WITS
UNIVERSITY

RESEARCH PROPOSAL

Representation Learning for Mass Spectrometry Imaging

Tapiwa Mazarura

2581366

School of Computer Science and Applied Mathematics

Faculty of Science

University of the Witwatersrand

Supervisors: Prof H Bau, Prof R Klein, Mr D Jarvis

18 May 2025

Declaration

I, Tapiwa Mazarura, declare that this report is my own, unaided work. It is being submitted for the degree of Honours in Computer Science at the University of the Witwatersrand, Johannesburg. It has not been submitted for any degree or examination at any other university.

Tapiwa Mazarura

18 May 2025

Abstract

Mass spectrometry imaging (MSI) generates high-dimensional molecular profiles that are invaluable for biomarker discovery in cancer diagnostics. However, the substantial dimensionality and inherent noise of MSI data pose significant challenges for traditional analysis pipelines. This research proposes a comprehensive computational framework that integrates classical dimensionality-reduction techniques (PCA, t-SNE, UMAP, non-negative matrix factorization) with a purpose-built 2D convolutional autoencoder (CNN-AE) enhanced by self-supervised contrastive pretraining. The objective is to extract compact, interpretable latent representations that retain discriminative cancer signatures while enabling robust downstream classification of malignant versus non-malignant tissue regions.

Following thorough data preprocessing—comprising mass alignment, baseline correction, total-ion current normalization, and selection of the top 200 m/z channels—baseline methods will establish performance benchmarks in terms of variance capture, clustering quality, and runtime. The CNN-AE framework then employs spatial patch extraction (64×64 pixels) and targeted augmentations (Poisson noise, pixel dropout, affine transforms) to train an encoder–decoder network with a 20-dimensional bottleneck. A subsequent self-supervised contrastive learning phase initializes the encoder to enhance feature invariance. Finally, a small multilayer perceptron classifier, either trained separately on fixed embeddings or jointly via a composite reconstruction-classification loss, evaluates the diagnostic utility of the learned representations.

Evaluation metrics include reconstruction mean squared error, clustering metrics (Adjusted Rand Index, purity, silhouette score), and classification performance (accuracy, F1 score, ROC-AUC), supplemented by visualizations such as confusion matrices, ROC curves, and UMAP scatter plots. We hypothesize that the CNN-AE—particularly when pretrained contrastively—will outperform classical methods by yielding lower reconstruction error, tighter class clusters, and superior classification metrics, while also revealing latent features that correspond to known and novel cancer biomarkers. This study will offer methodological guidelines for MSI data analysis, advancing both computational techniques and biomedical insights in cancer research

Contents

Chapter 1

Introduction

Globally, cancer persists as a major contributor to mortality, underscoring the necessity of timely and precise detection to enable effective therapies and enhance survival rates [19]. As computational techniques evolve, they present new opportunities for addressing some of the most pressing challenges in cancer diagnostics. One such challenge is the identification of biomarkers—molecular indicators of disease presence or progression—from high-dimensional and noisy datasets, such as those generated by mass spectrometry (MS) [2] .

MS data provide detailed molecular profiles by measuring thousands of m/z (mass-to-charge ratio) channels, each representing a potential molecular feature. However, this high dimensionality creates significant computational and analytical bottlenecks, including noise, redundancy, and difficulty in identifying the most informative features [1]. From a computer science perspective, this challenge can be reframed as a problem of efficient data reduction and feature selection. Can a computational pipeline identify meaningful features of mass spectrometry data that retain the discriminative power needed to classify cancer phenotypes accurately?

This research proposes a computational framework leveraging multiple dimensionality reduction and feature extraction techniques and classifiers to tackle this problem. Among the feature extraction and dimensionality reduction methods is the use of Autoencoders to effectively compress the high-dimensional MS data into a compact and informative latent space. Subsequently, machine learning classifiers, designed to analyze spatial and relational properties of data, are applied to the reduced feature set to achieve robust classification of cancerous versus non-cancerous phenotypes.

By adopting this hybrid approach, the project aims to enhance both the efficiency and interpretability of biomarker discovery. For computer scientists, this work demonstrates the application of advanced machine learning methodologies to a critical real-world problem, bridging the gap between theoretical model development and impactful biomedical applications. Moreover, the insights gained from the selected features could reveal biologically significant patterns, contributing to the broader understanding of cancer mechanisms.

The following sections provide a detailed roadmap for this research. The next section, **Background**, reviews related work in mass spectrometry data analysis and machine learning methodologies, focusing on dimensionality reduction, feature selection, and classification techniques. It establishes the context for the proposed computational framework and highlights existing gaps that this research aims to address.

The **Methods** section outlines the technical implementation of the proposed framework, including the design and training of the autoencoder, the application of machine learning classifiers, and the evaluation metrics for assessing performance. This section will detail the dataset preprocessing, the pipeline's workflow, and the experimental setup used to validate the approach.

The **Research Plan** provides a timeline and schedule of deliverables, detailing the milestones for each phase of the project. It also discusses potential limitations, such as computational resource constraints or challenges in generalizing results, and proposes strategies to mitigate these risks.

Finally, the **Conclusion** summarizes the contributions of this research, emphasizing its potential impact on the fields of machine learning and biomedical data analysis. It reiterates the significance of developing computational tools for high-dimensional data and their role in advancing cancer diagnostics.

Chapter 2

Literature Review

This chapter surveys computational strategies for feature extraction and dimensionality reduction in mass spectrometry (MS) data analysis, with an emphasis on autoencoder-based approaches, latent space interpretability, and downstream classification using learned representations. We focus on methods that balance information preservation with interpretability and evaluate their performance in constructing robust classifiers.

2.1 Dimensionality Reduction in MS Data

Mass spectrometry datasets are inherently high-dimensional, often comprising thousands of m/z features per sample, which complicates pattern recognition and model training [2]. Dimensionality reduction techniques aim to embed these features into lower-dimensional spaces while retaining critical biochemical information. Early work applied linear methods such as PCA (Principal Component Analysis) and NMF (Non-negative Matrix factorization) [2, 5, 21]. PCA maximizes variance capture but can yield components that mix unrelated signals, hindering interpretability. NMF enforces non-negativity and part-based representations, producing more localized features but often at the cost of higher reconstruction error in complex spectra [21].

Nonlinear methods, including t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP), were introduced to preserve local similarity relationships [15]. While effective for visualization and clustering, their non-parametric nature limits direct inversion for data reconstruction and complicates integration into end-to-end pipelines. These

limitations underscore the need for methods that jointly optimize embedding quality, reconstruction fidelity, and interpretability [**HuContrastive** , 15, 21].

2.2 Spatial–Spectral Embeddings for Classification

2.2.1 Supervised Spatial Colocalization (ColocML)

ColocML, a semi-supervised CNN, learns to rank ion-image pairs by spatial similarity against expert-curated gold standards. By encoding spatial tf-idf vectors and achieving a Spearman correlation of 0.797 with human scores, ColocML provides quantifiable spatial features that can serve as input to classifiers to discriminate tissue types or pathological states [13]. Its framework informs strategies for integrating spatial pattern features in end-to-end classification pipelines.

2.2.2 Spectral Embeddings (Spec2Vec and MS2DeepScore)

Inspired by Word2Vec, Spec2Vec embeds co-occurring MS/MS peaks into vectors that capture structural relationships, yielding 88% library-matching accuracy on GNPS data [7]. Complementarily, MS2DeepScore (using Siamese networks) learns low-dimensional spectral embeddings with 0.15 Root Mean Squared Error (RMSE) for predicting structural similarity [6]. These unsupervised spectral encodings can be concatenated with spatial embeddings or fed directly into downstream classifiers, improving molecular resolution for biomarker discrimination.

2.2.3 Hybrid Models

Building on separate spatial and spectral embeddings, a hybrid architecture, MS2Query, combines the Spec2Vec and MS2DeepScore vectors within a random forest to predict lipid classes, achieving 35% recall of high similarity analogs (Tanimoto ≤ 0.63) [9]. This ensemble approach demonstrates that integrating complementary embedding modalities can mitigate annotation gaps and enhance the discriminative power of classifiers trained on metabolomics data.

2.2.4 Self-Supervised Spatial Representation Learning (DeepION)

DeepION is a contrastive learning framework tailored for MSI that applies MSI-specific augmentations (e.g., Poisson noise, missing value masking) to train a ResNet18 encoder that produces robust 20-dimensional embeddings. DeepION outperforms traditional similarity metrics and SimSiam baselines by 25–60% in colocalization tasks, and its noise-invariant features are well suited for separating healthy vs. diseased samples in a classifier [3].

2.2.5 Contrastive and Metric-Learning Objectives

Both Siamese-network and SimCLR-inspired approaches have been shown to shape latent spaces for maximal class separability. By incorporating triplet or contrastive losses during embedding learning, these methods yield representations that align closely with classification boundaries, leading to classifiers robust to inter-sample heterogeneity and technical noise [4, 6].

2.3 Autoencoder Architectures for Latent Representation

Autoencoders (AEs) constitute a family of neural network models that learn compressed latent representations via encoder–decoder architectures. AEs impose a bottleneck to force information compression, optimizing reconstruction loss to approximate original spectra. Variational autoencoders (VAEs) introduce probabilistic latent variables, encouraging smooth manifold structures at the expense of increased complexity [8, 18]. Denoising autoencoders (DAEs) further enforce robustness by reconstructing clean inputs from noisy or partially masked spectra [4].

Recent adaptations tailor autoencoders to MS data characteristics. For example, convolutional AEs leverage local mass-to-charge patterns as analogous to spatial features in images, and graph-based classifiers model known biochemical relationships between metabolites [11]. Comparative studies demonstrate that autoencoders outperform PCA and NMF in minimizing reconstruction error [4]. However, challenges remain in interpreting latent features and ensuring they correspond to meaningful biochemical axes.

2.4 Interpretability of Latent Spaces

Interpretable latent representations facilitate hypothesis generation and biological insight. Approaches to increasing interpretability include sparsity constraints, disentangled representations, and post-hoc attribution methods. Sparse autoencoders impose L1 regularization on latent activations, promoting feature selection and facilitating mapping of latent nodes to specific mass spectral peaks [10]. Disentangled VAEs add regularization terms (e.g. β-VAE) to encourage independent latent factors, which can correspond to distinct biochemical processes [8].

2.5 Classifier Construction from Latent Embeddings

The second goal of this work is to build classifiers on latent embeddings that reliably distinguish sample classes (e.g., healthy vs. Pathological). Traditional pipelines extract latent vectors via unsupervised training, then fit separate classifiers such as random forests or support vector machines [4]. End-to-end frameworks jointly train encoder and classifier layers, optimizing a composite loss combining reconstruction and classification objectives. These hybrid models often outperform decoupled approaches by aligning latent spaces with discriminative boundaries [6].

Notably, deep metric learning approaches (e.g., Siamese networks, triplet losses) have been employed to shape latent spaces directly for classification tasks, enhancing inter-class separability and intra-class compactness [6]. Contrastive learning frameworks further leverage unlabeled data to refine embedding quality before supervised fine-tuning [4]. Comparative analyses indicate that classifiers built on autoencoder-derived embeddings achieve superior accuracy and robustness against noise compared to classifiers trained on raw spectral features or PCA projections.

2.6 Opportunities and Integration

While autoencoders excel in information retention and allow flexible architecture design, achieving interpretability and classification readiness simultaneously remains challenging. This project investigates:

- Comparative evaluation of linear (PCA, NMF) and nonlinear (UMAP, t-SNE) baselines against CNN AEs in terms of reconstruction error and latent interpretability.

- Integration of sparsity and disentanglement constraints into AE training to produce latent features with clear spectral associations.
- Development of an end-to-end classification framework coupling AE encoders with downstream classifiers under multi-objective loss, and benchmarking against decoupled pipelines.

By synthesizing best practices from existing literature, this study aims to establish methodological guidelines for constructing informative and interpretable latent spaces in MS data and deploying them effectively in classification tasks.

Chapter 3

Methods

This chapter details the methods used in the research.

3.1 Research Questions

1. How do two-dimensional and three-dimensional CNN-based autoencoder architectures compare with conventional dimensionality reduction methods (such as PCA and t-SNE) in their ability to preserve and highlight cancer-relevant molecular signatures within mass spectrometry imaging datasets?

2. What effects do variations in latent-space dimensionality and autoencoder design parameters (including network depth and convolutional kernel size) exert on both binary classification performance (e.g., accuracy, sensitivity, specificity) and the computational demands (training time, memory usage, inference speed) of the framework?

3. To what extent do the latent features extracted by the proposed CNN autoencoder framework correspond to established cancer biomarkers, and can they uncover previously unrecognized molecular spatial patterns of biological significance?

3.2 Hypothesis

We hypothesize that a purpose-built CNN autoencoder—configured with optimized latent-space dimensionality and architecture (depth and kernel size) and trained with combined denoising and

contrastive objectives—will outperform traditional dimensionality reduction techniques (PCA, t-SNE) by more faithfully preserving discriminative cancer signatures in MSI data, enabling a downstream binary classifier to achieve superior diagnostic performance while maintaining computational efficiency [4, 6, 15]. Furthermore, we anticipate that the resulting latent representations will not only recapitulate known cancer biomarkers but also reveal novel spatial–molecular patterns with potential biological relevance.

3.3 Dataset

Our dataset consists of high-resolution mass spectrometry imaging (MSI) of mouse tissue sections (brain and select organs), annotated at the pixel and regional levels [12, 14]. Each MSI frame comprises a three-dimensional data cube of size $H \times W \times M$, where:

- H, W (pixels): spatial dimensions (typically 100 μm per pixel).
- M: number of m/z channels (after channel selection, $M = 200$).

Annotations include:

- **Tumor Label** ($y_{\text{tumor}} \in \{0, 1\}$): binary mask marking malignant versus non-malignant regions, provided by expert pathologists.
- **Chemical Signatures** (y_{chem}): categorical labels for known compounds (e.g., blood–brain barrier penetrants) detected via targeted m/z windows.

Acquisition and File Formats

- **Imaging Modality**: MALDI-TOF MSI acquired on Bruker Autoflex III (spatial resolution 100 m, mass range m/z 100–700).
- **Data Files**: raw spectra stored in .RAW format; preprocessed spectra and annotations exported as HDF5 and PNG masks.
- **Spatial Registration**: tissue sections were aligned to histology slides via rigid registration, ensuring pixel-accurate tumor masks.

3.4 Data Preprocessing

To concentrate on tumor-associated patterns, we filter out all regions labelled as healthy, retaining only pixels or super-pixel clusters marked malignant. This strategy aligns with the focus on cancerous MSI clustering and avoids confounding background signals in downstream embedding and classification tasks [3, 4].

Baseline Correction and Mass Alignment We perform mass axis alignment to correct for instrument drift and inter-sample variability [8]. A reference spectrum is computed as the median spectrum over cancerous pixels. Each pixel's raw spectrum $x \in \mathbb{R}^M$ is aligned to this reference using dynamic time warping (tolerance 0.1Da). We then subtract baseline noise via a top-hat morphological filter (structuring element size 50), ensuring spectra have a zero baseline.

Total Ion Current (TIC) Normalization To reduce variation in overall signal intensities, we apply TIC normalization to each spectrum:

$$x' = \frac{x}{\sum_{i=1}^{M} x_i + \epsilon},$$

where $\epsilon = 10^{-6}$ safeguards against zero-division. After TIC scaling, each spectrum is min–max normalized to the $[0, 1]$ range per pixel, as recommended for contrastive MSI pipelines [3].

Channel Selection We restrict input dimensionality by selecting the top 200 most abundant m/z channels across all cancerous pixels (ranked by mean intensity). This reduces noise from rare peaks and standardizes our feature dimension $M = 200$.

Spatial Patch Extraction To leverage spatial context, we extract overlapping patches of size 64×64 pixels from the malignant regions, using a stride of 32 pixels. Only patches with $\geq$50% malignant pixels are retained to maintain label purity. This yields a large and diverse patch set for training both autoencoders and contrastive encoders.

On-the-Fly Augmentation Training incorporates MSI-specific augmentations to improve robustness [3] :

- Additive Poisson noise (5% of maximum intensity),

- Random pixel dropout (10% zeroed),
- Affine transforms: rotations $\pm 10^\circ$ and translations ± 5 px.

Such augmentations mimic experimental variability and encourage the model to learn invariant latent representations.

3.5 Experiment Design

3.5.1 Principal Component Analysis (PCA)

PCA seeks a linear subspace of dimension $d \ll M$ capturing maximal variance. Given an $N \times M$ data matrix X (each row a flattened ion image of size $H \times W = M$), the objective is:

$$\max_{W \in \mathbb{R}^{M \times d}} \operatorname{Tr}(W^T X^T X W) \quad \text{s.t. } W^T W = I_d, \tag{3.1}$$

where I_d is the $d \times d$ identity. The solution $W = [u_1, \ldots, u_d]$ comprises the top-d eigenvectors of the covariance $X^T X$. Each projected vector $z = xW \in \mathbb{R}^d$ retains maximum variance. The only hyperparameter is the target dimension d. In MSI, PCA highlights dominant co-localization patterns but neglects nonlinear structure [20, 15].

3.5.2 t-SNE

t-SNE embeds high-dimensional points into a low-dimensional space by preserving pairwise similarities.

- Input: data vectors $\{x_i\}_{i=1}^N$, each $x_i \in \mathbb{R}^M$.
- Kernel bandwidths $\{\sigma_i\}$ chosen so that conditional perplexity matches a user-specified *perplexity* P:

$$\text{Perplexity}(P_i) = 2^{-\sum_j p_{j|i} \log_2 p_{j|i}} = P. \tag{3.2}$$

- High-dim probabilities:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/(2\sigma_i^2))}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2/(2\sigma_i^2))}, \quad p_{ij} = \frac{p_{i|j} + p_{j|i}}{2N}. \tag{3.3}$$

- Low-dim similarity:
$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}, \quad y_i \in \mathbb{R}^d. \tag{3.4}$$

- Minimize KL divergence:
$$\mathcal{L}_{\text{tSNE}} = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{3.5}$$

Key hyperparameters:

- P: Perplexity (controls effective neighborhood size, e.g., 30)
- d: Embedding dimension (commonly 2 or 3 for visualization)
- Learning rate
- Number of iterations

In MSI, t-SNE reveals local clusters of spectra but at high computational cost and without an explicit mapping for new data [15, 20].

3.5.3 Uniform Manifold Approximation and Projection (UMAP)

UMAP builds a topological representation of data as a fuzzy simplicial set and optimizes a low-dim embedding [20]:

- Construct k-nearest neighbor graph on X, using metric (e.g., Euclidean) with $n_{\text{neighbors}}$.
- Compute high-dim fuzzy simplicial set weights p_{ij} via a smooth k-NN kernel with local connectivity and "smoothness" controlled by a parameter ρ_i and σ_i.
- Low-dim weights $q_{ij} = \left(1 + a\|y_i - y_j\|^{2b}\right)^{-1}$, with parameters a, b chosen to match the high-dim curve.
- Minimize cross-entropy:
$$\mathcal{L}_{\text{UMAP}} = \sum_{i \neq j} \Big[p_{ij} \log \frac{p_{ij}}{q_{ij}} + (1 - p_{ij}) \log \frac{1 - p_{ij}}{1 - q_{ij}} \Big]. \tag{3.6}$$

Hyperparameters:

- $n_{\text{neighbors}}$: neighborhood size (e.g., 15–50)
- `min_dist`: minimum distance between embedded points (controls cluster tightness)
- d: target dimension

For MSI, UMAP offers runtimes faster than t-SNE with global structure preservation and an explicit embedding function for new samples.

3.5.4 Non-negative Matrix Factorization (NMF)

NMF decomposes non-negative data $X \geq 0$ into $W, H \geq 0$:

$$\min_{W \in \mathbb{R}_+^{N \times d}, H \in \mathbb{R}_+^{d \times M}} \|X - WH\|_F^2 + \lambda \left(\|W\|_1 + \|H\|_1\right). \tag{3.7}$$

Here:

- $X \in \mathbb{R}^{N \times M}$: N samples (flattened images) with M features (ion channels)
- W: $N \times d$ coefficient matrix (sample loadings)
- H: $d \times M$ basis matrix (spectral patterns)
- λ: sparsity regularization weight (controls parts-based decomposition)

In MSI, columns of H correspond to co-localized ion clusters (molecular signatures). NMF offers interpretability but may struggle with noise and complex spatial structure.

3.5.5 2D Convolutional Autoencoder (CNN-AE)

Our autoencoder operates on normalized ion image patches $x \in [0,1]^{H \times W}$ (single-channel grayscale). The CNN-AE comprises:

- **Encoder** E_θ: a sequence of L convolutional blocks:

$$z^{(0)} = x, \quad z^{(l)} = \sigma\big(\text{BN}(\text{Conv}(z^{(l-1)}; K_l, S_l, P_l))\big), \quad l = 1 \dots L, \tag{3.8}$$

Where:

- K_l: kernel size (e.g., 3×3)
- S_l: stride (e.g., 2 for downsampling)
- P_l: padding (e.g., 1 for 'same' output size)
- BN: batch normalization
- $\sigma(\cdot)$: ReLU activation

The final encoder output is flattened to $z \in \mathbb{R}^d$, our latent code. We set $d = 20$ by default.

- **Decoder** D_θ: mirrors the encoder with L transpose-convolutional blocks:

$$\hat{z}^{(L+1)} = z, \quad \hat{z}^{(l)} = \sigma\big(\mathrm{BN}(\mathrm{ConvTranspose}(\hat{z}^{(l+1)}; K'_l, S'_l, P'_l))\big), \tag{3.9}$$

ending in $\hat{x} = D_\theta(z) \in [0,1]^{H \times W}$.

- **Reconstruction Loss**: mean squared error

$$\mathcal{L}_{\mathrm{rec}}(\theta) = \frac{1}{NHW} \sum_{i=1}^{N} \|x_i - D_\theta(E_\theta(x_i))\|_2^2. \tag{3.10}$$

- **Total Loss** (if training end-to-end with classifier):

$$\mathcal{L}_{\mathrm{AE}} = \mathcal{L}_{\mathrm{rec}} + \alpha\, \mathcal{L}_{\mathrm{cls}}, \tag{3.11}$$

where $\alpha \geq 0$ balances reconstruction vs. downstream classification (we often pretrain AE with $\alpha = 0$, then fine-tune with $\alpha > 0$).

Hyperparameters summary:

- H, W: Spatial dimensions (e.g., 64×64)
- L: Number of conv/transpose-conv blocks (e.g., 4)
- d: Latent dimension (e.g., 20)

- Learning rate η (e.g., 3×10^{-4})
- Batch size B (e.g., 64)
- α: Classification weight (during joint training, e.g., 0.1)
- Optimizer: Adam($\beta_1 = 0.9, \beta_2 = 0.999$)
- Epochs: 100–200

3.5.6 Latent-Space Classifier

Given encoder outputs $z_i = E_\theta(x_i) \in \mathbb{R}^d$, we train a classifier

$$C_\phi : \mathbb{R}^d \to \{1, \ldots, K\},$$

Parameterized as a two-layer MLP with hidden size h and dropout rate p:

$$\begin{aligned} u &= \text{ReLU}(W^{(1)} z + b^{(1)}), \\ u_d &= \text{Dropout}(u; p), \\ \hat{y} &= \text{Softmax}(W^{(2)} u_d + b^{(2)}), \end{aligned}$$

Trained with cross-entropy:

$$\mathcal{L}_{\text{cls}}(\phi) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} \log \hat{y}_{i,k}. \tag{3.12}$$

Hyperparameters:

- h: Hidden dimension (e.g., 64)
- p: Dropout probability (e.g., 0.5)
- Learning rate η_{cls} (e.g., 1×10^{-3})
- Epochs: 50–100

3.5.7 Self-Supervised Contrastive Pretraining (Optional extension)

Before AE training, one may initialize the encoder f_ψ via contrastive learning:

- Generate two augmentations (x_i^1, x_i^2) of each ion image x_i via random intensity scaling, Gaussian noise, and random occlusion.
- Compute embeddings $h_i^m = f_\psi(x_i^m)$, then projection $p_i^m = g(h_i^m)$ and prediction $q_i^m = q(p_i^m)$.
- Contrastive loss:

$$\mathcal{L}_{\text{con}} = -\frac{1}{2}\sum_{m=1}^{2} \text{sim}\big(q(p_i^m)),\ \text{stopgrad}(p_i^{3-m})\big), \tag{3.13}$$

  where $\text{sim}(u, v) = u^T v/(\|u\|\|v\|)$.

After pretraining, initialize $E_\theta = f_\psi$, then fine-tune with reconstruction/classification losses. This often accelerates convergence and improves embedding quality.

3.6 Evaluation Metrics

We will assess the CNN autoencoder features in the context of disease-state classification by training standard classifiers (SVM, linear classifier, and multilayer perceptron) on the extracted embeddings and comparing performance to baseline features (PCA, NMF, t-SNE, UMAP, spec2vec, etc.). Classification performance will be quantified by common metrics like as precision, recall, accuracy, and F1 score on held-out test data. In particular, overall classification accuracy (and error rate) will be computed via cross-validation to ensure robust estimates. We will present confusion matrices to inspect class-wise errors and imbalances, as has been done in prior MSI studies. We will determine classification accuracy from a linear evaluation to quantify representation quality, and also presented confusion matrices to illustrate clustering/classification results [15]. Likewise, we will plot ROC curves and compute AUC scores (for binary or one-vs-rest cases) to assess classifier discrimination as a function of decision threshold. These curves provide a threshold-independent view of sensitivity and specificity. In all cases, we will report the mean and standard deviation of metrics over folds. Figure-quality visualizations such as confusion-matrix heatmaps and ROC plots will be used to communicate classifier performance. See figures 3.1 and 3.2.

In addition to these supervised metrics, we will evaluate the feature representations via unsupervised clustering metrics. Since the ground-truth disease labels (e.g., cancer vs. normal tissue) are known, we can cluster the feature vectors and compare clusters to the true labels. To this end, we will compute the Adjusted Rand Index (ARI) between the clustering result and the true labels; ARI

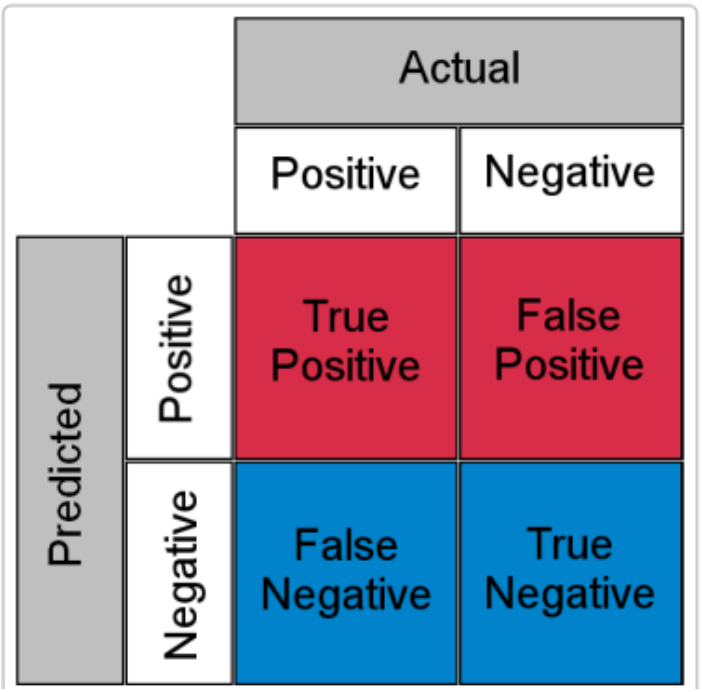

FIGURE 3.1: Example confusion matrix for a two-class classifier. Such matrices (true vs. predicted labels) will be generated to visualize per-class accuracy and misclassification patterns. The diagonal entries (True Positives and True Negatives) versus off-diagonals (False Positives/Negatives) inform sensitivity and precision.

measures how well the cluster assignment agrees with ground truth, corrected for chance [15]. A high ARI indicates that the feature space naturally separates the classes. We will also compute cluster purity, which assigns each cluster to its majority class and measures the fraction of correctly grouped samples; purity is the relative frequency of the most frequent annotation in a cluster [16]. Together, ARI and purity quantify external cluster quality against known labels. In addition, we will report the average silhouette score of the clustering to assess intra-cluster cohesion versus inter-cluster separation. The silhouette coefficient (ranging from 0 to 1) indicates how well each point fits its cluster compared to the next nearest cluster; high silhouette values suggest well-separated, tight clusters [17]. These unsupervised metrics complement the supervised metrics and follow the approaches recommended in MSI literature used to validate clustering of MSI spatial patterns [15].

For visualization, we will employ low-dimensional embeddings and heatmaps. We will generate t-SNE or UMAP scatter plots of the feature vectors (color-coded by true class) to qualitatively assess how well classes separate in 2D. Such plots have been widely used in MSI representation learning to illustrate class separation [4]. Likewise, we will inspect 2D projections of CNN-AE features versus

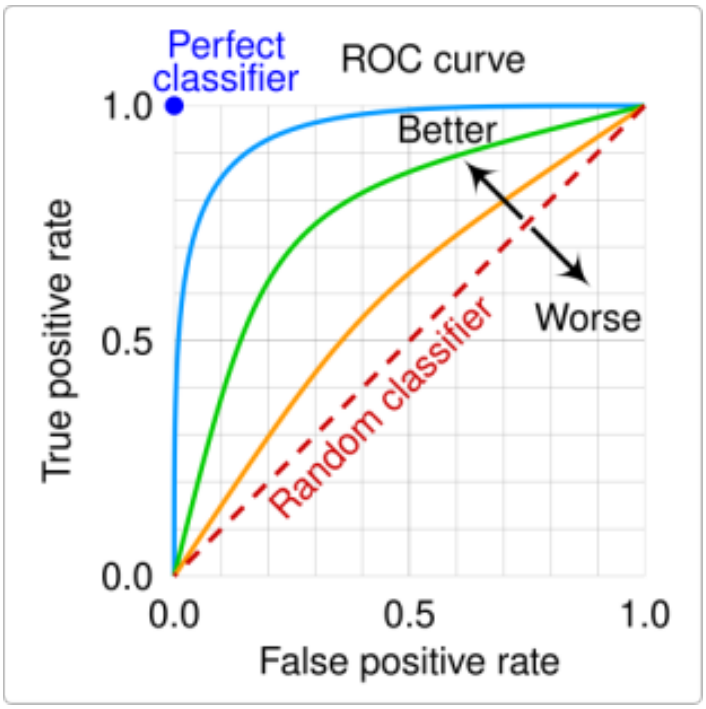

FIGURE 3.2: Example ROC curves for a binary classifier. We will similarly compute ROC curves and area-under-curve (AUC) for each classifier to evaluate sensitivity/specificity tradeoffs across thresholds.

baseline features to judge clustering structure. Finally, we will plot heatmaps of representative ion-intensity patterns or class-mean spectra for each cluster or class. Figure 3.3 (reproduced in) shows the average ion-intensity images for each cluster, which concisely summarizes spatial molecular patterns. We will create analogous heatmaps of cluster means or class means to highlight the spatial distribution of MSI signals for each category.

Overall, this multi-pronged evaluation strategy combines quantitative metrics and intuitive visualizations. Classification accuracy, ROC-AUC, and confusion matrices will directly compare how well features enable disease-state prediction. ARI, silhouette, and purity will reveal whether the learned feature space naturally groups samples of the same label together. t-SNE/UMAP plots will illustrate cluster separability in an accessible way (as in), and heatmaps of ion images or spectra will provide spatial context for each class. By reporting all these metrics and plots, we will thoroughly compare the CNN autoencoder embeddings against PCA, NMF, and other baselines, justifying our feature learning approach with strong quantitative and visual evidence.

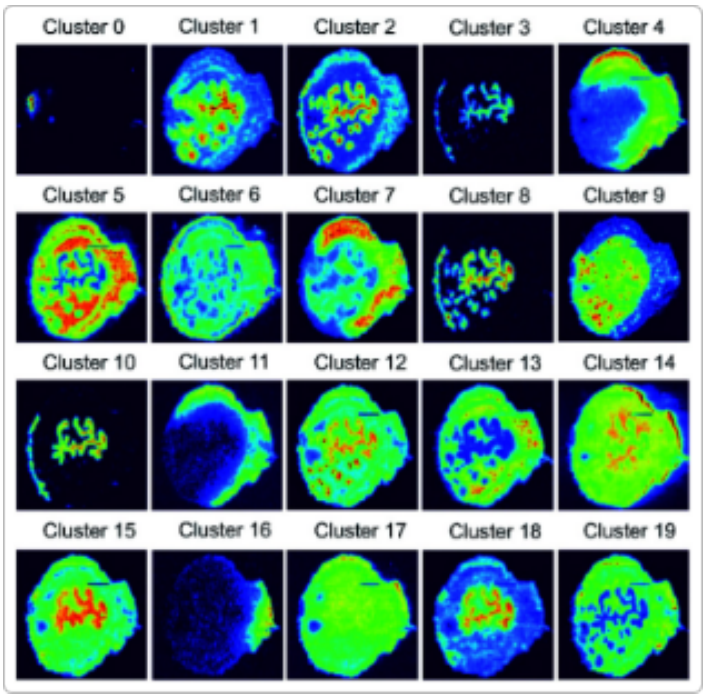

FIGURE 3.3: The average ion-intensity images for each cluster, which concisely summarize spatial molecular patterns.

3.7 Contribution to Feature Extraction

The feature extraction component of this study addresses three primary objectives: (1) reducing the dimensionality of high-dimensional MSI data, (2) uncovering informative structures in the latent space, and (3) generating embeddings suitable for downstream classification. The contributions of each approach are as follows:

- **Linear Baselines (PCA, NMF)**: Principal Component Analysis (PCA) [2, 15] and Non-Negative Matrix Factorization (NMF) [21, 15] provide interpretable, parts-based decompositions of spectral data. PCA captures maximal global variance, offering fast dimensionality reduction, while NMF yields non-negative, sparse factors that may correspond to biologically meaningful ion clusters. These linear methods serve as benchmarks against which more complex models can be compared.

- **Nonlinear Baselines (t-SNE, UMAP)**: Techniques such as t-SNE and UMAP preserve local and global data structures by modeling pairwise similarity relationships [20]. Embedding MSI pixels or patches into lower-dimensional manifolds reveals intrinsic clustering tendencies and

informs model design by highlighting underlying data topology that linear projections may miss.

- **Spatial–Spectral Autoencoder (CNN-AE)**: The 2D convolutional autoencoder integrates spatial context with spectral profiles [6, 4]. Convolutional layers encode local m/z co-localization within 64×64 patches, compressing each into a 20-dimensional latent vector. The decoder reconstructs the original ion-image patches, ensuring retention of spatial structure and molecular signatures. This end-to-end framework directly optimizes reconstruction loss (MSE), producing robust features that capture subtle, cancer-related patterns.

- **Clustering and Embedding Quality**: To assess the intrinsic structure of the latent spaces independent of supervised labels, we compute:

  - **Adjusted Rand Index (ARI)**: Ranges from -1 to 1. A value of 1 indicates perfect agreement between clusters and true labels; 0 indicates random assignments; and values below 0 indicate worse-than-random clustering [16].

  - **Cluster Purity**: Ranges from 0 to 1. A purity of 1 means each cluster contains samples from only one class (perfect homogeneity), while values near 0 indicate highly mixed clusters [16, 17].

  - **Silhouette Score**: Ranges from -1 to 1. Scores near 1 signify well-separated, cohesive clusters; scores around 0 indicate overlapping clusters; and negative scores suggest misassigned samples [17].

  These unsupervised metrics reveal the degree to which each feature extraction method captures class-relevant structure without direct supervision.

- **Classifier Readiness**: All extracted embeddings are evaluated via a small multilayer perceptron classifier trained to distinguish malignant from non-malignant regions. Comparative metrics—accuracy, F1 score, and ROC-AUC—quantify each method's ability to retain discriminative information, directly linking feature quality to the project's diagnostic objective.

Chapter 4

Research Plan

4.1 Timeline

TABLE 4.1: Research Timeline with Bi-Weekly Supervision Checkpoints

Phase	Dates	Activities
Setup and Data Acquisition	31 Jul–13 Aug	• Acquire raw mass spectrometry imaging (MSI) datasets and corresponding annotations, including cancer-annotated mouse brain and tissue samples. • Configure the computational environment, including setting up GPU nodes, Docker containers, and Python libraries (TensorFlow, PyTorch, Scikit-learn, etc.). Test cluster configurations to ensure compatibility with planned experiments. • Establish a version-controlled Git repository to manage code and experiments. Design a structured data storage format, specifying paths for raw data, processed data, and model outputs. Develop a metadata schema to track sample and preprocessing details for reproducibility.

Table continues on next page

Table 4.1 – Continued from previous page

Phase	Dates	Activities
Preprocessing and Exploration	14 Aug–3 Sep	• Filter MSI data to isolate tumor-positive regions or pixels using annotations. This step involves identifying regions corresponding to diseased versus healthy tissue. • Apply spectral preprocessing, including baseline correction to remove background noise and mass-to-charge (m/z) axis alignment to correct for instrumental drift. • Normalize spectral intensities using total ion current (TIC) scaling to account for sample-to-sample variability, followed by min-max scaling to standardize intensity values for downstream analysis. • Select the most informative m/z channels (top 200) by ranking them based on variance or domain-specific criteria. Extract fixed-size patches of 64×64 pixels (stride of 32) from spatial regions, preserving local spatial-spectral relationships. • Visualize the data to understand key patterns: plot spectral distributions, visualize spatial ion intensity maps, and inspect tumor versus non-tumor regions for differences.

Table continues on next page

Table 4.1 – Continued from previous page

Phase	Dates	Activities
Implementation of Baseline Feature Extraction Methods	4 Sep–24 Sep	• Implement dimensionality reduction techniques such as Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), Uniform Manifold Approximation and Projection (UMAP), and Non-Negative Matrix Factorization (NMF). Each method reduces the high-dimensional MSI data into compact embeddings for clustering and classification tasks. • Evaluate clustering quality using metrics such as Adjusted Rand Index (ARI), silhouette score, and purity. Compare the performance of each method in terms of clustering coherence and runtime efficiency. • Generate visualizations to assess the separability of the embeddings: use t-SNE and UMAP plots to depict clusters, overlaid with disease-state annotations for qualitative evaluation.

Table continues on next page

Table 4.1 – Continued from previous page

Phase	Dates	Activities
Training of CNN Autoencoder	25 Sep–8 Oct	• Design a 2D convolutional autoencoder (CNN-AE) architecture. The encoder extracts low-dimensional spatial-spectral features using stacked convolutional layers, while the decoder reconstructs the original MSI patches using transpose-convolutional layers. Define a latent dimension size of $d = 20$. • Train the CNN-AE using a reconstruction loss (mean squared error, MSE) to minimize the difference between the input and reconstructed patches. Augment the data with transformations (e.g., cropping, flipping, noise addition) to improve generalization. • Monitor reconstruction quality by visualizing reconstructed patches alongside the originals. Ensure the latent space captures key spectral-spatial information without redundancy.

Table continues on next page

Table 4.1 – Continued from previous page

Phase	Dates	Activities
Classifier Fine-Tuning and Evaluation	9 Oct–31 Oct	• Attach a Multilayer Perceptron (MLP) classifier to the pretrained encoder. Train the classifier using cross-entropy loss to map the learned embeddings to disease-state labels (e.g., cancer vs. healthy). • Evaluate classification performance using metrics such as recall, precision, accuracy and F1 score, and area under the ROC curve (AUC). Create confusion matrices to analyze per-class performance. • Compare the full workflow (CNN-AE + MLP classifier) against baseline methods (e.g., PCA + SVM) on metrics such as accuracy and runtime. • Prepare a detailed report with embedding visualizations, reconstructed ion image examples, clustering metrics, and classification results. Draft a LaTeX manuscript and presentation slide deck to disseminate findings.

4.2 Deliverables

1. **Data Package:** Filtered MSI patches, tumor masks, chemical-signature labels, and metadata.
2. **Preprocessing Scripts:** Python notebooks for baseline correction, alignment, normalization, and patch extraction.
3. **Baseline Analysis Report:** Embedding plots (PCA, t-SNE, UMAP, NMF), clustering purity tables, and runtime benchmarks.
4. **2D CNN-AE Implementation:** Modular code (PyTorch/TensorFlow) with pretrained weights and documentation.
5. **Classifier** & **VAE Extension:** Trained models, accuracy metrics, and comparison notes.
6. **Contrastive Module:** Augmentation pipeline, contrastive pretraining code, and evaluation results.
7. **Comprehensive Evaluation Notebook:** Combined metrics and visualization suite.
8. **Final Manuscript & Slides:** LaTeX paper draft and presentation slides.

4.3 Risk Factors

TABLE 4.2: Risk Factors and Mitigation Strategies

Risk	Impact	Mitigation Strategies
Data Quality & Label Noise	High	• Consult with domain experts on MSI data quality • Aggregate pixels into super-pixels to reduce noise • Implement a robust data validation pipeline

Continued on next page

Table 4.2 – Continued from previous page

Risk	**Impact**	**Mitigation Strategies**
Scope Limitations	High	• Focus on one linear method (PCA) and one nonlinear method (UMAP) as representatives to manage the scope • Document rationale for chosen methods and defer secondary analyses to future work
Compute Resource Limits	Medium	• Benchmark on representative subsets first • Employ mixed precision training
Hyperparameter Sensitivity	Medium	• Automated hyperparameter optimization • Early stopping with validation metrics • Adaptive learning rate scheduling
Integration Complexity	Medium	• Modular architecture design • Buffer time allocation in schedule
Reproducibility Issues	Medium	• Version control for all artifacts • Detailed documentation of parameters
Schedule Slippage	Low-Medium	• Early drafting of key components • Regular progress reviews • Parallel task execution

Continued on next page

Table 4.2 – Continued from previous page

Risk	Impact	Mitigation Strategies
Differences in Input Representations	Medium-High	• Clearly document and justify differences in inputs for different feature-extraction methods (PCA with 1D spectra versus CNN-AE with 2D patches) • Standardize inputs where possible, or interpret results within the context of input differences
Contrastive Learning Risks	Low	• Treat contrastive learning as prospective work, to be deprioritized if time or resources are constrained • Focus on simpler reconstruction-based approaches for the current scope

Chapter 5

Conclusion

The primary aim of this research is to develop and rigorously evaluate a computational framework for extracting informative and interpretable latent representations from high-dimensional mass spectrometry imaging (MSI) data, with the ultimate goal of enhancing cancer phenotype classification. By systematically comparing classical dimensionality-reduction methods (PCA, t-SNE, UMAP, NMF) against a purpose-built convolutional autoencoder (CNN-AE) architecture, this study addresses the dual challenges of preserving biologically meaningful molecular signatures and achieving robust downstream classification performance.

To achieve these objectives, the project employs a multi-stage methodology. First, extensive data preprocessing procedures—including baseline correction, total-ion-current normalization, and strategic channel selection—ensure that the input spectra are both standardized and focused on cancer-relevant features. Next, classical baselines (PCA, t-SNE, UMAP, NMF) are implemented to provide interpretable benchmarks in terms of variance capture, cluster purity, and computational efficiency. These methods serve to quantify the limitations of linear and non-parametric embeddings when confronted with complex, noisy MSI data.

Building upon these baselines, the research introduces a 2D CNN-AE designed to leverage spatial context in MSI patches. The encoder employs a sequence of convolutional blocks with batch normalization and ReLU activations to compress 64×64 pixel ion-image patches into a 20-dimensional latent space; the decoder mirrors this architecture via transpose convolutions to reconstruct the original

patches. Training proceeds via unsupervised reconstruction loss minimization with targeted augmentations (Poisson noise, pixel dropout, affine transforms) to build noise-resilient embeddings. The inclusion of supervised fine-tuning aligns latent features with tumor labels, optimizing the representations for downstream classification tasks.

The significance of these methods lies in their complementary strengths: classical techniques offer fast, interpretable projections but lack end-to-end adaptability, whereas the CNN-AE provides a flexible, spatially aware mechanism that can be directly coupled to classifiers. By comparing reconstruction error (MSE), clustering metrics (Adjusted Rand Index, purity, silhouette score), and classification outcomes (accuracy, precision, recall, F1, ROC-AUC) across methods, the study delineates the trade-offs between interpretability, computational cost, and predictive power.

Prospective results are expected to demonstrate that the CNN-AE—with optimized latent dimensionality and architecture depth—will (a) yield lower reconstruction error than PCA and NMF; (b) produce more tightly clustered representations of malignant versus non-malignant regions, as quantified by ARI and silhouette scores; and (c) support downstream classifiers that outperform those trained on classical embeddings, achieving higher accuracy and AUC in tumor detection tasks. Moreover, the sparsity and disentanglement constraints under consideration may reveal latent features that correspond to known cancer biomarkers, while also uncovering novel spatial–molecular associations warranting further biological investigation.

While contrastive learning techniques were initially considered to enhance the quality and invariance of learned embeddings, their implementation is deferred to future work. Contrastive pretraining offers a promising avenue for further improving feature separability and robustness, particularly in the presence of noise or variability in MSI data. As a prospective enhancement, it represents an opportunity to extend the current framework in subsequent studies.

In summary, this research aims to set methodological guidelines for MSI representation learning by integrating best practices from linear, manifold, and deep-learning approaches, thereby bridging the gap between theoretical model development and real-world cancer diagnostics applications. The anticipated outcomes will not only advance the technical state of the art in MSI analysis but also provide interpretable insights into the molecular underpinnings of cancer, guiding future biomarker discovery efforts.

Bibliography

[1] Theodore Alexandrov. "MALDI imaging mass spectrometry: statistical data analysis and current computational challenges". In: *BMC bioinformatics* 13.Suppl 16 (2012), S11.

[2] Theodore Alexandrov. "Spatial metabolomics and imaging mass spectrometry in the age of artificial intelligence". In: *Annual review of biomedical data science* 3.1 (2020), pp. 61–87.

[3] Lei Guo, Chengyi Xie, Rui Miao, Jingjing Xu, Xiangnan Xu, Jiacheng Fang, Xiaoxiao Wang, Wuping Liu, Xiangwen Liao, Jianing Wang, et al. "DeepION: A Deep Learning-Based Low-Dimensional Representation Model of Ion Images for Mass Spectrometry Imaging". In: *Analytical Chemistry* 96.9 (2024), pp. 3829–3836.

[4] Hang Hu, Jyothsna Padmakumar Bindu, and Julia Laskin. "Self-supervised clustering of mass spectrometry imaging data using contrastive learning". In: *Chemical science* 13.1 (2022), pp. 90–98.

[5] Hang Hu and Julia Laskin. "Emerging computational methods in mass spectrometry imaging". In: *Advanced Science* 9.34 (2022), p. 2203339.

[6] Florian Huber, Sven van der Burg, Justin JJ van der Hooft, and Lars Ridder. "MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra". In: *Journal of cheminformatics* 13.1 (2021), p. 84.

[7] Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H Spaaks, Faruk Diblen, Simon Rogers, and Justin JJ Van Der Hooft. "Spec2Vec: Improved mass spectral similarity scoring through the learning of structural relationships". In: *PLoS computational biology* 17.2 (2021), e1008724.

[8] Paolo Inglese, James L Alexander, Anna Mroz, Zoltan Takats, and Robert Glen. "Variational autoencoders for tissue heterogeneity exploration from (almost) no preprocessed mass spectrometry imaging data". In: *arXiv preprint arXiv:1708.07012* (2017).

[9] Niek F de Jonge, Joris JR Louwen, Elena Chekmeneva, Stephane Camuzeaux, Femke J Vermeir, Robert S Jansen, Florian Huber, and Justin JJ van der Hooft. "MS2Query: reliable and scalable MS2 mass spectra-based analogue search". In: *Nature Communications* 14.1 (2023), p. 1752.

[10] Anish Mudide, Joshua Engels, Eric J Michaud, Max Tegmark, and Christian Schroeder de Witt. "Efficient dictionary learning with switch sparse autoencoders". In: *arXiv preprint arXiv:2410.08201* (2024).

[11] Michael Murphy, Stefanie Jegelka, Ernest Fraenkel, Tobias Kind, David Healey, and Thomas Butler. "Efficiently predicting high resolution mass spectra with graph neural networks". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 25549–25562.

[12] Sphamandla Ntshangase, Sipho Mdanda, Tricia Naicker, Hendrik G Kruger, Sooraj Baijnath, and Thavendran Govender. "Spatial distribution of elvitegravir and tenofovir in rat brain tissue: Application of matrix-assisted laser desorption/ionization mass spectrometry imaging and liquid chromatography/tandem mass spectrometry". In: *Rapid Communications in Mass Spectrometry* 33.21 (2019), pp. 1643–1651.

[13] Katja Ovchinnikova, Lachlan Stuart, Alexander Rakhlin, Sergey Nikolenko, and Theodore Alexandrov. "ColocML: machine learning quantifies co-localization between mass spectrometry images". In: *Bioinformatics* 36.10 (2020), pp. 3215–3224.

[14] Annapurna Pamreddy, Sooraj Baijnath, Tricia Naicker, Sphamandla Ntshangase, Sipho Mdanda, Hlengekile Lubanyana, Hendrik G Kruger, and Thavendran Govender. "Bedaquiline has potential for targeting tuberculosis reservoirs in the central nervous system". In: *RSC advances* 8.22 (2018), pp. 11902–11907.

[15] Mridula Prasad, Geert Postma, Pietro Franceschi, Lutgarde MC Buydens, and Jeroen J Jansen. "Evaluation and comparison of unsupervised methods for the extraction of spatial patterns from mass spectrometry imaging data (MSI)". In: *Scientific reports* 12.1 (2022), p. 15687.

[16] Vera Rieder, Karin U Schork, Laura Kerschke, Bernhard Blank-Landeshammer, Albert Sickmann, and Jorg Rahnenfuhrer. "Comparison and evaluation of clustering algorithms for tandem mass spectra". In: *Journal of proteome research* 16.11 (2017), pp. 4035–4044.

[17] Ketan Rajshekhar Shahapure and Charles Nicholas. "Cluster quality analysis using silhouette score". In: *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*. IEEE. 2020, pp. 747–748.

[18] Aditya Divyakant Shrivastava, Neil Swainston, Soumitra Samanta, Ivayla Roberts, Marina Wright Muelas, and Douglas B Kell. "MassGenie: a transformer-based deep learning method for identifying small molecules from their mass spectra". In: *Biomolecules* 11.12 (2021), p. 1793.

[19] Rebecca L Siegel, Angela N Giaquinto, and Ahmedin Jemal. "Cancer statistics, 2024". In: *CA: a cancer journal for clinicians* 74.1 (2024), pp. 12–49.

[20] Tina Smets, Nico Verbeeck, Marc Claesen, Arndt Asperger, Gerard Griffioen, Thomas Tousseyn, Wim Waelput, Etienne Waelkens, and Bart De Moor. "Evaluation of distance metrics and spatial autocorrelation in uniform manifold approximation and projection applied to mass spectrometry imaging data". In: *Analytical chemistry* 91.9 (2019), pp. 5706–5714.

[21] Gustavo F Trindade, Marie-Laure Abel, and John F Watts. "Non-negative matrix factorization of large mass spectrometry datasets". In: *Chemometrics and Intelligent Laboratory Systems* 163 (2017), pp. 76–85.

Wits University Faculty of Science post-graduate student AI declaration

I understand that the use of generative AI tools (such as ChatGPT or similar) without explicitly declaring such use constitutes a form of plagiarism and is classified by Wits University as academic misconduct.

I declare that in the course of conducting the research towards my degree or in the preparation of this thesis/dissertation/research report (select one by marking with an X):

I **did not** make use of generative AI tools ☐

I **did** make use of generative AI tools for the following (tick all that apply):

1. Idea Generation (research problem/design, hypothesis) ☒
2. Sourcing Related Work (summarising, identifying sources) ☒
3. Methods and Experiment Design (experiment setup, model t[illegible]g) ☐
4. Data Analysis (presentation, coding, interpretation) ☐
5. Theoretical Development (theorem proving, conceptual analy[illegible] ☐
6. Code Development (generating algorithms, writing scripts) ☐
7. Presentation (rendering graphics, formatting) ☐
8. Editing (grammar, readability) ☒
9. Writing (text generation, document structuring) ☒
10. Citation Formatting (structuring, organising) ☐

If other uses were involved, please specify below:

Generative AI tool used (list all)	**Used for?**

If generative AI tools were used as an integral part of the experimental design or in the direct execution of my research, I confirm that details of this use are clearly outlined in the relevant experimental/methodology chapters of my thesis/dissertation/research report.

Student number: 2581366

Candidate signature: Inozarwa

Date: 20/04/2025