Oxford Academic Bioinformatics, Version 1.0, 2024, pp. 1–33

doi: DOI HERE
Advance Access Publication Date: In proceedings
Structural Bioinformatics

STRUCTURAL BIOINFORMATICS

SpectraFlow: A Novel Feature Selection Framework for Overcoming Challenges in 1D NMR Spectroscopy

Adrian Wesek,[1,*] Panteleimon G. Takis,[2] Caroline J. Sands,[1] Eric E.C. de Waal,[3] Zhong Chen,[4] Nandor Marczin[1] and Ling Li[5]

[1]Department of Surgery & Cancer, Imperial College London, United Kingdom, [2]Department of Chemistry, University of Ioannina, Greece, [3]Department of Anesthesiology, University Medical Center Utrecht, Netherlands, [4]Ashford and Saint Peter's Hospitals NHS Trust, United Kingdom and [5]Department of Engineering, City University, United Kingdom
[*]Corresponding author: adrianwesek@yahoo.com

Abstract

Motivation: The exploration of metabolic profiles through NMR spectroscopy has been dominated by targeted approaches, favoured for their specificity and direct relevance to known metabolic pathways. However, the potential of untargeted metabolomics to uncover novel biomarkers remains largely untapped due to its inherent challenges. These include the presence of noisy features, high dimensionality, and intercorrelation among features, which conventional analytical methods struggle to adequately address. The lack of robust analytical frameworks capable of navigating these complexities has hindered the full exploitation of untargeted metabolomics' potential to provide comprehensive insights into disease mechanisms and therapeutic targets.
Results: To bridge this gap, we introduce SpectraFlow, an innovative feature selection framework explicitly designed for the nuanced landscape of untargeted NMR metabolomics data. SpectraFlow excels in isolating noise-free metabolic features from global spectral data, demonstrating a keen ability to enhance predictive performance while ensuring clinical relevance. Our findings reveal that SpectraFlow not only corroborates several established biomarkers but also unveils novel metabolic features with potential implications for understanding and treating Vasoplegia syndrome.
Availability: SpectraFlow is freely available on GitHub at github.com/adigoryl/SpectraFlow.git.
Supplementary materials: Available at Bioinformatics online

Key words: Untargeted metabolomics, SpectraFlow, NMR spectroscopy, Feature selection, Vasoplegia syndrome, Sequential Attention, Wavelet Denoising, PCA-binning

Introduction

Vasoplegia syndrome, a formidable challenge observed post-cardiac surgery, is characterised by severe vasodilation leading to hypotension, despite maintaining normal or elevated cardiac output. This condition is closely associated with significant post-operative risks, such as multi-organ dysfunction or failure (1) (2) (3), underscoring the critical need for accurate prediction and early identification of at-risk patients. Despite its prevalence in nearly a third of patients undergoing cardiac surgery, the literature on predictive models for Vasoplegia is notably sparse. Traditional statistical methods employed in existing studies often fail to achieve predictive accuracies beyond 80% Area under the ROC Curve (4) (5), indicating a substantial room for improvement through advanced modelling techniques.

Metabolomics, utilising Nuclear Magnetic Resonance (NMR) spectroscopy and Liquid Chromatography–Mass Spectrometry (LC-MS) analysis, offers profound insights into metabolic processes, reflecting the organism's biochemical activity (6) (7) (8) (9). This field holds promise for early disease diagnosis and the discovery of new biomarkers. Yet, the inherent complexity of metabolomic datasets, characterised by high dimensionality and significant intercorrelation among variables, poses considerable analytical challenges (10) (6). These challenges necessitate sophisticated feature selection methods to distil meaningful insights effectively.

Historically, targeted metabolomics has been favoured for its direct approach and the simplicity of interpreting spectral data by focusing exclusively on known biomarkers. However, this method may overlook novel or unidentified features of potential significance (11) (12). In contrast, untargeted metabolomics provides a comprehensive view, examining a broad spectrum of metabolites without bias. Despite its advantages, untargeted metabolomics grapples with issues such as noise, the curse of dimensionality, and varied signal intensities, all of which require precise and robust analytical techniques to overcome (13).

In response to these challenges, we introduce the SpectraFlow pipeline, a novel feature selection framework designed expressly for the analysis of untargeted metabolomics data. Applied to a dataset on Vasoplegia, SpectraFlow aims not only to achieve superior predictive accuracy compared to traditional and targeted methods but also to identify novel biomarkers that could inform better post-surgical outcomes. This study aims to rigorously evaluate the SpectraFlow pipeline, emphasising its capability to

handle the complexities of untargeted metabolomics data and to discover clinically relevant biomarkers. We highlight two main contributions of this work: firstly, the introduction of SpectraFlow as an innovative combination of denoising, PCA-binning, advanced modelling techniques, rank concatenation, and a data-efficient rank evaluation strategy. This unique combination represents a significant advancement in the field of metabolomics feature selection. Secondly, by applying SpectraFlow to the study of post-operative Vasoplegia, we not only enhance our understanding of this complex condition but also reveal potential early-detection biomarkers, thus opening new avenues for disease management and research. Our findings underscore the transformative potential of SpectraFlow in metabolomic research, paving the way for novel insights into disease mechanisms and therapeutic interventions.

Materials and Methods

Dataset

In this study, we analyse findings from two datasets associated with a cohort of patients undergoing evaluation for Vasoplegia syndrome, comprising 147 patient samples with 39 positive and 108 negative cases for Vasoplegia.

The primary dataset, known as the 1D ^1H Carr-Purcell Meiboom-Gill (CPMG) dataset, was generated through NMR spectroscopy, offering comprehensive metabolic profiles represented by 18,637 variables for the entire cohort. This broad dataset served as the foundation for our untargeted analysis approach.

From this extensive dataset, a targeted subset was curated, focusing on 28 specific metabolic biomarkers identified using Small Molecule Enhancement Spectroscopy (SMolESY). The methodology behind the biomarkers' quantification and their selection criteria are detailed in (49) (50).

Metabolic phenotyping for both datasets was conducted at The National Phenome Centre (51), Imperial College, utilising blood plasma samples collected from patients prior to Left Ventricular Assist Device (LVAD) implantation, at UMC Utrecht, Netherlands.

The Origin and Challenges of Noise in NMR Data

NMR spectroscopy offers deep insights into molecular structures and interactions, but it's not without challenges. Noise in NMR data originates from various sources, including sample characteristics, equipment constraints, external interferences, and the random behaviours intrinsic to molecular processes (14). Even with meticulous instrument calibration, rigorous sample preparation, and optimised experimental setups, some noise remains, a consequence of the inherently stochastic nature of molecular motion.

The intrinsic noise complicates data scaling. However, for the optimal performance of machine learning models, scaling is essential. When data is correctly scaled, it ensures that all features adhere to a uniform scale, mitigating challenges like overfitting and ensuring each feature is given an equal opportunity during the feature selection process (15). This uniformity becomes especially vital in the context of global NMR spectra feature selection. While there can be significant differences in signal intensities, a feature with a lower intensity isn't necessarily less relevant.

Unit variance scaling, among various scaling methods for NMR data (16) (17) (18), stands out as notably beneficial for machine learning approaches relying on iterative optimisation methods like gradient descent. However, when applying scaling techniques, such as unit variance scaling, to datasets with inherent noise, a notable risk emerges: the inadvertent amplification of this noise. Such amplification can lead to the misinterpretation of noise as authentic data patterns, misleading feature selection processes and compromising model accuracy.

Given the criticality of data scaling, it's essential to augment it with potent noise reduction methods. This combination ensures that the data remains consistent in feature scales while also staying true to actual data patterns.

Wavelet Transform: A Solution to NMR Noise

Wavelet transform (WT) (19), a powerful mathematical tool, has been employed in various fields, ranging from signal processing to image compression (20). Its ability to decompose signals into different frequency components makes it particularly useful for signal compression and noise reduction tasks. Through the Discrete Wavelet Transform, NMR spectra can be broken down into a series of coefficients, reflecting its representation in the wavelet domain. When some of these coefficients have minimal values, they can be filtered out using thresholding, resulting in denoised data that maintains the integrity of the original structure (21).

Wavelets have been highlighted in several studies for their effectiveness in enhancing the clarity of NMR data by improving the signal-to-noise ratio (22) (23). More recently, WT has been recognised for its efficiency in removing NMR noise, outperforming methods such as Singular Value Decomposition and Hankel Matrix factorisation (24). However, no research has yet explored the significance of wavelet application as an NMR data preprocessing step for machine learning and feature selection. This paper seeks to briefly address that gap. For more details on the specific wavelet techniques we employed, please refer to Supplementary Material.

Addressing Dimensionality and Intercorrelation with PCA Binning

Selecting features in high-dimensional NMR datasets is notably challenging, particularly when the large number of data points or features greatly exceeds the sample size. This imbalance significantly increases the risk of overfitting, which can prevent models from accurately capturing and generalising real data patterns. A compounding issue is the inherent intercorrelations within NMR data: closely spaced metabolic features often share similar information due to the spectral data's dense structure (25) (26) (27). This closeness can lead to a higher chance of selecting metabolites that mirror the same metabolic activities, introducing redundancy and obscuring the identification of unique and insightful features. Such redundancy ultimately detracts from the model's performance and the precision of biomarker discovery.

To mitigate these issues, Principal Component Analysis (PCA) is deployed as a sophisticated tool for dimensionality reduction, widely recognised for its utility in various fields (28). Unlike traditional applications of PCA that process the dataset as a whole, our method applies PCA to discretely segmented ppm regions. This segmentation into bins, followed by PCA application, allows the first principal component of each bin to serve as a concise representation of its region, significantly reducing the dataset's complexity (29). This approach effectively addresses intercorrelations by ensuring each bin represented by PCA offers a unique snapshot of the dataset, thereby enhancing the selection process's specificity and relevance.

Maintaining a connection to the original ppm regions after feature selection is crucial for ensuring the interpretability and scientific validity of the findings. This traceability is essential for grounding any derived scientific hypotheses firmly within the dataset's original structure.

In deduction, PCA binning acts as an efficient preprocessing strategy that refines NMR datasets for enhanced feature selection. By tackling dimensionality and intercorrelations, it facilitates a more

detailed and meaningful analysis, paving the way for deeper insights into NMR data. Detailed methodology, including the PCA binning algorithm's pseudo-code, is provided in Supplementary Material.

Sequential Attention for Robust Feature Selection

Feature selection stands as a pivotal procedure in machine learning and statistics, aiming to select a subset of k essential features from a larger pool of d features that maximise a model's predictive performance (30). This undertaking becomes more demanding in biomedical settings. Beyond merely enhancing predictive accuracy, the selected features should align with practical biomedical interpretations, as their accurate identification can unveil insights into disease progression or potentially pave breakthroughs in drug discovery.

Historically, techniques such as Support Vector Machines, Decision Trees, Random Forests, and gradient-boosted models have led the charge in feature selection (31). While these methods have consistently demonstrated their reliability, the advancements in deep learning introduce potential avenues for refining feature selection through more complex data representations.

One noteworthy method within deep learning is the attention mechanism, which has revolutionised the computer vision and natural language processing domain (32). Its unique capability allows it to dynamically assign degrees of importance to contrast features, enabling models to "focus" on essential data points. Nevertheless, when deploying attention for feature selection, complexities emerge. Traditional attention mechanisms can overlook the marginal contributions a feature offers given the presence of other already selected features. The omission of these residual values can inadvertently lead to the selection of redundant features or potentially overlooking valuable synergies (33).

To address these challenges, we adopt the "Sequential Attention"[1] algorithm in our pipeline (33). The Sequential Attention algorithm ingeniously merges the strengths of the greedy forward selection method with the efficiency of deep learning. Traditional greedy forward selection, while thorough in assessing each feature's impact, is computationally demanding due to the need to train a model for each feature combination. This approach becomes inefficient, especially when dealing with a large number of features.

Building on the foundational concept of attention, the Sequential Attention mechanism provides a nuanced approach to feature selection. Unlike standard attention mechanisms that may allocate significance broadly across features, Sequential Attention employs a systematic process to iteratively discern and prioritise. Beginning with initial importance scores assigned to each feature, these scores undergo periodic adjustments as training advances. At specified intervals, features with the highest scores are locked in, and their scores are subsequently reset. This iterative process ensures that, over time, a refined subset of features emerges, each having proven its significance. This iterative approach mirrors the granularity of greedy forward selection but achieves it more efficiently, without the need for repeated model training sessions.

Furthermore, the design of the Sequential Attention algorithm ensures differentiability, facilitating smooth gradient-based optimisation. This versatility allows it to be integrated seamlessly as a component within larger models, including various neural network layers. When incorporated into models, this refined feature set becomes instrumental in driving accurate predictions and, thus, provides a potent and streamlined method for feature selection suitable for biomedical applications.

[1] We modified the original Sequential Attention algorithm to maintain the order in which features are selected.

SpectraFlow: Approach for Global NMR Feature Selection

The analysis and interpretation of global NMR data require robust preprocessing and feature selection strategies. Our pipeline synergistically combines denoising, PCA-binning, sequential attention, and an MLP model to address challenges specific to NMR data.

Denoising

Noise is an inherent aspect of 1D ^1H CPMG NMR spectra and can significantly influence downstream analyses. For a given spectra represented as $S_{\text{raw}}(p)$, where p denotes the chemical shift, denoising becomes essential to ensure data quality. The goal is to extract the inherent true signal $S(p)$ from the raw data. This is mathematically captured as:

$$S_{\text{raw}}(p) = S(p) + N(p) \tag{1}$$

In this equation, $N(p)$ denotes the noise component, and our denoised signal, after processing, closely approximates:

$$S_{\text{denoised}}(p) \approx S(p) \tag{2}$$

PCA-binning

After denoising, the next goal is to optimise the data's structure for more productive feature selection modelling. For this purpose, we employ the PCA-binning method, as detailed below:

- Let us represent each subject's global spectra as a sequence of values:

$$X(i) = [v_1, v_2, ..., v_N] \tag{3}$$

  where N signifies the total number of spectral points.
- Instead of considering each spectral point individually, we group them into bins according to the ppm regions. If we take a step size of 0.005 ppm and an overlap of 0.0025 ppm, we generate bins B_j where each B_j contains a subset of values sourced from $X(i)$, based on their ppm range.
- Every bin then goes through a PCA transformation. The first principal component obtained from the PCA captures the most variance and hence represents the entire bin. This process yields a new array of values for each subject, given as

$$Z(i) = [P(B_1), P(B_2), ..., P(B_K)] \tag{4}$$

  where K stands for the number of constructed bins and $P(B_j)$ is the score of the first principal component for bin B_j.

By employing this PCA-binning strategy, we not only decrease the complexity of the dataset but also secure a representation that captures the most critical information of each ppm region.

Sequential Attention

Upon obtaining the binned PCA features symbolised by $Z(i)$, it's vital to discern and emphasise the most informative bins. The sequential attention mechanism serves this exact purpose by attributing dynamic weights to these binned features:

$$\mathbf{A} = \text{Softmax}(\mathbf{W_a} Z(i) + \mathbf{b_a}) \tag{5}$$

Where:

- $\mathbf{W_a}$ and $\mathbf{b_a}$ denote trainable weights and biases of the attention layer, respectively.

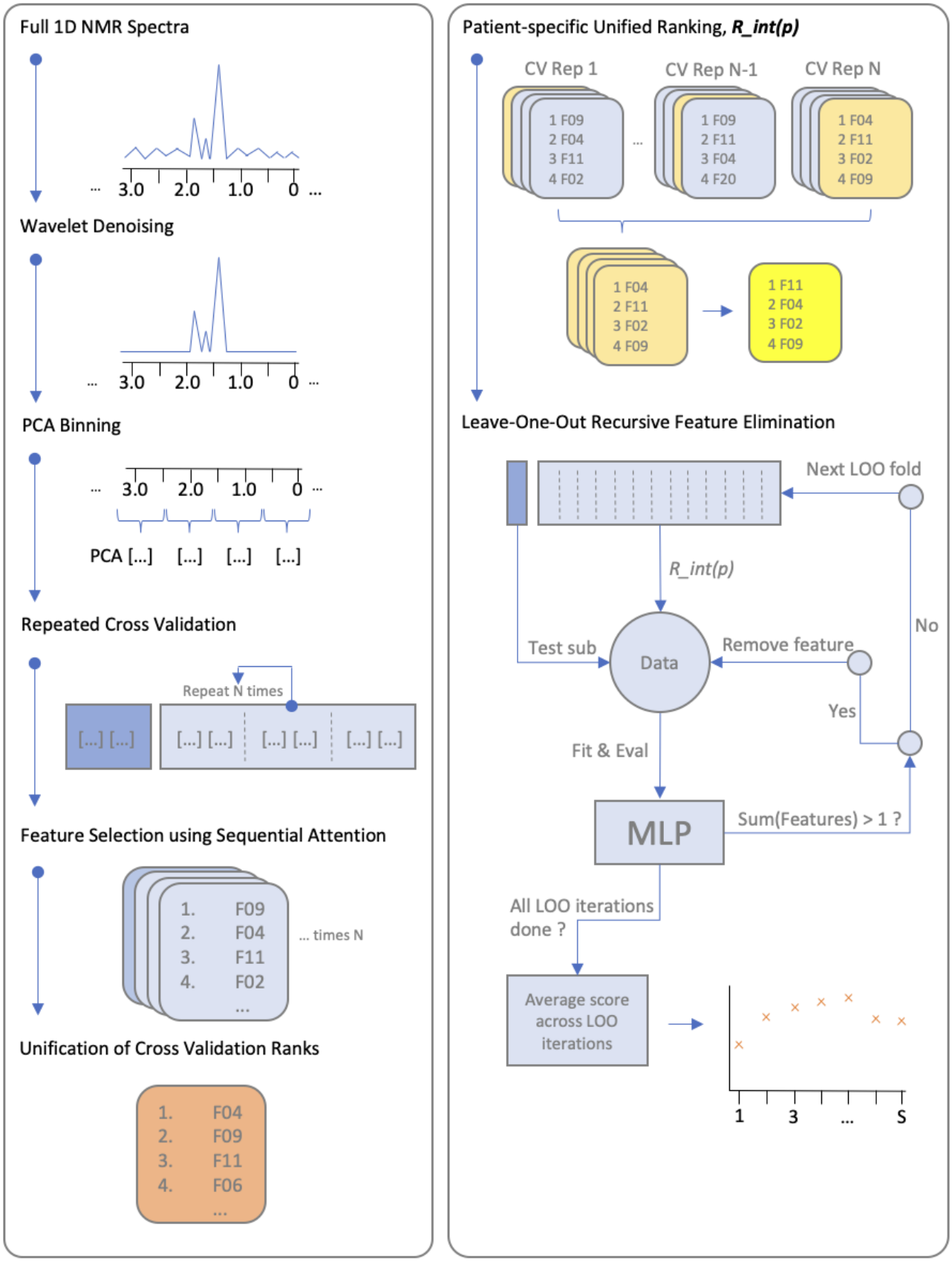


Fig. 1. An abstract representation of the SpectraFlow pipeline. On the left-hand side, the feature selection framework is detailed, and on the right-hand side, the evaluation of the Unified Ranking derived from Cross-Validation is presented.

- $\mathbf{A}$ represents the attention-weighted array.

The interaction between the weighted bins and the original binned features is described as:

$$\mathbf{Z}_{\text{attended}} = \mathbf{A} \odot Z(i) \tag{6}$$

Multilayer Perceptron (MLP)

The MLP is the backbone of our pipeline, effectively integrating and learning from the feature representations produced by the sequential attention mechanism. Each layer undergoes linear transformations and non-linear activations, successively transforming the attention-weighted features:

$$\mathbf{Z_1} = \mathbf{Z}_{\text{attended}}\mathbf{W_1} + \mathbf{b_1} \tag{7}$$

$$\mathbf{H_1} = \text{Activation}(\mathbf{Z_1}) \tag{8}$$

$$\vdots \tag{9}$$

$$\mathbf{Z_n} = \mathbf{H_{n-1}}\mathbf{W_n} + \mathbf{b_n} \tag{10}$$

$$\mathbf{H_n} = \text{Activation}(\mathbf{Z_n}) \tag{11}$$

During training, the model utilises gradient descent to iteratively adjust its weights. Given a batch size of m, the weight update is mathematically expressed as:

$$\mathbf{W}_{\text{new}} = \mathbf{W}_{\text{old}} - \frac{\alpha}{m}\sum_{i=1}^{m}\nabla_{\mathbf{W}} L_i \tag{12}$$

Where α represents the learning rate. The combination of sequential attention and MLP ensures the optimal selection and interpretation of NMR features, capturing the underlying complexities of the dataset.

Feature Rank Aggregation

Given a dataset with X features, the algorithmic pipeline is designed to select a subset of the top S features. Here, S is an externally defined parameter.

For a cross-validation scheme with C folds, repeated R times, the pipeline produces $C \times R$ distinct feature rankings, each of size S.

The inherent variability in training samples across different cross-validation folds, particularly pronounced in untargeted NMR datasets with a high feature-to-sample ratio, produces rankings that exhibit both consistency and variance. This mixture indicates some features' recurrent presence, while others differ across rankings.

This variance underscores the necessity for an aggregated ranking approach capable of merging these diverse outcomes into a singular, unified ranking. We implement the following steps:

- **Extraction of Unique Features:** Initially, we gather all unique features from the $C \times R$ distinct rankings, forming a set U. The size of U, denoted as $|U|$, could theoretically reach $S \times C \times R$, but typically, it's less due to the overlap of features across different rankings.
- **Frequency-based Selection:**
  Next, we calculate the occurrence frequency $f(F_i)$ for each feature F_i in the set U across all $C \times R$ rankings. The frequency is given by the sum of appearances of F_i in each ranking R_j:

$$f(F_i) = \sum_{j=1}^{C \times R} I(F_i \in R_j) \tag{13}$$

  Here, I is the indicator function that returns 1 if F_i is present in the ranking R_j, and 0 otherwise. Based on these frequencies, we select the top S features that appear most frequently across the rankings.
- **Average Position Ordering:**
  For these top S features, we calculate their average position $p(F_i)$ from the rankings where they are present:

$$p(F_i) = \frac{1}{f(F_i)} \sum_{\{j : F_i \in R_j\}} P(F_i, R_j) \tag{14}$$

  In this formula, $P(F_i, R_j)$ is the position of feature F_i in ranking R_j, and the summation runs over all j indices where F_i is included in R_j. The selected features are then ordered based on their average positions to finalise the aggregated ranking.

This streamlined approach ensures a more accurate and representative ranking of features, effectively capturing both the frequency of occurrence and the average positioning of features across multiple folds of cross-validation. Let's denote the unified ranking as R_u.

Unified Ranking and Evaluation Framework

Regardless of the specific task or dataset in consideration, a consistent evaluation framework is employed. We utilise a 10-fold cross-validation (CV) strategy, repeated 10 times. Each fold is stratified to maintain the balance between class labels. Prior to each repetition, the dataset undergoes a shuffle, followed by the generation of data folds. This scheme lays the foundation for our metabolite or ppm region selection analysis.

For every iteration of this setup, the SpectraFlow pipeline trains on $n - i$ folds, while the performance of the produced feature ranking is gauged using the test i fold, where n stands for all CV folds. As a direct consequence, executing a 10-fold CV, reiterated 10 times, generates 100 distinct feature rankings. Each of these rankings corresponds to a specific training partition within the overarching scheme.

With the primary aim of generating a singular, consolidated ranking from these 100, we employ the ranking concatenation method, as detailed in Section 2.7, resulting in a unified ranking, denoted as R_u. The performance and validity of R_u are then estimated using the Leave-One-Out Recursive Feature Elimination (LOO-RFE) methodology.

LOO-RFE

The Leave-One-Out method systematically evaluates feature relevance by training an MLP model on the dataset while excluding the data of the currently evaluated patient. This ensures each patient is individually tested once, with the rest of the dataset used for training. To prevent data contamination or leakage, for the patient under evaluation, we specifically extract rankings from cross-validation phases where this patient's data were not included in the training set. Such precautions ensure that no evaluation data samples were involved in generating these rankings, maintaining the integrity and unbiased nature of our evaluation. Subsequently, these rankings are aggregated to form a patient-specific unified ranking, $R_{int}(p)$, defined as:

$$R_{int}(p) = \bigcup_{k=1}^{C \times R} R_k(p) \tag{15}$$

where $R_{int}(p)$ represents this integrated ranking for patient p, aggregating the rankings $R_k(p)$ from each k^{th} iteration that excluded patient p's data.

The Recursive Feature Elimination process initiates with the entire set of S top features, as determined by the $R_{int}(p)$, for model

training. In each iteration of the LOO-RFE, the feature assessed as the least impactful—positioned at the bottom of the integrated ranking—is removed. This iterative elimination progresses from the bottom up, excluding features one at a time based on their order of significance within the integrated ranking, until only two features remain. Mathematically, this can be depicted as:

$$R_{\text{eval}}(p,t) = R_{\text{acc}}(p) - \{F_{\text{bottom}}(t)\} \tag{16}$$

Where:

- $R_{\text{eval}}(p,t)$ stands for the feature set relevant to patient p at iteration t.
- $F_{\text{bottom}}(t)$ signifies the bottom-ranked or least impactful feature during the t^{th} iteration.

As the iterative elimination progresses, the model's performance is assessed at each step. Once all features have been evaluated for all patients, the outcomes are averaged, providing a holistic LOO-RFE evaluation. This aggregated evaluation is defined as:

$$E_{\text{LOO-RFE}} = \frac{1}{|P|}\sum_{p=1}^{|P|}\text{Performance}(R_{\text{eval}}(p,t)) \tag{17}$$

Where:

- $|P|$ is the total count of patients.
- The "Performance" function quantifies the model's performance using the designated feature set.

Our decision to employ the Leave-One-Out Recursive Feature Elimination (LOO-RFE) methodology was strategically informed by the constrained sample size of our dataset. A conventional three-way data split would significantly diminish the effectiveness of feature selection due to the reduced data volume. The central objective of LOO-RFE within our study is to rigorously evaluate the performance of the unified feature ranking, denoted as R_u, under the challenging condition of having no separate validation dataset available. This approach facilitates an accurate estimation of the feature selection's efficacy in a manner that remains unbiased and robust against overfitting, which is critical in studies with limited data samples. For a detailed exposition of the rationale behind choosing LOO-RFE, including its methodological underpinnings and its advantage in ensuring the integrity of our findings, we encourage readers to consult the supplementary material provided. This expanded justification underscores our methodological rigor and the adaptability of LOO-RFE in addressing the unique challenges presented by our dataset, thus ensuring the reliability of our feature selection process.

Results & Discussion

In this section, we conduct a detailed examination of the SpectraFlow pipeline—a comprehensive collection of algorithms fine-tuned for precise feature selection—highlighting its potential to uncover findings of clinical significance.

We start by assessing the pivotal functions of Denoising and PCA-Binning in enhancing the pipeline's precision in identifying metabolically significant features from the intricacies of NMR spectra. This phase of analysis not only illustrates the value these preprocessing steps add to the overall model accuracy but also prepares the ground for a deeper investigation into the spectral regions SpectraFlow prioritises.

As we proceed, our focus shifts to a thorough validation of these pivotal spectral regions. This step confirms their significance as markers with potential relevance to Vasoplegia, ensuring that the identified features not only boost the performance of the predictive model but are also meaningful in a clinical context. By systematically validating both the preprocessing mechanisms and the significance of the resultant feature selection, we aim to demonstrate SpectraFlow's effectiveness in pinpointing clinically relevant metabolic predictors.

Impact of Denoising and PCA-Binning

This experiment investigates the influence of incorporating Denoising and PCA-Binning into the SpectraFlow analytical process, as detailed in Section 2.6. Our findings, summarised in Table 1, highlight the pivotal role of these preprocessing steps in enhancing the model's overall efficacy and ensuring the selection of noise-free features.

De	PCA-B	AUC	F1	Recall	Feature Selection
✓	✓	0.911	0.767	0.717	Noise-free
✓	-	0.885	0.686	0.615	Some noisy features
-	✓	0.909	0.676	0.585	Mostly noisy features
-	-	0.841	0.605	0.561	Mostly noisy features

Table 1. The impact of Denoising (DE) and PCA-Binning (PCA-B) on SpectraFlow performance. Performance metrics are based on the most optimal models derived from LOO-RFE, with the number of optimal training features ranging from 24 to 26. For the visual representation of the selected spectra regions, refer to the Supplementary material.

The data unequivocally show that integrating both Denoising and PCA-Binning significantly boosts model performance, leading to the selection of clean, noise-free features. Omitting the Denoising step markedly diminishes performance across all evaluated metrics, and negatively impacts the quality of feature selection. This is evidenced by the inclusion of noisy features within the top 50 selected features, as determined through manual inspection (refer to the Supplementary material for details on selected spectra regions for each experimental iteration). Additionally, employing SpectraFlow with only PCA-Binning as a preprocessing step not only reduces the model's performance relative to the baseline but also degrades the quality of feature selection, predominantly picking features from noisy regions of the spectra.

The absence of both PCA-Binning and Denoising from the SpectraFlow pipeline results in the lowest model performance, coupled with poor feature selection quality. This finding underscores the critical nature of the Denoising step in achieving noise-free feature selection. This is attributed to the fact that Denoising prevents noisy features from being amplified during data scaling, ensuring they are not mistaken for genuine features. Moreover, PCA-Binning's role in reducing the dimensionality of the data space is also highlighted. This reduction aids the model in better generalising the true underlying patterns, thereby enhancing predictive performance.

Untargeted: PPM Region Feature Selection

Starting with the original global NMR dataset, which has a resolution of 18,637, we implemented a series of preprocessing steps. Initially, denoising techniques were applied, followed by PCA binning. This process streamlined the dataset, reducing it to 4,095 features. With the modified dataset in place, we further adjusted the hyperparameters of our pipeline to select 50 ppm regions, which correspond to PCA-binned regions of interest. A comprehensive

listing of these regions, along with their selection frequencies, is presented in Figure 2.

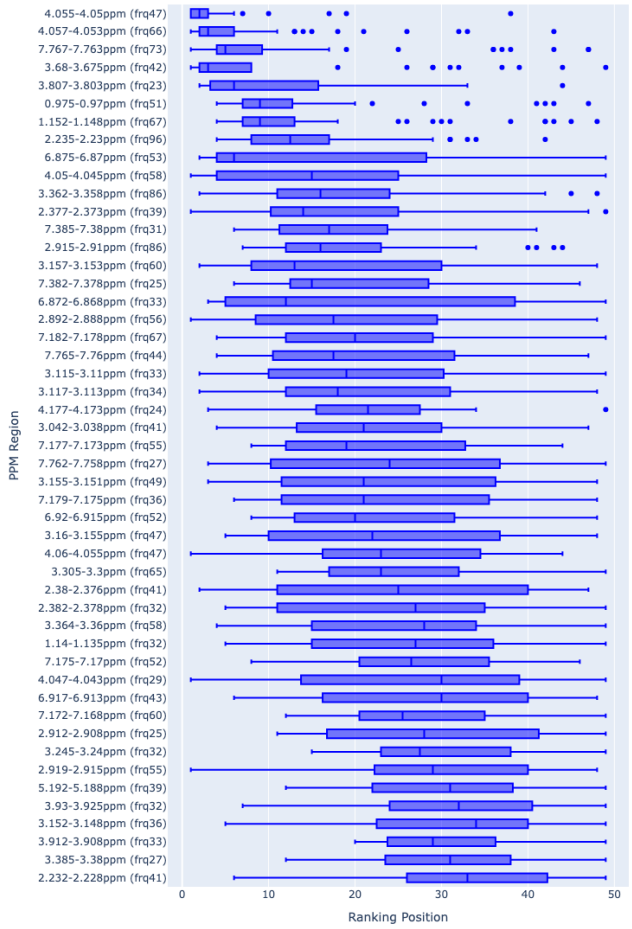


Fig. 2. Unified feature ranking representation of the top 50 ppm regions as selected by SpectraFlow. The ppm regions are arranged vertically with the most significant feature at the top and the least significant at the bottom. The horizontal blue bars indicate the interquartile positions of each ppm region, with the length of the bar representing the spread across respective cross-validation fold rankings. The term 'frq' denotes the count of rankings contributing to the interquartile position of a specific ppm region. For a comprehensive understanding, refer to Section 2.7.

To assess the efficacy of the selected ppm regions, we employed the LOO-RFE evaluation framework. This method helped identify the optimal number of features from the unified ranking that maximised prediction accuracy. Interestingly, performance peaked at 24 top features, wherein we achieved an F1 score of 77% and an ROC AUC of 91%. Table 2 expands on the performance metrics, and the entire LOO-RFE performance is summarised in Figure 3.

A crucial aspect of validating the robustness of our method involves discerning whether the selected ppm regions align with actual metabolic features as opposed to noisy areas. Given the expansive nature of the untargeted dataset, there's an inherent likelihood that some of the selected features might not align with the defined metabolite concentrations, such as those present in the

F1	ROC AUC	P. Prec.	P. Rec.	N. Prec.	N. Rec.
0.767	0.911	0.823	0.717	0.897	0.941

Table 2. Efficacy of the top 24 selected metabolic features derived from the untargeted dataset using SpectraFlow.

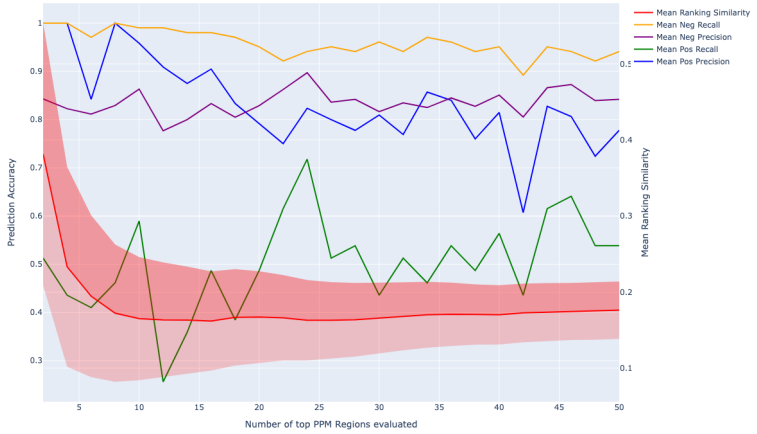


Fig. 3. The figure presents a visual representation of the LOO-RFE evaluation of the top 50 selected PPM regions (Unified ranking) from the targeted dataset by SpectraFlow. The Jaccard similarity quantifies the consistency among feature rankings obtained from each cross-validation fold. The figure presents a visual representation of positive recall, positive precision, negative recall, and negative precision metrics.

targeted dataset. Furthermore, a single metabolic concentration can be represented by multiple ppm peak regions. As such, the expectation for the algorithm to consistently pinpoint all relevant regions associated with a known metabolite is indeed demanding. Consequently, for the top 24 ppm regions, we engaged a spectroscopist to inspect and assign metabolic labels to the selected regions. The assignments can be viewed in Table 3, showcasing that nearly all regions correspond to a valuable metabolic feature.

Metabolic Predictors for Vasoplegia

To enhance our understanding of the metabolic predictors associated with Vasoplegia, we embarked on an analysis of the top 24 ppm regions, as identified by SpectraFlow for their collective predictive efficacy. Each region was subjected to a univariate ANOVA test to independently assess its predictive power for Vasoplegia, revealing several regions with significant p-values below 0.05.

These analyses highlighted Creatinine, Dimethyl-sulfone, Histidine, and 3-hydroxybutyrate as key biomarkers for Vasoplegia. It is important to note that not every region demonstrated statistical significance on an individual basis. Nonetheless, regions without significant p-values contribute to the collective predictive capability of the model, enhancing the differentiation of post-operative Vasoplegia cases and thus should not be disregarded in future work.

Additionally, we expanded our analysis to a targeted dataset, implementing both a partial SpectraFlow approach and PLS-DA, a state-of-the-art method in metabolome analysis. This comprehensive exploration, detailed in the Supplementary Material, consistently identified *Lysine, Phenylalanine, and Tyrosine* as significant metabolites across both targeted and untargeted selection methods. Consequently, we explore the clinical significance of these metabolites, alongside those listed in Table 4, to assess their implications for Vasoplegia, informed by current literature. This integrated approach not only validates the SpectraFlow algorithm's

PPM Range	Identified Metabolite
4.0548-4.0504	Creatinine (-CH2)
4.0574-4.0526	Creatinine (-CH2)
7.7674-7.763	*Tryptophane* (very low concentration)
3.6796-3.6752	Glycerol
3.8072-3.8028	Glucose
0.9746-0.9702	Leucine
1.1523-1.1479	*Propylene glycol*
2.2347-2.2303	Unsuppressed macromolecules
6.8747-6.8703	Tyrosine
4.0498-4.0454	Creatinine (-CH2)
3.3623-3.3579	*Methanol*
2.3772-2.3728	3-hydroxybutyrate
7.3846-7.3802	Phenylalanine
2.9145-2.9101	*N-dimethylglycine*
3.1571-3.1527	Dimethyl-sulfone
7.3824-7.378	Phenylalanine
6.872-6.8676	Tyrosine
2.892-2.8876	Unknown
7.1822-7.1778	Tyrosine
7.7647-7.7603	*Tryptophane* (very low concentration)
3.1148-3.1104	Unknown
3.117-3.1126	Unknown
4.1774-4.173	3-hydroxybutyrate
3.0422-3.0378	Lysine

Table 3. Metabolite assignments for the top 24 ppm regions derived from untargeted analysis with SpectraFlow. The selection of these 24 regions was influenced by the LOO-RFE criterion, optimising prediction performance. The metabolite assignments were performed by a spectroscopist, utilising a thorough examination of the full-resolution spectra and database matching. The italic font underlines the identified metabolic features that were not a part of the targeted dataset (See Supplementary Material for analysis on Targeted dataset).

PPM region	Label	p-value
3.1549-3.1505	Dimethyl-sulfone	0.021146
3.1571-3.1527	Dimethyl-sulfone	0.03644
4.0471-4.0427	Creatinine	0.010679
4.0498-4.0454	Creatinine	0.00040217
4.1774-4.1730	3-hydroxybutyrate	0.033959
7.7647-7.7603	Histidine	0.022911
7.7625-7.7575	Histidine	0.050899

Table 4. Metabolic Predictors with Significant p-values

predictive precision further but also advances our understanding of Vasoplegia's metabolic underpinnings.

Clinical Implications

This investigation into Vasoplegia unveils key metabolic markers from blood samples taken prior to the LVAD implantation in patients, providing a vital glimpse into their metabolic health before undergoing a significant medical intervention. The metabolic features identified herein could serve as potential early indicators or even preventive markers for Vasoplegia.

Creatinine, a byproduct of muscle metabolism, serves as a widely recognised marker for renal function. The elevated levels of creatinine associated with Vasoplegia in our study align with existing literature, hinting at a potential renal dysfunction or an increased susceptibility to Vasoplegia post-surgery (34) (35) (5) (36).

Dimethyl-sulfone, an organic sulfur compound known for its anti-inflammatory properties, is also implicated in renal dysfunction assessment. Although not directly linked to Vasoplegia, its role in renal health suggests a possible connection warranting further exploration (37).

Histidine, an essential amino acid with roles in proton buffering, metal ion chelation, and erythropoiesis, presents another area of interest. While the direct relationship with Vasoplegia remains uncharted, its physiological roles suggest a potential relevance (38) (39).

3-hydroxybutyrate, a ketone body generated during fatty acid oxidation, acts as an energy source in carbohydrate deficit conditions. Its identification may indicate a metabolic shift towards fat oxidation in Vasoplegia patients, especially under stress or catabolic states (40) (41) (42) (43).

Lysine, essential for protein synthesis, collagen formation, and fatty acid metabolism, may have an indirect correlation with Vasoplegia through its involvement in cardiovascular and renal health, although a direct correlation remains to be established (44).

Phenylalanine, pivotal in neurotransmitter synthesis, has its dysregulated catabolism implicated in myocardial senescence, hinting at possible cardiovascular implications relevant to Vasoplegia (45).

Tyrosine, a precursor to neurotransmitters like dopamine and norepinephrine, could influence vascular tone and blood pressure regulation—core components in Vasoplegia pathology (46) (47) (48).

The interplay among these metabolites primarily highlights the renal, cardiovascular, and metabolic health dimensions. The renal dysfunction, as indicated by creatinine and dimethyl-sulfone, alongside the cardiovascular implications from phenylalanine and lysine, may intersect with Vasoplegia's pathophysiology. Moreover, the metabolic shifts potentially signaled by 3-hydroxybutyrate could reflect underlying stress or catabolic states in Vasoplegia patients. This complex metabolic landscape underpins Vasoplegia, advocating for a comprehensive approach in future research to decode Vasoplegia's aetiology and devise prophylactic or therapeutic strategies.

Conclusion

In our pursuit to identify metabolic biomarkers predictive of Vasoplegia syndrome, we initiated our research by developing SpectraFlow. This algorithm is specifically designed to address the inherent challenges of untargeted NMR datasets, namely noise, complex intercorrelation, and the curse of dimensionality. Recognising the essentiality of overcoming these challenges for accurate feature selection, we structured our solution into a systematic pipeline. Through rigorous experiments, we demonstrated not only the robustness and efficacy of our method in terms of predictive performance but also its ability to reliably identify noise-free valid ppm regions that correlate with authentic metabolic features upon spectroscopic verification. Based on our comprehensive analysis, we have identified a set of metabolic features that serve as potential biomarkers. We further discuss the clinical implications of these identified metabolic features and suggest future research directions to extend this work.

Author contributions statement

A.W. was the principal researcher, leading the code development, and authored the manuscript. P.G.T. enhanced the study with spectroscopic analysis of the selected features. C.J.S. generated

NMR datasets from blood samples. E.E.C.W. provided pre-operative blood samples and post-operative class labels, essential for the project's clinical accuracy. Z.C. played a pivotal role in securing funding and additional support. N.M. was the project's driving force, initiating and overseeing its progress. L.L. directed the machine learning research components. All authors reviewed the manuscript.

Acknowledgments

This work was supported by LifeArc, a UK-registered charity number 1015243 (grant number not applicable).

Competing interests

No competing interest is declared.

References

1. Argenziano et al. "Management of vasodilatory shock after cardiac surgery: identification of predisposing factors and use of a novel pressor agent." *The Journal of thoracic and cardiovascular surgery*, 116(6):973–980, 1998.
2. Levin et al. "Methylene blue reduces mortality and morbidity in vasoplegic patients after cardiac surgery." *The Annals of thoracic surgery*, 77(2):496–499, 2004.
3. Omar et al. "Cardiac vasoplegia syndrome: pathophysiology, risk factors and treatment." *The American journal of the medical sciences*, 349(1):80–88, 2015.
4. Kolenbrander et al. "Predicting vasoplegia after continuous flow left ventricular assist device implantation, using a newly developed prediction score." *Journal of Cardiothoracic and Vascular Anesthesia*, 30:S42–S43, 2016.
5. De Waal et al. "Vasoplegia after implantation of a continuous flow left ventricular assist device: incidence, outcomes and predictors." *BMC anesthesiology*, 18(1):1–12, 2018.
6. Alonso et al. "Analytical methods in untargeted metabolomics: state of the art in 2015." *Frontiers in bioengineering and biotechnology*, 3:23, 2015.
7. Patterson et al. "Metabolomics reveals attenuation of the SLC6A20 kidney transporter in nonhuman primate and mouse models of type 2 diabetes mellitus." *Journal of biological chemistry*, 286(22):19511–19522, 2011.
8. Manna et al. "UPLC–MS-based urine metabolomics reveals indole-3-lactic acid and phenyllactic acid as conserved biomarkers for alcohol-induced liver disease in the Ppara-null mouse model." *Journal of proteome research*, 10(9):4120–4133, 2011.
9. Chan et al. "Metabolic profiling of human colorectal cancer using high-resolution magic angle spinning nuclear magnetic resonance (HR-MAS NMR) spectroscopy and gas chromatography mass spectrometry (GC/MS)." *Journal of proteome research*, 8(1):352–361, 2009.
10. Johnson et al. "Challenges and opportunities of metabolomics." *Journal of cellular physiology*, 227(8):2975–2981, 2012.
11. Remeseiro et al. "A review of feature selection methods in medical applications." *Computers in biology and medicine*, 112:103375, 2019.
12. Gorrochategui et al. "Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow." *TrAC Trends in Analytical Chemistry*, 82:425–442, 2016.
13. Schrimpe-Rutledge et al. "Untargeted metabolomics strategies—challenges and emerging directions." *Journal of the American Society for Mass Spectrometry*, 27(12):1897–1905, 2016.
14. Torres et al. "Common problems and artifacts encountered in solution-state NMR experiments." *Concepts in Magnetic Resonance Part A*, 45(2):e21387, 2016.
15. van den Berg et al. "Centering, scaling, and transformations: improving the biological information content of metabolomics data." *BMC genomics*, 7(1):1–15, 2006.
16. Karaman et al. "Preprocessing and pretreatment of metabolomics data for statistical analysis." *Metabolomics: From Fundamentals to Clinical Applications*, :145–161, 2017.
17. Mohammadkhani et al. "Effect of different pretreatment methods on classification of serum samples measured with 1 H-NMR." *No Journal*, :, 2022.
18. Gromski et al. "The influence of scaling metabolomics data on model classification accuracy." *Metabolomics*, 11(3):684–695, 2015.
19. Daubechies et al. "The wavelet transform, time-frequency localization and signal analysis." *IEEE transactions on information theory*, 36(5):961–1005, 1990.
20. Guo et al. "A review of wavelet analysis and its applications: Challenges and opportunities." *IEEE Access*, 10:58869–58903, 2022.
21. Srivastava et al. "A new wavelet denoising method for selecting decomposition levels and noise thresholds." *IEEE access*, 4:3862–3877, 2016.
22. Ge et al. "Noise reduction of nuclear magnetic resonance (NMR) transversal data using improved wavelet transform and exponentially weighted moving average (EWMA)." *Journal of Magnetic Resonance*, 251:71–83, 2015.
23. Monaretto et al. "Enhancing signal-to-noise ratio and resolution in low-field NMR relaxation measurements using post-acquisition digital filters." *Magnetic Resonance in Chemistry*, 57(9):616–625, 2019.
24. Altenhof et al. "DESPERATE: A Python library for processing and denoising NMR spectra." *Journal of Magnetic Resonance*, 346:107320, 2023.
25. Verleysen et al. "International work-conference on artificial neural networks." *Journal Unknown*, :758–770, 2005.
26. Altman et al. "The curse (s) of dimensionality." *Nat Methods*, 15(6):399–400, 2018.
27. Nagana Gowda et al. "Overview of NMR spectroscopy-based metabolomics: opportunities and challenges." *NMR-Based Metabolomics: Methods and Protocols*, :3–14, 2019.
28. Jolliffe et al. "Principal component analysis: a review and recent developments." *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.
29. Chai et al. "Combination of peak-picking and binning for NMR-based untargeted metabonomics study." *Journal of Magnetic Resonance*, 351:107429, 2023.
30. Remeseiro et al. "A review of feature selection methods in medical applications." *Computers in biology and medicine*, 112:103375, 2019.
31. Pudjihartono et al. "A review of feature selection methods for machine learning-based disease risk prediction." *Frontiers in Bioinformatics*, 2:927312, 2022.
32. Vaswani et al. "Attention is all you need." *Advances in neural information processing systems*, 30:, 2017.
33. Bateni et al. "Sequential Attention for Feature Selection." *arXiv preprint arXiv:2209.14881*, :, 2022.
34. Asleh et al. "Predictors and clinical outcomes of vasoplegia in patients bridged to heart transplantation with continuous-flow left ventricular assist devices." *Journal of the American Heart Association*, 8(22):e013108, 2019.
35. Chan et al. "Characterizing predictors and severity of vasoplegia syndrome after heart transplantation." *The Annals of thoracic surgery*, 105(3):770–777, 2018.
36. van Vessem et al. "Incidence and predictors of vasoplegia after heart failure surgery." *European Journal of Cardio-Thoracic Surgery*, 51(3):532–538, 2017.
37. Ehrich et al. "Serum Myo-Inositol, Dimethyl Sulfone, and Valine in Combination with Creatinine Allow Accurate Assessment of Renal Insufficiency—A Proof of Concept." *Diagnostics*, 11(2):234, 2021.
38. Moro et al. "Histidine: A systematic review on metabolism and physiological effects in human and different animal species." *Nutrients*, 12(5):1414, 2020.
39. Holeček et al. "Histidine in health and disease: metabolism, physiological importance, and use as a supplement." *Nutrients*, 12(3):848, 2020.
40. Wang et al. "3-hydroxybutyrate in the brain: Biosynthesis, function, and disease therapy." *Brain-X*, 1(1):e6, 2023.
41. Mierziak et al. "3-hydroxybutyrate as a metabolite and a signal molecule regulating processes of living organisms." *Biomolecules*, 11(3):402, 2021.
42. Nielsen et al. "Cardiovascular effects of treatment with the ketone body 3-hydroxybutyrate in chronic heart failure patients." *Circulation*, 139(18):2129–2141, 2019.
43. Horton et al. "The failing heart utilizes 3-hydroxybutyrate as a metabolic stress defense." *JCI insight*, 4(4):, 2019.
44. Tan et al. "Integrative physiology of lysine metabolites." *Physiological Genomics*, :, 2023.
45. Czibik et al. "Dysregulated phenylalanine catabolism plays a key role in the trajectory of cardiac aging." *Circulation*, 144(7):559–574, 2021.
46. Jongkees et al. "Effect of tyrosine supplementation on clinical and healthy populations under stress or cognitive demands." *Journal Unknown*, :, .
47. Hase et al. "Behavioral and cognitive effects of tyrosine intake in healthy human adults." *Pharmacology Biochemistry and Behavior*, 133:1–6, 2015.
48. Deijen et al. "Tyrosine improves cognitive performance and reduces blood pressure in cadets after one week of a combat training course." *Brain research bulletin*, 48(2):203–209, 1999.
49. Takis et al. "A Computationally Lightweight Algorithm for Deriving Reliable Metabolite Panel Measurements from 1D 1H

NMR." *Analytical Chemistry*, 93(12):4995–5000, 2021.
50. Takis et al. "SMolESY: An efficient and quantitative alternative to on-instrument macromolecular 1 H-NMR signal suppression." *Chemical Science*, 11(23):6000–6011, 2020.
51. Imperial College London "MRC-NIHR BRC National Phenome Centre."

SpectraFlow: A Novel Feature Selection Framework for Overcoming Challenges in 1D NMR Spectroscopy

Supplementary Material

Adrian Wesek et al.

Experimental Evidence

Untargeted Analysis

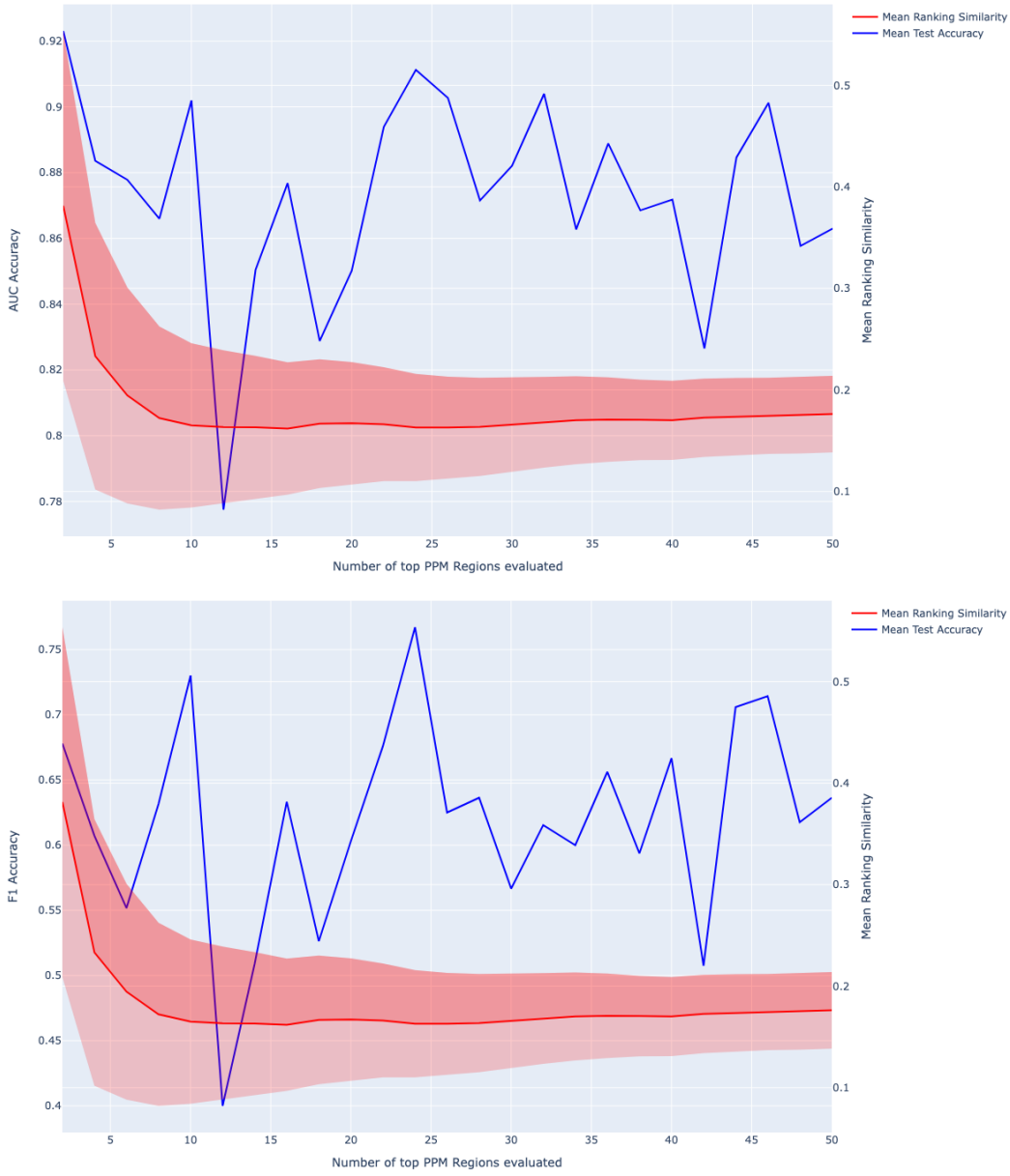


Fig. 4. The figure presents a visual representation of the LOO-RFE (Leave-One-Out Recursive Feature Elimination) evaluation of the top 50 selected PPM regions (Unified ranking) by SpectraFlow. Incorporated within is the average Jaccard similarity, a metric indicating the consistency of feature selection across each cross-valuation fold. Its value ranges from 0 (no shared features) to 1 (identical feature lists), with higher values denoting greater similarity. The top figure corresponds to AUC ROC, while the bottom figure reflects F1 prediction accuracy.

Targeted Analysis

From the CPMG dataset - as detailed in the original publication - a targeted subset was curated, focusing on 28 specific metabolic biomarkers identified using Small Molecule Enhancement Spectroscopy (SMolESY). The methodology behind the biomarkers' quantification and their selection criteria are detailed in (49) (50).

SpectraFlow Approach

In our quest to address the intricacies of untargeted NMR datasets, we developed SpectraFlow. However, for the scope of this analysis, we adapted segments of the pipeline to derive a biomarker ranking for the targeted dataset. This dataset comprised known metabolites, quantified by a spectroscopist. Notably, given that this dataset doesn't display noise or issues of high-dimensionality, we skipped the denoising and PCA binning stages. Instead, we incorporated sequential attention in combination with the LOO-RFE for the purpose of biomarker selection and performance evaluation.

Upon analysis, several biomarkers consistently emerged in the top 8 rankings. These included *Creatinine, Creatine, Isoleucine, Lysine, Phenylalanine, Acetone, Tyrosine, and Formic Acid*. Among them, Creatinine and Creatine were particularly prominent (see Fig. 5).

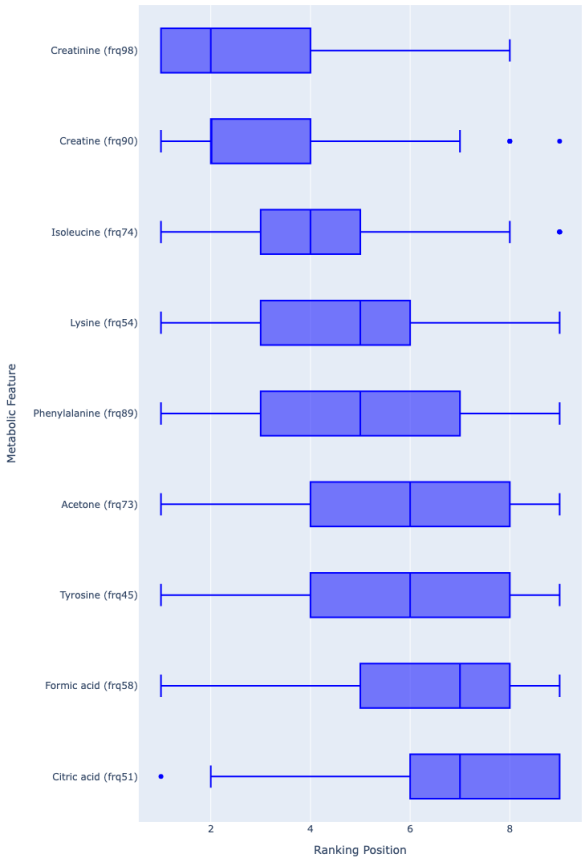


Fig. 5. Unified feature ranking representation of the top 9 metabolic features as selected by the partial implementation of the SpectraFlow model. The features are arranged vertically with the most significant feature at the top and the least significant at the bottom. The horizontal blue bars indicate the interquartile positions of each feature, with the length of the bar representing the spread across respective cross-validation fold rankings. The term 'frq' denotes the count of rankings contributing to the interquartile rank position of the specific metabolic features. For a comprehensive understanding, refer to Section 2.7.

A deeper dive into the results reveals that the model's predictions, when pinpointing positive cases, are robust and dependable (as evidenced by the high positive precision). However, the model has a shortcoming in its limited ability to detect all positive instances, as evidenced by the low positive recall. The commendable ROC AUC score suggests that with recalibration of the decision threshold, there's potential for significant enhancement in recall.

F1	ROC AUC	P. Prec.	P. Rec.	N. Prec.	N. Rec.
0.656	0.872	0.909	0.513	0.84	0.98

Table 5. Efficacy of top 8 selected metabolic features derived from the targeted dataset using SpectraFlow.

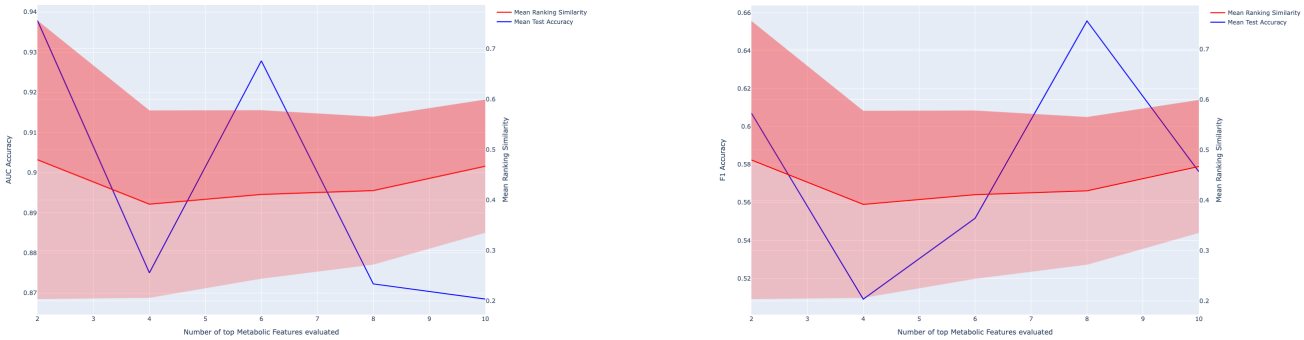


Fig. 6. The figure depicts the LOO-RFE assessment of the top 9 metabolic features (Unified ranking) from the targeted dataset using the SpectraFlow model. The Jaccard similarity quantifies the consistency among feature rankings obtained from each cross-validation fold. The left figure highlights AUC ROC, while the right focuses on F1 prediction accuracy.

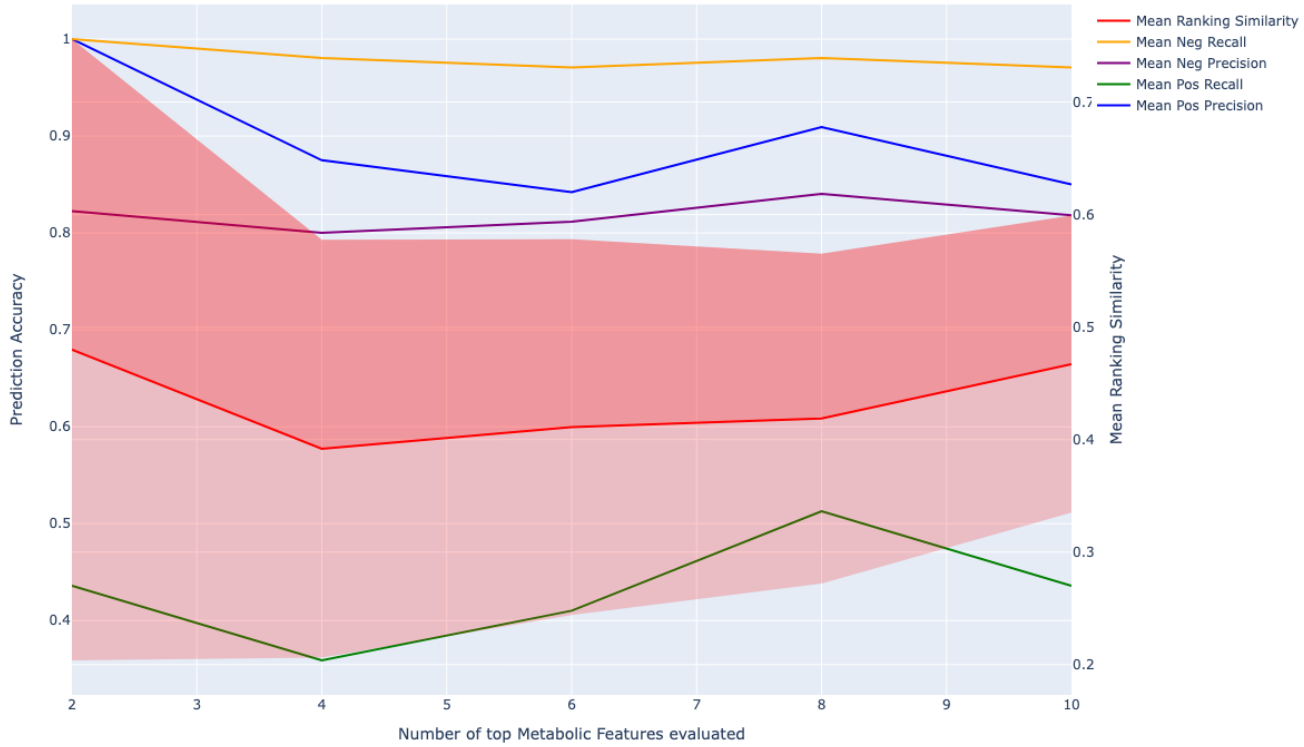


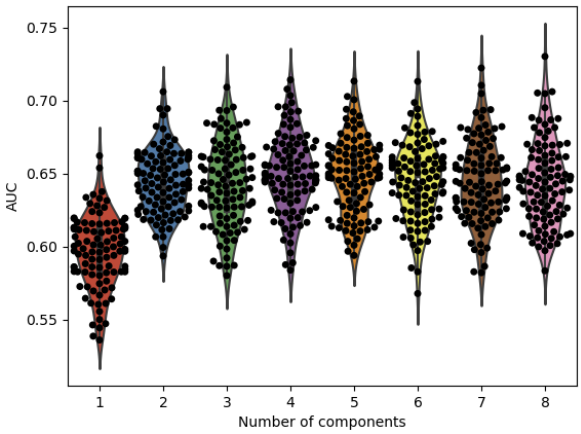
Fig. 7. The figure depicts the LOO-RFE assessment of the top 9 metabolic features (Unified ranking) from the targeted dataset using the SpectraFlow pipeline. The Jaccard similarity quantifies the consistency among feature rankings obtained from each cross-validation fold. The figure presents a visual representation of positive recall, positive precision, negative recall, and negative precision metrics.

To provide context and a comparative perspective, we juxtaposed the results obtained from our methodology against those derived from the widely-adopted Partial Least Squares Discriminant Analysis (PLS-DA).

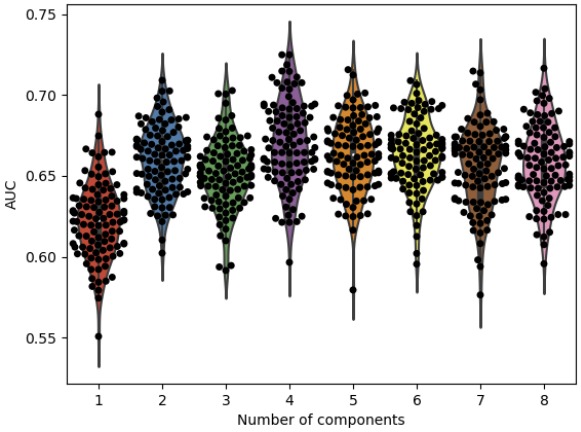
PLS-DA approach

The PLS-DA analysis commenced with the objective of determining the optimal number of components that would maximise data variance and AUC. This evaluation was carried out employing a 10-fold cross-validation, reiterated 10 times. Visualisation of the subjects was facilitated through PCA, enabling us to pinpoint and subsequently exclude outliers. Following the exclusion of outliers, we reiterated the

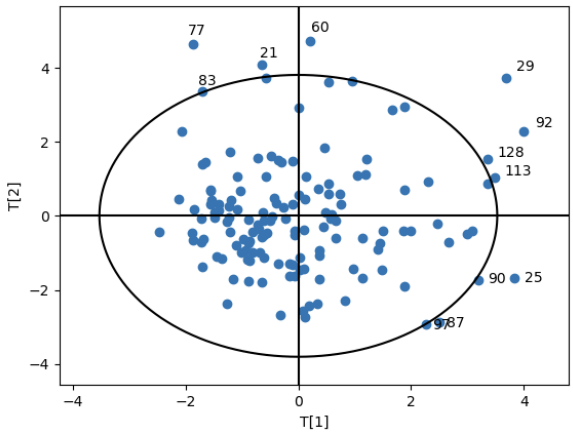
search for the optimal components. A manual experimentation phase followed, during which various outlier groups were eliminated, all with the aim of identifying the optimal configuration. All analyses were standardised using unit variance scaling.



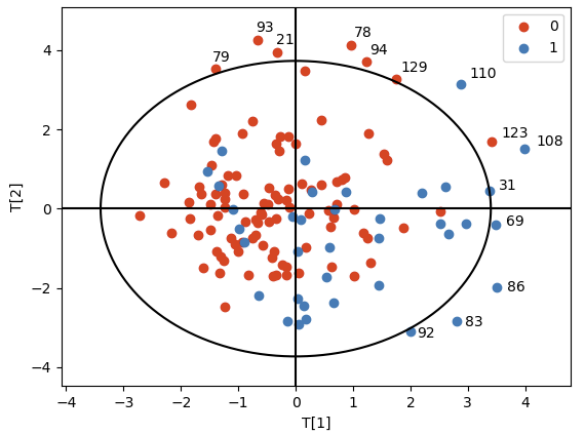
(a) ROC AUC performance of the initial optimal component search.



(b) ROC AUC performance of the optimal component search after outlier exclusion.



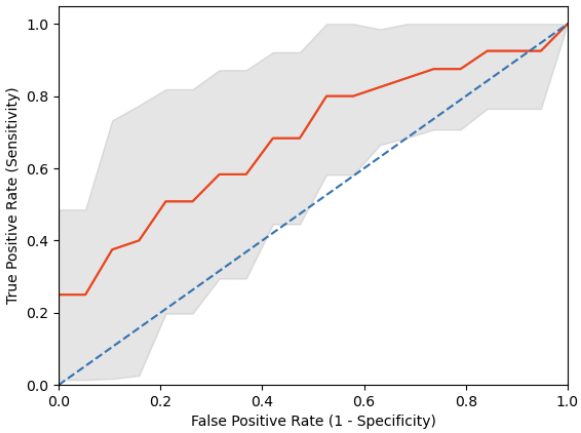
(c) Initial PCA for outlier identification.



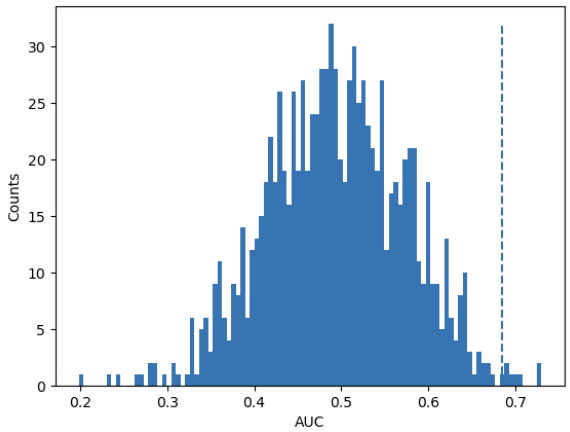
(d) PCA after excluding subjects referenced as 25, 29, 60, 77, 92.

Fig. 8. The figure illustrates the initial steps in PLS-DA, emphasising the determination of the optimal number of components and the exclusion of outliers.

Upon the finalisation of outliers' removal and having settled on the optimal components, PLS was executed, yielding a mean AUC of 68.4%. To fortify the validity of this performance metric, we undertook 1000 permutation tests. Within these tests, class labels were randomised, followed by the retraining of the PLS model using each permutation and the consequent computation of its performance. This procedure facilitated the calculation of the p-value linked to achieving the aforementioned 68.4% score with randomised labels. A p-value of 0.008 was derived, cementing the statistical significance of the score.



(a) ROC Curve for the Optimal PLS-DA Model. Mean ROC AUC across cross-validation is noted at 0.6842, though it may not be distinctly evident in the illustration.



(b) Distribution of ROC AUC scores over 1,000 repetitions with randomised class labels. The dotted vertical line represents the accuracy from the adjacent figure. A p-value of 0.008 suggests a significant difference in our model's AUC from random expectations.

Fig. 9. Performance of the PLS-DA model complemented by validation through the permutation randomisation test.

Post the PLS model training and validation, we proceeded to extract metabolites of statistical significance, relying on regression coefficients in tandem with the weights of the first component. Features that surpassed the significance threshold, set at a p-value of 10%, encompassed *Creatine, Creatinine, Acetone, Phenylalanine, and Lysine*. These are summarised in Table 6

Parameter	**p-value threshold**	**Statistically Significant Features**
Regression Coefficient	0.025	Creatinine, Acetone, Phenylalanine
Regression Coefficient	0.05	Creatinine, Acetone, Phenylalanine, Lysine
Regression Coefficient	0.1	Creatine, Creatinine, Acetone, Phenylalanine, Lysine
Regression Coefficient	0.15	Isoleucine, Valine, Glycine, Creatine, Creatinine, Dimethyl sulfone
Regression Coefficient	0.2	Isoleucine, Valine, Glycine, Creatine, Creatinine, Formic acid, ...
		... Dimethyl sulfone, Acetone, Phenylalanine, Lysine, Succinic acid
Weights	0.025	Creatinine
Weights	0.05	Creatinine, Phenylalanine, Lysine
Weights	0.1	Creatinine, Phenylalanine, Lysine
Weights	0.15	Creatinine, Phenylalanine, Lysine, Dimethylamine
Weights	0.2	Creatine, Creatinine, Citric acid, Phenylalanine, Lysine, Dimethylamine

Table 6. Metabolic features associated with Vasoplegia based on PLS-DA analysis

Comparison of Feature Selection Across Models

The predictive model derived from SpectraFlow has demonstrated superior performance. However, comparing it directly with PLS-DA is challenging due to the fundamentally different spectrum of techniques each employs. Despite this, we observed that features consistently ranked at the top by both methods bear a strong resemblance, suggesting they may represent reference metabolites. Consequently, we use these metabolites as a benchmark to evaluate whether SpectraFlow, when applied to a global spectrum, identifies ppm regions that, after labeling, correspond to metabolites showing overlap with those identified by both PLS-DA and SpectraFlow in the targeted dataset.

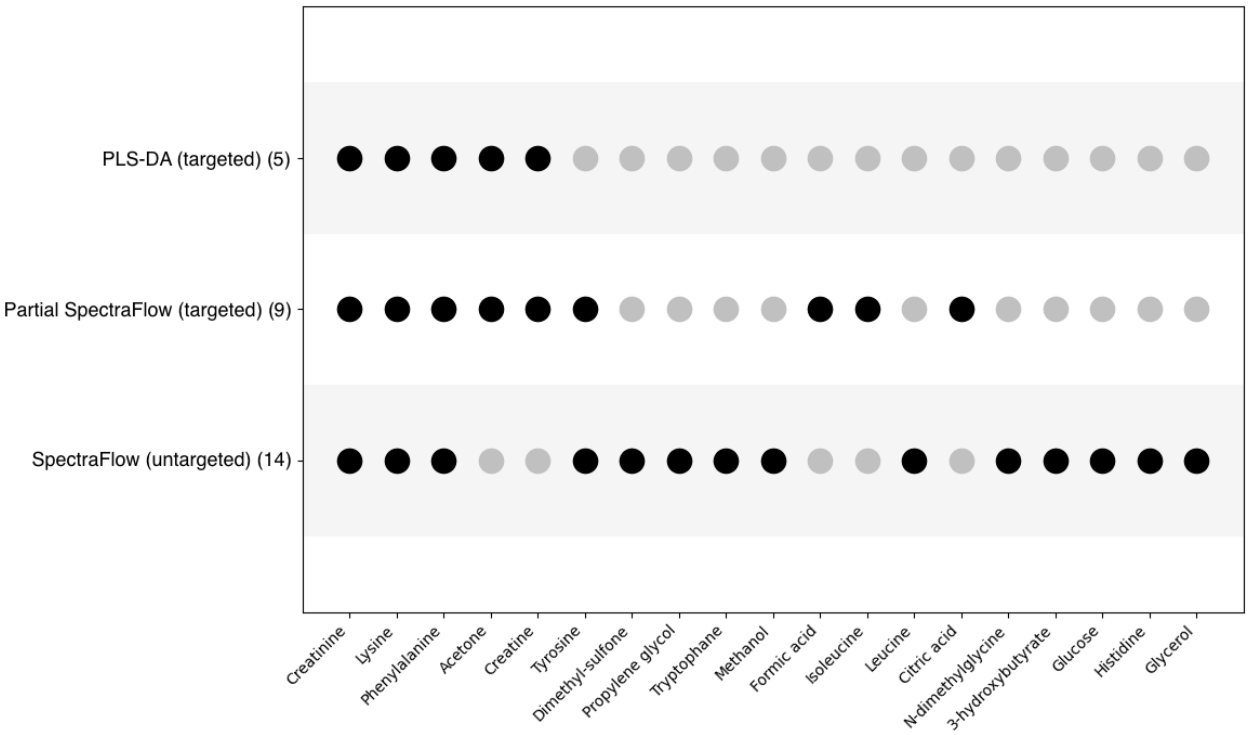


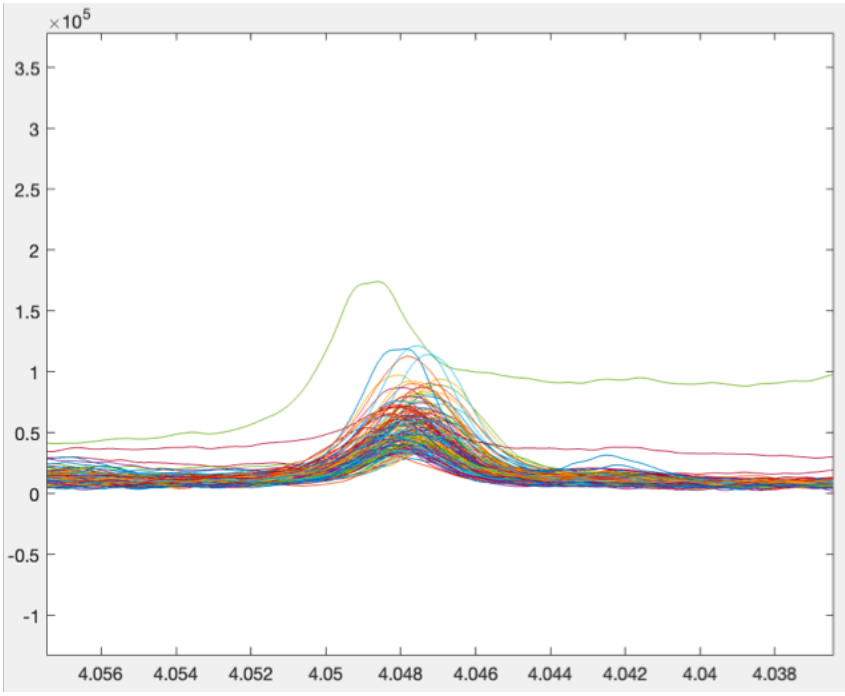
Fig. 10. Summary of metabolic features chosen using various methods. Please note that the features from SpectraFlow relate to specific ppm regions that have been identified and labeled.

Upon in-depth analysis of feature selections from both the targeted and untargeted NMR datasets, we identified significant metabolic overlaps. Among the top selections, metabolites such as *Creatinine, Lysine, Phenylalanine, and Tyrosine* were consistently highlighted in both datasets. These overlapping findings further attest to the robustness of our proposed pipeline. Despite the vastness of the global NMR dataset, with its inherently larger pool of potential features, the pipeline reliably identified regions consistent with the targeted dataset.

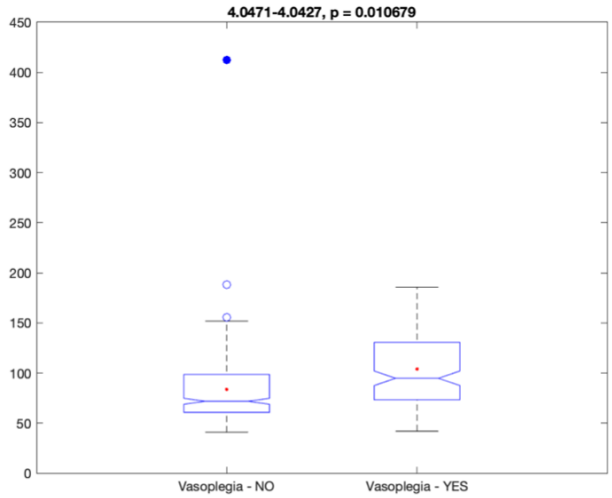
Moreover, the untargeted approach notably identified *Propylene Glycol, Tryptophan, Methanol, and N-dimethylglycine*, which were not included in the targeted dataset. Without the untargeted feature selection approach, these significant compounds would have been overlooked. This highlights the pivotal role of untargeted feature selection in achieving a more comprehensive feature identification.

PPM values with Significant p-values

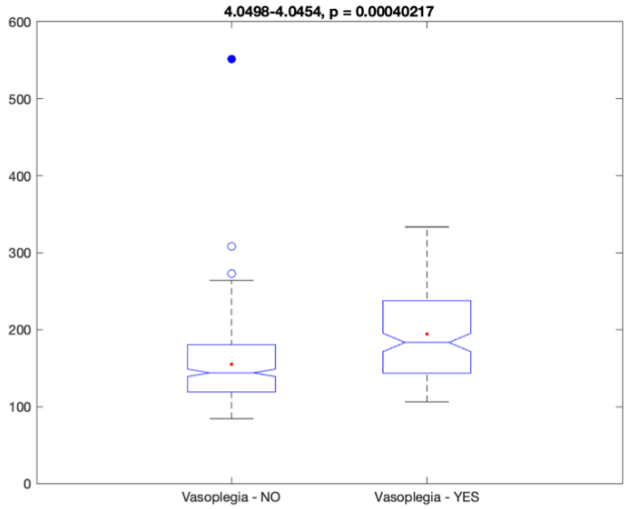
After selecting the top ppm regions, we conducted an ANOVA test on the top 24 regions to identify features that are significant in independently differentiating the Vasoplegia classes. For regions with a p-value less than 0.05, a spectroscopist meticulously examined the high-resolution spectra to quantify these ppm regions as metabolites.



(a) Assigned label: Creatinine.

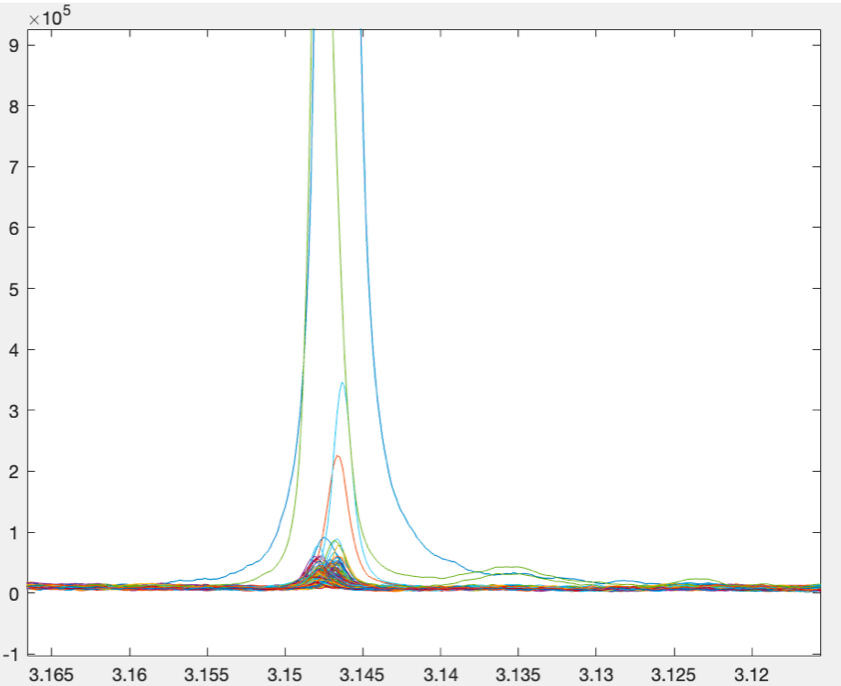


(b) ppm region 4.0471-4.0427, p=0.010679

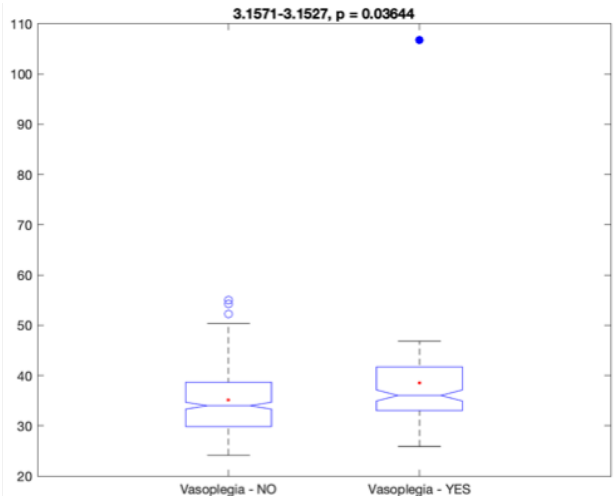


(c) ppm region 4.0498-4.0454, p=0.00040217

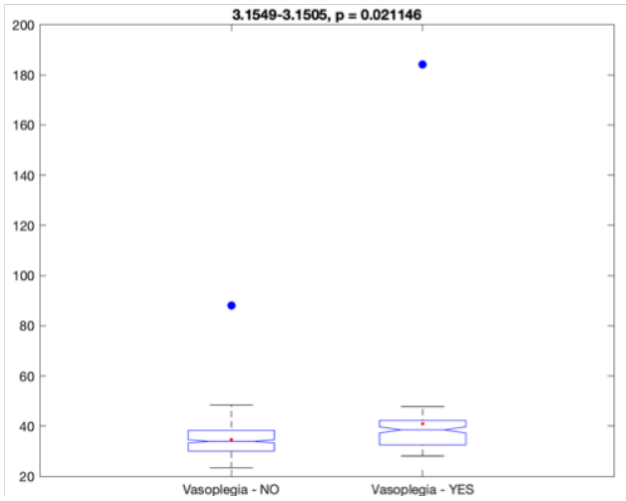
Fig. 11. Highlighted spectra region illustrating variance between subjects with and without Vasoplegia. The selected region of interest is based on feature selection from the untargeted dataset in SpectraFlow. The provided p-values are derived from ANOVA, indicating the independent significance of the region.



(a) Assigned label: Dimethyl-sulfone.

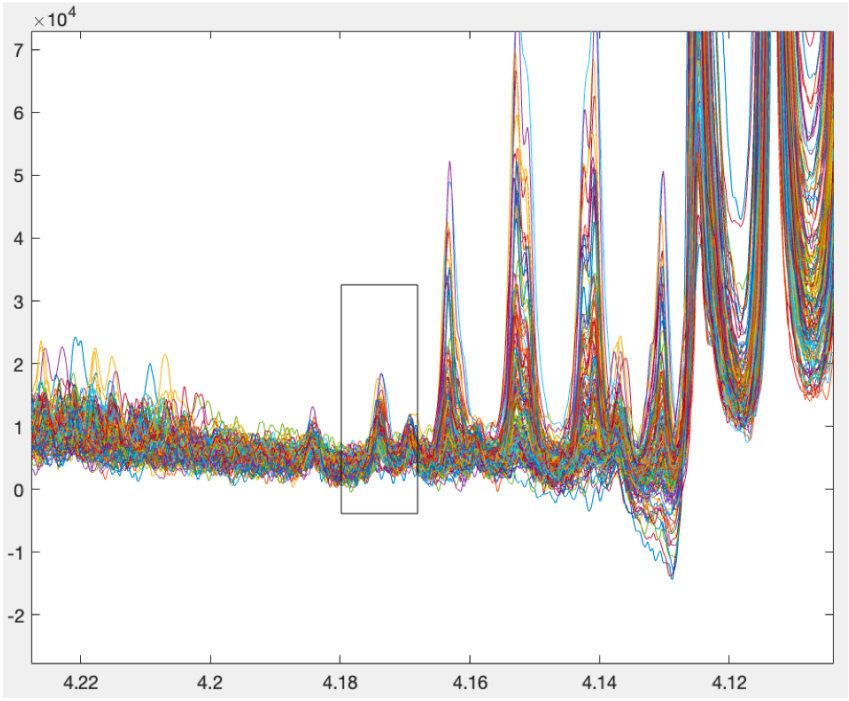


(b) ppm region 3.1571-3.1527, p=0.03644

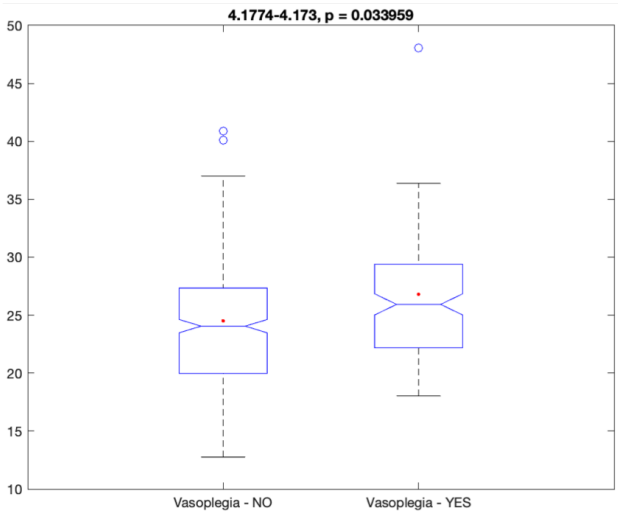


(c) ppm region 3.1549-4.1505, p=0.021146

Fig. 12. Highlighted spectra region illustrating variance between subjects with and without Vasoplegia. The selected region of interest is based on feature selection from the untargeted dataset in SpectraFlow. The provided p-values are derived from ANOVA, indicating the independent significance of the region.

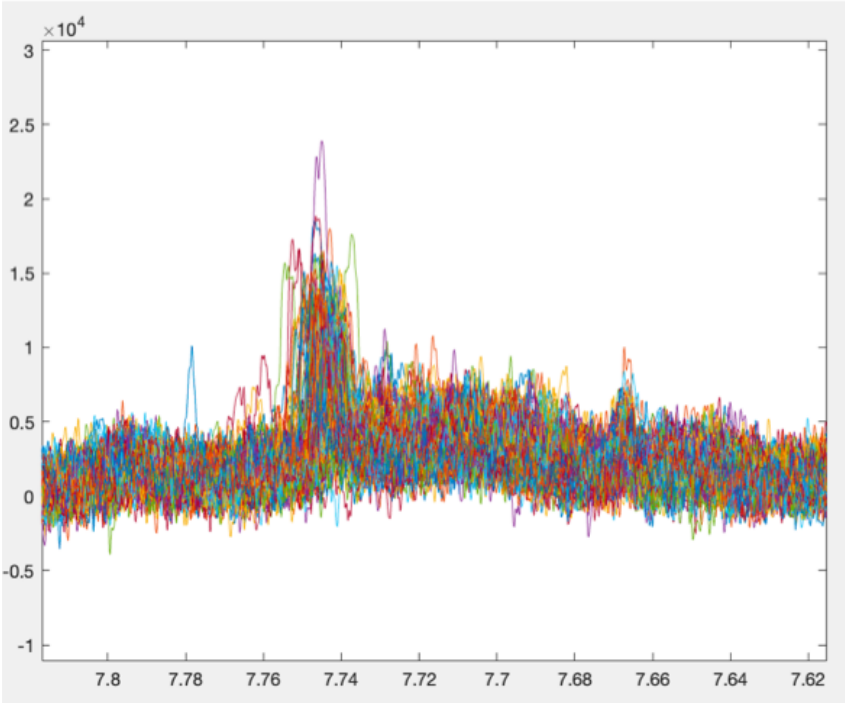


(a) Assigned label: 3-hydroxybutyrate.

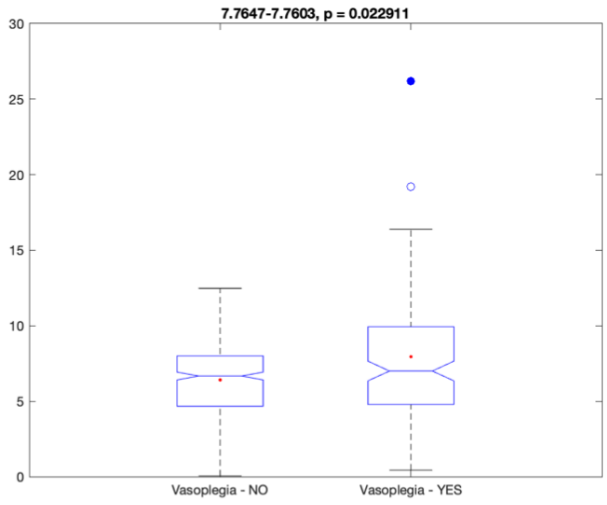


(b) ppm region 4.1774-4.173, p=0.033959

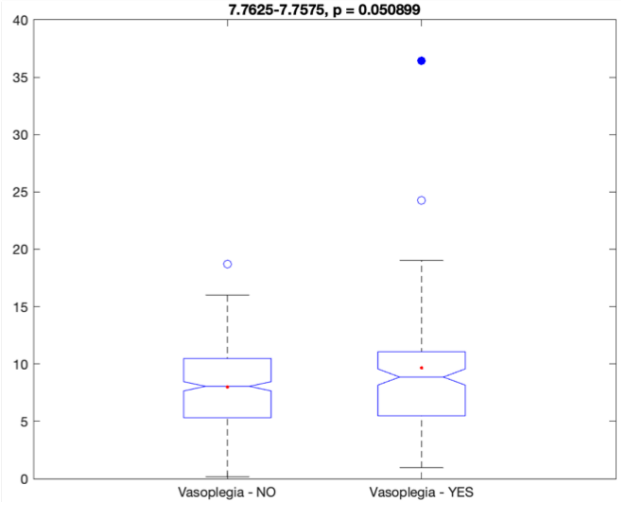
Fig. 13. Highlighted spectra region illustrating variance between subjects with and without Vasoplegia. The selected region of interest is based on feature selection from the untargeted dataset in SpectraFlow. The provided p-values are derived from ANOVA, indicating the independent significance of the region.



(a) Assigned label: Histidine.



(b) ppm region 7.7647-7.7603, p=0.022911



(c) ppm region 7.7625-7.7575, p=0.050899

Fig. 14. Highlighted spectra region illustrating variance between subjects with and without Vasoplegia. The selected region of interest is based on feature selection from the untargeted dataset in SpectraFlow. The provided p-values are derived from ANOVA, indicating the independent significance of the region.

Impact of Denoising and PCA-Binning - extension

In this supplement, we present additional details pertaining to the experiments described in the main paper, specifically under the section "Impact of Denoising and PCA-Binning."

We display the features chosen by the model in each respective configuration. This underscores the significance of denoising and PCA-binning as indispensable preprocessing steps, enhancing the model's ability to discern and select pertinent, noise-free features.

Figure 15 demonstrates SpectraFlow's application on data subjected to denoising followed by PCA-Binning transformation. This preprocessing sequence is crucial, as it ensures the feature selection process sidesteps noisy artifacts, yielding a selection that is entirely noise-free.

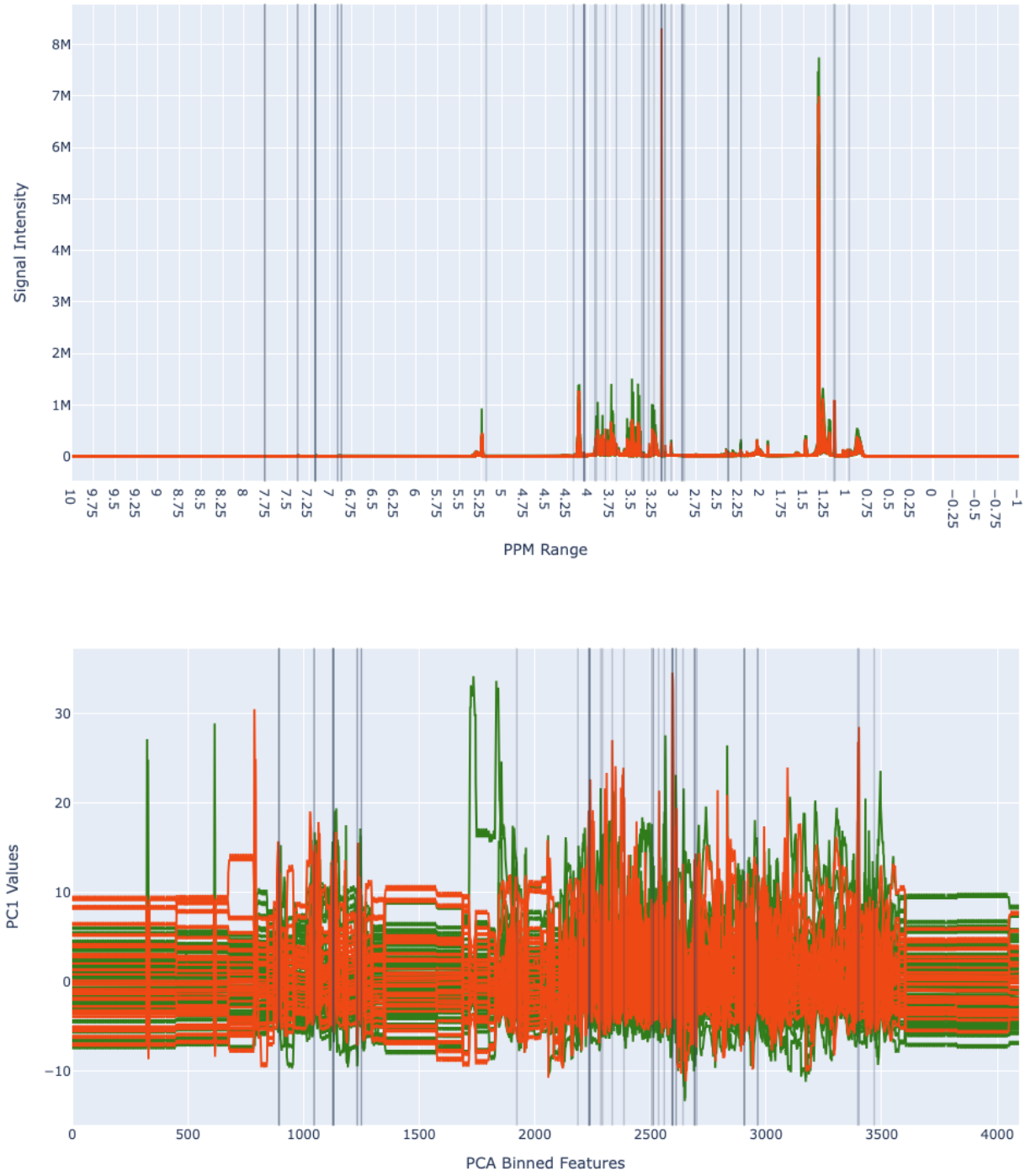


Fig. 15. The diagram showcases the denoised spectra alongside its PCA-binned transformation. The vertical lines inidcate the ppm regions ranked as top 50 by SpectraFlow. The colour differentiation, green and orange, represents Vasoplegia positive and negative patients, respectively.

In contrast, Figure 16 displays the application of SpectraFlow to data that underwent PCA-binning without the denoising step. Observe how the noisy areas become amplified and compete with genuine metabolic concentrations, resulting in a feature selection dominated by noisy features.

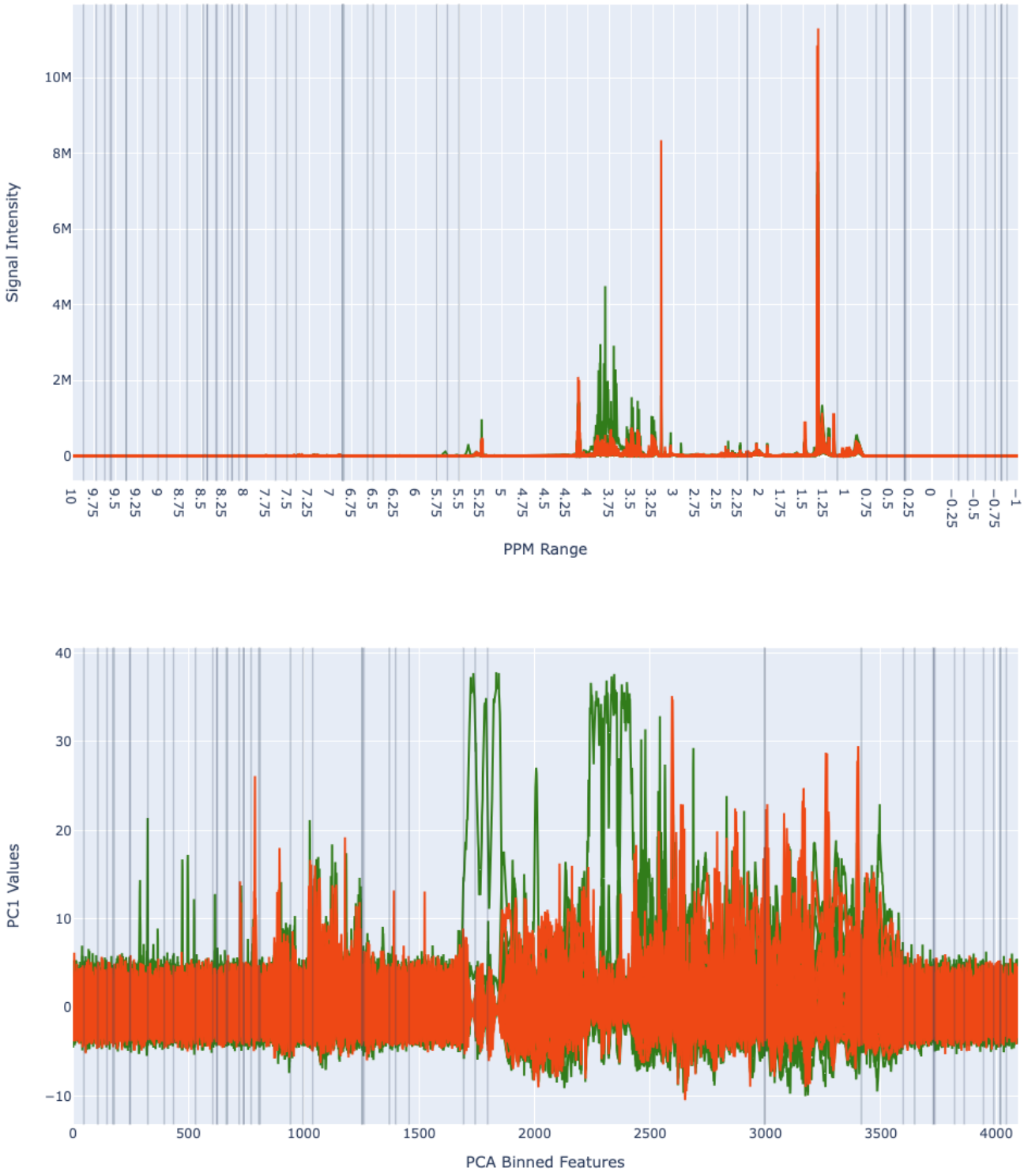
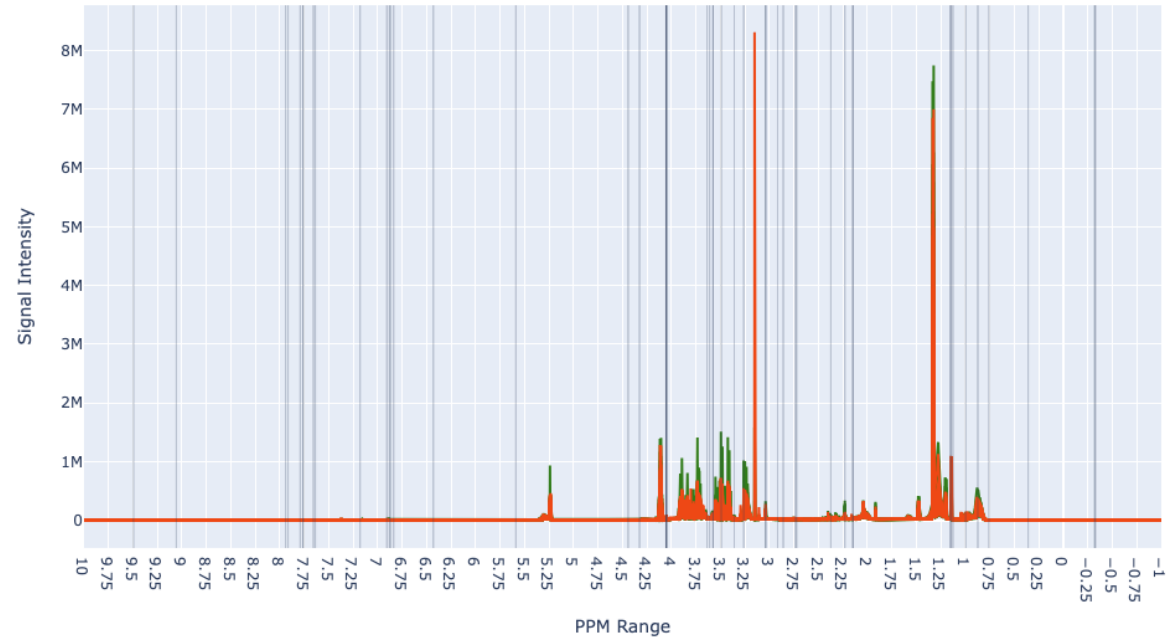


Fig. 16. The diagram showcases the data spectra that underwent PCA-binning transformation, without denoising. The vertical lines inidcate the ppm regions ranked as top 50 by SpectraFlow. The colour differentiation, green and orange, represents Vasoplegia positive and negative patients, respectively.

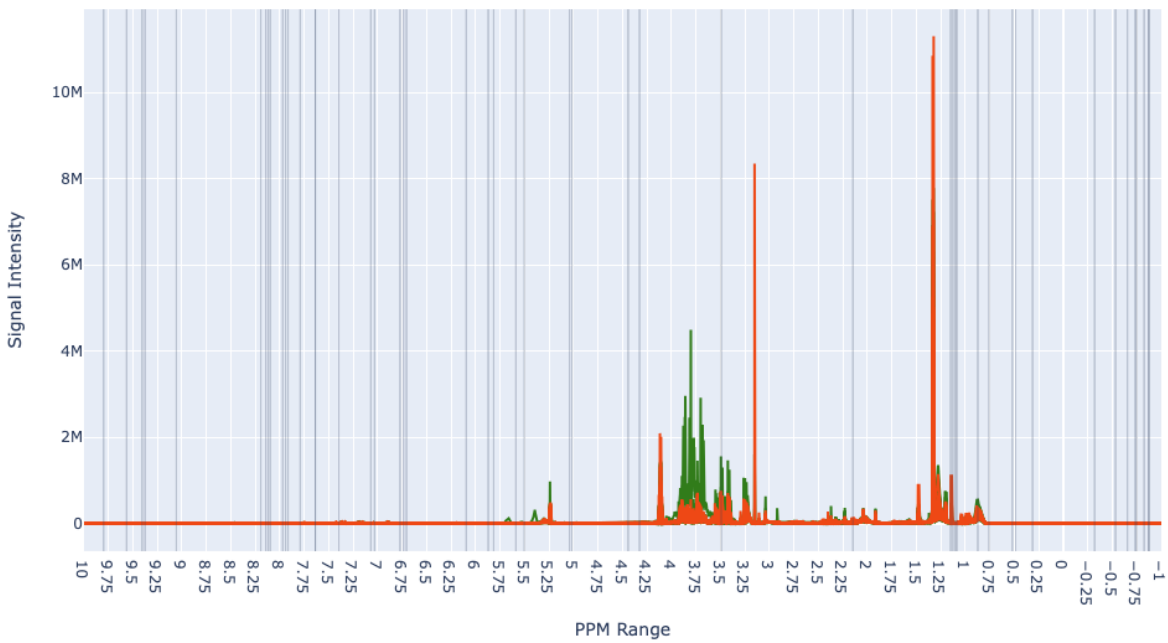
This highlights the efficiency and importance of wavelet denoising for identifying true metabolic features with real clinical significance. It's crucial to select the correct wavelet filter, in our case db1, along with an appropriate noise elimination threshold to remove noisy

features while preserving true metabolic concentrations. These concentrations, though exponentially small compared to other ppm regions, hold significant meaning. Investigating the impact of different wavelet filters, their sizes, and various filtering thresholds is a promising future research area for this method.

Moving on, inspect Figure 17. Feature selection without PCA-Binning generally results in noise-free selection, albeit with a few noisy features. However, the curse of dimensionality, which PCA-Binning addresses, affects the training/evaluation process, leading to lower performance and thus less reliable feature selection. Lastly, examine the SpectraFlow application without denoising and PCA-binning. Again, observe how the feature selection focuses entirely on noisy areas of the spectra.



(a) SpectraFlow feature selection without PCA-Binning.



(b) SpectraFlow feature selection without PCA-Binning and Denoising.

Fig. 17. Untargeted datasets (CPMG) spectra. The vertical lines inidcate the ppm regions ranked as top 50 by SpectraFlow. The colour differentiation, green and orange, represents Vasoplegia positive and negative patients, respectively.

Estimating the efficacy of the unified-ranking using LOO-RFE

The Leave-One-Out Recursive Feature Elimination (LOO-RFE) process, in the context of evaluating the unified ranking without directly employing the actual features specified by the unified ranking for each test patient, constitutes a sophisticated methodology that maximises data utility for feature selection while ensuring an unbiased validation framework. This approach is predicated on the generation of patient-specific rankings derived from portions of the dataset that exclude the data of the patient under evaluation. By doing so, it adheres to the principle of unbiased evaluation, as each patient-specific ranking is constructed from data entirely independent of the patient being tested, thus avoiding any potential bias that could arise from data leakage.

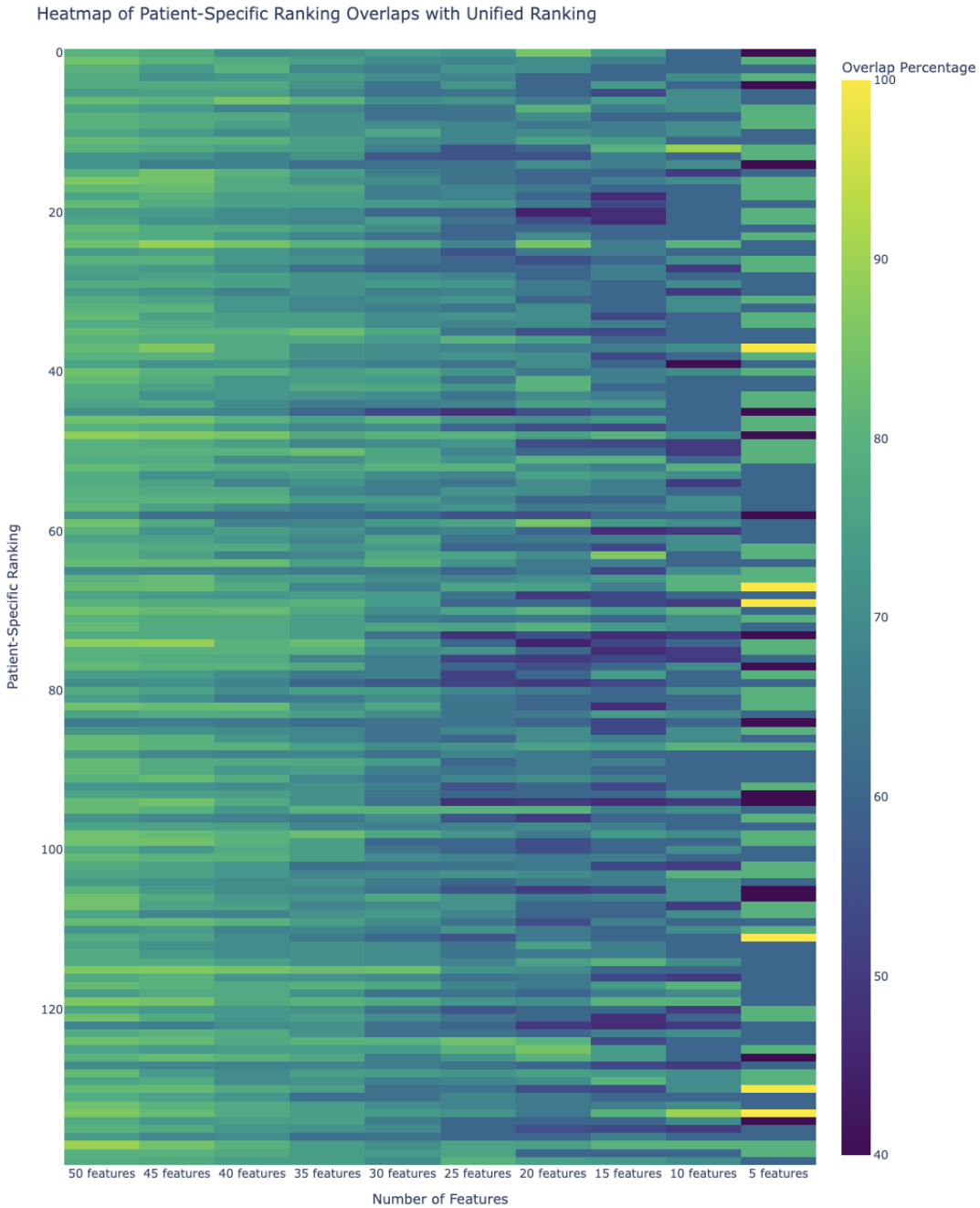


Fig. 18. Heatmap of overlap percentages between patient-specific and unified feature rankings for varying counts of top features. The x-axis represents the decreasing number of features considered, from 50 to 5 in steps of 5, while the y-axis corresponds to individual patient rankings. Colour intensity reflects the overlap percentage, with lighter shades indicating a higher degree of commonality.

The justification for this methodology's effectiveness in estimating the performance of the unified ranking, despite not using the unified ranking directly for each patient's evaluation, lies in the aggregate nature of the LOO-RFE process. Each patient-specific ranking is a subset

reflection of the broader unified ranking, conditioned on the exclusion of the patient's own data. By evaluating the model's performance across all patient-specific rankings and then aggregating these performances, the methodology indirectly assesses the robustness and relevance of the unified ranking. This aggregated performance metric offers a holistic view of the unified ranking's efficacy across the entire dataset, under the premise that a high-performing unified ranking should consistently produce high-performing patient-specific rankings when the corresponding patient data are excluded.

Furthermore, this process effectively leverages the entire dataset for feature selection, maximising the use of available data while maintaining the integrity of the evaluation process. The recursive feature elimination aspect, by systematically removing the least impactful features and reassessing performance, provides a dynamic mechanism to hone in on the most relevant features as defined by the unified ranking.

The strength of this approach lies in its ability to navigate the challenges posed by limited sample sizes and the high dimensionality of untargeted metabolomics data. It ensures that every data point contributes to the feature selection process while also serving as an independent validation point, thereby maximising data utility and maintaining a rigorous standard of unbiased evaluation. This dual utility not only enhances the reliability of the feature selection process but also validates the clinical relevance of the features identified by the unified ranking, making it a highly effective and unbiased methodology for evaluating the potential of untargeted metabolomics in personalised medicine.

Inspect Fig. 18. The graphic was generated through a process where initially, patient-specific feature rankings and a unified ranking of the top 50 features were computed from a dataset processed with denoising and PCA-Binning. For each patient-specific ranking, feature overlap percentages with the unified ranking were calculated at intervals, starting with the top 50 features and decreasing by removing the last 5 features at each step, down to the final 5 features. This iterative reduction allowed for the examination of how the concordance between patient-specific and unified feature selections evolved as fewer features were considered. The resultant overlap percentages were organised into a matrix, with rows representing individual patients and columns corresponding to the different feature counts.

The demonstrated overlap in Fig. 18 across all patient-specific rankings underlines the effectiveness of the Leave-One-Out Recursive Feature Elimination (LOO-RFE) methodology in providing a reliable measure of the unified ranking's performance. By illustrating how patient-specific rankings closely mirror the unified ranking, we can infer a strong correlation between the performance of individualised rankings and the overall efficacy of the unified approach. This concordance signifies that the unified ranking possesses a robust capacity to identify features of paramount importance across the diverse spectrum of patient data, thereby affirming its clinical relevance.

Dynamics of Sequential Attention-Guided Feature Selection in MLP Architecture

Sequence Attention

1. **Input Features Layer**:
   - This is the base layer where all the candidate input features (F1, F2, F3, ..., FN) are presented to the model.
2. **Sequential Attention Layer**:
   - Each input feature is connected to a corresponding attention node in this layer.
   - The attention nodes apply a trainable attention mechanism to assess the importance of each feature.
   - The attention scores influence the feature selection process and are updated based on training feedback.
3. **MLP Input Layer**:
   - The outputs of the attention nodes (which include the attention weights applied to the input features) are passed as inputs to the MLP.
   - The MLP's input layer size is the same as the number of features being considered.
4. **MLP Hidden Layer**:
   - The weighted input features are then processed through the MLP's hidden layers, which include non-linear activation functions, batch normalisation, and dropout regularisation.
   - The hidden layers enable the model to learn complex patterns and interactions between the features.
5. **MLP Output Layer**:
   - The top-most layer in the MLP produces the final prediction. For classification tasks, this layer typically has a softmax activation function; for regression tasks, it might have a linear activation.

Training Dynamics and Feature Selection

During training:

1. **Attention Weight Calculation**:
   - The model calculates attention weights for each feature based on their relevance to the output prediction.
   - These weights are not fixed; they evolve as the model learns from the data throughout the training epochs.
2. **Feature Selection**:
   - At specific intervals, a subset of features with the highest attention weights is selected.
   - This selection is influenced by the training progress, where early in training, the model explores the feature space more broadly, and later, it exploits the most informative features.

Post-feature Selection Dynamics

Once a feature is selected:

1. **Attention Weights Reinitialisation**:
   - The attention weights are reset with small random values, to prevent the model from becoming too reliant on the currently selected features and to encourage exploration of other features.
2. **Feature Masking**:
   - The selected features are masked to avoid being re-selected in subsequent rounds, allowing the model to focus on other informative features.
   - This is important for ensuring the diversity of the selected feature set and for preventing the redundancy of features.
3. **Intuition Behind Reinitialisation and Masking**:
   - Reinitialisation and masking help in redistributing the model's attention to potentially informative features that haven't been selected yet.
   - It prevents the model from "fixating" on early selections and helps in balancing exploration and exploitation, which is crucial in identifying a compact yet informative feature subset.

The described process iterates throughout the training epochs, continuously refining the feature subset and the prediction model for optimal performance.

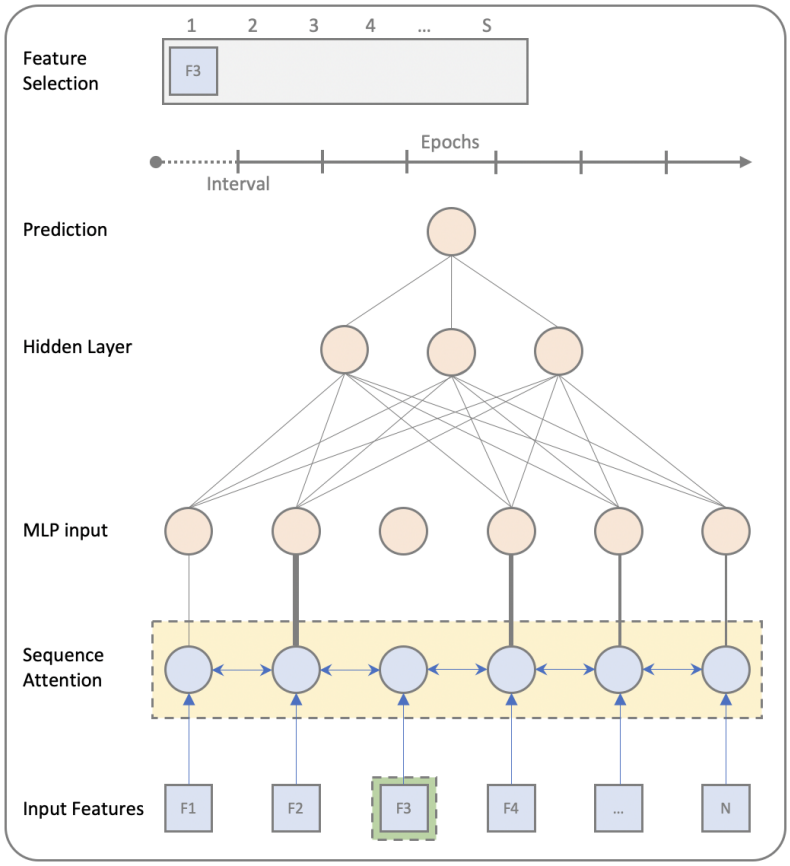


Fig. 19. A stochastic representation of Sequential Attention with MLP for feature selection.

Hyperparameters

The experimentation outlined in the original paper utilised a specific set of hyperparameters to configure the feature selection and learning process. The hyperparameters listed in Table 7 correspond directly to the argument names as they are defined within the code. These settings control various aspects of the feature selection architecture as well as the parameters for the MLP evaluator, determining the behaviour and performance of the model. The descriptions provided alongside each hyperparameter offer insights into their intended effect on the experimental setup.

Table 7. Hyperparameters for Feature Selection and MLP Evaluator

Hyperparameter	Sequential Attention + MLP	MLP Evaluator	Description
Seed	2024	2024	Random seed for reproducibility.
Data Pretreatments	'uv_scaler'	'uv_scaler'	Applies Unite Variance scaling.
Number of Jobs	90	90	Number of jobs to run in parallel.
CV Folds	10	LOO-RFE	Cross-validation folds to test.
CV Repetitions	10	N/A	Number of cross-validation repetitions.
Epochs	800	400	Number of epochs for the training phase.
Number of Selected Features	50	N/A	Number of features to select.
Number of Inputs to Select per Step	1	N/A	Number of features to select per interval.
Learning Rate	0.0002	0.0005	Learning rate for the model.
Number of Hidden Layers	1	1	Number of hidden layer.
Number of Hidden Unites	9	9	Number of nodes in the hidden layer
Decay Steps	250	N/A	Decay steps for the learning rate.
Decay Rate	0.96	N/A	Decay rate for the learning rate.
Alpha	0.01	0.005	Alpha value for L1 regularisation.
Beta	N/A	0.0005	Beta value for L2 regularisation.
Enable Batch Normalisation	True	N/A	Whether to enable batch normalisation.
Batch Size	20	N/A	Batch size for training and evaluation.

Data Preprocessing

PCA-binning

Algorithm 1: PCA Dimension Reduction on Bins of Spectroscopic Data (PCA-Binning)

Input : $file\_path \leftarrow$ path to dataset file (CSV)
Input : $step\_size \leftarrow 0.005$
Input : $overlap \leftarrow 0.0025$
Input : $n\_components \leftarrow 2$
Output: data_with_labels, ppm_feature_names_dict, variance_explained_dict

$df \leftarrow$ ReadData($file\_path$);

$pca\_components\_dict \leftarrow \{\}$;
$ppm\_feature\_names\_dict \leftarrow \{\}$;
$variance\_explained\_dict \leftarrow \{\}$;

$index \leftarrow 0$;
$start\_ppm \leftarrow 10$;
$last\_column\_ppm \leftarrow$ Min($df.columns$);
while $start\_ppm \geq last\_column\_ppm$ **do**
    $end\_ppm \leftarrow start\_ppm - step\_size$;
    $bin\_df \leftarrow df$ within $start\_ppm$ and end_ppm;

    if $bin\_df.columns \geq 8$ **then**
        $bin\_df\_scaled \leftarrow$ Standardize(bin_df);
        $pca\_components \leftarrow$ ApplyPCA(bin_df_scaled,$n\_components$);

        $pca\_components\_dict[index] \leftarrow pca\_components[:,0]$;
        $ppm\_feature\_names\_dict[index] \leftarrow bin\_df.columns$;
        $variance\_explained\_dict[index] \leftarrow$ GetVarianceExplained($pca\_components$);

        $index \leftarrow index + 1$;
        $start\_ppm \leftarrow end\_ppm + overlap$;
    end
    else
        Print("Not enough ppm values within the segment, merging with the next range");
        $start\_ppm \leftarrow start\_ppm - step\_size$;
    end
end
$data\_with\_labels \leftarrow$ Concatenate($columns$,$pca\_components\_dict$);
return $data\_with\_labels$,$ppm\_feature\_names\_dict$,$variance\_explained\_dict$;

Wavelet Thresholding for Denoising NMR Data

NMR spectra often encounter noise from various sources, potentially masking vital chemical shift information. To tackle this, we applied wavelet thresholding, an adept signal processing technique that can distinguish genuine signal features from noise.

Initially, the NMR signal undergoes decomposition into a set of wavelet coefficients using the Discrete Wavelet Transform (DWT):

$$s(t) \rightarrow c(j,k) \tag{18}$$

Here, $s(t)$ symbolises the NMR signal, and $c(j,k)$ denotes the wavelet coefficients at scale j and position k. In our study, the "db1" wavelet was chosen owing to its ability to accurately detect sharp peaks and nuanced details typical in NMR data.

Having the wavelet coefficients, the task then is to identify which of these likely represent noise. To achieve this, we use the Mean Absolute Deviation (MAD) to scale wavelet coefficients to derive a solid estimate of the noise standard deviation, σ:

$$\sigma = \frac{1}{0.6745} \times \mathrm{MAD}(c_{\mathrm{fine}}) \tag{19}$$

The component $1/0.6745$ in Equation 18 serves as a normalisation factor, estimating the standard deviation from MAD when confronted with Gaussian white noise.

With the value of σ, a universal threshold u_{thresh} is calculated as:

$$u_{thresh} = \sigma \times \sqrt{2 \times \ln(n)} \tag{20}$$

Here, n represents the signal length.

Each wavelet coefficient, $c(j,k)$, is then subjected to a process known as soft thresholding:

$$c'(j,k) = \begin{cases} \mathrm{sign}(c(j,k)) \times (|c(j,k)| - u_{thresh}), & \text{if } |c(j,k)| > u_{thresh} \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

Equation 21 elucidates that coefficients with relatively small magnitudes, which are likely indicative of noise, are suppressed. In contrast, significant coefficients, which correspond to genuine features, are not only retained but also undergo a reduction in magnitude.

Post-thresholding, the wavelet coefficients now depict a denoised version of the original NMR signal. By utilising the Inverse Discrete Wavelet Transform (IDWT), the denoised NMR signal $s'(t)$ is reconstructed using the thresholded coefficients $c'(j,k)$.

In summary, wavelet thresholding provides an advanced method for denoising NMR spectra. It accentuates the genuine chemical shift peaks while simultaneously minimising baseline noise. This strategy has significantly improved the clarity of our spectral analysis, paving the way for precise ppm region selections.

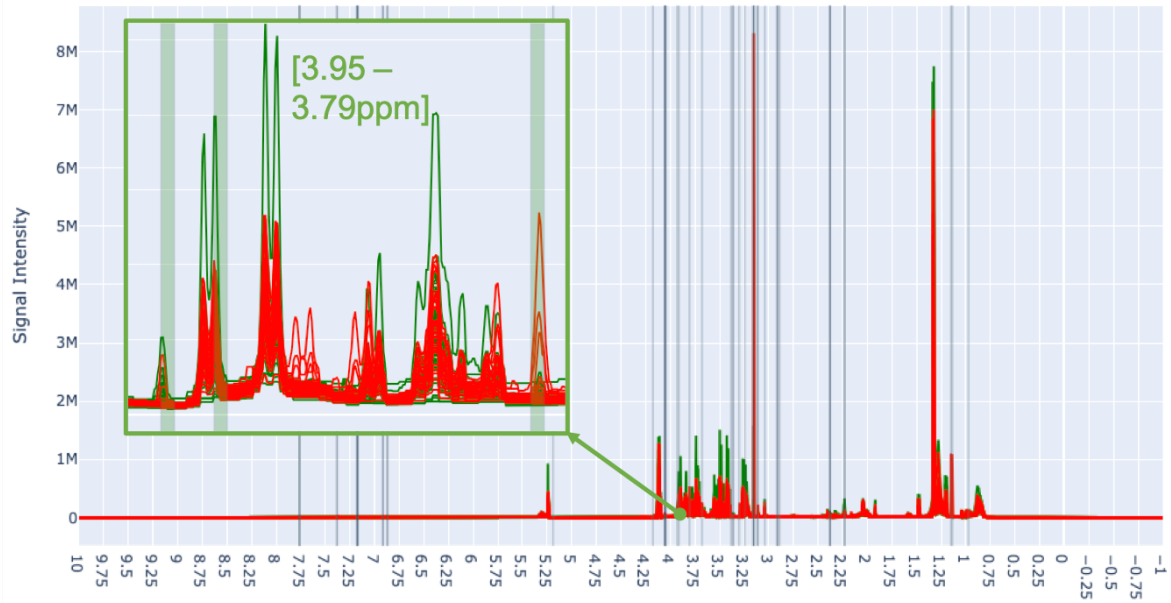
Influence of Wavelet Filter Size on Spectral PPM Region Selection

The size of the wavelet filter, represented as db1, db2, ... db38, significantly influences the transformation of datasets, consequently impacting feature selection. A notable correlation exists between the filter size and the noise introduced into the reconstructed denoised spectra. Preliminary observations reveal that as the filter size increases, the denoised spectra, particularly in peak regions, remain more reminiscent of the spectra prior to denoising. Conversely, a rise in filter size leads to the retention of greater fluctuations in noisy regions. This can detrimentally impact feature selection since these noisy areas may be erroneously interpreted as genuine features.

Refer to Figures 20 and 21 to visually comprehend the influence of wavelet filter sizes on the denoised spectra and the consequential effect on feature selection.

In our analysis, filters in the larger than db8 marginally enhanced prediction accuracy compared to db1. However, employing larger filters occasionally resulted in the incorrect selection of noisy ppm areas for top-ranked ppm regions. Given our primary aim to identify pertinent ppm regions, we favoured a db1 filter that maximally eliminated noise, even if this meant a slight compromise in accuracy. Importantly, the optimal filter choice remains contingent on the specific objective and dataset characteristics. Thus, no universal "correct" choice exists; the decision hinges on the specific application and dataset properties.

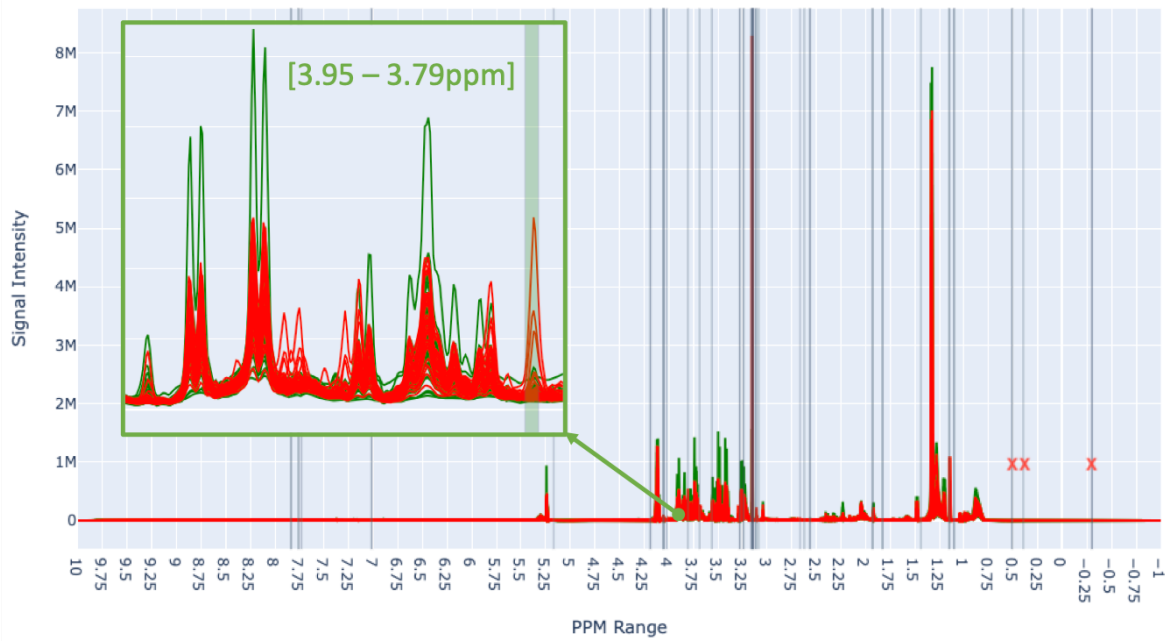
It is pertinent to note that our findings are preliminary. As such, they should be interpreted with caution, and additional experimentation is warranted to robustly validate our observations.



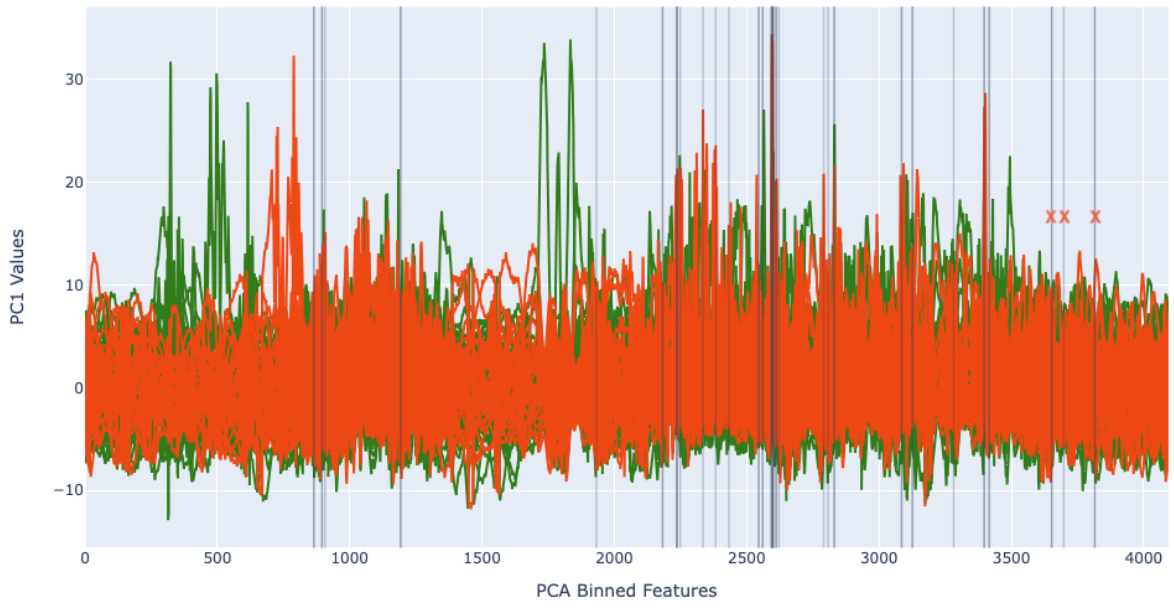
(a) Wavelet db1 Denoised Untargeted NMR Spectra. The range between 3.95 to 3.79 ppm is enlarged, showcasing data transformation in greater detail.

(b) PCA-Bin Transformed Representation of the spectra above.

Fig. 20. The diagram showcases the denoised spectra alongside its PCA-binned transformation. The vertical lines inidcate the ppm regions ranked as top 50 by SpectraFlow. The colour differentiation, green and orange, represents Vasoplegia positive and negative patients, respectively.



(a) Wavelet db20 Denoised Untargeted NMR Spectra. The range between 3.95 to 3.79 ppm is enlarged, showcasing data transformation in greater detail.



(b) PCA-Bin Transformed Representation of the spectra above.

Fig. 21. The diagram showcases the denoised spectra alongside its PCA-binned transformation. The vertical lines inidcate the ppm regions ranked as top 50 by SpectraFlow. The colour differentiation, green and orange, represents Vasoplegia positive and negative patients, respectively. Red crosses underscore noisy features that were included in the top selection.