

Politecnico di Torino

Master's Degree in Computer Engineering

Specialization in AI & Data Analytics

Academic Year 2025/2026

Graduation Session September 2026

Self-Supervised Pretraining and Transfer Learning for Biomedical Classification from Raman Data

Supervisors:

Prof. Lia Morra

Prof. Alfonso Rodríguez-Patón

Candidate:

Alex Ferroni

Abstract

Raman spectroscopy provides label-free molecular fingerprints of biological samples and is therefore a promising tool for rapid biomedical diagnosis. However, deep learning applications in this field are limited by the scarcity of large annotated datasets, since reliable clinical labels are expensive and difficult to obtain. This thesis investigates self-supervised pre-training as a strategy to learn transferable Raman representations from unlabeled spectra and reduce the dependence on task-specific labels.

Two complementary self-supervised approaches were studied. Barlow Twins was adapted to one-dimensional Raman spectra to learn representations invariant to realistic perturbations such as noise, scaling, and baseline variations. In parallel, a Spectral Masked Autoencoder (SMAE) was implemented to reconstruct masked spectral regions and capture local-to-global biochemical structure. The resulting pre-trained encoders were transferred to downstream biomedical classification tasks, including inflammatory bowel disease plasma spectra and BACTERIA-ID pathogen identification.

The experiments show that self-supervised initialization generally improves downstream performance compared with training the same architectures from scratch. SMAE achieved the strongest results on the IBD task, while Barlow Twins provided a robust full-spectrum representation for bacterial identification. To assess whether the best IBD model relied on meaningful spectral evidence, Integrated Gradients and SHAP were used for interpretability analysis. The explanations highlighted localized Raman bands associated with plausible protein, carotenoid, lipid/carboxylate, aromatic-amino-acid, and phosphate or saccharide-related contributions.

Overall, this thesis demonstrates that self-supervised Raman encoders can provide transferable, data-efficient, and interpretable representations under limited-label conditions. Future work will focus on adapting these models to labeled breast cancer tissue data as they become available, particularly within the SPECTRA-BREAST project, and on improving architecture design, hyperparameter tuning, validation protocols, and explainability methods for clinical translation.

Contents

Abstract **i**

1 Introduction **1**
1.1 The Clinical Problem: Breast-Conserving Surgery 1
1.2 Principles of Raman Spectroscopy 1
1.3 Raman Spectroscopy for Biomedical Diagnostics 3
1.4 The SPECTRA-BREAST Solution 3
1.5 Challenges in AI for Spectroscopy: Pre-training and Interpretability . 4
1.6 Objectives of the Thesis 4
1.7 Thesis Structure 5

2 State of the Art **6**
2.1 From Chemometrics to Deep Learning in Spectral Analysis 6
2.2 Convolutional Architectures for Raman Spectroscopy 6
2.2.1 Standard 1D-CNNs 7
2.2.2 ResNet-based Developments 7
2.3 Transformers and Self-Attention 7
2.4 Self-Supervised Pre-training for Spectral Data 8
2.4.1 Contrastive Learning: SimCLR and Barlow Twins 8
2.4.2 Masked Autoencoders and Foundation Models 9
2.5 Interpretability for Biomarker Discovery 10
2.6 Summary and Motivation for the Thesis 11

3 Dataset Preparation and Preprocessing **12**
3.1 Introduction to Data Collection 12
3.2 Pre-Training Corpus Components 13
3.2.1 Dataset 1: MDA-MB-231 (DeepR Pre-training Corpus for Denoising) 13
3.2.2 Dataset 2: Cancer Metastasis (Single-Cell Resolution - "Cell" Dataset) 14
3.2.3 Dataset 3A: Cancer Metastasis / TWIST (In Vivo Mouse Models) 15
3.2.4 Dataset 3B: Immunotherapy Response (IMMUNO) 16
3.2.5 TropHY Dataset: Intraoperative Tissue Mappings 16
3.2.6 SpectraBreast: Highly Raw (RAW) Raman Maps and Isolated Pre-Processing 17

3.3 Datasets for Downstream Classification and Fine-Tuning 20
3.3.1 IBD Dataset: Blood Plasma Liquid Biopsy 20
3.3.2 BACTERIA-ID Dataset: Open-Source Pathogen Identification 20
3.4 Theoretical Background on Raman Preprocessing 21
3.5 Spectral Processing and Data Harmonization 22
3.5.1 Uniformation Philosophy and Feature Space Harmonization . 22
3.5.2 Amplitude-Dependent Normalization Strategies 23
3.6 Summary of the Pre-Training Corpus 25

4 Pre-Training Foundation Models for Raman Spectroscopy 28
4.1 Introduction to Self-Supervised Learning 28
4.2 Representation Learning and Transfer Learning Rationale 28
4.3 Barlow Twins for Spectral Invariance 29
4.3.1 Backbone Selection: 1D-ResNet Variants 30
4.3.2 Domain-Specific Spectral Augmentations 32
4.3.3 The Projector Network and Cross-Correlation Objective . . . 35
4.3.4 Barlow Twins: Loss Convergence 35
4.4 Masked Autoencoders (MAE) for Spectral Reconstruction 37
4.4.1 1D Patching, Masking, and Positional Embeddings 37
4.4.2 Selected Best MAE Configuration 39
4.4.3 The Asymmetric Architecture: Encoder vs. Decoder 40
4.4.4 Pre-Training Objective and Denoising Properties 40
4.5 Pre-Training Dynamics and Temporal Reconstruction 41
4.5.1 MAE: Training and Spectral Reconstruction 41

5 Downstream Fine-Tuning and Experimental Results 48
5.1 Fine-Tuning Methodology 48
5.1.1 Architectural Adaptation 48
5.1.2 Optimization Strategy 49
5.1.3 Training-Time Augmentation 50
5.2 Evaluation Metrics and Diagnostic Reading 50
5.3 Inflammatory Bowel Disease (IBD) Classification 51
5.3.1 Task Definition and Evaluation Protocol 51
5.3.2 Fine-Tuning Strategies and Data Augmentation 52
5.3.3 Hold-Out Evaluation 53
5.3.4 Training Curves, Confusion Matrices, and Probability Diagnostics 54
5.3.5 Fixed-Test Cross-Validation Results 54

5.3.6 Model-Level Interpretation 59
5.4 Pathogen Identification (BACTERIA-ID) 60
5.4.1 Task Definition and Evaluation Protocol 60
5.4.2 BACTERIA-ID Fine-Tuning Setup 61
5.4.3 Fold-Based Results 61
5.4.4 Training Dynamics 62
5.4.5 Confusion Matrix and Best-Model Analysis 63
5.4.6 Bacteria Result Interpretation 63
5.5 Discussion 67

6 Model Interpretability and Explainable AI (XAI) 69
6.1 Purpose of the Interpretability Analysis 69
6.2 Explainability Methods 70
6.2.1 Integrated Gradients 70
6.2.2 SHAP 70
6.3 Single-Spectrum Explanations 70
6.4 Class-Level Average Explanations 72
6.5 Top Raman Peaks Identified by IG and SHAP 73
6.6 Literature-Based Biomarker Assignment 74
6.7 SHAP Directionality 75
6.8 Biochemical Interpretation 75
6.9 Discussion 77

7 Conclusion and Future Perspectives 79
7.1 Summary of Contributions 79
7.2 Limitations 80
7.3 Future Perspectives 81

Chapter 1

Introduction

1.1 The Clinical Problem: Breast-Conserving Surgery

Breast-conserving surgery (BCS) is a primary surgical treatment aiming to remove cancerous tissue while preserving healthy breast tissue [1]. However, achieving clear surgical margins remains a major clinical challenge. The majority of studies report positive or close resection margins in 20% to 40% of patients who undergo the initial procedure [1], [2]. This high rate of positive margins necessitates additional re-excision surgeries to prevent local tumor recurrence.

The need for a second operation significantly increases healthcare costs and inflicts severe emotional and physical distress on patients [3]. Current intraoperative margin assessment methods are often slow, subjective, and lack the required accuracy, highlighting the urgent need for fast, reliable, and objective technologies to guide surgeons during the initial operation.

1.2 Principles of Raman Spectroscopy

Raman spectroscopy is an optical technique that probes the molecular composition of a sample by measuring inelastic light scattering. When monochromatic light, usually generated by a laser, interacts with matter, most photons are elastically scattered and preserve the same energy as the incident beam. This dominant process is known as Rayleigh scattering. A much smaller fraction of photons exchanges energy with the vibrational modes of the molecules in the sample. These photons are scattered with a slightly different energy, producing the Raman effect. The difference between the incident and scattered photon energy is reported as a Raman shift, commonly expressed in cm^{-1}, and corresponds to specific molecular vibrations.

The central value of Raman spectroscopy in biomedical applications is that each biological sample produces a spectrum that can be interpreted as a molecular fingerprint. Peaks in the spectrum are associated with vibrational modes of biochemical constituents such as proteins, lipids, nucleic acids, carbohydrates, pigments, and other metabolites. For this reason, Raman spectroscopy does not require external labels or staining agents: the contrast comes directly from the intrinsic chemistry of the sample. This makes the technique particularly attractive for medical diagnostics,

where preserving tissue integrity and minimizing sample preparation are important requirements [4], [5].

A Raman spectrum is usually represented as intensity as a function of Raman shift. The exact position, height, and shape of the peaks depend on the molecular composition and physical state of the sample. In biological tissue, for example, protein-rich regions may show strong Amide-related bands, lipid-rich regions may emphasize CH stretching and deformation modes, and nucleic-acid or carbohydrate contributions may appear in specific fingerprint regions. Since disease processes often alter cell metabolism, extracellular matrix organization, lipid content, protein conformation, or inflammatory status, these biochemical changes can be reflected in the Raman spectrum. The diagnostic task therefore becomes a pattern-recognition problem: identifying whether the spectral fingerprint corresponds to healthy tissue, malignant tissue, inflammatory plasma, or a bacterial class.

For biomedical Raman analysis, the so-called fingerprint region is especially important. This region, approximately between 650 and 1750 cm^{-1}, contains many diagnostically relevant vibrational bands related to proteins, lipids, amino acids, nucleic acids, and carbohydrates. Although higher-wavenumber regions may also contain useful information, the fingerprint interval is often preferred because it provides a dense biochemical description of the sample. In this thesis, this region plays a central role in the Spectral Masked Autoencoder experiments and in the interpretability analysis, where attribution peaks are mapped back to physically meaningful Raman shifts.

Despite its advantages, Raman spectroscopy also presents important technical challenges. Raman scattering is intrinsically weak, so spectra can be affected by low signal-to-noise ratio, fluorescence background, baseline drift, cosmic spikes, and instrument-dependent variations. Biological samples add further complexity because they are heterogeneous, and their spectra often contain overlapping contributions from many molecular species. A single Raman band is rarely unique to one molecule, especially in plasma or tissue. For this reason, biomedical interpretation must be cautious: Raman peaks can suggest biochemical families or molecular regions, but they should not be overinterpreted as definitive molecular identification without additional validation.

These challenges explain why preprocessing and machine learning are so important in modern Raman spectroscopy. Before spectra can be used by a model, they often require interpolation onto a common spectral axis, normalization, baseline management, and careful range selection. Once standardized, the spectra can be analyzed by deep learning models capable of capturing subtle multivariate pat-

terns that may be difficult to identify manually. However, the same complexity also creates the risk that a model learns instrumental artifacts or dataset-specific biases. This motivates the two main methodological directions of this thesis: self-supervised pre-training, to learn robust and transferable spectral representations from large unlabeled datasets, and explainable AI, to verify whether the resulting classifiers focus on plausible Raman bands rather than noise.

1.3 Raman Spectroscopy for Biomedical Diagnostics

To overcome the limitations of traditional pathology, Raman spectroscopy has emerged as a highly informative, optical molecular fingerprinting tool [5]. Based on the inelastic scattering of monochromatic light, this technique is label-free, non-destructive, and highly sensitive to biochemical composition.

In oncological applications, Raman signals can detect subtle biochemical variations in tissue. Changes in the lipid-to-protein ratio and specific alterations in collagen structures and Amide bands act as indicators of malignant transformation, assisting in the differentiation between normal, benign, and malignant tissues [4]. Because it provides a real-time, molecular-level snapshot of the tissue without requiring exogenous dyes, Raman spectroscopy is uniquely positioned to assist in rapid intraoperative diagnostics.

1.4 The SPECTRA-BREAST Solution

To address the clinical need for clear surgical margins, the European SPECTRA-BREAST project aims to revolutionize tumor resection assessment during BCS [6]. The project introduces an innovative, multimodal technology combining Hyperspectral Imaging (HSI) and fiber-optic Raman Spectroscopy with advanced Artificial Intelligence (AI) algorithms.

In the proposed workflow, an AI algorithm first analyzes HSI data to rapidly scan the tissue and point out potentially positive margins. Subsequently, these flagged areas are analyzed by a Raman probe. A secondary AI model processes the Raman signals to definitively identify true positive margins with high specificity, offering real-time, highly accurate guidance in the operating room.

1.5 Challenges in AI for Spectroscopy: Pre-training and Interpretability

Developing robust deep learning models to classify medical spectral data presents significant challenges. Standard supervised approaches, such as 1D Convolutional Neural Networks (CNNs), require massive amounts of expertly annotated data to avoid overfitting. However, acquiring labeled clinical datasets is incredibly expensive and hindered by inter-patient variability [7].

To overcome this data scarcity bottleneck, modern AI is shifting towards *Self-Supervised Learning (SSL)*. By pre-training foundation models on large corpora of unlabelled Raman spectra, the model can implicitly learn the underlying "physics" and structural dependencies of Raman scattering. This generalized knowledge can then be fine-tuned on specific clinical tasks. Furthermore, in clinical settings, "black-box" predictions are insufficient. It is crucial to explore Explainable AI (XAI) techniques to ensure that the models are relying on genuine biochemical biomarkers rather than instrumental artifacts [8].

1.6 Objectives of the Thesis

This Master's Thesis focuses on testing, evaluating, and interpreting modern deep learning architectures for Raman spectral analysis. Rather than solely building models from scratch, the primary goal is to systematically explore pre-training algorithms and self-supervised foundation models, assessing their viability for clinical data. The specific objectives are as follows:

- **Data Standardization:** To aggregate and preprocess a large-scale, heterogeneous corpus of biomedical Raman spectra (both public unlabelled data and proprietary clinical datasets), mapping them onto a common spectral axis.
- **Baseline Evaluation:** To implement and evaluate standard supervised models, particularly 1D-CNNs, to establish a robust performance baseline for breast cancer tissue classification.
- **Testing Pre-Trained Architectures:** To explore and evaluate state-of-the-art self-supervised pre-training paradigms (such as Masked Autoencoders and Vision Transformers). The objective is to test whether transferring representations from unlabelled data mitigates clinical data scarcity.

- **Biomarker Discovery via XAI:** To explore interpretability techniques—specifically Integrated Gradients (IG) and SHapley Additive exPlanations (SHAP)—to map the networks' mathematical weights back to physical vibrational bands, validating the models' clinical reliability.

1.7 Thesis Structure

The remainder of this document is organized as follows. **Chapter 2** reviews the state-of-the-art of deep learning applied to spectral analysis. **Chapter 3** details the data collection and the mathematical preprocessing pipeline. **Chapter 4** presents the core methodology, detailing the supervised baselines and the tested self-supervised pre-training algorithms. **Chapter 5** discusses the experimental setup, fine-tuning strategies, and clinical classification results. **Chapter 6** delves into Explainable AI, utilizing interpretation frameworks to validate the models' biochemical findings. Finally, **Chapter 7** draws the conclusions of this work.

Chapter 2

State of the Art

2.1 From Chemometrics to Deep Learning in Spectral Analysis

The analysis of Raman spectra has traditionally relied on chemometric and classical machine learning approaches. Methods such as Principal Component Analysis coupled with Linear Discriminant Analysis (PCA-LDA), Partial Least Squares Discriminant Analysis (PLS-DA), Random Forests, and Support Vector Machines (SVM) have been widely used for decades to classify tissues, identify chemical compounds, and detect pathogens [5].

While these classical methods are highly interpretable and well-understood by clinicians, they pose several practical and computational challenges. They rely entirely on handcrafted pipelines, requiring extensive manual preprocessing—including baseline correction, smoothing, and feature selection—to extract meaningful data [5]. Furthermore, they struggle significantly when the discriminative structure of the problem is highly non-linear or when the necessary biochemical markers are distributed across complex, overlapping, and noisy spectral regions [7].

To overcome these limitations, Deep Learning (DL) has progressively emerged as the dominant paradigm in biomedical spectroscopy. Deep networks automate the feature extraction process, learning task-specific hierarchical representations directly from raw or minimally preprocessed spectra. This drastically reduces the reliance on manually engineered descriptors, allowing models to automatically identify the most predictive structural features for complex diagnostic tasks [5].

2.2 Convolutional Architectures for Raman Spectroscopy

Because Raman spectra can be naturally treated as continuous sequences of intensities along the wavenumber axis, one-dimensional convolutional neural networks (1D-CNNs) were among the earliest and most successful deep architectures applied to these signals [9].

2.2.1 Standard 1D-CNNs

At their core, 1D-CNNs operate by sliding learnable mathematical filters (kernels) across the spectral sequence. These filters act as highly specialized feature detectors, activating when they encounter specific local patterns such as peak widths, curve slopes, and the neighborhood relationships between adjacent wavenumbers. Unlike manual peak-picking, the network learns which filter shapes are most useful for the specific clinical task.

In biomedical applications, CNN-based pipelines have consistently shown strong performance in supervised settings. For example, previous works have successfully utilized 1D-CNNs to distinguish breast tissue categories, cancer cell lines, and molecular subtypes from Raman spectra, establishing them as a robust and reliable reference baseline in oncology [10], [11].

2.2.2 ResNet-based Developments

As models grew deeper in an attempt to capture more complex biological patterns, they encountered the vanishing gradient problem, where learning signals decay before reaching the initial layers of the network. To solve this, researchers moved toward residual architectures, originally inspired by the 2D ResNet framework used for image recognition [12].

By adapting these models into **1D-ResNets**, studies introduced "skip connections" that allow gradients to bypass certain computational layers, enabling the training of much deeper networks without losing information. Today, 1D-ResNet architectures are highly popular not only for end-to-end classification but also as robust feature-extraction backbones in more complex predictive frameworks.

Nevertheless, purely convolutional approaches retain an intrinsic structural limitation: their receptive field is localized. While they excel at identifying proximal motifs (peaks close to one another), CNNs inherently struggle to model long-range dependencies between peaks located at opposite ends of the spectral axis—a common occurrence when a biological condition is defined by the ratio of distant molecular vibrations [13].

2.3 Transformers and Self-Attention

To capture these global spectral dependencies without relying solely on infinitely deep convolutional layers, researchers have adapted the Transformer architecture to **1D spectral data**. Originally developed for natural language processing and

later popularized in vision through Vision Transformers (ViTs) [14], Transformers abandon convolutions entirely in favor of a mechanism known as "attention."

Instead of sliding a local filter over the signal, a 1D-ViT divides the continuous Raman spectrum into a sequence of non-overlapping 1D segments (patches). Through the **Self-Attention** mechanism, the model computes pairwise relationships across the entire input sequence simultaneously. Essentially, every patch "looks" at every other patch to determine which distant spectral regions are chemically related [13].

Recent works have proposed 1D Raman Transformer architectures for highly accurate disease detection, leveraging this ability to encode distributed combinations of peaks [15]. However, the literature also notes a major drawback: lacking the intrinsic inductive biases of CNNs (which naturally assume that close points are related), Transformers are highly data-hungry and prone to severe overfitting on small clinical datasets. This limitation heavily motivates the recent shift toward self-supervised pre-training.

2.4 Self-Supervised Pre-training for Spectral Data

A recurring problem in clinical spectroscopy is the imbalance between the relative ease of acquiring raw spectra and the high cost, time, and ethical constraints of obtaining expert clinical annotations [7]. Self-supervised learning (SSL) provides an elegant solution. SSL allows models to learn foundational representations from vast amounts of unlabeled data by solving mathematical "pretext tasks" before being fine-tuned on specific, data-scarce diagnostic applications.

2.4.1 Contrastive Learning: SimCLR and Barlow Twins

Contrastive learning aims to learn representations by teaching the network to recognize identical samples despite artificial variations. A prominent early framework is **SimCLR** [16]. It works by generating two differently augmented "views" of the same spectrum (e.g., adding noise, scaling amplitude, or shifting the baseline). The network is trained to maximize the mathematical similarity between these two views (positive pairs) while pushing them apart from all other different spectra in the training batch (negative pairs).

However, SimCLR requires massive batch sizes to supply enough negative samples to prevent the network from collapsing into trivial solutions. To overcome this computational bottleneck, recent literature has shifted towards more elegant architectures like **Barlow Twins** [17]. Instead of relying on negative pairs, Barlow

Twins feeds two augmented views of the same spectrum into identical 1D networks and computes the cross-correlation matrix of their output features. The training objective forces this matrix to be as close to the identity matrix as possible. This novel mechanism ensures that the same features across the two views correlate perfectly (values of 1 on the diagonal) while different feature dimensions decorrelate entirely (values of 0 off the diagonal), effectively learning robust, non-redundant representations without needing large batches.

A recent breakthrough in this direction is the **Raman Spectroscopy Pre-Trained Encoder (RSPTE)**, proposed by Eranti et al. in 2026 [18]. RSPTE adapts the Barlow Twins framework into a true domain-independent foundation model. By pre-training a 1D-ResNet backbone on a massive, highly heterogeneous corpus of unlabeled Raman spectra encompassing medical, biological, and mineralogical domains, RSPTE proved capable of learning fundamental spectral primitives (e.g., peak shapes, widths, and intensity ratios) [18]. Crucially, when fine-tuned on downstream data-scarce medical tasks—such as melanoma and COVID-19 detection—RSPTE demonstrated remarkable transferability, dramatically improving classification accuracy compared to models trained from scratch [18].

Despite these successes, a major challenge in contrastive frameworks like RSPTE remains: they depend heavily on domain-specific data augmentation. In spectroscopy, artificial baseline shifts, zeroing, or noise injection risk distorting the authentic chemical information if not carefully designed [19].

2.4.2 Masked Autoencoders and Foundation Models

To bypass the augmentation bottleneck entirely, researchers have adapted the **Masked Autoencoder (MAE)** paradigm [20] to 1D spectroscopy. MAEs utilize a reconstruction-based pretext task that requires no artificial data distortion.

In a 1D MAE framework, the spectrum is divided into patches, and a large subset (often 50% to 75%) is randomly masked. A 1D-ViT encoder processes only the visible patches, learning a dense, abstract latent representation of the remaining data. A decoder then attempts to reconstruct the missing biochemical context based solely on that latent understanding. By forcing the network to predict missing peaks from available context, the model intrinsically learns the physical relationships between vibrational modes.

Recent works, such as SMAE and RamanMAE, demonstrate that masked pre-training significantly improves downstream classification performance when labeled data is scarce [19], [21]. Reconstructing the signal also acts as an implicit denoising filter, establishing 1D-MAEs as a highly promising, augmentation-free foundation

strategy for spectroscopy.

2.5 Interpretability for Biomarker Discovery

In biomedical applications, deep learning models are expected not only to yield high diagnostic accuracy but also to assist in the discovery of novel clinical biomarkers [5]. Explainable AI (XAI) is essential to verify that predictions align with known physical chemistry rather than instrumental artifacts or dataset biases, ensuring clinical trust.

Recent studies have started incorporating XAI to bridge the gap between algorithmic decisions and biochemical meaning. In the context of 1D Raman spectroscopy, this means mapping the mathematical weights of the network directly back to specific vibrational bands. Two major techniques dominate recent literature:

- **Integrated Gradients (IG):** IG is an axiomatic attribution method originally developed for vision tasks [22], but effectively adapted to 1D spectral signals. It works by accumulating the gradients of the model's output along a straight mathematical path from a "blank" baseline—typically defined as a flat, zero-intensity spectrum—up to the actual patient's spectrum. This technique yields a high-resolution 1D attribution map. Recently, IG has been successfully applied to Raman deep learning by researchers such as Yan et al. (2022) [23], who used an attentional neural network combined with an IG interpretation algorithm to uncover the specific spectral markers underlying the discrimination of pathogen-derived extracellular vesicles. This demonstrated that IG can reliably highlight subtle, biologically significant peaks without human prompting.

- **SHAP (SHapley Additive exPlanations):** Rooted in cooperative game theory [24], SHAP calculates the marginal contribution of each input feature to the final prediction. When applied to Raman data, each wavenumber or spectral segment acts as a "player" contributing to the final diagnostic decision. A recent prominent study by Lin et al. successfully applied SHAP combined with spectral segmentation to Raman classification tasks [8]. Their work demonstrated that SHAP reveals not only *where* the model looks, but the *direction* of the impact. For example, it can explicitly prove whether an increased intensity in a specific lipid or Amide band actively pushes the model's diagnosis toward a "cancerous" rather than a "healthy" prediction [8].

By quantifying the exact contribution of different spectral regions, these techniques transform the neural network from a "black-box" classifier into a powerful, transparent tool for biomarker exploration and clinical validation [8].

2.6 Summary and Motivation for the Thesis

The literature reviewed in this chapter highlights a clear evolutionary trajectory in Raman spectral analysis: from classical chemometrics to deep convolutional networks, and recently toward transformer architectures and self-supervised learning.

While methods like Masked Autoencoders (MAE) and contrastive learning offer compelling theoretical solutions to the pervasive issue of data scarcity, the literature also reveals notable limitations. Most current studies evaluate these advanced paradigms in isolation or within highly controlled, single-domain datasets. Furthermore, the adoption of increasingly complex models often exacerbates the "black-box" problem, making their clinical translation difficult without the integration of robust Explainable AI (XAI) frameworks to verify the biological validity of the predictions.

Therefore, the current state of the art presents a distinct methodological challenge: bridging the gap between theoretical representation learning and transparent clinical application on highly heterogeneous data. Addressing this open challenge—by systematically investigating how self-supervised foundation models generalize across diverse Raman datasets, and by validating their learned representations through post-hoc biochemical interpretability—constitutes the fundamental motivation for the experimental work detailed in the subsequent chapters.

Chapter 3

Dataset Preparation and Preprocessing

3.1 Introduction to Data Collection

One of the most crucial and complex challenges in applying self-supervised deep learning architectures to Raman spectroscopy lies in the need for massive, highly heterogeneous data corpora. Unlike traditional guided classification approaches, the primary objective of this research phase is not the simple categorization of a sample, but rather the extraction and modeling of the universal physical and chemical laws governing Raman scattering. The computational system must be able to learn a true "molecular grammar," understanding the geometric and quantitative relationships between spectral peaks, chemical bond vibrations, and the optical responses of biological tissues, regardless of the specific experimental setup or pathology examined.

To meet this fundamental requirement and support the various experimental phases described in this thesis, a large-scale data aggregation and standardization pipeline was designed and implemented. The collected data is logically structured to support two fundamentally different and complementary training paradigms:

- **Unannotated Pre-Training Corpus:** A massive aggregation of Raman spectra originating from widely diversified domains (in vitro cell lines, in vivo mouse models, and human intraoperative tissue mappings). This corpus, intentionally stripped of discrete clinical labels, is used to force the neural network (utilizing mechanisms such as Masked Autoencoders or spectral regularization criteria like Barlow Twins) to reconstruct or align signals, thereby internalizing the intrinsic physics of scattering and the structural variance of the chemical background.

- **Datasets for Downstream Classification and Fine-Tuning:** Highly specialized datasets provided with precise diagnostic, microbiological, or histopathological labels. These are employed to evaluate the model's ability to transfer the learned physical knowledge towards fine diagnostic classification tasks, biomarker extraction, and intraoperative tissue segmentation.

3.2 Pre-Training Corpus Components

In order to guarantee the learning of robust physical representations that are invariant to experimental fluctuations, data from multiple biological domains was unified. This heterogeneity prevents the model from overfitting to systematic artifacts specific to a single spectrometer or biological cell line.

3.2.1 Dataset 1: MDA-MB-231 (DeepR Pre-training Corpus for Denoising)

The first core of the corpus consists of data extracted from the DeepR framework workspaces [25]. The biological source of these spectra is the **MDA-MB-231** cell line, a widely used in vitro model of human breast adenocarcinoma. Biologically, this is a triple-negative cell line, characterized by a highly aggressive, invasive, and phenotypically unstable profile.

This specific sub-dataset does not contain discrete clinical labels but is natively structured as dual-SNR (Signal-to-Noise Ratio) spectral pairs. As demonstrated in the RamanMAE [21] and DeepR [25] frameworks, this structure is specifically designed to guide self-supervised algorithms in the domain of chemical denoising:

- **Global Matrix:** A massive archive comprising 172,312 spectral pairs, rigidly split into a training set (159,618 pairs) and a testing set (12,694 pairs).
- **Low SNR Spectra (Inputs):** Signals acquired with a short integration time of only 0.1 seconds. These represent the degraded experimental condition, where molecular vibrational peaks are partially submerged by electronic and photonic noise.
- **High SNR Spectra (Targets):** Signals acquired on the exact same spatial point with a longer integration time (1.0 seconds), serving as the clean, high-chemical-resolution ground truth.

The primary goal of this dataset is purely physical reconstruction. The excellent original quality control of this matrix guarantees the total absence of missing values (`NaN`), allowing the direct application of stochastic masking to induce the network to reconstruct the morphology of the missing molecular peaks, effectively functioning as a powerful non-linear spectral smoothing technique.

3.2.2 Dataset 2: Cancer Metastasis (Single-Cell Resolution - "Cell" Dataset)

Although this dataset shares the same biological root as Dataset 1 (the MDA-MB-231 cell line), it introduces a radical methodological variation [21], justifying its identification as the **"Cell" Dataset**.

While Dataset 1 aggregates highly repetitive spectral pairs optimized for pure signal reconstruction and noise suppression, the "Cell" Dataset was acquired by applying a Raman microscopy technique at **single-cell resolution** to evaluate downstream classification. The explicit choice to isolate individual cellular bodies allows the exploration of the fine biochemical transitions that occur during tumor progression. The original matrix has a dimension of $23,530 \times 871$, with a wavenumber axis extending from 600.46 cm^{-1} to 1963.25 cm^{-1}.

The preliminary analysis of the matrix required careful physical discrimination of the signal:

- **Signal vs. Background:** The original matrix contains both authentic cellular spectra and pure background acquisitions. Spectra identified as background-only measurements (Status 0) were excluded from the chemically informative subset.

- **Missing Values Handling:** The matrix also contained `NaN` values due to reading anomalies of the spectrometer's CCD camera. Background-only spectra and rows affected by missing values were removed together, for a total of 8,101 discarded spectra. This filtering step leaves 15,429 authentic cellular spectra (Status 1), valid for chemical modeling.

The 15,429 valid spectra describe the molecular evolution across three sequential phenotypic stages derived from a genetically matched mouse model: parental cells confined to the primary tumor (P231), circulating tumor cells isolated from the bloodstream (CTC), and metastatic cells colonizing the lung tissue (LM). To correctly map this biological variability without incurring data leakage, a strict Leave-One-Cell-Out cross-validation approach is enforced (Table 3.1).

Table 3.1: Class, percentage, and biological replicate distribution for Dataset 2 (Authentic Cellular Spectra Only).

Class ID	**Phenotype / Metastatic Stage**	**Spectra**	**%**	**Unique Cell IDs**
1	P231 (Parental - Primary Tumor)	5,387	34.9%	69 cells (IDs: 137–205)
2	CTC (Circulating Tumor Cells)	5,413	35.1%	69 cells (IDs: 1–69)
3	LM (Lung Metastasis)	4,629	30.0%	67 cells (IDs: 70–136)
Global Distribution		**15,429**	**100.0%**	**205 Total Unique Cells**

3.2.3 Dataset 3A: Cancer Metastasis / TWIST (In Vivo Mouse Models)

This dataset extends the study of metastasis to an *in vivo* tissue context, collecting single-point Raman spectra directly from tumor nodules developed within 67 genetically modified mice [21]. The dataset maps 9 cellular variants engineered to express different levels of the TWIST transcription factor, a key regulator of the epithelial-mesenchymal transition (EMT).

Out of 18,092 originally recorded spectra, a quality control pipeline removed 1,596 saturated signals or those lacking optical coupling, stabilizing the dataset at 16,496 valid spectra.(Table 3.2).

Table 3.2: Class and biological murine replicate distribution for Dataset 3A.

Class ID	**Cell / Tumor Type**	**Spectra**	**%**	**Unique Mouse IDs**
1	S4T1CT	1,657	10.0%	8 mice (IDs: 21–28)
2	S4T1TW	1,909	11.6%	8 mice (IDs: 41–48)
3	S67NR	2,209	13.4%	8 mice (IDs: 61–68)
4	S4T07	2,522	15.3%	10 mice (IDs: 81–85, 89–93)
5	SFARN	1,519	9.2%	7 mice (IDs: 105–111)
6	SCX1KD	1,403	8.5%	6 mice (IDs: 121–126)
7	SCX1VC	1,485	9.0%	6 mice (IDs: 141–146)
8	SFX1KD	1,933	11.7%	7 mice (IDs: 161–167)
9	SFX1VC	1,859	11.3%	7 mice (IDs: 181–187)
Global Distribution		**16,496**	**100.0%**	**67 Total Unique Mice**

3.2.4 Dataset 3B: Immunotherapy Response (IMMUNO)

Designed to map the kinetic modifications of the tumor microenvironment in response to therapeutic regimens, this dataset collects 11,099 Raman spectra from 38 mice affected by breast carcinoma and treated with different immunotherapy protocols (IGGC Control, Anti-CTLA4, Anti-PDL1, and COMB combination therapy) [21].

Table 3.3: Class and treatment arm distribution for Dataset 3B (IMMUNO).

Class ID	**Clinical Treatment Group**	**Spectra**	**%**	**Unique Mouse IDs**
1	IGGC (Control Group)	2,935	26.4%	10 mice (IDs: 21–30)
2	CT4A (Anti-CTLA4 Immunotherapy)	2,692	24.3%	9 mice (IDs: 41–49)
3	PD1L (Anti-PDL1 Immunotherapy)	2,164	19.5%	7 mice (IDs: 61–67)
4	COMB (Combination Therapy)	3,308	29.8%	12 mice (IDs: 81–92)
Global Distribution		**11,099**	**100.0%**	**38 Total Unique Mice**

3.2.5 TropHY Dataset: Intraoperative Tissue Mappings

The clinical TropHY dataset provides hyperspectral Raman maps acquired ex-vivo immediately following surgical resection on oncological patients. Structurally, each map is stored in a MATLAB file containing a three-dimensional tensor, where the first two dimensions define the spatial coordinate (X, Y) of the pixel on the tissue sample, and the third dimension represents the continuous Raman spectrum sampled over 820 points (extending from 648.7 to 3050.4 cm^{-1}).

After applying local morphological filters to eliminate pixels burned by high laser power density or empty areas of the slide, a total of **115,913 valid tissue spectra** were extracted. Depending on the clinical objective, the dataset supports multiple classification schemes, such as a 4-class morphological classification to discriminate the fundamental histological constituents: Muscle (0), Stroma (1), Tumor (2), and Vessel (3) (Table 3.4).

TropHY was initially considered as a natural candidate for downstream fine-tuning and external validation, because its annotations describe clinically meaningful tissue classes in an intraoperative setting. However, the available cohort is limited to only two patients, and the resulting biological variability is therefore too narrow to support a robust patient-independent validation protocol. For this reason, using TropHY as a decisive fine-tuning benchmark would be methodologically fragile: performance estimates could be dominated by patient-specific acquisition patterns rather than by generalizable histopathological information. In this thesis, the dataset is therefore treated primarily as a valuable source of heterogeneous

tissue spectra for representation learning and corpus enrichment, while its use for supervised validation is interpreted with caution.

Table 3.4: Detailed map-level and patient-level distribution for the TropHY dataset (4-Class Task).

MapID	Name	Patient	Muscle (0)	Stroma (1)	Tumor (2)	Vessel (3)	Total
M1	1a	Paz 1	0	4,953	9,275	0	14,228
M2	1b	Paz 1	0	6,370	7,836	0	14,206
M3	3a	Paz 1	0	0	11,059	0	11,059
M4	5a	Paz 1	0	0	10,743	0	10,743
M5	5b	Paz 1	0	0	11,205	0	11,205
M6	5c	Paz 1	0	0	12,487	0	12,487
M7	7a	Paz 2	3,494	0	2,940	0	6,434
M8	7b	Paz 2	3,669	0	4,429	1,841	9,939
M9	8a	Paz 2	0	9,848	2,822	0	12,670
M10	9a	Paz 2	0	10,234	795	1,913	12,942
Global Distribution			**7,163**	**31,405**	**73,591**	**3,754**	**115,913**

3.2.6 SpectraBreast: Highly Raw (RAW) Raman Maps and Isolated Pre-Processing

Provided within the European research project SpectraBreast, this dataset represents the most complex spectral benchmark of the thesis. Unlike all previously analyzed datasets (DeepR, Cell, TWIST, IMMUNO), which were delivered in an already partially filtered and stabilized format by their respective spectroscopy laboratories, SpectraBreast introduces data in a completely **RAW** state in the form of spatial hyperspectral mappings.

The original SpectraBreast tensors massively suffer from all the physical perturbations typical of the uncontrolled intraoperative environment. Within the raw signal coexist saturated fluorescence emission (which overwhelms the intensity of Raman scattering), high-energy cosmic rays (spikes), and instrumental alignment artifacts. This radically raw nature imposes the need to apply a dedicated and isolated initial preprocessing pipeline before these spectra can be utilized for deep learning.

In accordance with the logical architecture implemented in the initial operational cells of the computational environment, the raw spatial data of SpectraBreast are processed pixel by pixel through three fundamental steps, aimed at isolating Raman information from instrumental artifacts:

1. **Cosmic Ray and Isolated Spike Removal:** High-energy cosmic rays appear as

narrow intensity jumps affecting one or a few adjacent channels. Detection is therefore performed on the local first-order difference signal rather than on the absolute intensity alone. Defining

$$d_i = x_i - x_{i-1}, \tag{3.1}$$

a channel is marked as a spike candidate when its derivative-like deviation is abnormally large with respect to the robust dispersion of the local neighborhood:

$$|d_i - \text{median}(d_{i-w:i+w})| > \tau \cdot \text{MAD}(d_{i-w:i+w}). \tag{3.2}$$

Here, w defines the local window, MAD is the Median Absolute Deviation, and τ controls detector sensitivity. This formulation highlights abrupt discontinuities while ignoring broad Raman peaks and smooth fluorescence trends. Corrupted points are replaced via local linear interpolation from adjacent clean channels.

2. **Background and Fluorescence Subtraction:** To extract molecular vibrational peaks from the fluorescence background, an Asymmetric Least Squares (AsLS) correction or an iterative high-degree polynomial fit is applied. The baseline profile $\mathbf{z}$ is estimated by minimizing:

$$S = \sum_i w_i (x_i - z_i)^2 + \lambda \sum_i (\Delta^2 z_i)^2 \tag{3.3}$$

The parameter λ controls smoothness, while the asymmetric weights w_i (typically p for $x_i > z_i$ and $1 - p$ for $x_i < z_i$, with $p \ll 1$) force the model to follow the lower fluorescence trend rather than the positive Raman peaks.

3. **Savitzky–Golay Smoothing and Residual Noise Filtering:** After artifact removal, a Savitzky–Golay filter attenuates high-frequency fluctuations while preserving Raman-band morphology. Unlike a moving average, it fits a low-degree polynomial within a sliding window, retaining local curvature and narrow peak height. For a window centered on x_i:

$$\hat{x}_i = \sum_{k=-m}^{m} c_k x_{i+k} \tag{3.4}$$

where $2m+1$ is the window length and c_k are the polynomial-order coefficients. This step removes residual noise without the distortions caused by excessive smoothing or arbitrary signal truncation.

The visual impact of this isolated preprocessing stage is documented in Figures 3.1 and 3.2.

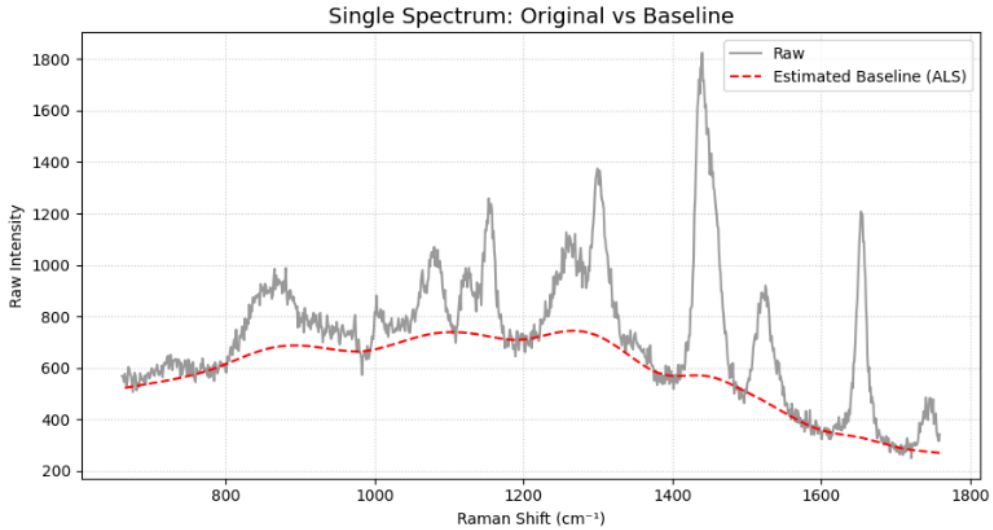


Figure 3.1: Representative RAW spectrum extracted from a SpectraBreast hyperspectral map before correction, showing the dominant fluorescence contribution and isolated high-intensity artifacts.

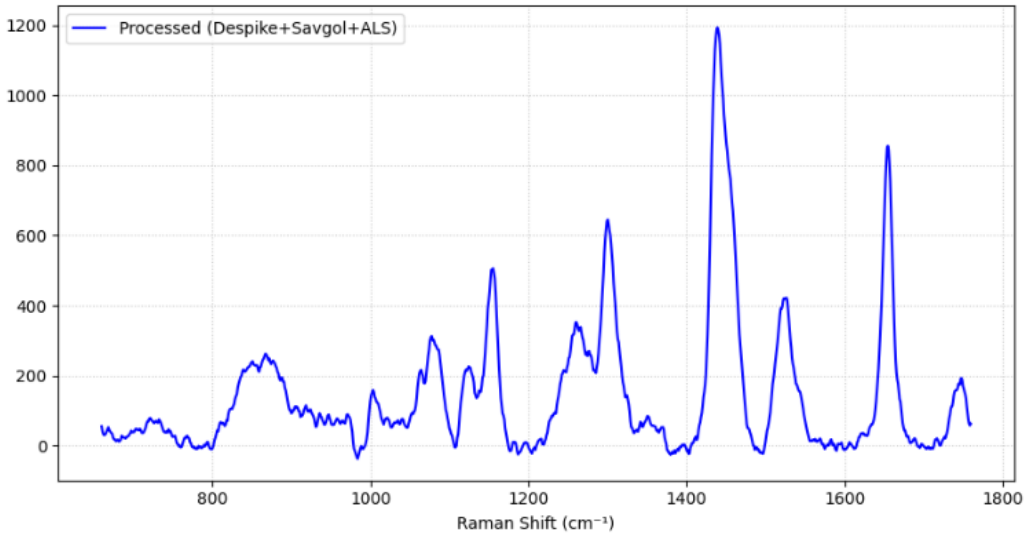


Figure 3.2: The same representative SpectraBreast spectrum after preprocessing, where the removal of the background component produces a cleaner vibrational profile suitable for subsequent harmonization.

3.3 Datasets for Downstream Classification and Fine-Tuning

The following datasets were primarily employed in the fine-tuning phase to test the effectiveness of the learned physical representations on specific, clinically relevant scenarios. In contrast to TropHY, whose patient-level variability is too limited to support a robust independent validation, these datasets provide more suitable supervised targets for evaluating diagnostic transfer, microbiological discrimination, or clinically oriented classification.

3.3.1 IBD Dataset: Blood Plasma Liquid Biopsy

This dataset translates the modeling towards the analysis of systemic biofluids. It collects the spectral profiles from blood plasma liquid biopsies performed on a clinical cohort of 228 patients, covering the spectral axis from 403.89 cm^{-1} to 1754.26 cm^{-1}.

The dataset is accompanied by patient metadata that provides the diagnostic ground truth, securely anonymized to prevent data leakage across model splits. The diagnostic labels map the inflammatory architecture, classifying the 228 patients into Healthy Controls (HC) and individuals suffering from Inflammatory Bowel Diseases (IBD), specifically Crohn's Disease and Ulcerative Colitis.

3.3.2 BACTERIA-ID Dataset: Open-Source Pathogen Identification

The BACTERIA-ID dataset used in this thesis is based on the open Raman pathogen-identification resource introduced by Ho et al., who demonstrated rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning [26].

This dataset introduces a microbiological domain to the modeling process, mapping the spectroscopic signatures of prevalent bacterial classes, including widely studied pathogens such as *Staphylococcus aureus*, *Escherichia coli*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa*. The dataset is structurally divided into two distinct splits:

- **"References" Split:** Contains approximately 60,000 spectra acquired under highly controlled and stabilized laboratory conditions. Due to its vast scale and minimal biological background, this subset possesses the ideal characteristics to be retroactively integrated into the large unannotated Pre-training Corpus.

- **"Fine-Tune" Split:** Contains exactly 3,000 spectra of clinical isolates exhibiting a much more realistic and challenging degree of variance (e.g., bacteria treated with antibiotics or cultured in complex blood media). This subset is utilized strictly in the downstream phase for the final evaluation of the model's diagnostic classification performance on real-world pathogen samples.

3.4 Theoretical Background on Raman Preprocessing

Raman spectra cannot usually be passed directly to a learning algorithm without a harmonization step. Even when different spectra describe the same biological phenomenon, they may be acquired with different spectrometers, laser wavelengths, spectral resolutions, integration times, or preprocessing histories. As a consequence, the raw numerical vectors may differ not only because of biology, but also because of acquisition and instrumentation. If these effects are not controlled, a neural network may learn to separate datasets, instruments, or preprocessing artifacts rather than clinically meaningful biochemical patterns.

A first issue is the alignment of the spectral axis. Raman spectra are functions of Raman shift, but different datasets often provide measurements on different wavenumber grids. Deep learning models, however, generally require fixed-length inputs where each index corresponds to the same physical region across all samples. Interpolation therefore becomes necessary to map all spectra onto a common axis. In this thesis, harmonization to a 1000-point representation allows the models to compare spectra consistently and ensures that a convolutional filter, Transformer token, or attribution score refers to the same Raman shift across samples.

A second issue is intensity scaling. Absolute Raman intensity can vary for reasons that are not directly related to the biochemical state of the sample, including laser power, acquisition time, focusing conditions, optical coupling, and sample thickness. Normalization is used to reduce these global amplitude effects and make spectra more comparable. This is particularly important in multi-domain pre-training, where spectra from different experiments must be combined. Without normalization, the model could treat acquisition intensity as a shortcut feature instead of learning the relative morphology of Raman peaks.

A third issue is background and baseline variability. Biological Raman spectra are often affected by fluorescence, broad background contributions, and slowly varying baseline drift. These effects can dominate the raw intensity scale and partially obscure weaker Raman peaks. Although baseline correction is not always applied

in the same way across all datasets, the theoretical objective is clear: the learning algorithm should focus as much as possible on vibrational structure rather than on broad non-specific background trends. This is also why robust pre-training is useful: a model exposed to heterogeneous spectra and realistic perturbations can learn representations less sensitive to nuisance variability.

Finally, preprocessing must preserve chemical information. Excessive smoothing, aggressive baseline subtraction, or inappropriate normalization can remove or distort diagnostically relevant peaks. For this reason, preprocessing is not merely a technical cleaning step, but a methodological compromise between standardization and biochemical fidelity. The pipeline adopted in this thesis is designed to make heterogeneous Raman spectra compatible with deep learning while retaining the spectral morphology needed for downstream classification and interpretability.

3.5 Spectral Processing and Data Harmonization

3.5.1 Uniformation Philosophy and Feature Space Harmonization

While the isolated pre-processing pipeline resolved the severe instrumental artifacts of the raw datasets (such as SpectraBreast), the coexistence of structurally disparate datasets requires a secondary, global mathematical strategy. Even after cleaning the signals, the consolidated datasets still present unaligned wavenumber axes and different native resolutions (e.g., 871 channels for Cell, 820 for TropHY). However, the projection of all spectra onto a single spectral window is not imposed universally for every model: it is mainly required when the architecture assumes a fixed positional structure.

For the Vision Transformer (ViT), the reference spectral axis is defined in the *"biological fingerprint"* region, extending uniformly from **650 to 1750 cm^{-1}**. The main configuration discretizes this interval into **1000 equidistant points**, because reference works reported stronger results with this resolution; however, the **500-point** interpolation also achieved competitive performance and is therefore considered as an additional experimental setting. For clarity, the figures in this section report only the 1000-point case, but both resolutions are explored during model evaluation. The fixed cut is necessary because the ViT associates each input patch with a positional embedding; if spectra from different datasets covered different Raman-shift ranges or had incompatible sampling grids, the same embedding index could correspond to different physical wavenumbers. Such inconsistency would make the attention

mechanism compare chemically non-equivalent positions. For this reason, the ViT pipeline uses a fixed spectral axis and projects all compatible spectra onto it via **Cubic-Spline Interpolation**.

This restriction is less rigid for Barlow Twins and for the 1D-CNN baselines. In those cases, spectra can be kept closer to their native acquisition ranges when the experimental protocol benefits from preserving dataset-specific information. This strategy is consistent with the reference methodological literature, where different spectral ranges are sometimes exploited rather than forcibly discarded, provided that the comparison is performed within a model or training setup able to handle such variability. Therefore, the 650–1750 cm^{-1} cut should be interpreted as an architectural requirement of the ViT harmonization pipeline, not as an absolute preprocessing rule applied to every experiment.

Unlike simple linear interpolation, which joins points via straight line segments introducing discontinuities in the first derivatives and artificially smoothing the cusps of narrow Raman peaks, the cubic spline models the signal using piecewise third-degree polynomials. Given the original spectral nodes, the spline guarantees the continuity of the function, its first derivative, and its second derivative at every junction point:

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \quad \text{for } x \in [x_i, x_{i+1}] \tag{3.5}$$

This mathematical rigor preserves the area under the molecular peaks and the non-linear curvatures of the signals. The interpolation behavior is first inspected locally in Figure 3.3, where the original samples and the newly interpolated points are shown in a restricted spectral interval. The global result is then reported in Figure 3.4, while the effect of amplitude normalization across spectra is shown in Figure 3.5.

3.5.2 Amplitude-Dependent Normalization Strategies

The final step of the pipeline concerns amplitude standardization. The absolute intensity of the scattered photons via the Raman effect is intrinsically arbitrary, being influenced by external non-biological parameters such as microscopic variations in laser power, optical focus stability on the sample, and integration times.

To evaluate the impact of amplitude standardization on the learning process, three normalization strategies are explored rather than assuming a single optimal choice a priori. All three transformations preserve the relative spectral morphology and therefore do not remove the chemical information encoded in peak positions

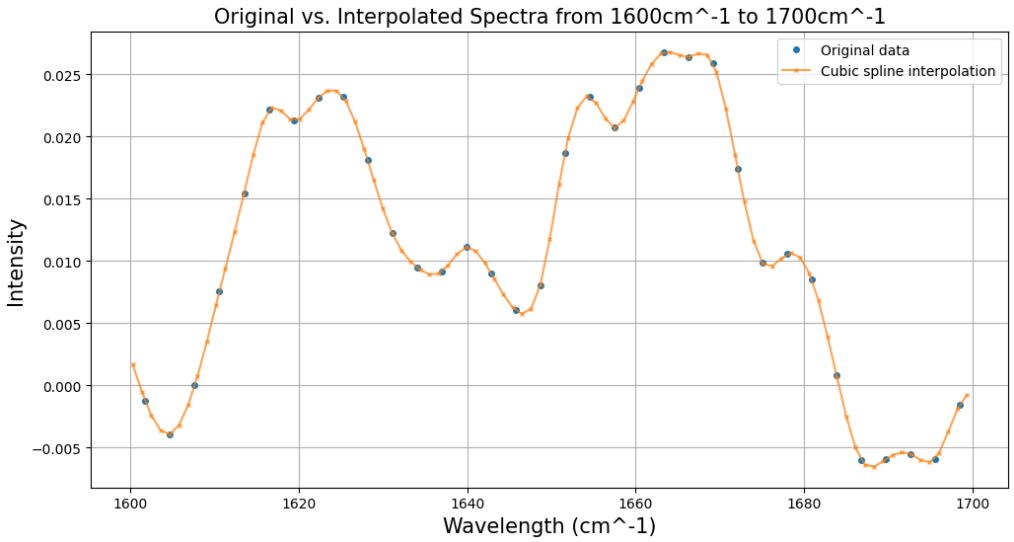


Figure 3.3: Generic visual example of the cubic-spline interpolation mechanism. The original sampled points are shown in blue, while the yellow points represent the values reconstructed on a denser target grid. This local view illustrates how the spline follows the curvature of the signal between measured samples without reducing the interpolation to straight linear segments.

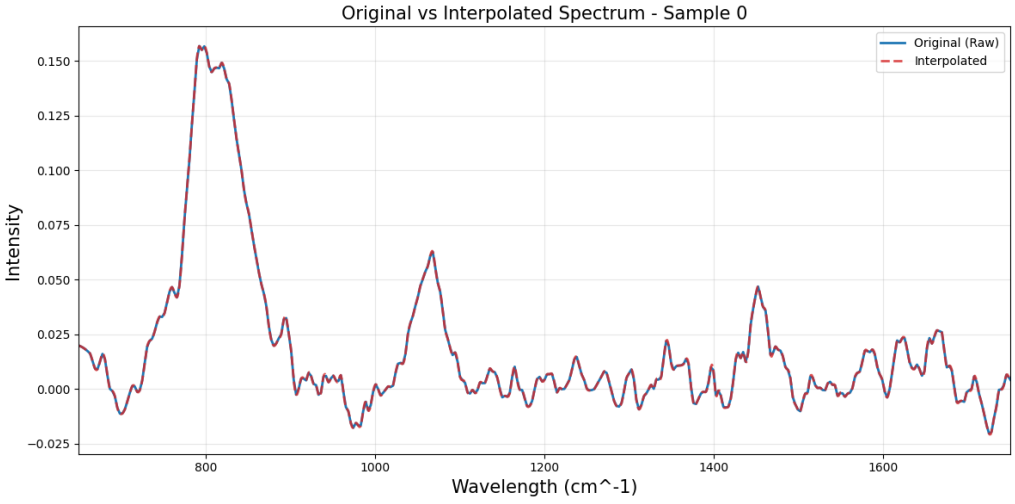


Figure 3.4: Global overlay between the original spectrum and its cubic-spline projection onto the common 650–1750 cm^{-1} ViT window. The overlap verifies that the harmonization step preserves the global spectral profile while enforcing a fixed input length and shared physical axis.

and peak relationships; instead, they modify the scale or centering of the intensity vector so that the model can focus less on acquisition-dependent amplitude effects.

- **Min-Max Normalization:** Each spectrum is linearly rescaled inside the $[0, 1]$ interval:

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3.6}$$

  This strategy removes absolute intensity differences while preserving the shape of the spectrum and the relative height of its peaks.

- **Standard Normal Variate (SNV):** Each spectrum is centered and divided by its own standard deviation:

$$x_{SNV} = \frac{x - \mu_x}{\sigma_x} \tag{3.7}$$

  where μ_x and σ_x are the mean and standard deviation of the individual spectrum. SNV is useful for reducing multiplicative scattering effects and baseline-related intensity offsets.

- L_2 **Normalization:** Each spectrum is projected onto the unit hypersphere:

$$x_{L_2} = \frac{x}{\|x\|_2} \tag{3.8}$$

  This approach keeps the direction of the spectral vector while forcing all spectra to have the same Euclidean norm.

The three alternatives are tested experimentally to verify whether one amplitude representation leads to more stable or more discriminative downstream performance. For simplicity, Figure 3.5 visualizes only the Min-Max case, but the same principle applies to SNV and L_2 normalization: the numerical scale changes, while the Raman peak positions and the structural information of the spectrum remain available to the model.

3.6 Summary of the Pre-Training Corpus

The completion of the individual pre-processing and global harmonization pipelines allowed the construction of the unified spectral corpus for pre-training. Having reduced all signals to homogeneous vectors of 1000 points sampled in the same physical window, the computational system can map chemical bonds transversally

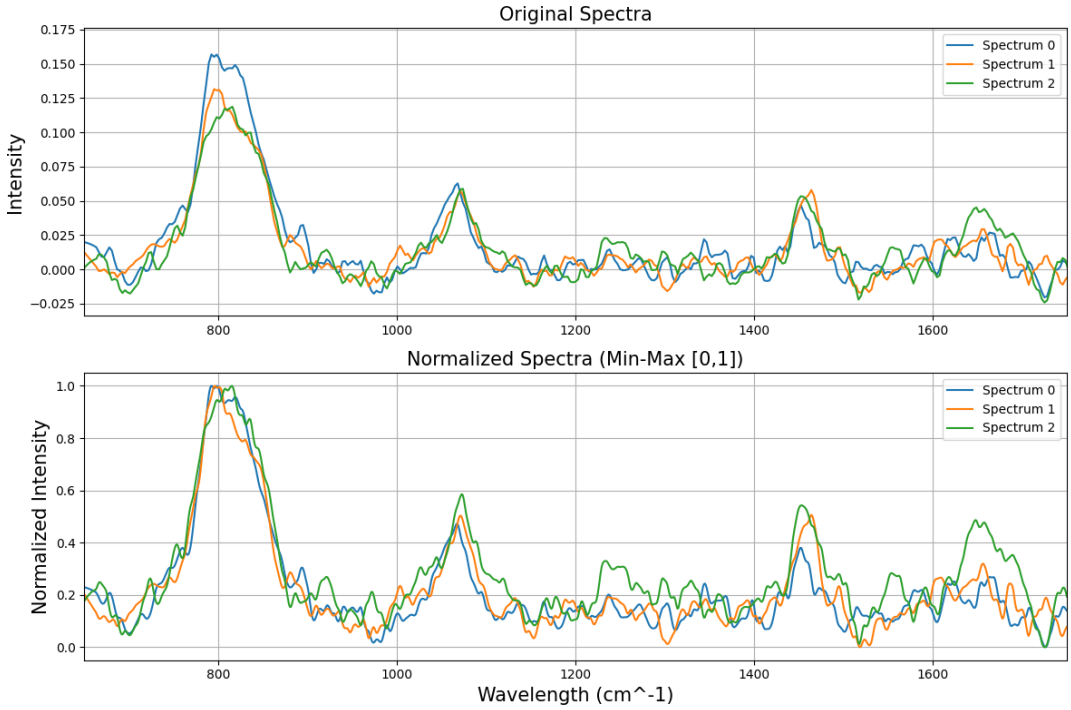


Figure 3.5: Example of Min-Max amplitude normalization across spectra. The upper panel shows spectra with different offsets and intensity scales, while the lower panel shows the same signals mapped to a shared $[0, 1]$ dynamic range without altering the relative position of the Raman bands.

across biological domains. The overall dimensional scale of the unified information asset is summarized in Table 3.5.

Table 3.5: Quantitative summary of the datasets harmonized to 1000 points and fused for the learning of universal physical laws. (Note: The BACTERIA-ID reference split, not listed here, provides an additional ∼60,000 spectra available for retrospective pre-training enrichment).

Dataset	**Biological Context / Data Source**	**Total Valid Spectra**
DeepR (MDA-MB-231)	Pre-filtered dual-SNR pairs for chemical denoising	172,312
TropHY (Unlabeled Subset)	Human intraoperative histological maps examined ex-vivo	115,913
TWIST	In vivo mouse models for induced EMT expression	16,496
Cell (MDA-MB-231)	Single-cell resolution for phenotypic heterogeneity	15,429
Immuno	In vivo mouse models for oncological therapy monitoring	11,099
SpectraBreast	Raw maps (RAW) from the operating room *(Processed via isolated pre-processing)*	*Added progressively*
Total Harmonized Corpus for Pre-training		**> 331,249**

Chapter 4

Pre-Training Foundation Models for Raman Spectroscopy

4.1 Introduction to Self-Supervised Learning

The successful deployment of deep learning in clinical and biological spectroscopy is fundamentally bottlenecked by the scarcity of annotated data. As detailed in Chapter 3, while acquiring raw Raman spectra is relatively fast, obtaining expert-level histopathological or biochemical annotations for each pixel is extremely expensive, requires the intervention of domain experts (such as pathologists), and is highly time-consuming. To overcome this critical limitation, this thesis leverages the paradigm of Self-Supervised Learning (SSL).

Instead of relying on human-provided labels, SSL constructs "pretext tasks" directly from the intrinsic structure of the unlabeled data. By forcing the neural network to solve these tasks across the massive harmonized corpus of multidomain spectra (comprising over 330,000 samples from varying biological and chemical backgrounds), the model learns a robust, generalized latent representation of the physical laws governing Raman scattering.

In this chapter, two radically different but highly effective self-supervised paradigms are explored, implemented, and compared: **Barlow Twins**, a contrastive-style architecture focusing on redundancy reduction and instrumental invariance, and **Masked Autoencoders (MAE)**, a generative architecture focusing on the physical reconstruction of missing chemical information.

4.2 Representation Learning and Transfer Learning Rationale

Self-supervised learning can be understood as a way to transform unlabelled data into a source of supervision. Instead of requiring a human diagnostic label for each spectrum, the model is trained to solve an auxiliary task whose target can be generated directly from the data. In Raman spectroscopy, this is particularly attractive because unlabelled spectra are easier to collect than clinically annotated spectra. A model can therefore learn general spectral regularities before being adapted to a

smaller supervised dataset.

The goal of pre-training is not to solve the final clinical task directly, but to learn an encoder that maps spectra into useful latent representations. A good representation should preserve chemically meaningful information while discarding nuisance variability such as noise, baseline changes, or acquisition-specific intensity fluctuations. During downstream fine-tuning, a supervised classifier can then use this representation as a starting point. This typically improves data efficiency because the model does not need to learn low-level spectral structure from scratch using only a small labelled set.

Two broad families of self-supervised objectives are especially relevant for Raman data. Contrastive or redundancy-reduction methods, such as Barlow Twins, train the model to produce similar embeddings for different augmented views of the same spectrum. These methods emphasize invariance: the representation should remain stable when the input is perturbed in ways that do not change the underlying chemical identity. Reconstruction-based methods, such as Masked Autoencoders, instead remove part of the input and train the model to reconstruct it. These methods emphasize contextual understanding: the representation should contain enough information to infer missing spectral regions from the rest of the signal.

These two objectives are complementary. Invariance learning is useful for handling acquisition variability, while reconstruction learning is useful for capturing relationships between Raman bands. The experimental design of this thesis therefore compares Barlow Twins and SMAE not as interchangeable algorithms, but as two different hypotheses about what a Raman foundation model should learn.

4.3 Barlow Twins for Spectral Invariance

Many conventional metric-learning algorithms require the use of positive and negative pairs, training models to create similar embeddings for identical classes and divergent embeddings for different ones. However, these approaches often require massive batch sizes to function correctly or make assumptions about data similarity that may be incorrect when dealing with highly heterogeneous spectral datasets.

To circumvent this, the **Barlow Twins** approach was adapted for 1D Raman spectral data. The primary objective of this architecture is to learn an embedding space where identical biological samples, even when subjected to severe instrumental perturbations, produce the exact same mathematical representation. It achieves this by explicitly focusing on redundancy reduction, eliminating the need for negative pairs entirely.

The central idea is illustrated in Figure 4.1. Starting from a single Raman spectrum, the pipeline generates two independent augmented views that should preserve the same biochemical identity while altering only acquisition-related nuisance factors. Both views are processed by the same encoder and projector, producing two latent representations. The training objective then aligns the corresponding dimensions of these representations while simultaneously decorrelating different dimensions, so that the model learns invariant but non-redundant spectral descriptors.

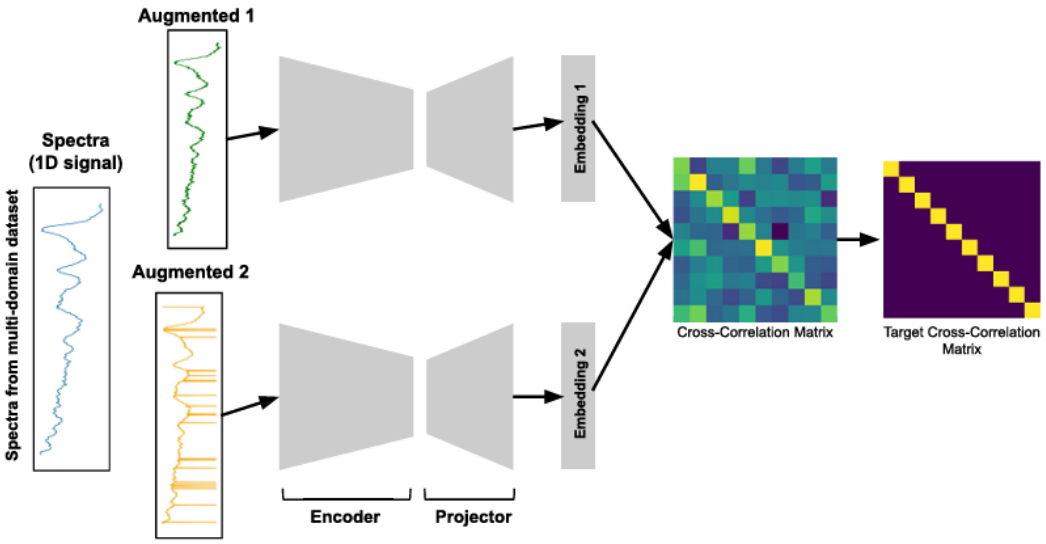


Figure 4.1: Visual demonstration of the Barlow Twins pipeline adapted for 1D Raman spectra. A single input spectrum is mathematically augmented into two distinct versions. Both are passed through a shared Encoder and Projector network, yielding two embeddings. A cross-correlation matrix is then derived and optimized to approximate the identity matrix.

4.3.1 Backbone Selection: 1D-ResNet Variants

Because Barlow Twins operates on the holistic view of the entire spectrum to compute correlations, a 1D Convolutional Neural Network (CNN) based on the Residual Network (ResNet) topology is deployed as the foundational backbone. Following the design philosophy of the original Barlow Twins framework [17], the main candidate was a **1D-ResNet-34**. This architecture provides enough depth to capture complex biochemical multiplets and long-range peak relationships without moving to a much heavier ResNet-50-like model.

For comparison, a lighter **1D-ResNet-18** was also implemented and tested. The purpose of this experiment was to verify whether a smaller encoder could provide more stable representation learning on one-dimensional Raman spectra. ResNet-18

contains fewer residual blocks, requires less memory, and is faster to train, but it also has lower representational capacity. In practice, this makes it less suitable for subtle spectral structures such as weak peaks, shoulders, and fluorescence-modulated bands.

Both backbones were tested using the flattened spectral representation before the projector, consistently with the Barlow-style objective. A global average pooling variant was also explored during preliminary experiments, but it substantially degraded the quality of the learned representations. For this reason, it was not retained as a meaningful final configuration: the pooling operation compressed the spectral feature map too aggressively, removing local information that is important for Raman peak morphology.

In standard deep CNNs, increasing depth often leads to vanishing gradients, making training unstable. The ResNet architecture mitigates this by introducing *Skip Connections*, allowing the network to learn residual functional mappings. To process spectral data properly, the base architecture was completely adapted to 1D by substituting all 2D spatial layers with their sequential equivalents (`Conv1D`, `MaxPool1D`, and `ZeroPadding1D`). These 1D convolutions sweep across the wavenumber axis, extracting hierarchical physical features. The early layers learn to identify simple baseline slopes and broad fluorescence drifts, while the deeper layers focus on complex, high-frequency multiplet peak formations.

Table 4.1: Backbone variants retained for the Barlow Twins comparison on Raman spectra.

Variant	**Main Advantage**	**Observed Limitation**
1D-ResNet-18	Faster and lighter; lower memory footprint.	Less expressive for weak Raman bands and subtle peak morphology.
1D-ResNet-34	Higher representational capacity; better suited to complex Raman patterns.	Heavier than ResNet-18, but still manageable for the pretraining corpus.

Based on these comparisons, the final Barlow Twins experiments retained the **1D-ResNet-34 without global average pooling** as the backbone. This configuration performed better because it preserved richer local spectral information before the projector while still benefiting from the stability of residual learning. The ResNet-18 variant remained computationally attractive, but its reduced depth limited the quality of the learned biochemical representation.

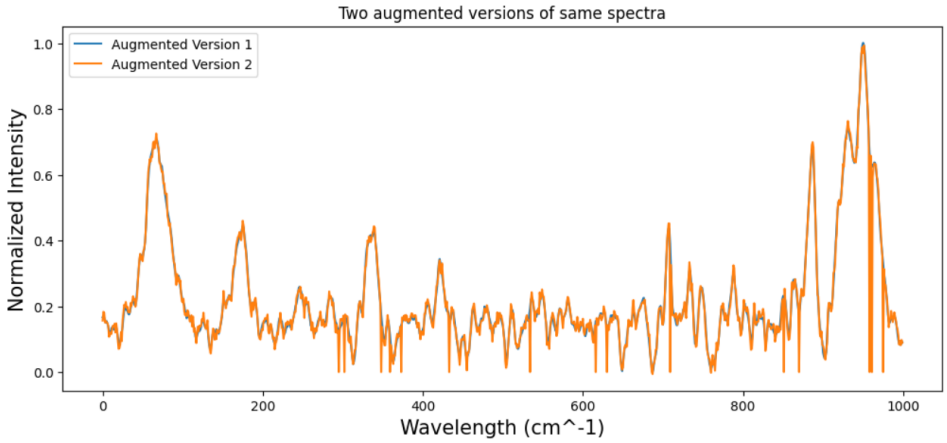
4.3.2 Domain-Specific Spectral Augmentations

A critical component of the Barlow Twins methodology is the generation of two distorted views (View A and View B) from the exact same sample. Conventional image augmentations (such as rotation, cropping, or color jittering) have no physical analogue in 1D spectral data. Therefore, a specialized TensorFlow-based augmentation pipeline was implemented for 1D Raman spectra. At each call, the pipeline stochastically applies only a subset of transformations, ensuring that the two views are different but still chemically consistent. The default configuration applies augmentation to every spectrum, while the individual operations are selected through random probabilities and physically interpretable parameter ranges.

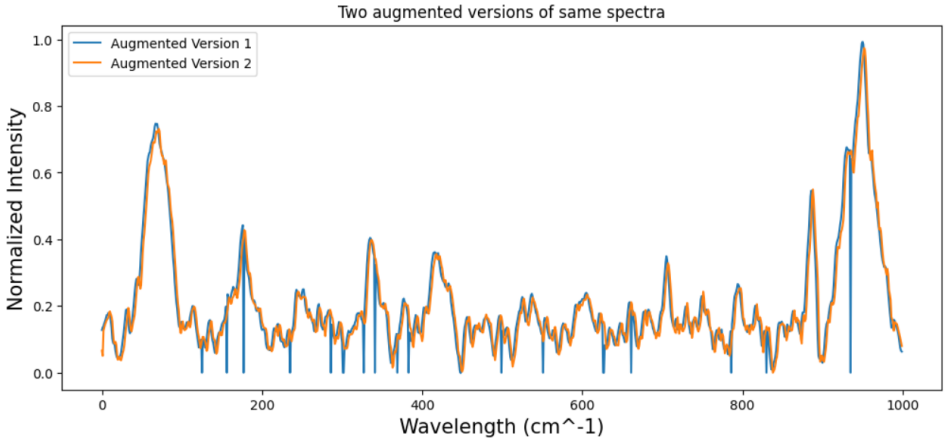
In the final pre-training setup, the active augmentations were the following:

- **Gaussian Noise Injection:** Applied when the first random branch selects the noise path, with a standard deviation of 0.01. This simulates additive detector noise and small stochastic fluctuations in the measured Raman intensity.
- **Average Blurring:** Applied as an alternative branch to noise using a one-dimensional average pooling filter with a random kernel size between 2 and 5 spectral points. This mimics mild instrumental broadening or reduced optical resolution.
- **Wavenumber-Axis Shift:** Applied with a discrete random shift within ± 5 spectral points. This simulates small calibration offsets between instruments and prevents the model from relying too strongly on absolute index positions.
- **Random Zeroing:** Applied by independently setting approximately 2% of spectral points to zero. This acts as a sparse 1D dropout mechanism and makes the representation robust to missing or corrupted detector channels.
- **Random Intensity Scaling:** Applied as an alternative to random zeroing by multiplying the whole spectrum by a factor sampled between 0.9 and 1.1. This models acquisition-dependent intensity variations caused by laser power, exposure time, or sample coupling.

The implementation also included optional routines for Poisson shot noise, peak broadening, and contiguous region dropout. These were kept as configurable extensions but were not activated in the final Barlow Twins augmentation branch used for the reported experiments. This choice kept the perturbations strong enough to promote invariance, while avoiding excessive spectral distortion that could erase diagnostically meaningful Raman peaks.

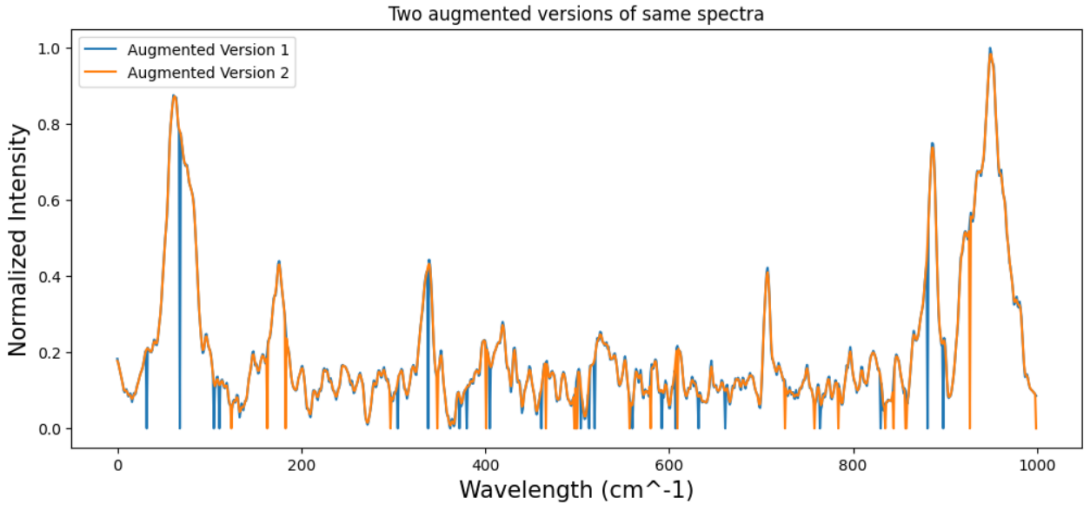


(a) Augmentation Comparison 1

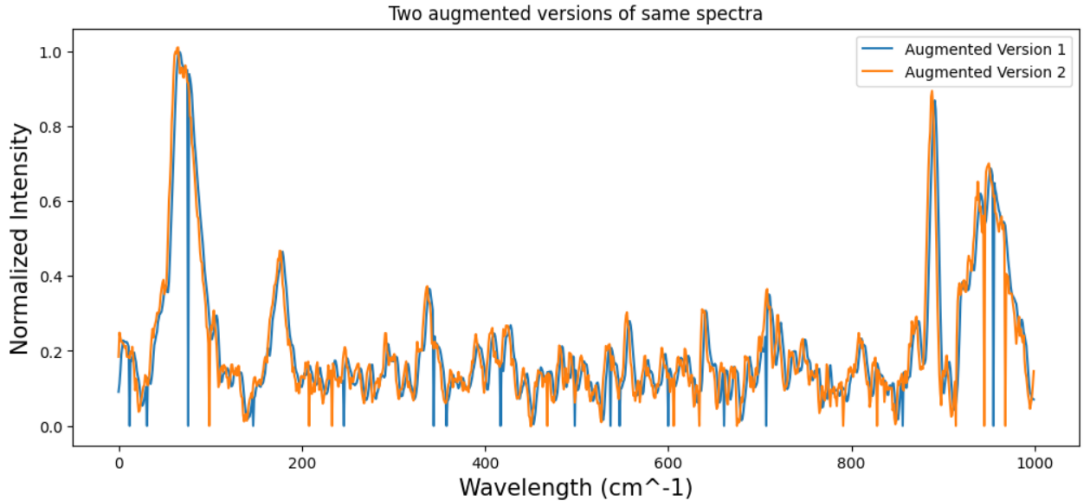


(b) Augmentation Comparison 2

Figure 4.2: First two examples of the stochastic Barlow Twins data augmentation pipeline applied to native Raman spectra. Each original signal is independently perturbed to create distinct augmented views while preserving its chemical identity.



(c) Augmentation Comparison 3



(d) Augmentation Comparison 4

Figure 4.2: Additional examples of the stochastic Barlow Twins augmentation pipeline. The network is trained to map all augmented views of the same spectrum to the same latent representation, achieving invariance to instrumental artifacts (continued).

4.3.3 The Projector Network and Cross-Correlation Objective

The intermediate representations extracted by the ResNet-34 backbone are fed into a Projector Network. This network consists of sequential `Dense-BatchNorm-ReLU` blocks, which expand the embeddings into a massive, high-dimensional latent space (e.g., 2048 dimensions). This expansion is a critical hyperparameter in Barlow Twins: projecting the features into such a vast space provides the necessary degrees of freedom to separate and decorrelate independent biochemical features without compression bottlenecks.

Given a batch of augmented spectral pairs, the network yields two distinct embedding matrices, one for each augmented view. After normalizing these embeddings along the batch dimension, their empirical cross-correlation matrix $\mathcal{C}$ is computed. Each element $\mathcal{C}_{ij}$ measures how strongly the i-th projected feature from the first view correlates with the j-th projected feature from the second view. The Barlow Twins loss function ($\mathcal{L}_{BT}$) drives this cross-correlation matrix to approximate the identity matrix:

$$\mathcal{L}_{BT} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2 \tag{4.1}$$

The first term of the equation is the invariance term: it pushes the diagonal entries of $\mathcal{C}$ toward one, meaning that corresponding latent dimensions should respond similarly for the two augmented versions of the same spectrum. The second term is the redundancy-reduction term: weighted by the hyperparameter λ, it penalizes non-zero off-diagonal correlations, discouraging different latent dimensions from encoding the same information. Together, these two terms prevent representation collapse while promoting compact, non-redundant biochemical descriptors.

4.3.4 Barlow Twins: Loss Convergence

The Barlow Twins network was pre-trained using an AdamW optimizer with a large batch size (e.g., 512) to ensure statistically stable cross-correlation matrices across the batch.

As shown in Figure 4.3, the loss smoothly converges over the training epochs. This steady, progressive descent is an excellent indicator of successful training: it demonstrates that the optimization of the decorrelation objective is working effectively, and that the network is successfully learning to mathematically align the heavily augmented views into invariant chemical representations without suffering from representation collapse.

More specifically, the decrease of the Barlow Twins loss indicates that the two components of the objective are being optimized jointly. The diagonal terms of the cross-correlation matrix are progressively encouraged to approach one, meaning that the two augmented versions of the same spectrum produce increasingly similar latent responses. At the same time, the off-diagonal terms are penalized, reducing redundancy between different embedding dimensions. Therefore, convergence does not simply mean that the network is minimizing a generic numerical error: it suggests that the encoder is learning representations that are simultaneously invariant to the selected spectral augmentations and less redundant across latent features.

The absence of sudden instability or divergence in the curve is also important. In self-supervised learning, a poorly balanced objective or overly destructive augmentations can lead to unstable optimization or uninformative representations. In this experiment, the smooth loss trajectory supports the choice of the augmentation pipeline and confirms that the Barlow Twins objective remained compatible with the physical structure of the Raman spectra.

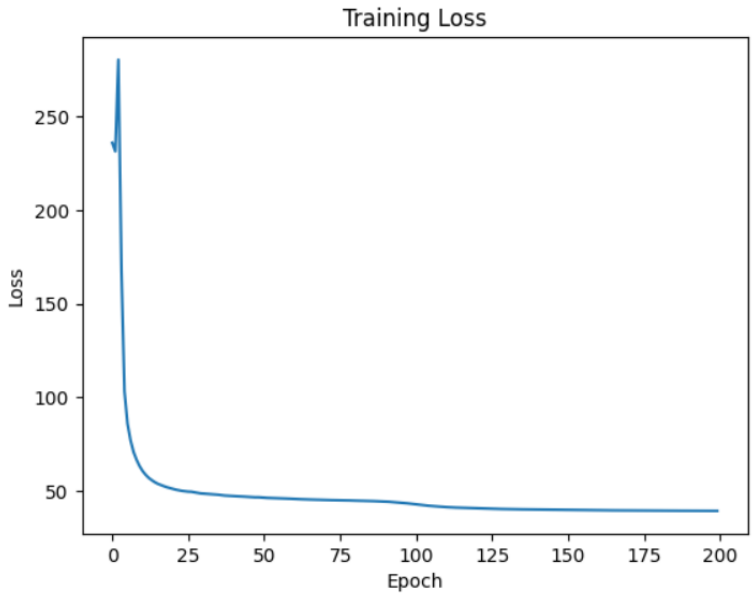


Figure 4.3: Evolution of the Barlow Twins loss over the pre-training epochs. The steady convergence indicates the progressive alignment of the augmented views and the successful reduction of latent redundancy.

4.4 Masked Autoencoders (MAE) for Spectral Reconstruction

The Spectral Masked Autoencoder (SMAE) operates on a completely different learning paradigm. Rather than learning invariances through contrastive augmentation, SMAE is a purely generative architecture that learns the physical correlations between chemical bonds by predicting missing data. Inspired by the Vision Transformer (ViT), the architecture was structurally adapted to interpret 1D Raman spectroscopy.

In the context of Raman signals, this reconstruction task has a direct physical interpretation. A spectrum is not a generic numerical vector: neighboring and distant wavenumber regions are linked by molecular vibrations, biochemical composition, fluorescence background, and acquisition noise. By hiding a large fraction of the input and asking the model to reconstruct it, the network is forced to infer these dependencies instead of memorizing isolated peak positions. This makes the pre-training objective especially suitable for a foundation model, because the encoder learns reusable representations before any diagnostic label is introduced.

Importantly, the decoder is used only during the self-supervised pre-training stage. Its role is to provide a reconstruction signal that teaches the encoder how to organize the spectral information. Once pre-training is completed, the decoder is discarded and only the trained encoder is retained as the transferable feature extractor for downstream classification tasks. In this sense, reconstruction is not the final objective of the thesis pipeline, but the mechanism through which the encoder learns chemically meaningful Raman representations.

4.4.1 1D Patching, Masking, and Positional Embeddings

Unlike CNNs, which process data locally through sliding windows, Transformers process data as a sequence of discrete tokens. In the executed MAE pre-training notebook, each harmonized Raman spectrum is represented as a 1000-point vector and divided into 10 consecutive, non-overlapping patches of 100 points each. Each patch therefore corresponds to a relatively broad spectral interval, rather than to a single isolated peak. This design forces each token to summarize a local portion of the Raman fingerprint while still keeping the sequence short enough for efficient Transformer self-attention.

A random masking operation is then applied at the patch level with a mask ratio of 50%. Instead of corrupting individual intensity values, the model removes

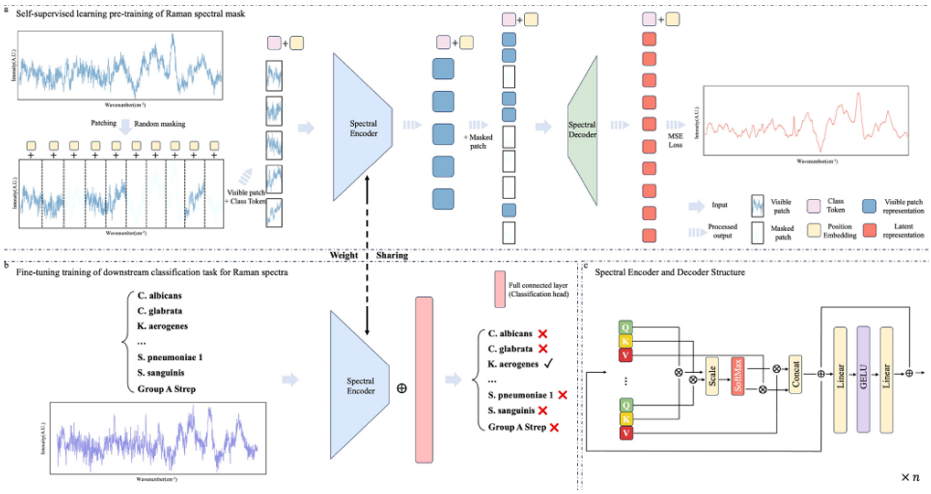


Figure 4.4: Workflow of the Spectral Masked Autoencoder (SMAE). The Raman spectrum is divided into 1D patches, partially masked, and encoded using only the visible tokens. A lightweight Decoder reconstructs the missing signal during pre-training; afterward, the Decoder is discarded and only the Encoder is transferred to downstream tasks.

entire 100-point spectral segments. For a 1000-point spectrum, this means that approximately 5 out of 10 patches are hidden at each forward pass. The missing content may include complete Raman bands, shoulders, or portions of the local baseline, making the task substantially harder than point-wise interpolation. The encoder receives only the visible patches, which reduces the computational cost of self-attention and forces the network to infer the missing regions from broader spectral context.

Because the self-attention mechanism processes all visible patches simultaneously without an inherent sense of order, the network would normally not know whether a token belongs to the beginning, middle, or end of the Raman fingerprint region. To solve this, 1D positional embeddings are added to the patch embeddings. This crucial step anchors each token to its specific spectral location, allowing the model to distinguish chemically different regions and to learn long-range relationships between distant Raman bands.

After encoding, the missing locations are reintroduced through learnable mask tokens and passed to a lightweight decoder. The decoder reconstructs the full spectrum, while the loss is computed only on the masked patches. In this way, the model is rewarded for predicting physically plausible missing Raman information, not for copying the visible input.

Table 4.2: Best Spectral MAE pre-training configuration selected for Raman reconstruction.

Parameter	**Value**
Spectrum size	1000 points
Patch size	100 points
Number of patches	10
Mask ratio	50%
Training epochs	500
Encoder dimension	256
Encoder depth	8 layers
Encoder heads	8
Encoder MLP dimension	512
Decoder dimension	128
Decoder depth	1 layer
Decoder heads	8
Optimizer	AdamW, learning rate 1×10^{-4}
Scheduler	Cosine annealing
Total parameters	4,485,604

4.4.2 Selected Best MAE Configuration

The best-performing MAE model was obtained after a dedicated pre-training run of 500 epochs on the 1000-point fingerprint corpus, comprising 331,249 spectra. This configuration was not selected arbitrarily: it was retained after several comparative trials involving different patch sizes, masking ratios, and decoder capacities. In particular, patch sizes of 20, 25, 50, and 100 spectral points were explored, together with more aggressive masking ratios such as 75% and decoder variants with greater depth, including a 2-layer decoder.

The final setting reported in this thesis corresponds to the configuration that produced the best downstream classification behavior after transfer, not merely the lowest reconstruction loss during pre-training. Empirically, the 100-point patch size provided a better balance between local Raman morphology and global fingerprint structure, while the 50% masking ratio made the reconstruction task sufficiently challenging without removing too much chemical context. Similarly, the 1-layer decoder was preferred because it preserved the asymmetric MAE principle: the decoder remained weak enough to prevent trivial reconstruction, forcing the encoder to learn more transferable spectral representations. This final design is also conceptually close to the configuration proposed in the original MAE paper [20], which emphasizes an expressive encoder, a lightweight decoder, and a reconstruction objective applied only to masked content. The complete configuration extracted from the executed notebook is summarized in Table 4.2.

4.4.3 The Asymmetric Architecture: Encoder vs. Decoder

As illustrated in Figure 4.4, a defining and highly engineered feature of the MAE is its heavily asymmetric design, which separates the processing into two vastly different networks:

- **The Spectral Encoder:** During pre-training, a 50% masking ratio is uniformly applied to the 10-patch sequence. The masked patches are entirely discarded from the encoder input, so the encoder processes only about five visible spectral tokens per spectrum. These tokens are passed through an 8-layer Transformer Encoder with 256-dimensional embeddings, 8 attention heads, and a 512-dimensional MLP block. Processing only the visible fraction of the spectrum reduces the computational overhead and forces the encoder to model relationships between distant Raman regions. This attention mechanism allows the network to learn that the presence or absence of a molecular band in one interval can be correlated with another interval, even when intermediate spectral patches are hidden.

- **The Lightweight Decoder:** Once the visible patches are encoded into a latent representation, they are combined with "mask tokens"—shared, learnable vectors that indicate the positions of the missing 100-point segments. This restored sequence is then fed into a much smaller decoder, using 128-dimensional embeddings and only 1 Transformer layer. This structural bottleneck is intentional: if the decoder were too powerful, it could solve much of the task through local interpolation. By keeping the decoder shallow and narrow, the deep encoder is forced to compress the physically relevant Raman structure needed to reconstruct the masked patches.

4.4.4 Pre-Training Objective and Denoising Properties

The network is optimized using the Mean Squared Error (MSE) computed *only* on the masked patches:

$$\mathcal{L}_{MAE} = \frac{1}{|M|} \sum_{i \in M} (x_i - \hat{x}_i)^2 \tag{4.2}$$

Beyond pure feature extraction, this generative pretext task imparts a powerful and inherent denoising capability to the model. Because stochastic CCD noise is entirely random and uncorrelated across the wavenumber axis, the network cannot predict it. By forcing the network to predict missing intensities based on global physical rules, the MAE inherently learns to ignore the noise and output only clean theoretical

peak intensities, substantially increasing the Signal-to-Noise Ratio (SNR) of the reconstructed output.

4.5 Pre-Training Dynamics and Temporal Reconstruction

The models were trained on the unified multi-domain pre-training corpus. The following sections detail the convergence of the loss functions and provide visual proofs of the successful representation learning process.

4.5.1 MAE: Training and Spectral Reconstruction

For the MAE, tracking the MSE loss and masked shape similarity over the full 500-epoch pre-training run (Figures 4.5 and 4.6) reveals the network's generative learning dynamics. The steep initial decline reflects the rapid acquisition of the low-frequency structure of the spectra, while the slower subsequent phase corresponds to the more difficult refinement of local Raman morphology inside the masked patches. This behavior is expected for a reconstruction-based pretext task: learning the global envelope of a spectrum is easier than predicting chemically meaningful peak details from incomplete context.

The most direct visualization of this physical learning process is the temporal evolution of the spectral reconstruction. Figures 4.7–4.10 report the same reconstruction task at four representative checkpoints of the 500-epoch training run: the first epoch, two intermediate stages, and the final epoch. This sequence makes the learning dynamics visually explicit, showing how the decoder evolves from a coarse reconstructor into a model capable of recovering increasingly meaningful Raman morphology in the masked regions.

The comparison is intentionally performed on spectra originating from different sources, so that the qualitative validation is not restricted to a single acquisition domain. In each plot, the blue curve represents the original spectrum, while the red curve represents the MAE reconstruction obtained from the visible context. The yellow shaded vertical regions correspond to the masked spectral intervals, i.e., the parts of the signal that are removed from the encoder input and must be reconstructed by the decoder. Therefore, good agreement between the red and blue curves inside the yellow areas indicates that the model is not simply copying the input, but is inferring the missing signal from the learned relationships between neighboring and distant Raman bands.

Before inspecting these qualitative reconstructions, the global training curves provide a compact summary of the optimization process. The MSE curve quantifies the numerical reconstruction error on the masked content, while the masked cosine similarity evaluates whether the reconstructed regions preserve the overall spectral shape. Reading these two metrics together is important: a decreasing MSE indicates better point-wise reconstruction, whereas an increasing cosine similarity confirms that the reconstructed masked regions progressively become more consistent with the morphology of the original Raman signal.

The four checkpoints highlight complementary aspects of the learning process. At the beginning of training, the decoder mainly captures smooth low-frequency trends and tends to miss narrow vibrational structures. By the intermediate stages, the reconstructed signal starts to align with the principal peaks and baseline morphology, although some fine amplitude errors remain. At the final epoch, the reconstructed curves show a much tighter agreement with the original spectra, especially in the masked intervals, suggesting that the encoder has internalized a more transferable representation of Raman spectral structure.

Overall, the progression from Epoch 1 to Epoch 500 shows a clear transition from coarse baseline recovery to more chemically informative reconstruction. Early predictions mainly reproduce the smooth envelope of the spectrum, intermediate predictions recover the principal Raman bands, and final predictions refine the local morphology of masked peaks. This behavior supports the use of the pre-trained MAE encoder as a transferable representation extractor for downstream diagnostic tasks.



Figure 4.5: Reconstruction loss (MSE) of the Spectral MAE during pre-training. The rapid initial decrease indicates that the model quickly learns the global spectral structure, while the slower convergence phase corresponds to the refinement of sharper Raman features.

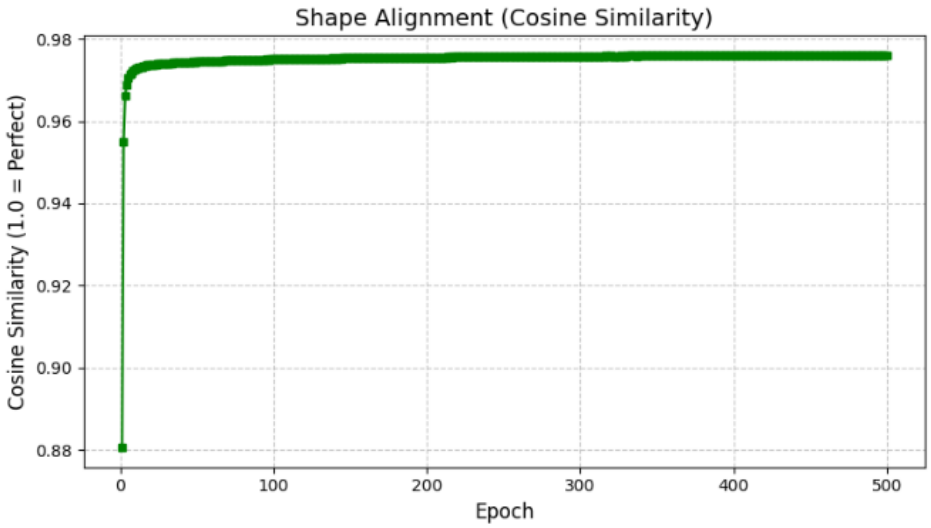


Figure 4.6: Shape-alignment metric for the Spectral MAE during pre-training. The increasing cosine similarity indicates that the reconstructed spectra progressively preserve the morphology of the original Raman signals.

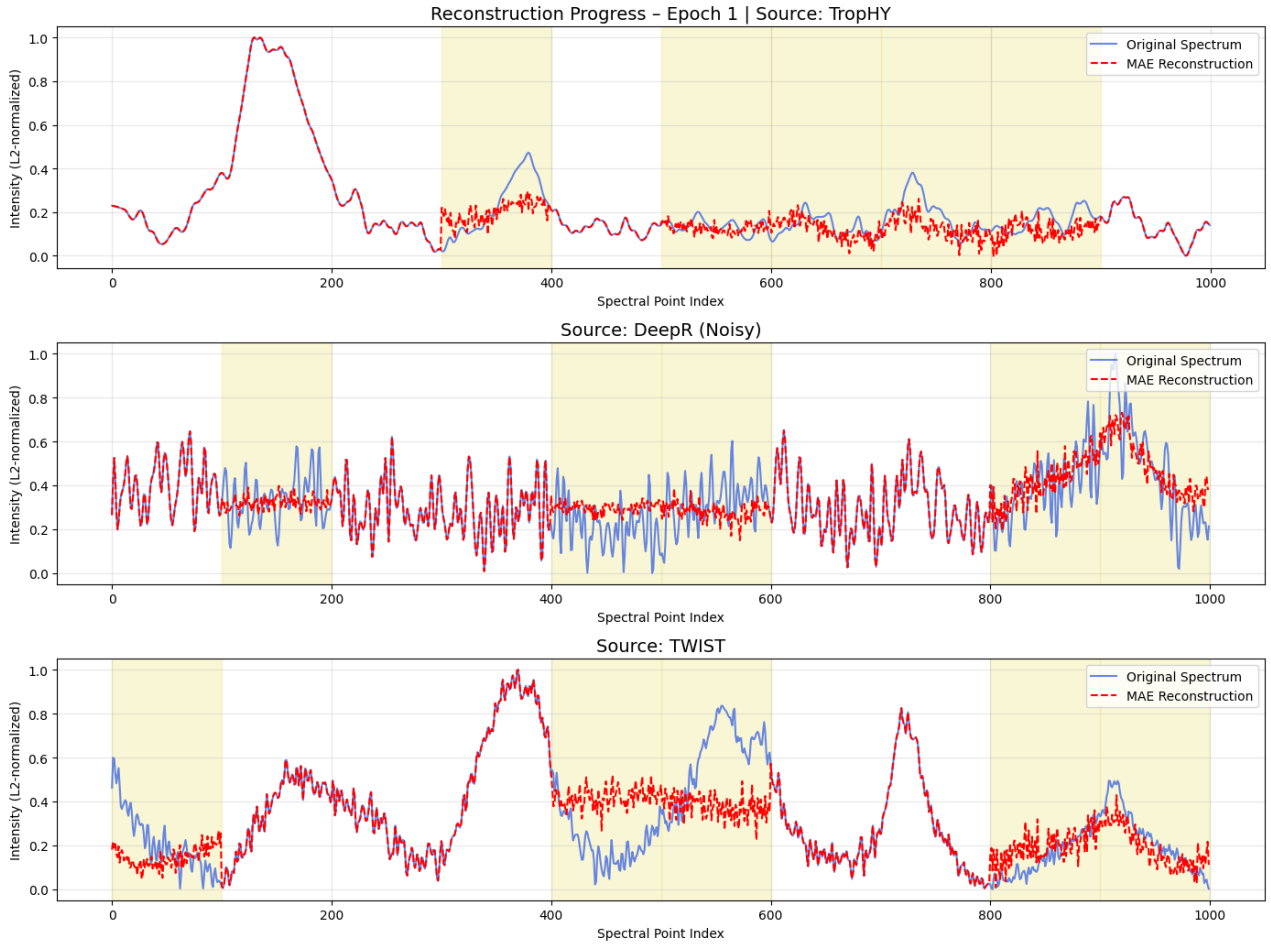


Figure 4.7: MAE spectral reconstruction after Epoch 1. The prediction captures only a coarse approximation of the spectral morphology.

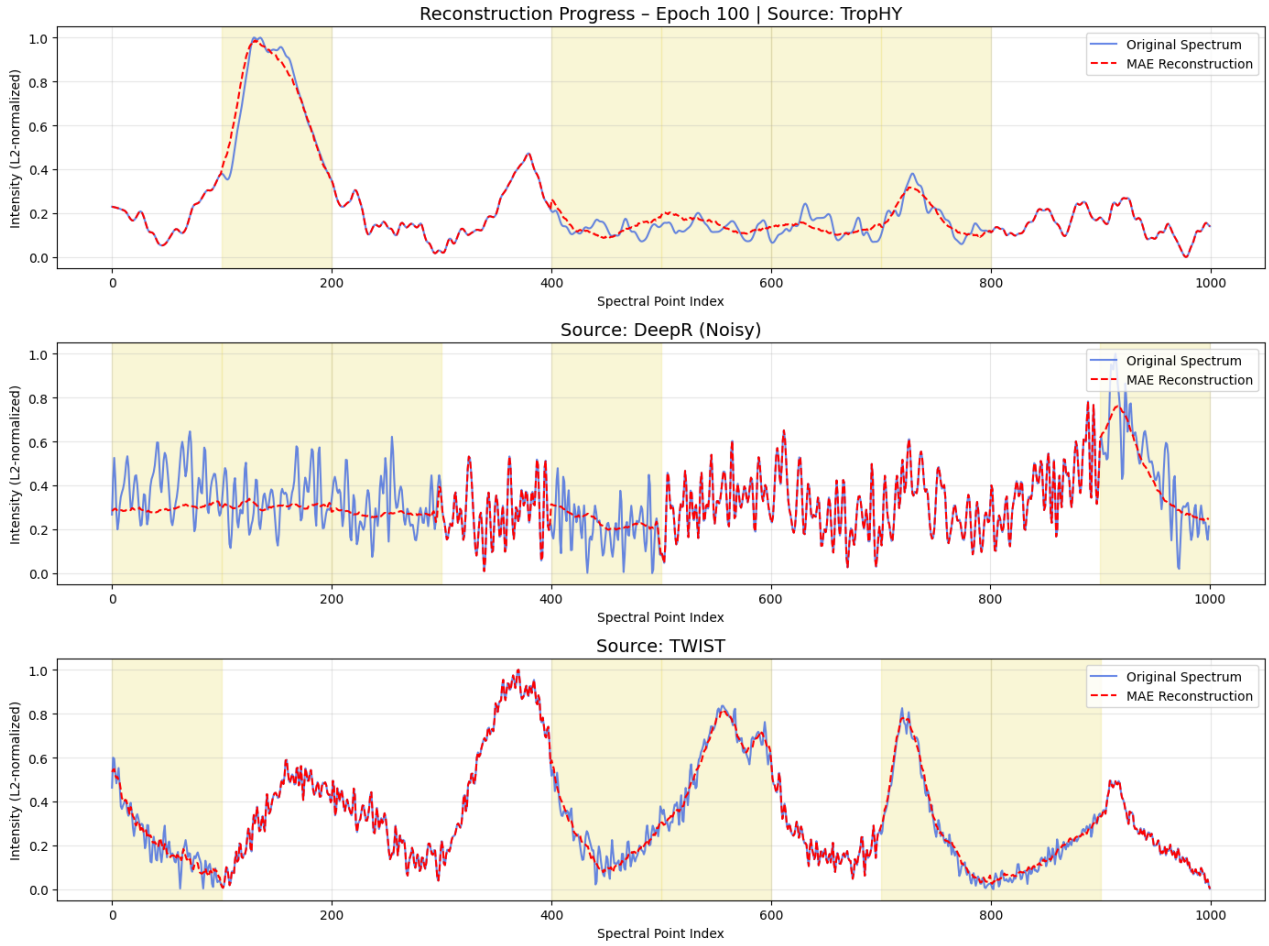


Figure 4.8: MAE spectral reconstruction after Epoch 100. The reconstruction shows improved peak localization and a more stable baseline.

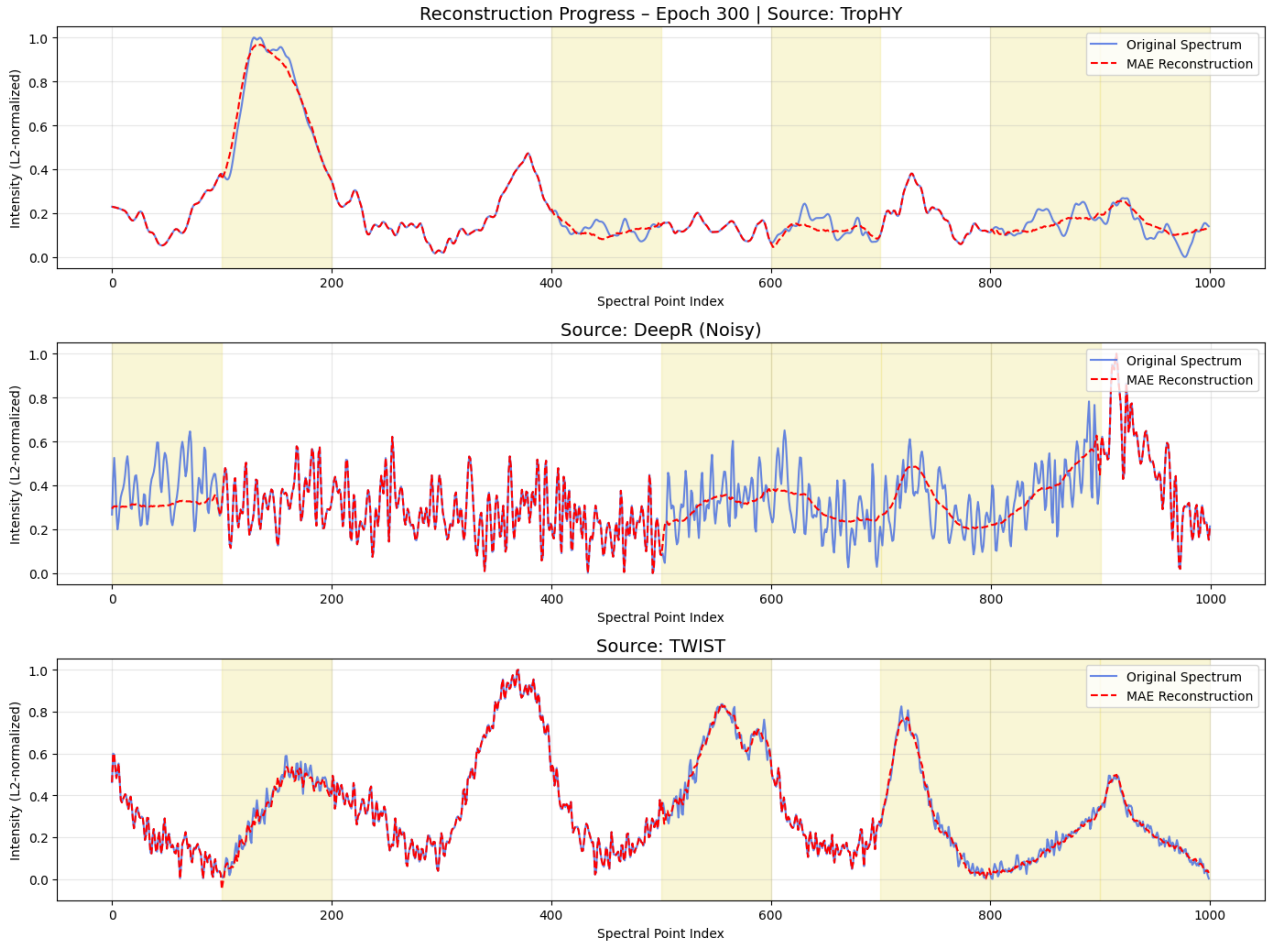


Figure 4.9: MAE spectral reconstruction after Epoch 300. This late-stage checkpoint shows a close agreement with the original spectra, including masked regions.

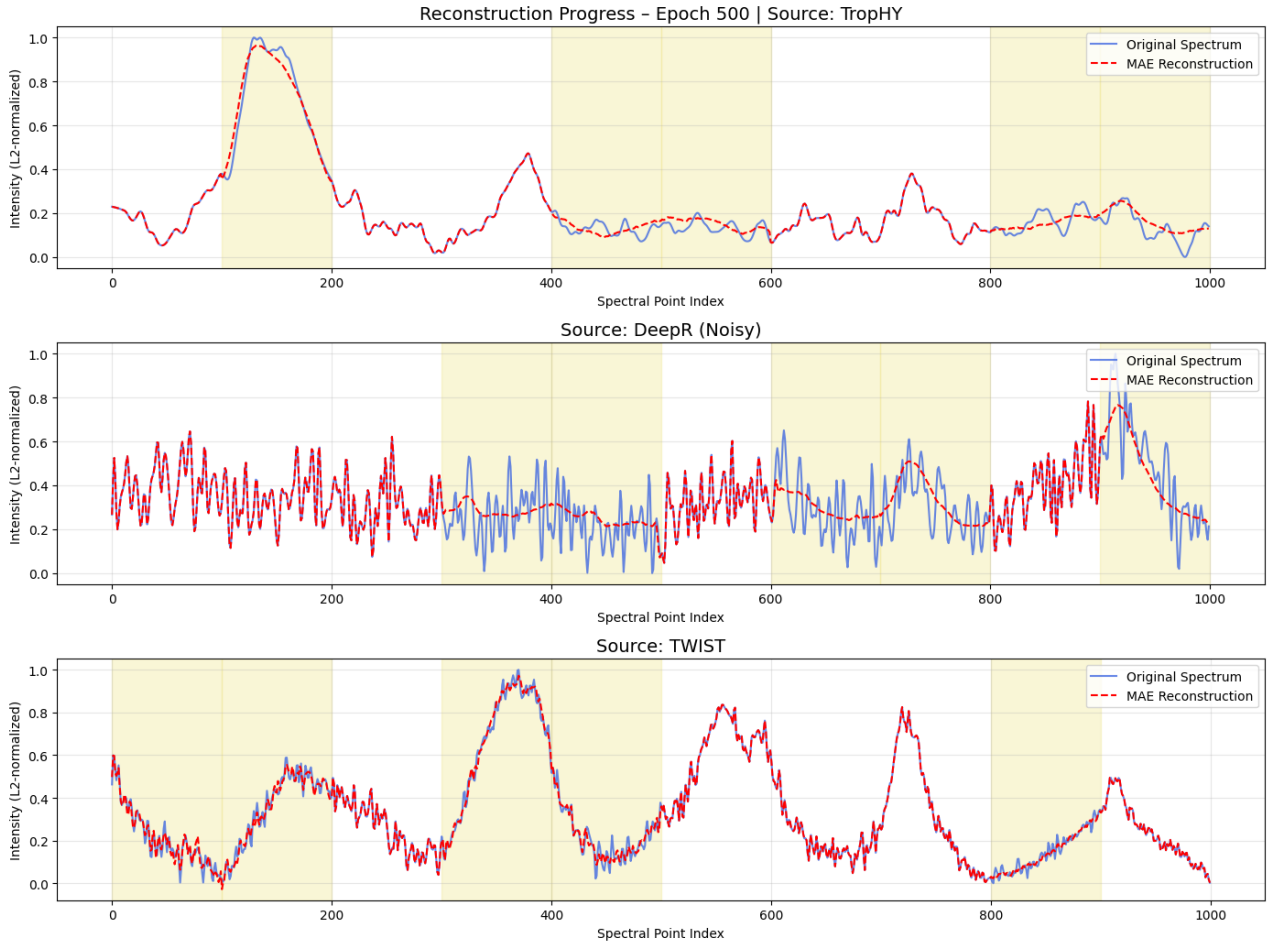


Figure 4.10: MAE spectral reconstruction after Epoch 500. The final checkpoint illustrates the reconstruction quality reached at the end of the pre-training run.

Chapter 5

Downstream Fine-Tuning and Experimental Results

5.1 Fine-Tuning Methodology

Following the self-supervised pre-training phase, the learned encoders were transferred to supervised downstream tasks in order to evaluate whether the representations acquired from unlabeled Raman spectra were useful for diagnostic classification. Two foundation models were considered throughout this chapter: **Barlow Twins**, representing the contrastive redundancy-reduction family, and the **Spectral Masked Autoencoder (SMAE)**, representing the generative masked-reconstruction family.

The comparison was designed to answer two complementary questions. First, each architecture was evaluated against its own randomly initialized counterpart, in order to measure whether self-supervised pre-training provided an advantage over training the same model from scratch. Second, the two pre-trained models were compared against each other, in order to understand whether contrastive invariance learning or masked spectral reconstruction produced more transferable Raman representations.

5.1.1 Architectural Adaptation

For downstream classification, the pre-trained models were not optimized with their original self-supervised objectives. Instead, the representation learned during pre-training was reused as the transferable backbone and a new supervised head was attached for the target task. In Barlow Twins, the transferred component is the 1D-ResNet representation learned through the redundancy-reduction objective; during fine-tuning, the model is trained with a supervised classification loss rather than with the Barlow Twins cross-correlation loss. In SMAE, the Transformer encoder learned through masked reconstruction is reused analogously, while the reconstruction-specific decoder is not part of the downstream classifier. Therefore, the downstream models are best described as pre-trained backbones adapted with task-specific supervised heads, rather than as the original self-supervised networks used unchanged.

The supervised head was kept lightweight to reduce overfitting on limited biomedical datasets, but its output dimensionality and exact configuration depended on the downstream task. In the IBD experiments the head mapped the representation to two diagnostic classes, whereas in BACTERIA-ID it mapped the representation to the multiclass pathogen-identification labels. Across tasks, the head generally combined regularization layers and a final classification layer, so that most of the representational burden remained in the transferred backbone.

A key difference between the two backbones concerns the spectral range used at fine-tuning time. The **Barlow Twins model was trained and evaluated on the full available Raman range**, consistently with its convolutional architecture and its pre-training strategy based on whole-spectrum augmented views. In contrast, the **SMAE model was evaluated on the fingerprint region only**, matching the 1000-point fingerprint setting used during the best masked-autoencoder pre-training run. This distinction is important because it means that the two models do not simply differ architecturally: they also exploit different spectral views of the same downstream tasks.

5.1.2 Optimization Strategy

All downstream experiments followed the same general transfer-learning principle, but the exact optimization schedule varied across architecture and task. For pre-trained models, the backbone was initialized from the self-supervised checkpoint and the supervised head was initialized from scratch. In the runs where it was useful, training followed a two-stage Linear Probing to Fine-Tuning strategy:

1. **Linear Probing (LP):** the pre-trained backbone was frozen and only the supervised head was trained. This stage tests whether the frozen representation is already informative for the downstream labels.

2. **Fine-Tuning (FT):** part or all of the backbone was subsequently unfrozen and optimized with the supervised head, usually using a lower learning rate to avoid destroying the pre-trained representation.

This LP–FT schedule was not imposed universally. Some downstream configurations used only direct supervised training or a simplified fine-tuning schedule, depending on the notebook and on the stability of the specific task. The non-pre-trained controls were handled differently from the transfer-learning models: they were initialized randomly and trained directly with the supervised objective, using the corresponding preprocessing, classifier design, regularization, and data augmentation. Therefore, the comparison *within* each model family isolates the effect of

pre-training, while the comparison *between* Barlow Twins and SMAE evaluates the relative transferability of the two self-supervised objectives.

5.1.3 Training-Time Augmentation

To reduce overfitting, data augmentation was applied only to the training split of each downstream fold. The augmentations were designed to reproduce realistic experimental variability in Raman acquisition:

- **Intensity scaling**, simulating variations in laser power, exposure time, or sample coupling;
- **Gaussian noise injection**, reproducing small detector-level fluctuations;
- **Spectral shifting**, modeling minor calibration offsets along the wavenumber axis;
- **Random masking/cutout**, forcing the classifier to rely on multiple biochemical regions instead of overfitting to a single dominant band.

5.2 Evaluation Metrics and Diagnostic Reading

The downstream tasks in this thesis are evaluated using complementary metrics because accuracy alone is not always sufficient to describe biomedical classification behavior. Accuracy measures the fraction of correctly classified samples and is intuitive when classes are balanced. However, in small biomedical datasets, and especially in multiclass settings, accuracy can hide uneven performance across classes. A model may obtain a reasonable accuracy while still performing poorly on minority or clinically important categories.

For this reason, Macro-F1 is used whenever possible. The F1-score combines precision and recall for each class, and the macro average assigns the same weight to every class regardless of its frequency. This is useful for Raman datasets where class imbalance may occur or where all diagnostic categories should be treated as equally relevant. In the BACTERIA-ID task, for example, a macro-averaged score is more informative than accuracy alone because it penalizes models that perform well only on the most frequent or easiest bacterial classes.

Confusion matrices provide a more detailed class-level view. Each row corresponds to the true class and each column to the predicted class, making it possible to identify systematic errors. In binary IBD classification, the confusion matrix reveals whether the model confuses healthy controls with IBD spectra or whether

errors are concentrated in one direction. In BACTERIA-ID, it shows whether the model learns a broad pathogen separation or whether specific bacterial strains remain difficult to distinguish.

For binary classification, probability-based diagnostics are also important. The ROC curve evaluates how well the model ranks positive samples above negative samples across decision thresholds, while the distribution of predicted probabilities shows whether the classifier is confident or whether many samples lie close to the decision boundary. These analyses are particularly useful for small clinical datasets, where a single threshold can give an incomplete picture of model behavior.

The evaluation strategy therefore combines aggregate metrics, class-level matrices, probability diagnostics, and fixed-test cross-validation. This combination reduces the risk of drawing conclusions from a single favorable split and provides a more reliable assessment of whether self-supervised pre-training improves downstream Raman classification.

5.3 Inflammatory Bowel Disease (IBD) Classification

5.3.1 Task Definition and Evaluation Protocol

The IBD experiment evaluates binary classification between healthy controls (HC) and inflammatory bowel disease (IBD) spectra. The dataset contains 228 spectra in total, with 93 HC spectra and 135 IBD spectra. This task is clinically challenging because the spectral differences are subtler than in pathogen identification: systemic plasma alterations related to inflammation are expected to produce weaker and more heterogeneous biochemical signatures than species-level bacterial fingerprints.

A specific evaluation protocol was adopted to make the estimates more stable. First, a single stratified split was created and a final test set was separated at the beginning of the experiment. This test set was never used for model selection. Then, instead of reporting only one train/validation split, the remaining data were repeatedly divided into training and validation folds by stratified 5-fold cross-validation. Each fold was trained on a different train/validation partition, but all folds were evaluated on the same fixed test set. Therefore, the reported cross-validation statistics measure how sensitive the fine-tuning procedure is to the choice of supervised training and validation samples, while keeping the final test population constant.

The two architectures were evaluated with the spectral representation used in their corresponding pre-training configuration. Barlow Twins was fine-tuned on the

full 1000-point Raman range. SMAE was fine-tuned on the 1000-point fingerprint representation. For this reason, the comparison is not only architectural: it also reflects the best spectral view selected for each self-supervised strategy.

Table 5.1: IBD dataset and evaluation protocol used for downstream fine-tuning. The fixed test set was held out before cross-validation and reused only for final evaluation of each fold.

Model	**Input Range**	**Total**	**Fixed Test**	**CV Pool**	**CV Strategy**
Barlow Twins	Full range	228	35	193	Stratified 5-fold
SMAE	Fingerprint	228	35	193	Stratified 5-fold

For SMAE, the initial hold-out split was explicitly defined as 70% training, 15% validation, and 15% test, corresponding to 159 training spectra, 34 validation spectra, and 35 test spectra. The train and validation subsets were then merged to form the 193-sample cross-validation pool. The same conceptual protocol was used for Barlow Twins: 35 samples were held out as the final test set, while the remaining 193 samples were used for stratified train/validation folds.

5.3.2 Fine-Tuning Strategies and Data Augmentation

The fine-tuning procedure was architecture-specific. For Barlow Twins, the pre-trained convolutional encoder was loaded from the self-supervised checkpoint and followed by a lightweight supervised classifier. The pre-trained run used a two-stage strategy. In the first stage, the backbone was frozen and only the final classification layer was trained for a warm-up linear-probing phase. In the second stage, the backbone was partially unfrozen: only the last 10% of the encoder layers were made trainable, while Batch Normalization layers were kept frozen to avoid destabilizing the pre-trained statistics on the small IBD dataset. The classifier included Gaussian noise with standard deviation 0.01, dropout of 0.6, and an L2 penalty of 0.05 on the final dense layer. Class weights, early stopping on validation loss, and learning-rate reduction on plateau were used to reduce overfitting. The randomly initialized Barlow Twins control used the same downstream setting but without loading the self-supervised backbone.

For SMAE, the generative decoder was discarded and the Transformer encoder was connected to a compact MLP head. The head applied LayerNorm, a hidden linear layer of size 128, GELU activation, dropout of 0.4, and a final binary classification layer. The pre-trained SMAE was also optimized with a two-stage LP–FT strategy: first the encoder was frozen and only the head was trained, then the full

model was unfrozen and fine-tuned with a lower learning rate. AdamW optimization, class-balanced cross-entropy, ReduceLROnPlateau, and early stopping were used during training. The from-scratch SMAE control used the same classifier but initialized the ViT encoder randomly.

Training-time data augmentation was applied only to the training subset of each split or fold. For SMAE, the augmentation pipeline included random intensity scaling between 0.95 and 1.05, light Gaussian noise with scale 10^{-4}, one-dimensional cutout with 50% probability masking a continuous 5% region of the spectrum, and random spectral shifting by -2 to $+2$ points with 50% probability. For Barlow Twins, the downstream classifier used Gaussian noise directly in the model as an additional regularization layer, together with strong dropout and L2 regularization.

5.3.3 Hold-Out Evaluation

The first evaluation step was a single hold-out experiment. This step was useful as an initial diagnostic check: it verified that the fine-tuning code, the supervised head, the class weighting, and the data augmentation strategy were able to learn from the IBD labels before moving to the more stable cross-validation protocol. The numerical results are reported only in Table 5.3, while the corresponding training behavior and confusion matrices are shown in the following figures.

Table 5.2: IBD hold-out split used as the initial single-split evaluation. For SMAE, the split was explicitly 70% training, 15% validation, and 15% test. For Barlow Twins, the same final-test logic was then retained for the fixed-test cross-validation protocol.

Model	**Input Range**	**Train**	**Validation**	**Test**
SMAE	Fingerprint	159	34	35
Barlow Twins	Full range	–	–	35

Table 5.3: IBD single hold-out results. Numerical performance values are reported here rather than in the text. The Barlow Twins row corresponds to the single fixed-test evaluation associated with the confusion matrix in Figure 5.2.

Architecture	**Init.**	**Input**	**Accuracy**	**Macro-F1**
Barlow Twins	Pre-trained	Full range	82.86%	81.64%
Barlow Twins	Scratch	Full range	60.00%	37.50%
SMAE	Pre-trained	Fingerprint	94.29%	94.05%
SMAE	Scratch	Fingerprint	51.43%	47.11%

5.3.4 Training Curves, Confusion Matrices, and Probability Diagnostics

After the initial hold-out check, the diagnostic plots were used to understand how each architecture behaved during fine-tuning and testing. These figures are reported before the cross-validation summary because they explain why the fixed-test repeated-split protocol was adopted: IBD is a small and biologically subtle task, so a single accuracy value is not sufficient to determine whether the model is learning a robust disease-related signal or only adapting to a favorable split.

For Barlow Twins, Figure 5.1 is used to inspect optimization stability on the full Raman range. The two diagnostics are arranged vertically to make the training–validation gap easier to compare across epochs. This visualization is particularly useful in IBD because the model can fit the supervised folds quickly, while validation behavior reveals whether the learned decision rule is likely to generalize.

The class-level Barlow Twins test behavior is reported in Figure 5.2. This matrix is useful because it shows where the aggregate hold-out score comes from: correct HC and IBD predictions appear on the diagonal, while the off-diagonal cells show the residual confusion between the two diagnostic plasma-spectrum classes. The figure refers to the selected fixed-test Barlow Twins result used in Table 5.3, not to an average matrix across folds.

For SMAE, the following figures summarize the hold-out evaluation from three complementary viewpoints. Figure 5.3 first shows the training behaviour and is used to check whether optimization remains stable. Figure 5.4 then reports the final class assignments on the test set. Finally, Figures 5.5 and 5.6 look at the predicted probabilities: the ROC curve evaluates ranking performance across thresholds, while the histogram shows how the test spectra are distributed with respect to the default 0.5 decision threshold.

5.3.5 Fixed-Test Cross-Validation Results

The final IBD comparison used the more stable evaluation protocol: the initial test set was kept fixed, while the remaining spectra were repeatedly split into train and validation folds. This avoids basing the conclusion on a single favorable split and provides a clearer estimate of how much each method depends on the selected training and validation samples. The aggregate cross-validation results are summarized in Table 5.4, and the fold-wise Barlow Twins test accuracies are reported in Table 5.5 because the best Barlow Twins confusion matrix comes from one specific fold.

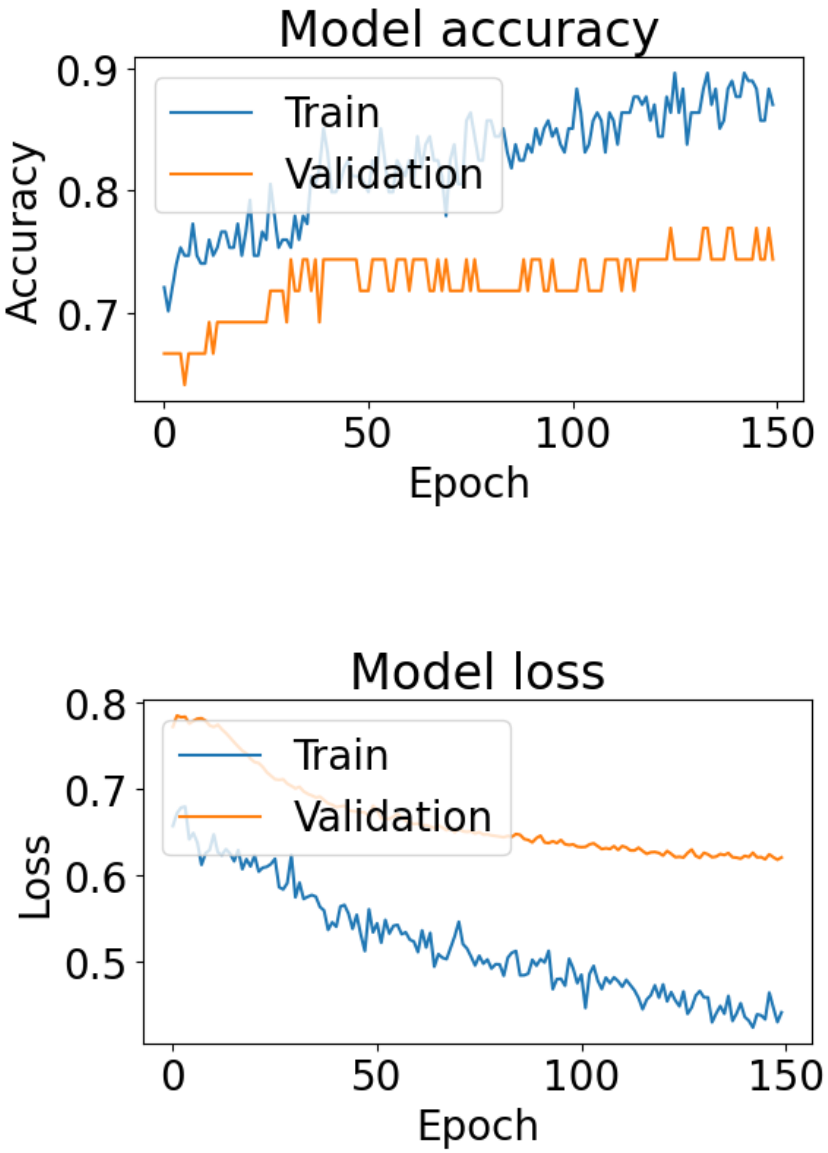


Figure 5.1: Barlow Twins IBD training diagnostics arranged vertically. The upper panel reports accuracy and the lower panel reports loss for the training and validation subsets during full-range fine-tuning.

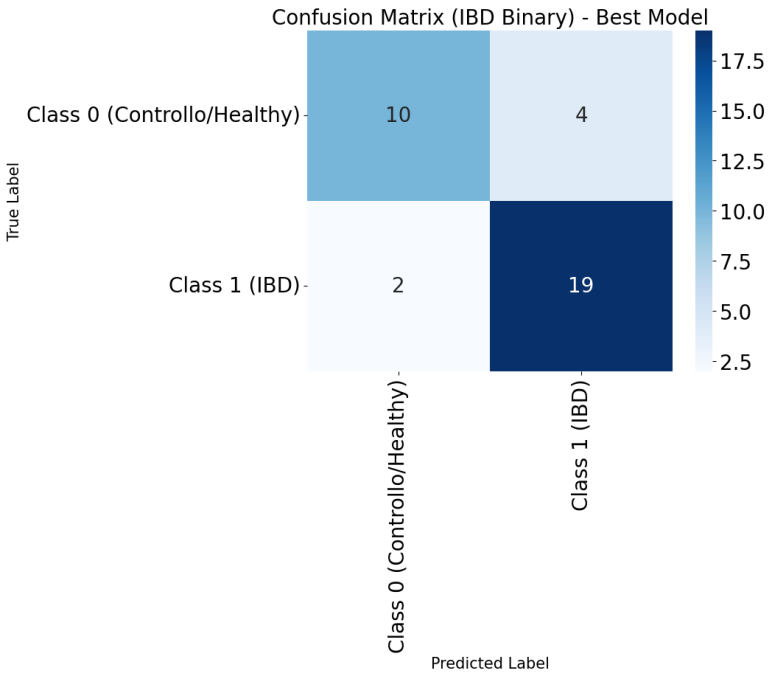


Figure 5.2: Barlow Twins confusion matrix on the IBD fixed test set. Rows indicate true labels and columns indicate predicted labels.

Table 5.4: IBD fixed-test cross-validation results. Each fold uses a different train/validation split from the 193-sample cross-validation pool and is evaluated on the same 35-sample test set.

Architecture	**Init.**	**Input**	**Accuracy**	**Macro-F1**
Barlow Twins	Pre-trained	Full range	79.43% ± 2.14	–
Barlow Twins	Scratch	Full range	60.00% ± 0.00	–
SMAE	Pre-trained	Fingerprint	89.14% ± 3.33	88.73% ± 3.69
SMAE	Scratch	Fingerprint	64.00% ± 9.66	58.79% ± 12.55

Table 5.5: IBD fold-wise test accuracy for Barlow Twins under the fixed-test protocol. The best pre-trained fold is the one used for Figure 5.2.

Initialization	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**
Pre-trained	77.14%	80.00%	82.86%	77.14%	80.00%
Scratch	60.00%	60.00%	60.00%	60.00%	60.00%

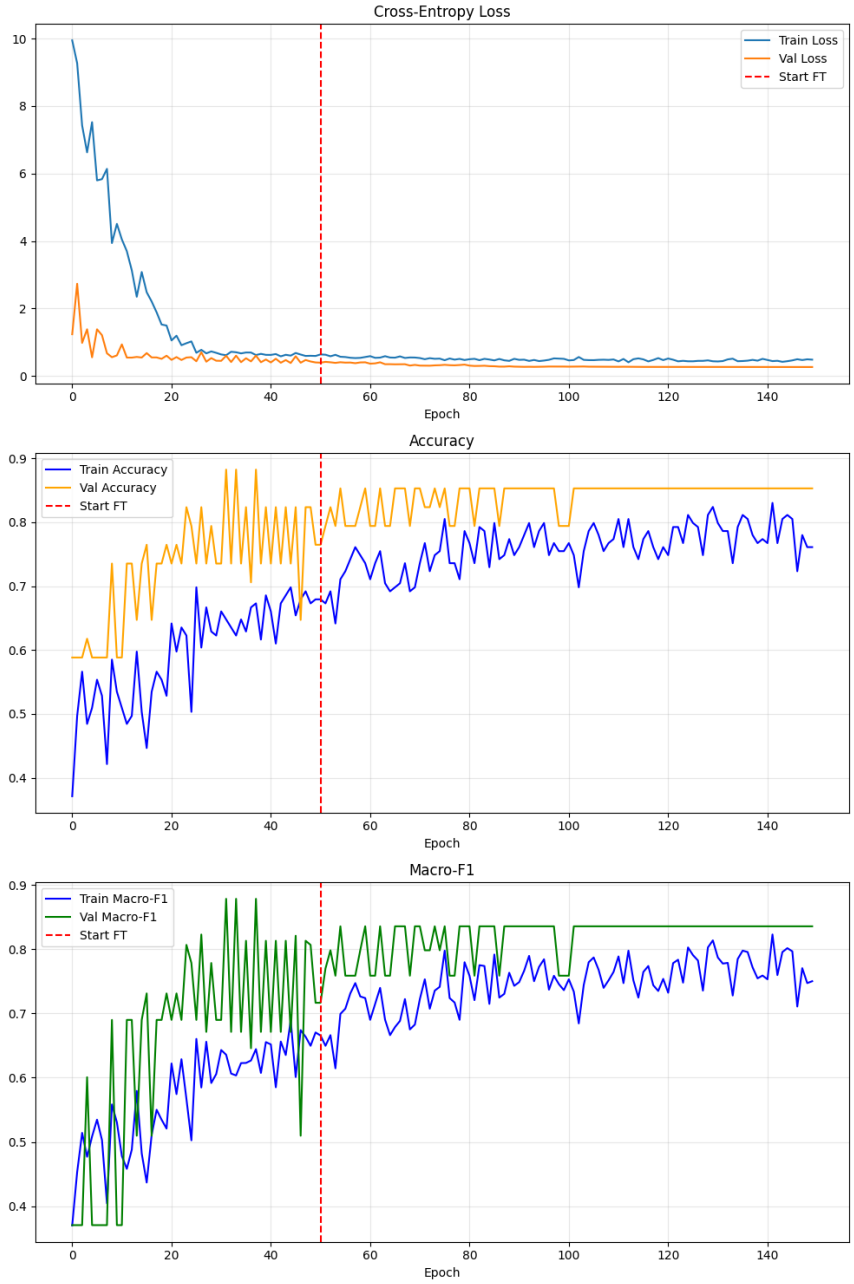


Figure 5.3: SMAE IBD training diagnostics on the fingerprint region during supervised adaptation.

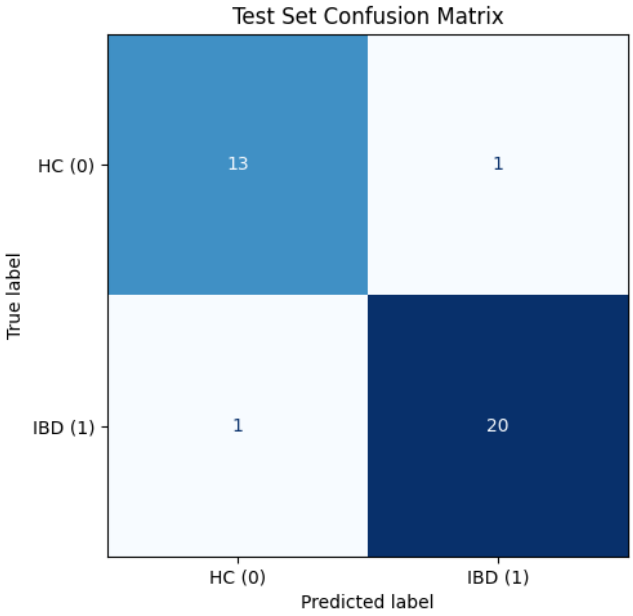


Figure 5.4: SMAE confusion matrix on the IBD hold-out test set using the fingerprint region.

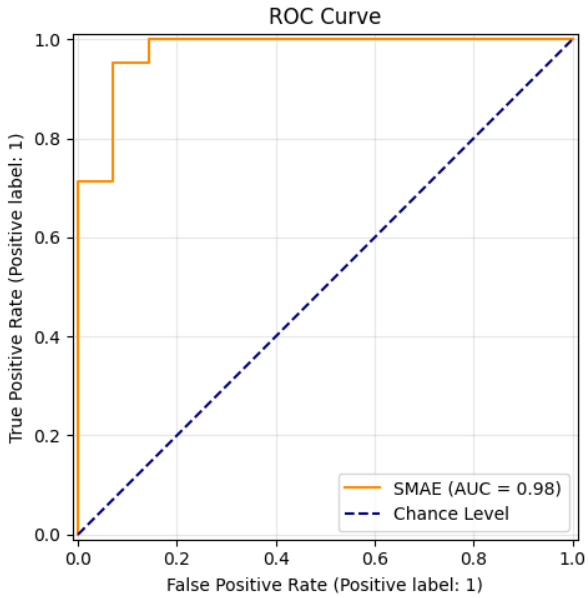


Figure 5.5: SMAE ROC curve on the IBD hold-out test set. The dashed diagonal line represents chance-level discrimination.

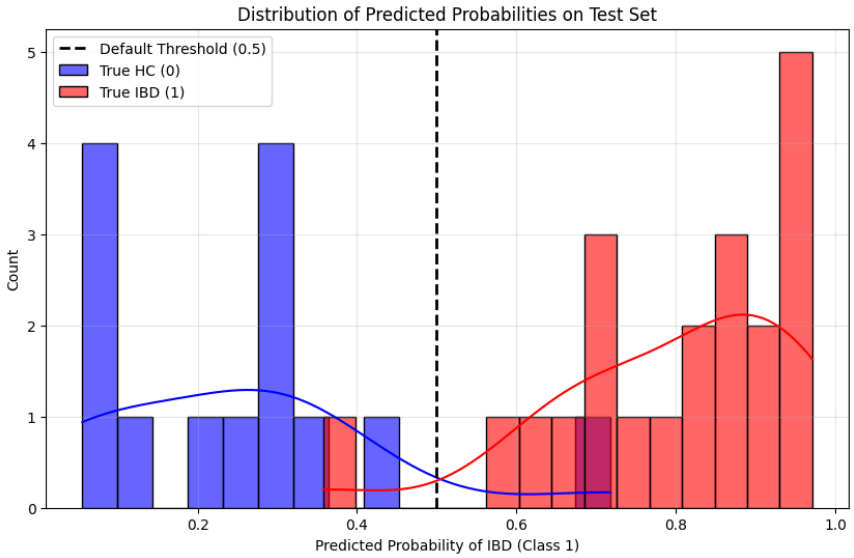


Figure 5.6: Distribution of SMAE predicted IBD probabilities on the hold-out test set. The vertical dashed line marks the default threshold of 0.5.

5.3.6 Model-Level Interpretation

The IBD experiment shows that self-supervised pre-training is beneficial for both model families. In the hold-out comparison, the pre-trained Barlow Twins model clearly improves over the corresponding scratch baseline, and the same pattern is observed again in the fixed-test cross-validation setting. This is important because the Barlow Twins scratch model remains close to a majority-class behavior, whereas the pre-trained model produces a more meaningful separation between HC and IBD spectra. The confusion matrix confirms that the full-range contrastive representation is not simply predicting the dominant class, but is able to recover both diagnostic categories with a substantially better balance.

The SMAE results are particularly strong on this IBD test set. In the hold-out evaluation, the pre-trained SMAE obtains the best overall performance among the tested IBD configurations, and the confusion matrix, ROC curve, and probability distribution all support the same interpretation: the model assigns low IBD probabilities to most HC spectra and high IBD probabilities to most IBD spectra, with only limited overlap around the decision threshold. This indicates that the fingerprint-region masked-reconstruction pre-training learned features that transfer well to the inflammatory classification task.

The fixed-test cross-validation results provide the more conservative interpretation. Even when the train/validation partition is changed across folds, both pre-trained models remain clearly above their scratch controls. SMAE keeps the strongest mean performance, while Barlow Twins shows a consistent gain over ran-

dom initialization. Therefore, the main conclusion is not only that pre-training improves the best single split, but that it also stabilizes downstream behavior when the supervised IBD data are resampled.

The difference between the two pre-trained approaches should be interpreted together with their input ranges and objectives. Barlow Twins uses the full Raman range and learns invariance to acquisition-like perturbations through contrastive redundancy reduction. SMAE uses the fingerprint region and learns to reconstruct masked spectral intervals, forcing the Transformer encoder to model local-to-global biochemical relationships. For this IBD dataset, the SMAE fingerprint representation appears especially effective, but the Barlow Twins results still demonstrate that contrastive full-range pre-training provides a substantially better starting point than training the same architecture from scratch. Overall, IBD remains a subtle and data-limited task, so the most reliable conclusion comes from combining hold-out diagnostics, confusion matrices, probability-based plots, and fixed-test cross-validation rather than relying on a single metric alone.

5.4 Pathogen Identification (BACTERIA-ID)

5.4.1 Task Definition and Evaluation Protocol

The BACTERIA-ID experiment evaluates pathogen identification as a 30-class Raman classification task. This setting is substantially different from IBD: instead of separating two clinically related plasma-spectrum classes, the model must distinguish bacterial strains whose spectra contain more explicit biochemical fingerprints. For this reason, BACTERIA-ID is useful for testing whether the self-supervised encoders can scale from binary diagnosis to fine-grained multiclass discrimination.

The notebooks used for this experiment directly implemented repeated fold-based evaluation. The available training set contained 3000 spectra, split into 2400 training spectra and 600 validation spectra inside each fold, while the predefined test set contained 3000 spectra and was kept fixed for final evaluation. Barlow Twins was evaluated on the full Raman range, whereas SMAE was evaluated on the 1000-point fingerprint representation between approximately 650 and 1750 cm^{-1}. The comparison therefore reflects both the different self-supervised objectives and the spectral view selected for each architecture.

5.4.2 BACTERIA-ID Fine-Tuning Setup

For Barlow Twins, the BACTERIA-ID notebook used the RamanFoundation/Barlow Twins full-range backbone trained on the 1000-point min–max normalized spectra. In each fold, the pre-trained backbone was loaded from the self-supervised checkpoint and placed inside a supervised Keras classifier with a final 30-class softmax layer. Unlike the IBD Barlow configuration, the BACTERIA-ID experiment did not freeze the backbone: the entire model was fine-tuned end-to-end so that the convolutional representation could adapt to the bacterial strain labels.

Regularization was applied directly to the supervised classification layer. The final dense layer used an L1–L2 penalty with coefficients 0.05 and 0.05, respectively, which discourages overly large or overly dense class-specific weights. Training used AdamW with learning rate 5×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, sparse categorical cross-entropy, batch size 256, and a maximum of 200 epochs. Early stopping monitored validation accuracy with patience 30 and restored the best weights. A ReduceLROnPlateau callback monitored validation loss, reduced the learning rate by a factor of 0.1 after 30 stagnant epochs, and used a minimum learning rate of 5×10^{-7}. The same fold procedure was repeated for the scratch control, replacing the pre-trained backbone with a randomly initialized twin network.

For SMAE, the BACTERIA-ID notebook followed the LP–FT strategy used for the Transformer encoder. Linear probing first trained only the supervised head with the encoder frozen; full fine-tuning then unfroze the model and optimized the encoder and classifier together. This difference is important when reading the training curves: the SMAE figure explicitly shows the transition from linear probing to full fine-tuning, whereas the Barlow Twins curve corresponds to direct end-to-end supervised fine-tuning of the full pre-trained backbone.

5.4.3 Fold-Based Results

Unlike the initial IBD hold-out diagnostic, the BACTERIA-ID analysis was organized directly around five independent folds. Each fold trained a model on a different train/validation partition and then evaluated it on the same predefined test set. This makes the reported mean and standard deviation a measure of how stable the downstream performance remains when the supervised training data are resampled.

For Barlow Twins, the pre-trained full-range encoder achieved an average test accuracy of 80.67% $\pm$ 0.61 across the five folds. The corresponding randomly initialized control reached 52.81% $\pm$ 24.78, showing both lower performance and much higher instability. The very low score of one scratch fold indicates that training the

full Barlow architecture from random initialization can be unreliable in this multi-class setting, whereas the pre-trained representation provides a consistently useful starting point.

For SMAE, the pre-trained fingerprint encoder achieved a mean test accuracy of 76.97% ± 0.97 and a Macro-F1 score of 76.16% ± 1.01. The from-scratch ViT control reached 75.00% ± 0.73 accuracy and 73.91% ± 1.00 Macro-F1. Thus, SMAE pre-training improves both reported metrics, although the gain is more moderate than for Barlow Twins because the scratch ViT already learns a strong fingerprint-region classifier on this dataset.

Table 5.6: BACTERIA-ID fold-wise test performance for pre-trained and scratch models. Barlow Twins reports accuracy from the full-range notebook; SMAE reports accuracy and Macro-F1 from the fingerprint-region notebook.

Model	Input Range	Initialization	Accuracy	Macro-F1
Barlow Twins	Full range	Scratch	52.81% ± 24.78	–
Barlow Twins	Full range	Pre-trained	80.67% ± 0.61	–
SMAE	Fingerprint	Scratch	75.00% ± 0.73	73.91% ± 1.00
SMAE	Fingerprint	Pre-trained	76.97% ± 0.97	76.16% ± 1.01

Table 5.7: BACTERIA-ID fold-wise test accuracy. The best Barlow Twins fold is Fold 1, while the best SMAE fold is also Fold 1.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Pre-trained Barlow Twins	81.83%	80.60%	80.50%	80.23%	80.17%
Scratch Barlow Twins	63.63%	3.33%	67.50%	63.83%	65.73%
Pre-trained SMAE	78.37%	76.60%	77.77%	75.67%	76.47%
Scratch SMAE	75.30%	75.37%	73.60%	75.03%	75.70%

5.4.4 Training Dynamics

The training curves in Figures 5.7 and 5.8 show the typical behavior observed in the best-performing runs. Barlow Twins reaches high training accuracy rapidly and then stabilizes with a validation accuracy around the 80% range. This is consistent with the fold results: the pre-trained convolutional encoder already provides a useful full-range representation, but the validation gap shows that the supervised classifier still has to manage overfitting in a 30-class setting.

SMAE shows a clearer two-stage transfer-learning profile. During linear probing, the frozen encoder improves gradually but remains limited. After full fine-tuning begins, validation accuracy increases sharply and the loss decreases substantially,

confirming that the fingerprint encoder benefits from being adapted to the bacterial labels rather than being used only as a fixed feature extractor.

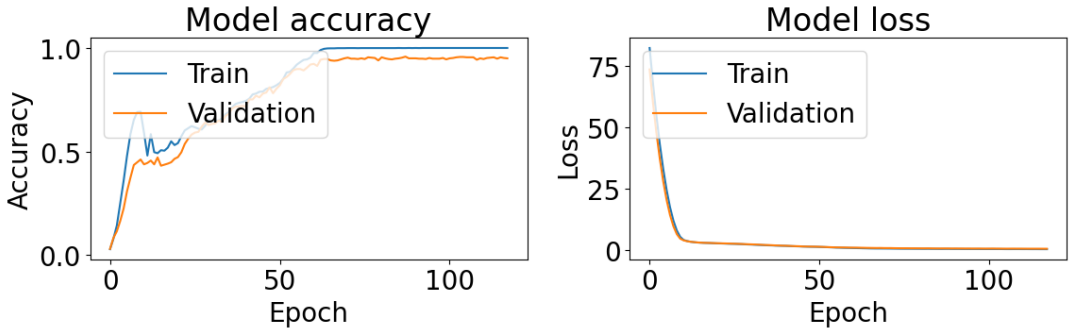


Figure 5.7: BACTERIA-ID training and validation curves for the best pre-trained Barlow Twins fold using the full Raman range.

5.4.5 Confusion Matrix and Best-Model Analysis

The confusion matrices provide the class-level interpretation of the best models. In both cases, the dominant diagonal structure indicates that most bacterial classes are correctly recovered on the predefined test set. This is important because aggregate accuracy alone could hide uneven behavior across the 30 strains; the matrices instead show that the models learn a broad multiclass separation rather than succeeding only on a small subset of easy classes.

The best Barlow Twins fold achieved 81.83% test accuracy, which is the highest single-fold result among the reported BACTERIA-ID experiments. This suggests that the full-range contrastive representation is particularly effective for bacterial identification, likely because useful discriminative information is distributed across the spectrum and not only inside the compact fingerprint interval. However, the scratch Barlow results also show that this architecture depends strongly on pre-training: without the self-supervised initialization, performance is unstable and can collapse on some folds.

The best SMAE fold reached 78.37% accuracy, with fold-level Macro-F1 of 77.60%. The confusion matrix remains strongly diagonal. Compared with Barlow Twins, SMAE is slightly lower in accuracy but provides a stable fingerprint-only solution and a consistent improvement over its scratch baseline.

5.4.6 Bacteria Result Interpretation

Overall, BACTERIA-ID confirms that self-supervised Raman pre-training is most beneficial when the downstream task contains structured spectral differences. Bar-

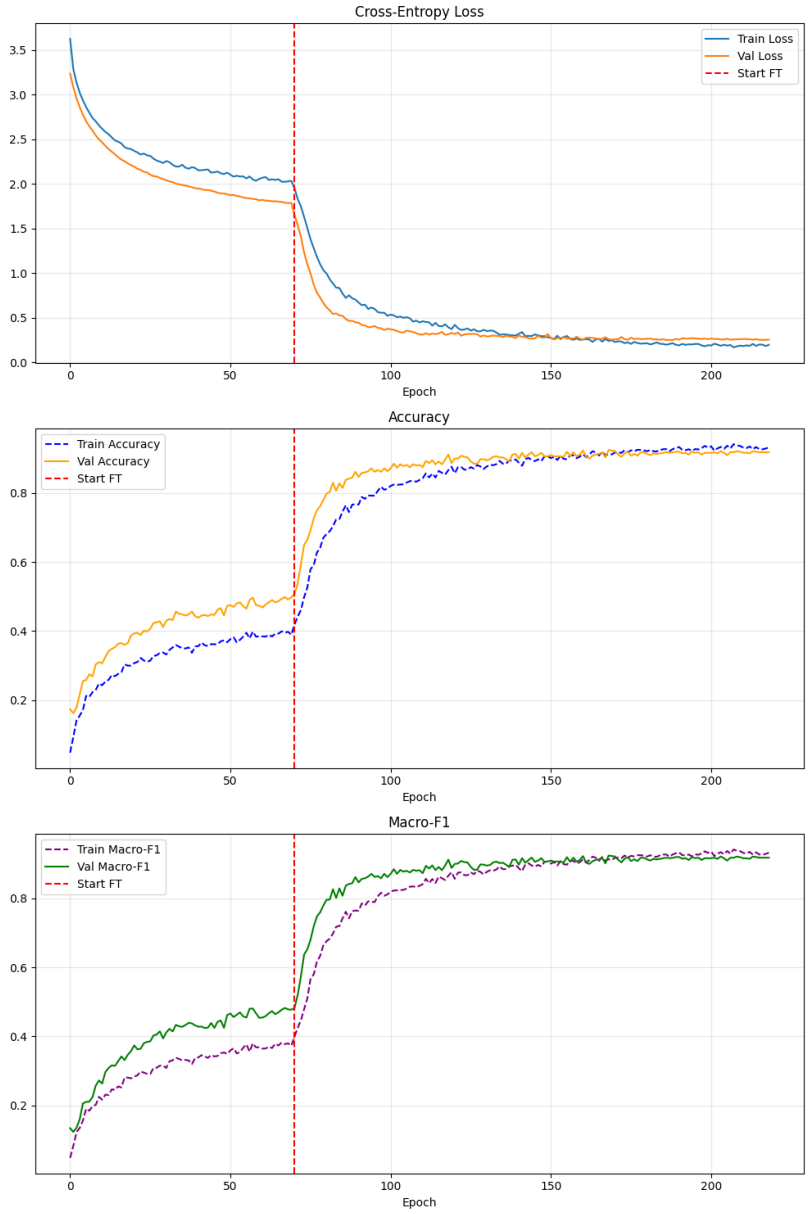


Figure 5.8: BACTERIA-ID training and validation curves for the pre-trained SMAE classifier using the fingerprint region. The vertical line marks the transition from linear probing to full fine-tuning.

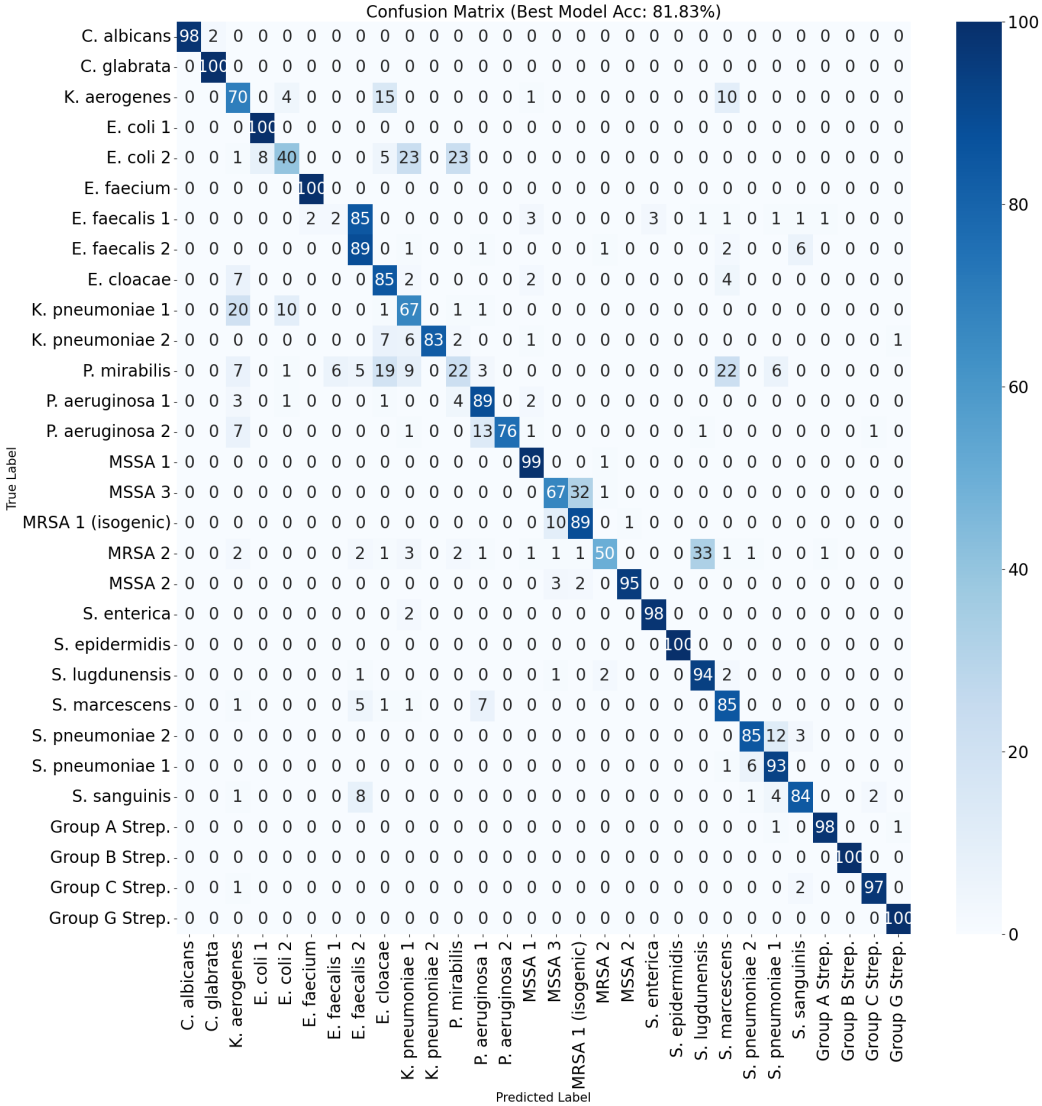


Figure 5.9: Confusion matrix for the best BACTERIA-ID Barlow Twins model. The best pre-trained fold reached 81.83% test accuracy using the full Raman range.

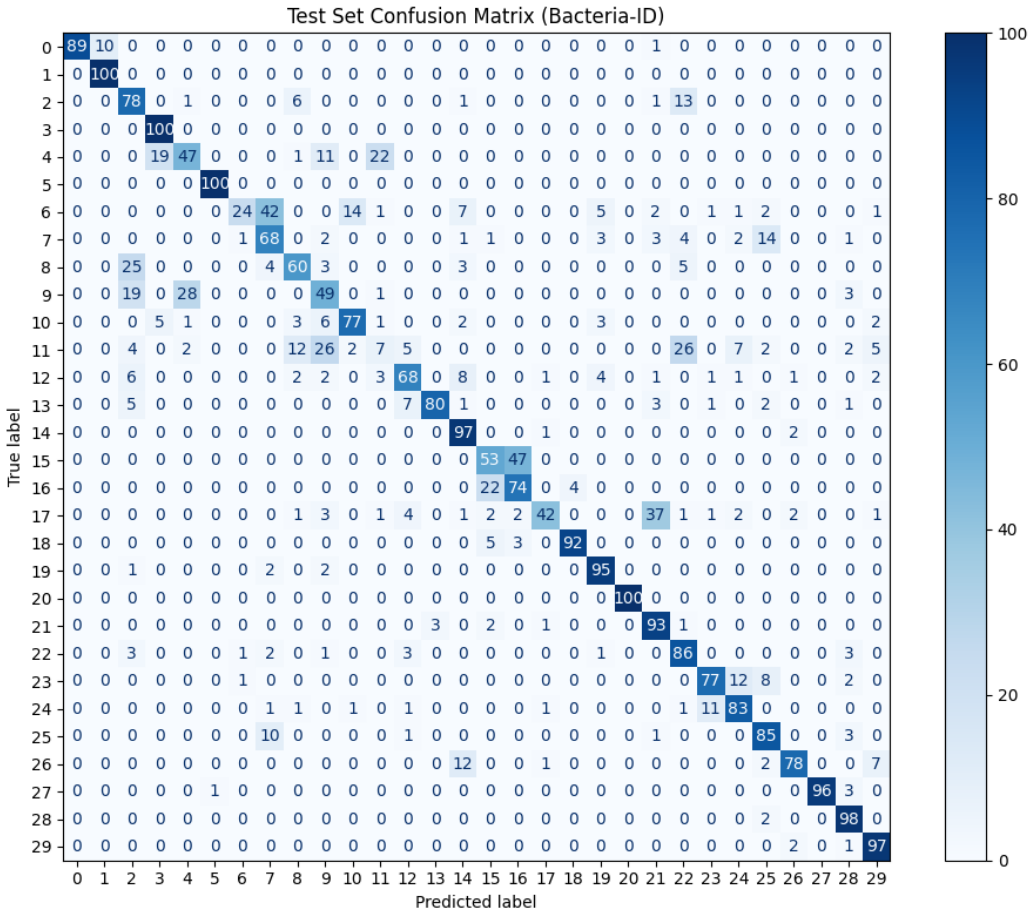


Figure 5.10: Confusion matrix for the BACTERIA-ID SMAE model using the fingerprint region. The best pre-trained fold reached 78.37% test accuracy and 77.60% Macro-F1.

low Twins gives the strongest absolute accuracy and the largest gain over scratch, showing that full-range contrastive pre-training can produce a highly transferable bacterial representation. SMAE is slightly behind in accuracy but remains very competitive, especially considering that it uses only the fingerprint region; its low fold-to-fold variability indicates stable behavior across classes.

The comparison also shows that the two architectures fail in different ways. Barlow Twins benefits greatly from the pre-trained backbone but its scratch control is fragile. SMAE is more robust from scratch, probably because the fingerprint representation and Transformer classifier are already well matched to the task, but masked-reconstruction pre-training still improves accuracy and Macro-F1. Therefore, the BACTERIA-ID results support the use of self-supervised initialization not only as a way to raise the best score, but also as a way to make fold-based evaluation more reliable.

5.5 Discussion

The downstream experiments demonstrate that the value of self-supervised Raman foundation models depends strongly on the biological task. In IBD classification, the models operate in a subtle binary regime: the spectral differences between healthy controls and inflammatory disease are weaker, the supervised dataset is smaller, and the main benefit of pre-training is to stabilize learning and reduce the tendency toward majority-class or split-dependent behavior. In this setting, the SMAE fingerprint representation gives the strongest IBD results, while Barlow Twins still shows that full-range contrastive pre-training is clearly preferable to training the same architecture from scratch.

In BACTERIA-ID, the role of representation learning is more directly visible because bacterial spectra contain stronger class-specific biochemical fingerprints. Both pre-trained models produce stable fold-based performance, but they do so in different ways. Barlow Twins achieves the best absolute accuracy, with 80.67% $\pm$ 0.61 across folds and a best fold of 81.83%, indicating that full-range contrastive representations are highly transferable for pathogen identification. SMAE remains slightly lower in accuracy but very competitive, reaching 76.97% $\pm$ 0.97 accuracy while using only the fingerprint region.

Taken together, the two downstream tasks suggest that there is no universally optimal Raman foundation model. SMAE is particularly effective when the fingerprint region contains compact diagnostic information and masked reconstruction can teach meaningful local-to-global spectral dependencies. Barlow Twins is especially

attractive when broader full-range spectral cues are useful and when contrastive invariance can produce a robust initialization for difficult multiclass discrimination. The best downstream strategy therefore depends on the clinical or microbiological target, the spectral range available, and the amount of labeled data used for fine-tuning.

Chapter 6

Model Interpretability and Explainable AI (XAI)

6.1 Purpose of the Interpretability Analysis

The interpretability analysis was performed on the fine-tuned Spectral Masked Autoencoder (SMAE) used for IBD classification. This choice was deliberate. Among the downstream IBD experiments, SMAE produced one of the strongest single-split results, reaching 94.29% test accuracy and 94.05% Macro-F1 on the hold-out test set. Because the objective of explainability is to understand what a reliable model has learned, the XAI analysis was applied to this high-performing SMAE configuration rather than to a weaker or unstable baseline.

The aim of this chapter is therefore not to introduce a new classifier, but to inspect the decision logic of the already fine-tuned SMAE model. In particular, the analysis asks whether the model bases its HC/IBD predictions on meaningful Raman regions or whether it may be relying on accidental artifacts. To answer this question, local single-spectrum explanations and class-level average explanations were computed with two complementary methods: Integrated Gradients (IG) and SHAP.

The model was trained on the 1000-point fingerprint representation between 650 and 1750 cm^{-1}. The interpretability results are consequently expressed directly in Raman shift units, making it possible to compare the most important features with known biochemical regions in plasma spectroscopy. The test set contained 35 spectra, with 14 healthy controls and 21 IBD samples.

Table 6.1: SMAE model used for the interpretability analysis. The XAI study was applied to the high-performing hold-out model because it provides a meaningful target for explaining the learned diagnostic strategy.

Model	**Input Range**	**Test Samples**	**Accuracy**	**Macro-F1**
Fine-tuned SMAE	Fingerprint, 650–1750 cm^{-1}	35	94.29%	94.05%

6.2 Explainability Methods

Two independent attribution methods were used. Integrated Gradients was used as the main gradient-based method, while SHAP was used as an independent additive explanation framework. Using both methods is important because agreement between them gives stronger evidence that the highlighted Raman bands are genuinely relevant for the classifier.

6.2.1 Integrated Gradients

Integrated Gradients attributes a prediction to the input features by accumulating gradients along a path from a baseline spectrum to the observed spectrum. In this experiment, IG was applied to the fine-tuned SMAE classifier and computed over the Raman fingerprint axis. For single spectra, the resulting attribution curve highlights local regions that drive the specific prediction. For class-level analysis, IG attributions were averaged separately over the HC and IBD test samples, producing one mean importance profile per class.

6.2.2 SHAP

SHAP estimates the contribution of each Raman feature to the model output using an additive attribution formulation. A background set from the training data was used to define the reference distribution, and SHAP values were then computed for the test spectra. As for IG, SHAP was inspected both locally, for individual spectra, and globally, by averaging feature importance across class-specific test samples. The SHAP beeswarm plot additionally preserves the sign of the attribution, showing whether high or low intensity at a Raman shift pushes the model output toward the IBD class.

6.3 Single-Spectrum Explanations

The first level of analysis considers individual spectra. These plots are useful because they show whether a single prediction is supported by localized Raman bands rather than by diffuse noise across the entire signal. To make the individual attribution points visible, Figures 6.1 and 6.2 are placed directly after this short introductory text and enlarged as much as possible. The first figure reports the Integrated Gradients heatmaps, while the second shows an importance-focused SHAP/attribution

view rather than the signed positive–negative SHAP map. This layout emphasizes where the model is looking most strongly in representative HC and IBD spectra.

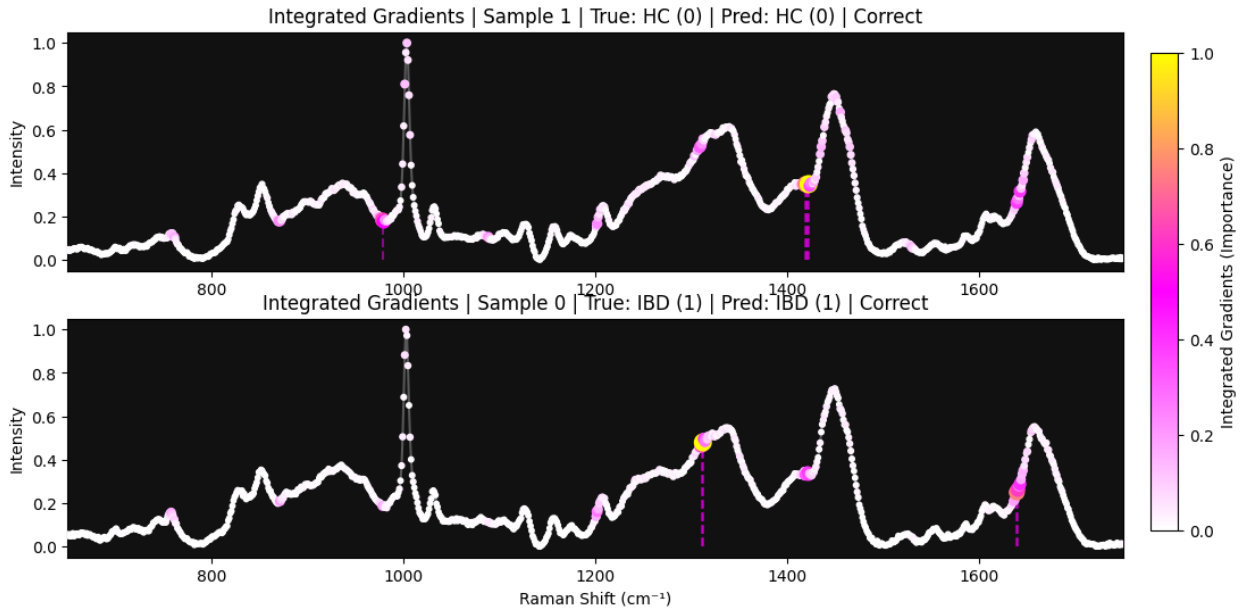


Figure 6.1: Single-spectrum Integrated Gradients explanations for representative HC and IBD test spectra.

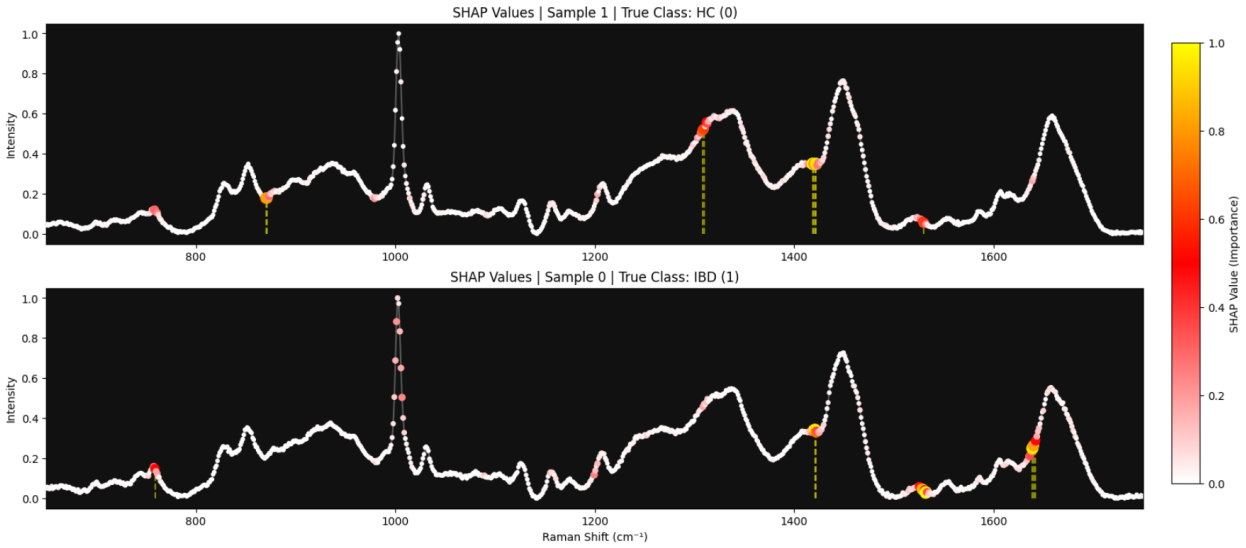


Figure 6.2: Single-spectrum SHAP attribution-importance view for representative HC and IBD spectra.

6.4 Class-Level Average Explanations

Single-spectrum explanations are useful for inspecting individual predictions, but they can be sensitive to the exact spectrum selected. For this reason, the attributions were also averaged separately across the HC and IBD test samples. These class-level plots provide a more stable view of the Raman regions repeatedly used by the model.

Figure 6.3 shows the mean IG profiles for the two classes. The HC average profile is dominated by strong regions around 1421.9, 978.1, 1638.8, 1311.8, and 1088.2 cm^{-1}. The IBD average profile emphasizes a partially overlapping but differently ranked set of bands, with the strongest peaks at 1311.8 and 1638.8 cm^{-1}, followed by 1421.9, 1201.7, and 871.3 cm^{-1}.

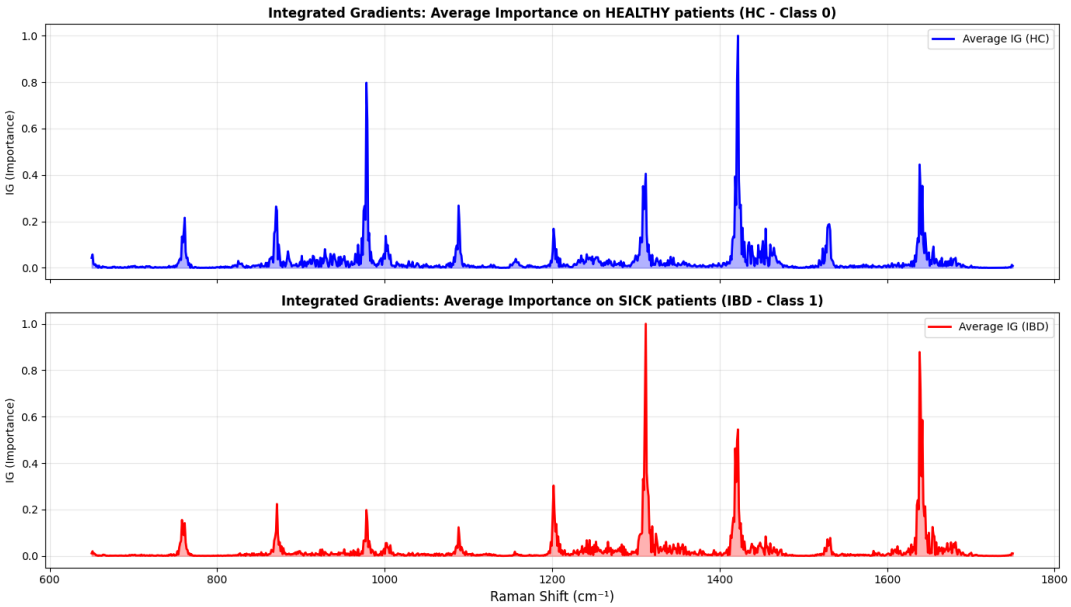


Figure 6.3: Average Integrated Gradients importance profiles for HC and IBD test spectra. The profiles summarize the spectral regions repeatedly used by the SMAE model for each class.

Figure 6.4 shows the analogous class-level SHAP importance profiles. The SHAP analysis again highlights a compact set of fingerprint-region bands, including 1421.9, 1638.8, 1529.8, 1201.7, and bands around 870–871 cm^{-1}. The partial overlap between IG and SHAP supports the reliability of the identified regions, while the differences in ranking are expected because the two methods measure attribution through different mathematical definitions.

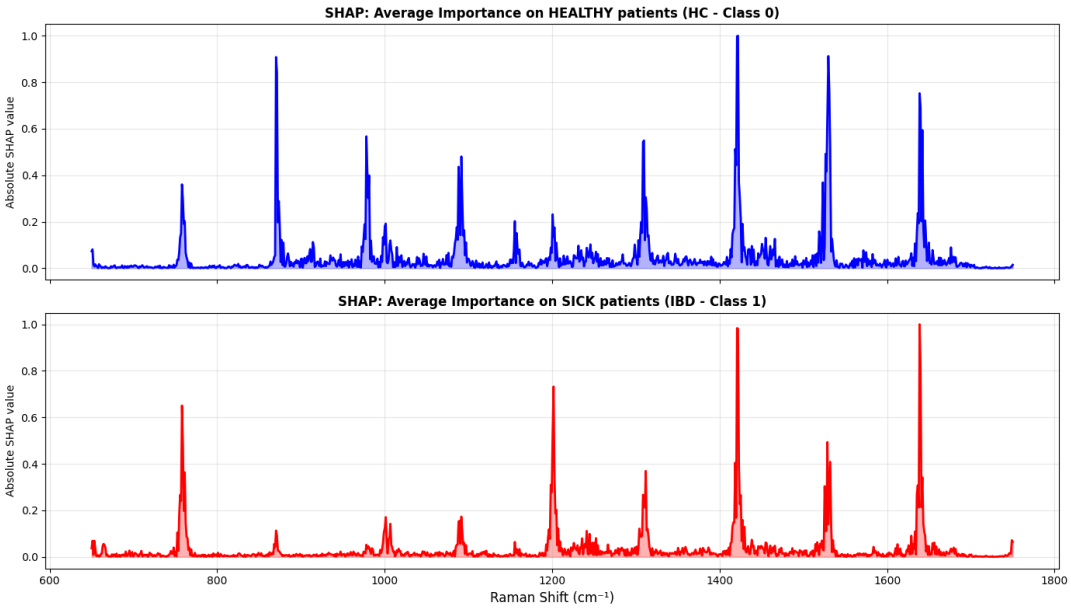


Figure 6.4: Average SHAP importance profiles for HC and IBD test spectra. The repeated peaks indicate Raman regions with stable global influence on the SMAE decision function.

6.5 Top Raman Peaks Identified by IG and SHAP

The top five peaks were extracted from the class-level IG and SHAP profiles using a local peak selection strategy. A minimum distance constraint was used to avoid selecting multiple adjacent points from the same spectral band. The resulting tables report normalized importance values, so the highest peak within each method and class is set to 1.000.

Table 6.2: Top five Raman shifts identified by Integrated Gradients for the HC and IBD classes. Importance values are normalized within each class-specific profile.

	HC – IG		**IBD – IG**	
Rank	**Shift (cm^{-1})**	**Importance**	**Shift (cm^{-1})**	**Importance**
1	1421.87	1.000	1311.76	1.000
2	978.13	0.797	1638.79	0.878
3	1638.79	0.445	1421.87	0.545
4	1311.76	0.405	1201.65	0.303
5	1088.24	0.268	871.32	0.224

The most important observation is that several bands recur across methods or classes. The region around 1421–1422 cm^{-1} appears in both IG and SHAP, and in both class-level analyses. The 1638.8 cm^{-1} band is also repeatedly selected and

Table 6.3: Top five Raman shifts identified by SHAP for the HC and IBD classes. Importance values are normalized within each class-specific profile.

	HC – SHAP		IBD – SHAP	
Rank	**Shift (cm^{-1})**	**Importance**	**Shift (cm^{-1})**	**Importance**
1	1421.87	1.000	1638.79	1.000
2	1529.78	0.912	1420.77	0.983
3	870.22	0.908	1201.65	0.732
4	1638.79	0.753	757.91	0.651
5	978.13	0.567	1528.68	0.494

is the highest-ranked SHAP feature for IBD. Additional recurrent regions include 1201.7 cm^{-1}, 1528–1530 cm^{-1}, and the 870–871 cm^{-1} region. This recurrence suggests that the SMAE model learned a compact biochemical fingerprint rather than a purely method-specific attribution pattern.

6.6 Literature-Based Biomarker Assignment

The top IG and SHAP peaks were taken from the notebook outputs reported in Tables 6.2 and 6.3. The literature was then used only to support the biochemical writing of nearby Raman regions, not to introduce additional peaks. Tables 6.4 and 6.5 therefore retain only assignments whose reference band is close to the model-selected wavenumber or belongs to a broad, explicitly reported Raman region. Peaks without a sufficiently close assignment are kept as model evidence but are not forced into a molecular interpretation.

Table 6.4: Healthy-control biomarker assignments for all unique top IG and SHAP peaks returned by the notebook. "Both" indicates that the feature was selected by both attribution families or appeared as a consensus band across the class-level analyses.

Model shift	**Method**	**Ref. band**	**Biochemical assignment**	**Source**
1421.87 cm^{-1}	Both	1420–1450 cm^{-1} region	CH_2/CH_3 deformation or carboxylate-related contribution	[27]
978.13 cm^{-1}	Both	–	Unassigned structural band retained from the notebook output	Model-derived
1638.79 cm^{-1}	Both	Amide I region	Protein C=O stretching / secondary structure	[27], [28]
1311.76 cm^{-1}	IG	1304 cm^{-1}	Lipids, CH twisting/bending	[29]
1088.24 cm^{-1}	IG	1080 cm^{-1}	Phosphate or glycogen-related vibration	[27]
1529.78 cm^{-1}	SHAP	1524 cm^{-1}	Carotenoids	[28]
870.22 cm^{-1}	SHAP	860–880 cm^{-1}	Proline / saccharides	[27]

Table 6.5: IBD biomarker assignments for all unique top IG and SHAP peaks returned by the notebook. The recurrent overlap between protein, carotenoid, lipid, and aromatic-amino-acid regions supports a multi-molecular inflammatory fingerprint.

Model shift	Method	Ref. band	Biochemical assignment	Source
1638.79 cm^{-1}	Both	Amide I region	Protein C=O stretching / secondary structure	[27], [28]
1528.68 cm^{-1}	SHAP	1524 cm^{-1}	Carotenoids	[28]
1421.87 cm^{-1}	IG	1420–1450 cm^{-1} region	CH_2/CH_3 deformation or carboxylate-related contribution	[27]
1420.77 cm^{-1}	SHAP	1420–1450 cm^{-1} region	CH_2/CH_3 deformation or carboxylate-related contribution	[27]
1311.76 cm^{-1}	IG	1304 cm^{-1}	Lipids, CH twisting/bending	[29]
1201.65 cm^{-1}	Both	1200–1210 cm^{-1}	Tyrosine / phenylalanine, Amide III region	[28]
871.32 cm^{-1}	IG	860–880 cm^{-1}	Proline / saccharides	[27]
757.91 cm^{-1}	SHAP	750–760 cm^{-1}	Tryptophan ring breathing	[28]

6.7 SHAP Directionality

The SHAP beeswarm plot in Figure 6.5 adds directional information to the global analysis. Each point corresponds to a test spectrum, and the horizontal axis shows whether a feature pushes the model output toward or away from the IBD class. This is useful because mean absolute importance identifies relevant notebook-derived bands, while the beeswarm also shows whether higher or lower intensity at those bands contributes to the pathological prediction.

6.8 Biochemical Interpretation

The biochemical interpretation is based on the attribution peaks produced by the notebook, not on peaks introduced from external tables. The literature is used only to describe nearby Raman regions when the match is plausible. For this reason, some model-selected bands are discussed as recurrent spectral evidence without assigning them to a specific molecule.

The most recurrent feature is the 1638.79 cm^{-1} region, selected by both IG and SHAP for both classes. This peak lies within the broad Amide I protein region, so it is interpreted as evidence that the classifier is sensitive to protein-related spectral structure in plasma. The assignment is kept broad because the model peak is not exactly identical to a single reference value, but it is close enough to the Amide I envelope to support a protein secondary-structure interpretation.

The 1528.68–1529.78 cm^{-1} region is selected by SHAP and is close to the reported carotenoid band around 1524 cm^{-1}. This is one of the clearest literature matches

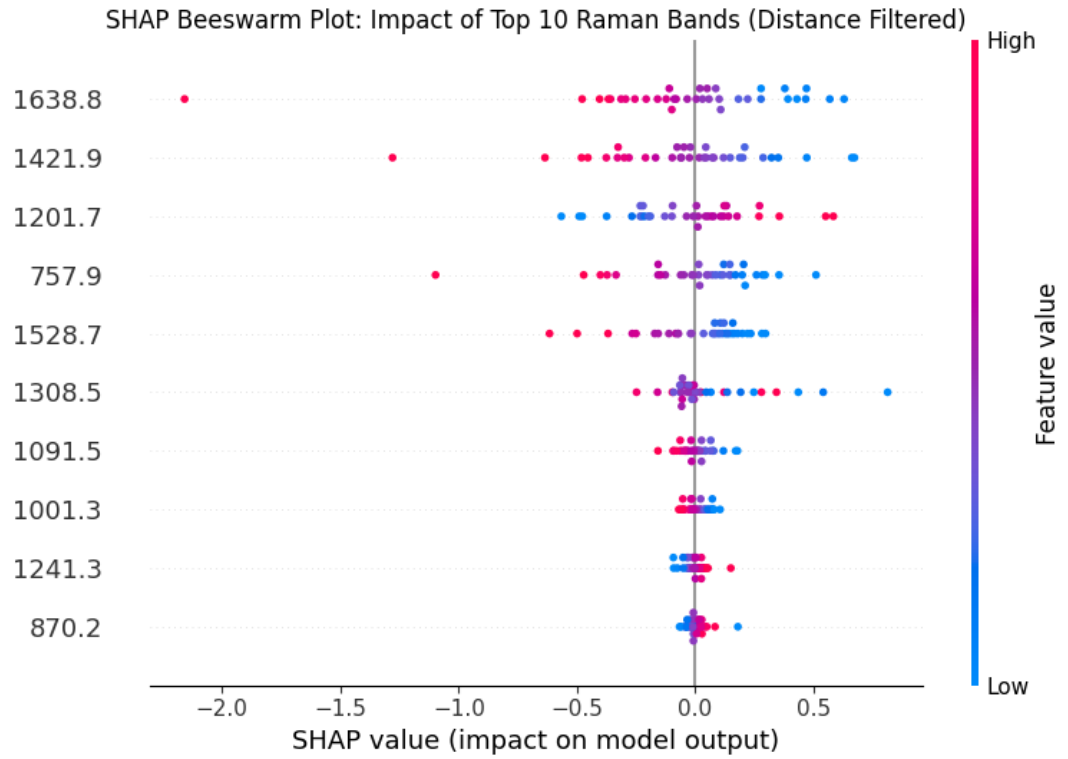


Figure 6.5: SHAP beeswarm plot for the top Raman bands selected from the notebook outputs. The plot shows both the magnitude and direction of each feature contribution across the test set.

among the notebook-derived peaks. It suggests that the model may use carotenoid- or antioxidant-related spectral information as part of the HC/IBD separation, while still avoiding a stronger claim than the spectral proximity supports.

The 1311.76 cm^{-1} IBD peak is close to the reported 1304 cm^{-1} lipid-related band and can therefore be interpreted as a lipid CH twisting or bending contribution. The adjacent 1420.77 and 1421.87 cm^{-1} outputs are kept as separate notebook peaks because they come from different attribution profiles, but they are interpreted as the same local Raman region. This region is compatible with CH_2/CH_3 deformation or carboxylate-related contributions in plasma spectra, so it is discussed as a protein/lipid-related structural band rather than as a completely unassigned feature.

Several amino-acid-related matches are also sufficiently close. The 1201.65 cm^{-1} band falls within the 1200–1210 cm^{-1} region associated with tyrosine, phenylalanine, and Amide III contributions. The 757.91 cm^{-1} SHAP peak is close to the 750–760 cm^{-1} tryptophan ring-breathing region. These bands suggest that aromatic amino-acid and protein-related contributions may participate in the inflammatory plasma fingerprint learned by SMAE.

The 870.22–871.32 cm^{-1} region is close to the 860–880 cm^{-1} interval associated with proline or saccharide contributions. The 1088.24 cm^{-1} HC peak is retained because it is close to the 1080 cm^{-1} phosphate/glycogen reference region. By contrast, 978.13 cm^{-1} remains unassigned because no sufficiently close literature band was retained; it is still reported because it comes directly from the notebook.

Overall, the interpretation supports the central XAI conclusion without overclaiming: the SMAE classifier appears to rely on a distributed Raman fingerprint made of notebook-derived bands, including protein/Amide I, carotenoid, lipid-related, aromatic-amino-acid, proline/saccharide, and phosphate/glycogen-related regions where the reference bands are close enough. Peaks without a close match are left as model-derived spectral evidence rather than forced biochemical biomarkers.

6.9 Discussion

The interpretability analysis provides evidence that the high-performing SMAE model does not rely on shortcut learning. Both IG and SHAP identify localized Raman bands in the fingerprint region, and several of these bands recur across HC and IBD class-level explanations. The strongest recurring regions include approximately 1421–1422, 1638.8, 1528–1530, 1201.7, 1088.2, and 870–871 cm^{-1}. These

regions form a multi-band spectral signature rather than a single-feature rule.

The single-spectrum plots show that individual predictions are supported by structured local evidence. The average IG and SHAP profiles then demonstrate that this evidence is not limited to one example, but persists across the test samples of each class. The SHAP beeswarm adds directionality, showing whether the top notebook-derived bands push the model output toward or away from the IBD class. Together, these local, global, and directional explanations strengthen the biological plausibility of the SMAE model while keeping the biomarker interpretation conservative.

Overall, Chapter 6 shows that the SMAE representation is not only accurate for IBD classification, but also interpretable in the original Raman domain. The model appears to learn a distributed biochemical fingerprint involving protein-related, lipid-related, and mixed molecular regions of the plasma spectrum. This supports the broader conclusion of the thesis: self-supervised Raman foundation models can provide useful diagnostic performance while still allowing their decisions to be inspected through physically meaningful spectral features.

Chapter 7

Conclusion and Future Perspectives

7.1 Summary of Contributions

This thesis investigated the use of self-supervised learning as a practical route toward Raman foundation models for biomedical spectroscopy. The central motivation was the mismatch between the large amount of Raman spectra that can be collected without labels and the much smaller number of spectra that are usually available with reliable clinical annotations. Instead of relying only on fully supervised training, this work explored whether unlabeled spectra could be used to pre-train general spectral encoders and then adapt them to downstream diagnostic tasks with limited labeled data.

The first contribution was the construction of a harmonized multi-domain Raman pre-training resource. Spectra from different biological settings were brought into a common processing pipeline, including interpolation, normalization, and range selection. This step was essential because Raman datasets often differ in acquisition range, resolution, baseline behavior, and noise level. By organizing the data into a unified representation, the thesis created the conditions for training models that learn reusable spectral structure rather than dataset-specific artifacts.

The second contribution was the implementation and evaluation of two complementary self-supervised strategies. Barlow Twins was used as a contrastive redundancy-reduction framework, encouraging the model to produce stable representations under realistic spectral augmentations such as noise, scaling, and baseline perturbations. In parallel, the Spectral Masked Autoencoder (SMAE) was used to learn Raman structure through masked reconstruction, forcing the encoder to infer missing spectral regions from their biochemical context. These two approaches captured different aspects of the Raman signal: Barlow Twins emphasized invariance, while SMAE emphasized local-to-global spectral reconstruction.

The third contribution was the downstream evaluation of the pre-trained representations across multiple biomedical tasks. The models were transferred to classification settings involving inflammatory bowel disease plasma spectra, and bacterial pathogen identification. Across these experiments, pre-training generally provided a stronger and more stable starting point than training the same architectures from scratch. The IBD experiments were particularly informative because the labeled

dataset was small and clinically subtle: the pre-trained SMAE achieved the strongest performance, while Barlow Twins still showed a clear improvement over its scratch baseline. These results support the main hypothesis of the thesis, namely that self-supervised Raman encoders can improve data efficiency when annotated biomedical spectra are limited.

The fourth contribution was the interpretability analysis of the best-performing IBD model. Integrated Gradients and SHAP were used to inspect both single-spectrum predictions and class-level attribution profiles. The explanations showed that the SMAE classifier focused on localized Raman bands rather than diffuse noise. The most recurrent regions included bands around 1421–1422, 1638.8, 1528–1530, 1201.7, 1088.2, and 870–871 cm^{-1}. These bands were then interpreted conservatively against nearby Raman assignments, supporting a distributed fingerprint involving protein structure, carotenoid-related signals, lipid or carboxylate-related deformation modes, aromatic amino acids, and phosphate or saccharide-related contributions. This step was important because it connected model performance with physically meaningful spectral evidence, reducing the risk that the classifier was relying on an uninterpretable shortcut.

Overall, the thesis demonstrates that Raman self-supervised learning is not only a technical pre-training exercise, but a realistic framework for building transferable and interpretable diagnostic models. The results suggest that spectral foundation models can help bridge the gap between data-rich, label-poor spectroscopy and the label-efficient models required for clinical translation.

7.2 Limitations

Despite these encouraging results, several limitations remain. First, labeled tissue data for the breast cancer use case were not yet available in this thesis. For this reason, the downstream validation could only be performed on the currently accessible labeled datasets, mainly the IBD plasma classification task and the BACTERIA-ID pathogen identification task. The results should therefore be interpreted as evidence that the pre-trained Raman encoders transfer well to available biomedical classification problems, rather than as a final validation for breast cancer tissue diagnosis. Once labeled breast tissue spectra become available, larger and more clinically targeted experiments will be needed to test generalization across instruments, acquisition protocols, hospitals, and patient populations.

Second, the pre-training corpus was harmonized into common spectral representations, but Raman spectroscopy is highly sensitive to preprocessing choices.

Baseline correction, normalization, interpolation, and spectral-range selection can all affect the information available to the model. Although the thesis adopted consistent preprocessing strategies, future work should evaluate how robust the learned representations are to alternative preprocessing pipelines and to rawer forms of spectral input.

Third, the explored model families represent only part of the design space. Barlow Twins and SMAE were selected because they are well suited to contrastive invariance learning and masked reconstruction, respectively. However, other architectures and objectives may capture complementary information. The results therefore provide a strong starting point, but not an exhaustive comparison of all possible Raman foundation-model designs.

Finally, the interpretability analysis remains a model-based explanation rather than a definitive biochemical validation. IG and SHAP can highlight influential spectral regions, and the selected bands can be compared with known Raman assignments, but plasma spectra are highly overlapping and many bands are not molecule-specific. The biochemical interpretation should therefore be viewed as a plausible and conservative reading of the model evidence, not as a complete molecular identification.

7.3 Future Perspectives

A first direction for future work is the adaptation of these pre-trained encoders to larger labeled breast-cancer datasets as they become available. This is particularly relevant for the SPECTRA-BREAST project, where Raman spectroscopy is being developed for intraoperative margin assessment during breast-conserving surgery [6]. The pre-trained models developed in this thesis could serve as initialization backbones for future supervised fine-tuning once larger, well-curated breast cancer annotations are collected. In this scenario, the self-supervised encoders would provide a chemically informed starting representation, while the new labels would specialize the model toward clinically actionable breast tissue discrimination.

A second direction is to improve the Barlow Twins architecture. In this thesis, Barlow Twins relied on a convolutional backbone and a projection head to learn invariant spectral embeddings. Future work could explore replacing or augmenting the projector with Transformer-based modules, allowing the model to capture longer-range dependencies between distant Raman bands before the redundancy-reduction objective is applied. A Transformer projector or hybrid CNN–Transformer contrastive architecture may be especially useful for spectra where diagnostic infor-

mation is distributed across multiple separated biochemical regions.

A third direction is systematic hyperparameter optimization for each dataset and task. The optimal masking ratio, patch size, encoder depth, augmentation strength, learning rate, and spectral range may differ substantially between plasma, tissue, and bacterial spectra. Rather than using a single configuration across all settings, future studies could perform dataset-specific searches to tune the pre-training and fine-tuning pipelines. This may improve performance further, especially on subtle tasks where small spectral differences are clinically important.

A fourth direction is to extend the interpretability framework. Integrated Gradients and SHAP provided useful first-order explanations, but future work could compare them with additional methods such as occlusion sensitivity, counterfactual spectral perturbations, concept-based explanations, attention rollout for Transformer models, or prototype-based interpretation. Combining several explanation families would make it easier to distinguish robust biochemical evidence from method-specific attribution artifacts. In addition, interpretability should eventually be connected to experimental validation, for example by checking whether highlighted bands remain stable across cohorts, instruments, and preprocessing variants.

A fifth direction is external and prospective validation. Before these models can be considered for clinical deployment, they should be tested on data acquired from independent centers and under realistic acquisition conditions. This includes evaluating robustness to instrument drift, sample preparation differences, fluorescence background, and patient heterogeneity. Domain adaptation and calibration methods may also be required so that a model trained on one acquisition setting remains reliable when transferred to another.

Finally, future work should consider deployment constraints. For intraoperative use, models must be lightweight, fast, and reliable enough to provide predictions in real time. The encoders studied in this thesis are promising because they can be pre-trained offline and then fine-tuned into compact downstream classifiers. With additional labeled breast cancer data, optimized architectures, stronger validation, and clinically grounded interpretability, Raman foundation models could become a useful component of decision-support systems for rapid, label-free biomedical diagnostics.

Bibliography

[1] R. G. Pleijhuis, M. Graafland, J. de Vries, J. Bart, J. S. de Jong, and G. M. van Dam, "Obtaining adequate surgical margins in breast-conserving therapy for patients with early-stage breast cancer: Current modalities and future directions," *Annals of Surgical Oncology*, vol. 16, no. 10, pp. 2717–2730, 2009.

[2] L. Omar et al., "Positive margin rates after breast-conserving surgery by histologic subtype: A systematic review and meta-analysis evaluating the impact of oncoplastic surgery," *Annals of Surgical Oncology*, vol. 32, no. 7, pp. 4899–4909, 2025.

[3] European Commission, *Spectra-breast: Spectroscopic intraoperative margin assessment in breast-conserving surgery*, https://cordis.europa.eu/project/id/101187508, 2024.

[4] K. Hanna et al., "Raman spectroscopy of breast tissues," *Analyst*, vol. 140, pp. 55–62, 2015.

[5] C.-S. Ho et al., "Deep learning for raman spectroscopy: A review," *Analytica Chimica Acta*, vol. 1143, pp. 1–18, 2021.

[6] SPECTRA-BREAST Consortium, "Spectra-breast proposal: Spectroscopic intraoperative margin assessment in breast-conserving surgery," *European Commission Horizon Europe Grant Agreement 101187508*, 2024.

[7] J. Langer et al., "Deep learning for raman spectroscopy: A review," *Leibniz IPHT Publications*, 2021.

[8] Y. Lin et al., "Enhancing explainability in raman spectroscopy classification with shap and spectral segmentation," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 344, p. 126 394, 2025.

[9] J. Liu et al., "Deep learning for raman spectrum recognition," *Optics Express*, vol. 25, no. 23, pp. 28 059–28 070, 2017.

[10] Y. Li et al., "Classifying breast cancer tissue by raman spectroscopy with one-dimensional convolutional neural network," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2021.

[11] W. Li et al., "Raman spectroscopy combined with convolutional neural network for rapid identification of breast cancer molecular subtypes," *Biomedical Signal Processing and Control*, 2024.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] Y. Yao et al., "Deep learning-based raman spectroscopy qualitative analysis algorithm: A convolutional neural network and transformer approach," *Talanta*, vol. 275, pp. 126–135, 2025.

[14] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[15] C. Min et al., "Rat: Raman transformer for highly accurate melanoma detection with critical features visualization," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 2024.

[16] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PMLR, 2020, pp. 1597–1607.

[17] J. Zbontar et al., "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*, PMLR, 2021, pp. 12 310–12 320.

[18] A. Eranti, Y. Tewari, R. Palacios, and A. Gupta, "Raman spectroscopy pretrained encoder: A self-supervised learning approach for data-efficient domain-independent spectroscopy analysis," *IEEE Access*, vol. 14, pp. 40 311–40 327, 2026.

[19] P. Ren, R.-g. Zhou, and Y. Li, "A self-supervised learning method for raman spectroscopy based on masked autoencoders," *Expert Systems with Applications*, 2025.

[20] K. He et al., "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[21] A. Ahmed et al., "Ramanmae: Masked autoencoders enable efficient molecular imaging by learning biologically meaningful spectral representations," *Analytical Chemistry*, 2025.

[22] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*, PMLR, 2017, pp. 3319–3328.

[23] Z. e. o. Yan, "Deep learning-enabled raman spectroscopic identification of pathogen-derived extracellular vesicles and the biogenesis process," *Analytical Chemistry*, vol. 94, no. 36, pp. 12 416–12 426, 2022.

[24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, vol. 30, 2017.

[25] C. C. Horgan et al., "Deepr: A deep learning framework for raman spectroscopy," *Analytical Chemistry*, vol. 93, no. 48, pp. 15 850–15 860, 2021.

[26] C.-S. Ho et al., "Rapid identification of pathogenic bacteria using raman spectroscopy and deep learning," *Nature Communications*, vol. 10, no. 1, p. 4927, 2019. DOI: 10.1038/s41467-019-12898-9

[27] E. Buchan et al., "Raman spectroscopic molecular fingerprinting of biomarkers for inflammatory bowel disease," *Clinical and Translational Medicine*, vol. 13, no. 11, e1345, 2023.

[28] C. Morasso et al., "Raman analysis reveals biochemical differences in plasma of crohn's disease patients," *Journal of Crohn's and Colitis*, vol. 14, no. 11, pp. 1572–1580, 2020.

[29] M. D. Ghauri et al., "Biomolecular fingerprint of crohn's disease: A comparative raman spectroscopic study of blood and tissue samples," *Journal of Biophotonics*, vol. 19, e70219, 2026.