Project Report

Marco Savastano
Carmine Vardaro

Information Engineering for Digital Medicine
Artificial Intelligence for Omics Data Analysis Course 2025-2026

Contents

1 Introduction 2

1.1 Background Clinico 2

1.2 La Metabolomica Untargeted 2

1.3 Problematiche Aperte 2

1.4 Scopo del Lavoro 2

2 Materials and Methods 2

2.1 Descrizione del Dataset 2

2.2 Quality Assessment and Data Cleaning 3

2.3 Data Pre-Processing 4

2.3.1 Data Normalization 4

2.3.2 Data Transformation 7

2.3.3 Data Scaling 9

2.3.4 Global Assessment 9

2.4 Low-Level Data Fusion 11

2.4.1 Fusion Strategy and Block Scaling 11

2.4.2 SUM-PCA: Exploratory Analysis of Fused Data 11

2.5 Anomaly Detection 13

2.6 Dataset Splitting 16

2.7 Analisi Statistica e Machine Learning 16

2.8 Stack Tecnologico 16

3 Results and Discussion 17

3.1 Performance dei Modelli su Singoli Dataset (ESI+ / ESI-) 17

3.2 Interpretabilità e Potenziali Biomarcatori 18

3.3 Univariate analysis sui biomarcatori 18

4 Conclusions 18

References 18

Abstract

The short abstract (50-80 words) is intended to give the reader an overview of the work.

1 Introduction

1.1 Background Clinico

Breve panoramica sulla patologia (CHD) e importanza di trovare nuovi biomarcatori non invasivi.

1.2 La Metabolomica Untargeted

Perché la LC-MS è la scelta giusta qui (visione olistica del fenotipo).

1.3 Problematiche Aperte

Qui introduci il ”problema” del tuo progetto: la complessità dei dati, la necessità di integrare due modi di ionizzazione (ESI+/ESI-) e la scelta del miglior metodo di preprocessing (non esiste una ”ricetta unica”).

1.4 Scopo del Lavoro

Valutare e ottimizzare un workflow chemiometrico (dal preprocessing alla Data Fusion) per distinguere soggetti Sani vs Patologici e identificare le feature biologicamente rilevanti.

2 Materials and Methods

2.1 Descrizione del Dataset

The dataset is composed of metabolomic data acquired by LC–MS in both positive (ESI+) and negative (ESI-) ionisation mode.

There are no missing values and zeros, due to an value inputation phase already done focused on the replacement with one-fifth of the minimum value recorded in the dataset for that molecule. [1]

The following tables give an overview on the dataset composition, for both ESI+ and ESI- with the corresponding values:

TABLE I: ESI- Dataset Distribution and Characteristics

DESCRIPTION	VALUE
Total Samples	219
Total Features (Metabolites)	52
Class Count: CTRL	107
Class Count: CHD	104
Class Count: QC	8
Samples with suffix ’_00’	28
Samples with suffix ’_01’ (Tech Replicate)	14
Samples without suffix	177
Estimated Unique Biological Samples	205
CTRL - Biological Samples	100
CTRL - Technical Replicates	7
CHD - Biological Samples	97
CHD - Technical Replicates	7
QC - Total Samples	8
Negative Values Present	No

TABLE II: ESI+ Dataset Distribution and Characteristics

DESCRIPTION	VALUE
Total Samples	219
Total Features (Metabolites)	98
Class Count: CTRL	107
Class Count: CHD	104
Class Count: QC	8
Samples with suffix '_00'	28
Samples with suffix '_01' (Tech Replicate)	14
Samples without suffix	177
Estimated Unique Biological Samples	205
CTRL - Biological Samples	100
CTRL - Technical Replicates	7
CHD - Biological Samples	97
CHD - Technical Replicates	7
QC - Total Samples	8
Negative Values Present	No

2.2 Quality Assessment and Data Cleaning

This preliminary phase is crucial to validate the technical quality of the experiment before proceeding with biological interpretation.

To this end, we monitored the behavior of Quality Control (QC) samples (pooled aliquots injected periodically) and technical replicates (samples analyzed in duplicate, denoted with suffixes _00 and _01). The evaluation was performed using Principal Component Analysis (PCA) applied separately to the raw data of Negative (ESI-) and Positive (ESI+) ionization modes. Autoscaling was applied prior to PCA to ensure all metabolites contributed equally to the model, regardless of their absolute intensity.

PCA was employed as an unsupervised exploratory technique to visualize the intrinsic structure of the data variance. To facilitate the inspection of instrumental stability and technical reproducibility, specific sub-plots were generated from the same global PCA model to isolate QC samples and technical replicates. The results for the Negative and Positive ionization modes are presented in Figure 1 and Figure 2, respectively.

a) Variance and Model Structure

The Scree Plots (Fig. 1c, 2c) display the percentage of variance explained by each Principal Component (PC). The Loadings Plots (Fig. 1b, 2b) and Profiles (Fig. 1d-e, 2d-e) provide an overview of the features driving the separation.

b) Instrumental Stability (QC Analysis)

The stability of the LC-MS system was evaluated by isolating the QC samples in the PCA space. As shown in the QC-specific score plots (Fig. 1f, 2f), only the QC samples are visualized to assess their compactness. In the **Negative Ionization mode (ESI-)**, the QCs form a tight, well-defined cluster, indicating high instrumental stability. In the **Positive Ionization mode (ESI+)**, the QCs show a slightly higher dispersion. This behavior is attributed to the inherent characteristics of Hydrophilic Interaction Liquid Chromatography (HILIC). The HILIC separation mechanism relies on a water-enriched layer on the stationary phase, making the partitioning equilibrium more sensitive to minor fluctuations in column conditioning compared to reversed-phase chromatography. However, despite this inherent dispersion, the QC cluster remains distinct from the biological variability.

c) Technical Reproducibility (Replicates Analysis)

Reproducibility was assessed by projecting the technical replicates onto the PCA space (Fig. 1g, 2g). In these plots, non-replicated samples are hidden to highlight the distance between paired measurements (_00 and _01). In both ionization modes, the pairs of replicates are projected in close proximity, often overlapping, confirming that the analytical workflow yields consistent results for the same biological sample.

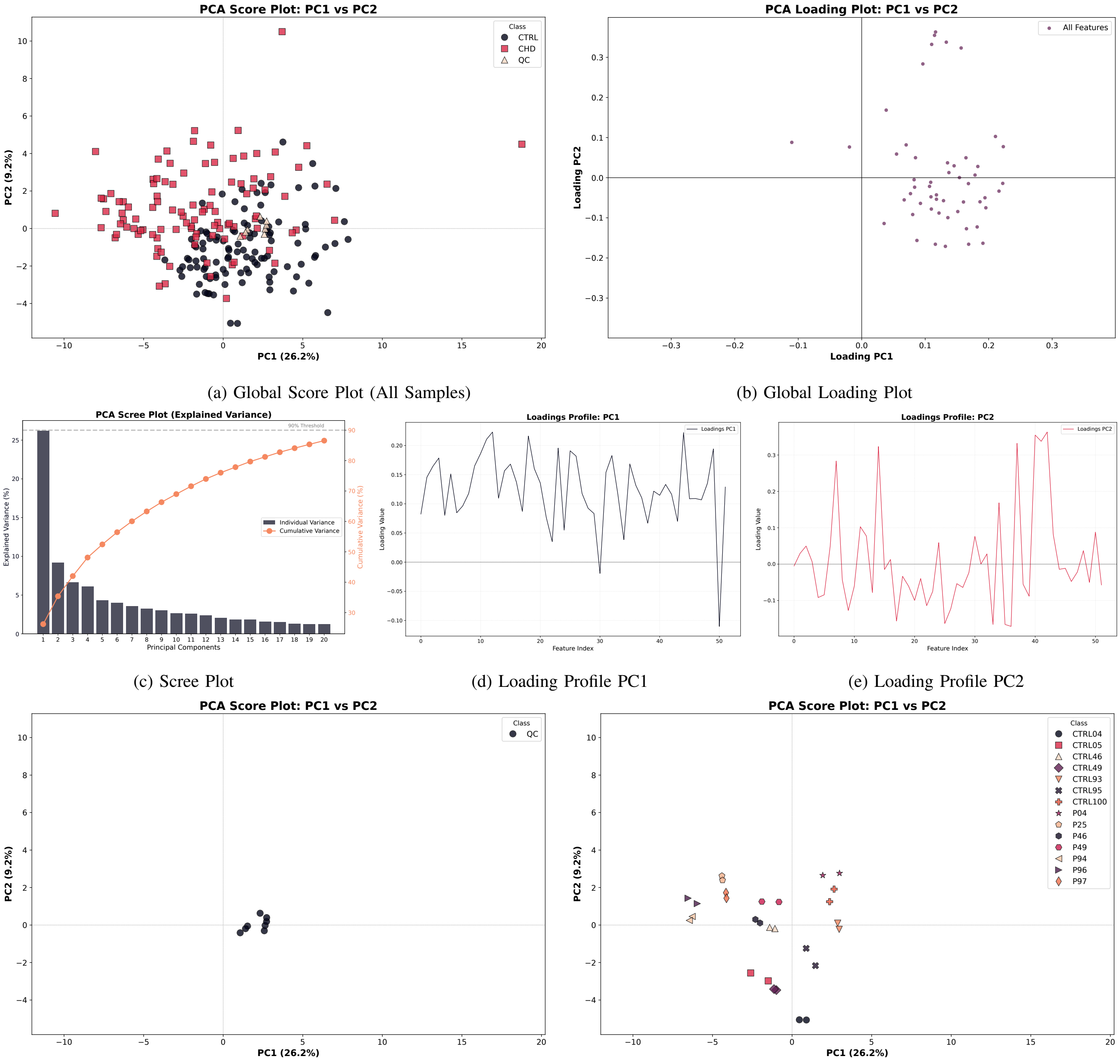
Having confirmed the technical robustness of the experiment, specific data cleaning steps were implemented to prepare the dataset for biological modeling.

d) Removal of Quality Control Samples

QC samples were removed from the final dataset as they have fulfilled their purpose of monitoring instrumental stability. Retaining QCs in downstream supervised analysis (e.g., PLS-DA) would introduce an artificial class that does not reflect a biological phenotype. Furthermore, due to their chemical homogeneity, QCs would form a dense cluster accounting for a large portion of the total variance, potentially masking the subtler biological differences between CHD and CTRL groups.

e) Removal of Technical Duplicates

To ensure the statistical independence of observations, technical duplicates were handled by retaining only one measurement per biological subject (samples with suffix _00). Including both replicates would violate the assumption of independence required by most statistical tests, artificially inflating the sample size and underestimating the intra-class variance. As duplicates were not available for all samples, averaging was avoided to prevent inconsistency in the data structure. Therefore, the removal of the second replicate (_01) ensures a homogeneous dataset where each sample represents a unique biological entity.



(a) Global Score Plot (All Samples)

(b) Global Loading Plot

(c) Scree Plot

(d) Loading Profile PC1

(e) Loading Profile PC2

(f) Instrumental Stability (QC only)

(g) Reproducibility (Replicates only)

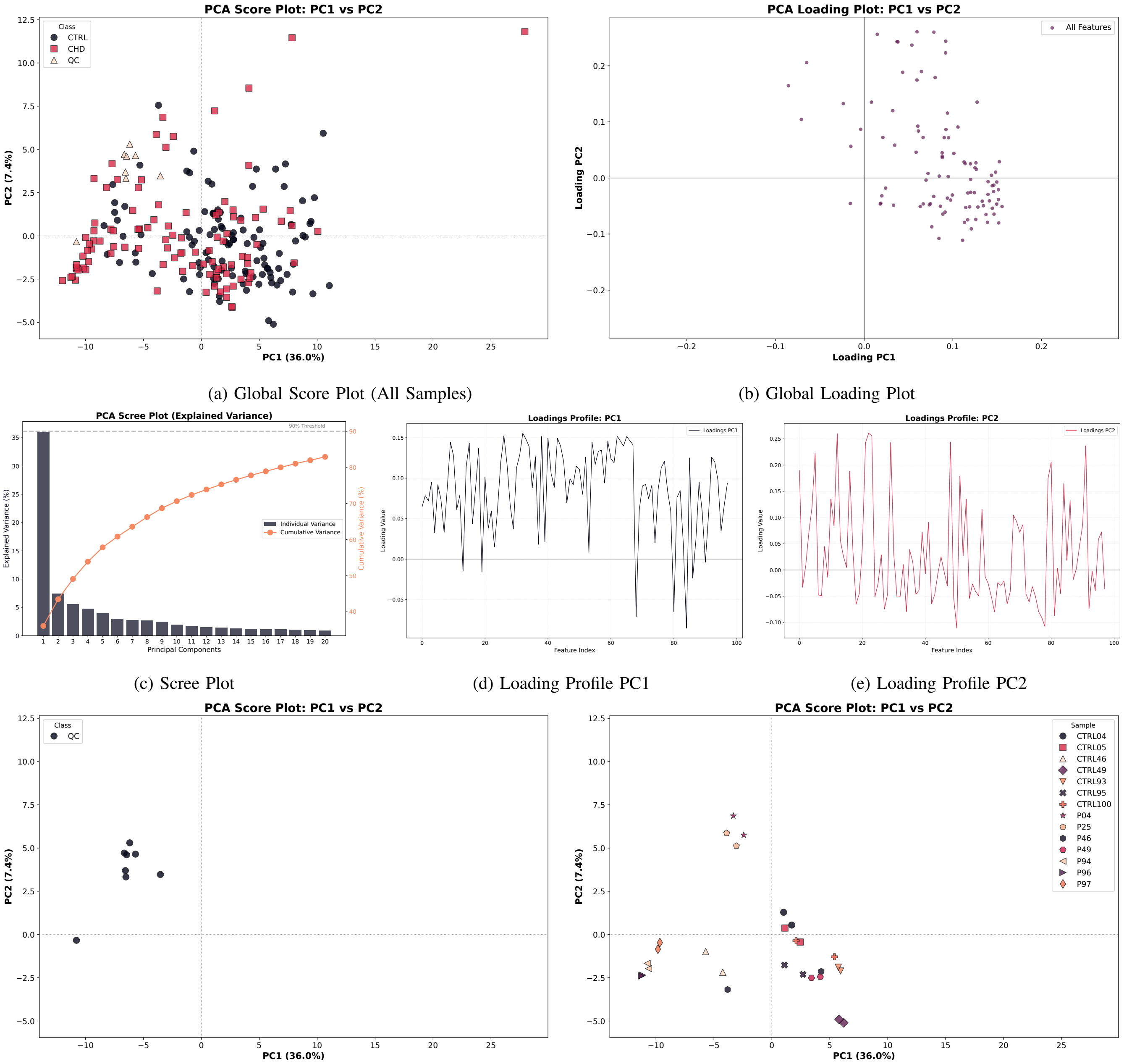
Fig. 1: **Quality Assessment for ESI- Dataset.** (a) Global PCA Score plot showing the distribution of all classes (CHD, CTRL, QC). (b) Loading plot showing feature contributions. (c-e) Variance analysis and loading profiles. (f) Zoom on QC samples: the tight cluster confirms high stability. (g) Zoom on technical replicates: paired samples show high overlap, confirming reproducibility.

2.3 Data Pre-Processing

2.3.1 Data Normalization

The LC-MS untargeted analysis of biological fluids is inherently subject to systematic variations unrelated to the biological problem, such as differences in sample dilution (e.g., hydration status of the subjects) and fluctuations in ionization efficiency. As evidenced by the raw data distribution (Fig. 3a and Fig. 4a), a significant variability in the median intensity across samples was observed, necessitating a normalization step to render the samples comparable.

We evaluated multiple normalization strategies, ranging from global intensity corrections (TIC, Mean, Median) to distribution-based methods (Quantile) and robust probabilistic approaches (Probabilistic Quotient Normalization - PQN). The selection of the optimal method was driven by a dual criterion: (i) qualitative inspection of sample distributions via boxplots, and (ii) quantitative assessment of the Coefficient of Variation (CV%) calculated on the technical replicates and across the biological


PCA Score Plot: PC1 vs PC2

PCA Loading Plot: PC1 vs PC2

(a) Global Score Plot (All Samples)

(b) Global Loading Plot

PCA Scree Plot (Explained Variance)

Loadings Profile: PC1

Loadings Profile: PC2

(c) Scree Plot

(d) Loading Profile PC1

(e) Loading Profile PC2

PCA Score Plot: PC1 vs PC2

PCA Score Plot: PC1 vs PC2

(f) Instrumental Stability (QC only)

(g) Reproducibility (Replicates only)

Fig. 2: **Quality Assessment for ESI+ Dataset.** (a-b) Global PCA model overview. (c-e) Variance and Loadings. (f) Stability check: QCs show a wider dispersion compared to negative mode, consistent with HILIC sensitivity, but remain distinct from biological variance. (g) Reproducibility check showing paired replicates.

groups.

Figures 3 and 4 illustrate the effect of selected normalization algorithms on the Negative (ESI-) and Positive (ESI+) datasets, respectively. While the raw data showed pronounced ”batch-like” or dilution-related fluctuations, all normalization methods improved the alignment of sample medians. Specifically:

- **Total Ion Current (TIC):** Provided a standard correction based on the total signal sum, effectively reducing global differences but potentially sensitive to high-intensity artifacts.
- **Quantile Normalization:** Resulted in perfectly aligned distributions (Fig. 3c, 4c). However, visual inspection suggests this approach may be overly aggressive, forcing all samples to conform to an identical distribution and potentially suppressing genuine biological heterogeneity.
- **Probabilistic Quotient Normalization (PQN):** Demonstrated a robust alignment of medians and interquartile ranges (Fig. 3d, 4d) without imposing the artificial uniformity observed with Quantile normalization.

To objectively quantify the reduction in technical variance, the median Coefficient of Variation (CV%) was calculated for all features across the Control (CTRL) and Disease (CHD) groups. The results are summarized in Table III and Table IV.

In both ionization modes, the absence of normalization yielded the highest variability (Avg CV $\approx$ 58% for ESI- and 68% for ESI+). Consistent with the visual inspection, **Quantile normalization** achieved the lowest numerical CV values (Avg CV $\approx$ 50% and 57%, respectively). However, minimal variance is not solely indicative of data quality; it may also reflect overfitting and loss of biological signal. **PQN** consistently ranked among the top performing methods, achieving a substantial reduction in variance (Avg CV $\approx$ 51.5% for ESI- and 59.7% for ESI+) comparable to Quantile and Median normalization, while theoretically preserving the relative abundance ratios of metabolites better than global sum methods.

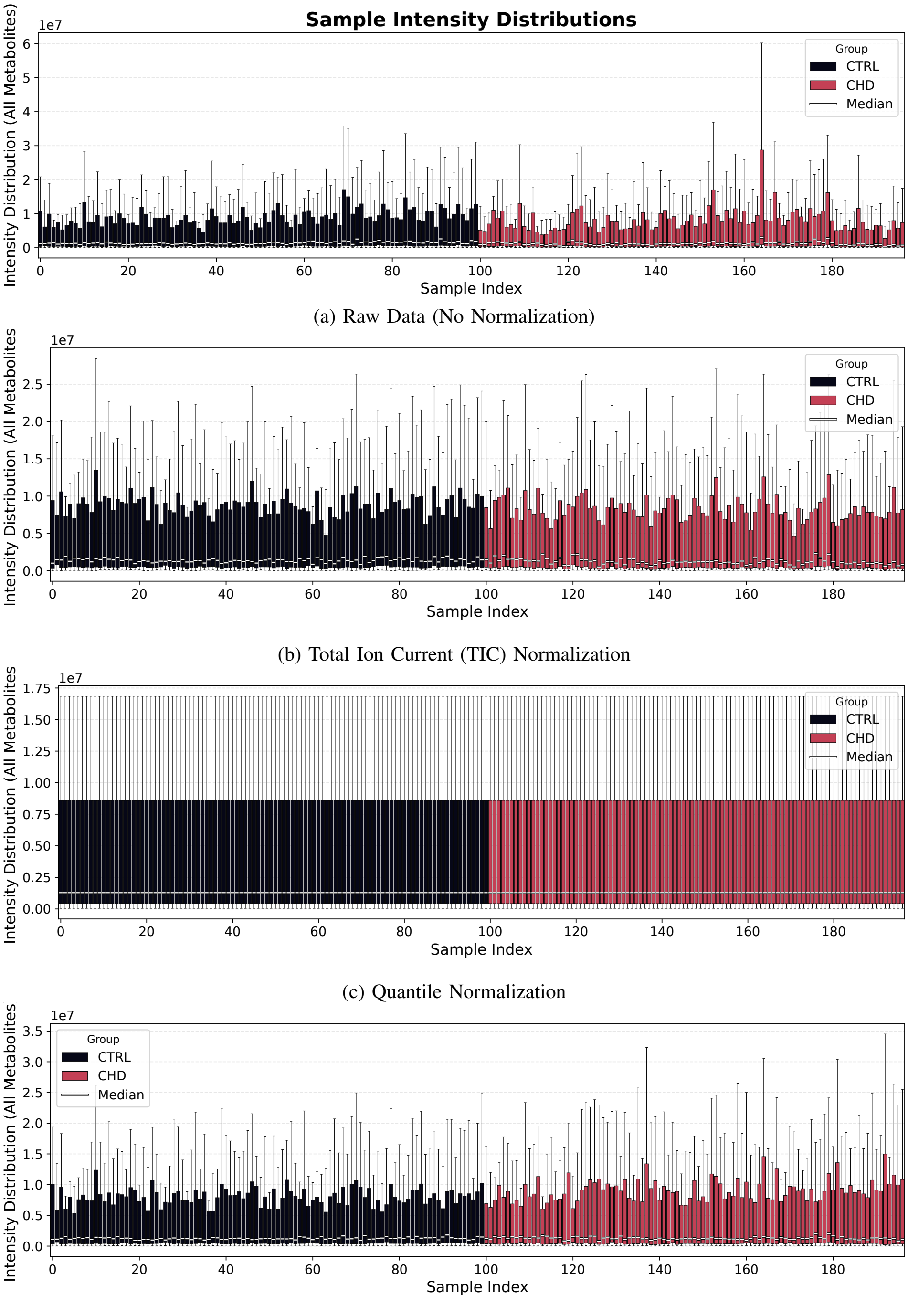
Based on the combined evidence, **PQN (Probabilistic Quotient Normalization)** was selected as the optimal strategy for this study. It provides the best trade-off between the reduction of systematic error (comparable to the most aggressive methods) and the preservation of biological variance required for the subsequent biomarker discovery phase.

TABLE III: Comparison of Normalization Methods by Coefficient of Variation (CV%) - ESI Negative Dataset.

Normalization Method	**Median CV CTRL (%)**	**Median CV CHD (%)**	**Average CV (%)**
Quantile	45.46	55.10	50.28
PQN	**44.94**	**58.06**	**51.50**
Mean / TIC	45.91	59.66	52.79
Median	45.48	63.94	54.71
Range / Max	51.28	63.22	57.25
None (Raw)	49.11	67.40	58.26

TABLE IV: Comparison of Normalization Methods by Coefficient of Variation (CV%) - ESI Positive Dataset.

Normalization Method	**Median CV CTRL (%)**	**Median CV CHD (%)**	**Average CV (%)**
Quantile	53.35	61.62	57.48
Median	54.48	60.69	57.58
Range / Max	53.84	64.46	59.15
PQN	**56.12**	**63.38**	**59.75**
Mean / TIC	54.44	65.09	59.77
None (Raw)	58.43	77.84	68.14



(a) Raw Data (No Normalization)

(b) Total Ion Current (TIC) Normalization

(c) Quantile Normalization

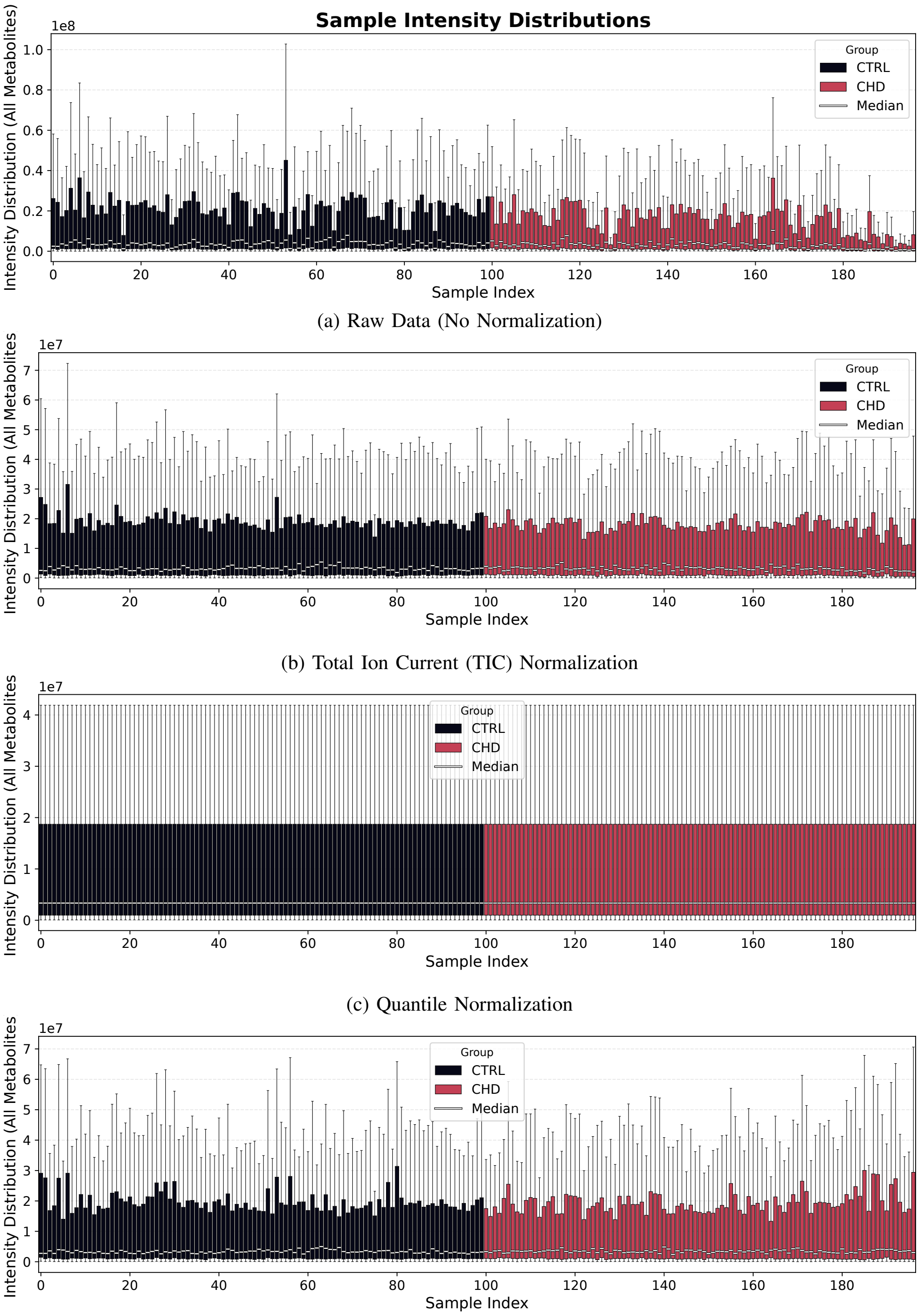
(d) Probabilistic Quotient Normalization (PQN)

Fig. 3: **Comparison of Normalization Strategies for ESI- Dataset.** Boxplots representing the global intensity distribution of all samples. (a) Raw data showing significant systematic variation (e.g., dilution effects). (b) TIC normalization, acting on the total sum. (c) Quantile normalization, forcing identical distributions potentially suppressing biological signal. (d) PQN, the selected method, which effectively reduces technical variance while preserving biological information.

2.3.2 Data Transformation

Following the assessment of normalization strategies, the distribution of the intensity values was evaluated to satisfy the assumptions of normality required by multivariate statistical methods (e.g., PCA, PLS-DA) and parametric univariate tests (e.g., t-test). LC-MS metabolomics data typically exhibit a right-skewed distribution, where the majority of signals have low intensity, while a few highly abundant metabolites stretch the dynamic range, potentially dominating the variance.

We evaluated different variance-stabilizing transformations, including power transformations (Square Root, Cube Root) and logarithmic transformations (Log2, Natural Logarithm, Log10). Visual inspection of the global density plots revealed that power



(a) Raw Data (No Normalization)

(b) Total Ion Current (TIC) Normalization

(c) Quantile Normalization

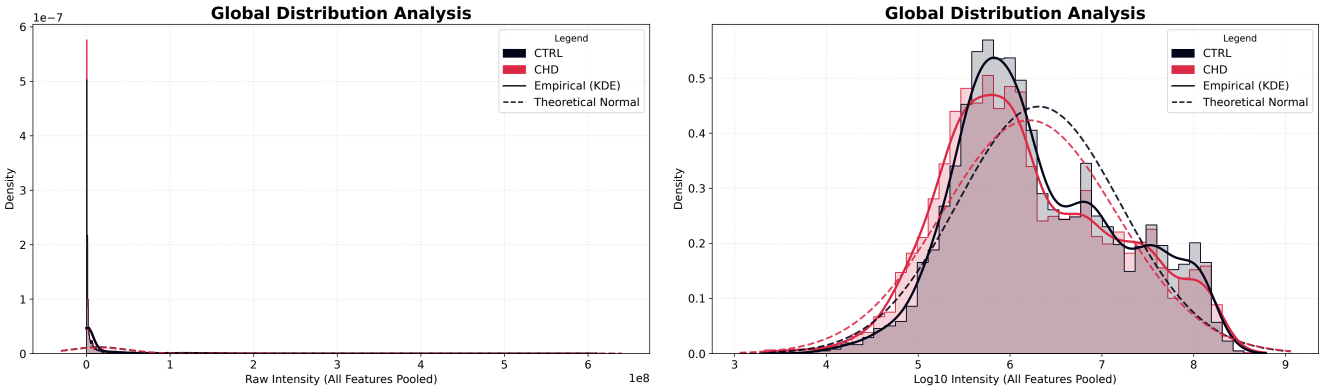
(d) Probabilistic Quotient Normalization (PQN)

Fig. 4: **Comparison of Normalization Strategies for ESI+ Dataset.** (a) Raw data distribution. (b) TIC normalization results. (c) Quantile normalization results showing aggressive distribution alignment. (d) PQN results, selected as the optimal compromise for downstream analysis.

transformations were insufficient to correct the skewness of the distributions (data not shown). Conversely, all logarithmic transformations effectively compressed the high-intensity values and expanded the low-intensity range, resulting in a distribution approximating a Gaussian curve.

Since Log2, Ln, and Log10 produced equivalent distributional shapes differing only in scale, **Log10 transformation** was selected as the standard method for this study. This transformation is widely accepted in mass spectrometry as it renders orders of magnitude easily interpretable while effectively symmetrizing the data distribution.

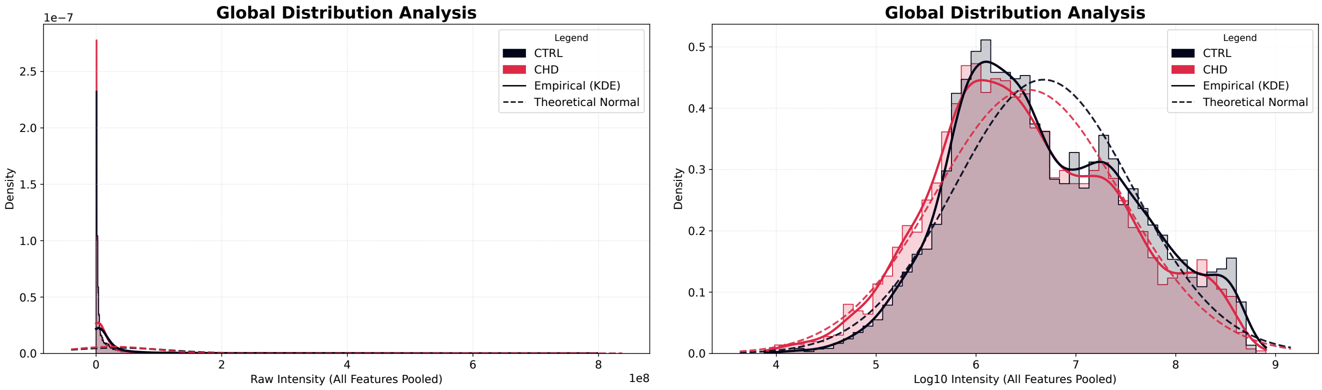
Figure 5 and Figure 6 illustrate the comparative analysis between the non-transformed data and the Log10-transformed data for the ESI- and ESI+ datasets, respectively. In the transformed data, the empirical density (solid line) shows a significant overlap with the theoretical Gaussian distribution (dashed line), confirming the efficacy of the transformation.



(a) Original Distribution (Skewed)

(b) After Log10 Transformation

Fig. 5: **Effect of Log10 Transformation on ESI- Dataset.** Global density plots pooling all feature intensities. (a) The original data distribution is highly right-skewed, deviating significantly from the theoretical normal distribution (dashed line). (b) Log10 transformation successfully centers the distribution, achieving a Gaussian-like shape suitable for multivariate analysis.



(a) Original Distribution (Skewed)

(b) After Log10 Transformation

Fig. 6: **Effect of Log10 Transformation on ESI+ Dataset.** (a) The raw positive mode data exhibits a pronounced right-skewness. (b) Log10 transformation corrects the skewness, aligning the empirical density (solid line) with the theoretical Gaussian curve (dashed line).

2.3.3 Data Scaling

The final step of the pre-processing pipeline involved data scaling. In metabolomics, metabolite intensities can span several orders of magnitude. Without proper scaling, variables with high abundance and large variance would naturally dominate multivariate models based on variance maximization (e.g., PCA), biasing the results and potentially masking significant biological variations present in low-abundance metabolites.

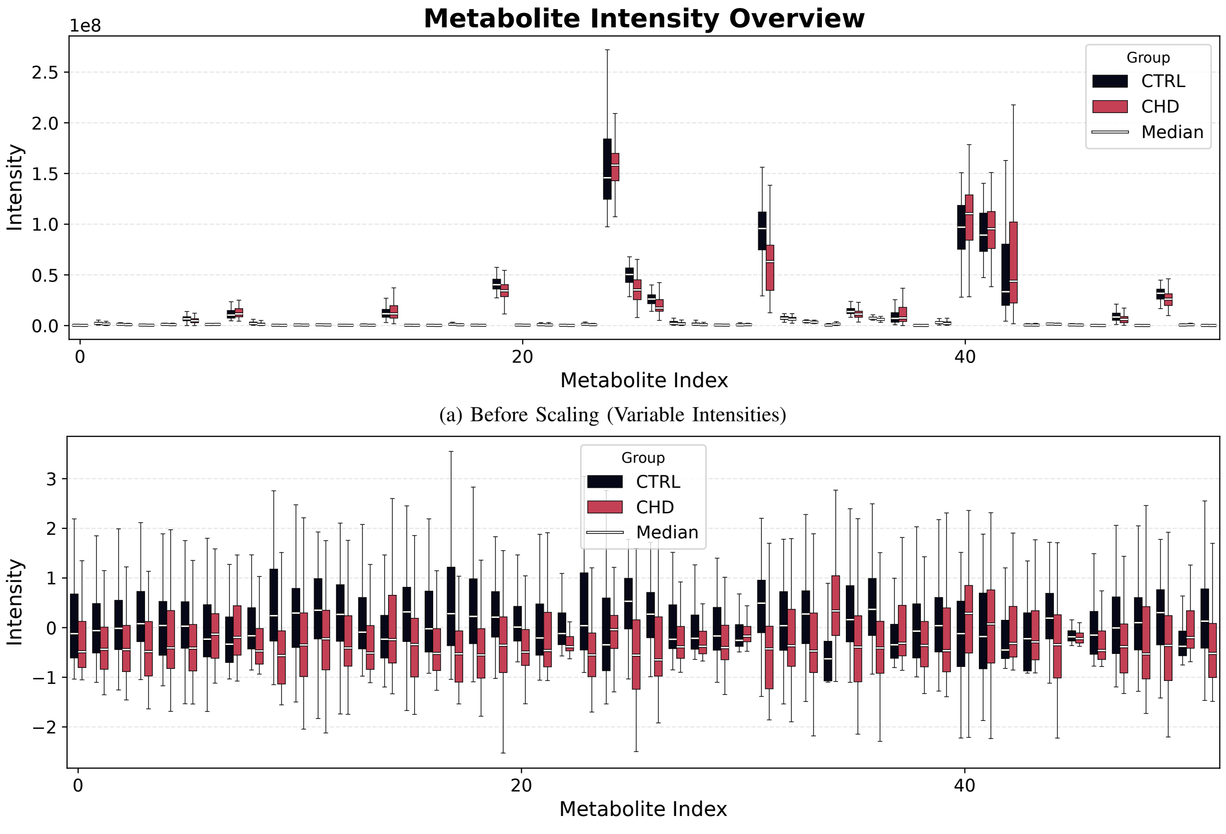
To address this issue, **Autoscaling** (Unit Variance Scaling) was applied to the normalized and transformed data. This method involves mean-centering each variable and dividing it by its standard deviation. As a result, all metabolites are scaled to have a mean of zero and a standard deviation of one, ensuring that each feature contributes equally to the statistical model regardless of its absolute concentration.

Figure 7 and Figure 8 demonstrate the effect of autoscaling on the feature distributions for ESI- and ESI+ datasets, respectively. Before scaling (Panel a), the variables exhibit disparate ranges of intensity. After autoscaling (Panel b), all features are comparable, centered around zero with standardized variance, making the dataset suitable for unsupervised and supervised modeling.

2.3.4 Global Assessment

To validate the efficacy of the entire pre-processing pipeline (PQN normalization, Log10 transformation, and Autoscaling), a comparative Principal Component Analysis (PCA) was performed on the dataset before and after data pretreatment. This step is essential to confirm that the applied corrections have successfully removed systematic bias and magnitude-dependent effects without distorting the biological signal.

Figure 9 illustrates the dramatic shift in data structure for the **Negative (ESI-) dataset**. In the **Raw Data** (Panels a-c), the variance is dominated by high-intensity compounds. The Loading Plot (c) shows that a restricted number of fatty acids



(a) Before Scaling (Variable Intensities)

(b) After Autoscaling (Unit Variance)

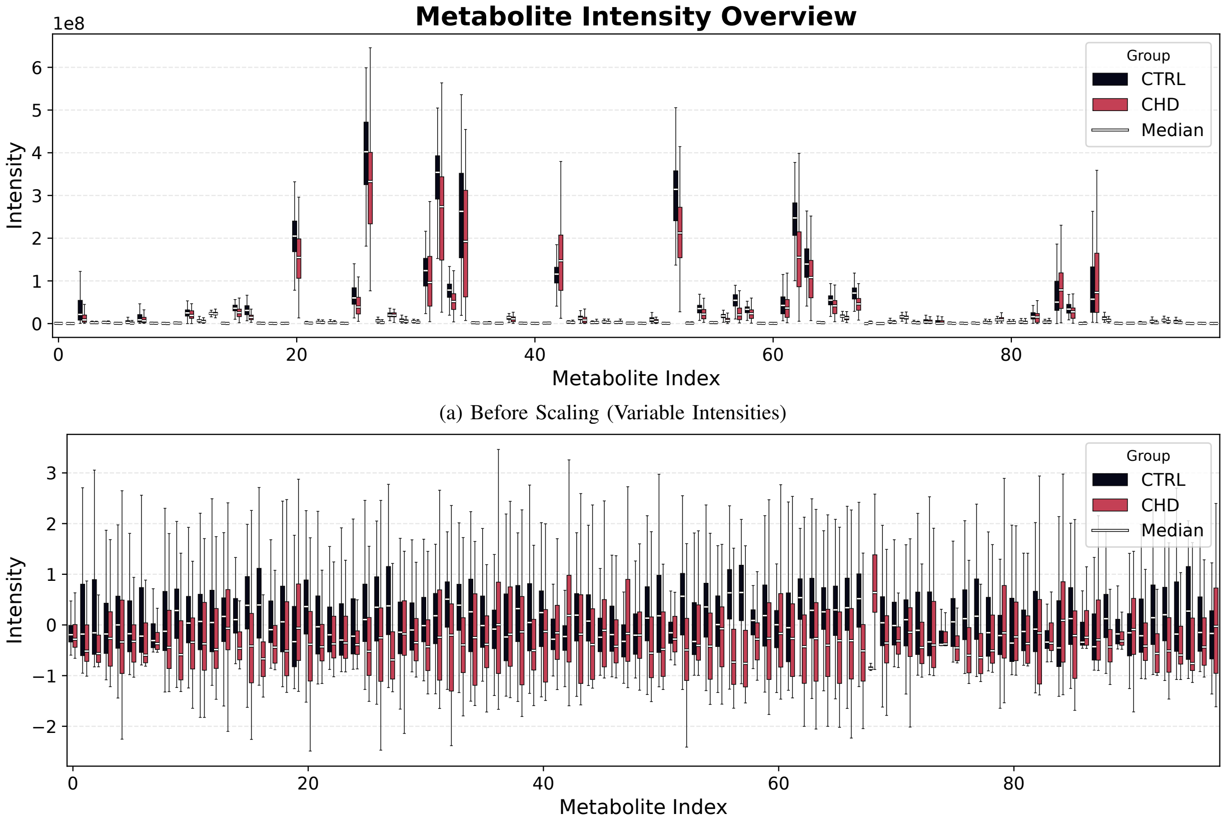
Fig. 7: **Effect of Autoscaling on ESI- Features.** Boxplots representing the distribution of individual metabolites (features). (a) Original features show highly variable ranges and variances. (b) After autoscaling, all features are mean-centered with unit variance, ensuring equal weight in multivariate analysis.

(e.g., Palmitoleic acid, Oleic acid) possess extremely high loading values, effectively driving the separation solely based on their abundance magnitude rather than biological class differences. Consequently, the Score Plot (b) shows a clustering pattern heavily influenced by these dominant features.

In the **Pre-processed Data** (Panels d-f), the variance is more evenly distributed across components (Scree Plot d), reflecting the effect of autoscaling which gives equal weight to all metabolites. The Score Plot (e) reveals a more homogeneous distribution of samples centered at the origin. Crucially, the Loading Plot (f) now highlights a diverse range of metabolites (e.g., Indolelactic acid) contributing to the model, ensuring that the subsequent multivariate analysis searches for biomarkers across the entire dynamic range of the metabolome.

A similiar trend was observed for the **Positive (ESI+) dataset** (Figure 10). In the raw state (Panels a-c), the variance was heavily skewed by high-abundance lipid species, specifically medium-chain acylcarnitines (e.g., C8-Carnitine, Decanoylcarnitine). The Loading Plot (c) clearly indicates that the separation along PC1 was driven almost exclusively by the magnitude of these compounds. Following the application of the optimized pre-processing pipeline (Panels d-f), the variance distribution was normalized. The Loading Plot (f) shows a re-balancing of feature importance: while acylcarnitines remain relevant, other metabolite classes (e.g., indolic compounds like trans-3-Indoleacrylic acid) now contribute significantly to the model components. This confirms that the pre-processing successfully revealed the latent biological information previously masked by the dominant lipid signals.

Having independently optimized the data quality for both ionization modes, the separate processing workflows were concluded. To exploit the complementary nature of the ESI+ and ESI- profiles and capture comprehensive metabolic signatures, the subsequent phase involves the integration of the two datasets. Specifically, a **Low-Level Data Fusion** strategy will be applied prior to outlier detection. This approach allows for the identification of potential anomalies not only within a single analytical domain but also arising from inconsistencies between the two ionization modes, ensuring a robust dataset for the final classification modeling.



(a) Before Scaling (Variable Intensities)

(b) After Autoscaling (Unit Variance)

Fig. 8: **Effect of Autoscaling on ESI+ Features.** (a) The disparate scales of metabolite intensities in the positive mode. (b) The homogenized feature space achieved through autoscaling.

2.4 Low-Level Data Fusion

To obtain a holistic view of the metabolic phenotype and exploit the complementary nature of the two ionization modes, a **Low-Level Data Fusion** strategy was implemented. This approach involves the concatenation of the pre-processed datasets into a single super-matrix, allowing for the simultaneous analysis of interactions between all variables.

2.4.1 Fusion Strategy and Block Scaling

Prior to concatenation, a critical **Block Scaling** step was performed to ensure fairness between the two analytical blocks (ESI- and ESI+). Although both datasets were individually autoscaled, differences in the number of variables (P) and the intrinsic numerical redundancy could lead to one block dominating the multivariate model. To prevent this, each block X_k was scaled by its Frobenius norm ($||X_k||_F$), defined as the square root of the sum of the squared elements:

$$X_{k,scaled} = \frac{X_k}{\sqrt{\sum_{i,j} x_{k,ij}^2}} \tag{1}$$

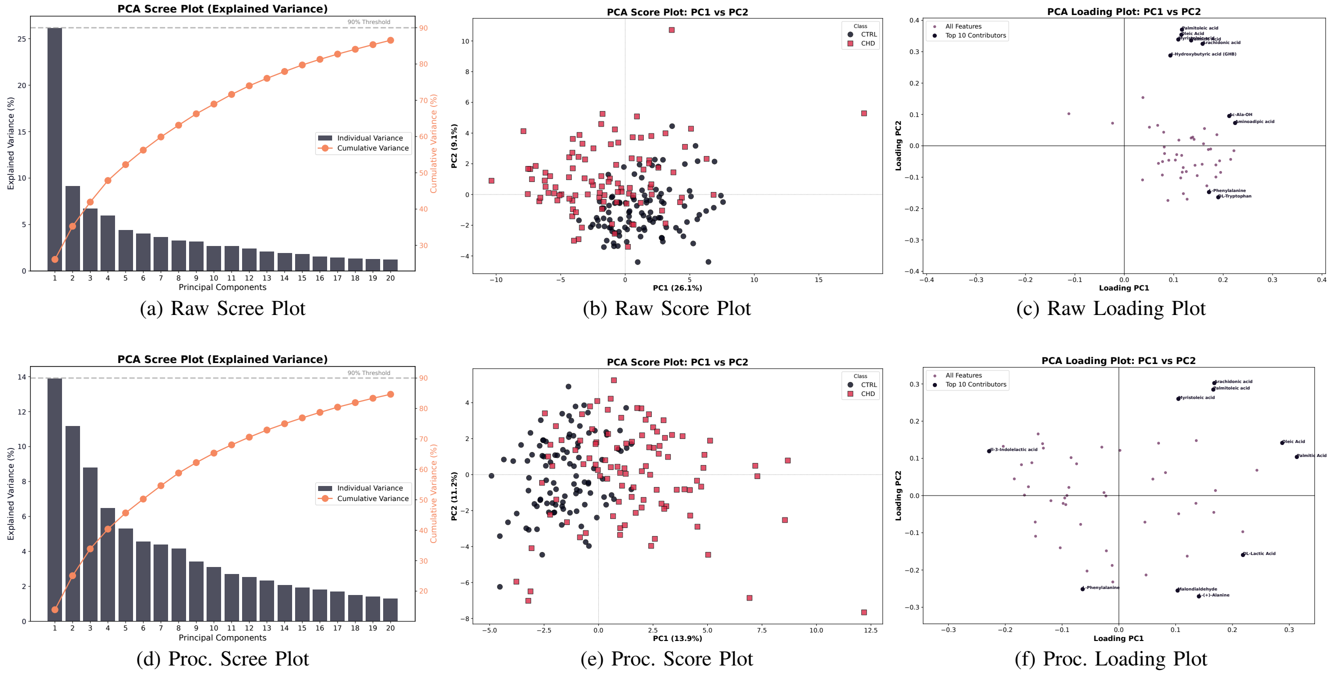
This operation normalizes the total variance (energy) of each matrix, ensuring that both ionization modes contribute equitably to the fused model. Subsequently, the blocks were concatenated horizontally to form the super-matrix $X_{fus} = [X_{neg}|X_{pos}]$, preserving the sample alignment.

2.4.2 SUM-PCA: Exploratory Analysis of Fused Data

Principal Component Analysis applied to the fused super-matrix (defined as **SUM-PCA**) was utilized to decompose the global variance into Super Scores (representing the consensus sample trajectory) and Super Loadings (representing the contribution of features from both blocks).

The results of the SUM-PCA are presented in Figure 11. The **Scree Plot (g)** reveals the variance contribution of each block to the principal components. A clear complementarity is observed: PC1 (13.8%) is predominantly driven by Block 2 (ESI+), indicating that the positive dataset accounts for the majority of the global variation. In contrast, PC2 (7.7%) and PC3 (7.2%) show a substantial recovery of contribution from Block 1 (ESI-), balancing the model.

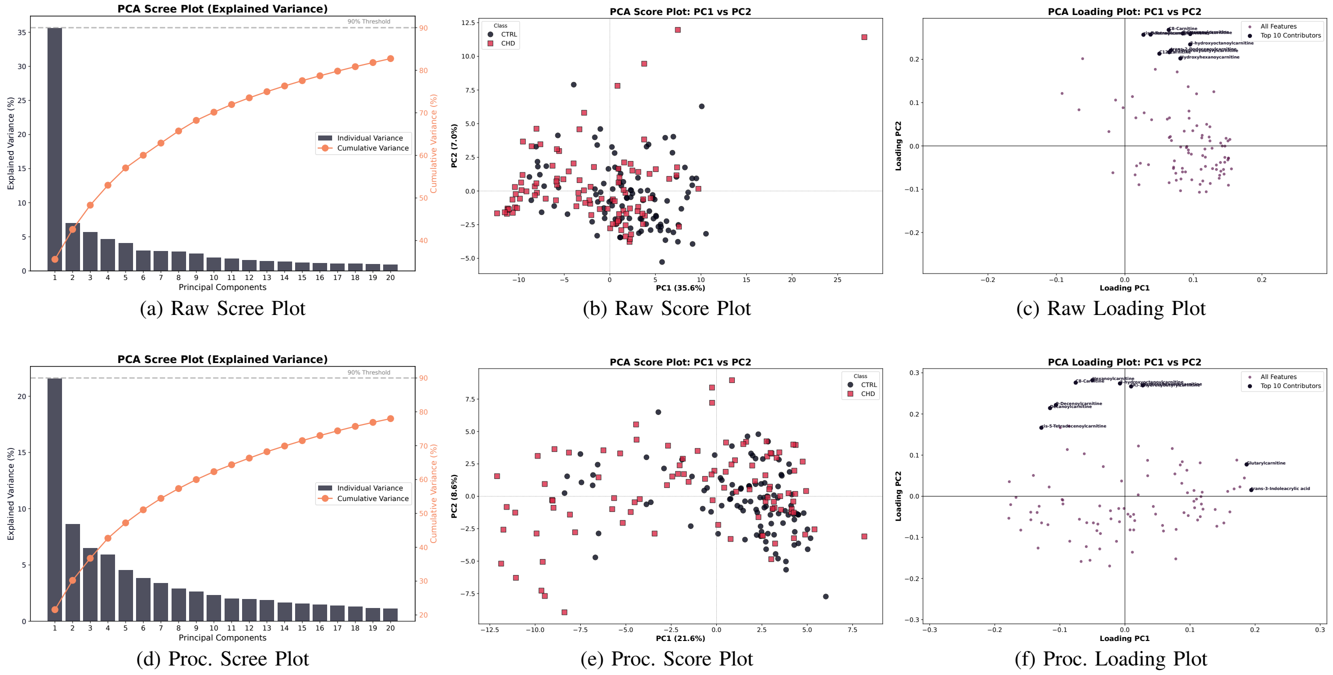
The **Score Plots (a-c)** show the distribution of samples in the integrated latent space, suggesting a trend of biological distinction driven by the combined metabolic profile.



(a) Raw Scree Plot (b) Raw Score Plot (c) Raw Loading Plot

(d) Proc. Scree Plot (e) Proc. Score Plot (f) Proc. Loading Plot

Fig. 9: **Impact of Pre-processing on ESI- Data Structure. Top Row (a-c):** PCA on raw data. The model is dominated by high-abundance features (fatty acids) explaining a large portion of variance (26% on PC1), masking subtle biological signals. **Bottom Row (d-f):** PCA on pre-processed data (PQN + Log10 + Autoscaling). The variance is democratized, and the influence of dominant metabolites is rescaled, revealing a more complex biological structure suitable for biomarker discovery.



(a) Raw Scree Plot (b) Raw Score Plot (c) Raw Loading Plot

(d) Proc. Scree Plot (e) Proc. Score Plot (f) Proc. Loading Plot

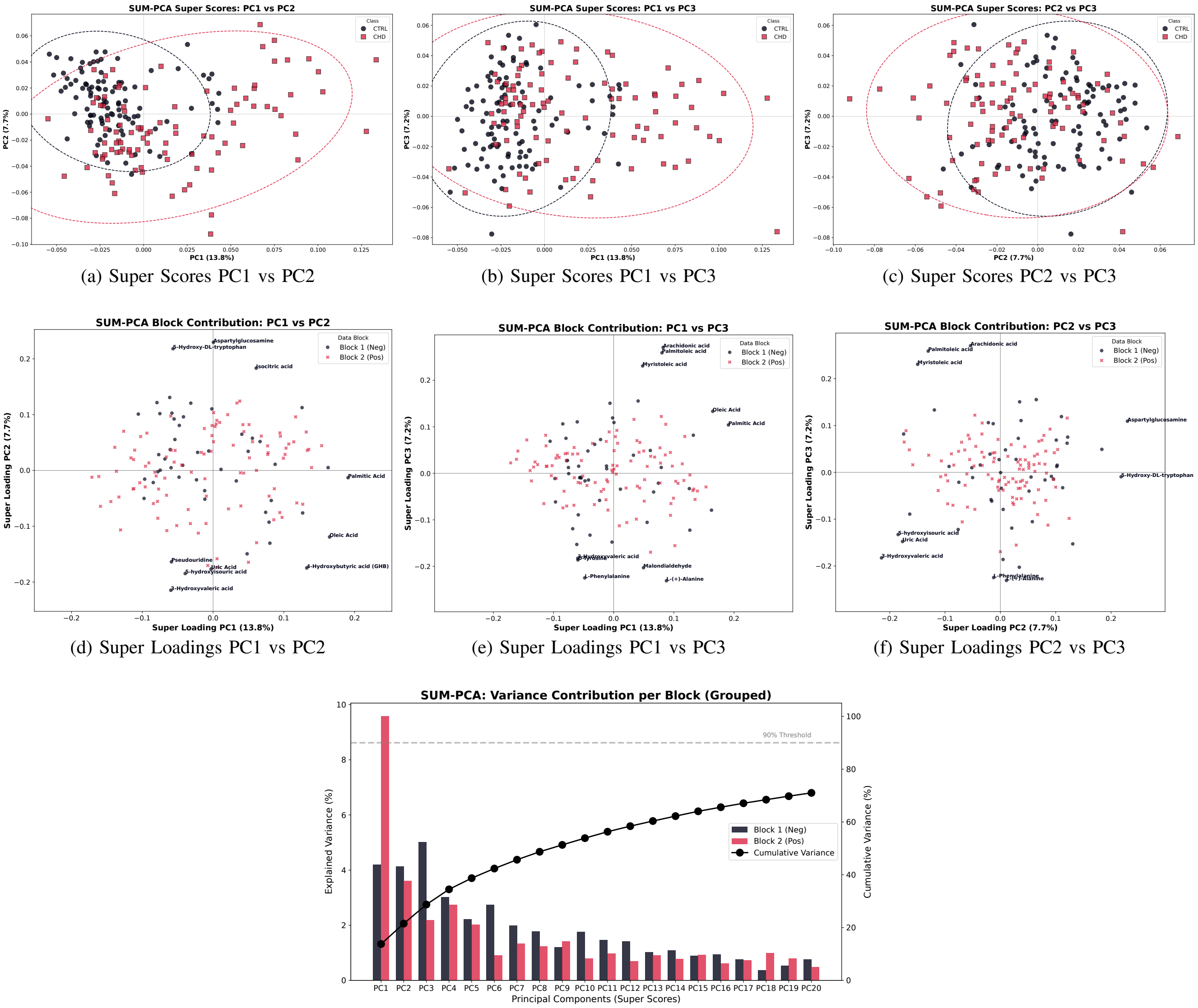
Fig. 10: **Impact of Pre-processing on ESI+ Data Structure. Top Row (a-c):** PCA on raw data. The model is biased by high-intensity acylcarnitines. **Bottom Row (d-f):** PCA on pre-processed data. The influence of dominant lipids is rescaled, allowing for a more comprehensive representation of the metabolome, including amino acid derivatives and indoles.

The **Super Loadings Plots (d-f)** provide insight into the specific features driving these components. Interestingly, while the Positive block dominates the global variance on PC1 (as seen in the Scree Plot), the top contributing features (highest loading values) identified in the plots belong primarily to the **Negative Block (ESI-)**.

- **PC1 vs PC2 (d):** The loading space is characterized by high-ranking negative mode metabolites, including fatty acids like Oleic Acid and Palmitic Acid, and polar compounds such as Uric Acid. This suggests that while the Positive block

provides the ”bulk” of the information, specific Negative block markers are highly specific and hold the strongest weights in defining the sample separation along these axes.
- **PC3 contributions:** Similarly, the third component is heavily influenced by specific markers from the Negative dataset, such as Arachidonic Acid and amino acid derivatives (e.g., L-Phenylalanine), further confirming the crucial role of ESI- features in characterizing the fine structure of the data.

This integrated analysis demonstrates that Low-Level Fusion successfully combined the broad variance coverage of the ESI+ mode with the high feature specificity of the ESI- mode.



(a) Super Scores PC1 vs PC2

(b) Super Scores PC1 vs PC3

(c) Super Scores PC2 vs PC3

(d) Super Loadings PC1 vs PC2

(e) Super Loadings PC1 vs PC3

(f) Super Loadings PC2 vs PC3

(g) Variance Contribution per Block (Scree Plot)

Fig. 11: **SUM-PCA Results on Low-Level Fused Data.** (a-c) Super Score plots showing sample distribution in the integrated space. (d-f) Super Loading plots. Notably, despite the high global variance of the Positive block, the top ranking contributors (labeled features) are primarily from the **ESI- Block** (e.g., Palmitic Acid, Uric Acid), highlighting their specific relevance. (g) Scree plot highlighting the variance contribution: PC1 is dominated by Block 2 (Pos), while subsequent components show increased contribution from Block 1 (Neg).

2.5 Anomaly Detection

Following data fusion, a rigorous anomaly detection phase was conducted to identify samples presenting extreme deviant behaviors or technical artifacts. This step was performed on the **Low-Level Fused dataset** to capture potential inconsistencies arising not only within a single ionization mode but also from the interaction between the two platforms (e.g., a sample appearing normal in ESI- but aberrant in ESI+).

Crucially, the detection was performed **independently for each class** (CTRL and CHD). This stratified approach ensures that the definition of ”normality” respects the specific biological topology of each group, preventing the algorithm from flagging a genuine pathological variation as a technical outlier.

A multi-methodological consensus strategy was adopted, combining multivariate statistics and machine learning algorithms to evaluate the samples from different topological perspectives (distance, density, and isolation).

a) Multivariate Statistical Profiling

First, Principal Component Analysis (PCA) was used to define the boundaries of the model space. Two complementary metrics were calculated for each sample:

- **Hotelling's T^2:** Measures the distance of a sample from the center of the model within the orthogonal subspace defined by the Principal Components (PCs). It identifies ”extreme” samples that follow the model but possess exaggerated leverage.
- **Q-Residuals (Squared Prediction Error):** Measures the distance of a sample from the model plane (i.e., the error between the raw data and the PCA reconstruction). It identifies samples with a chemical composition inconsistent with the correlation structure of the majority.

To ensure mathematical robustness, the number of PCs used for the T^2 calculation was dynamically selected to explain 90% of the variance, respecting the degrees of freedom constraints.

b) Machine Learning-Based Detection

To complement the statistical metrics, three unsupervised Machine Learning algorithms were deployed. The input space for these models was carefully selected to mitigate the ”Curse of Dimensionality”:

1) **Isolation Forest (iForest):** Applied directly to the **Raw Fused Data**. Since iForest relies on random partitioning rather than distance calculations, it is intrinsically robust to high-dimensional spaces. It identifies anomalies as points that are ”few and different,” requiring fewer random splits to be isolated from the rest of the data.
2) **One-Class SVM (OC-SVM):** Applied to the **PCA Scores (95% variance)**. This algorithm maps the data into a high-dimensional feature space using an RBF kernel to find the smallest hypersphere enclosing the ”normal” observations ($\nu = 0.05$). The use of PCA scores ensures the model focuses on biological variance rather than noise.
3) **Local Outlier Factor (LOF):** Applied to the **PCA Scores (95% variance)**. Unlike the global approach of iForest and OC-SVM, LOF evaluates the local density of a sample compared to its k-nearest neighbors ($k = 20$). It is particularly effective at identifying samples located in sparse regions or at the edges of the class cluster.

c) Visualization of Decision Boundaries

To visually interpret the behavior of the ML algorithms, the decision boundaries were projected onto the first two Principal Components (Figure 12 and Figure 13). These plots illustrate the different topological assumptions of the methods: iForest and OC-SVM define a global envelope around the core distribution, while LOF adapts to the local density variations.

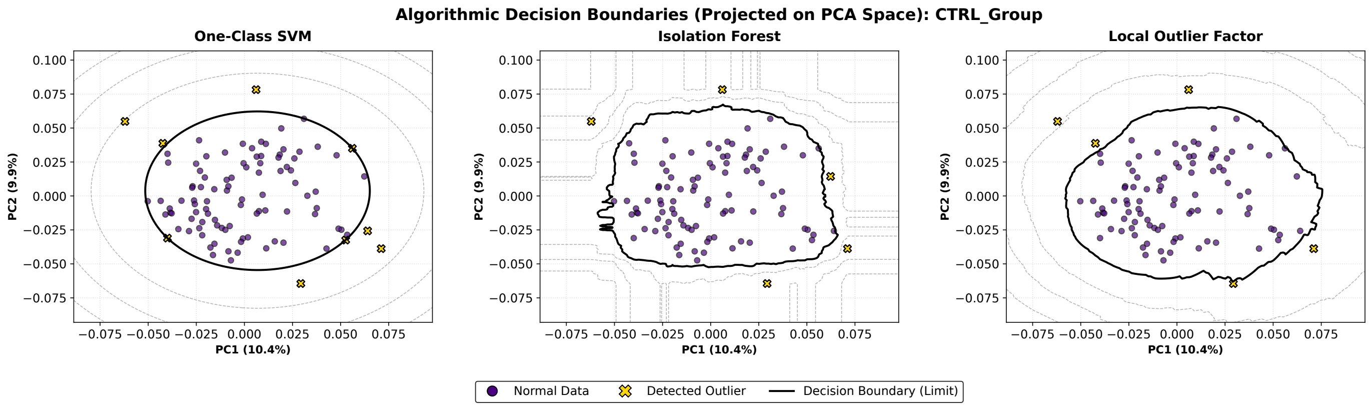


Fig. 12: **Algorithmic Decision Boundaries for Control Group (CTRL).** 2D projection of the decision boundaries generated by One-Class SVM, Isolation Forest, and Local Outlier Factor on the Control dataset. The colored regions and contours illustrate the area considered ”normal” by each algorithm. Samples falling outside these boundaries (marked with 'X') are flagged as potential outliers. Note that actual detection was performed in the high-dimensional space (or raw space for iForest) to preserve information; this 2D representation is for visualization of the topological approach only.

d) Consensus Evaluation and Sample Filtering

Since the primary objective of omics studies is to identify robust biomarkers, the exclusion of samples must be carefully balanced to avoid reducing statistical power while ensuring data quality. A consensus approach was adopted, integrating the multivariate metrics (T^2 vs Q) with the alerts generated by the Machine Learning algorithms and a visual inspection of the intensity distributions.

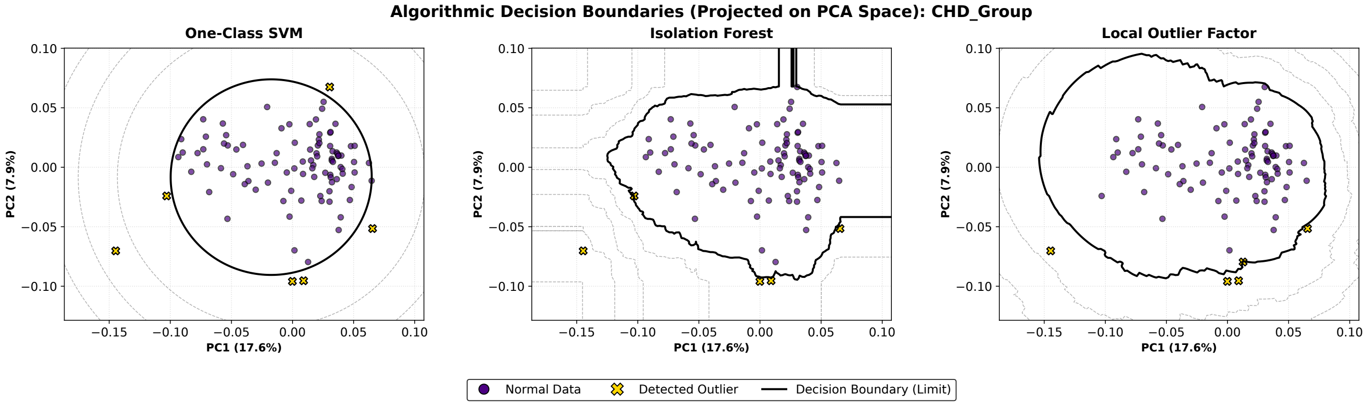


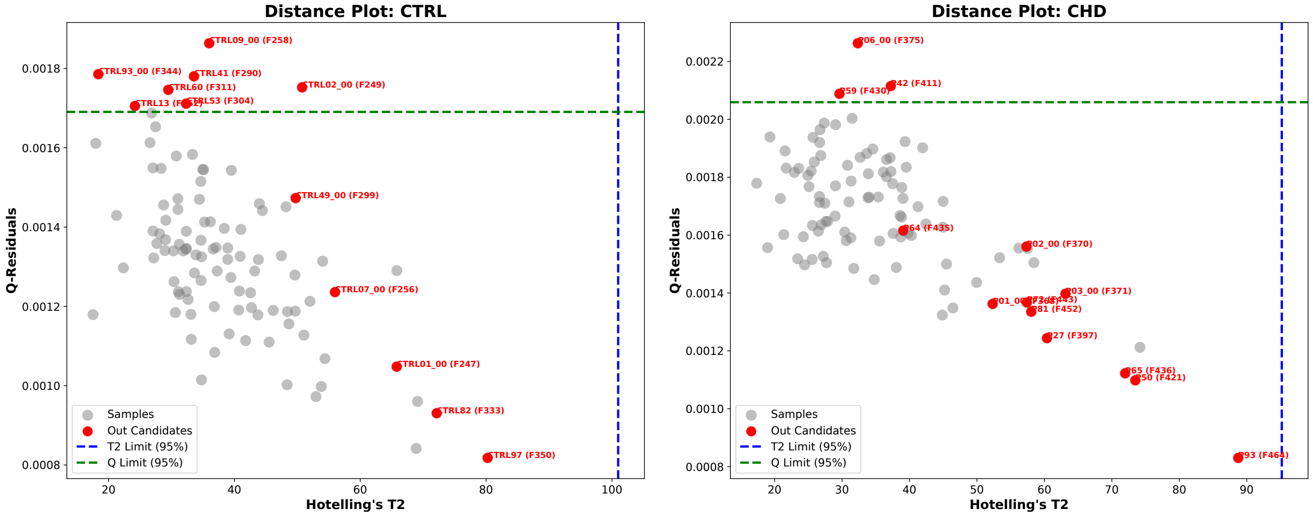
Fig. 13: **Algorithmic Decision Boundaries for Disease Group (CHD).** Visualization of anomaly detection results for the CHD class. The algorithms successfully identify samples that deviate from the main distribution cluster. The consensus among these methods, combined with the statistical metrics (T^2 and Q), forms the basis for the final exclusion criteria.

The **Distance Plots** (Figure 14) provide a comprehensive view of the outlier candidates. Notably, the majority of flagged samples exhibited high **Q-Residuals** (y-axis), indicating a significant deviation from the correlation structure of the model (Model Mismatch), likely due to technical artifacts or qualitative inconsistencies in sample composition. Conversely, no extreme biological outliers (high T^2) were detected in the Control group, whereas the CHD group showed specific deviant behaviors.

To validate these statistical alerts, the raw intensity distributions of the candidate samples were inspected via **Boxplots** (Figure 15 and Figure 16). This step was decisive for borderline cases.

- **Confirmed Exclusions:** Samples *CTRL09_00*, *CTRL93_00*, and *CTRL41* were removed due to severe violation of the Q-residuals limit. Similarly, *CTRL60*, *CTRL02_00*, and *CTRL53* were excluded as their statistical anomaly was corroborated by visible distributional artifacts in the boxplots. In the pathological group, *P06_00*, *P42*, and *P59* were removed for high Q-residuals. Sample *P93* was excluded as a unique "biological outlier," being unanimously flagged by all Machine Learning algorithms (iForest, SVM, LOF) and located at the extreme of the model space (T^2).
- **Saved Samples:** Sample *CTRL13*, despite having a Q-residual value near the threshold, was retained. Its raw intensity profile appeared consistent with the healthy population distribution, suggesting that the statistical deviation was not sufficient to justify data loss.

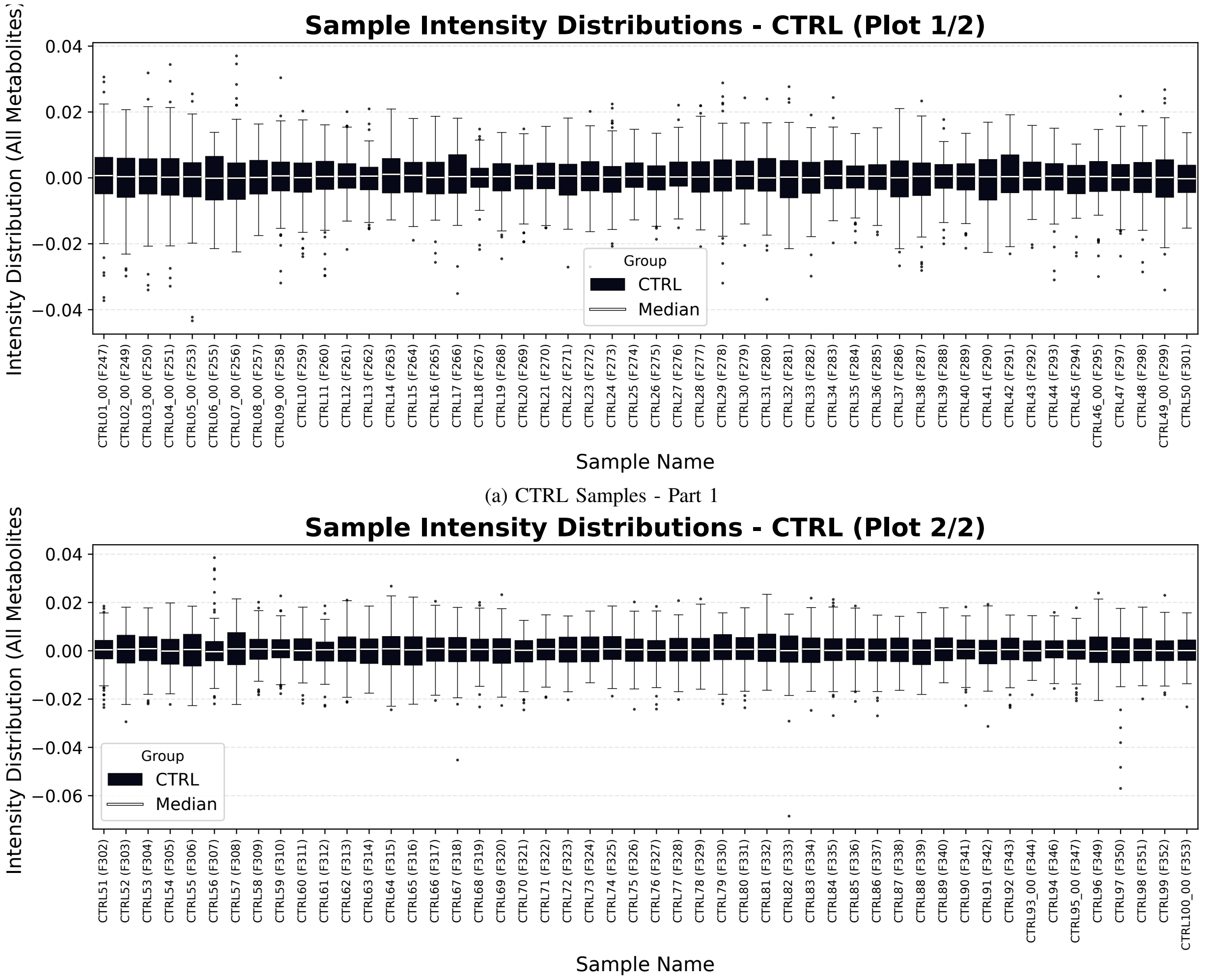
In total, **10 samples** (approx. 5% of the dataset) were removed: *CTRL02_00, CTRL09_00, CTRL41, CTRL53, CTRL60, CTRL93_00* from the Control group, and *P06_00, P42, P59, P93* from the CHD group.



(a) Control Group

(b) CHD Group

Fig. 14: **Distance Plots (T^2 vs Q-Residuals).** Gold standard visualization for outlier diagnosis. Samples flagged by at least two Machine Learning algorithms are highlighted in red. (a) Controls show anomalies primarily in the Q-residual space (vertical axis), indicating technical mismatch. (b) CHD samples show a more complex pattern, with sample P93 appearing as a distinct outlier consistent with ML detection.



(a) CTRL Samples - Part 1

(b) CTRL Samples - Part 2

Fig. 15: **Intensity Distributions of Control Samples.** Visual inspection used to validate statistical outliers. Deviant distributions confirmed the exclusion of samples like CTRL02_00.

e) Final Dataset Re-Processing and Validation

Following the removal of the identified outliers, the dataset was re-processed from the ground up. This step is critical: removing samples alters the global parameters (mean, standard deviation, Frobenius norm) used for Normalization, Autoscaling, and Block Scaling. Therefore, to ensure mathematical rigor, the remaining samples were re-normalized and re-scaled independently in their respective blocks (ESI- and ESI+) before being fused again.

A final PCA was performed on the cleaned and re-fused dataset (Figure 17) to verify the data structure. The Score Plot (b) shows a more homogeneous distribution without the compressing effect of extreme outliers. The Scree Plot (a) confirms a balanced variance distribution, and the Loading Plot (c) highlights that the separation is now driven by relevant biological features (e.g., C8-Carnitine, Indole derivatives) rather than technical artifacts. This curated dataset constitutes the input for the subsequent Data Splitting and Classification phases.

2.6 Dataset Splitting

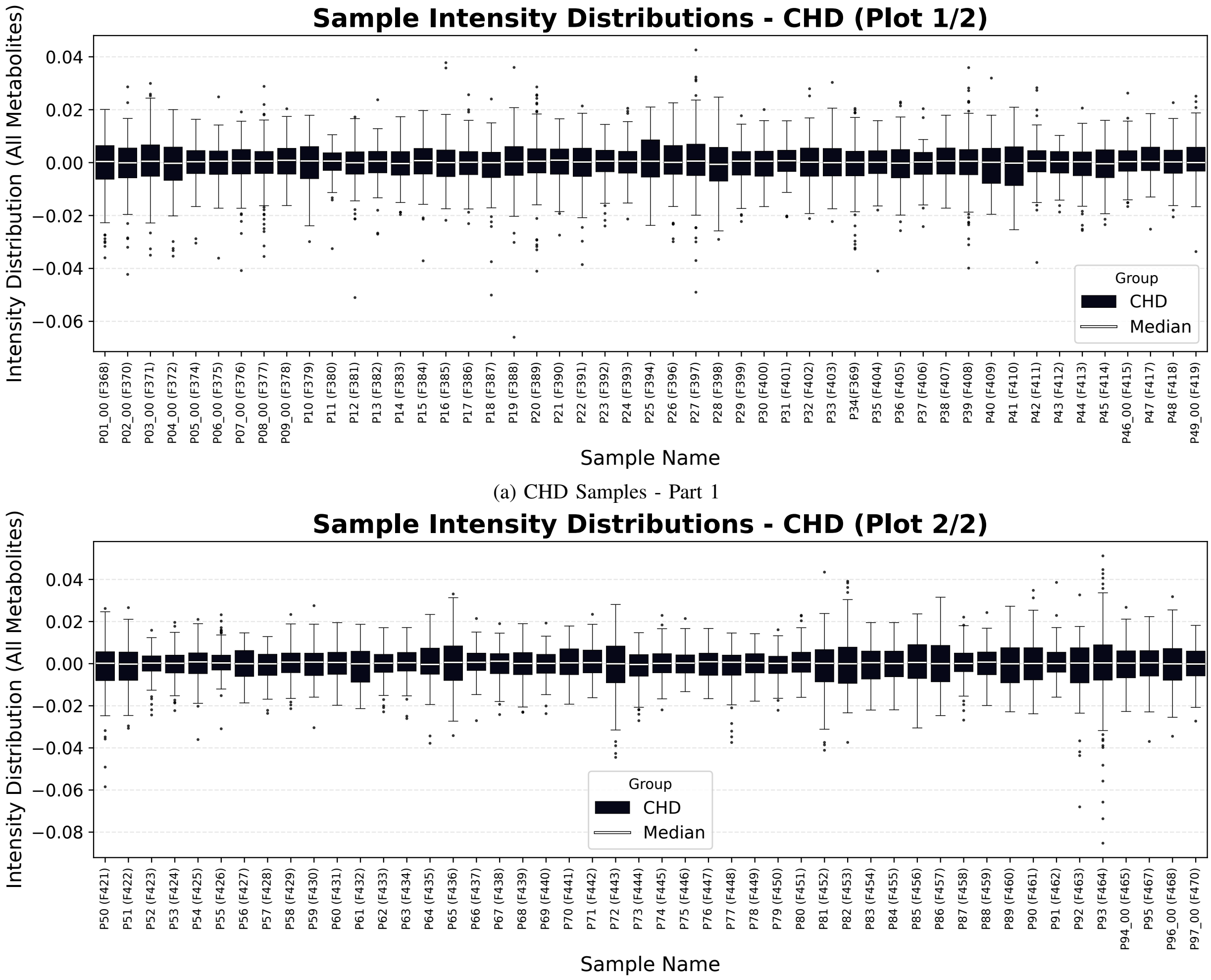
Descrizione del metodo usato per identificare e rimuovere gli outlier (fondamentale per la pulizia del dato).

2.7 Analisi Statistica e Machine Learning

- **Unsupervised:** PCA (per l'esplorazione).
- **Supervised:** PLS-DA, SVM, Random Forest, Logistic Regression.
- **Validazione:** Descrizione rigorosa dello split Training Set vs Validation Set (o Cross-Validation) per evitare l'overfitting.

2.8 Stack Tecnologico

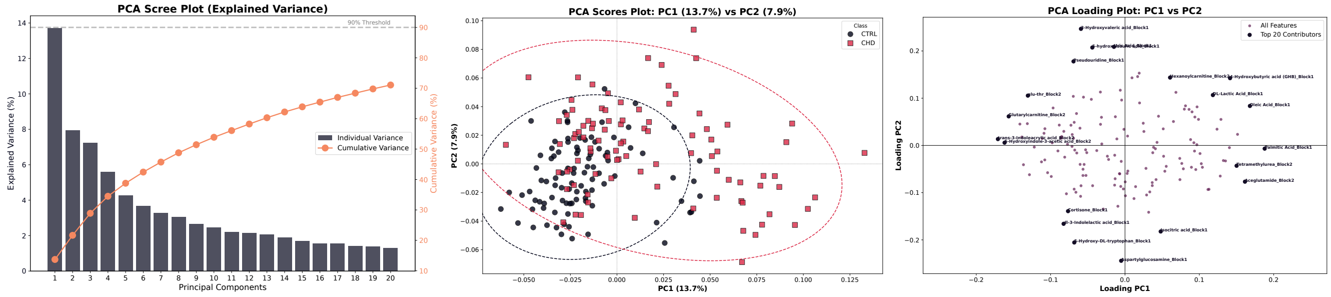
Breve paragrafo sulle librerie Python utilizzate (Pandas, Scikit-learn, ecc.) per garantire la riproducibilità.



(a) CHD Samples - Part 1

(b) CHD Samples - Part 2

Fig. 16: **Intensity Distributions of CHD Samples.** Used to verify the consistency of pathological samples before removal.



(a) Final Scree Plot

(b) Final Score Plot

(c) Final Loading Plot

Fig. 17: **PCA of the Curated Fused Dataset (Post-Outlier Removal).** (a) Variance is well-distributed. (b) Samples are homogeneously distributed in the score space, with no extreme outliers distorting the axes. (c) Loadings confirm the contribution of relevant biological markers (e.g., Acylcarnitines, Indoles) to the data variance.

3 Results and Discussion

3.1 Performance dei Modelli su Singoli Dataset (ESI+ / ESI-)

Confronto delle metriche (Accuratezza, Specificità, Sensibilità) tra PLS-DA, SVM, RF, LR. Quale modello performa meglio sui dati positivi? E sui negativi?

3.2 Interpretabilità e Potenziali Biomarcatori

Analisi delle Feature Importances e Analisi Univariata (Volcano Plot). Identificazione/interpretazione biologica dei top-metaboliti.

3.3 Univariate analysis sui biomarcatori

Metodi statistici/algoritmici utilizzati per ridurre la dimensionalità e rimuovere il rumore prima del modeling.

4 Conclusions

Sintesi del miglior workflow identificato. Considerazioni sull'interpretabilità biologica e limiti dello studio (es. numero di campioni, assenza di validazione esterna).

References

[1] Mires, Stuart, et al. "Plasma metabolomic and lipidomic profiles accurately classify mothers of children with congenital heart disease: an observational study." Metabolomics 20.4 (2024): 70.

TABLE V: Simulation Parameters

Information message length	$k = 16000$ bit
Radio segment size	$b = 160$ bit
Rate of component codes	$R_{cc} = 1/3$
Polynomial of component encoders	$[1, 33/37, 25/37]_8$