**Supplementary Information**

**Multi-excitation Raman spectroscopy for label-free, strain-level characterisation of bacterial pathogens in artificial sputum media**

Adam. P. Lister†1,5, Callum. J. Highmore†,2,4, Niall Hanrahan1,5, James Read1,5, Alasdair P. S. Munro3,4, Samuel Tan4, Raymond. N. Allan2,3,6, Saul N. Faust\*3,4,5, Jeremy. S. Webb\*2,5 and Sumeet Mahajan\*1,5

1: School of Chemistry, Faculty of Engineering and Physical Sciences, University of Southampton, SO17 1BJ, Southampton, United Kingdom

2: School of Biological Sciences, Faculty of Environmental and Life Sciences, University of Southampton, SO17 1BJ, Southampton, United Kingdom

3: NIHR Southampton Clinical Research Facility and Biomedical Research Centre, University Hospital Southampton NHS Foundation Trust, Southampton, UK SO16 6YD

4: Faculty of Medicine and Institute for Life Sciences, University of Southampton, Southampton, United Kingdom

5: National Biofilms Innovation Centre (NBIC) and Institute for Life Sciences, University of Southampton, Southampton, United Kingdom

6:School of Pharmacy, Faculty of Health and Life Sciences, De Montfort University, Leicester, LE1 9BH, UK

\*Corresponding author(s): Sumeet Mahajan, [s.mahajan@soton.ac.uk](mailto:s.mahajan@soton.ac.uk); Jeremy S Webb, [j.s.webb@soton.ac.uk](mailto:j.s.webb@soton.ac.uk)

**Table of Contents.**

**Figure S1.** Workflows for multivariate analytical methods

**Figure S2.** Class mean raw spectra for bacteria, taken at 532nm, 633nm, and 785nm.

**Figure S3.** Effects of pre-processing steps on spectra prior to classification.

**Figure S4.** Raman spectra and UV-Visible absorption spectra of porphyrins and carotenoids.

**Figure S5.** Images of *S. aureus* strains.

**Figure S6.** Table of peak assignments for main Raman peaks.

**Figure S7.** PCA of bacterial spectra after polynomial background subtraction.

**Figure S8.** Comparison of classification accuracies between the luminescent background and the Raman spectra.

**Figure S9.** Classification accuracies for PCA-SVM of pure bacterial pellets.

**Figure S10.** Preparation of mixed samples of bacteria and ASM.

**Figure S11.** Raman spectra of bacteria in ASM.

**Figure S12.** Classification accuracies for SVM of pure bacterial pellets.

**Figure S13.** Classification accuracies for PCA-SVM of bacteria in ASM.

**Figure S1:** Workflows for PCA, SVM, and PCA-SVM in iRootLab

**Step-by-step multivariate statistical methods performed in iRootLab (v0.17.8.22-d)**

**Principal component analysis:**

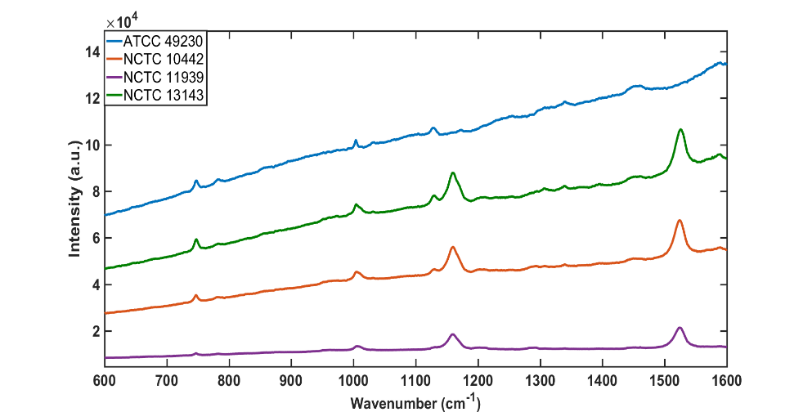
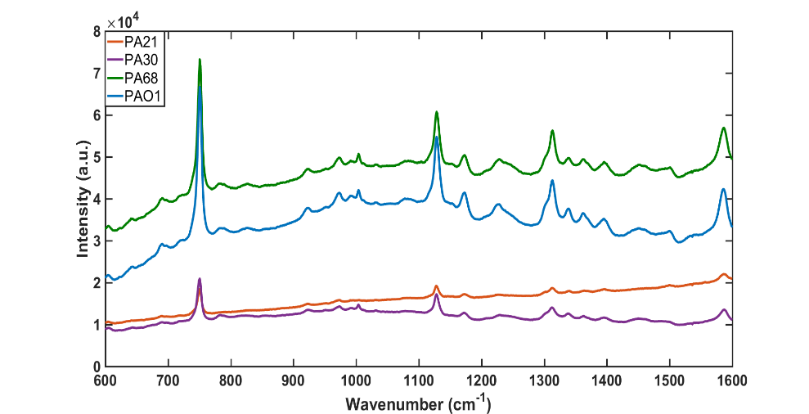
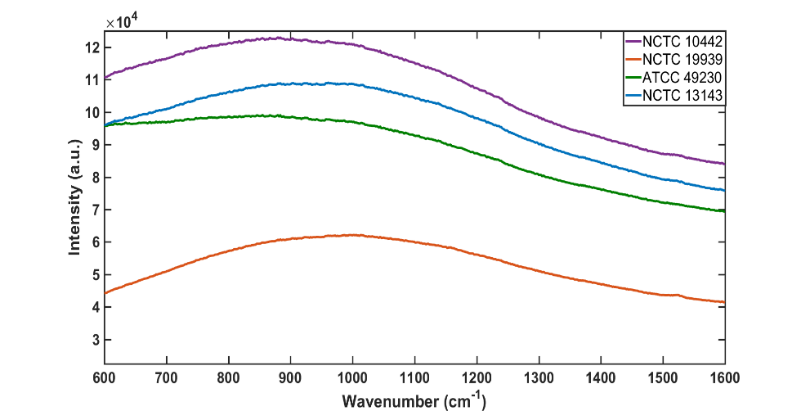
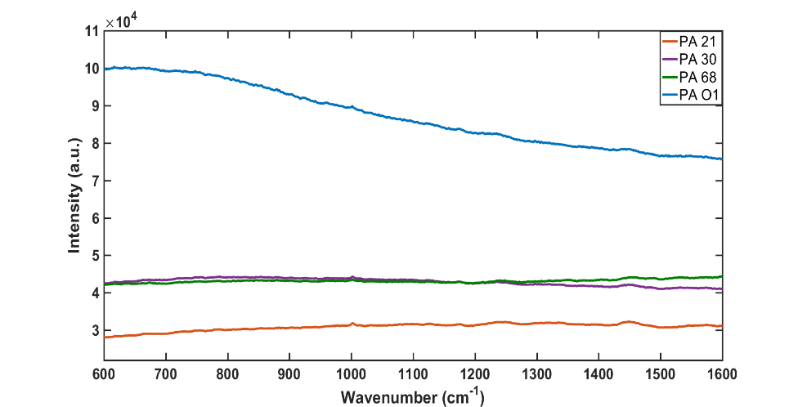
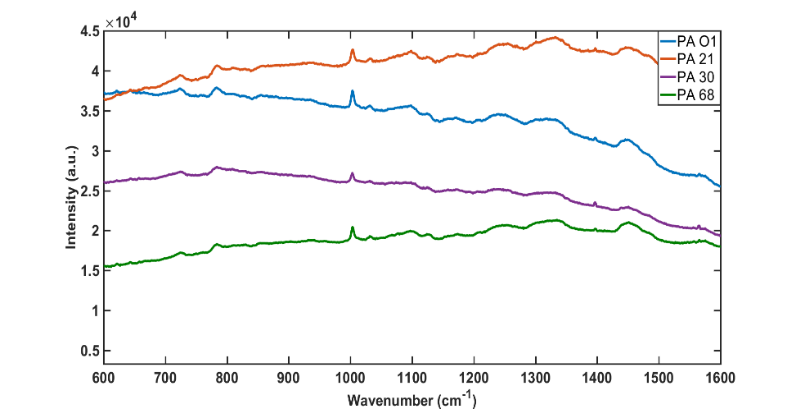
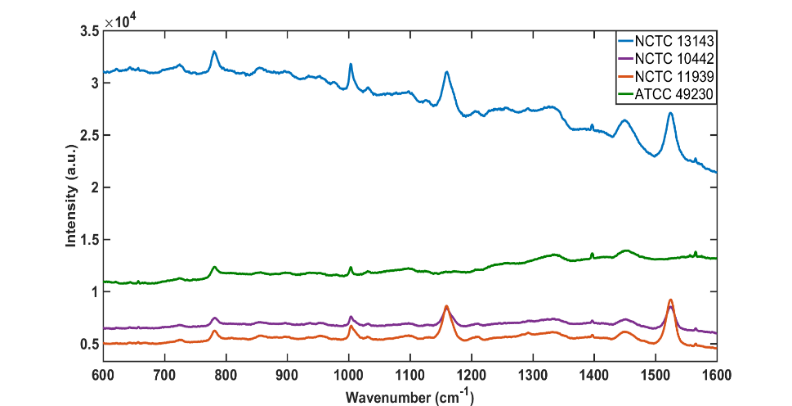
1. Load combined dataset
2. (If desired), select ‘Pre’ and polynomial baseline correction (5th order polynomial, epsilon = 0, no background spectra, index of background [1])
3. Select ‘Pre’ and wavelet de-noising. Use default settings (wavelet name: ‘haar’, 6 decompression levels, thresholds: [0, 0, 0, 100, 1000, 1000])
4. Select ‘Pre’ and normalization. Use default settings (Normalises to maximum intensity)
5. Select ‘Fcon’ and Principal Component Analysis. Use default settings (10 principal components, no rotation of factors)
6. Extract the scores from the ds01\_pca01 variable in the Matlab workspace.
7. Calculate centroids for each strain.
8. Determine Euclidean distance of each spectrum from each centroid, and assign spectral identity based on the shortest distance to a centroid.

**Support Vector Machine:**

1. Load combined dataset
2. (If desired), select ‘Pre’ and polynomial baseline correction (5th order polynomial, epsilon = 0, no background spectra, index of background [1])
3. Select ‘Pre’ and wavelet de-noising. Use default settings (wavelet name: ‘haar’, 6 decompression levels, thresholds: [0, 0, 0, 100, 1000, 1000])
4. Select ‘Pre’ and normalization. Use default settings (Normalises to maximum intensity)
5. Select ‘Sub-dataset Generation Specs’ on the left
6. Create a new K-fold cross-validation (used 10-fold)
7. Select ‘Block’ on the left
8. Create a new support vector machine, using the default parameters (c=1, gamma =1. iRootLab will alter these during the analysis).
9. Return to ‘Dataset’ on the left, and select your pre-processed data.
10. Select ‘AS’ and Grid Search.
11. In the menu that pops up, use the following parameters:
    1. SGS - SGS\_crossval01
    2. Classifier – clssr\_svm01
    3. Templates: SVM
    4. Test post-processor, chooser, and estimation post-processor are left as default settings.
12. Allow the grid search to run, then select ‘Log’ on the left
13. Select log\_gridsearch\_gridsearch01 and extract\_block
14. Return to ‘Dataset’ and import your validation data
15. Select the validation data and select ‘AS’ and then Rater.
16. In the menu that pops up, use the following parameters:
    1. Classifier - Clssr\_svm01
    2. SGS - SGS\_crossval01
    3. As before, leave other parameters as the defaults
17. Allow to run, then select ‘Log’ and estlog\_classxclass\_rater01.
18. Select confusion matrices and hit okay to retrieve the results.

**Principal Component Analysis-Support Vector Machine:**

1. Load combined dataset
2. (If desired), select ‘Pre’ and polynomial baseline correction (5th order polynomial, epsilon = 0, no background spectra, index of background [1])
3. Select ‘Pre’ and wavelet de-noising. Use default settings (wavelet name: ‘haar’, 6 decompression levels, thresholds: [0, 0, 0, 100, 1000, 1000])
4. Select ‘Pre’ and normalization. Use default settings (Normalises to maximum intensity)
5. Select ‘Fcon’ and Principal Component Analysis. Use default settings (10 principal components, no rotation of factors)
6. Select ‘Sub-dataset Generation Specs’ on the left
7. Create a new K-fold cross-validation (used 10-fold)
8. Select ‘Block’ on the left
9. Create a new support vector machine, using the default parameters (c=1, gamma =1. iRootLab will alter these during the analysis).
10. Return to ‘Dataset’ on the left, and select your pre-processed data.
11. Select ‘AS’ and Grid Search.
12. In the menu that pops up, use the following parameters:
    1. SGS - SGS\_crossval01
    2. Classifier – clssr\_svm01
    3. Templates: SVM
    4. Test post-processor, chooser, and estimation post-processor are left as default settings.
13. Allow the grid search to run, then select ‘Log’ on the left
14. Select log\_gridsearch\_gridsearch01 and extract\_block
15. Return to ‘Dataset’ and import your validation data
16. Repeat steps 2-5.
17. Select the PCA of the validation data and select ‘AS’ and then Rater.
18. In the menu that pops up, use the following parameters:
    1. Classifier - Clssr\_svm01
    2. SGS - SGS\_crossval01
    3. As before, leave other parameters as the defaults
19. Allow to run, then select ‘Log’ and estlog\_classxclass\_rater01.
20. Select confusion matrices and hit okay to retrieve the results.



**A**

**B**

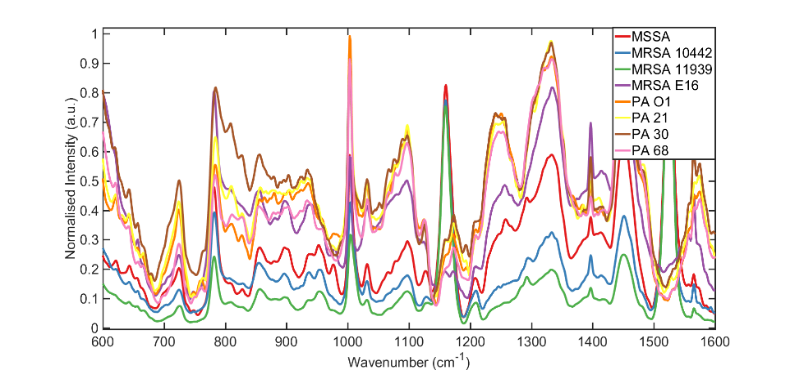
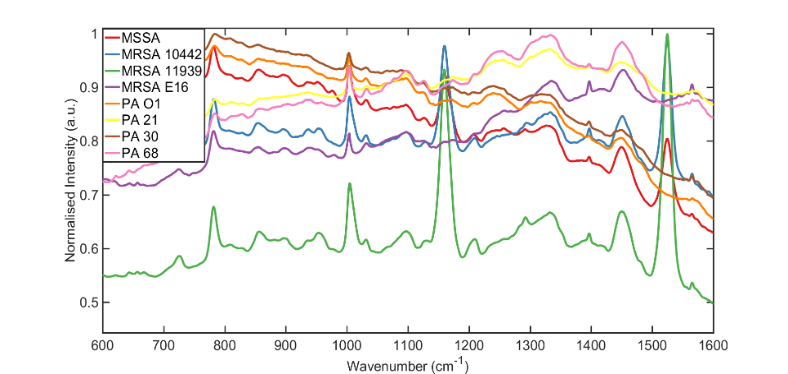
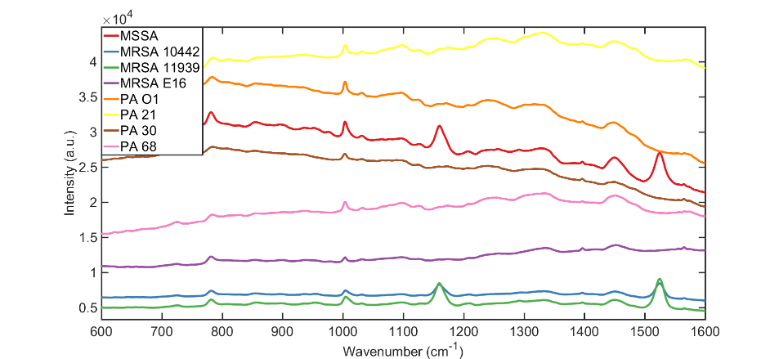
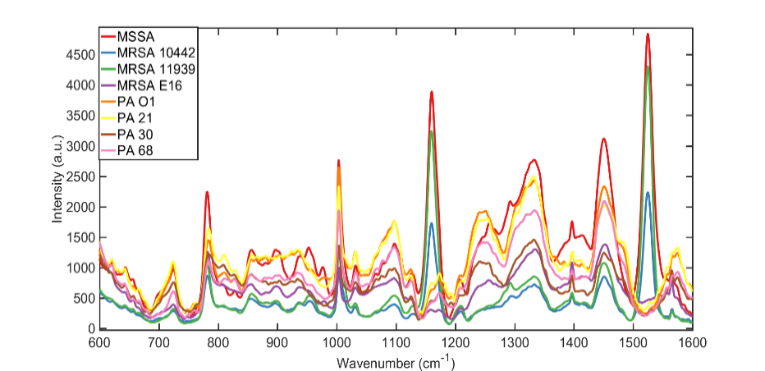
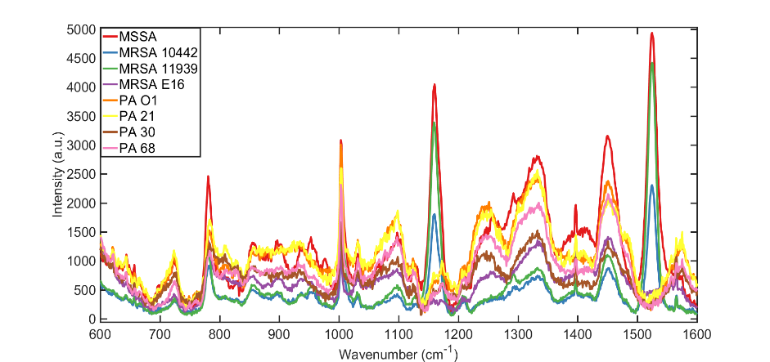
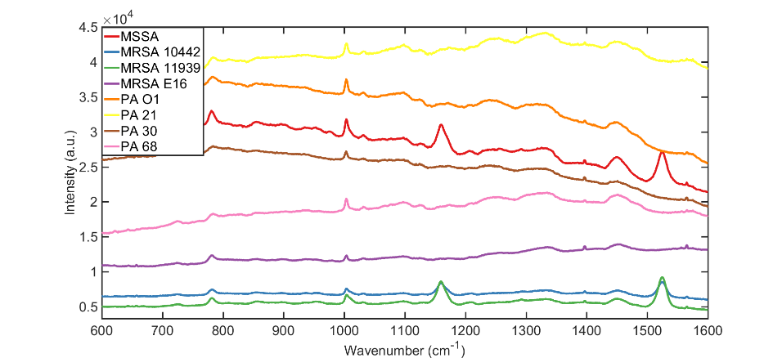
**C**

**D**

**E**

**F**

**Figure S2:** Class means of the unprocessed Raman data for the eight strains of S. aureus and P. aeruginosa used in this study at **(A-B)** 532nm excitation **(C-D)** 633nm excitation and **(E-F)** 785nm excitation.



No background subtraction

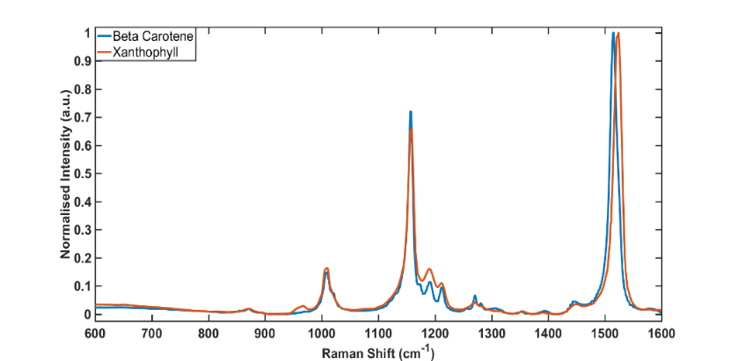
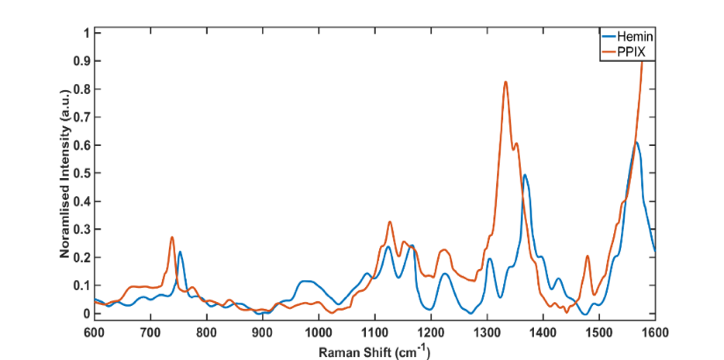
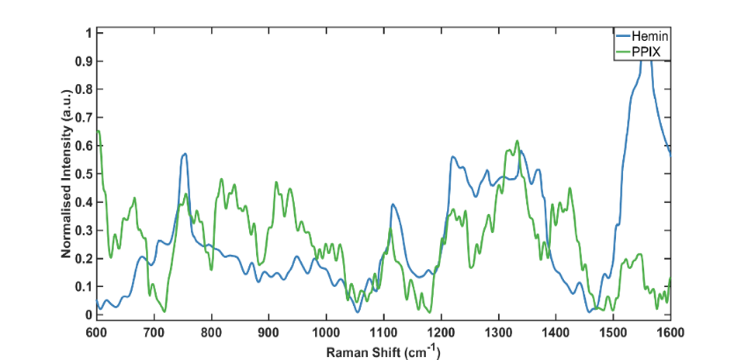
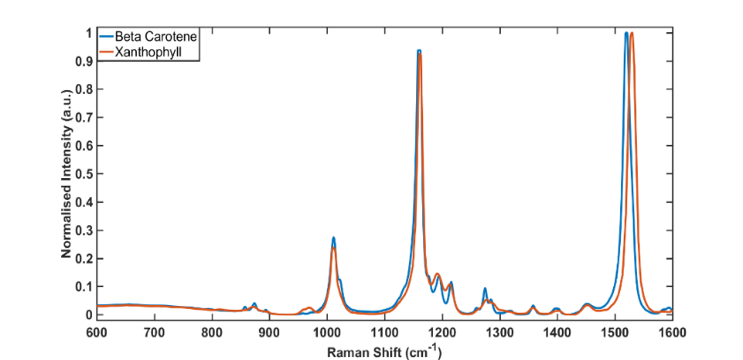
Background subtraction

Denoising

Normalisation

**Figure S3:** Depiction of the changes to the raw Raman spectra used in this study as they undergo each step of pre-processing before being presented to the PCA and SVM algorithms used in this study. Raw spectra are either smoothed using a wavelet denoising algorithm and then normalised to their maximum intensity (left branch), or fitted with a polynomial to remove the broad fluorescent background, and then smoothed via wavelet denoising and then normalised to their maximum intensity (right branch). This normalised data is what is fed into the classifiers.

**Figure S4:** Raman spectra of **(A)** two porphyrin molecules: hemin and protoporphyrin IX, showing peaks around 730-750cm-1 and 1120cm-1, assigned to ring breathing and half-ring modes of the backbone, respectively. **(B)** two carotenoid molecules: Beta-carotene, and xanthophyll, showing characteristic peaks at ~1160cm-1 and 1520cm-1, arising from vibrations of the conjugated backbone of these molecules. Spectra taken at 532 nm excitation. **(C)** two porphyrin molecules: hemin and protoporphyrin IX, showing peaks around 730-750cm-1 and 1120cm-1, assigned to ring breathing and half-ring modes of the backbone, respectively. **(D)** two carotenoid molecules: Beta-carotene, and xanthophyll, showing characteristic peaks at ~1160cm-1 and 1520cm-1, arising from vibrations of the conjugated backbone of these molecules. Spectra taken at 785nm excitation. **(E)** Normalised UV-Visible absorbance spectra of beta-carotene and xanthophyll, with absorbance peaks at 455nm and 484nm, and 425nm, 448nm and 478nm, respectively. **(F)** Normalised UV-Visible absorbance spectra of hematin and protoporphyrin IX, with absorbance peaks at 385nm, and 353nm, respectively.

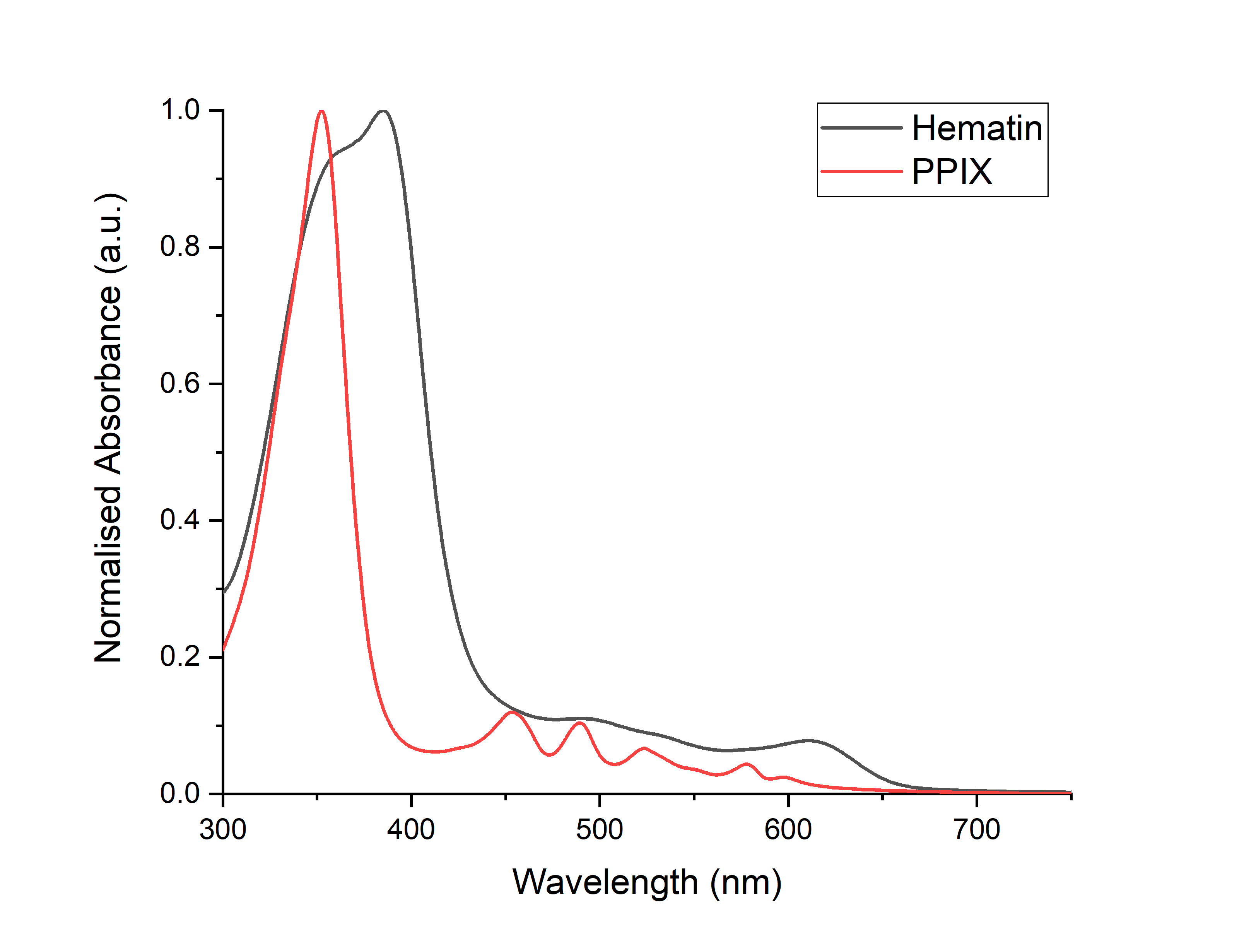
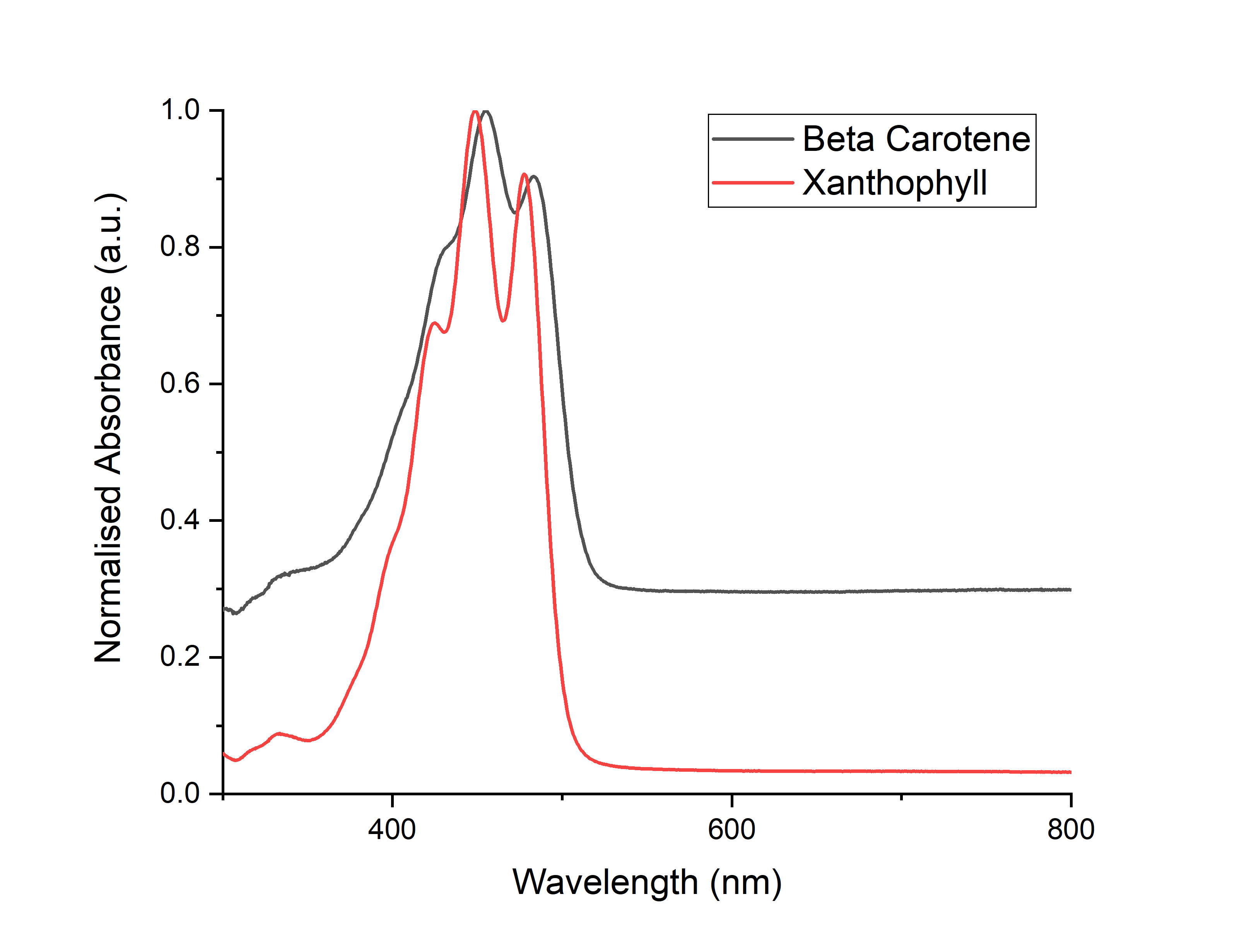


C

D

A

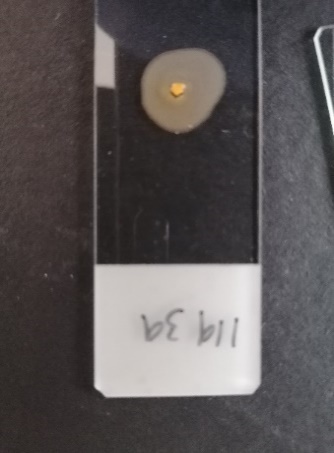
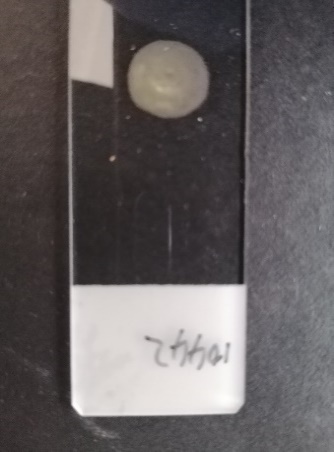
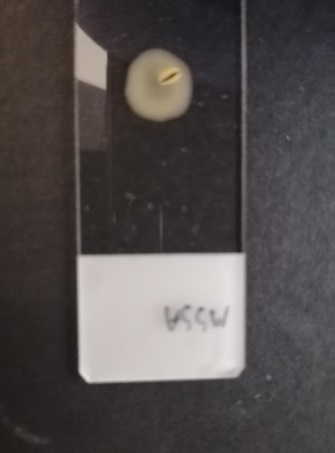
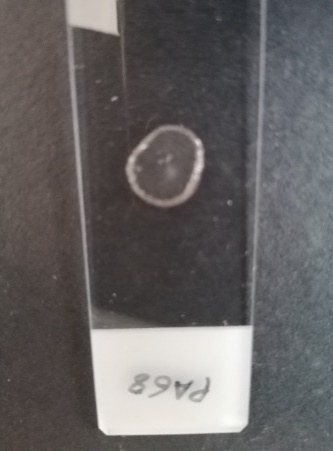
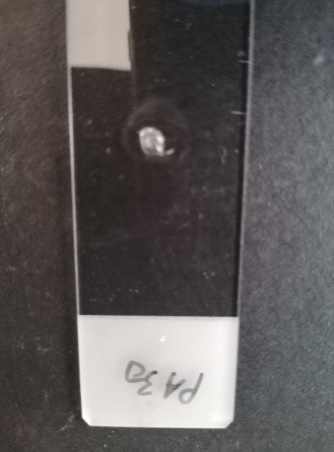
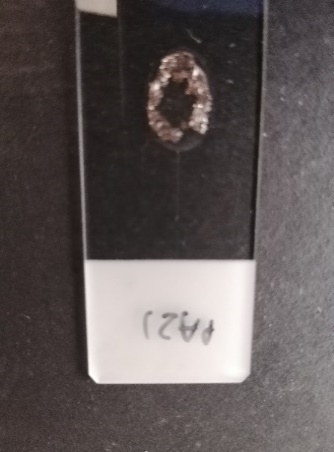
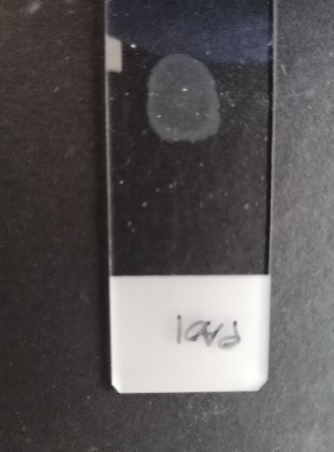
B



E

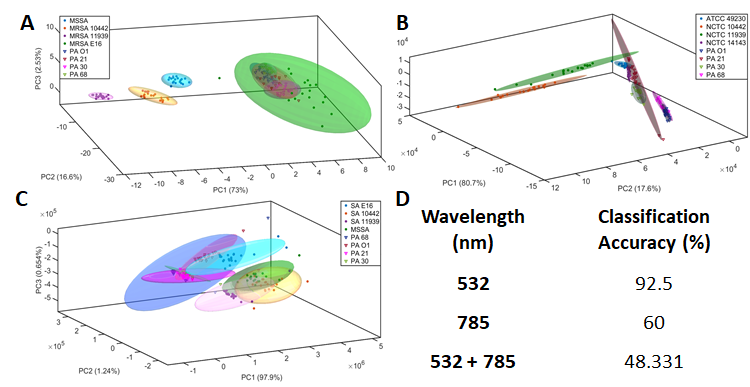
F

**Figure S5:** Images of dried pellets of the 8 strains used in this study. **Top row (Left to right):** ATCC 49230, NCTC 10442, NCTC 11939, and NCTC 13143. **Bottom row (Left to right):** PA O1, PA 21, PA 30, PA 68. Top row shows the yellow pigmentation of the three S. aureus strains with large peaks in locations associated with carotenoid vibrations, and the white colouration of the S. aureus strain without them.



|  |  |
| --- | --- |
| **Raman shift (cm-1)** | **Chemical species** |
| 620 | C-C twisting mode – phenylalanine |
| 640 | C-S stretching & C-C twisting proteins- tyrosine |
| 752 | ν15 breathing mode of porphyrins |
| 780-880 (Broad) | Fused quartz (substrate) |
| 850-855 | Part of the Fermi doublet for tyrosine, tyrosine ring breathing mode |
| 1005 | Ring breathing mode of phenylalanine |
| 1122 | ν22 porphyrin half ring |
| 1168 | ν(C=C) δ(COH) lipids  ν(C-C) carotenoid |
| 1220-1300 | Proteins - Amide III: α-helix, random coil, & β-sheet. |
| 1340-1360 | Tryptophan |
| 1340 | Guanine - DNA |
| 1518 | ν(C=C) in porphyrin  C-C & conjugated C=C band stretch in carotenoids |
| 1582 | δ(C=C), phenylalanine |

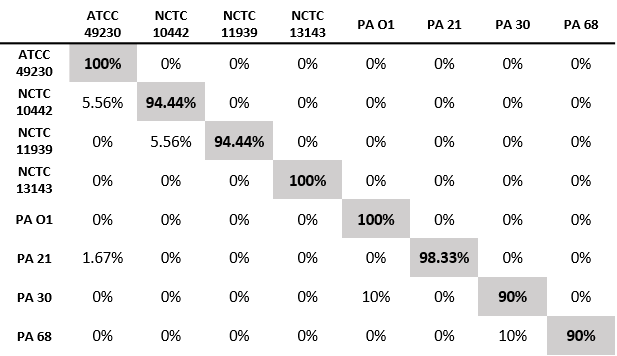
**Figure S6:** Table of peak assignments observed in the spontaneous Raman spectra of both bacterial species under 532nm and 785nm excitation wavelengths.



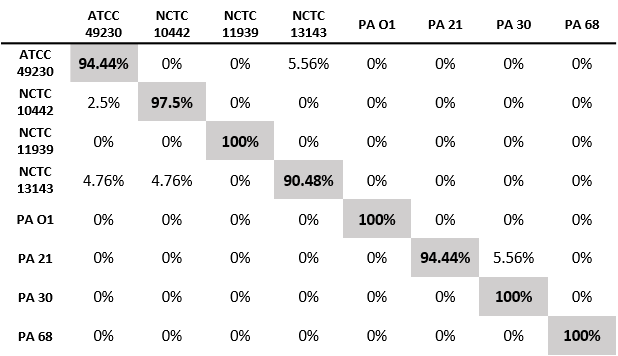
**Figure S7: (A)** Projection of the first three principal components (PCs) for the spontaneous Raman spectra of 4 strains of S. aureus and 4 strains of P. aeruginosa taken with 785nm excitation and with polynomial background subtraction. Good separation of the classes belonging to S. aureus, but poor separation of P. aeruginosa (**B)** Projection of the first three principal components of the spontaneous Raman spectra of 4 strains of S. aureus and 4 strains of P. aeruginosa taken with 532nm excitation and without polynomial background subtraction. **(C)** Combination of the data used for A and B, showing improved separation. **(D)** Table of classification accuracies for projects A-C.

|  |  |  |  |
| --- | --- | --- | --- |
| **Excitation (nm)** | **Accuracy (%)** | | |
| **Raw spectrum** | **Processed spectrum** | **Background only** |
| 532 | 88.75 | 92.50 | 73.00 |
| 785 | 60.00 | 60.00 | 59.38 |

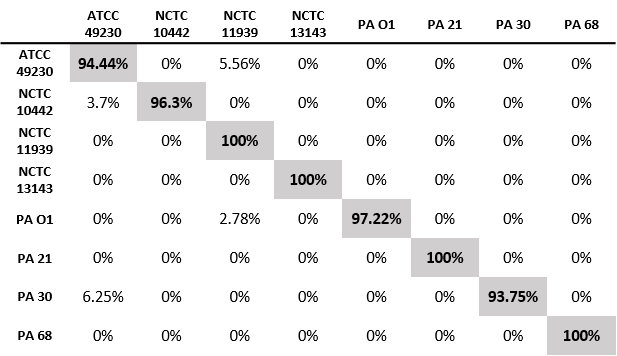
**Figure S8:** Classification accuracies for 532 and 785nm excitation for the eight strains, using PCA and Euclidian distances when processed differently. (1) Raw spectrum: no smoothing and no polynomial background subtraction, (2) Processed spectrum: background subtracted spectra with normalisation, and smoothing (3) Background only: spectra with all Raman peaks removed, leaving just the luminescent background. It can be seen that although classification accuracy is lower when using only the background spectra nevertheless, it does have information to achieve accuracies of 73 and 59% that are comparable with the raw and processed spectra.



A



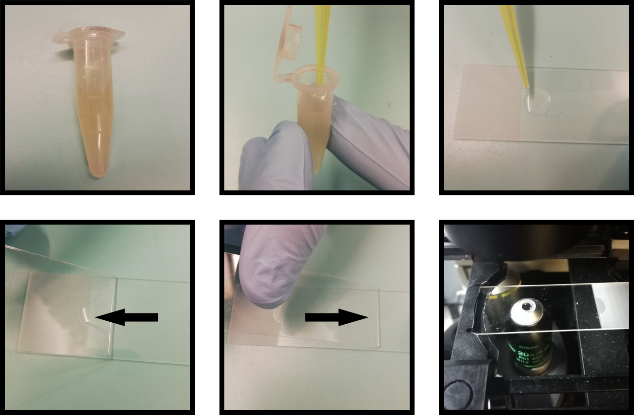
B



C

**Figure S9:** Table of percentage classification accuracies for PCA-SVM of pure bacterial strains for **(A)** 785nm, and **(B)** 532nm excitations, as well as **(C)** concatenated spectra. Correct classifications are shown in the highlighted boxes for each strain, with numbers in the remaining boxes indicating the percentage of samples for that species that were incorrectly identified as another strain. Perfect classification would be represented by scores of 100% across all the highlighted boxes, with scores of 0% in all other boxes.

**Figure S10:** Illustration of sample preparation, as used to make samples of infected sputum ready for Raman spectroscopic analysis. **(A)** Infected sputum is collected from a patient **(B-C)** 10μL aliquot of the sample is taken and deposited onto a clean quartz microscope slide. **(D-E)** A second clean slide (spreader slide) is drawn back towards the droplet at a 30-40 degree angle, ensuring even contact with the lower slide until the droplet attaches to the spreader slide.The spreader slide is pushed back away from the droplet in a smooth motion, drawing the droplet out across the quartz slide in a thin smear. **(F)** The smeared slide is mounted in the Raman spectrometer, ready for spectral acquisitions, as documented in the methods section.



A

B

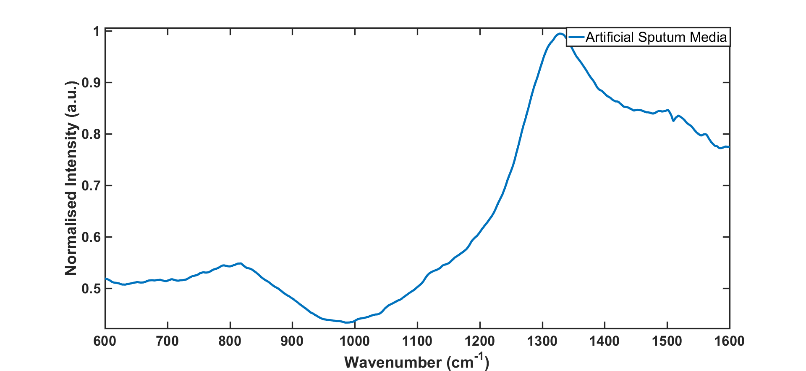
C

D

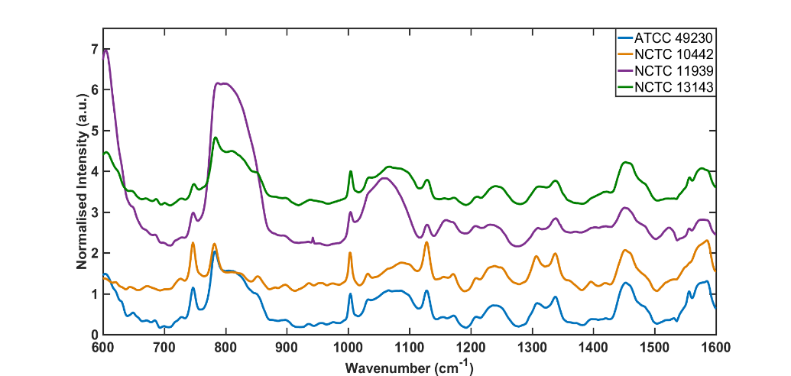
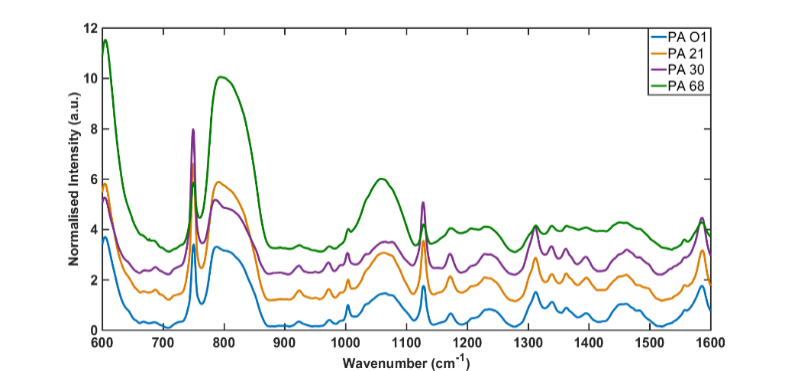
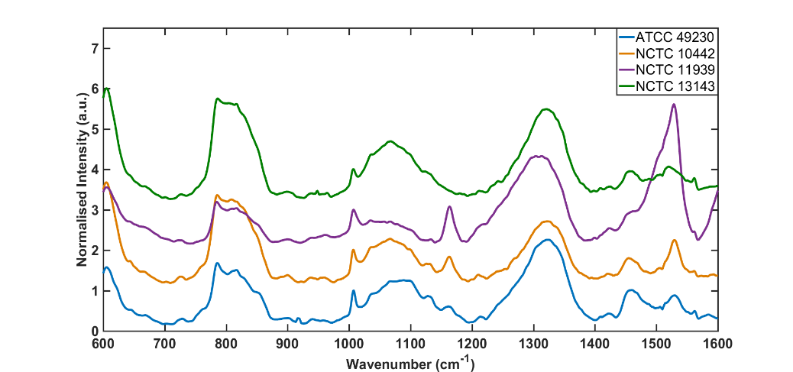
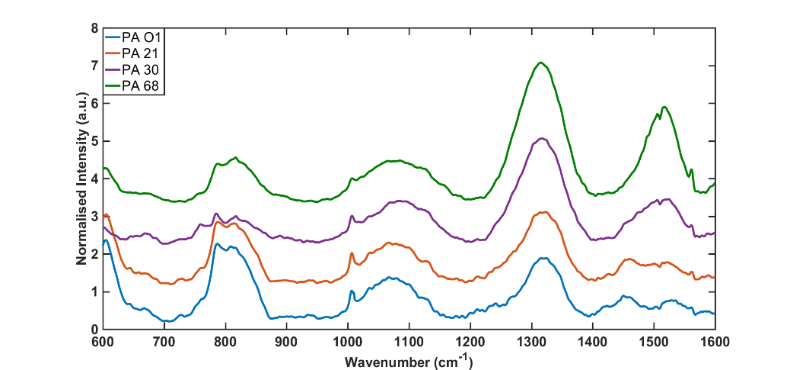
E

F

**Figure S6:** Illustration of sample preparation, as used to make samples of infected sputum ready for Raman spectroscopic analysis. **(A)** Infected sputum is collected from a patient **(B-C)** 10μL aliquot of the sample is taken and deposited onto a clean quartz microscope slide. **(D-E)** A second clean slide (spreader slide) is drawn back towards the droplet at a 30-40 degree angle, ensuring even contact with the lower slide until the droplet attaches to the spreader slide.The spreader slide is pushed back away from the droplet in a smooth motion, drawing the droplet out across the quartz slide in a thin smear. **(F)** The smeared slide is mounted in the Raman spectrometer, ready for spectral acquisitions, as documented in the methods section.



***Figure S11:*** *Class mean spectra of bacterial samples in artificial sputum media at different wavelengths (spectra are offset for clarity). Spectra are normalised to the intensity of phenylalanine at 1004cm-1.* ***(A-B)***S. aureus *at 785nm and 532nm* ***(C-D)***P. aeruginosa *at 785nm and 532nm. Peak assignments can be found in the table in Figure S3.* ***(E-F)*** *Artificial Sputum Media (ASM) blank spectra at 785nm and 532nm.*



**785nm**

**532nm**

**A**

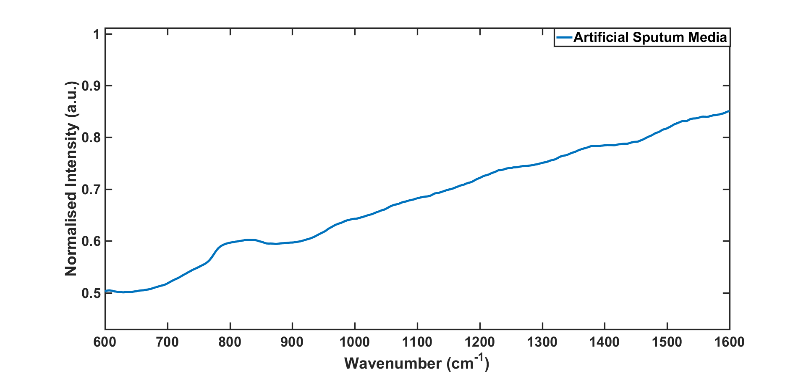
**B**

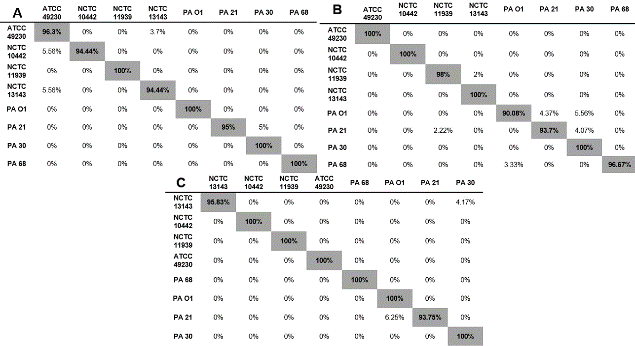
**C**

**D**

**E**

**F**





**Figure S12:** Table of percentage classification accuracies for SVM of pure bacterial strains for **(A)** 785nm, and **(B)** 532nm excitations, as well as **(C)** concatenated spectra. Correct classifications are shown in the highlighted boxes for each strain, with numbers in the remaining boxes indicating the percentage of samples for that species that were incorrectly identified as another strain. Perfect classification would be represented by scores of 100% across all the highlighted boxes, with scores of 0% in all other boxes.

**Figure S13:** Tables of percentage classification accuracies for PCA-SVM of bacterial strains in artificial sputum media for **(A)** 785nm, and **(B)** 532nm excitations, as well as **(C)** concatenated spectra. Correct classifications are shown in the highlighted boxes for each strain, with numbers in the remaining boxes indicating the percentage of samples for that species that were incorrectly identified as another strain. Perfect classification would be represented by scores of 100% across all the highlighted boxes, with scores of 0% in all other boxes.

