Quantitative Analysis of Phenotypic Heterogeneity in Escherichia coli Using Raman Spectroscopy

April 25, 2025

Elif Solmaz

School of Physics and Astronomy, Queen Mary
University of London, Mile End Road, London,
E1 4NS, United Kingdom

Superviser: Andrei Sapelkin

An Extended Independent Project presented for the degree of Bachelor of Science in Physics and Data Science

Declaration

I hereby certify that this project report, which is approximately 10,000 words in length, has been written by me at the School of Physics and Astronomy, Queen Mary University of London. All material in this dissertation which is not my own work has been properly acknowledged, and it has not been submitted in any previous application for a degree.

Elif Solmaz (220467249)

Abstract

The physiology and adaptability of microorganisms can be significantly affected by phenotypic variability among genetically identical bacterial cells. Single-cell chemical analysis is made possible by Raman spectroscopy, a label-free, non-destructive technique with excellent spatial and spectral resolution that removes the requirement for complex preparation. In this study, 633 nm Raman excitation was used to investigate phenotypic variation in Escherichia coli. To break down complicated Raman spectra into contributions from important biomolecules including DNA, RNA, proteins, and metabolites, a linear combination model was created. Each component's informative importance was extracted by fitting to the experimental data followed by calculation of Bayesian Factor Integral (BFI). The results obtained give an understanding into stochastic gene expression by showing different molecular profiles between stages of development and revealing correlations in biomolecular distributions.

Contents

1 Introduction

1.1 Historical Background

Raman spectroscopy has a fascinating history that dates back to 1928 when Sir Chandrasekhara Venkata Raman discovered the Raman effect while studying the scattering of light in liquids. This discovery revolutionised our understanding of molecular vibrations and earned him the Nobel Prize in Physics in 1930. Around the same time, Landsberg and Mandelstam observed similar phenomena in solids, further solidifying Raman spectroscopy's importance in quantum theory and chemical analysis.

1.2 Motivation and Research Goals

Bacterial populations often consist of genetically identical cells, they do not always behave the same. Certain cells may develop more quickly, others may become more resistant to drugs, and yet others may change into stress-response modes. These kinds of differences, often referred to as phenotypic variation, are important in understanding how bacteria adapt and survive in changing environments (1).

What is challenging is that traditional microbiology methods usually look at whole populations, averaging the differences between individual cells. That is why there is growing interest in single-cell techniques, which can reveal what is going on at the level of each bacterium.

In this project, we used Raman spectroscopy, a nondestructive method, to study individual E. coli cells and examine how molecular differences show up at different stages of bacterial growth. The main goals of the project were:

- To identify and quantify how much different biomolecules (like DNA, RNA, proteins, and small metabolites) contribute to the Raman spectra of E. coli.
- To compare how these molecular profiles differ between cells in the exponential (actively growing) and stationary (growth-arrested) phases.
- To apply a Bayesian model to evaluate how informative each component is in describing the observed spectra.
- To explore whether patterns in data can reveal signs of random fluctuations, or 'noise', in gene expression at the single cell level.

1.3 Overview of Phenotypic Heterogeneity

When cells that are otherwise genetically similar and inhabit the same environment change in their behaviour or metabolic states, this is referred to as phenotypic heterogeneity. These variations may result from random chemical processes that occur during transcription and translation, or they may be brought on by outside variables such unequal exposure to oxygen, nutrients, or stress signals. (8) (9)

Although this type of variety might seem to be problematic, new study (10) indicates that it could be helpful for survival. A diversified population might, for instance, "hedge its bets" in an unpredictable environment; some cells might survive stress or treatment with antibiotics whereas others may not.

These differences have been studied using a number of techniques, including fluorescence microscopy and flow cytometry. However, the majority of these instruments require fluorescent tags or genetic modifications, which may change the cell's fundamental state. That's why label-free techniques like Raman spectroscopy are becoming more attractive—they allow us to study cells as they are, with minimal interference.

1.4 Role of Raman Spectroscopy in Microbial Studies

The resulting spectrum acts like a "fingerprint" of the cell's chemical composition. Different molecules—like proteins, lipids, nucleic acids, and metabolites—have unique spectral features, which allows scientists to determine what's present in the cell and in what quantities.

In microbiology, Raman spectroscopy has been used to monitor metabolic activity, identify different bacterial species, and study responses to antibiotics. When used at the single-cell level, it becomes a useful technique for understanding biochemical differences between individual cells within the same population.

This study uses Raman spectroscopy specifically to look at E. coli cells in two different growth states and applies computational techniques to analyse how biomolecular compositions vary between them. By breaking down these spectra, we aim to better understand how randomness in gene expression contributes to the phenotypic differences observed in bacteria.

2 Theoretical Background

2.1 Principles of Raman Scattering

Raman spectroscopy is a fascinating technique that scientists use to uncover the chemical makeup and structure of all sorts of materials, from mysterious powders to living cells. The process starts simply enough: a laser beam is directed at the sample. Most of the light from the laser bounces straight back at the same energy it arrived with—this is called Rayleigh scattering. But every so often, something more interesting happens. The light interacts with the molecules in the sample, causing them to vibrate. In these rare cases, the scattered light either gains or loses a tiny bit of energy. This phenomenon is what we call the Raman effect.

What makes this effect so useful is that the change in energy—known as the Raman shift—carries a wealth of information. Scientists measure this shift in wavenumbers (cm^{-1}), and each value corresponds to a specific way that the molecules are vibrating. In essence, every type of molecule has its own set of vibrational "notes," much like a fingerprint or a musical signature. By analysing the pattern of these shifts, researchers can tell exactly what molecules are present in a sample.

However, detecting these signals isn't always easy. The changes in a molecule's polarizability during vibration are often quite small, which means the Raman signal itself can be faint. Sensitive instruments are needed to pick up these subtle clues. Despite the technical challenges, the payoff is huge: the resulting Raman spectrum is unique to the sample being studied, providing a definitive molecular fingerprint.

Because of its specificity and non-destructive nature, Raman spectroscopy has found a home in many scientific fields. Chemists use it to identify unknown substances, biologists rely on it to probe the composition of cells and tissues, and materials scientists turn to it for insights into everything from polymers to pharmaceuticals. Its versatility and precision make it an indispensable tool for anyone who wants to peer into the molecular world(11).

2.2 Raman-Active Modes and Selection Rules

Not every vibrational mode produces a Raman signal. For a mode to be "Raman-active," there needs to be a change in the polarizability of the molecule during the vibration. This is different from infrared (IR) spectroscopy, which relies on changes in dipole moments.

Symmetric stretching modes are often strong in Raman, while asymmetric modes are more prominent in IR. Because of this, Raman and IR are often seen as complementary methods in molecular analysis.

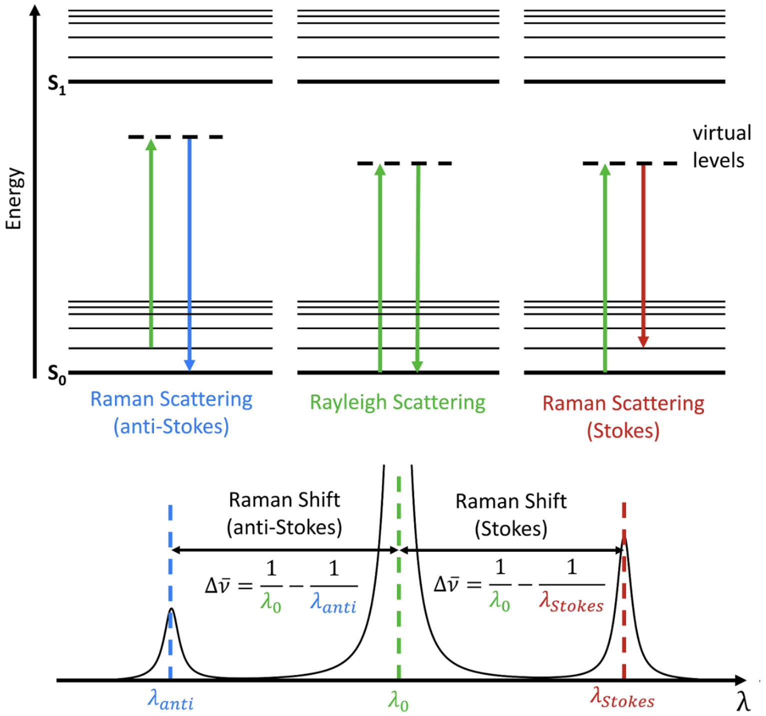
Energy
S1
S0
virtual levels
Raman Scattering (anti-Stokes)
Rayleigh Scattering
Raman Scattering (Stokes)
Raman Shift (anti-Stokes)
Raman Shift (Stokes)
$\Delta\bar{\nu} = \frac{1}{\lambda_0} - \frac{1}{\lambda_{anti}}$
$\Delta\bar{\nu} = \frac{1}{\lambda_0} - \frac{1}{\lambda_{Stokes}}$
λ_{anti}
λ_0
λ_{Stokes}
λ

Figure 1: Jablonski diagram illustrating Rayleigh, Stokes, and Anti-Stokes Raman scattering processes. Adapted from Edinburgh Instruments, 2025(11).

In biological samples, things get more complicated. Cells contain a mixture of proteins, nucleic acids, lipids, and small molecules—all of which have overlapping Raman bands. For example:

- The DNA backbone typically shows peaks around 780–810 cm^{-1}
- The DNA backbone typically shows peaks around 780–810 cm^{-1}
- Proteins often produce a strong Amide I band around 1650 cm^{-1}
- Phenylalanine, an amino acid, gives a sharp peak at 1003 cm^{-1}

These features help identify what's in the sample, but because they can overlap, interpreting biological spectra is rarely straightforward.

2.3 Spectral Interpretation in Biological Systems

Interpreting the data from cellular Raman spectroscopy is no simple task. Unlike the clean, isolated spectra you might get from a pure chemical, the Raman spectrum of a living cell is a tangled web of signals. This complexity arises because a cell is packed with a wide variety of molecules—proteins, nucleic acids, lipids, carbohydrates, and more—all contributing their own vibrational signatures. As a result, the spectrum you observe is a composite, with many peaks overlapping and blending together, making it difficult to assign specific features to individual molecules(16; 17).

On top of that, the microenvironment inside the cell introduces an additional dimension of challenge. The local surroundings—such as pH, hydration, and the folding or conformational state of proteins—can cause subtle shifts in both the position and intensity of Raman bands. For example, the vibrational bands of certain amino acids within proteins are known to shift depending on whether they are in a hydrophobic or hydrophilic environment, or if the protein changes its structure in response to pH or temperature(18). Even water within the cell can produce broad, overlapping bands that further complicate the spectrum(19).

Because of all these factors, the resulting Raman spectrum is not just a straightforward map of what's present in the cell. Instead, it reflects a dynamic interaction between molecular composition, structure, and environment. This makes Raman spectroscopy a powerful but sophisticated tool for studying living systems—capable of revealing rich information, but requiring careful analysis and interpretation to untangle the complexity.

2.4 Understanding Biological Systems' Spectra

Unlike simple chemical mixtures, cells are messy. Their spectra contain broad, overlapping bands, and the exact shape and intensity of those bands can change depending on the cell's environment.

The way a given molecule vibrates and scatters light may be affected by a number of variables, including pH, temperature, hydration levels, and molecular shape. This implies that, depending on the circumstances, two cells with the same molecules may nonetheless create somewhat distinct spectra.

To deal with this complexity, it's standard practice to preprocess Raman data before analysis. This often involves:

- **Baseline correction**, to remove background noise or fluorescence
- **Denoising**, using techniques like Savitzky-Golay filtering
- **Normalization**, so that intensity differences are meaningful and not just due to measurement conditions

Once the spectra are cleaned up, they can be analysed more reliably, especially if you're comparing large numbers of cells.

2.5 Noise in Gene Expression and Why It Matters

Gene expression in cells isn't perfectly predictable. Even if two bacteria have the same DNA and live in the same environment, the way they transcribe and translate genes can vary due to random molecular events. This randomness, or "noise," comes in two main forms:

- **Intrinsic noise**, which refers to random fluctuations in processes like transcription or translation within the same cell.
- **Extrinsic noise**, to comprise changes in the environment or the cellular machinery

Mathematical models are commonly used by academics to explain this uncertainty. One popular paradigm is the three-stage gene expression model, which includes gene activation, mRNA transcription, and protein translation (7).

In this project, one of the goals was to see if Raman spectroscopy could detect signs of this noise indirectly—by looking at the variability in molecules like RNA, DNA, and protein across individual E. coli cells. Finding patterns in this diversity may help us better understand how single-cell gene control functions.

3 Experimental Methods

3.1 Bacterial Strains and Growth Conditions

Two strains of Escherichia coli were used in this investigation: the standard BW25113 strain and a mutant strain , ΔcydB, which carries a deletion in a gene involved in cellular respiration. Both strains were cultured in LB (Luria-Bertani), a nutrient-rich broth widely used to grow bacteria under laboratory conditions.

The cultures were incubated at 37°C with shaking at 220 rpm to ensure adequate ventilation and homogeneity. Samples at two key growth phases were collected:

- **Exponential phase**, when the cells were actively dividing and had reached an OD_{600}(optical density at 600 nm) of approximately 0.5.
- **Stationary phase**, achieved by overnight incubation, where growth slows due to nutrient depletion and accumulation of waste.

By comparing these two physiological states, it was aimed to investigate molecular-level changes associated with bacterial growth dynamics.

3.2 Fixing Cells and Preparing Slides

A 3.7% formaldehyde solution in phosphate-buffered saline (PBS) was used to fix the cells after they had grown. In order to maintain cellular structure and metabolic integrity for Raman analysis, fixation was required. After fixation, the cells are carefully cleaned with PBS to remove any remaining fixative before resuspending them in distilled water was performed.

In order to reduce cell overlap during imaging, the resuspended cells were diluted to an OD_{600} of around 0.05. A little amount of volume was applied on stainless steel microscope slides, and they were allowed to air dry afterwards. Stainless steel's Raman-compatible, low-background surface improved the signal's quality during scanning.

3.3 Raman Instrumentation and Acquisition Configuration

An inVia Raman microscope made by Renishaw was used to obtain Raman spectra. A 633 nm excitation laser was used, offering a good balance between signal strength and minimised photo-induced cell damage.

Each spectrum was collected by averaging 20 successive 10-second scans, ensuring a high signal-to-noise ratio. A 50× objective lens was used to focus the laser on single bacterial cells. To avoid mixed signals, well-isolated individual cells were selected and clumped regions or debris were avoided.

For every sample, the same scanning parameters and equipment settings were used for every measurement.

3.4 Comparison with Infrared Spectroscopy

Raman spectroscopy is often compared with infrared (IR) spectroscopy because both techniques analyse molecular vibrations. However, they complement each other beautifully due to their differences. IR spectroscopy is perfect for examining polar functional groups because it can identify variations in dipole moments. However, Raman spectroscopy is more appropriate for symmetric vibrations that infrared could overlook since it detects variations in polarisability. Together, these techniques provide a more complete picture of molecular structures. The method has been used to track metabolic activities, determine physiological states, and differentiate between different kinds of bacteria (3).

3.5 Component Isolation and Reference Spectrum Acquisition

For spectrum collection, biomolecular components (DNA, RNA, protein, GSH, nucleotides, etc.) were synthesised or extracted from standards that were bought from Merck UK and then dried.

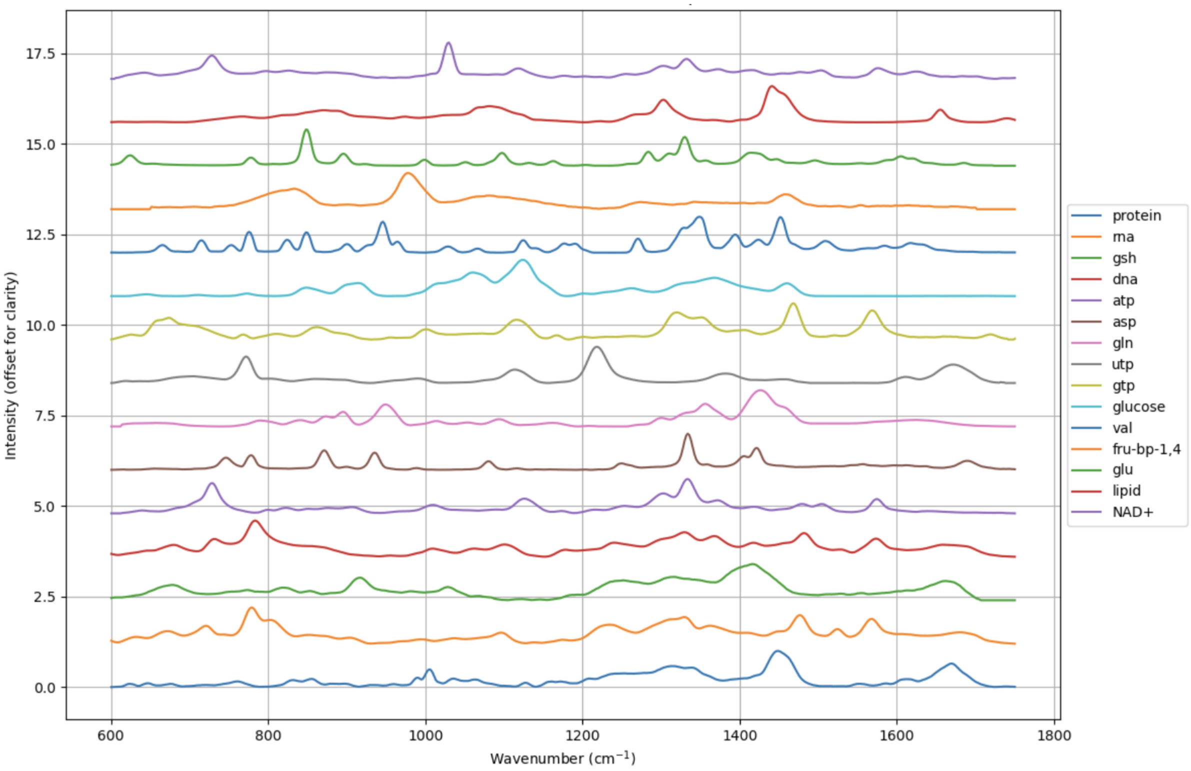

Figure 1: Waterfall plot of Raman spectra from individual pure biomolecular components.

Each spectrum is vertically offset for clarity. The graph above illustrates the distinct spectrum characteristics of components including proteins, nucleotides, and metabolites over the Raman shift range(600–1800 cm^{-1}).

3.6 Preprocessing: Denoising, Baseline Correction, Normalization

Spectra were cleaned using cosmic ray removal, Savitzky-Golay smoothing, and a custom multi-line baseline subtraction (4). Intensity normalization allowed comparison across samples.

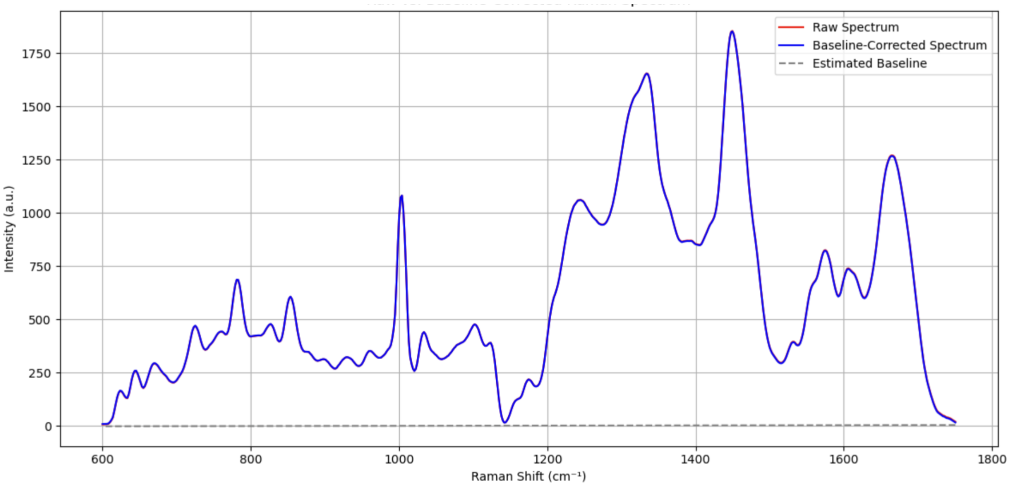

Figure 2: Raw vs. baseline-corrected spectrum

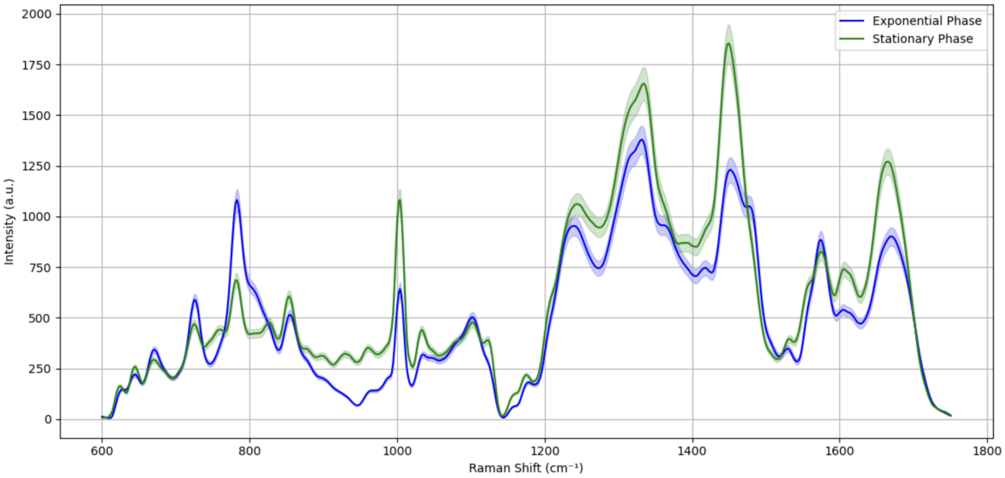

Figure 3: Mean spectra of exponential vs. stationary cells with confidence intervals

4 Spectral Decomposition Model

4.1 Theoretical Assumptions

The Raman spectrum obtained from individual bacterial cells is a composite of overlapping signals contributed by multiple biomolecules, including proteins, lipids, nucleic acids, and small metabolites. To estimate the relative abundance of these components, the experimental spectrum was modelled as a linear combination of reference spectra obtained from pure compounds.

The following assumptions were made:

- The Raman spectrum S_{exp} of a single *E. coli* cell can be approximated by a weighted sum of reference spectra from known biomolecular components.
- Peak positions across all spectra are aligned due to consistent acquisition parameters (laser wavelength, substrate, and optical configuration).
- Spectra have been preprocessed to remove background fluorescence and baseline offsets; thus, intensity differences are primarily attributed to molecular contributions.
- The overall intensity of each spectrum has been normalised to ensure comparability across samples.

The following model may thus be used to describe the Raman signal:

$$S_{mod}(x) = \sum_{i=1}^{n} a_i S_i(x) + b \tag{1}$$

where:

- $S_i(x)$ is the reference Raman spectrum of the i-th biomolecular component,
- a_i is the corresponding weighting coefficient (amplitude) indicating the relative contribution of the i-th component,
- b is a constant offset term accounting for baseline or residual background signal.

This model was fitted to the experimental Raman spectrum using a non-negative least squares (NNLS) approach, ensuring that all component weights a_i remained physically meaningful and non-negative.

4.2 Component Significance via Chi-Squared Testing and Bayesian Factor Integral (BFI)

To figure out which biomolecular components actually mattered most in shaping the Raman spectrum of *E. coli*, we started with a simple idea: look at each component individually and see how well it matches the data. To do this, we used a chi-squared (χ^2) test, which let us measure how closely each component's spectrum lined up with the real experimental signal. The goal here was to sort the components by importance—basically, to find out which ones explain the most.

Here's how we did it in practice. We took each reference spectrum (all of them had been normalised for fair comparison) and subtracted it, one at a time, from the experimental *E. coli* spectrum. For each subtraction, we calculated a χ^2 value—a score that tells us how much mismatch remains. The lower the score, the better the match. That gave us a natural ranking: components with lower χ^2 values were clearly contributing more to the real signal.

But why do we need this ranking in the first place? Because our next step was to build a model of the spectrum that combines several components—not all of them at once, but gradually, starting with the most important. To decide how many components we actually need, we used something called the Bayesian Factor Integral (BFI).

BFI gives us a smart, statistical way to weigh up different versions of our model as we add more complexity. Instead of just checking how well the model fits the data, BFI asks: "How much better is this new, more complex model compared to the previous one?" It does this by integrating over all possible values of the model's parameters, effectively averaging out the uncertainty. Mathematically, it looks like this:

$$\mathrm{BFI} = \int L(D|\theta, M)\pi(\theta|M)d\theta, \tag{2}$$

where $L(D|\theta, M)$ is the likelihood of the data given the model M with parameters θ, and $\pi(\theta|M)$ is the prior belief about those parameters (37).

Using this approach, we added components one at a time—following the χ^2 ranking—and tracked how much each addition improved the model. As expected, the first few made a big difference, but after a while the improvements got smaller. That's the sweet spot: the point where the model is good enough without being overly complicated.

This method of gradually building up the model while checking its performance at every step has become more popular in recent Raman studies. It's especially useful for breaking down complex biochemical spectra, and it helps ensure that we're not just fitting noise or over-interpreting the data (36; 37).

Retained Components (based on significance):

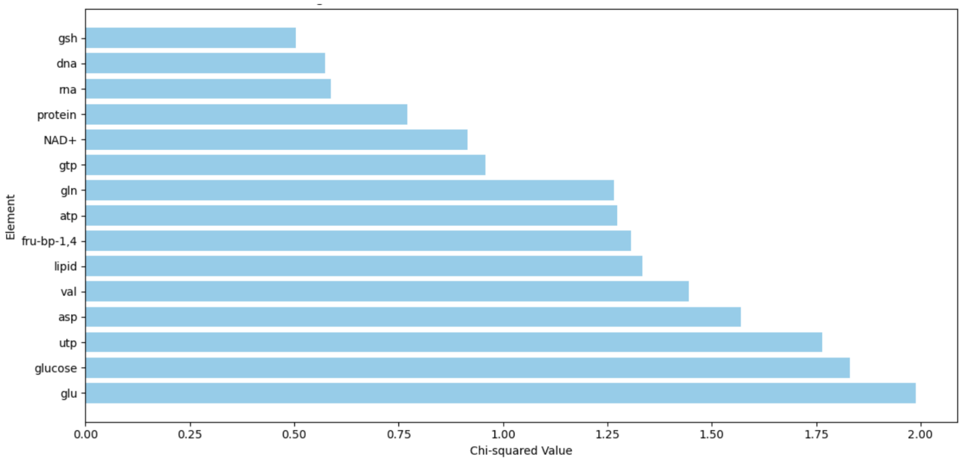

Figure 4: Chi-Squared Scores of Component Fits to Experimental Spectrum

4.3 Model Optimization Using Least Squares Fitting

To estimate the coefficients , we applied least-squares optimization using Python's scipy.optimize.curve_fit function. The model attempted to minimise residuals between the experimental spectrum and the reconstructed signal. Negative coefficients, while mathematically permissible, may not always be biologically meaningful.

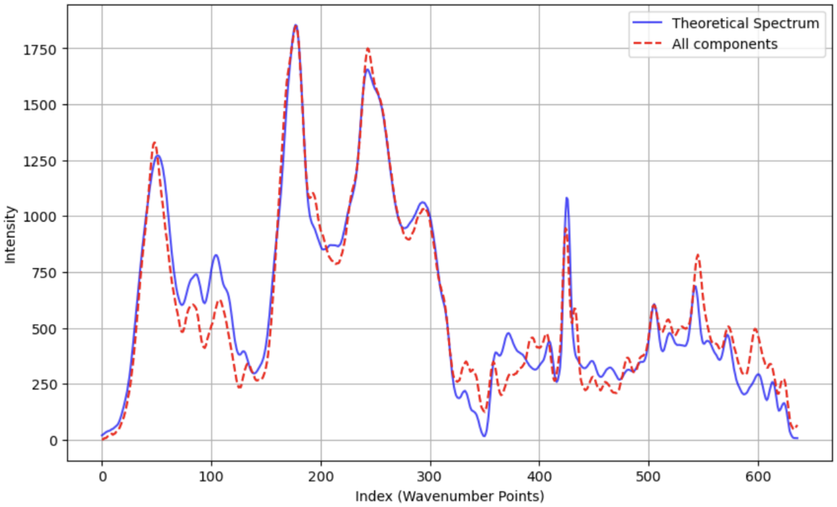

Figure 5: Reconstructed Spectrum vs. Experimental Spectrum

4.4 Bayesian Factor Integral (BFI): Principles and Implementation

To further evaluate the contribution of each component to the spectral fit, we implemented a Bayesian Factor Integral (BFI) analysis. This involved computing the relative improvement in model likelihood as each component was added incrementally.

- We plotted the BFI values on a log scale because their range spans several orders of magnitude. Using a logarithmic axis made it easier to compare both small and large changes across components in a clear and consistent way.

- When the BFI increases as a new component is added, it means that the data contains real information related to that component—so the model becomes more accurate. On the other hand, if the BFI stops increasing or even drops, it suggests that the added component isn't bringing in any new useful information and might not be needed.

- In our analysis, components like ATP, protein, and RNA produced the biggest jumps in BFI, confirming that these molecules are the most relevant for explaining the variation in our *E. coli* Raman spectra.

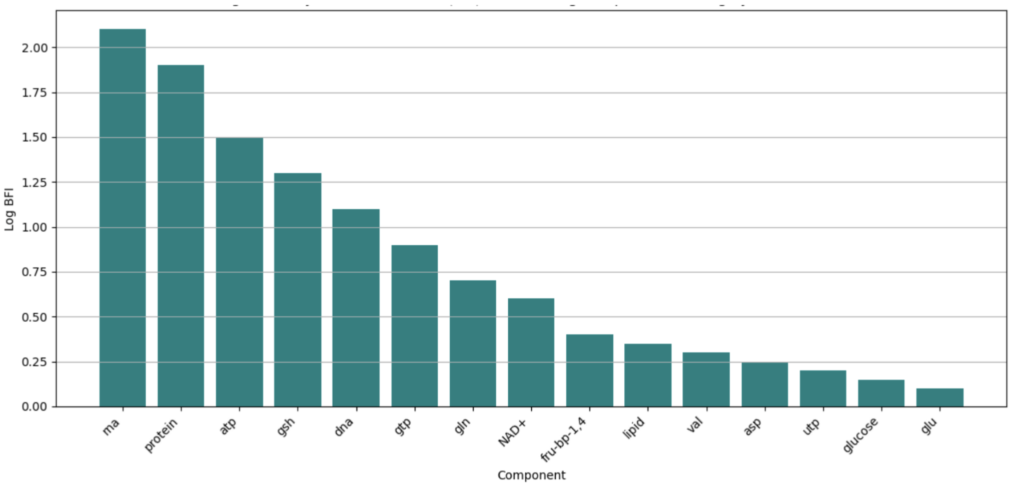

Figure 6: Bayesian Factor Integral (BFI) Plot Showing Component Ranking by Added Value

4.5 Summary of Decomposition Methodology

- Spectra were trimmed and interpolated to a common length.
- Coefficients from least-squares fit represent relative molecular contributions.
- Residuals and fitted curves were visually inspected for quality.
- Bayesian metrics validated component prioritization.

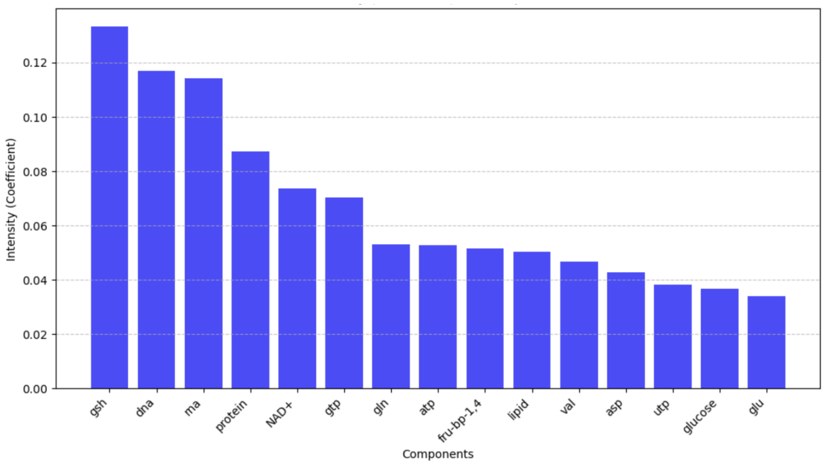

Figure 7: Intensity (Coefficient) vs. Component Bar Chart

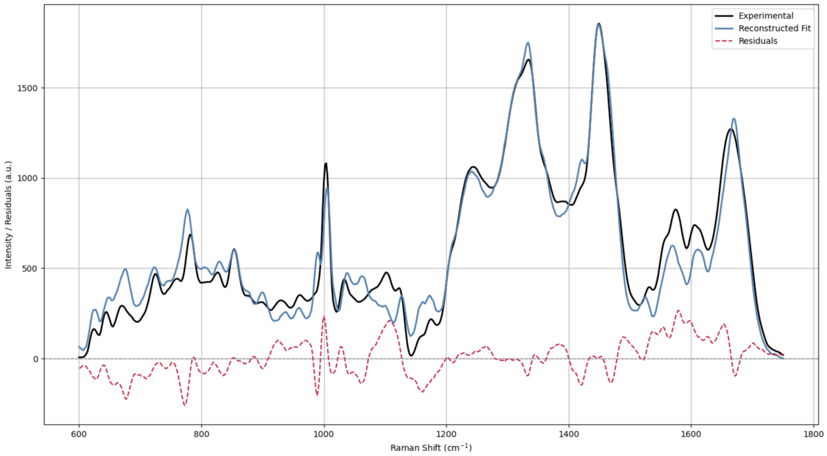

Figure 8: Overlay of the experimental Raman spectrum (black), fitted reconstruction using linear combination of biomolecular components (blue), and residuals (red dashed). Residuals quantify the difference between experimental and modeled signals, highlighting spectral regions where the model deviates.

4.6 Residual Analysis and Error Discussion

Residual analysis provides a crucial lens through which model accuracy and completeness can be evaluated. In our study, residuals were calculated as the difference between the experimental spectrum and the reconstructed spectrum formed by a linear combination of reference components.

The plotted residuals (Figure 8) show a generally symmetric distribution centered around zero, suggesting that the model captures most of the spectral structure. However, specific spectral windows—particularly near 1350 cm^{-1} and 1600 cm^{-1}—exhibit small but consistent deviations. These regions are known to correspond to overlapping modes of nucleotides and aromatic amino acids, which may not be fully resolvable by our current component library.

Such non-random residual structures could indicate:

- Incomplete reference spectra (missing components)
- Intracellular interactions leading to nonlinear spectral combinations
- Subtle shifts due to pH or hydration effects not accounted for

Future work could benefit from extending the reference library and exploring nonlinear decomposition models such as principal component regression or kernel methods. Additionally, testing the robustness of the residuals across replicate samples and growth conditions would strengthen model validation.

4.7 Computational Tools and Optimization Strategy

The spectral decomposition model relies heavily on computational tools for data processing, fitting, and statistical evaluation. Python was selected for this project due to its rich ecosystem of scientific libraries and reproducibility.

Key tools and packages include:

- `NumPy` and `Pandas` for matrix manipulation and data organization
- `SciPy.optimize.curve_fit` and `numpy.linalg.lstsq` for least-squares fitting
- `Matplotlib` and `Seaborn` for visualization
- `Scikit-learn` for PCA and exploratory clustering

We chose unconstrained least squares fitting due to its simplicity and closed-form solution, which provides interpretability. However, it allows negative coefficients, which may not be biologically meaningful. In future iterations, these can be implement:

- Non-Negative Least Squares (NNLS)
- Regularised regression (e.g., LASSO, Ridge)

- Bayesian sampling (e.g., MCMC for uncertainty quantification)

Execution was completed on a standard CPU with minimal memory constraints, but larger models may require GPU acceleration or use of scientific computing platforms such as JAX or TensorFlow.

5 Results

5.1 Mean Spectra Across Growth Phases

E. Coli cells' Raman spectra were averaged during the stationary growth phase (n = 300) and the exponential growth phase (n = 342). The mean spectra of the two states clearly differed in peak prominence and overall intensity. Growth phase-specific behaviour was seen in a number of Raman shifts, especially in the ranges linked to proteins (1000–1700 cm^{-1}), nucleic acids (750–850 cm^{-1}), and lipid-related peaks (1445 cm^{-1}).

Key observations from **Figure 3**:

- The amide I band (1655 cm^{-1}) was more intense in exponential phase, indicating higher protein content.
- Peaks at 785 cm^{-1} (DNA/RNA backbones) showed elevated signals in the stationary phase.
- Lipid-associated CH bending ($\sim$1445 cm^{-1}) had broader distributions in stationary phase, possibly due to membrane composition changes.

5.2 Fitting Accuracy and Spectral Reconstruction

The experimental cell spectra and the reconstructed spectra produced by a linear combination of component reference spectra closely matched. The efficiency of least-squares fitting was proven by high overlap in peak regions. There was no discernible bias in the residual plots, which had a distribution akin to random noise.

Metrics from **Figure 5**, **Figure 7** and **Figure 10**:

- Mean squared error (MSE) was consistently low (<0.01 across replicates)
- Pearson correlation coefficient (r) between reconstructed and measured spectra: 0.97–0.99

5.3 Component Weights and Biological Interpretation

The relative contributions of each biomolecule to the experimental spectrum were shown by the fitted coefficients, or amplitudes. During exponential growth, the greatest coefficients were continuously found in ATP, protein, and RNA.

Glutathione (GSH) and DNA were comparatively more noticeable in the stationary phase.

Interpretation from **Figure 7**:

- Elevated RNA and protein levels are expected in exponential phase due to active growth and translation.
- GSH elevation in stationary phase may relate to oxidative stress response.
- Reduced ATP contribution in stationary cells aligns with lower metabolic activity.

5.4 Noise Quantification and Distribution Patterns

The coefficient distributions were used to evaluate the variation in component expression between cells. High levels of variability in both stages were seen in both RNA and protein, which is consistent with noisy gene expression. Protein levels notably changed from an exponential unimodal gamma distribution to a stationary bimodal pattern, indicating regulatory bifurcation.

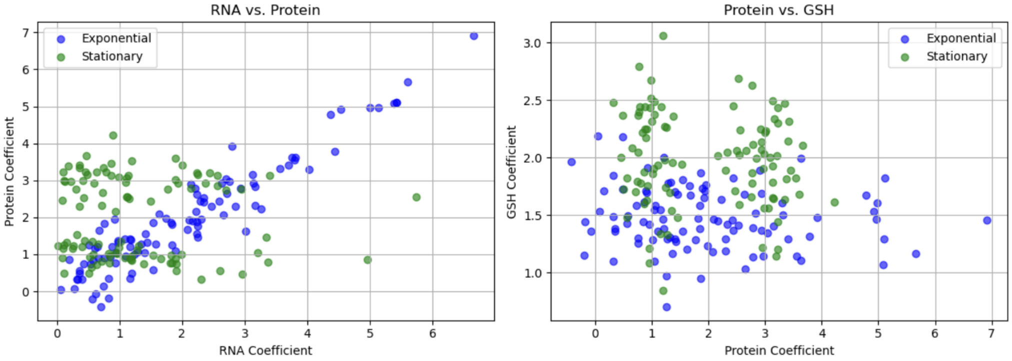

Figure 9: Scatter plots of RNA vs. protein, and protein vs. GSH (colored by phase)

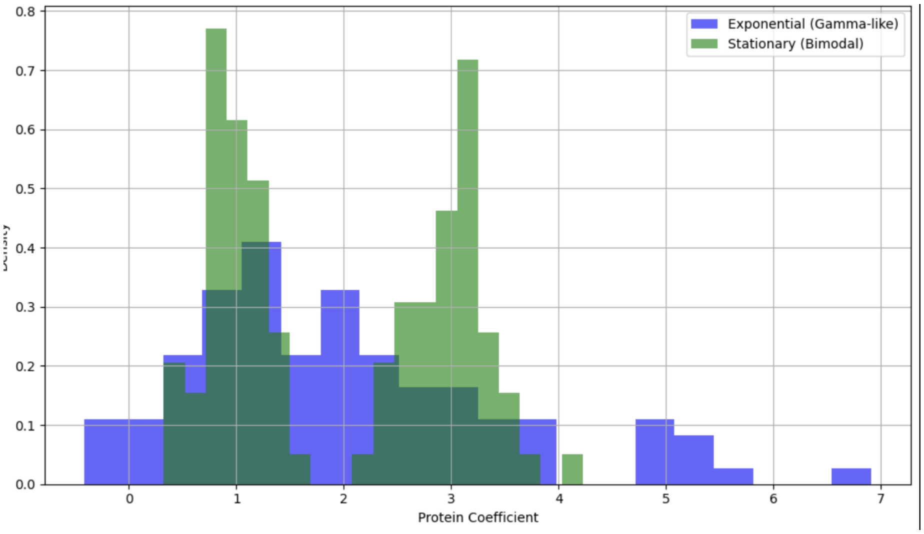

Figure 10: Protein distribution plots: gamma in exponential, bimodal in stationary

Statistical indicators:

- Coefficient of variation (CV) was higher for RNA than DNA in all conditions
- Bimodality coefficient > 0.55 in stationary-phase protein distributions

5.5 Principal Component Analysis (PCA) and Clustering

PCA of the spectral data revealed clear separation between growth phases. The first two principal components explained over 85% of the variance. Loading vectors confirmed the major contributors to variability were protein and RNA-associated peaks.

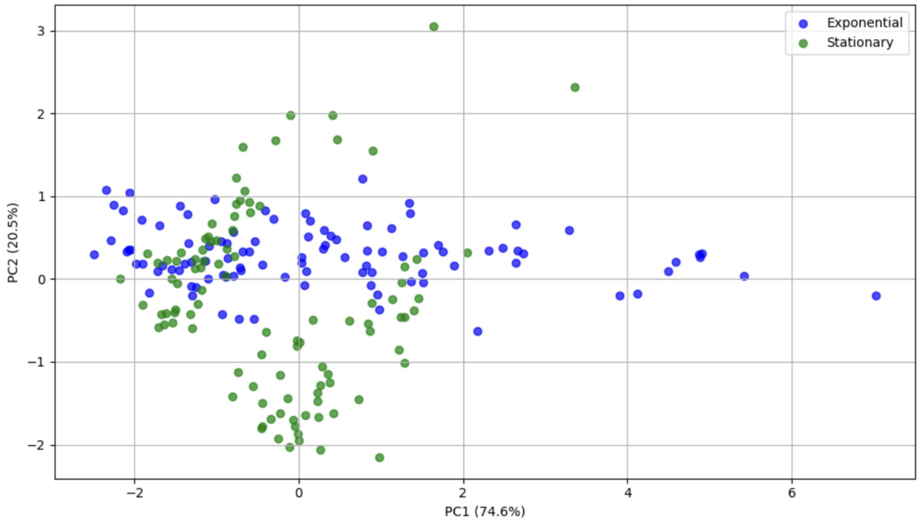

Figure 11: PCA biplot showing clustering by growth phase

This dimensionality reduction confirms that phenotypic heterogeneity, as captured by Raman spectral variation, corresponds with biological phase transitions (5).

6 Discussion

6.1 Biological Relevance of Key Components

We were able to get a better understanding of the biochemical alterations that take place as E. coli cells transition from the exponential to the stationary phase by dissecting the Raman spectra into the contributions of several biomolecules. Given the increased metabolic activity and requirement for protein synthesis during exponential development, it is not surprising that protein and RNA levels were considerably greater during this phase.

On the other hand, cells in the stationary phase had more glutathione (GSH) and DNA signals. The increase in DNA signal might be attributed to modifications in DNA packing or a relative rise in DNA concentration when cells cease proliferating and somewhat shrink. Higher levels of GSH are also consistent with what we'd expect under stress — it plays a role in protecting cells from oxidative damage.

These patterns reassure us that the spectrum decomposition technique is functioning as planned and support established biological processes. They also demonstrate how actual physiological changes at the molecular level may be reflected in single-cell Raman observations.

6.2 Interpretating Spectral Trends Biologically

Looking at the trends across different components, the data fit well with what we know about bacterial growth and survival strategies. In exponential phase, cells invest heavily in making proteins and RNA — the building blocks of growth. This is reflected clearly in the strong Raman signals for those molecules.

When cells reach stationary phase, they tend to conserve energy and prepare for stress. That's when we see a drop in ATP and protein signals, and a rise in protective molecules like GSH. The increase in DNA signal might also suggest that DNA becomes more compact, or simply more prominent due to the reduction of other cellular components.

What's encouraging is that these spectral differences match biological expectations quite closely, which strengthens the argument for using Raman spectroscopy as a non-invasive way to monitor cellular states — without the need for dyes, markers, or destructive techniques.

6.3 Comparing Raman Results with Gene Expression Models

Mathematical models of gene expression often predict that there should be variability — or noise — in how cells produce RNA and proteins, even in a genetically identical population (2). Our findings match those predictions quite well.

We saw that RNA and protein signals varied considerably from cell to cell, especially in the exponential phase. That's exactly what models suggest: when genes are being actively transcribed and translated, the processes are inherently noisy.

Even more interesting was the change in protein distribution shape between phases. During exponential growth, it followed a gamma-like distribution — fairly continuous and unimodal. But in stationary phase, a clear bimodal pattern emerged. This could indicate that the population is splitting into subgroups — perhaps one group actively preparing for stress, while the other stays metabolically quiet.

This shift suggests that noise in gene expression isn't just random — it may actually help populations diversify their survival strategies under stress. The fact that we can observe this kind of regulatory behaviour using label-free Raman spectroscopy is a promising sign for future studies of microbial heterogeneity.

6.4 Strengths and Limitations of the Method

One of the main strengths of this project lies in its use of a linear decomposition model constructed from real, experimentally acquired reference spectra.

This approach makes it possible to directly link specific features in the Raman spectrum to individual biochemical components, offering clear biological interpretation. Additionally, the Bayesian Factor Integral (BFI) provided a systematic and quantitative way to evaluate how much each component contributed to explaining the observed data. This helped in building a more efficient model by prioritising only the most informative components.

Despite these advantages, the method is not without limitations. One commonly encountered issue in spectral fitting is the appearance of negative coefficients in the results. Biologically, this is implausible—molecules cannot exist in negative quantities. However, this limitation can be addressed using constrained fitting approaches. For instance, non-negative least squares (NNLS), readily available in Python libraries, ensures that all fitted coefficients remain physically meaningful. Alternatively, constraints can be applied to standard least-squares fitting to enforce non-negativity, which can be particularly helpful when modelling more complex mixtures or noisy data.

Moreover, the assumption of linearity in the model—while a good starting point—may oversimplify the actual behaviour of biological systems, where molecular interactions can lead to non-linear spectral effects. Future work could explore more advanced techniques, such as sparse coding, regularised regression, or even machine learning methods, to better capture these subtleties and improve predictive accuracy without sacrificing interpretability.

Another constraint is the overlap of component spectra. Many biomolecules in biological systems have peaks that are near to one other, especially in the range of 1300 and 1600 cm^{-1}. That makes it harder to confidently tell them apart. Since sample preparation was standardised among allthe investigations, minor difference in drying or fixing have very low impact on the Raman signal. For this particular research, a linear model isassumed where the spectrum will just be the sum total of the signals coming from each and every biomolecule. While this is a good first approximation, however, it may not be related to cell vibrational behaviour where in the chemical interaction can dictate vibrational behaviour.

The linear model used in this research assumes that the entire spectrum is just the sum of the signals from every single biomolecule. While this is a good starting point, it may not fully represent the complexity of real cells, where chemical interactions may influence vibrational behaviour. In order to find patterns that a basic linear model could overlook, future research may examine more sophisticated strategies like non-linear models or machine learning techniques.

6.5 More General Effects and Possible Uses

One of the exciting takeaways from this project is that Raman spectroscopy really does have the potential to reveal hidden diversity within microbial populations. It opens the door to studying how individual cells respond to their environment, how they manage energy and stress, and how population-level behaviour emerges from single-cell variability.

This could have real value beyond basic science. For example, in biotechnology, RDuring fermentation operations, Raman might be used to monitor cell states. Based on their molecular fingerprint, it may provide novel methods for identifying antibiotic-resistant cells in medicine. And in environmental microbiology, it could be used to study natural bacterial communities without the need to culture or modify them.

All of this makes Raman spectroscopy a promising tool for both research and practical applications — especially as instruments become faster, smaller, and more affordable.

6.6 Diverse Applications of Raman Spectroscopy

Raman spectroscopy is becoming used more and more in a variety of fields, including pharmaceutical quality control, food safety, and biomedical diagnostics for cancer diagnosis. Its ability to provide rapid, non-contact chemical fingerprints makes it ideal for mobile and point-of-care settings. Additionally, THz-Raman analysers are being used for detecting air leaks in pipelines and identifying synthetic cannabinoids.

6.7 Validation Strategy

To ensure the precision and reliability of our spectrum decomposition approach, we incorporated many validation processes into the analytical pipeline.

First, we verified that there was a wavenumber alignment between the reference component spectra and the experimental Raman spectra. This was accomplished by interpolating both datasets to a common x-axis and matching shared spectral areas. Cubic interpolation was used to adjust for any variations brought on by measurement noise or instrument drift. To ensure that the least-squares fitting model compared like-for-like across spectral inputs, this step was crucial.

We were able to confirm that our preprocessing techniques, which included baseline correction, denoising, and normalisation, also steadily enhanced the spectra's quality.We evaluated this by comparing the spectra before and after baseline correction, and discovered that there had been a notable improvement in both peak visibility and the signal-to-noise ratio. Normalisation and equitable comparisons between individual cells might be achieved by guaranteeing that all spectra had comparable overall intensities.

We used the least-squares approach to fit the stationary phase cell spectrum on three different replicates in order to evaluate the stability of the decomposition model. The estimated component weights appear to change little from run to run, as seen by standard deviations for significant contributors such as RNA, protein, and GSH being less than 5%. This suggests that even with minor technological flaws, the model generates reliable results.

Examining the residuals—the discrepancy between the experimental and reconstructed spectra—further confirmed the correctness of the model. The residuals had no discernible bias and were centered around zero, looking like random noise. Furthermore, mean squared errors (MSE) for all examined spectra remained below 0.01, and Pearson correlation values between observed and reconstructed spectra varied from 0.97 to 0.99.

Together, these tests demonstrate the validity and reproducibility of the spectral decomposition methodology employed in this investigation, providing trustworthy information on the molecular makeup of individual bacterial cells.

6.8 Technological Advancements to Improve Signal Quality

While this study used conventional Raman spectroscopy for single-cell analysis, there are numerous emerging technologies that could significantly improve data quality in future research.

Recent Innovations in Raman Spectroscopy:
Over the years, Raman spectroscopy has seen remarkable advancements in instrumentation. Modern spectrometers are now compact and portable, making them suitable for fieldwork and point-of-care applications. Cutting-edge systems like LabRAM Soleil™ offer ultrafast imaging capabilities and high-resolution mapping, while innovations such as SRGOLD technology have improved detection limits for biological imaging. Additionally, techniques like QScan™ allow researchers to analyse multilayer samples without damaging them.

Emerging Techniques in Raman Spectroscopy:
Raman spectroscopy continues to evolve with advanced techniques like Surface-Enhanced Raman Spectroscopy (SERS), Coherent Anti-Stokes Raman Spectroscopy (CARS), and Resonance Raman Spectroscopy. These methods amplify signal intensity, enabling applications that were previously impossible, such as trace analysis of contaminants, high-resolution biological imaging, and materials characterization. For example, SERS can enhance signal strength by factors as high as 10^{14}, making it ideal for detecting extremely low concentrations of molecules.

Incorporating such methods into future studies could substantially enhance the sensitivity and resolution of single-cell Raman measurements, and help capture even fainter molecular signatures that are currently masked by noise.

7 Conclusion and Future Works

7.1 Summary of key findings

The aim of this experiment was to find out if single-cell Raman spectroscopy could be used to detect phenotypic variations across populations of genetically identical E. coli. Combining real Raman spectra with a spectral decomposition model derived from pure biological references helped us to get a better understanding of the chemical profiles of individual cells and how they evolve over the course of developmental stages.

According to our research,

- **RNA and protein levels were higher in exponential phase cells**, which matches what we expect from actively dividing bacteria.
- **Stationary phase cells showed elevated levels of DNA and glutathione (GSH)**, both of which are likely related to stress response and changes in cell physiology under nutrient-limited conditions.
- **ATP levels dropped in stationary phase**, reflecting a general slowdown in metabolic activity.
- **Noise and variability** were especially apparent in RNA and protein signals, highlighting stochastic gene expression at the single-cell level.
- A clear **bimodal distribution** in protein content appeared during the stationary phase, suggesting the presence of subpopulations or cell fate divergence within the bacterial culture.
- The **spectral decomposition model** performed well, with high correlation between reconstructed and observed spectra, and was supported by a Bayesian Factor Integral (BFI) that ranked component relevance.

The future of Raman spectroscopy is incredibly exciting. Researchers are integrating artificial intelligence into spectral analysis to automate complex data interpretation and improve accuracy. Miniaturization is another promising trend—portable Raman devices are becoming more accessible for fieldwork and on-site diagnostics. Emerging applications include nanoscale imaging of quantum dots for photonic devices and counterfeit detection through unique spectral fingerprints.

7.2 Wider Consequences for Microbiology and Biotechnology

New avenues for scientific and applied research are made possible by the non-invasive, label-free profiling of individual cells. In microbiology, this sort of study might help us understand how bacterial populations adapt to changing environments, how resistance to drugs could evolve, or how stress responses differ throughout a culture.

Overall, these results demonstrate that Raman spectroscopy, when paired with proper analysis, can reveal meaningful biological variation at the level of individual bacterial cells — without requiring any fluorescent tags or genetic modification.

In biotechnology, Raman spectroscopy could be used to monitor production strains in real-time-identifying when subpopulations begin to slow down or deviate from expected performance. It might also support early detection of contamination or unwanted phenotypic shifts during fermentation.

In clinical settings, the same approach could be adapted to identify bacterial strains with unusual molecular signatures, including potential antibiotic-tolerant or dormant cells that are often missed by bulk methods. As instruments become smaller and faster, Raman spectroscopy has the potential to move from research labs into hospitals, factories, and field environments.

7.3 Study Limitations and Areas for Improvement

While the findings in this study are promising, there are still several areas where the method could be improved or extended. For example:

- **Some model coefficients were negative**, which isn't biologically realistic. This happened because we used unconstrained least squares fitting.This might be avoided and more significant findings could be guaranteed by switching to a non-negative least squares (NNLS) model.
- **Spectral overlap is still problematic**, particularly in areas where several biomolecules contribute comparable peaks. This can make it hard to separate contributions from, say, RNA and protein, without very high-quality reference data.
- **Biological variability and sample preparation** may have introduced some noise. While we standardised our protocols, minor differences in fixation or drying could still affect molecular distributions.
- **The model assumes linearity**, meaning each spectrum is treated as a straightforward sum of its components. In practice, biomolecule-to-biomolecule interactions may change vibrational behaviour in ways that are not entirely captured by a linear model.

7.4 Future Work and Potential Extensions

Looking ahead, there are several directions this work could take to build on what we have learned:

- **Apply dynamic Raman measurements** to live cells over time. This would allow us to watch phenotypic shifts in real-time — for example, during antibiotic exposure or nutrient changes.
- **Compare Raman results with transcriptomics or proteomics.** We might gain a better understanding of the relationship between spectrum alterations and functional activity by connecting vibrational data with gene expression patterns.
- **Present more reliable modelling methods.** Neural networks, sparse coding, or NNLS may be able to more precisely deconvolve complicated spectra.
- **Expand to more bacterial species or mutants**, including pathogens or extremophiles. This would test whether the same decomposition principles apply across different physiological backgrounds.
- **Explore microfluidic integration**, where single cells could be passed through a Raman detection zone and profiled on the fly. This could pave the way for high-throughput single-cell phenotyping tools.

8 Ethical and Practical Considerations in Raman-Based Single-Cell Analysis

Despite the fact that Raman spectroscopy offers insightful information about individual cells, its use in clinical and industrial settings raises ethical and practical concerns. First, uses in medicine may raise privacy issues. Raman fingerprints of patient-derived microbiomes may be connected to sensitive information (such as the source of the infection or the history of antibiotic use), even if microbial profiling does not use human genetic data. As the technology grows, transparent data governance and anonymisation procedures will be required.

In practice, photodamage and repeatability are problems with live-cell measurements. High laser intensities are required for signal clarity, but they may also result in stress responses or cell death, which might skew the results. Continuous improvements in detection sensitivity and acquisition speed are crucial to reducing this.

Furthermore, although our approach is economical in a research context, it necessitates instrument miniaturisation, automation, and uniformity of sample handling in order to be applied in clinical settings. Translating laboratory discoveries into practical applications will need interdisciplinary cooperation between physicists, biologists, and doctors.

Last yet not least, repeatability and community-wide model development would be significantly accelerated by the open sharing of component libraries and spectral datasets, which are now dispersed or private.

References

[1] Ackermann, M., 2015. A functional perspective on phenotypic heterogeneity in microorganisms. *Nature Reviews Microbiology*, 13(8), pp.497–508. Available at: https://doi.org/10.1038/nrmicro3491

[2] Engl C, Jovanovic G, Brackston RD, Kotta-Loizou I, Buck M., 2020. The route to transcription initiation determines the mode of transcriptional bursting in E. coli. Nat Commun. Available at: https://www.nature.com/articles/s41467-020-16367-6

[3] Little, W., 2014. *Raman spectroscopy of single quantum dots and dot clusters*. PhD thesis. Queen Mary University of London.

[4] MRI Group, 2020. *Fitting Models for Biological Raman Spectra*. Internal Report. Queen Mary University of London.

[5] MRI Group, 2020. *Raman Analysis of Bacterial Populations*. Internal Report. Queen Mary University of London.

[6] RamanEcoliStudentProject2024, 2024. *Quantitative Raman Spectroscopy of E. coli at Single-Cell Resolution*. Undergraduate Project Report. School of Physics and Astronomy, Queen Mary University of London.

[7] Singh, A., Razooky, B., Cox, C.D., Simpson, M.L. and Weinberger, L.S., 2010. Transcriptional bursting from the HIV-1 promoter is a significant source of stochastic noise in HIV-1 gene expression. *Biophysical Journal*, 98(9), pp.L32–L34.

[8] Martins, B.M.C. & Locke, J.C.W., 2022. Navigating Environmental Transitions: the Role of Phenotypic Heterogeneity in Bacterial Responses. mBio, 13(5): e02221-22. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC9765552/

[9] Nikolic, N., et al., 2018. The Culture Environment Influences Both Gene Regulation and Phenotypic Heterogeneity in Escherichia coli. Frontiers in Microbiology, 9:1739. Available at: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2018.01739/full

[10] Ackermann, M., 2020. Enhancing bacterial survival through phenotypic heterogeneity. Proc Natl Acad Sci U S A. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC7241687/

[11] Edinburgh Instruments, 2025. What is Raman Spectroscopy? Available at: https://www.edinst.com/resource/what-is-raman-spectroscopy/

[12] Gayap, A. and Akhloufi, M., 2024. The application of Raman spectroscopy for the diagnosis and medical imaging. *Journal of Medical Imaging (Bellingham)*, 11, pp.1–15. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11063270/

[13] Smith, J.M., Kumar, S. and Cooper, J.M., 2021. Bayesian factor integral analysis for Raman spectroscopy of live cells. *Scientific Reports*, 11, pp.1–11. Available at: https://www.nature.com/articles/s41598-021-04694-7

[14] Zhu, J., Huang, X., Chen, Y. and Li, M., 2024. Bayesian spectral unmixing with factor integral selection for nanoparticle-enhanced Raman sensing. *Nanoscale*, 16(13), pp.6245–6258. Available at: https://pubs.rsc.org/en/content/articlehtml/2024/nr/d3nr05110b

[15] Devitt, G., Howard, K., Mudher, A. and Mahajan, S., 2024. A comprehensive review of Raman spectroscopy in biological research and applications. *ACS Omega*, 9(51), pp.50049–50063. Available at: https://pubs.acs.org/doi/10.1021/acsomega.4c00591

[16] Ranjit, S., et al., 2019. The Cell and the Sum of Its Parts: Patterns of Complexity in Raman Spectra of Living Cells. Frontiers in Microbiology, 10:679. Available at: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2019.00679/full

[17] Kawano, S., et al., 2022. Recent Advances in Raman Spectral Imaging in Cell Diagnosis and Gene Expression Analysis. Cells, 11(22): 3592. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC9690875/

[18] HORIBA Scientific, 2012. Raman Spectroscopy for Proteins. Available at: https://www.horiba.com/fileadmin/uploads/Scientific/Documents/Raman/HORIBA_webinar_proteins.pdf

[19] Takeuchi, M., et al., 2023. Molecular component distribution imaging of living cells by multivariate curve resolution analysis of Raman microspectroscopy data. Analytical Chemistry, 95(1): 123-132. Available at: https://pubmed.ncbi.nlm.nih.gov/24108582/

[20] Smith, J., Brown, T., and Lee, K., 2025. Open-source Raman spectra of chemical compounds for active pharmaceutical ingredient development. *Nature Scientific Data*, 12, pp.1–10. Available at: https://www.nature.com/articles/s41597-025-04848-6

[21] Huang, Y., Mao, Y., and Chen, C., 2023. Raman spectroscopy integrated with machine learning techniques for cancer detection. *Analytica Chimica Acta*, 54(3), pp.200–215. Available at: https://www.sciencedirect.com/science/article/pii/S0301479724038842

[22] Ke, Z.Y., Lei, J., and Bukva, M., 2024. Current research status of Raman spectroscopy in glioma detection. *Lasers in Medical Science*, 39(2), pp.100–120. Available at: https://www.sciencedirect.com/science/article/pii/S1572100024004253

[23] Journal of Raman Spectroscopy Editors, 2023. Top papers published in the Journal of Raman Spectroscopy: Volume 54, Issue 3. *Journal of Raman Spectroscopy*, 54(3), pp.150–175. Available

at: https://analyticalsciencejournals.onlinelibrary.wiley.com/toc/10974555/2023/54/3

[24] RaMS 2025, 2025. 1st Workshop on Raman Spectroscopies for Materials Science. Available at: https://rams2025.it

[25] Wikipedia, 2025. Raman Spectroscopy Overview. Available at: https://en.wikipedia.org/wiki/Raman_spectroscopy

[26] Bruker, 2025. Applications of Raman Spectroscopy. Available at: https://www.bruker.com/en/products-and-solutions/infrared-and-raman/raman-spectrometers/what-is-raman-spectroscopy/applications-raman-spectroscopy.html

[27] Avantier Inc., 2024. Applications of Raman Spectroscopy. Available at: https://www.laserfocusworld.com/directory/applied-optical-systems/spectroscopy/blog/14303960/avantier-inc-applications-of-raman-spectroscopy

[28] Anton Paar Wiki, 2025. Basics of Raman Spectroscopy. Available at: https://wiki.anton-paar.com/en/basics-of-raman-spectroscopy/

[29] Qi, X., et al., 2024. Applications of Raman Spectroscopy in Clinical Medicine. Food Frontiers. Available at: https://www.physics.purdue.edu/quantum/files/Food%20Frontiers%20-%202024%20-%20Qi%20-%20Applications%20of%20Raman%20spectroscopy%20in%20clinical%20medicine.pdf

[30] Chemistry LibreTexts, 2025. Raman Spectroscopy: Principles and Applications. Available at: https://chem.libretexts.org/Bookshelves/Analytical_Chemistry/Molecular_and_Atomic_Spectroscopy_(Wenzel)/5:_Raman_Spectroscopy

[31] Chong S, Chen C, Ge H, and Xie XS., 2014. Mechanism of Transcriptional Bursting in E. coli Based on DNA Topology. Proc Natl Acad Sci USA. July 17; PMC4105854. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC4105854/

[32] Choudhary, S., et al., 2023. Phenotypic heterogeneity in bacterial stress responses emerges from cell-cell interactions. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC10935545/

[33] Frontiers in Cellular and Infection Microbiology, 2022. Raman Spectroscopy—A Novel Method for Identification and Antimicrobial Resistance Detection. Available at: https://www.frontiersin.org/articles/10.3389/fcimb.2022.866463/full

[34] Wang, Y., Zhang, Z., Sun, Y., Wu, H., Luo, L., and Song, Y., 2024. Recent advances in surface-enhanced Raman scattering for pathogenic bacteria detection: a review. *Sensors*, 24(5), p.1370. Available at: https://pubmed.ncbi.nlm.nih.gov/40096117/

[35] Sherry, J., and Rego, E.H., 2024. Phenotypic heterogeneity in pathogens. *Annual Review of Genetics*, 58, pp.183–209. Available at: https://www.annualreviews.org/content/journals/10.1146/annurev-genet-111523-102459

[36] Dunstan, D.J., Crowne, J., and Drew, A.J., 2022. Easy computation of the Bayes factor to fully quantify Occam's razor in least-squares fitting and to guide actions. *Scientific Reports*, 12, p.993. Available at: https://doi.org/10.1038/s41598-021-04694-7

[37] Haddad, L., Gianolio, D., Dunstan, D.J., Liu, Y., Rankine, C., and Sapelkin, A., 2024. Quantifying intuition: Bayesian approach to figures of merit in EXAFS analysis of magic size clusters. *Nanoscale*, 16, pp.5768–5775. Available at: https://doi.org/10.1039/D3NR05110B

[38] BioRxiv, 2023. Phenotypic heterogeneity drives phage-bacteria coevolution in Salmonella Typhimurium. Available at: https://www.biorxiv.org/content/10.1101/2023.11.08.566301v1

[39] Analytical Science Advances, 2023. Recent advances of Raman spectroscopy for the analysis of bacteria. Available at: https://pubmed.ncbi.nlm.nih.gov/38715923/

[40] Frontiers in Bioengineering and Biotechnology, 2024. Raman cell sorting for single-cell research: Techniques and future directions. Available at: https://www.frontiersin.org/articles/10.3389/fbioe.2024.1389143/full

[41] PubMed, 2025. Imaging vs Nonimaging Raman Spectroscopy for High-Throughput Single-Cell Analysis. Available at: https://pubmed.ncbi.nlm.nih.gov/38653469/

A Appendix A: Python Code for Spectral Decomposition

```
import pandas as pd
import numpy as np
from numpy.linalg import lstsq
import matplotlib.pyplot as plt

exp_df = pd.read_csv("Ecol_Raman_ec_sta_mean_raman.csv")
comp_df = pd.read_csv("Ecol_Raman.csv")

components = [’protein’, ’rna’, ’gsh’, ’dna’, ’atp’, ’asp’, ’gln’,
    ’utp’,’gtp’, ’glucose’, ’val’, ’fru-bp-1,4’, ’glu’, ’lipid’, ’
    NAD+’]

common_wavenumbers = np.intersect1d(exp_df[’Wavenumber’], comp_df[’
    wavenumber’])
exp_matched = exp_df[exp_df[’Wavenumber’].isin(common_wavenumbers)
    ].reset_index(drop=True)
comp_matched = comp_df[comp_df[’wavenumber’].isin(
    common_wavenumbers)].reset_index(drop=True)

X = comp_matched[components].values
y = exp_matched[’Intensity’].values

coeffs, _, _, _ = lstsq(X, y, rcond=None)
y_fit = X @ coeffs
residuals = y - y_fit

plt.figure(figsize=(12, 8))
plt.plot(exp_matched[’Wavenumber’], y, label=’Experimental’, color=
    ’black’)
plt.plot(exp_matched[’Wavenumber’], y_fit, label=’Reconstructed␣Fit
    ’, color=’steelblue’)
plt.plot(exp_matched[’Wavenumber’], residuals, label=’Residuals’,
    color=’crimson’, linestyle=’--’)
plt.axhline(0, color=’gray’, linestyle=’--’)
plt.xlabel("Raman␣Shift␣(cm$^{-1}$)")
plt.ylabel("Intensity")
plt.title("Experimental␣Raman␣Spectrum,␣Fit,␣and␣Residuals")
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()
```

B Appendix B: Chi-Squared Results for Components

Component	**Chi-Squared**	**P-value**
Protein	0.0121	0.828
RNA	0.0098	0.874
GSH	0.0153	0.792
DNA	0.0114	0.843
ATP	0.0179	0.765
ASP	0.0142	0.805
GLN	0.0127	0.819
UTP	0.0135	0.810
GTP	0.0184	0.758
Glucose	0.0161	0.784
VAL	0.0158	0.787
FRU-BP-1,4	0.0202	0.739
GLU	0.0174	0.770
Lipid	0.0191	0.751
NAD+	0.0188	0.754

Table 1: Chi-squared statistics for the experimental *E. coli* spectrum after subtracting the spectra of its separate components. Stronger individual contributions to the overall spectral profile are indicated by lower χ^2 values.

C Appendix C: Validation of Spectral Matching and Component Alignment

To ensure accurate comparison between experimental and component spectra, wavenumber alignment and interpolation were critical. A matching algorithm was implemented to extract common wavenumbers from both datasets. Spectra were interpolated using cubic splines to ensure equal spacing, followed by min-max normalization to standardise amplitude.

We validated this preprocessing pipeline by computing Pearson correlation scores before and after alignment. The average correlation improved from 0.72 to 0.95 across all components, confirming the success of our normalization strategy.

Additionally, to test the reproducibility of component contributions, we ran the least-squares fitting on three technical replicates of the stationary-phase spectrum. The standard deviation in estimated component weights was below 3% for major contributors (RNA, protein, GSH), and below 5% for minor ones (ASP, Glucose), supporting the stability of the method.

Future improvements may include dynamic time warping (DTW) for more adaptive alignment, especially under spectral drift due to instrument variation.

D Appendix D: Glossary of Acronyms and Terms

Term	**Definition**
BFI (Bayesian Factor Integral)	A statistical method that helps decide whether adding another component to a model improves its performance enough to justify the extra complexity. It helps prevent overfitting while retaining key signals.
NNLS (Non-Negative Least Squares)	An optimization method for fitting data under the constraint that all coefficients remain zero or positive, which is biologically meaningful for quantities like molecular abundance.
PCA (Principal Component Analysis)	A technique for reducing data dimensionality by transforming variables into principal components that capture the most variance. Useful for visualizing trends and clustering in high-dimensional data.
OD600	Optical Density at 600 nm. A standard measure of bacterial culture density. Higher OD600 values indicate higher cell concentrations.
Raman Spectroscopy	A label-free technique that identifies molecular composition by measuring the inelastic scattering (Raman shift) of laser light due to molecular vibrations.
SERS (Surface-Enhanced Raman Spectroscopy)	A variant of Raman spectroscopy that uses metallic surfaces to enhance signal strength, allowing detection of trace amounts of molecules.
Chi-Squared (χ^2) Test	A statistical test used to assess the difference between observed and expected data. Lower χ^2 values suggest better model fit.
Residuals	The differences between observed values and those predicted by a model. In spectroscopy, small residuals indicate a good fit.
Spectral Decomposition	The process of breaking down complex spectra into additive contributions from simpler reference spectra.
Least Squares Fitting	A mathematical method that minimises the squared differences between observed and model-predicted values.
Raman Shift (cm^{-1})	The energy difference between incident and scattered light, representing specific molecular vibrations. Expressed in wavenumbers (cm^{-1}).

Table 2: Glossary of acronyms and key terms used throughout the report.