicine. *Annual Review of Biomedical Engineering* **17**, 415–445 (2015).

33. Palonpon, A. F. *et al.* Raman and SERS microscopy for molecular imaging of live cells. *Nature Protocols* **8**, 677–692. ISSN: 1750-2799 (Apr. 2013).
34. Cheng, J.-X. & Xie, X. S. Coherent Anti-Stokes Raman Scattering Microscopy: Instrumentation, Theory, and Applications. *The Journal of Physical Chemistry B* **108**, 827–840. ISSN: 1520-6106, 1520-5207 (Jan. 2004).
35. Zhang, D., Wang, P., Slipchenko, M. N. & Cheng, J.-X. Fast vibrational imaging of single cells and tissues by stimulated Raman scattering microscopy. *Accounts of Chemical Research* **47**, 2282–2290. ISSN: 1520-4898 (Aug. 19, 2014).
36. Lorenz, B., Wichmann, C., Stöckel, S., Rösch, P. & Popp, J. Cultivation-Free Raman Spectroscopic Investigations of Bacteria. *Trends in Microbiology* **25**, 413–424. ISSN: 0966842X (May 2017).
37. Maurer, F. P., Christner, M., Hentschke, M. & Rohde, H. Advances in Rapid Identification and Susceptibility Testing of Bacteria in the Clinical Microbiology Laboratory: Implications for Patient Care and Antimicrobial Stewardship Programs. *Infectious Disease Reports* **9**. ISSN: 2036-7430. doi:10.4081/idr.2017.6839. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5391540/> (2019) (Mar. 30, 2017).
38. Vartoukian, S. R., Palmer, R. M. & Wade, W. G. Strategies for culture of ‘unculturable’ bacteria. *FEMS Microbiology Letters* **309**, 1–7. ISSN: 0378-1097 (Aug. 1, 2010).
39. Ramamurthy, T., Ghosh, A., Pazhani, G. P. & Shinoda, S. Current Perspectives on Viable but Non-Culturable (VBNC) Pathogenic Bacteria. *Frontiers in Public Health* **2**. ISSN: 2296-2565. doi:10.3389/fpubh.2014.00103. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4116801/> (2019) (July 31, 2014).
40. Lidstrom, M. E. & Konopka, M. C. The role of physiological heterogeneity in microbial population behavior. *Nature Chemical Biology* **6**, 705–712. ISSN: 1552-4469 (Oct. 2010).
41. Cleary, J. L., Condren, A. R., Zink, K. E. & Sanchez, L. M. Calling All Hosts: Bacterial Communication In Situ. *Chem* **2**, 334–358. ISSN: 2451-9294 (Mar. 9, 2017).
42. Lu, F.-K. *et al.* Multicolor stimulated Raman scattering (SRS) microscopy. *Molecular physics* **110**, 1927–1932. ISSN: 0026-8976 (Aug. 2012).
43. Zhang, X. *et al.* Label-free Live Cell Imaging of Nucleic Acids using Stimulated Raman Scattering (SRS) Microscopy. *Chemphyschem : a European journal of chemical physics and physical chemistry* **13**, 1054–1059. ISSN: 1439-4235 (Mar. 2012).
44. Ji, M. *et al.* Rapid, Label-Free Detection of Brain Tumors with Stimulated Raman Scattering Microscopy. *Science Translational Medicine* **5**, 201ra119–201ra119. ISSN: 1946-6234, 1946-6242 (Sept. 4, 2013).
45. Lu, F.-K. *et al.* Label-free DNA imaging in vivo with stimulated Raman scattering microscopy. *Proceedings of the National Academy of Sciences* **112**, 11624–11629. ISSN: 0027-8424, 1091-6490 (Sept. 15, 2015).
46. Rygula, A. *et al.* Raman spectroscopy of proteins: a review. *Journal of Raman Spectroscopy* **44**, 1061–1076. ISSN: 1097-4555 (2013).
47. Lippert, J. L., Tyminski, D. & Desmeules, P. J. Determination of the secondary structure of proteins by laser Raman spectroscopy. *Journal of the American Chemical Society* **98**, 7075–7080. ISSN: 0002-7863 (Oct. 1976).

48. Czamara, K. *et al.* Raman spectroscopy of lipids: a review. *Journal of Raman Spectroscopy* **46**, 4–20. ISSN: 1097-4555 (2015).
49. Nieva, C. *et al.* The lipid phenotype of breast cancer cells characterized by Raman microspectroscopy: towards a stratification of malignancy. *PloS One* **7**, e46456. ISSN: 1932-6203 (2012).
50. De Gelder, J., De Gussem, K., Vandenabeele, P. & Moens, L. Reference database of Raman spectra of biological molecules. *Journal of Raman Spectroscopy* **38**, 1133–1147. ISSN: 03770486 10974555 (2007).
51. Herget, S. *et al.* Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): Characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans. *BMC Structural Biology* **8**, 35. ISSN: 1472-6807 (Aug. 11, 2008).
52. Wiercigroch, E. *et al.* Raman and infrared spectroscopy of carbohydrates: A review. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **185**, 317–335. ISSN: 1386-1425 (Oct. 5, 2017).
53. Wilson, W. A. *et al.* Regulation of glycogen metabolism in yeast and bacteria. *FEMS Microbiol Rev* **34**, 952–85. ISSN: 1574-6976 (Electronic) 0168-6445 (Linking) (Nov. 2010).
54. Hankermeyer, C. R. & Tjeerdema, R. S. Polyhydroxybutyrate: plastic made and degraded by microorganisms. *Reviews of Environmental Contamination and Toxicology* **159**, 1–24. ISSN: 0179-5953 (1999).
55. Macernis, M. *et al.* Resonance Raman Spectra of Carotenoid Molecules: Influence of Methyl Substitutions. *The Journal of Physical Chemistry A* **119**, 56–66. ISSN: 1089-5639 (Jan. 8, 2015).
56. N, V. K. B., Kampe, B., Rösch, P. & Popp, J. Characterization of carotenoids in soil bacteria and investigation of their photodegradation by UVA radiation via resonance Raman spectroscopy. *Analyst* **140**, 4584–4593. ISSN: 1364-5528 (June 15, 2015).
57. Pahlow, S. *et al.* Isolation and identification of bacteria by means of Raman spectroscopy. *Advanced Drug Delivery Reviews. Pharmaceutical applications of Raman spectroscopy – from diagnosis to therapeutics* **89**, 105–120. ISSN: 0169-409X (July 15, 2015).
58. Stöckel, S., Kirchhoff, J., Neugebauer, U., Rösch, P. & Popp, J. The application of Raman spectroscopy for the detection and identification of microorganisms. *Journal of Raman Spectroscopy* **47**, 89–109. ISSN: 1097-4555 (2016).
59. De Siqueira e Oliveira, F. S., Giana, H. E. & Silveira, L. Discrimination of selected species of pathogenic bacteria using near-infrared Raman spectroscopy and principal components analysis. *Journal of Biomedical Optics* **17**, 107004. ISSN: 1560-2281 (Oct. 2012).
60. Wang, Y., Lee, K. & Irudayaraj, J. Silver Nanosphere SERS Probes for Sensitive Identification of Pathogens. *The Journal of Physical Chemistry C* **114**, 16122–16128. ISSN: 1932-7447 (Oct. 7, 2010).
61. Cheng, I.-F., Chang, H.-C., Chen, T.-Y., Hu, C. & Yang, F.-L. Rapid (<5 min) Identification of Pathogen in Human Blood by Electrokinetic Concentration and Surface-Enhanced Raman Spectroscopy. *Scientific Reports* **3**, 2365. ISSN: 2045-2322 (Aug. 6, 2013).

62. Almarashi, J. F. M., Kapel, N., Wilkinson, T. S. & Telle, H. H. *Raman Spectroscopy of Bacterial Species and Strains Cultivated under Reproducible Conditions* Journal of Spectroscopy. <https://www.hindawi.com/journals/jspec/2012/540490/> (2019).
63. Samek, O. *et al.* Identification of individual biofilm-forming bacterial cells using Raman tweezers. *Journal of Biomedical Optics* **20**, 051038. ISSN: 1083-3668, 1560-2281 (Mar. 2015).
64. Hamasha, K. *et al.* Sensitive and specific discrimination of pathogenic and nonpathogenic *Escherichia coli* using Raman spectroscopy—a comparison of two multivariate analysis techniques. *Biomedical Optics Express* **4**, 481–489. ISSN: 2156-7085 (Apr. 1, 2013).
65. Osorio-Román, I. O. *et al.* Characterization of bacteria using its O-antigen with surface-enhanced Raman scattering. *Analyst* **135**, 1997–2001. ISSN: 1364-5528 (July 19, 2010).
66. Münchberg, U., Rösch, P., Bauer, M. & Popp, J. Raman spectroscopic identification of single bacterial cells under antibiotic influence. *Analytical and Bioanalytical Chemistry* **406**, 3041–3050. ISSN: 1618-2650 (May 2014).
67. Haider, S. *et al.* Raman microspectroscopy reveals long-term extracellular activity of chlamydiae. *Molecular Microbiology* **77**, 687–700. ISSN: 1365-2958 (2010).
68. Silge, A. *et al.* Shedding light on host niches: label-free in situ detection of *Mycobacterium gordonae* via carotenoids in macrophages by Raman microspectroscopy. *Cellular Microbiology* **17**, 832–842. ISSN: 1462-5822 (2015).
69. Tang, M., McEwen, G. D., Wu, Y., Miller, C. D. & Zhou, A. Characterization and analysis of mycobacteria and Gram-negative bacteria and co-culture mixtures by Raman microspectroscopy, FTIR, and atomic force microscopy. *Analytical and Bioanalytical Chemistry* **405**, 1577–1591. ISSN: 1618-2650 (Feb. 1, 2013).
70. Harz, M. *et al.* Direct analysis of clinical relevant single bacterial cells from cerebrospinal fluid during bacterial meningitis by means of micro-Raman spectroscopy. *J Biophotonics* **2**, 70–80. ISSN: 1864-0648 (Electronic) 1864-063X (Linking) (Feb. 2009).
71. Kloss, S., Rosch, P., Pfister, W., Kiehntopf, M. & Popp, J. Toward culture-free Raman spectroscopic identification of pathogens in ascitic fluid. *Anal Chem* **87**, 937–43. ISSN: 1520-6882 (Electronic) 0003-2700 (Linking) (Jan. 20, 2015).
72. Kloss, S. *et al.* Culture independent Raman spectroscopic identification of urinary tract infection pathogens: a proof of principle study. *Anal Chem* **85**, 9610–6. ISSN: 1520-6882 (Electronic) 0003-2700 (Linking) (Oct. 15, 2013).
73. Kloss, S. *et al.* Destruction-free procedure for the isolation of bacteria from sputum samples for Raman spectroscopic analysis. *Anal Bioanal Chem* **407**, 8333–41. ISSN: 1618-2650 (Electronic) 1618-2642 (Linking) (Nov. 2015).
74. Boardman, A. K. *et al.* Rapid Detection of Bacteria from Blood with Surface-Enhanced Raman Spectroscopy. *Analytical Chemistry* **88**, 8026–8035. ISSN: 0003-2700 (Aug. 16, 2016).
75. Moritz, T. J. *et al.* Effect of cefazolin treatment on the nonresonant Raman signatures of the metabolic state of individual *Escherichia coli* cells. *Analytical Chemistry* **82**, 2703–2710. ISSN: 1520-6882 (Apr. 1, 2010).

76. Meisel, S., Stöckel, S., Rösch, P. & Popp, J. Identification of meat-associated pathogens via Raman microspectroscopy. *Food Microbiology* **38**, 36–43. ISSN: 0740-0020 (Apr. 1, 2014).
77. Meisel, S. *et al.* Raman Spectroscopy as a Potential Tool for Detection of *Brucella* spp. in Milk. *Appl. Environ. Microbiol.* **78**, 5575–5583. ISSN: 0099-2240, 1098-5336 (Aug. 15, 2012).
78. Meisel, S., Stöckel, S., Elschner, M., Rösch, P. & Popp, J. Assessment of two isolation techniques for bacteria in milk towards their compatibility with Raman spectroscopy. *Analyst* **136**, 4997–5005. ISSN: 1364-5528 (Nov. 7, 2011).
79. Assaf, A., Cordella, C. B. Y. & Thouand, G. Raman spectroscopy applied to the horizontal methods ISO 6579:2002 to identify *Salmonella* spp. in the food industry. *Analytical and Bioanalytical Chemistry* **406**, 4899–4910. ISSN: 1618-2650 (Aug. 1, 2014).
80. Vossenberg, J. v. d. *et al.* Identification of bacteria in drinking water with Raman spectroscopy. *Analytical Methods* **5**, 2679–2687. ISSN: 1759-9679 (May 16, 2013).
81. Silge, A. *et al.* Identification of water-conditioned *Pseudomonas aeruginosa* by Raman microspectroscopy on a single cell level. *Systematic and Applied Microbiology* **37**, 360–367. ISSN: 0723-2020 (July 1, 2014).
82. Kusić, D., Kampe, B., Rösch, P. & Popp, J. Identification of water pathogens by Raman microspectroscopy. *Water Research* **48**, 179–189. ISSN: 0043-1354 (Jan. 1, 2014).
83. Jorge-Villar, S. E. & Edwards, H. G. Microorganism Response to Stressed Terrestrial Environments: A Raman Spectroscopic Perspective of Extremophilic Life Strategies. *Life : Open Access Journal* **3**, 276–294. ISSN: 2075-1729 (Mar. 13, 2013).
84. Jehlička, J. *et al.* Potential and limits of Raman spectroscopy for carotenoid detection in microorganisms: implications for astrobiology. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* **372**. ISSN: 1364-503X. doi:10.1098/rsta.2014.0199. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4223861/> (2019) (Dec. 13, 2014).
85. Jehlička, J. & Oren, A. Raman spectroscopy in halophile research. *Frontiers in Microbiology* **4**. ISSN: 1664-302X. doi:10.3389/fmicb.2013.00380. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3857566/> (2019) (Dec. 10, 2013).
86. Jehlička, J., Edwards, H. G. M. & Oren, A. Raman Spectroscopy of Microbial Pigments. *Applied and Environmental Microbiology* **80** (ed Müller, V.) 3286–3295. ISSN: 0099-2240, 1098-5336 (June 1, 2014).
87. Okada, K. *et al.* Nondestructive Structural Analysis of Photosynthetic Pigments in Living *Rhodobacter sphaeroides* Mutants by Near-Infrared Fourier Transform Raman Spectroscopy. *Applied Spectroscopy* **46**, 518–523 (Mar. 1, 1992).
88. Koyama, Y., Umemoto, Y., Akamatsu, A., Uehara, K. & Tanaka, M. Raman spectra of chlorophyll forms. *Journal of Molecular Structure* **146**, 273–287. ISSN: 0022-2860 (Aug. 1, 1986).
89. Edwards, H. G. *et al.* Raman Spectroscopic Analysis of Geological and Biogeological Specimens of Relevance to the ExoMars Mission. *Astrobiology* **13**, 543–549. ISSN: 1531-1074 (June 1, 2013).

90. *Habitability of other planets and satellites* (eds De Vera, J.-P. & Seckbach, J.) *Cellular origin, life in extreme habitats and astrobiology* **volume 28**. OCLC: ocn828487709 (Springer, Dordrecht ; New York, 2013). 419 pp. ISBN: 978-94-007-6545-0.
91. Trevisan, J., Angelov, P. P., Carmichael, P. L., Scott, A. D. & Martin, F. L. Extracting biological information with computational analysis of Fourier-transform infrared (FTIR) biospectroscopy datasets: current practices to future perspectives. *Analyst* **137**, 3202–3215. ISSN: 1364-5528 (June 18, 2012).
92. Jolliffe, I. T. *Principal Component Analysis* 2nd ed. ISBN: 978-0-387-95442-4. <https://www.springer.com/gb/book/9780387954424> (2019) (Springer-Verlag, New York, 2002).
93. McLachlan, G. J. *Discriminant analysis and statistical pattern recognition* Google-Books-ID: 0PFQAAAAMAAJ. 552 pp. ISBN: 978-0-471-61531-6 (Wiley, 1992).
94. Martin, F. L. *et al.* Distinguishing cell types or populations based on the computational analysis of their infrared spectra. *Nature Protocols* **5**, 1748–1760. ISSN: 1754-2189, 1750-2799 (Nov. 2010).
95. De Sousa Marques, A., de Melo, M. C. N., Cidral, T. A. & de Lima, K. M. G. Feature selection strategies for identification of *Staphylococcus aureus* recovered in blood cultures using FT-IR spectroscopy successive projections algorithm for variable selection: A case study. *Journal of Microbiological Methods* **98**, 26–30. ISSN: 0167-7012 (Mar. 1, 2014).
96. Lasch, P., Haensch, W., Naumann, D. & Diem, M. Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **1688**, 176–186. ISSN: 0925-4439 (Mar. 2, 2004).
97. Butler, H. J. *et al.* Using Raman spectroscopy to characterize biological materials. *Nature Protocols* **11**, 664–687. ISSN: 1750-2799 (Apr. 2016).
98. Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C. & Popp, J. Sample size planning for classification models. *Analytica Chimica Acta* **760**, 25–33. ISSN: 1873-4324 (Jan. 14, 2013).
99. Esbensen, K. H. & Geladi, P. Principles of Proper Validation: use and abuse of re-sampling for validation. *Journal of Chemometrics* **24**, 168–187. ISSN: 1099-128X (2010).
100. Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician* **46**, 175–185. ISSN: 0003-1305 (1992).
101. Yao, Z. & Ruzzo, W. L. A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. *BMC Bioinformatics* **7**, S11. ISSN: 1471-2105 (Mar. 20, 2006).
102. Parry, R. M. *et al.* k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. *The Pharmacogenomics Journal* **10**, 292–309. ISSN: 1470-269X (Aug. 2010).
103. Wagner, F., Yan, Y. & Yanai, I. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *bioRxiv*, 217737 (Apr. 9, 2018).
104. Cui, X. *et al.* Analysis and classification of kidney stones based on Raman spectroscopy. *Biomedical Optics Express* **9**, 4175–4183. ISSN: 2156-7085 (Aug. 9, 2018).

105. Li, B., Wei, Y., Duan, H., Xi, L. & Wu, X. Discrimination of the geographical origin of *Codonopsis pilosula* using near infrared diffuse reflection spectroscopy coupled with random forests and k-nearest neighbor methods. *Vibrational Spectroscopy* **62**, 17–22. ISSN: 0924-2031 (Sept. 1, 2012).
106. Tin Kam Ho. *Random decision forests* in *Proceedings of 3rd International Conference on Document Analysis and Recognition* Proceedings of 3rd International Conference on Document Analysis and Recognition. **1** (Aug. 1995), 278–282 vol.1. doi:10.1109/ICDAR.1995.598994.
107. Díaz-Uriarte, R. & Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3. ISSN: 1471-2105 (Jan. 6, 2006).
108. Zucknick, M., Richardson, S. & Stronach, E. A. Comparing the Characteristics of Gene Expression Profiles Derived by Univariate and Multivariate Classification Methods. *Statistical applications in genetics and molecular biology* **7**, Article7. ISSN: 1544-6115 (2008).
109. Beckmann, M. *et al.* Metabolite fingerprinting of urine suggests breed-specific dietary metabolism differences in domestic dogs. *The British Journal of Nutrition* **103**, 1127–1138. ISSN: 1475-2662 (Apr. 2010).
110. Huynh, T. *et al.* Estimating CT Image From MRI Data Using Structured Random Forest and Auto-Context Model. *IEEE Transactions on Medical Imaging* **35**, 174–183. ISSN: 0278-0062 (Jan. 2016).
111. Amjad, A., Ullah, R., Khan, S., Bilal, M. & Khan, A. Raman spectroscopy based analysis of milk using random forest classification. *Vibrational Spectroscopy* **99**, 124–129. ISSN: 0924-2031 (Nov. 1, 2018).
112. Qi, Y. Random Forest for Bioinformatics, 18.
113. Seo, T.-K. Classification of nucleotide sequences using support vector machines. *Journal of Molecular Evolution* **71**, 250–267. ISSN: 1432-1432 (Oct. 2010).
114. Öz, E. & Kaya, H. Support vector machines for quality control of DNA sequencing. *Journal of Inequalities and Applications* **2013**, 85. ISSN: 1029-242X (Mar. 4, 2013).
115. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & Jones, D. T. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *Journal of Molecular Biology* **337**, 635–645. ISSN: 0022-2836 (Mar. 26, 2004).
116. Zavaljevski, N., Stevens, F. J. & Reifman, J. Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics (Oxford, England)* **18**, 689–696. ISSN: 1367-4803 (May 2002).
117. Devos, O., Ruckebusch, C., Durand, A., Duponchel, L. & Huvenne, J.-P. Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation. *Chemometrics and Intelligent Laboratory Systems* **96**, 27–33. ISSN: 0169-7439 (Mar. 15, 2009).
118. Sattlecker, M., Bessant, C. M., Smith, J. A. & Stone, N. Investigation of support vector machines and Raman spectroscopy for lymph node diagnostics. *The Analyst* **135**, 895–901 (2010).
119. Morais, C. L. M., Costa, F. S. L. & Lima, K. M. G. Variable selection with a support vector machine for discriminating *Cryptococcus* fungal species based on ATR-FTIR spectroscopy. *Analytical Methods* **9**, 2964–2970. ISSN: 1759-9679 (May 25, 2017).

120. Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning* **20**, 273–297. ISSN: 1573-0565 (Sept. 1, 1995).
121. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* **5**, 115–133. ISSN: 1522-9602 (Dec. 1, 1943).
122. Lancashire, L. J., Lemetre, C. & Ball, G. R. An introduction to artificial neural networks in bioinformatics—application to complex microarray and mass spectrometry datasets in cancer studies. *Briefings in Bioinformatics* **10**, 315–329. ISSN: 1467-5463 (May 1, 2009).
123. Radajewski, S., Ineson, P., Parekh, N. R. & Murrell, J. C. Stable-isotope probing as a tool in microbial ecology. *Nature* **403**, 646–649. ISSN: 1476-4687 (Feb. 2000).
124. Abraham, W.-R. Applications and impacts of stable isotope probing for analysis of microbial interactions. *Applied Microbiology and Biotechnology* **98**, 4817–4828. ISSN: 1432-0614 (June 1, 2014).
125. Tang, J. Microbial metabolomics. *Current Genomics* **12**, 391–403. ISSN: 1875-5488 (Sept. 2011).
126. Wang, Y., Huang, W. E., Cui, L. & Wagner, M. Single cell stable isotope probing in microbiology using Raman microspectroscopy. *Current Opinion in Biotechnology* **41**, 34–42. ISSN: 1879-0429 (2016).
127. Huang, W. E., Griffiths, R. I., Thompson, I. P., Bailey, M. J. & Whiteley, A. S. Raman Microscopic Analysis of Single Microbial Cells. *Analytical Chemistry* **76**, 4452–4458. ISSN: 0003-2700 (Aug. 1, 2004).
128. Li, M., Huang, W. E., Gibson, C. M., Fowler, P. W. & Jousset, A. Stable Isotope Probing and Raman Spectroscopy for Monitoring Carbon Flow in a Food Chain and Revealing Metabolic Pathway. *Analytical Chemistry* **85**, 1642–1649. ISSN: 0003-2700 (Feb. 5, 2013).
129. Kubryk, P. *et al.* Exploring the Potential of Stable Isotope (Resonance) Raman Microspectroscopy and Surface-Enhanced Raman Scattering for the Analysis of Microorganisms at Single Cell Level. *Analytical Chemistry* **87**, 6622–6630. ISSN: 0003-2700 (July 7, 2015).
130. Eichorst, S. A. *et al.* Advancements in the application of NanoSIMS and Raman microspectroscopy to investigate the activity of microbial cells in soils. *FEMS Microbiology Ecology* **91**. ISSN: 0168-6496. doi:10.1093/femsec/fiv106. <https://academic.oup.com/femsec/article/91/10/fiv106/2467303> (2019) (Oct. 1, 2015).
131. Zhang, D. *et al.* Magnetic nanoparticle-mediated isolation of functional bacteria in a complex microbial community. *The ISME Journal* **9**, 603–614. ISSN: 1751-7370 (Mar. 2015).
132. Kumar B. N., V., Guo, S., Bocklitz, T., Rösch, P. & Popp, J. Demonstration of Carbon Catabolite Repression in Naphthalene Degrading Soil Bacteria via Raman Spectroscopy Based Stable Isotope Probing. *Analytical Chemistry* **88**, 7574–7582. ISSN: 0003-2700 (Aug. 2, 2016).
133. Li, M. *et al.* Rapid resonance Raman microspectroscopy to probe carbon dioxide fixation by single cells in microbial communities. *The ISME journal* **6**, 875–885. ISSN: 1751-7370 (Apr. 2012).

134. Huang, W. E., Li, M., Jarvis, R. M., Goodacre, R. & Banwart, S. A. in *Advances in Applied Microbiology* 153–186 (Academic Press, Jan. 1, 2010).
doi:10.1016/S0065-2164(10)70005-8.
<http://www.sciencedirect.com/science/article/pii/S0065216410700058> (2019).
135. Wang, Y. *et al.* Raman Activated Cell Ejection for Isolation of Single Cells. *Analytical Chemistry* **85**, 10697–10701. ISSN: 0003-2700 (Nov. 19, 2013).
136. Muhamadali, H., Chisanga, M., Subaihi, A. & Goodacre, R. Combining Raman and FT-IR Spectroscopy with Quantitative Isotopic Labeling for Differentiation of *E. coli* Cells at Community and Single Cell Levels. *Analytical Chemistry* **87**, 4578–4586. ISSN: 0003-2700, 1520-6882 (Apr. 21, 2015).
137. Cui, L., Yang, K., Zhou, G., Huang, W. E. & Zhu, Y.-G. Surface-Enhanced Raman Spectroscopy Combined with Stable Isotope Probing to Monitor Nitrogen Assimilation at Both Bulk and Single-Cell Level. *Analytical Chemistry* **89**, 5793–5800. ISSN: 0003-2700 (June 6, 2017).
138. Chisanga, M., Muhamadali, H., Kimber, R. & Goodacre, R. Quantitative detection of isotopically enriched *E. coli* cells by SERS. *Faraday Discussions* **205**, 331–343. ISSN: 1364-5498 (2017).
139. Van Manen, H.-J., Lenferink, A. & Otto, C. Noninvasive Imaging of Protein Metabolic Labeling in Single Human Cells Using Stable Isotopes and Raman Microscopy. *Analytical Chemistry* **80**, 9576–9582. ISSN: 0003-2700 (Dec. 15, 2008).
140. Wei, L., Yu, Y., Shen, Y., Wang, M. C. & Min, W. Vibrational imaging of newly synthesized proteins in live cells by stimulated Raman scattering microscopy. *Proceedings of the National Academy of Sciences* **110**, 11226–11231. ISSN: 0027-8424, 1091-6490 (July 9, 2013).
141. Stiebing, C. *et al.* Complexity of fatty acid distribution inside human macrophages on single cell level using Raman micro-spectroscopy. *Analytical and Bioanalytical Chemistry* **406**, 7037–7046. ISSN: 1618-2650 (Nov. 2014).
142. Li, J. & Cheng, J.-X. Direct Visualization of De novo Lipogenesis in Single Living Cells. *Scientific Reports* **4**, 6807 (Oct. 29, 2014).
143. Miyagi, M. & Kasumov, T. Monitoring the synthesis of biomolecules using mass spectrometry. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* **374**. ISSN: 1364-503X. doi:10.1098/rsta.2015.0378 (Oct. 28, 2016).
144. Shi, L. *et al.* Optical imaging of metabolic dynamics in animals. *Nature Communications* **9**. ISSN: 2041-1723. doi:10.1038/s41467-018-05401-3.
<http://www.nature.com/articles/s41467-018-05401-3> (2019) (Dec. 2018).
145. Berry, D. *et al.* Tracking heavy water (D₂O) incorporation for identifying and sorting active microbial cells. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E194–203. ISSN: 1091-6490 (Jan. 13, 2015).
146. Huang, W. E. *et al.* Raman-FISH: combining stable-isotope Raman spectroscopy and fluorescence in situ hybridization for the single cell analysis of identity and function. *Environmental Microbiology* **9**, 1878–1889. ISSN: 1462-2912 (Aug. 2007).

147. Tao, Y. *et al.* Metabolic-Activity-Based Assessment of Antimicrobial Effects by D₂O-Labeled Single-Cell Raman Microspectroscopy. *Analytical Chemistry* **89**, 4108–4115. ISSN: 0003-2700 (Apr. 4, 2017).
148. Song, Y. *et al.* Raman-Deuterium Isotope Probing for in-situ identification of antimicrobial resistant bacteria in Thames River. *Scientific Reports* **7**, 16648. ISSN: 2045-2322 (Nov. 30, 2017).
149. Song, Y. *et al.* Single-cell genomics based on Raman sorting reveals novel carotenoid-containing bacteria in the Red Sea. *Microbial Biotechnology* **10**, 125–137. ISSN: 1751-7915 (Oct. 17, 2016).
150. Atzrodt, J., Derdau, V., Fey, T. & Zimmermann, J. The renaissance of H/D exchange. *Angewandte Chemie (International Ed. in English)* **46**, 7744–7765. ISSN: 1433-7851 (2007).
151. Herridge, D. F., Peoples, M. B. & Boddey, R. M. Global inputs of biological nitrogen fixation in agricultural systems. *Plant and Soil* **311**, 1–18 (2008).
152. Udvardi, M. & Poole, P. S. Transport and metabolism in legume-rhizobia symbioses. *Annu Rev Plant Biol* **64**, 781–805. ISSN: 1545-2123 (Electronic) 1543-5008 (Linking) (2013).
153. Long, S. R. Rhizobium-legume nodulation: life together in the underground. *Cell* **56**, 203–214. ISSN: 0092-8674 (Jan. 27, 1989).
154. Zaidi, A., Khan, M. S. & Musarrat, J. *Microbes for legume improvement* ISBN: 978-3-319-59173-5 (Springer Berlin Heidelberg, New York, NY, 2017).
155. Beringer, J. E., Brewin, N., Johnston, A. W. B., Schulman, H. M. & Hopwood, D. A. The Rhizobium-Legume Symbiosis. *Proceedings of the Royal Society of London. Series B, Biological Sciences* **204**, 219–233. ISSN: 0080-4649 (1979).
156. Den Herder, G. & Parniske, M. The unbearable naivety of legumes in symbiosis. *Curr Opin Plant Biol* **12**, 491–9. ISSN: 1879-0356 (Electronic) 1369-5266 (Linking) (2009).
157. Batut, J. *et al.* Localization of a symbiotic fix region on Rhizobium meliloti pSym megaplasmid more than 200 kilobases from the nod-nif region. *MGG Molecular & General Genetics* **199**, 232–239. ISSN: 0026-3925, 1617-4623 (1985).
158. Gubler, M., Zürcher, T. & Hennecke, H. The Bradyrhizobium japonicum fixBCX operon: identification of fixX and of a 5' mRNA region affecting the level of the fixSCX transcript. *Molecular Microbiology* **3**, 141–148. ISSN: 0950-382X, 1365-2958 (Feb. 1989).
159. Kaminski, P. A. *et al.* Characterization of the fixABC region of Azorhizobium caulinodans ORS571 and identification of a new nitrogen fixation gene. *Molecular & general genetics: MGG* **214**, 496–502. ISSN: 0026-8925 (Nov. 1988).
160. Lodwig, E. M. *et al.* Role of polyhydroxybutyrate and glycogen as carbon storage compounds in pea and bean bacteroids. *Mol Plant Microbe Interact* **18**, 67–74. ISSN: 0894-0282 (Print) 0894-0282 (Linking) (2005).
161. Terpolilli, J. J. *et al.* Lipogenesis and redox balance in nitrogen-fixing pea bacteroids. *J Bacteriol* **198**, 20. ISSN: 1098-5530 (Electronic) 0021-9193 (Linking) (2016).
162. Madison, L. L. & Huisman, G. W. Metabolic engineering of poly(3-hydroxyalkanoates): from DNA to plastic. *Microbiol Mol Biol Rev* **63**, 21–53. ISSN: 1092-2172 (Print) 1092-2172 (Linking) (1999).

163. Anderson, A. J. & Dawes, E. A. Occurrence, metabolism, metabolic role, and industrial uses of bacterial polyhydroxyalkanoates. *Microbiol Rev* **54**, 450–72. ISSN: 0146-0749 (Print) 0146-0749 (Linking) (1990).
164. Bergersen, F. J., Peoples, M. B. & Turner, G. L. A Role for poly-Formula-hydroxybutyrate in Bacteroids of Soybean Root Nodules. *Proc. R. Soc. B* **245**, 59–64. ISSN: 0962-8452 1471-2954 (1991).
165. Bode, G., Mauch, F., Ditschuneit, H. & Malfertheiner, P. Identification of structures containing polyphosphate in *Helicobacter pylori*. *J Gen Microbiol* **139**, 3029–33. ISSN: 0022-1287 (Print) 0022-1287 (Linking) (1993).
166. Pallerla, S. R. *et al.* Formation of volutin granules in *Corynebacterium glutamicum*. *FEMS Microbiol Lett* **243**, 133–40. ISSN: 0378-1097 (Print) 0378-1097 (Linking) (2005).
167. Bond, P. L., Erhart, R., Wagner, M., Keller, J. & Blackall, L. L. Identification of some of the major groups of bacteria in efficient and nonefficient biological phosphorus removal activated sludge systems. *Applied and Environmental Microbiology* **65**, 4077–4084. ISSN: 0099-2240 (1999).
168. Majed, N. & Gu, A. Z. Application of Raman microscopy for simultaneous and quantitative evaluation of multiple intracellular polymers dynamics functionally relevant to enhanced biological phosphorus removal processes. *Environ Sci Technol* **44**, 8601–8. ISSN: 1520-5851 (Electronic) 0013-936X (Linking) (2010).
169. Beringer, J. E. R factor transfer in *Rhizobium leguminosarum*. *J Gen Microbiol* **84**, 188–98. ISSN: 0022-1287 (Print) 0022-1287 (Linking) (1974).
170. Johnston, A. W. & Beringer, J. E. Identification of the rhizobium strains in pea root nodules using genetic markers. *J Gen Microbiol* **87**, 343–50. ISSN: 0022-1287 (Print) 0022-1287 (Linking) (1975).
171. Poole, P. S., Schofield, N. A., Reid, C. J., Drew, E. M. & Walshaw, D. L. Identification of chromosomal genes located downstream of *dctD* that affect the requirement for calcium and the lipopolysaccharide layer of *Rhizobium leguminosarum*. *Microbiology* **140**, 2797–809. ISSN: 1350-0872 (Print) 1350-0872 (Linking) (1994).
172. Gotz, R., Evans, I. J., Downie, J. A. & Johnston, A. W. B. Identification of the host-range DNA which allows *Rhizobium leguminosarum* strain TOM to nodulate cv. Afghanistan peas. *Mol.Gen. Genet.* **201**, 296–300. ISSN: 0026-3925 1617-4623 (1985).
173. Nikel, P. I., Chavarria, M., Martinez-Garcia, E., Taylor, A. C. & de Lorenzo, V. Accumulation of inorganic polyphosphate enables stress endurance and catalytic vigour in *Pseudomonas putida* KT2440. *Microb Cell Fact* **12**, 50. ISSN: 1475-2859 (Electronic) 1475-2859 (Linking) (2013).
174. Xu, J., Webb, I., Poole, P. & Huang, W. E. Label-Free Discrimination of Rhizobial Bacteroids and Mutants by Single-Cell Raman Microspectroscopy. *Analytical Chemistry* **89**, 6336–6340. ISSN: 0003-2700 (June 20, 2017).
175. Poole, P. S., Blyth, A., Reid, C. J. & Walters, K. myo-Inositol catabolism and catabolite regulation in *Rhizobium leguminosarum* bv. *viciae*. *Microbiology* **140**, 2787–2795. ISSN: 1350-0872 1465-2080 (1994).

176. Hu, S., Morris, I. K., Singh, J. P., Smith, K. M. & Spiro, T. G. Complete assignment of cytochrome c resonance Raman spectra via enzymic reconstitution with isotopically labeled hemes. *Journal of the American Chemical Society* **115**, 12446–12458. ISSN: 0002-7863 (1993).
177. Okada, M. *et al.* Label-free Raman observation of cytochrome c dynamics during apoptosis. *Proc Natl Acad Sci U S A* **109**, 28–32. ISSN: 1091-6490 (Electronic) 0027-8424 (Linking) (2012).
178. Maquelin, K. *et al.* Identification of medically relevant microorganisms by vibrational spectroscopy. *J Microbiol Methods* **51**, 255–71. ISSN: 0167-7012 (Print) 0167-7012 (Linking) (2002).
179. Ruiz-Chica, A. J., Medina, M. A., Sánchez-Jiménez, F. & Ramírez, F. J. Characterization by Raman spectroscopy of conformational changes on guanine–cytosine and adenine–thymine oligonucleotides induced by aminooxy analogues of spermidine. *Journal of Raman Spectroscopy* **35**, 93–100. ISSN: 0377-0486 1097-4555 (2004).
180. Preisig, O., Zufferey, R., ThonyMeyer, L., Appleby, C. A. & Hennecke, H. A high-affinity cbb(3)-type cytochrome oxidase terminates the symbiosis-specific respiratory chain of *Bradyrhizobium japonicum*. *J. Bacteriol.* **178**, 1532–1538. ISSN: 0021-9193 (1996).
181. Lopez, O. *et al.* Regulation of gene expression in response to oxygen in *Rhizobium etli*: role of FnrN in fixNOQP expression and in symbiotic nitrogen fixation. *J Bacteriol* **183**, 6999–7006. ISSN: 0021-9193 (Print) 0021-9193 (Linking) (2001).
182. PaaU, A. S., Lee, D. & Cowles, J. R. Comparison of Nucleic Acid Content in Populations of Free-Living and Symbiotic *Rhizobium meliloti* by Flow Microfluorometry. *Journal of Bacteriology*, 1156–1158 (1977).
183. Reijnders, L., Visser, L., Aalbers, A. M., Van Kammen, A. & Houwers, A. A comparison of DNA from free living and endosymbiotic *Rhizobium leguminosarum* (strain PRE). *Biochim Biophys Acta* **414**, 206–16. ISSN: 0006-3002 (Print) 0006-3002 (Linking) (1975).
184. Sutton, W. D. Some features of the DNA of *Rhizobium* bacteroids and bacteria. *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis* **366**, 1–10. ISSN: 00052787 (1974).
185. Bergersen, F. J. The bacterial component of soybean root nodules; changes in respiratory activity, cell dry weight and nucleic acid content with increasing nodule age. *J Gen Microbiol* **19**, 312–23. ISSN: 0022-1287 (Print) 0022-1287 (Linking) (1958).
186. Dilworth, M. J. & Williams, D. C. Nucleic Acid Changes in Bacteroids of *Rhizobium lupini* During Nodule Development. *Journal of General Microbiology* **48**, 31–36. ISSN: 0022-1287 (1967).
187. PAAU, A. S., ORO, J. & COWLES, J. R. DNA Content of Free Living Rhizobia and Bacteroids of Various *Rhizobium*-Legume Associations. *Plant Physiol.* **63**, 402–405 (1979).
188. Sutton, W. D., Van Den Bos, R. C. & Bisseling, T. The DNA content of *Rhizobium* (strain NZP 2257) bacteroids and bacteria. *Plant Science Letters* **12**, 145–149. ISSN: 03044211 (1978).
189. Ding, C. & Cantor, C. R. Quantitative analysis of nucleic acids—the last few years of progress. *J Biochem Mol Biol* **37**, 1–10. ISSN: 1225-8687 (Print) 1225-8687 (Linking) (2004).

190. Nielsen, K. *et al.* Comparison of five DNA quantification methods. *Forensic Sci Int Genet* **2**, 226–30. ISSN: 1878-0326 (Electronic) 1872-4973 (Linking) (2008).
191. Desjardins, P. & Conklin, D. NanoDrop microvolume quantitation of nucleic acids. *J Vis Exp*. ISSN: 1940-087X (Electronic) 1940-087X (Linking). doi:10.3791/2565. <https://www.ncbi.nlm.nih.gov/pubmed/21189466> (2010).
192. Gałat, A. Study of the Raman scattering and infrared absorption spectra of branched polysaccharides. *Acta Biochimica Polonica* **27**, 135–142. ISSN: 0001-527X (1980).
193. De Gelder, J. *et al.* Monitoring Poly(3-hydroxybutyrate) Production in *Cupriavidus necator* DSM 428 (H16) with Raman Spectroscopy. *Analytical Chemistry* **80**, 2155–2160. ISSN: 0003-2700, 1520-6882 (Mar. 2008).
194. Achbergerova, L. & Nahalka, J. Polyphosphate—an ancient energy source and active metabolic regulator. *Microb Cell Fact* **10**, 63. ISSN: 1475-2859 (Electronic) 1475-2859 (Linking) (2011).
195. Cevallos, M. A., Encarnacion, S., Leija, A., Mora, Y. & Mora, J. Genetic and physiological characterization of a *Rhizobium etli* mutant strain unable to synthesize poly-beta-hydroxybutyrate. *J Bacteriol* **178**, 1646–54. ISSN: 0021-9193 (Print) 0021-9193 (Linking) (1996).
196. Senior, P. J., Beech, G. A., Ritchie, G. A. & Dawes, E. A. The role of oxygen limitation in the formation of poly- β -hydroxybutyrate during batch and continuous culture of *Azotobacter beijerinckii*. *The Biochemical Journal* **128**, 1193–1201. ISSN: 0264-6021 (Aug. 1972).
197. Rao, N. N., Liu, S. & Kornberg, A. Inorganic Polyphosphate in *Escherichia coli*: the Phosphate Regulon and the Stringent Response. *Journal of Bacteriology* **180**, 2186–2193. ISSN: 0021-9193 (Apr. 1998).
198. Gant, T. G. Using Deuterium in Drug Discovery: Leaving the Label in the Drug. *Journal of Medicinal Chemistry* **57**, 3595–3611. ISSN: 0022-2623 (May 8, 2014).
199. Junk, T. & Catallo, W. J. Hydrogen isotope exchange reactions involving C–H (D, T) bonds. *Chemical Society Reviews* **26**, 401–406. ISSN: 1460-4744 (Jan. 1, 1997).
200. Nikel, P. I., Chavarría, M., Fuhrer, T., Sauer, U. & de Lorenzo, V. *Pseudomonas putida* KT2440 Strain Metabolizes Glucose through a Cycle Formed by Enzymes of the Entner-Doudoroff, Embden-Meyerhof-Parnas, and Pentose Phosphate Pathways. *The Journal of Biological Chemistry* **290**, 25920–25932. ISSN: 1083-351X (Oct. 23, 2015).
201. McClure, N. C., Weightman, A. J. & Fry, J. C. Survival of *Pseudomonas putida* UWC1 containing cloned catabolic genes in a model activated-sludge unit. *Applied and Environmental Microbiology* **55**, 2627–2634. ISSN: 0099-2240 (Oct. 1989).
202. Simon, M. J. *et al.* Sequences of genes encoding naphthalene dioxygenase in *Pseudomonas putida* strains G7 and NCIB 9816-4. *Gene* **127**, 31–37. ISSN: 0378-1119 (May 15, 1993).
203. Huang, W. E. *et al.* Resolving Genetic Functions within Microbial Populations: In Situ Analyses Using rRNA and mRNA Stable Isotope Probing Coupled with Single-Cell Raman-Fluorescence In Situ Hybridization. *Appl. Environ. Microbiol.* **75**, 234–241. ISSN: 0099-2240, 1098-5336 (Jan. 1, 2009).

204. Marchesi, J. R. *et al.* Design and evaluation of useful bacterium-specific PCR primers that amplify genes coding for bacterial 16S rRNA. *Applied and Environmental Microbiology* **64**, 795–799. ISSN: 0099-2240 (Feb. 1998).
205. Wang, Y. *et al.* Reverse and Multiple Stable Isotope Probing to Study Bacterial Metabolism and Interactions at the Single Cell Level. *Analytical Chemistry* **88**, 9443–9450. ISSN: 0003-2700 (Oct. 4, 2016).
206. Babul, J., Clifton, D., Kretschmer, M. & Fraenkel, D. G. Glucose metabolism in *Escherichia coli* and the effect of increased aldolase. *Biochemistry* **32**, 4685–4692. ISSN: 0006-2960 (May 4, 1993).
207. Franco, M. I., Turin, L., Mershin, A. & Skoulakis, E. M. C. Molecular vibration-sensing component in *Drosophila melanogaster* olfaction. *Proceedings of the National Academy of Sciences* **108**, 3797–3802. ISSN: 0027-8424, 1091-6490 (Mar. 1, 2011).
208. Romano, A. H. & Conway, T. Evolution of carbohydrate metabolic pathways. *Research in Microbiology* **147**, 448–455. ISSN: 0923-2508 (Sept. 1996).
209. Sudarsan, S., Dethlefsen, S., Blank, L. M., Siemann-Herzberg, M. & Schmid, A. The Functional Structure of Central Carbon Metabolism in *Pseudomonas putida* KT2440. *Appl. Environ. Microbiol.* **80**, 5292–5303. ISSN: 0099-2240, 1098-5336 (Sept. 1, 2014).
210. *Pseudomonas: Molecular and Applied Biology* (ed Kahlon, R. S.) (Springer International Publishing, 2016). ISBN: 978-3-319-31197-5.
<https://www.springer.com/gp/book/9783319311975> (2019).
211. Selvarasu, S. *et al.* Characterizing *Escherichia coli* DH5 α growth and metabolism in a complex medium using genome-scale flux analysis. *Biotechnology and Bioengineering* **102**, 923–934. ISSN: 1097-0290 (2009).
212. Entner, N. & Doudoroff, M. Glucose and gluconic acid oxidation of *Pseudomonas saccharophila*. *The Journal of Biological Chemistry* **196**, 853–862. ISSN: 0021-9258 (May 1952).
213. Vicente, M. & Cánovas, J. L. Glucolysis in *Pseudomonas putida*: physiological role of alternative routes from the analysis of defective mutants. *Journal of Bacteriology* **116**, 908–914. ISSN: 0021-9193 (Nov. 1973).
214. Castillo, T. d. *et al.* Convergent Peripheral Pathways Catalyze Initial Glucose Catabolism in *Pseudomonas putida*: Genomic and Flux Analysis. *Journal of Bacteriology* **189**, 5142–5152. ISSN: 0021-9193, 1098-5530 (July 15, 2007).
215. Pui, C. F. *et al.* Review Article *Salmonella*: A foodborne pathogen. *Int. Food Res. J.* **18**, 465–473 (2011).
216. Dougan, G. & Baker, S. *Salmonella enterica* serovar Typhi and the pathogenesis of typhoid fever. *Annual Review of Microbiology* **68**, 317–336. ISSN: 1545-3251 (2014).
217. Gordon, M. A. *Salmonella* infections in immunocompromised adults. *J Infect* **56**, 413–22. ISSN: 1532-2742 (Electronic) 0163-4453 (Linking) (June 2008).
218. Hohmann, E. L. Nontyphoidal salmonellosis. *Clin Infect Dis* **32**, 263–9. ISSN: 1058-4838 (Print) 1058-4838 (Linking) (Jan. 15, 2001).
219. Graham, S. M. Nontyphoidal salmonellosis in Africa. *Curr Opin Infect Dis* **23**, 409–14. ISSN: 1473-6527 (Electronic) 0951-7375 (Linking) (Oct. 2010).

220. Gordon, M. A. *et al.* Epidemics of invasive *Salmonella enterica* serovar enteritidis and *S. enterica* Serovar typhimurium infection associated with multidrug resistance among adults and children in Malawi. *Clin Infect Dis* **46**, 963–9. ISSN: 1537-6591 (Electronic) 1058-4838 (Linking) (Apr. 1, 2008).
221. Kingsley, R. A. *et al.* Epidemic multiple drug resistant *Salmonella* Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. *Genome Res* **19**, 2279–87. ISSN: 1549-5469 (Electronic) 1088-9051 (Linking) (Dec. 2009).
222. Leekitcharoenphon, P. *et al.* Genomics of an emerging clone of *Salmonella* serovar Typhimurium ST313 from Nigeria and the Democratic Republic of Congo. *J Infect Dev Ctries* **7**, 696–706. ISSN: 1972-2680 (Electronic) 1972-2680 (Linking) (Oct. 15, 2013).
223. Bumann, D. & Schothorst, J. Intracellular *Salmonella* metabolism. *Cell Microbiol* **19**. ISSN: 1462-5822 (Electronic) 1462-5814 (Linking). doi:10.1111/cmi.12766. <https://www.ncbi.nlm.nih.gov/pubmed/28672057> (Oct. 2017).
224. Thompson, A., Fulde, M. & Tedin, K. The metabolic pathways utilized by *Salmonella* Typhimurium during infection of host cells. *Environ Microbiol Rep* **10**, 140–154. ISSN: 1758-2229 (Electronic) 1758-2229 (Linking) (Apr. 2018).
225. Faucher, S. P., Porwollik, S., Dozois, C. M., McClelland, M. & Daigle, F. Transcriptome of *Salmonella enterica* serovar Typhi within macrophages revealed through the selective capture of transcribed sequences. *Proc Natl Acad Sci U S A* **103**, 1906–11. ISSN: 0027-8424 (Print) 0027-8424 (Linking) (Feb. 7, 2006).
226. Liu, Y. *et al.* Proteomic Analyses of Intracellular *Salmonella enterica* Serovar Typhimurium Reveal Extensive Bacterial Adaptations to Infected Host Epithelial Cells. *Infect Immun* **83**, 2897–906. ISSN: 1098-5522 (Electronic) 0019-9567 (Linking) (July 2015).
227. Shi, L. *et al.* Proteomic analysis of *Salmonella enterica* serovar typhimurium isolated from RAW 264.7 macrophages: identification of a novel protein that contributes to the replication of serovar typhimurium inside macrophages. *J Biol Chem* **281**, 29131–40. ISSN: 0021-9258 (Print) 0021-9258 (Linking) (Sept. 29, 2006).
228. Keller, M. A., Piedrafita, G. & Ralser, M. The widespread role of non-enzymatic reactions in cellular metabolism. *Curr Opin Biotechnol* **34**, 153–61. ISSN: 1879-0429 (Electronic) 0958-1669 (Linking) (Aug. 2015).
229. Helaine, S. *et al.* Dynamics of intracellular bacterial replication at the single cell level. *Proc Natl Acad Sci U S A* **107**, 3746–51. ISSN: 1091-6490 (Electronic) 0027-8424 (Linking) (Feb. 23, 2010).
230. Bumann, D. Heterogeneous host-pathogen encounters: act locally, think globally. *Cell Host Microbe* **17**, 13–9. ISSN: 1934-6069 (Electronic) 1931-3128 (Linking) (Jan. 14, 2015).
231. Knodler, L. A. *Salmonella enterica*: living a double life in epithelial cells. *Curr Opin Microbiol* **23**, 23–31. ISSN: 1879-0364 (Electronic) 1369-5274 (Linking) (Feb. 2015).
232. Alakomi, H. L. *et al.* Lactic acid permeabilizes gram-negative bacteria by disrupting the outer membrane. *Applied and Environmental Microbiology* **66**, 2001–2005. ISSN: 0099-2240 (May 2000).
233. Kollipara, L. & Zahedi, R. P. Protein carbamylation: in vivo modification or in vitro artefact? *Proteomics* **13**, 941–944. ISSN: 1615-9861 (Mar. 2013).

234. Schmidt, F. *et al.* Time-resolved quantitative proteome profiling of host-pathogen interactions: the response of *Staphylococcus aureus* RN1HG to internalisation by human airway epithelial cells. *Proteomics* **10**, 2801–2811. ISSN: 1615-9861 (Aug. 2010).
235. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367–72. ISSN: 1546-1696 (Electronic) 1087-0156 (Linking) (Dec. 2008).
236. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & cellular proteomics: MCP* **13**, 2513–2526. ISSN: 1535-9484 (Sept. 2014).
237. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (Apr. 20, 2015).
238. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* **44**, D471–80. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (D1 Jan. 4, 2016).
239. Góral, J. & Zichy, V. Fourier Transform Raman studies of materials and compounds of biological importance. *Spectrochimica Acta Part A: Molecular Spectroscopy. Applications of Fourier Transform Raman Spectroscopy* **46**, 253–275. ISSN: 0584-8539 (Jan. 1, 1990).
240. Hernández, B., Pflüger, F., Kruglik, S. G. & Ghomi, M. Characteristic Raman lines of phenylalanine analyzed by a multiconformational approach. *Journal of Raman Spectroscopy* **44**, 827–833. ISSN: 1097-4555 (2013).
241. Xu, J., Webb, I., Poole, P. & Huang, W. E. Label-Free Discrimination of Rhizobial Bacteroids and Mutants by Single-Cell Raman Microspectroscopy. *Analytical Chemistry* **89**, 6336–6340. ISSN: 0003-2700 (June 20, 2017).
242. Langridge, G. C. *et al.* Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. *Proc Natl Acad Sci U S A* **112**, 863–8. ISSN: 1091-6490 (Electronic) 0027-8424 (Linking) (Jan. 20, 2015).
243. Matono, T. *et al.* Emergence of Resistance Mutations in *Salmonella enterica* Serovar Typhi Against Fluoroquinolones. *Open Forum Infect Dis* **4**, ofx230. ISSN: 2328-8957 (Print) 2328-8957 (Linking) (2017).
244. Needham, B. D. & Trent, M. S. Fortifying the barrier: the impact of lipid A remodelling on bacterial pathogenesis. *Nat Rev Microbiol* **11**, 467–81. ISSN: 1740-1534 (Electronic) 1740-1526 (Linking) (July 2013).
245. Nuccio, S. P. & Baumler, A. J. Comparative analysis of *Salmonella* genomes identifies a metabolic network for escalating growth in the inflamed gut. *MBio* **5**, e00929–14. ISSN: 2150-7511 (Electronic) (Mar. 18, 2014).
246. Xu, J. *et al.* Raman Deuterium Isotope Probing Reveals Microbial Metabolism at the Single-Cell Level. *Analytical Chemistry* **89**, 13305–13312. ISSN: 0003-2700 (Dec. 19, 2017).
247. Lathrop, S. K. *et al.* Replication of *Salmonella enterica* Serovar Typhimurium in Human Monocyte-Derived Macrophages. *Infection and Immunity* **83**, 2661–2671. ISSN: 1098-5522 (July 2015).

248. Forest, C. G., Ferraro, E., Sabbagh, S. C. & Daigle, F. Intracellular survival of *Salmonella enterica* serovar Typhi in human macrophages is independent of *Salmonella* pathogenicity island (SPI)-2. *Microbiology (Reading, England)* **156**, 3689–3698. ISSN: 1465-2080 (Pt 12 Dec. 2010).
249. Nagl, M. *et al.* Phagocytosis and killing of bacteria by professional phagocytes and dendritic cells. *Clinical and Diagnostic Laboratory Immunology* **9**, 1165–1168. ISSN: 1071-412X (Nov. 2002).
250. Savina, A. & Amigorena, S. Phagocytosis and antigen presentation in dendritic cells. *Immunological Reviews* **219**, 143–156. ISSN: 0105-2896 (Oct. 2007).
251. Cheminay, C., Mohlenbrink, A. & Hensel, M. Intracellular *Salmonella* Inhibit Antigen Presentation by Dendritic Cells. *The Journal of Immunology* **174**, 2892–2899. ISSN: 0022-1767 1550-6606 (2005).
252. Bueno, S. M. *et al.* The capacity of *Salmonella* to survive inside dendritic cells and prevent antigen presentation to T cells is host specific. *Immunology* **124**, 522–533. ISSN: 1365-2567 (Aug. 2008).
253. Fu, G., Wijburg, O. L. C., Cameron, P. U., Price, J. D. & Strugnell, R. A. *Salmonella enterica* Serovar Typhimurium infection of dendritic cells leads to functionally increased expression of the macrophage-derived chemokine. *Infection and Immunity* **73**, 1714–1722. ISSN: 0019-9567 (Mar. 2005).
254. Swart, A. L. & Hensel, M. Interactions of *Salmonella enterica* with dendritic cells. *Virulence* **3**, 660–7. ISSN: 2150-5608 (Electronic) 2150-5594 (Linking) (Nov. 15, 2012).
255. Garcia-Del Portillo, F., Jungnitz, H., Rohde, M. & Guzman, C. A. Interaction of *Salmonella enterica* serotype typhimurium with dendritic cells is defined by targeting to compartments lacking lysosomal membrane glycoproteins. *Infection and Immunity* **68**, 2985–2991. ISSN: 0019-9567 (May 2000).
256. Niedergang, F., Sirard, J. C., Blanc, C. T. & Kraehenbuhl, J. P. Entry and survival of *Salmonella typhimurium* in dendritic cells and presentation of recombinant antigens do not require macrophage-specific virulence factors. *Proc Natl Acad Sci U S A* **97**, 14650–5. ISSN: 0027-8424 (Print) 0027-8424 (Linking) (Dec. 19, 2000).
257. Bonafonte, M. A. *et al.* The relationship between glycogen synthesis, biofilm formation and virulence in *Salmonella enteritidis*. *Fems Microbiology Letters* **191**, 31–36. ISSN: 0378-1097 (Oct. 1, 2000).
258. Diacovich, L., Lorenzi, L., Tomassetti, M., Meresse, S. & Gramajo, H. The infectious intracellular lifestyle of *Salmonella enterica* relies on the adaptation to nutritional conditions within the *Salmonella*-containing vacuole. *Virulence* **8**, 975–992. ISSN: 2150-5608 (Electronic) 2150-5594 (Linking) (Aug. 18, 2017).
259. Okoro, C. K. *et al.* Signatures of adaptation in human invasive *Salmonella Typhimurium* ST313 populations from sub-Saharan Africa. *PLoS Negl Trop Dis* **9**, e0003611. ISSN: 1935-2735 (Electronic) 1935-2727 (Linking) (Mar. 2015).
260. Spector, M. P. *et al.* The medium-/long-chain fatty acyl-CoA dehydrogenase (*fadF*) gene of *Salmonella typhimurium* is a phase 1 starvation-stress response (SSR) locus. *Microbiology* **145** (Pt 1), 15–31. ISSN: 1350-0872 (Print) 1350-0872 (Linking) (Jan. 1999).
261. Guo, L. *et al.* Lipid A acylation and bacterial resistance against vertebrate antimicrobial peptides. *Cell* **95**, 189–98. ISSN: 0092-8674 (Print) 0092-8674 (Linking) (Oct. 16, 1998).

262. Syed, A. & Smith, E. A. Raman Imaging in Cell Membranes, Lipid-Rich Organelles, and Lipid Bilayers. *Annual Review of Analytical Chemistry (Palo Alto, Calif.)* **10**, 271–291. ISSN: 1936-1335 (2017).
263. Parsons, J. B. & Rock, C. O. Bacterial lipids: metabolism and membrane homeostasis. *Progress in Lipid Research* **52**, 249–276. ISSN: 1873-2194 (July 2013).
264. Mantena, R. K. *et al.* Reactive oxygen species are the major antibacterials against *Salmonella Typhimurium* purine auxotrophs in the phagosome of RAW 264.7 cells. *Cell Microbiol* **10**, 1058–73. ISSN: 1462-5822 (Electronic) 1462-5814 (Linking) (May 2008).
265. Holzer, S. U. & Hensel, M. Divergent roles of *Salmonella* pathogenicity island 2 and metabolic traits during interaction of *S. enterica* serovar typhimurium with host cells. *PLoS One* **7**, e33220. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking) (2012).
266. Sprenger, M., Kasper, L., Hensel, M. & Hube, B. Metabolic adaptation of intracellular bacteria and fungi to macrophages. *Int J Med Microbiol*. ISSN: 1618-0607 (Electronic) 1438-4221 (Linking). doi:10.1016/j.ijmm.2017.11.001. <https://www.ncbi.nlm.nih.gov/pubmed/29150190> (Nov. 7, 2017).
267. Dandekar, T. *et al.* *Salmonella*-how a metabolic generalist adopts an intracellular lifestyle during infection. *Front Cell Infect Microbiol* **4**, 191. ISSN: 2235-2988 (Electronic) 2235-2988 (Linking) (2014).
268. Alvarez-Ordóñez, A., Fernandez, A., Lopez, M., Arenas, R. & Bernardo, A. Modifications in membrane fatty acid composition of *Salmonella typhimurium* in response to growth conditions and their effect on heat resistance. *Int J Food Microbiol* **123**, 212–9. ISSN: 0168-1605 (Print) 0168-1605 (Linking) (Apr. 30, 2008).
269. Hurdle, J. G., O'Neill, A. J., Chopra, I. & Lee, R. E. Targeting bacterial membrane function: an underexploited mechanism for treating persistent infections. *Nat Rev Microbiol* **9**, 62–75. ISSN: 1740-1534 (Electronic) 1740-1526 (Linking) (Jan. 2011).
270. Frawley, E. R. *et al.* Iron and citrate export by a major facilitator superfamily pump regulates metabolism and stress resistance in *Salmonella Typhimurium*. *Proc Natl Acad Sci U S A* **110**, 12054–9. ISSN: 1091-6490 (Electronic) 0027-8424 (Linking) (July 16, 2013).
271. Muhlig, A., Behr, J., Scherer, S. & Muller-Herbst, S. Stress response of *Salmonella enterica* serovar typhimurium to acidified nitrite. *Appl Environ Microbiol* **80**, 6373–82. ISSN: 1098-5336 (Electronic) 0099-2240 (Linking) (Oct. 2014).
272. Starosta, A. L., Lassak, J., Jung, K. & Wilson, D. N. The bacterial translation stress response. *FEMS Microbiol Rev* **38**, 1172–201. ISSN: 1574-6976 (Electronic) 0168-6445 (Linking) (Nov. 2014).
273. Eisenreich, W., Heesemann, J., Rudel, T. & Goebel, W. Metabolic host responses to infection by intracellular bacterial pathogens. *Front Cell Infect Microbiol* **3**, 24. ISSN: 2235-2988 (Print) 2235-2988 (Linking) (2013).
274. Biedzka-Sarek, M. & El Skurnik, M. How to outwit the enemy: dendritic cells face *Salmonella*. *APMIS* **114**, 589–600. ISSN: 0903-4641 (Print) 0903-4641 (Linking) (Sept. 2006).
275. Fabrik, I., Hartlova, A., Rehulka, P. & Stulik, J. Serving the new masters - dendritic cells as hosts for stealth intracellular bacteria. *Cell Microbiol* **15**, 1473–83. ISSN: 1462-5822 (Electronic) 1462-5814 (Linking) (Sept. 2013).

276. Malik-Kale, P. *et al.* Salmonella - at home in the host cell. *Front Microbiol* **2**, 125. ISSN: 1664-302X (Electronic) 1664-302X (Linking) (2011).
277. Garcia-Gutierrez, E. *et al.* A Comparison of the ATP Generating Pathways Used by *S. Typhimurium* to Fuel Replication within Human and Murine Macrophage and Epithelial Cell Lines. *PLoS One* **11**, e0150687. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking) (2016).
278. Liu, Z., Que, S., Xu, J. & Peng, T. Alanine aminotransferase-old biomarker and new concept: a review. *Int J Med Sci* **11**, 925–35. ISSN: 1449-1907 (Electronic) 1449-1907 (Linking) (2014).
279. Bowden, S. D., Ramachandran, V. K., Knudsen, G. M., Hinton, J. C. & Thompson, A. An incomplete TCA cycle increases survival of *Salmonella Typhimurium* during infection of resting and activated murine macrophages. *PLoS One* **5**, e13871. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking) (Nov. 8, 2010).
280. O'Callaghan, D., Maskell, D., Liew, F. Y., Easmon, C. S. & Dougan, G. Characterization of aromatic- and purine-dependent *Salmonella typhimurium*: attention, persistence, and ability to induce protective immunity in BALB/c mice. *Infect Immun* **56**, 419–23. ISSN: 0019-9567 (Print) 0019-9567 (Linking) (Feb. 1988).
281. Liu, Y. *et al.* Quantitative Proteomics Charts the Landscape of *Salmonella* Carbon Metabolism within Host Epithelial Cells. *J Proteome Res* **16**, 788–797. ISSN: 1535-3907 (Electronic) 1535-3893 (Linking) (Feb. 3, 2017).
282. Andino, A. *et al.* Survival of *Salmonella enterica* in poultry feed is strain dependent. *Poult Sci* **93**, 441–7. ISSN: 0032-5791 (Print) 0032-5791 (Linking) (Feb. 2014).
283. Feng, Y. & Cronan, J. E. *Escherichia coli* unsaturated fatty acid synthesis: complex transcription of the *fabA* gene and in vivo identification of the essential reaction catalyzed by FabB. *J Biol Chem* **284**, 29526–35. ISSN: 1083-351X (Electronic) 0021-9258 (Linking) (Oct. 23, 2009).
284. Savitzky, A. & Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* **36**, 1627–1639. ISSN: 0003-2700 (July 1, 1964).
285. Eilers, P. H. C. & Boelens, H. F. M. Baseline Correction with Asymmetric Least Squares Smoothing, 24.
286. Coombes, K. R. *et al.* Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics* **5**, 4107–4117. ISSN: 1615-9853 (Nov. 2005).
287. Friedrichs, M. S. A model-free algorithm for the removal of baseline artifacts. *Journal of Biomolecular NMR* **5**, 147–153. ISSN: 1573-5001 (Feb. 1, 1995).
288. Liland, K. H., Almøy, T. & Mevik, B.-H. Optimal Choice of Baseline Correction for Multivariate Calibration of Spectra. *Applied Spectroscopy* **64**, 1007–1016. ISSN: 0003-7028, 1943-3530 (Sept. 2010).
289. Kneen, M. A. & Annegarn, H. J. Algorithm for fitting XRF, SEM and PIXE X-ray spectra backgrounds. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* **109-110**, 209–213. ISSN: 0168-583X (Apr. 1, 1996).

290. Lieber, C. A. & Mahadevan-Jansen, A. Automated Method for Subtraction of Fluorescence from Biological Raman Spectra. *Applied Spectroscopy* **57**, 1363–1367 (Nov. 1, 2003).
291. Baker, M. J. *et al.* Using Fourier transform IR spectroscopy to analyze biological materials. *Nature Protocols* **9**, 1771–1791. ISSN: 1750-2799 (Aug. 2014).
292. Lasch, P. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemometrics and Intelligent Laboratory Systems. Special Issue Section: Selected Papers from the 1st African-European Conference on Chemometrics, Rabat, Morocco, September 2010 Special Issue Section: Preprocessing methods Special Issue Section: Spectroscopic imaging* **117**, 100–114. ISSN: 0169-7439 (Aug. 1, 2012).
293. Maaten, L. v. d. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605. ISSN: ISSN 1533-7928 (Nov 2008).
294. Abdelmoula, W. M. *et al.* Data-driven identification of prognostic tumor subpopulations using spatially mapped t-SNE of mass spectrometry imaging data. *Proceedings of the National Academy of Sciences* **113**, 12244–12249. ISSN: 0027-8424, 1091-6490 (Oct. 25, 2016).
295. Heuer, H. Text comparison using word vector representations and dimensionality reduction. *arXiv:1607.00534 [cs]*. arXiv: 1607 . 00534.
<http://arxiv.org/abs/1607.00534> (2019) (July 2, 2016).
296. Panta, S. R. *et al.* A Tool for Interactive Data Visualization: Application to Over 10,000 Brain Imaging and Phantom MRI Data Sets. *Frontiers in Neuroinformatics* **10**. ISSN: 1662-5196. doi:10.3389/fninf.2016.00009.
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4791544/> (2019) (Mar. 15, 2016).
297. Li, J., Chen, X., Hovy, E. & Jurafsky, D. Visualizing and Understanding Neural Models in NLP. *arXiv:1506.01066 [cs]*. arXiv: 1506.01066. <http://arxiv.org/abs/1506.01066> (2019) (June 2, 2015).
298. Li, W., Cerise, J. E., Yang, Y. & Han, H. Application of t-SNE to human genetic data. *Journal of Bioinformatics and Computational Biology* **15**, 1750017. ISSN: 1757-6334 (Aug. 2017).
299. Kobak, D. & Berens, P. *The art of using t-SNE for single-cell transcriptomics* preprint (Bioinformatics, Oct. 25, 2018). doi:10.1101/453449.
<http://biorxiv.org/lookup/doi/10.1101/453449> (2019).
300. Kyriakides, A., Kastanos, E. & Pitris, C. *Classification of Raman Spectra using Support Vector Machines* in 2009 9th International Conference on Information Technology and Applications in Biomedicine 2009 9th International Conference on Information Technology and Applications in Biomedicine (Nov. 2009), 1–4. doi:10.1109/ITAB.2009.5394428.
301. Hunter, R. & Anis, H. Genetic support vector machines as powerful tools for the analysis of biomedical Raman spectra. *Journal of Raman Spectroscopy* **49**, 1435–1444. ISSN: 1097-4555 (2018).
302. Ullah, R. *et al.* Raman spectroscopy combined with a support vector machine for differentiating between feeding male and female infants mother's milk. *Biomedical Optics Express* **9**, 844–851. ISSN: 2156-7085 (Jan. 26, 2018).

303. Xu, J. *et al.* A new approach to find biomarkers in chronic fatigue syndrome/myalgic encephalomyelitis (CFS/ME) by single-cell Raman micro-spectroscopy. *Analyst*. ISSN: 1364-5528. doi:10.1039/C8AN01437J.
<https://pubs.rsc.org/en/content/articlelanding/2019/an/c8an01437j>
(2019) (Aug. 22, 2018).
304. Liang, B. *et al.* Raman profiling of embryo culture medium to identify aneuploid and euploid embryos. *Fertility and Sterility*. ISSN: 0015-0282.
doi:10.1016/j.fertnstert.2018.11.036.
<http://www.sciencedirect.com/science/article/pii/S0015028218322374>
(2019) (Jan. 22, 2019).
305. Shipp, D. W., Sinjab, F. & Notingher, I. Raman spectroscopy: Techniques and applications in the life sciences. *Advances in Optics and Photonics*, 75.