Improved Raman data analysis for non-invasive, label-free biomolecular characterisation of single bacterial cells.

Zechuan Gong[1], Andrei Sapelkin[2], Christoph Engl[1,3]

[1]Department of Biochemistry, Centre for Molecular Cell Biology, School of Biological and Behavioural Sciences, Queen Mary University of London, UK.

[2]Department of Physics and Astronomy, School of Physical and Chemical Sciences, Queen Mary University of London, UK.

[3]corresponding author.

Supplementary Information

Supplementary Note: Optimisation of Raman data analysis from single cells

We cultured *E. coli* bacteria under various growth conditions before fixing their physiological state through exposure to formaldehyde prior to Raman microspectroscopy. Formaldehyde treatment had a negligible impact (i.e. within errors) on the Raman spectrum (Fig. S1). We recorded Raman spectra using an upright Renishaw inVia™ confocal Raman microscope with a 50x objective (0.8 numerical aperture), HeNe laser at 633 nm excitation wavelength and 200 seconds exposure time (Fig. S2). The optical resolution (xy: 500 nm; z: 1 μm) enabled us to collect Raman spectra from individual cells. The location of the laser beam centre within the cell body had no effect on the Raman spectrum (Fig. S3 and S4). The cells were suspended on stainless-steel microscope slides to reduce background and enhance the Raman signal (Fig. S5). After subtracting the fluorescence background via multiple-line fitting (Fig. S6), we applied a Savitzky-Golay filter to smoothen the spectra and extract the background noise within the data (Fig. S7). The intensity threshold at which a signal was considered to be significant (Fig. 1f) was set as the upper limit of the 95% confidence interval of the mean standard deviation of the background noise (Fig. S8). We fit each cellular spectrum using reference spectra from a library of purified biomolecules (Fig. S9). Each biomolecule in our library generates a unique Raman signature that aids its identification within the cellular spectrum. For example, RNA[1,2] can be distinguished from DNA (Fig. S10) through an approximate 4 cm^{-1} red-shift and apparent shoulder at 808 cm^{-1}. The location and shape of the peaks within the Raman spectra of nucleotides are similar to previously published results[3]. It is however challenging to distinguish between their different phosphorylation states (Fig. S11). Inside the cell we therefore regard the Raman signal from each nucleotide as the combination of all its phosphorylation states and hence use the collective terms ANP, UNP and GNP, respectively. We performed the fitting via a linear combination of the reference spectra to quanyify their contribution to the overall cell spectrum and fitted the results by the least squares method. The Raman intensity of each biomolecule measured was quantified by determining its fitting parameter p_i using the following equation:

$$cell\ spectrum = p_1x_1 + p_2x_2 + \cdots + p_nx_n$$

x_1 is the normalised Raman spectrum of each purified biomolecule; p_i is the fitting parameter (i.e. the relative contribution) of each reference spectrum. The standard error and confidence interval of each parameter is determined by computing the residuals and parameter covariance matrix.

Kullback-Leibler divergence (D_{KL})[4] was used to evaluate how the linear combination spectrum is different from the actual cell spectrum, with smaller D_{KL} indicating a better result (Fig S12). Within the range of 800-1000 cm^{-1}, 1200-1500 cm^{-1}, and 1600-1750 cm^{-1}, our results show a strong alignment of the fitted spectrum with the actual cellular spectrum in both exponential and stationary growth. In regions below 800 cm^{-1}, around 1000 cm^{-1}, and between 1500-1600 cm^{-1}, the fitted spectrum also displays spectral features consistent with those found in the cell spectrum, with minor variations in amplitudes (Fig 1c,d). The outcomes of the fitting clearly demonstrate that the cellular peaks are comprised of signals from multiple biomolecules, hence assigning Raman peaks to a biomolecule, as is frequently done, is insufficient. Our approach instead can determine the contribution of multiple biomolecules to each peak (Fig. 1e) and therefore enables a more accurate quantification of the cellular content (Fig. 1f). The intensity of the Raman signal from each biomolecule is governed by the Raman activity of its chemical structure and proportional to its cellular content. We therefore assume that only those biomolecules that exhibit both strong Raman activity and high cellular content significantly contribute to the total spectrum of the cell. The results indicate that only around a third of the reference biomolecules in our library met these criteria generating signals above the background noise threshold (Fig. 1f). We found that the relative differences in the abundance of the detected biomolecules between exponential and stationary phase is similar to the data obtained using established disruptive techniques (Table S1).

Supplementary Figures and Table

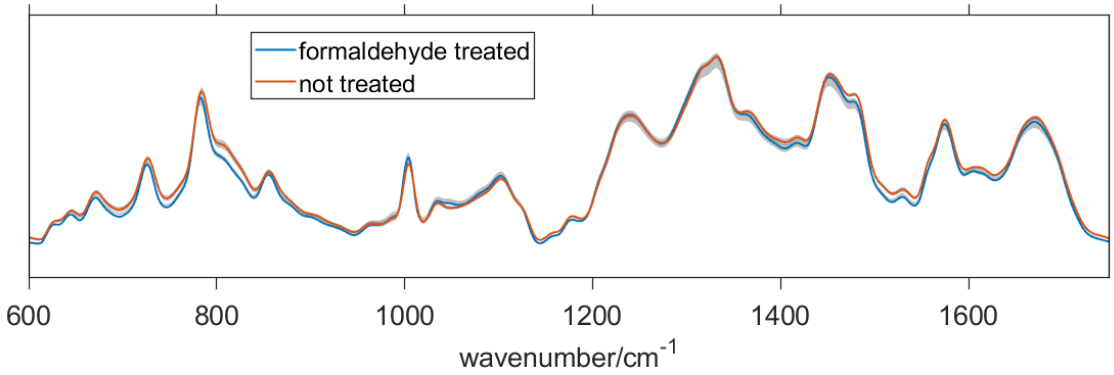

Figure S1. Impact of formaldehyde fixation on the cellular Raman spectrum. The average Raman spectra of formaldehyde treated (n = 342) and untreated (n = 20) *E. coli* cells. The grey areas indicate 0.95 confidential intervals of mean calculated by z-test.

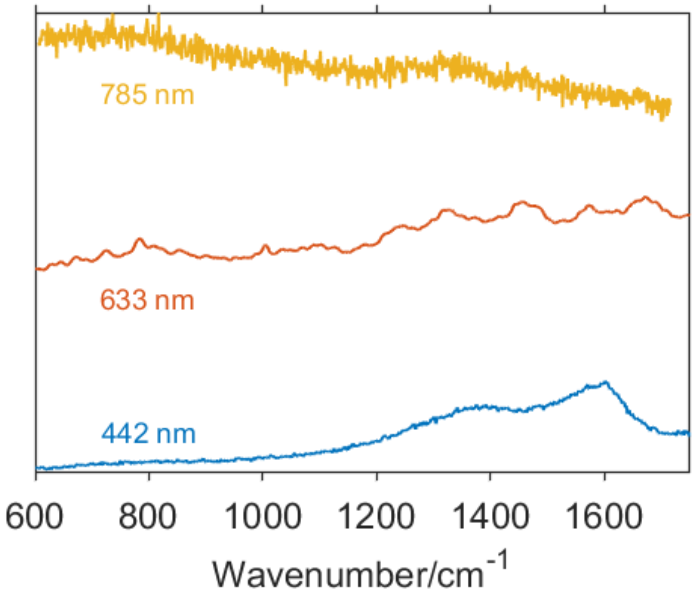

Figure S2. Comparison of normalised Raman spectra using under different excitation wavelengths at 200 seconds exposure time. In our current setup, the strongest Raman signal is obtained at 633 nm. A wavelength of 785 nm is unable to excite sufficient Raman signals from *E. coli* cells, while 442 nm results in apparent signal loss potentially due to cell damage by the higher energy of the photons associated with the shorter wavelength.

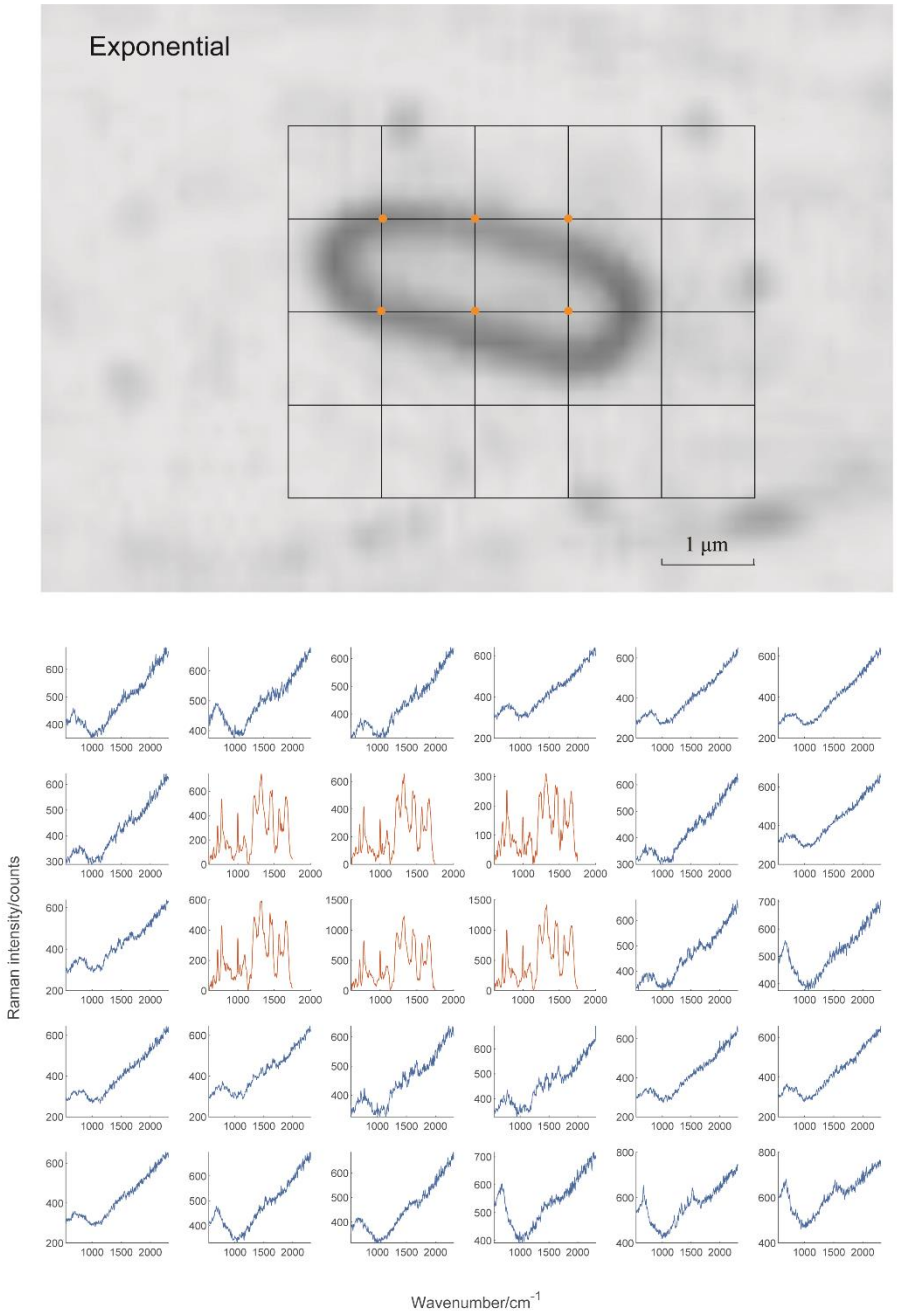

Figure S3. Raman mapping of a single *E. coli* cell in exponential phase. The subplots show the Raman spectra of the square vertices (locations of the laser beam centre), indicated as orange dots on the grid in the microscope image above with the step of 1 µm. The Raman spectrum with bacterial cellular information is highlighted in orange. The intensity of the spectrum increases when the beam is focused onto the cell body and the spectral signal disappears when the beam is focused approximately 1 µm outside of the cell. Importantly, the profile of the Raman spectrum (and hence the biomolecular composition) does not change when the beam is focussed onto different subcellular locations.

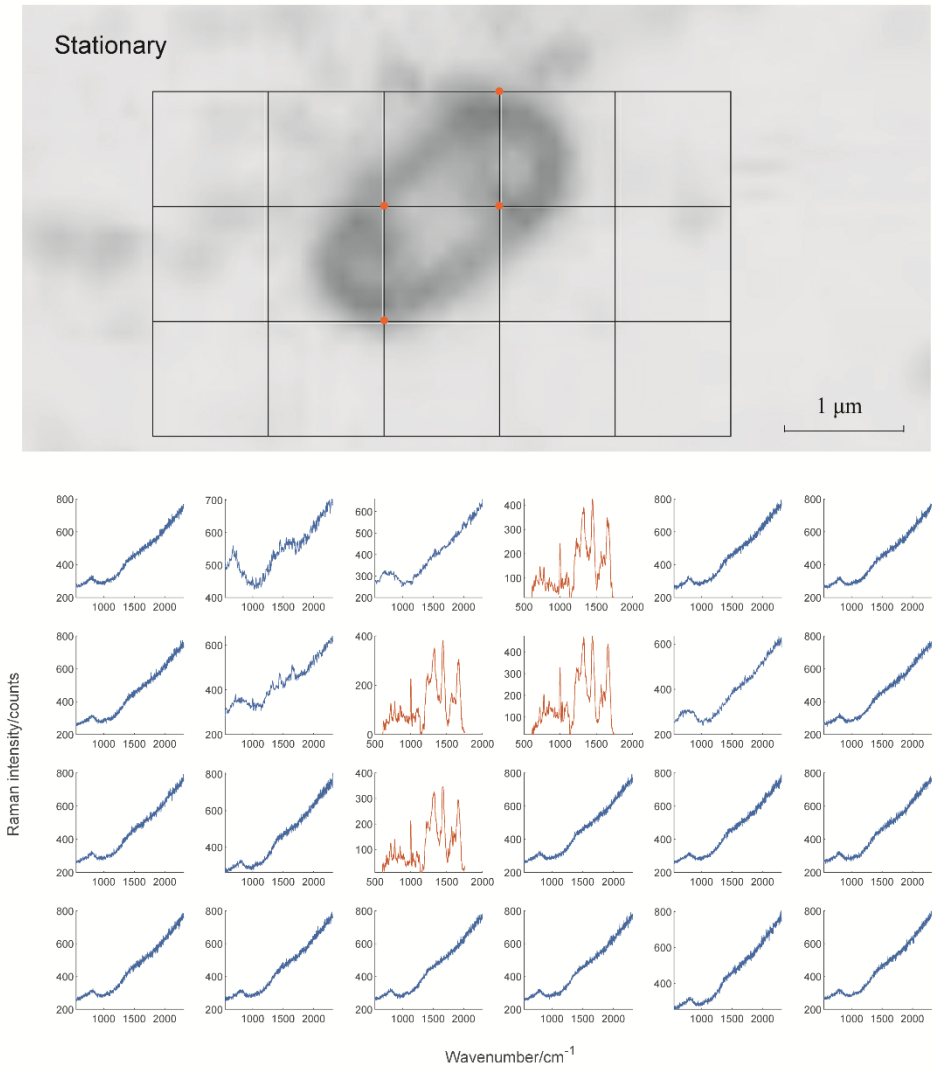

Figure S4. Raman mapping of a single *E. coli* cell in stationary phase. The subplots show the Raman spectra of the square vertices (locations of the laser beam centre), indicated as orange dots on the grid in the microscope image above with the step of 1 µm. The Raman spectrum with bacterial cellular information is highlighted in orange. The data shows that the Raman profile is similar throughout the cell and independent of the subcellular location of the laser beam.

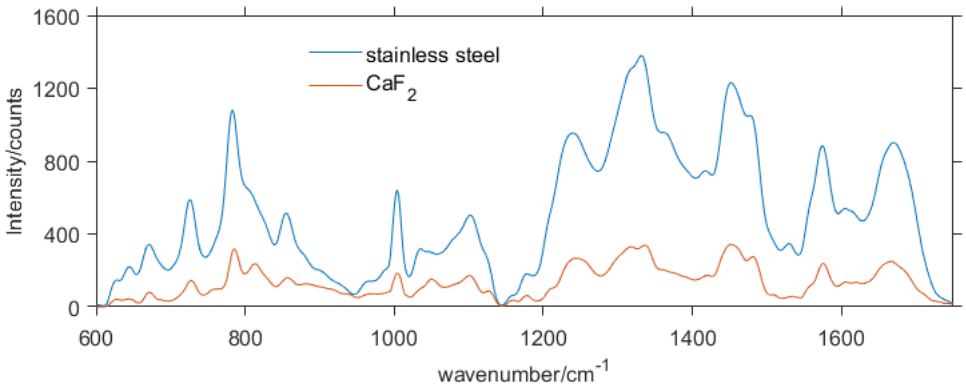

Figure S5. Substrate selection for Raman microspectroscopy. Shown are average Raman spectra of single *E. coli* cells collected on either stainless steel (blue) or calcium fluoride CaF_2 substrate (red). The intensity of the Raman signal is approximately 4-fold higher when cells are mounted on stainless steel substrate compared to CaF_2.

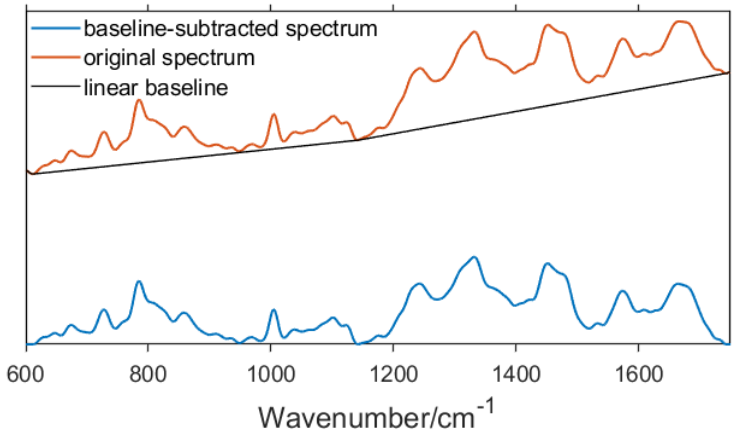

Figure S6. Comparison of the Raman spectrum before and after baseline subtraction. A linear baseline (black line) was subtracted from the original spectrum (orange) using multiple line fitting. The baseline-subtracted spectrum is shown in blue.

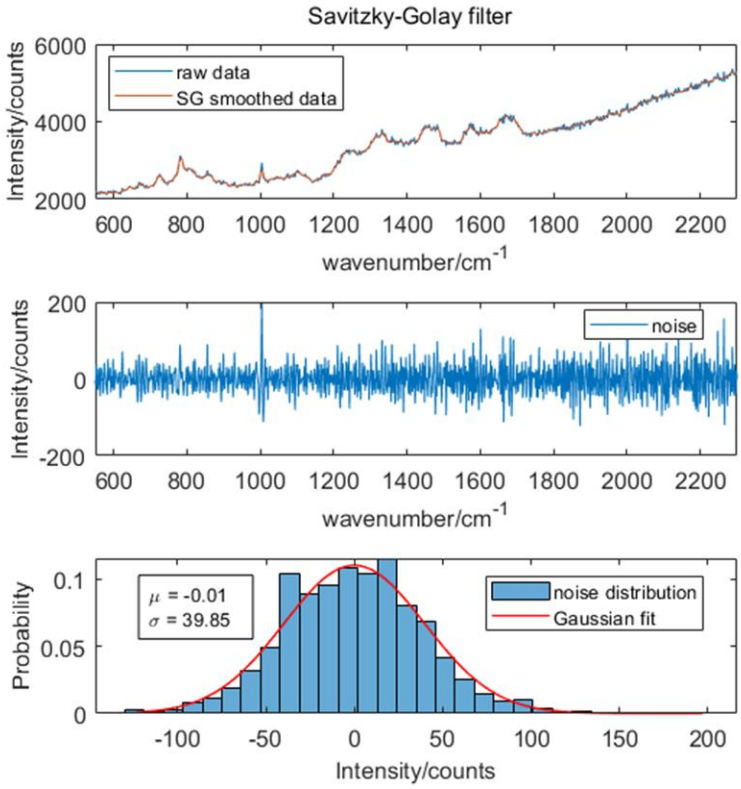

Figure S7. Application of a Savitzky-Golay filter to the Raman spectra. Shown is an example of smoothing the raw Raman spectrum by Savitzky-Golay (SG) filter (window size = 11, polynomial order = 3) and examination of the background noise within the data. The probability histogram of the noise distribution is simulated by a Gaussian function with a mean (μ) of -0.01 and a standard deviation (σ) of 39.85.

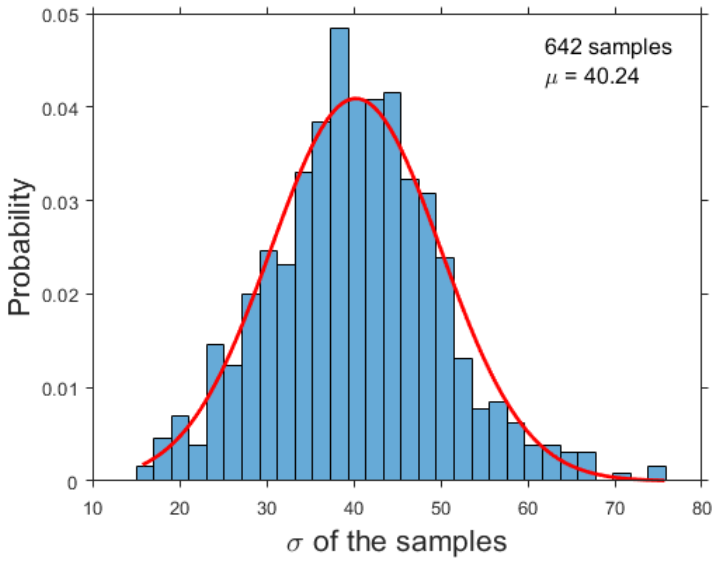

Figure S8. Probability histogram of the σ values of background noise from the SG filter. The data is obtained after analysing the Raman spectra collected from 342 cells in exponential and 300 cells in stationary phase. The probability distribution of σ can be simulated by a Gaussian function with a mean of μ = 40.24, signifying the average σ of the background noise. The 95% confidential interval is calculated by $1.96 * \sigma$, where 1.96 is the 95% quantile in z-test. Consequently, we only considered Raman intensities above a threshold of 78.84 as a significant contribution to the cellular Raman spectrum.

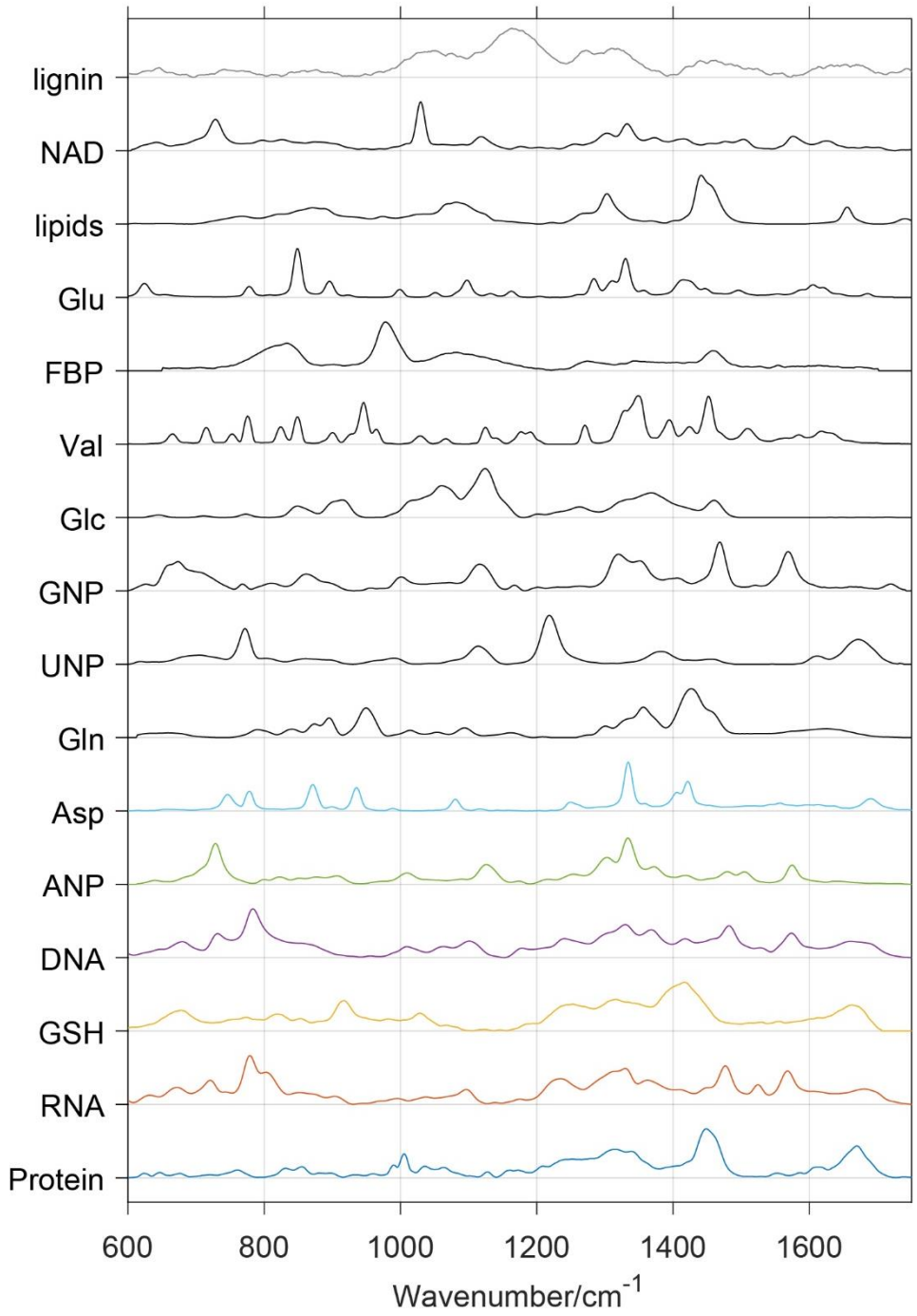

Figure S9. Reference library of Raman spectra from purified biomolecules. The spectra were collected using the same Raman microspectroscopy set-up as described for cells. The biomolecules were selected based on their absolute concentrations in *E. coli* cells[18]. Lignin is a polymer of plant cell walls[29] and served as a negative control. DNA, RNA and protein were purified from *E. coli* cells using the AllPrep Bacterial DNA/RNA/Protein Kit from Qiagen (cat no. 47054). The remaining biomolecules were ordered from commercial suppliers (e.g. Sigma-Aldrich).

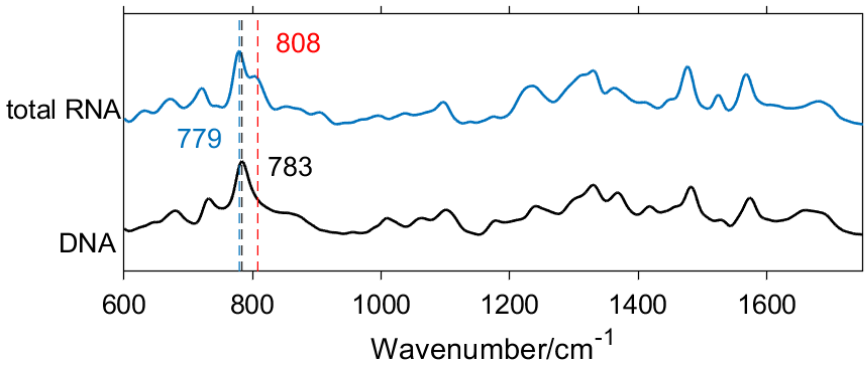

Figure S10. Raman spectra of total DNA and RNA extracted from *E. coli* cells. We detected an overall red-shift of approximately 4 cm^{-1} between the RNA and DNA spectra. This may be a consequence of the formation of hydrogen bonds between double strands in DNA that decreases vibrational energy of the X-H stretching frequency.[1] An apparent shoulder at 808 cm^{-1} in RNA further aids its distinction from DNA[2].

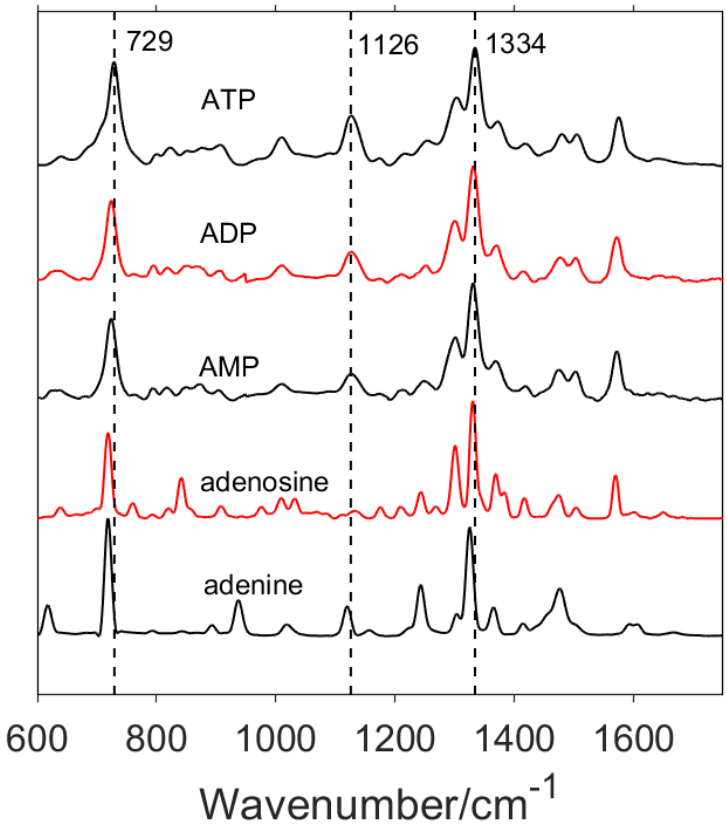

Figure S11. Comparison of the Raman spectra of adenine, adenosine, AMP, ADP, ATP. The two predominant peaks at around 729 and 1334 cm^{-1} are evident in all derivatives. The key difference between ATP, ADP and AMP is the size of the peak at 1126 cm^{-1}.

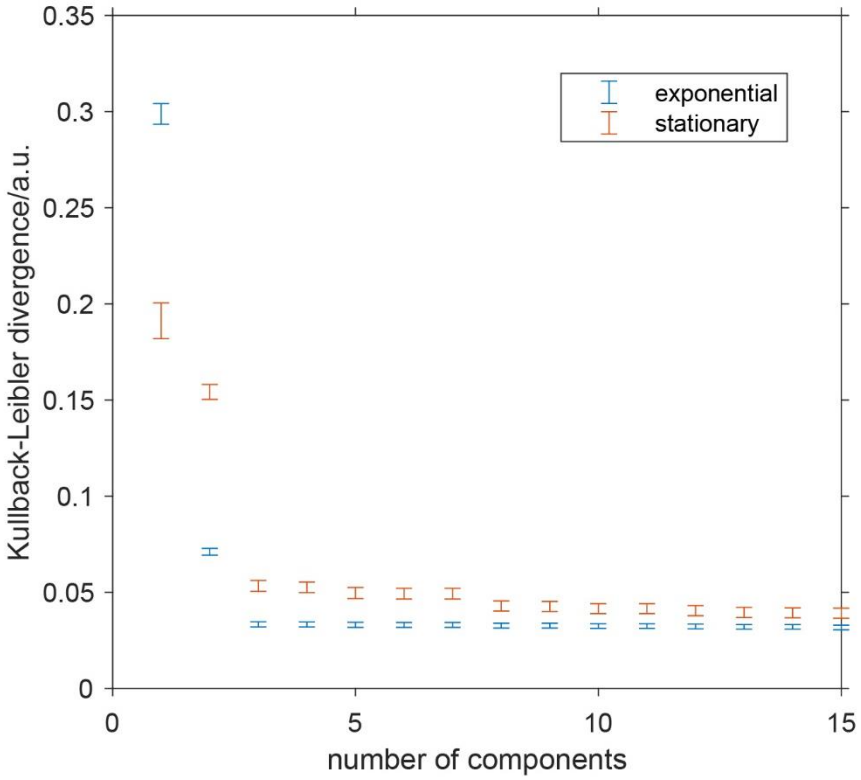

Figure S12. Kullback-Leibler divergence plot. The plot shows how the difference between the cellular and the fitted Raman spectra decreases as components (i.e. spectra from the reference library) are added during fitting. The components are ordered by the magnitude of their fitting parameters from maximum to minimum (protein, RNA, GSH, DNA, ANP, Asp, Gln, UNP, GNP, Glc, Val, FBP, Glu, lipid, NAD+).

	Exponential phase established techniques	Stationary phase established techniques	Ratio established techniques	Ratio Raman Microspectroscopy
DNA	6.59 µg/10^8CFU	2.60 µg/10^8CFU	39.5%	21.22 ± 2.65%
RNA	26.5 ± 4.1 mg/mL	18.3 ± 1.5 mg/mL	69.1 ± 7.1%	70.08 ± 6.04%
Protein	65.2 mg/mL	97.8 mg/mL	150%	174 ± 18.38%
Glutathione	36.7 pmol/mL	82.3 pmol/mL	224.3%	264.3 ± 33.21%

Table S1. Relative differences in total DNA, RNA, protein, and glutathione abundance of *E. coli* cells in exponential and stationary phase. The abundance of these biomolecules is dependent on the growth phase. The ratio between exponential and stationary phase measured by Raman microspectroscopy is similar to previous reports[5,6] using established disruptive techniques. Adenine nucleotides are not included in this table due to the fluctuation in ATP concentrations during centrifugation[7].

Supplementary References

1. Falamas, A., Kalra, S., Chis, V., Notingher, I. Monitoring the RNA distribution in human embryonic stem cells using Raman micro-spectroscopy and fluorescence imaging. *AIP Conf Proc* **1565**, 43-47 (2013).

2. Wright, A. M., Howard, A. A., Howard, C. et al. Charge Transfer and Blue Shifting of Vibrational Frequencies in a Hydrogen Bond Acceptor. *J Phys Chem A* **117**, 5435-5446 (2013).

3. Ostovarpour, S., Blanch, E. W. Phosphorylation Detection and Characterization in Ribonucleotides Using Raman and Raman Optical Activity (ROA) Spectroscopies. *Applied Spectroscopy* **66**, 289-293 (2012).

4. Chen, X., Shen, J., Liu, C. et al. Applications of Data Characteristic AI-Assisted Raman Spectroscopy in Pathological Classification. *Anal Chem* **96**, 6158–6169 (2024).

5. Zimmerman, S. B., Trach, S. O. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of *Escherichia coli*. *J Mol Biol* **222**, 599-620 (1991).

6. Loewen, P. C. Levels of glutathione in *Escherichia coli*. *Can J Biochem* **57**, 107-111 (1979).

7. Lundin, A. & Thore, A. Comparison of methods for extraction of bacterial adenine nucleotides determined by firefly assay. *Appl Microbiol* **30**, 713-721 (1975).