Project Title Page

MSc Data Science Project Report

Infrared Analysis of Blood Serum

prepared by

Niyas Yasin

Project Supervisor Professor Neil Hunt

Date 25/08/2024

No. of Words 7450

2023/24 Entry Cohort

MSc Data Science: MSc Project

Declaration of Help Received

Name of Student: Niyas Yasin

Project Supervisor(s): Professor Neil Hunt, Dr Alan Lewis

This form should be completed by each student and submitted as part of the report. The information provided allows proper acknowledgement of the help you have received and will help the markers in evaluating your performance.

1. Did anyone other than the project supervisor(s) assist you in carrying out your project?

   Tick the appropriate box:

   ☐ Yes — ***Complete Section 2***

   ☒ No — ***Go to section 3***

2. Detail the help received (Suggested categories are given below the table):

Complete in block capitals

Name of Helper	Category (i to v)	Details (if necessary)

(i) Day to day help
(ii) Initial training in a technique or operation
(iii) Specific assistance (e.g. correction of code)
(iv) Help with interpretation of results
(v) Any other category of help

3. I have read the Assessment Information in the MSc Handbook on academic misconduct and plagiarism. I have completed the Academic Integrity Tutorial.
   I declare that this assessment is a presentation of original work, I am the sole author and I am not attempting to pass off work created by generative AI content tools as my own.

Name of student (in lieu of signature)

Niyas Yasin Date 25/08/2024

Contents

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to everyone who has contributed to this project from its inception.

Firstly, I extend my heartfelt thanks to Professor Hunt, whose unwavering support and invaluable guidance provided the foundation for this project. His insights and encouragement were instrumental throughout the entire process.

I would also like to give special thanks to Alan Lewis and Connor, who offered crucial assistance with the coding aspects. Their willingness to help and share their expertise made navigating the technical challenges much smoother.

Finally, I am deeply grateful to Neil Hunt for his patience, continuous guidance, and constructive feedback, especially as the project neared completion. His input was vital to the project's success.

To all those mentioned, your support has been greatly appreciated, and this project would not have been possible without your contributions.

Abstract

This study explores the application of two-dimensional infrared (2D-IR) spectroscopy combined with machine learning to analyse protein secondary structures, specifically focusing on the α-helix and β-sheet content. Various machine learning models, including linear regression, support vector regression (SVR), kernel ridge regression (KRR), random forests, and feedforward neural networks, were evaluated to predict these structural elements from spectroscopic data. The study's results, validated against an independent dataset, demonstrate the effectiveness of integrating spectroscopy and machine learning for detailed protein structure analysis.

Chapter 1: Introduction

1.1 Background

The Role of Spectroscopy in Molecular Analysis

Spectroscopy, particularly infrared (IR) spectroscopy, is an essential analytical technique used to probe molecular structures and interactions (Rutherford, Greetham, Parker, et al., 2022). It plays a critical role in various scientific fields, including chemistry, biology, and materials science. IR spectroscopy involves the interaction of infrared radiation with matter, leading to the absorption of specific wavelengths that correspond to the vibrational modes of the molecules. By analysing the absorbed wavelengths, researchers can deduce detailed information about the molecular structure, functional groups, and bonding environments of a sample (Hunt, 2009).

In the biomedical field, IR spectroscopy is a powerful tool for examining the composition and structure of biological molecules, such as proteins, nucleic acids, and lipids (Bellisola & Sorio, n.d.). It is particularly valuable for studying biofluids like blood serum, where it provides insights into the molecular components present (Voronina et al., 2021). However, traditional IR spectroscopy faces significant challenges when applied to aqueous solutions, as water strongly absorbs infrared light, often obscuring the signals from the molecules of interest (Hume et al., 2019).

Challenges in Biomedical Applications

When studying proteins in biofluids, the strong water absorption presents a major obstacle (Fritzsch et al., 2020). Proteins, which are the workhorses of the cell, have complex structures that are crucial for their function. These structures are typically organized into four levels:

primary, secondary, tertiary, and quaternary structures. The secondary structure, which includes α-helices and β-sheets, is of particular interest because it forms the foundation of the protein's overall shape and stability (Kabsch & Sander, 1983).

The α-helix and β-sheet are the two most common types of secondary structure, and their content within a protein can influence its biological activity, stability, and interaction with other molecules (Yang et al., 2022). Accurate analysis of these secondary structures in a physiologically relevant environment (i.e., in water-rich solutions) is essential for understanding protein function in vivo. However, traditional IR methods struggle to distinguish these structures due to the overlapping absorption bands of water and the protein (Hunt, 2024b).

1.2 Advancements with Two-Dimensional Infrared (2D-IR) Spectroscopy

The Need for Enhanced Spectral Resolution

Two-dimensional infrared (2D-IR) spectroscopy is a recent advancement that addresses the limitations of traditional IR spectroscopy by providing enhanced spectral resolution (Hunt, 2009). This technique involves the interaction of two infrared pulses with the sample, creating a two-dimensional spectrum that offers more detailed information about the molecular vibrations. By separating the signals in two dimensions, 2D-IR spectroscopy can disentangle overlapping absorption bands, allowing for clearer identification of the molecular structures present (Shim et al., 2007).

Application to Protein Secondary Structure Analysis

2D-IR spectroscopy is particularly well-suited for studying protein secondary structures in water-rich environments, as it can effectively isolate the signals from the protein's amide I band, which is sensitive to secondary structure (Hume et al., 2019). The amide I band arises primarily from the C=O stretching vibrations of the protein backbone, and its position and shape in the spectrum provide insights into the presence of α-helices, β-sheets, and other secondary structural elements (Yang et al., 2022).

By applying 2D-IR spectroscopy to proteins dissolved in water, researchers can obtain detailed information about the secondary structure in a state that closely resembles the natural physiological environment (Hunt, 2024a). This capability represents a significant advantage over other structural biology techniques such as X-ray crystallography or cryo-electron

microscopy (cryo-EM), which require the protein to be in a non-native state (e.g., crystalline form or frozen) (Fritzsch et al., 2020).

Bottom-Up vs. Top-Down Approaches in Spectral Analysis

In analysing the data from 2D-IR spectroscopy, there are two primary approaches:

1. **Bottom-Up Approach**: This method involves analysing each spectrum individually, often requiring detailed modelling of the spectrum to understand the relationship between the spectral features and the underlying structure (Rutherford et al., 2023). This approach demands a deep understanding of the spectrum-structure relationship and is computationally intensive.
2. **Top-Down Approach**: This approach leverages the high-throughput capabilities of modern 2D-IR spectrometers to measure large libraries of protein spectra (Rutherford, Greetham, Parker, et al., 2022). By employing machine learning techniques, researchers can identify patterns and relationships between the spectral features and the secondary structures without needing to model each spectrum in detail (Ye et al., 2020). This approach is particularly powerful when dealing with large datasets, as it can reveal complex relationships that might be missed by traditional methods.

This dissertation adopts a top-down approach, using machine learning to analyse the 2D-IR spectra and predict the α-helix and β-sheet content of the proteins (Ren et al., 2022). This method represents a novel application of 2D-IR spectroscopy in combination with advanced computational techniques, aiming to provide a more efficient and accurate analysis of protein secondary structures in physiologically relevant conditions (Hunt, 2024b).

1.3 Objectives

The primary objective of this dissertation is to explore the application of 2D-IR spectroscopy combined with machine learning to analyse protein secondary structures. Specifically, the study focuses on predicting the α-helix and β-sheet content in proteins based on their 2D-IR spectra (Rutherford, Greetham, Parker, et al., 2022). The key goals of this research are as follows:

- **Develop a Methodology**: To establish a robust methodology for preprocessing and analysing 2D-IR spectral data, ensuring that the relevant structural information is accurately captured and interpreted (Rutherford, Greetham, Parker, et al., 2022).

- **Evaluate Machine Learning Models**: To evaluate a range of machine learning models, including linear regression, kernel ridge regression, and support vector machines, for their ability to predict protein secondary structure from the spectral data (Mckenna & Dubey, 2022), (Muggleton et al., n.d.).
- **Assess Model Performance**: To assess the performance of these models in terms of accuracy, robustness, and computational efficiency, with a particular focus on their ability to handle the non-linear and complex nature of the 2D-IR spectral data (Ren et al., 2022).
- **Provide a Comparative Analysis**: To provide a comparative analysis of the different models, highlighting their strengths and limitations in the context of protein secondary structure analysis (Wardah et al., 2019).

This study represents a first step in applying 2D-IR spectroscopy and machine learning to protein structure analysis, laying the groundwork for future research that could extend these methods to more complex systems, such as protein-drug interactions (Rutherford, Greetham, Towrie, et al., 2022).

Chapter 2: Data

2.1 Overview of Data

The dataset employed in this study comprises protein samples obtained from commercial suppliers, carefully selected to ensure a broad representation of secondary structures (Hunt, 2024a). These proteins were dissolved in water and subjected to Two-Dimensional Infrared (2D-IR) spectroscopic measurements, with the resulting raw spectra files stored in CSV format. Each protein is identified by its unique Protein Data Bank (PDB) ID, providing a reference to its three-dimensional structural data (Berman, 2000).

The primary focus of this research is the analysis of the α-helix and β-sheet content within the protein samples. **α-helices** and **β-sheets** are the two predominant types of protein secondary structures (Barth, 2007). The α-helix is a right-handed coil structure stabilized by hydrogen bonds between the backbone amides of amino acids, resulting in a rod-like appearance. It is a common motif in the structure of proteins, contributing to the overall stability and function of the protein (Haris & Chapman, 1995). On the other hand, β-sheets consist of β-strands connected laterally by at least two or three backbone hydrogen bonds, forming a twisted,

pleated sheet. The strands can be arranged in parallel or antiparallel orientations, influencing the protein's overall architecture and stability (Kabsch & Sander, 1983).

These secondary structures are crucial because they define the protein's 3D conformation, directly influencing its biological function (Dill & MacCallum, 2012). The α-helix and β-sheet structures can be considered the "fingerprints" of a protein, as their specific arrangements and proportions are unique to each protein, affecting how it interacts with other molecules (Voronina et al., 2021).

To ensure a comprehensive analysis, the dataset includes 33 proteins selected from a well-documented and published library specifically designed to encompass a diverse range of structural types. This selection ensures that the study covers a wide spectrum of protein conformations, enabling a thorough exploration of spectroscopic analysis methods. Each protein in this library was chosen for its distinctive structural characteristics, which are representative of various protein families and functions.

In addition to the main dataset, three additional proteins were reserved as a validation set, providing an independent dataset to assess the robustness and generalizability of the predictive models developed in this study. Each protein sample yielded approximately 200 spectra files, resulting in a substantial and diverse collection of spectral data for subsequent analysis.

The rigorous selection of proteins, coupled with the focus on α-helix and β-sheet content, underscores the study's objective to develop machine learning models that can accurately predict protein secondary structures from 2D-IR spectral data. This approach not only enhances our understanding of protein structures in a physiologically relevant environment but also contributes to the broader field of protein spectroscopy by providing a framework for future studies (Rutherford, Greetham, Towrie, et al., 2022).

2.2 Spectral Data Processing

To analyse protein secondary structures using 2D-IR spectroscopy data, it was critical to preprocess the spectral data to ensure accuracy and reliability. The following preprocessing steps were conducted:

Normalization: The intensity values across all spectra were normalized to a common scale using the Standard Scaler from scikit-learn. This step was crucial to reduce variability between different measurements, particularly given that the measurements were conducted in a water-rich environment where the solvent could introduce significant noise and variation. By

normalizing the data, with a mean of 0 and a standard deviation of 1, we ensured that the spectral intensities were directly comparable across different protein samples, which is essential for accurate downstream analysis.

Selection of Probe Frequency Range:

Rationale: The probe frequency range of 1600-1700 cm^{-1} was chosen for detailed analysis, focusing on the amide I band. This band is well-known for its sensitivity to protein secondary structures, particularly the α-helix and β-sheet components (Barth & Zscherp, 2002). The strong signals within this range make it ideal for analysing these specific structures, providing clear and distinct spectral features that can be used for model training (Rutherford, Greetham, Parker, et al., 2022).

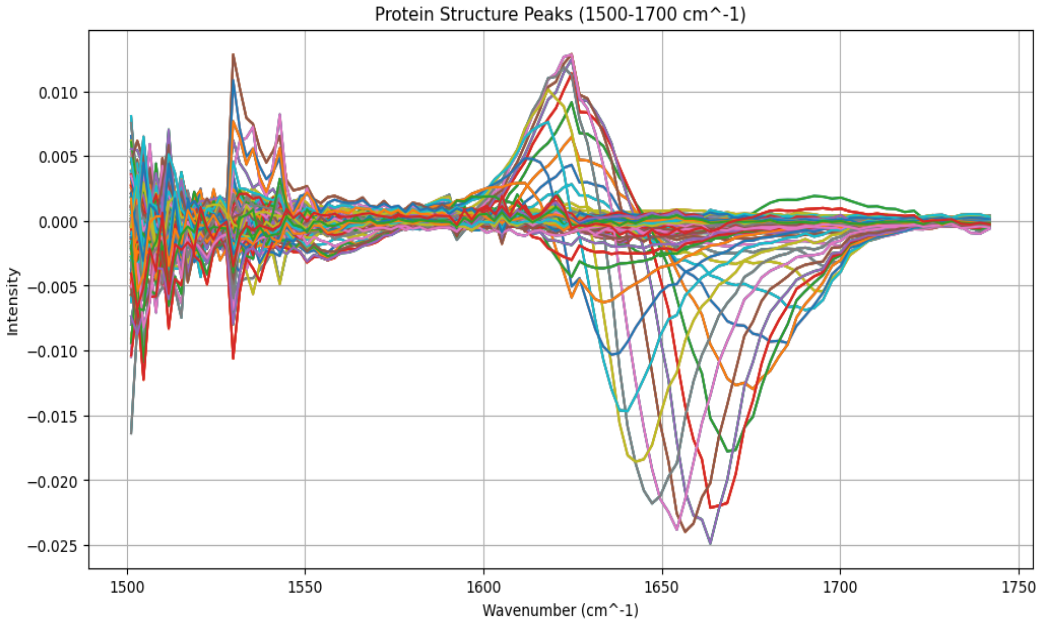

Figure 1

Figure 1: Spectral peaks within the 1600-1700 cm^{-1} probe frequency range, showing the distinct features associated with α-helix and β-sheet structures. The intensity variations across different wavenumbers are crucial for accurate model training and secondary structure prediction.

Feature Extraction

Following preprocessing, key features were extracted from the spectral data to represent the α-helix and β-sheet content of the proteins:

Extraction of Specific Frequency Ranges:

- **β-sheet Content**: The 1630-1650 cm^{-1} range was selected to extract features related to β-sheet structures. This range was chosen because it is associated with characteristic

vibrational modes of β-sheets, which are critical for understanding the secondary structure of proteins (Barth, 2007).

- **α-helix Content**: The 1650-1670 cm^{-1} range was used to extract features related to α-helix structures. This range corresponds to the vibrational modes typically observed for α-helices, making it a reliable indicator of this secondary structure component (Barth & Zscherp, 2002).

Feature Variables:

- **alpha_max**: The maximum intensity value within the α-helix range (1650-1670 cm^{-1}) was extracted as a key feature. This value indicates the peak intensity associated with the α-helix structure, providing a quantitative measure of its prominence in the spectrum (Hunt, 2024b).
- **alpha_min**: The minimum intensity value within the α-helix range was also extracted. This feature helps to characterize the depth of the spectral troughs, which can be indicative of the structural environment of the α-helix.
- **beta_max**: Similarly, the maximum intensity within the β-sheet range (1630-1650 cm^{-1}) was extracted. This feature captures the peak intensity associated with β-sheet structures, which is crucial for identifying the presence and extent of β-sheets in the protein (Hunt, 2024b).

- **beta_min**: The minimum intensity within the β-sheet range was extracted to further describe the spectral characteristics of β-sheets.

These features were selected based on their relevance to the protein secondary structures of interest. They provided the input variables for training various machine learning models aimed at predicting protein secondary structure from the spectral data (Ren et al., 2022).

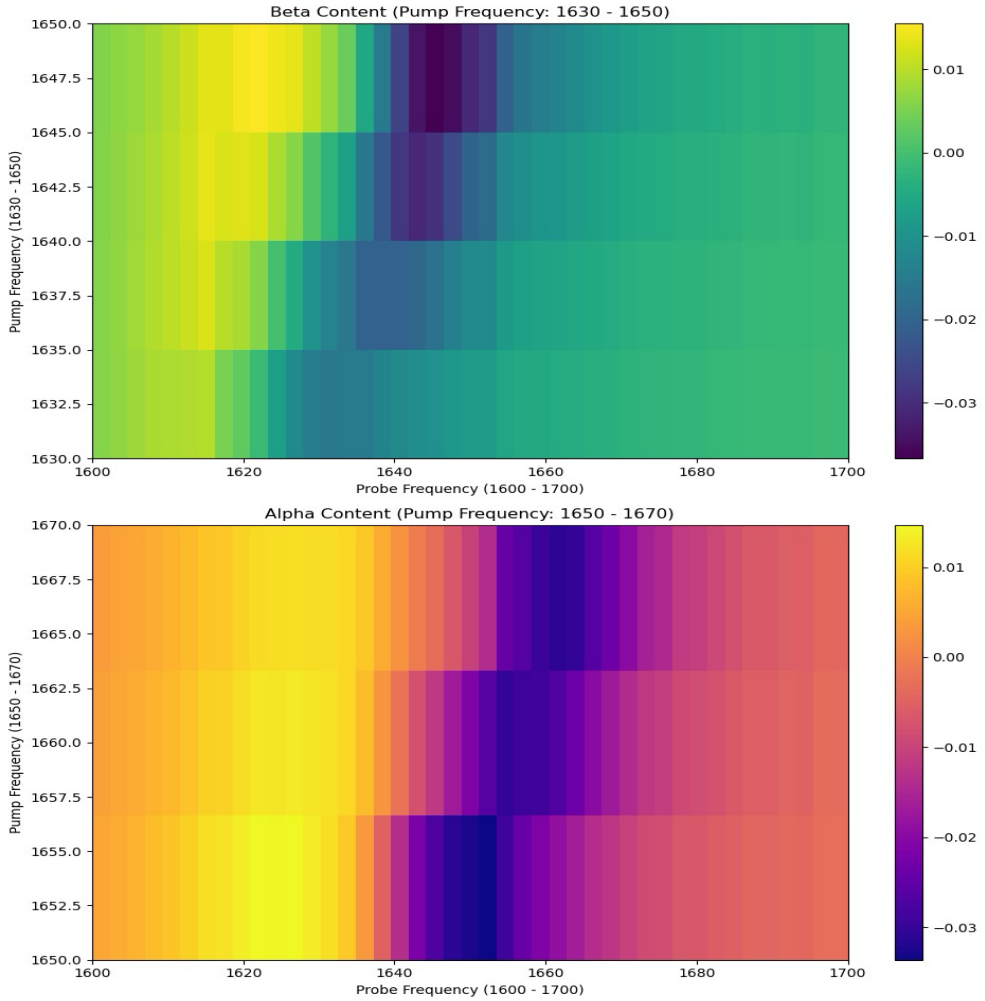

Figure 2

Figure 2: Heatmap representations of (top) β-sheet content in the 1630-1650 cm^{-1} range and (bottom) α-helix content in the 1650-1670 cm^{-1} range, across the probe frequency range of 1600-1700 cm^{-1}. These heatmaps illustrate the specific spectral regions analysed for secondary structure determination.

Chapter 3: Methodology And Results

3.1 Machine Learning Models

3.1.1 Overview of Data Analysis Methods

In machine learning, data analysis methods are broadly categorized into supervised and unsupervised learning techniques. These two paradigms fundamentally differ in how they approach data, their learning processes, and the type of problems they solve (Jianlin Cheng et al., 2008).

Supervised Learning:

Supervised learning relies on training models using a labelled dataset where each data point is paired with its corresponding output. The goal is for the model to learn the relationship between the input data (features) and the output (target), such that it can make accurate predictions when faced with new, unseen data (Jianlin Cheng et al., 2008), (AlQuraishi, 2021). In supervised learning, the quality of predictions is improved through exposure to more training examples and through iterative tuning of model parameters.
In this study, supervised learning is employed to predict the α-helix and β-sheet contents of proteins based on spectral data obtained via 2D-IR spectroscopy. The problem formulation is naturally suited to supervised learning because we are provided with spectral data (input) and corresponding α and β contents (output), which we aim to predict (Ye et al., 2020), (Rutherford, Greetham, Parker, et al., 2022).
Models such as Linear Regression, Kernel Ridge Regression (KRR), and Support Vector Machines (SVM)are examples of supervised learning models used in this research. These models differ in complexity, assumptions, and ability to model non-linear relationships, which will be discussed in greater detail in the following sections (Ye et al., 2020), (AlQuraishi, 2021).

Unsupervised Learning:

Unsupervised learning, on the other hand, deals with unlabelled data. It is often used to identify hidden patterns, groupings, or structures within the dataset (Jianlin Cheng et al., 2008). The model is not provided with any explicit output during training, but instead, it discovers inherent structures on its own. Common applications include clustering, anomaly detection, and dimensionality reduction (AlQuraishi, 2021).
While unsupervised learning methods can be powerful for exploratory data analysis, they were not the focus of this study. Given the availability of labelled data with specific target variables (α and β contents), **supervised learning** was the natural choice for the predictive models used in this study (Jianlin Cheng et al., 2008), (Ye et al., 2020). However, unsupervised methods could potentially complement this approach in future work, for instance, by clustering proteins with similar spectral characteristics or performing unsupervised feature extraction before applying supervised methods (AlQuraishi, 2021), (Ren et al., 2022).

3.1.2 Supervised Learning Models

3.1.2.1 Linear Regression

Linear Regression was chosen as the initial model to serve as a baseline for understanding the relationship between the extracted spectral features and the target variables—Total α and Total β protein content. The assumption was that there might be a linear relationship between the spectral data and the secondary structure elements of the proteins.

Data Preparation

To simplify the analysis and ensure consistency, we selected only one spectral file per protein. This selection was based on the shortest Tw value and was marked as "Cycle1," ensuring the data came from the first cycle of the experiment. This approach was intended to reduce the dataset's complexity while providing a representative sample for each protein.

Given that this was the initial approach, we considered additional features beyond the basic ones. The selected features for this model included **Beta_Max_Intensity**, **Beta_Max_Intensity_Freq**, **Beta_Min_Intensity**, **Beta_Min_Intensity_Freq**, **Alpha_Max_Intensity**, **Alpha_Max_Intensity_Freq**, **Alpha_Min_Intensity**, and **Alpha_Min_Intensity_Freq**. These features were chosen to capture both the amplitude and frequency characteristics of the spectral data, providing a comprehensive input for the regression model (Jianlin Cheng et al., 2008), (Ye et al., 2020).

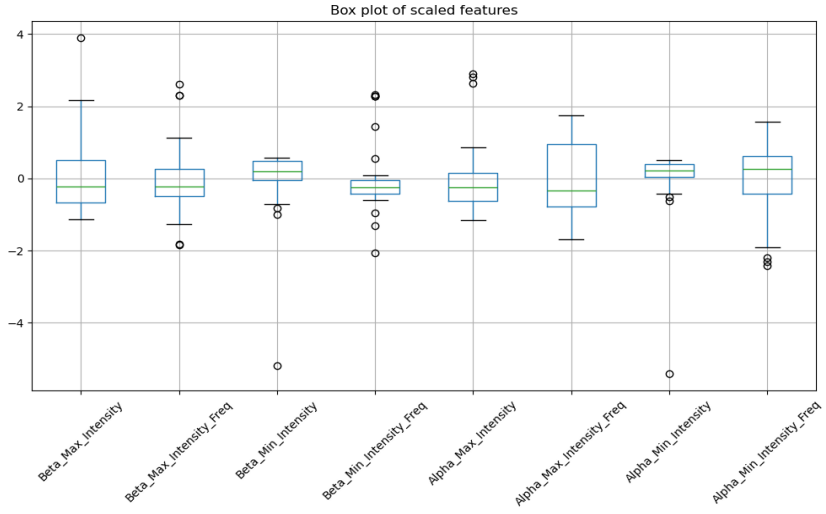

Figure 3

Figure 3: Box plot for checking the outliers

To further examine the distribution of these features and identify any outliers, a box plot of the scaled features was generated (Figure 3). Although some outliers were observed, we decided

not to remove them due to the small size of the dataset, opting to preserve as much data as possible for the analysis (Kohavi, 1995).

Model Training

Initially, the dataset was split into 80% for training and 20% for testing, a common practice in supervised learning (Xu & Goodacre, 2018). The model's effectiveness was evaluated using Mean Squared Error (MSE) and R^2 score (Weisberg & Sanford, 2005). However, due to the relatively small dataset, we noticed that the model's performance could be sensitive to the specific random split of the data. Small variations in how the data was divided could lead to significant differences in model performance, raising concerns about the model's generalizability (Hawkins, 2004).

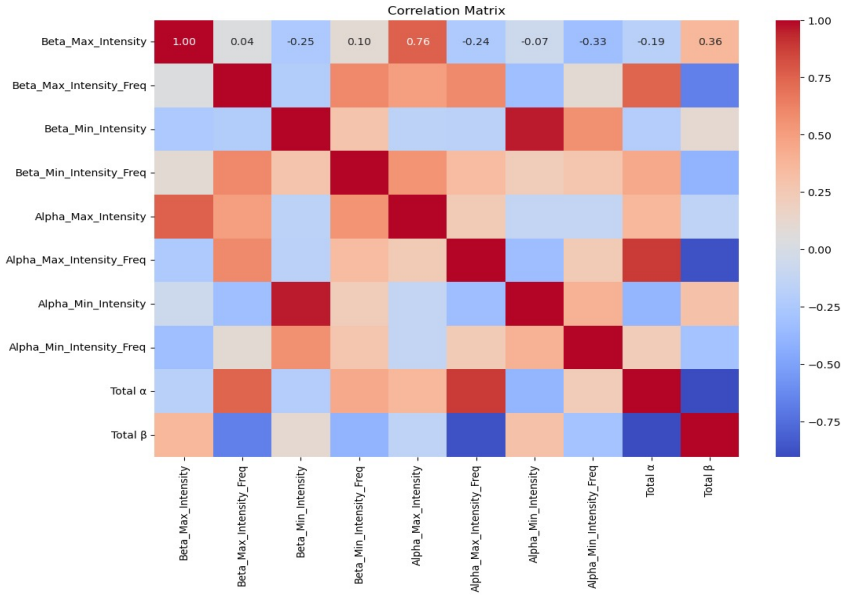

Figure 4

Figure 4: Correlation matrix for eight feature variables

To better understand the relationships between the features and the target variables, a correlation matrix was constructed (Figure 4). This matrix highlighted the correlations among the spectral features and the protein content, providing insight into which features might be most predictive (Jianlin Cheng et al., 2008).

Cross-Validation for Robust Evaluation

To address the sensitivity to data splitting, we implemented **5-Fold Cross-Validation**. The dataset was divided into five equal parts, or "folds," and the model was trained and tested five

times. Each time, a different fold was used as the test set while the remaining four folds were used for training. This method provided a more comprehensive evaluation by averaging the results across all folds, reducing the impact of any single train-test split, and offering a more robust assessment of the model's generalization ability (Kuhn & Johnson, 2013).

Two Approaches: Single and Dual Linear Regression Models

Given that the task involved predicting two separate targets—Total α and Total β—we explored two modelling strategies:

1. **Single Model for Both Targets**: A single Linear Regression model was trained to predict both Total α and Total β simultaneously. While this approach simplified the process, it was sensitive to the randomness in data splits, especially given the small dataset, leading to variability in the model's predictions (Montgomery & Peck, 2012).
2. **Separate Models for Each Target**: Two separate Linear Regression models were developed, one for predicting Total α and another for predicting Total β. This allowed for a more focused approach for each target, potentially improving prediction accuracy by tailoring the model to the specific characteristics of each protein structure component(Jianlin Cheng et al., 2008).

For both approaches, **5-Fold Cross-Validation** was applied. This process helped identify the variability in the model's predictions due to the random state used during data splitting. The results indicated that while Linear Regression provided an important baseline, its limitations became evident, particularly in capturing non-linear relationships within the spectral data (AlQuraishi, 2021).

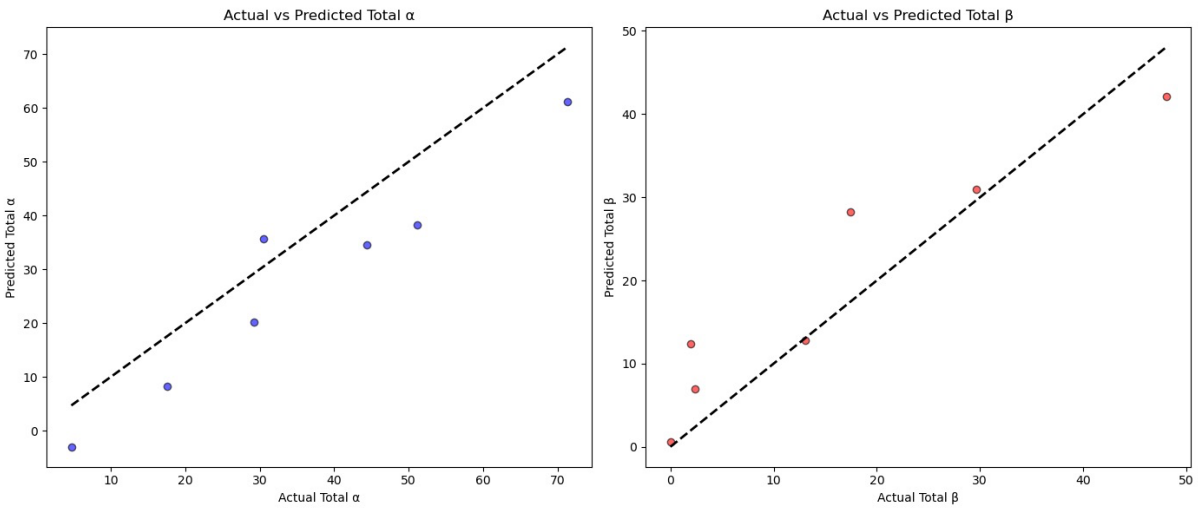

Figure 5

Figure 5: Visualising the model's performance, we plotted the predicted versus actual values for both Total α and Total β. This visual comparison highlights the alignment between the predicted and actual values, illustrating the model's effectiveness and its limitations.

3.1.2.2 Comprehensive Linear Regression Analysis Using Full Dataset

After the initial exploration using a simplified dataset, we expanded our study to incorporate the full range of available spectral data. This section outlines the steps taken to handle the complete dataset, identify and manage outliers, standardize features, evaluate the model's performance, and discuss the limitations of using linear regression in this context.

Data Collection and Feature Extraction

We began by gathering all spectral files available for each protein from our dataset. These files were processed to extract key features—maximum and minimum intensity values within specific frequency ranges associated with α-helix (1650-1670 cm^{-1}) and β-sheet (1630-1650 cm^{-1}) structures (Barth & Zscherp, 2002). This ensured that the model had a comprehensive set of input variables reflecting the relevant spectral characteristics.

Outlier Detection and Management

Given the comprehensive dataset, the detection and management of outliers were crucial to avoid skewing the model's predictions. We first generated a box plot of the standardized features to visually inspect for any extreme values (Figure 6) (Kohavi, 1995).

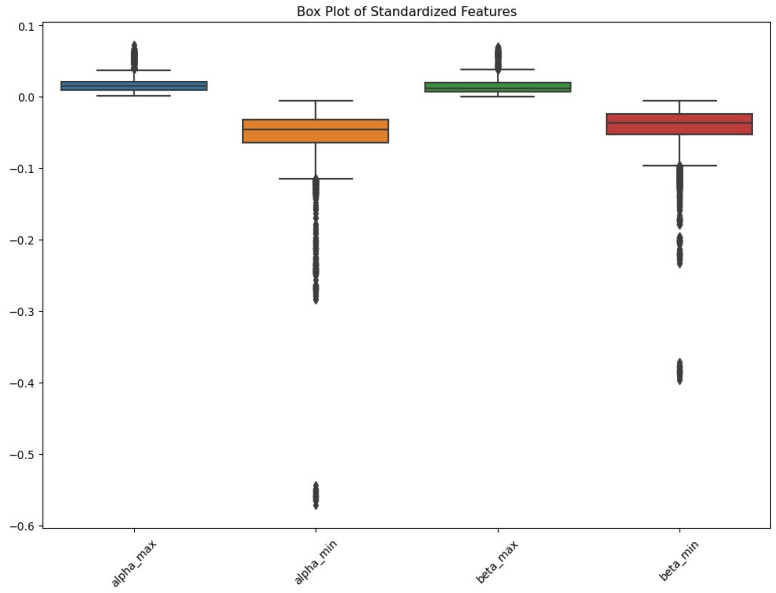

Figure 6

Figure 6: Box plot before cleaning the outliers for the whole dataset

Next, outliers were identified using the Interquartile Range (IQR) method. A pair plot (Figure 7) was then used to confirm the presence of these outliers visually. The identified outliers were subsequently removed from the dataset to improve model reliability.

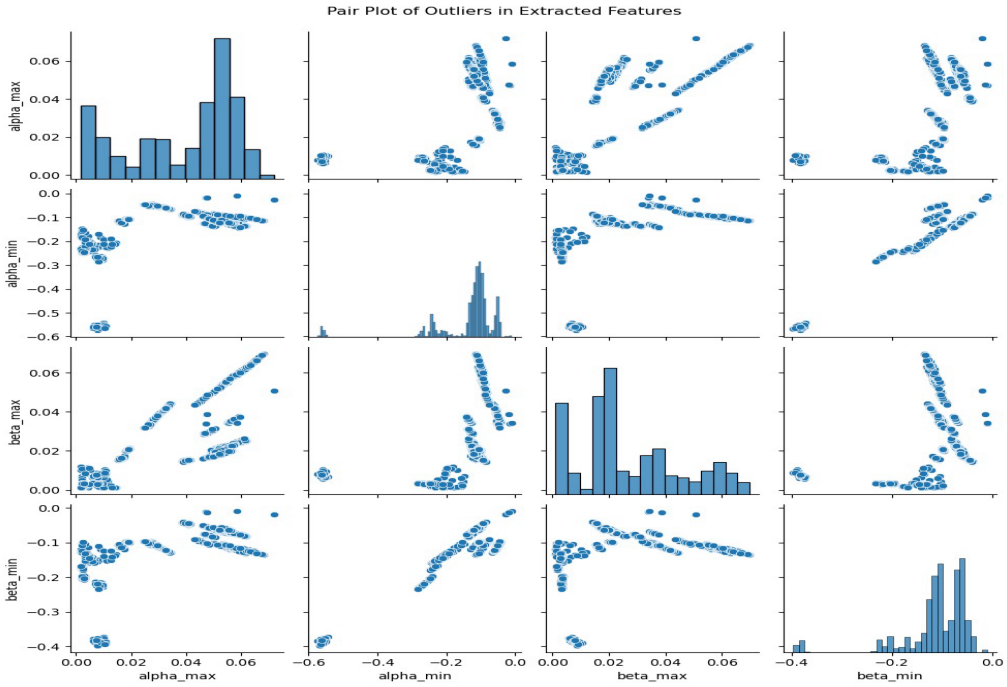

Figure 7

Figure 7: Pair plot of the extracted features (alpha_max, alpha_min, beta_max, and beta_min) illustrating the relationships and distributions of these variables. The diagonal histograms represent the distribution of each feature, while the scatter plots show pairwise relationships. This visualization is critical for identifying potential outliers and understanding the interactions between features, which can impact the performance and accuracy of the machine learning models used in the study.

After removing the outliers, another box plot was created to validate the cleaned dataset (Figure 8). This step confirmed that the remaining data was more consistent and suitable for analysis.

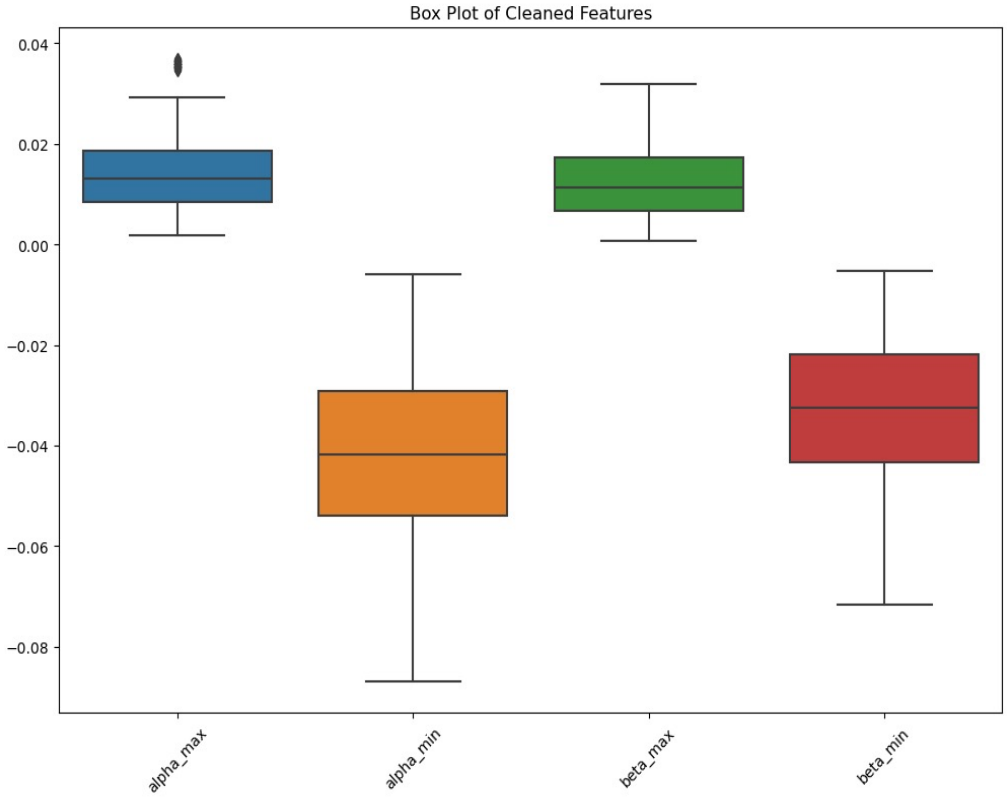

Figure 8

Figure 8: Box plot after cleaning the outliers for the whole dataset.

Feature Scaling

Following outlier removal, the cleaned data underwent standardization using the Standard Scaler. This process was essential to bring all features to a common scale, with a mean of 0 and a standard deviation of 1(Geladi & Kowalski, 1986). Proper scaling ensures that the model treats all features equally, thus improving its performance.

Limitations of Linear Regression

While linear regression was initially chosen for its simplicity and interpretability, it's important to acknowledge its inherent limitations in this context. Linear regression assumes a linear relationship between input features and output variables. However, this assumption is problematic when dealing with spectral data and protein content. The absorption process in chemical terms is inherently nonlinear, which suggests that a linear model might not capture the complexity of the relationship between spectral signals and protein secondary structure.

This inherent mismatch between the model's assumptions and the nature of the data likely contributes to suboptimal performance, as evidenced by the results (AlQuraishi, 2021).

Linear Regression Modelling

Despite the limitations, we applied the linear regression model to the pre-processed dataset. The dataset was divided into training and testing subsets, and the model was trained on the scaled features. The model's effectiveness was assessed using Mean Squared Error (MSE) and R^2 scores for both Total α and Total β protein content (Weisberg & Sanford, 2005).

Linear Regression Model Results:

- Mean Squared Error (MSE) for Total α: 73.5110

- R^2 Score for Total α: 0.6432

- Mean Squared Error (MSE) for Total β: 33.8607

- R^2 Score for Total β: 0.7807

These results indicated improved performance compared to the initial simplified model, particularly in predicting β-sheet content. However, the results also reflect the limitations of using linear regression for such complex data.

Incorporating the full dataset into our linear regression model provided a more comprehensive understanding of the relationship between spectral features and protein secondary structures. However, the inherent limitations of linear regression—particularly its assumption of linearity—suggest that this model may not fully capture the nonlinear complexities present in the data. To address these limitations, subsequent sections will explore more advanced modelling techniques, including Kernel Ridge Regression (KRR), Gradient Boosting, and Support Vector Machines (SVM which are better suited for handling the nonlinearities inherent in spectral data (Jianlin Cheng et al., 2008), (AlQuraishi, 2021).

3.1.2.3 Comprehensive Kernel Ridge Regression (KRR) Analysis

After recognizing the limitations of linear regression, particularly its inability to model non-linear relationships in the spectral data, we turned to Kernel Ridge Regression (KRR). KRR combines the principles of ridge regression with the flexibility of kernel functions, allowing the model to handle non-linear patterns by mapping the input features into a higher-dimensional space where linear regression can be effectively applied (Rutherford, Greetham, Parker, et al., 2022).

Data Utilization

Unlike the linear regression model, which was initially limited to a subset of the data, KRR was applied to the full dataset, encompassing all available spectral files for each protein. This comprehensive approach enabled the model to leverage a richer set of information, potentially capturing more subtle patterns in the spectral data. The inclusion of all available data was particularly important for a model like KRR, which relies on the detailed mapping of input features to accurately model complex relationships (Ye et al., 2020).

Hyperparameter Tuning

A critical component of KRR is the selection of hyperparameters that govern the model's behaviour. Specifically, we focused on tuning two key hyperparameters:

- **Alpha:** This parameter controls the regularization strength, which helps prevent the model from overfitting by penalizing excessively large coefficients (Kuhn & Johnson, 2013).

- **Gamma:** This parameter is associated with the RBF kernel and determines the influence of individual data points on the model. Higher values of gamma lead to tighter, more localized models, while lower values create smoother, more generalized models (Kuhn & Johnson, 2013).

To identify the optimal values for these parameters, we employed a grid search strategy. This involved systematically testing a range of values for both alpha and gamma, using cross-validation to assess the model's performance for each combination. The use of cross-validation was crucial in ensuring that the chosen hyperparameters would generalize well to unseen data, not just perform well on the training set (Kohavi, 1995).

Model Training and Performance

The KRR model was trained separately for predicting Total α and Total β content, using the RBF kernel to capture non-linearities in the data. The performance of the model was evaluated using cross-validation, specifically focusing on the mean squared error (MSE) as the primary metric. The results were promising, with the model showing a significant reduction in MSE compared to the linear regression baseline:

- **Kernel Ridge Regression Alpha Mean Squared Error:** 15.8332 ± 0.4919
- **Kernel Ridge Regression Beta Mean Squared Error:** 10.5820 ± 0.5168

These results underscore the superiority of KRR in handling non-linear relationships in spectral data, providing a more accurate and reliable prediction of protein secondary structure content.

3.1.2.4 Comprehensive Ensemble Learning Models Analysis

Following the Kernel Ridge Regression (KRR) analysis, we employed ensemble learning techniques, which combine the predictions of multiple models to improve performance and robustness. The primary models explored were Gradient Boosting, Random Forest, and Decision Tree Regressors. Ensemble methods are powerful because they reduce overfitting and improve generalization by averaging out the biases of individual models.

Data Preparation and Model Setup

The features used in this analysis were the same as in previous models: alpha_max_scaled, alpha_min_scaled, beta_max_scaled, and beta_min_scaled, while the target variables were Total α and Total β. The dataset was split into training (80%) and testing (20%) sets using a random state to ensure reproducibility (Kohavi, 1995).

Each model was initialized using the Multioutput Regressor wrapper, allowing us to predict multiple target variables simultaneously. This approach was crucial because Total α and Total β are related but distinct measures, and predicting them together can lead to better overall performance (Breiman, 2001).

Model Evaluation

1. **Gradient Boosting Regressor**:

- Gradient Boosting is a sequential ensemble method where each new model attempts to correct the errors of the previous ones. This model showed strong performance:
- **MSE for Total α**: 14.1937, **R² Score for Total α**: 0.9311
- **MSE for Total β**: 8.2785, **R² Score for Total β**: 0.9464
- These results indicate that Gradient Boosting effectively captures the complex relationships in the data, though it may still have some limitations in modelling very subtle variations (Friedman, 2001).

2. **Random Forest Regressor**:
   - Random Forest is an ensemble of Decision Trees, where each tree is built on a random subset of data and features, thus reducing variance and improving generalization (Breiman, 2001).
   - **MSE for Total α**: 1.5815, **R² Score for Total α**: 0.9923
   - **MSE for Total β**: 0.9811, **R² Score for Total β**: 0.9936
   - The Random Forest model outperformed both the Gradient Boosting and Decision Tree models, showing superior accuracy and robustness, making it the best performer in this analysis.
3. **Decision Tree Regressor**:
   - Decision Trees are simple yet powerful models that split the data into subsets based on feature values, leading to a hierarchical decision structure. (Quinlan, 1986).
   - **MSE for Total α**: 5.1288, **R² Score for Total α**: 0.9751
   - **MSE for Total β**: 2.5959, **R² Score for Total β**: 0.9832
   - While Decision Trees provided reasonable predictions, they were more prone to overfitting compared to the ensemble methods, leading to lower performance than the Random Forest and Gradient Boosting models.

The results from the ensemble learning models confirm the superiority of Random Forest in handling the complexity of the spectral data, providing both high accuracy and generalization. Gradient Boosting also performed well, but the trade-off between bias and variance was better managed by Random Forest. Decision Trees, while useful for interpretation, did not achieve the same level of performance. These findings suggest that ensemble learning methods, particularly Random Forest, are well-suited for predicting protein secondary structure content from spectral data.

3.1.2.5 Comprehensive Support Vector Machines (SVM) Analysis

Support Vector Machines (SVM) were selected as the next modelling technique to improve prediction accuracy, particularly for non-linear data. SVM is highly effective for high-dimensional datasets and tasks involving non-linear relationships, making it an ideal choice for our spectral data analysis (Vapnik, 1995).

Data Preparation and Feature Selection

The features used for this analysis were the same as in the previous models:

- alpha_max_scaled
- alpha_min_scaled
- beta_max_scaled
- beta_min_scaled

The target variables were `Total α` and `Total β`. The data was split into training (80%) and testing (20%) sets, ensuring a consistent evaluation framework across models (Xu & Goodacre, 2018).

Model Training and Hyperparameter Tuning

SVM models require careful tuning of hyperparameters to achieve optimal performance. The two key hyperparameters tuned were:

- **C**: The regularization parameter that controls the trade-off between achieving a low training error and a low testing error.
- **Gamma**: Defines how far the influence of a single training example reaches, affecting the decision boundary.

A grid search approach was utilized, testing a range of values for these hyperparameters:

- **C**: [0.1, 1, 10, 100]
- **Gamma**: [1, 0.1, 0.01, 0.001]

Cross-validation was employed to assess the model's performance, ensuring that the selected hyperparameters generalized well to unseen data (Kohavi, 1995).

The best parameters identified were:

- **For Total α**: C = 100, Gamma = 1
- **For Total β**: C = 100, Gamma = 1

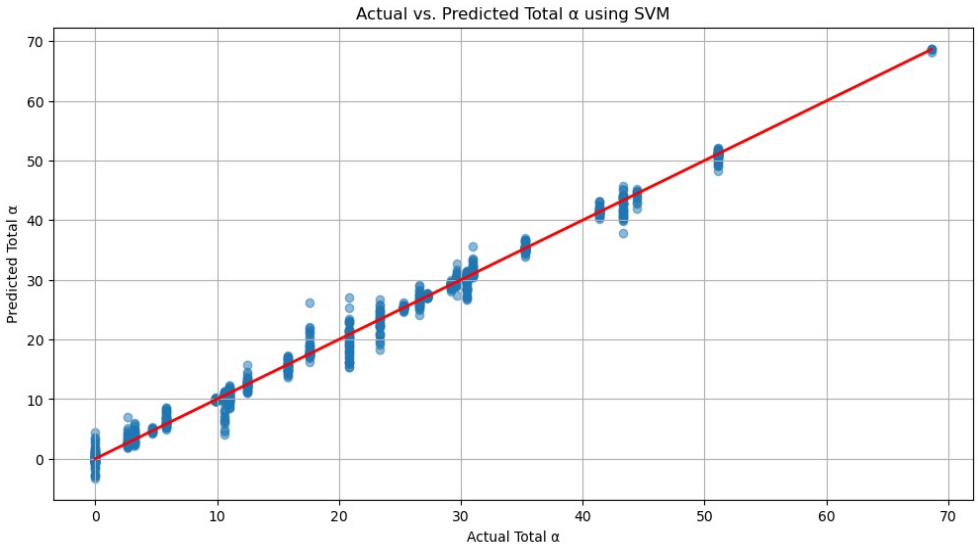

Figure 9

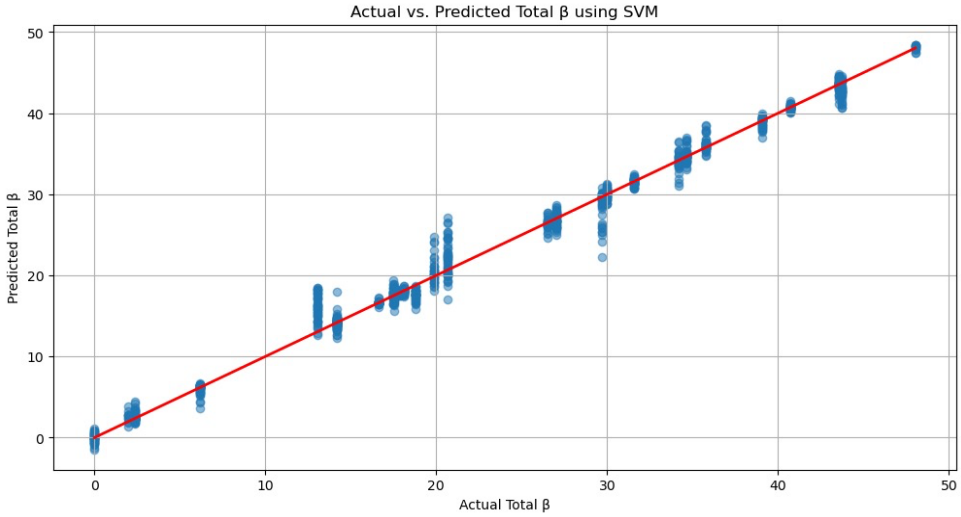

Figure 10

After training with these optimal hyperparameters, the model's predictions were plotted against the actual values (Figures 9 and 10). These plots clearly show a strong alignment along the diagonal, indicating accurate predictions for both Total α and Total β.

Model Evaluation

The final models were trained using the best hyperparameters and evaluated on the test set. The Mean Squared Error (MSE) was the primary metric used to assess the model's performance.

The results were as follows:

- **Best SVM Test MSE for Total α**: 1.7577
- **Best SVM Test MSE for Total β**: 1.6163

To further refine the model, an expanded grid search was conducted with the following parameters:

- **C**: [10, 50, 100, 200]
- **Gamma**: [0.1, 1, 10]

The expanded grid search yielded even better results:

- **Best expanded SVM Test MSE for Total α**: 0.2123
- **Best expanded SVM Test MSE for Total β**: 0.3465

These results highlighted SVM's superior ability to model the non-linear relationships in our spectral data compared to previous models like Linear Regression and Kernel Ridge Regression.

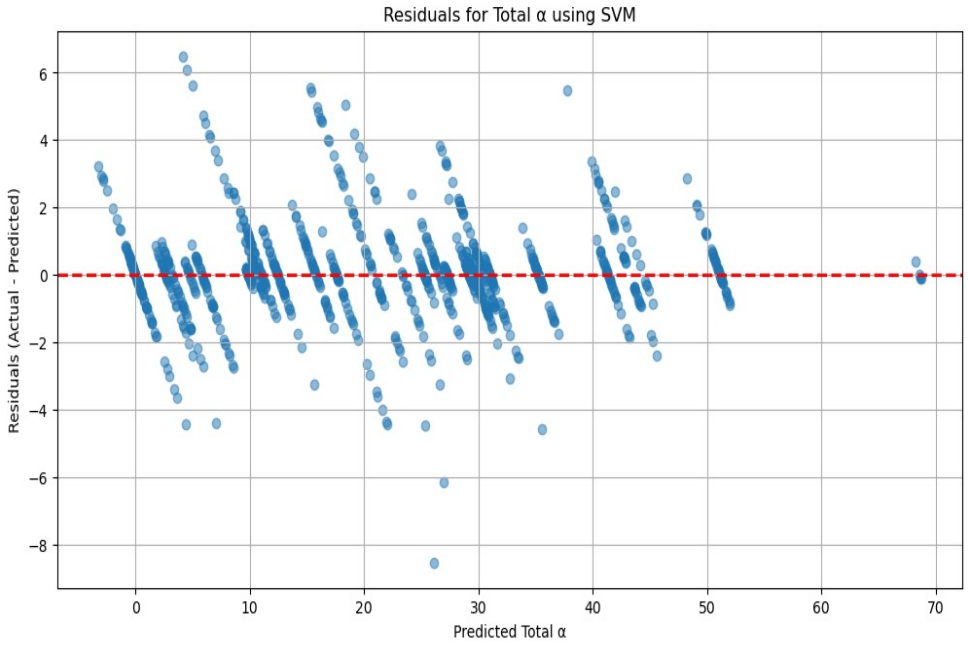

Figure 11

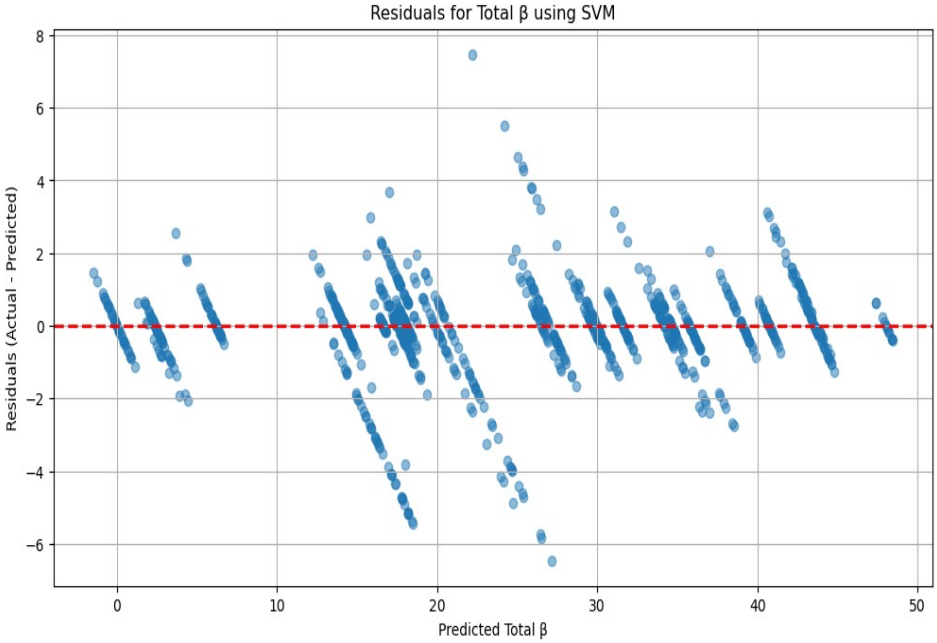

Figure 12

Residual plots (Figures 11 and 12) further confirmed the model's accuracy. These plots displayed the differences between the actual and predicted values, showing that most residuals were centred around zero, indicating minimal errors and thus high prediction accuracy.

Validation Data

To validate the model's performance, a new validation dataset was introduced, following the same feature extraction and scaling processes. When the best-expanded SVM model was applied to this validation set, the results were notably less accurate than those observed during cross-validation:

- **Validation Metrics for Total Alpha:**
  - MAE: 6.8093
  - MSE: 47.5694
  - RMSE: 6.8971
  - R^2: 0.5425
- **Validation Metrics for Total Beta:**
  - MAE: 6.5999
  - MSE: 60.9431
  - RMSE: 7.8066
  - R^2: 0.3430

Overfitting

The primary reason a model performs well on training data but poorly on unseen data is overfitting. Overfitting occurs when a model learns not just the underlying patterns but also the noise and specific quirks of the training data (Hawkins, 2004). This can happen due to:

- **High Model Complexity:** SVMs with high values of C and gamma (like C=200 and gamma=10) can tightly fit the training data, minimizing errors but making the model sensitive to slight variations in the validation data.
- **Insufficient Regularization:** The parameter C controls the balance between low training error and regularization. A very high C value can lead to insufficient regularization, enhancing the model's complexity and overfitting.

Dataset Shift

If the characteristics of the validation dataset differ significantly from those in the training and testing datasets, the model may not perform well due to:

- **Different Distributions:** The validation data might come from a different distribution or have been collected under different conditions, causing the model to struggle because it has learned features that don't generalize well outside the training set.

- **Changes Over Time:** If the data evolves over time (known as concept drift), models trained on historical data might not perform well on newer data (Widmer & Kubat, 1996).

These points highlight the challenges of model generalization, especially when applying a model trained on one dataset to a new one. Further refinement and additional data may be needed to improve the model's ability to generalize.

3.1.2.6 Neural Networks (MLP) Analysis

In our exploration to enhance prediction accuracy for Total α and Total β content in proteins, we employed a Neural Network model, specifically a Multi-Layer Perceptron (MLP). The MLP is well-suited for complex, non-linear relationships within the data due to its ability to model interactions between features through hidden layers (AlQuraishi, 2021).

Data Preparation and Feature Selection

The dataset used for training the MLP model included all the relevant features except the non-feature columns, 'Protein', 'Total α', and 'Total β'. The target variables for prediction were 'Total α' and 'Total β'. The dataset was split into training (80%) and testing (20%) sets, a consistent approach used across all models (Voronina et al., 2021).

Model Training and Hyperparameter Configuration

The MLP model was configured with two hidden layers containing 50 and 25 neurons, respectively. The activation function used was ReLU (Rectified Linear Unit), which introduces non-linearity, allowing the network to model more complex patterns (Senior et al., 2020). The model was trained using the Adam optimizer, an adaptive learning rate optimization algorithm, with an initial learning rate of 0.0001 and a regularization parameter (alpha) of 0.01. The model was set to train for up to 10,000 iterations to ensure convergence (Friedman, 2001).

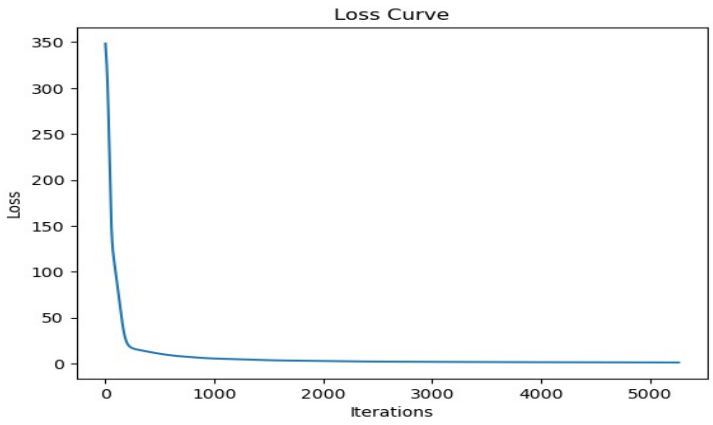

Figure 13

After training, the model was evaluated on the test set, yielding a Mean Squared Error (MSE) of **2.4470**, indicating a reasonable level of prediction accuracy. The loss curve (Figure 13) shows the gradual decrease in loss over iterations, reflecting the model's learning process.

Cross-Validation and Model Performance

To further validate the model, we employed 5-fold cross-validation, a technique that splits the data into five parts, training the model on four and testing on the fifth, rotating through all parts. This approach helps ensure that the model's performance is consistent across different subsets of the data.

The cross-validation results produced the following MSE scores:

- **CV MSE Scores:** [2.2602, 1.8620, 2.5621, 2.5158, 2.3401]
- **Average CV MSE:** 2.3080

These scores confirm that the model performs consistently well across different splits of the data.

Model Evaluation on the Validation Set

To evaluate the model's performance in a real-world scenario, predictions were made on a separate validation set. The predicted Total α and Total β values for three proteins, along with their actual values, are summarized as follows:

Protein	Predicted Total Alpha	Actual Total Alpha	Predicted Total Beta	Actual Total Beta
Chymotrypsinogen A (2cga)	7.3188	7.35	37.8360	32.24
DT-diaphorase (1d4a)	24.2736	28.94	20.0674	11.36
Ovalbumin (1ova)	25.2332	29.02	21.9618	31.31

The metrics calculated for these predictions revealed the following:

- **Mean Absolute Error (MAE) for Total Alpha:** 2.8281
- **Mean Absolute Error (MAE) for Total Beta:** 7.8838
- **Mean Squared Error (MSE) for Alpha:** 12.0387
- **Root Mean Squared Error (RMSE) for Alpha:** 3.4697
- **R^2 for Alpha:** 0.8842
- **Mean Squared Error (MSE) for Beta:** 64.8407
- **Root Mean Squared Error (RMSE) for Beta:** 8.0524
- **R^2 for Beta:** 0.3010

Analysis and Conclusion

The results show that the MLP model performs reasonably well for predicting Total α content with an R^2 of 0.8842, indicating a good fit. However, the model struggles with predicting Total β content, as reflected by the lower R^2 of 0.3010 and higher error metrics (Senior et al., 2020).

The performance differences between α and β predictions might be due to the complexity of the relationships within the β content, which the current model configuration could not fully

capture. These results suggest that while Neural Networks provide a powerful tool for complex data modelling, further optimization or alternative approaches may be needed to improve β content predictions.

The overall findings emphasize the importance of selecting appropriate model architectures and hyperparameters tailored to the specific data characteristics to achieve optimal performance.

Chapter 4: Discussion & Future Works

This study represents a pioneering effort in integrating two-dimensional infrared (2D-IR) spectroscopy with machine learning to analyse protein secondary structures, particularly focusing on the prediction of α-helix and β-sheet content. Our research explored a range of machine learning models, including linear regression, kernel ridge regression (KRR), support vector machines (SVM), random forests, gradient boosting, and neural networks, to map the complex relationships inherent in 2D-IR spectral data to the secondary structural elements of proteins (Rutherford, Greetham, Parker, et al., 2022, p. 202). The findings of this research have significant implications for both the fields of spectroscopy and computational biology, marking a substantial step forward in the application of machine learning techniques to protein structure analysis.

Model Performance and Key Insights:

- The ensemble learning models, particularly Random Forest, emerged as the most effective in predicting the secondary structure components, demonstrating a robust ability to handle the non-linear, complex nature of the spectral data. Random Forest's capacity to aggregate the predictions from multiple decision trees mitigated the risk of overfitting, which was a notable challenge in other models such as SVM and neural networks (Quinlan, 1986).
- Linear regression, while providing a straightforward and interpretable baseline, was limited in its ability to capture the non-linear relationships between the spectral features and the protein structures. This limitation highlighted the necessity for more advanced models like KRR and SVM, which are designed to address non-linearity and complex data interactions (Montgomery & Peck, 2012). However, these advanced models also encountered challenges, particularly with overfitting when handling high-dimensional features (Senior et al., 2020).

- The study underscored the importance of feature selection and preprocessing, specifically the role of normalization and standardization. The spectral data, characterized by varying scales and magnitudes, required robust standardization to prevent the disproportionate influence of certain features on the model's learning process (Kuhn & Johnson, 2013). This was crucial for improving the models' performance and ensuring the stability of predictions across different experimental conditions.

Challenges Encountered:

- One of the significant challenges encountered in this study was overfitting, particularly with more complex models like SVM and neural networks. This issue highlighted the delicate balance between model complexity and generalization, emphasizing the need for careful model selection and hyperparameter tuning to avoid learning noise rather than meaningful patterns (Dietterich, 2000).
- Another challenge was the difficulty in accurately predicting β-sheet content, as opposed to α-helix content. The complexity of β-sheets, involving intricate hydrogen-bonding patterns and the potential for parallel and antiparallel arrangements, likely contributed to this discrepancy. This suggests that while our models were effective for certain structural elements, they struggled with the nuanced features of β-sheets, indicating a potential area for further refinement (Kabsch & Sander, 1983).

Advanced Neural Networks and Convolutional Approaches: This study represents the initial foray into applying 2D-IR spectroscopy and machine learning for protein structure analysis. While we explored simpler models to avoid overfitting, there is significant potential in investigating more advanced neural network architectures, such as convolutional neural networks (CNNs), tailored specifically to the structure of spectral data (AlQuraishi, 2021). However, given the complexity and potential for overfitting with CNNs, especially with the relatively small feature set used in this study, we opted to focus on models that provided a good balance between simplicity and performance. This decision underscores the importance of matching the model complexity to the problem at hand, ensuring that we don't over-complicate the solution.

4.1 Future works

Building on the findings of this study, several avenues can be pursued to further enhance the application of 2D-IR spectroscopy and machine learning in protein structure analysis:

Exploration of Principal Component Analysis (PCA):

- While this study did not incorporate PCA, future research could benefit from this dimensionality reduction technique. PCA could help identify the most significant features from the spectral data, potentially improving the performance of machine learning models by focusing on the most informative aspects of the data (Widmer & Kubat, 1996). PCA could also assist in visualizing the spectral data's underlying structure, providing deeper insights into the relationships between different protein features.

Application of Advanced Neural Networks:

- Future research could explore the application of advanced neural network architectures, such as CNNs or even recurrent neural networks (RNNs), particularly in cases where the spectral data might benefit from a more sophisticated analysis of its spatial or temporal dimensions (AlQuraishi, 2021). However, it is crucial to balance model complexity with the risk of overfitting. This study lays the groundwork for such future exploration, ensuring that any increase in model complexity is justified by a corresponding improvement in predictive accuracy and robustness.

Development of Hybrid Models:

- The creation of hybrid models that combine the strengths of different machine learning techniques could provide a more comprehensive analysis of protein structures. For example, integrating ensemble methods with neural networks could yield models that balance the robustness of ensemble methods with the deep learning capabilities of neural networks, potentially leading to improved accuracy in complex structure prediction task.

Refined Feature Engineering:

- Future studies should focus on refining the feature engineering process, potentially exploring more sophisticated spectral features that could capture subtle aspects of protein secondary structure. This might involve experimenting with different spectral

preprocessing techniques or leveraging machine learning to discover new features that are strongly correlated with specific structural elements (Kuhn & Johnson, 2013).

Enhanced Model Validation Strategies:

- Another avenue for future work involves enhancing model validation strategies. While our study used standard cross-validation methods, future research could explore more complex validation techniques that better account for the unique characteristics of protein spectral data (Xu & Goodacre, 2018). Techniques such as nested cross-validation or Bayesian optimization could be used to fine-tune model parameters more effectively, ensuring that the models generalize well across different datasets and experimental conditions.

Chapter 5: Conclusion

In conclusion, this research marks a significant advancement in the application of 2D-IR spectroscopy combined with machine learning for protein structure analysis. The integration of these techniques has demonstrated the potential for detailed, accurate predictions of protein secondary structures, specifically α-helix and β-sheet content, within physiologically relevant environments. Through rigorous evaluation of various machine learning models, this study identified Random Forest as a particularly effective approach for handling the complexities of spectral data.

While challenges such as overfitting and the accurate prediction of β-sheets were encountered, the study also highlighted the importance of feature selection, preprocessing, and the careful balancing of model complexity. The potential of advanced neural networks, although not fully explored in this study, represents an exciting avenue for future research, particularly as the field continues to evolve.

The use of PCA to identify key points of interest in the spectral data, combined with advanced interpolation techniques, offers a promising direction for future studies aiming to refine feature extraction processes and enhance model accuracy. This research lays the groundwork for more sophisticated analyses that can delve deeper into the molecular intricacies of proteins.

Ultimately, this work not only contributes to our understanding of protein structures but also sets the stage for broader applications in structural biology, drug discovery, and beyond. The journey from spectral data to structural insight is complex, but with the tools and approaches

developed in this study, we are well-positioned to continue unravelling the molecular mysteries that underpin life itself.

6. References

AlQuraishi, M. (2021). Machine learning in protein structure prediction. *Current Opinion in Chemical Biology*, *65*, 1–8. https://doi.org/10.1016/j.cbpa.2021.04.005

Barth, A. (2007). Infrared spectroscopy of proteins. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, *1767*(9), 1073–1101. https://doi.org/10.1016/j.bbabio.2007.06.004

Barth, A., & Zscherp, C. (2002). What vibrations tell about proteins. *Quarterly Reviews of Biophysics*, *35*(4), 369–430. https://doi.org/10.1017/S0033583502003815

Bellisola, G., & Sorio, C. (n.d.). *Infrared spectroscopy and microscopy in cancer research and diagnosis*.

Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In G. Goos, J. Hartmanis, & J. Van Leeuwen (Eds.), *Multiple Classifier Systems* (Vol. 1857, pp. 1–15). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

Dill, K. A., & MacCallum, J. L. (2012). The Protein-Folding Problem, 50 Years On. *Science*, *338*(6110), 1042–1046. https://doi.org/10.1126/science.1219021

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Fritzsch, R., Hume, S., Minnes, L., Baker, M. J., Burley, G. A., & Hunt, N. T. (2020). Two-dimensional infrared spectroscopy: An emerging analytical tool? *The Analyst*, *145*(6), 2014–2024. https://doi.org/10.1039/C9AN02035G

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, *185*, 1–17. https://doi.org/10.1016/0003-2670(86)80028-9

Haris, P. I., & Chapman, D. (1995). The conformational analysis of peptides using fourier transform IR spectroscopy. *Biopolymers*, *37*(4), 251–263. https://doi.org/10.1002/bip.360370404

Hawkins, D. M. (2004). The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1–12. https://doi.org/10.1021/ci0342472

Hume, S., Hithell, G., Greetham, G. M., Donaldson, P. M., Towrie, M., Parker, A. W., Baker, M. J., & Hunt, N. T. (2019). Measuring proteins in H_2O with 2D-IR spectroscopy. *Chemical Science*, *10*(26), 6448–6456. https://doi.org/10.1039/C9SC01590F

Hunt, N. T. (2009). 2D-IR spectroscopy: Ultrafast insights into biomolecule structure and function. *Chemical Society Reviews*, *38*(7), 1837. https://doi.org/10.1039/b819181f

Hunt, N. T. (2024a). Biomolecular infrared spectroscopy: Making time for dynamics. *Chemical Science*, *15*(2), 414–430. https://doi.org/10.1039/D3SC05223K

Hunt, N. T. (2024b). Using 2D-IR Spectroscopy to Measure the Structure, Dynamics, and Intermolecular Interactions of Proteins in H_2O. *Accounts of Chemical Research*, *57*(5), 685–692. https://doi.org/10.1021/acs.accounts.3c00682

Jianlin Cheng, Tegge, A. N., & Baldi, P. (2008). Machine Learning Methods for Protein Structure Prediction. *IEEE Reviews in Biomedical Engineering*, *1*, 41–49. https://doi.org/10.1109/RBME.2008.2008239

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. https://doi.org/10.1002/bip.360221211

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:2702042

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Mckenna, A., & Dubey, S. (2022). Machine learning based predictive model for the analysis of sequence activity relationships using protein spectra and protein descriptors. *Journal of Biomedical Informatics*, *128*, 104016. https://doi.org/10.1016/j.jbi.2022.104016

Montgomery, D. C., & Peck, E. A. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.

Muggleton, S., King, R. D., & Sternberg, M. J. E. (n.d.). *Protein secondary structure prediction using logic-based machine learning*.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/BF00116251

Ren, H., Zhang, Q., Wang, Z., Zhang, G., Liu, H., Guo, W., Mukamel, S., & Jiang, J. (2022). Machine learning recognition of protein secondary structures based on two-dimensional spectroscopic descriptors. *Proceedings of the National Academy of Sciences*, *119*(18), e2202713119. https://doi.org/10.1073/pnas.2202713119

Rutherford, S. H., Greetham, G. M., Parker, A. W., Nordon, A., Baker, M. J., & Hunt, N. T. (2022). Measuring proteins in H2O using 2D-IR spectroscopy: Pre-processing steps and applications toward a protein library. *The Journal of Chemical Physics*, *157*(20), 205102. https://doi.org/10.1063/5.0127680

Rutherford, S. H., Greetham, G. M., Towrie, M., Parker, A. W., Kharratian, S., Krauss, T. F., Nordon, A., Baker, M. J., & Hunt, N. T. (2022). Detection of paracetamol binding to albumin in blood serum using 2D-IR spectroscopy. *The Analyst*, *147*(15), 3464–3469. https://doi.org/10.1039/D2AN00978A

Rutherford, S. H., Hutchison, C. D. M., Greetham, G. M., Parker, A. W., Nordon, A., Baker, M. J., & Hunt, N. T. (2023). Optical Screening and Classification of Drug Binding to

Proteins in Human Blood Serum. *Analytical Chemistry*, *95*(46), 17037–17045. https://doi.org/10.1021/acs.analchem.3c03713

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

Shim, S.-H., Strasfeld, D. B., Ling, Y. L., & Zanni, M. T. (2007). Automated 2D IR spectroscopy using a mid-IR pulse shaper and application of this technology to the human islet amyloid polypeptide. *Proceedings of the National Academy of Sciences*, *104*(36), 14197–14202. https://doi.org/10.1073/pnas.0700804104

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer New York. https://doi.org/10.1007/978-1-4757-2440-0

Voronina, L., Leonardo, C., Mueller-Reif, J. B., Geyer, P. E., Huber, M., Trubetskov, M., Kepesidis, K. V., Behr, J., Mann, M., Krausz, F., & Žigman, M. (2021). Molecular Origin of Blood-Based Infrared Spectroscopic Fingerprints**. *Angewandte Chemie International Edition*, *60*(31), 17060–17069. https://doi.org/10.1002/anie.202103272

Wardah, W., Khan, M. G. M., Sharma, A., & Rashid, M. A. (2019). Protein secondary structure prediction using neural networks and deep learning: A review. *Computational Biology and Chemistry*, *81*, 1–8. https://doi.org/10.1016/j.compbiolchem.2019.107093

Weisberg, & Sanford. (2005). *Applied Linear Regression* (3rd ed.). Wiley.

Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning*, *23*(1), 69–101. https://doi.org/10.1007/BF00116900

Xu, Y., & Goodacre, R. (2018). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the

Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*, *2*(3), 249–262. https://doi.org/10.1007/s41664-018-0068-2

Yang, S., Zhang, Q., Yang, H., Shi, H., Dong, A., Wang, L., & Yu, S. (2022). Progress in infrared spectroscopy as an efficient tool for predicting protein secondary structure. *International Journal of Biological Macromolecules*, *206*, 175–187. https://doi.org/10.1016/j.ijbiomac.2022.02.104

Ye, S., Zhong, K., Zhang, J., Hu, W., Hirst, J. D., Zhang, G., Mukamel, S., & Jiang, J. (2020). A Machine Learning Protocol for Predicting Protein Infrared Spectra. *Journal of the American Chemical Society*, *142*(45), 19071–19077. https://doi.org/10.1021/jacs.0c06530