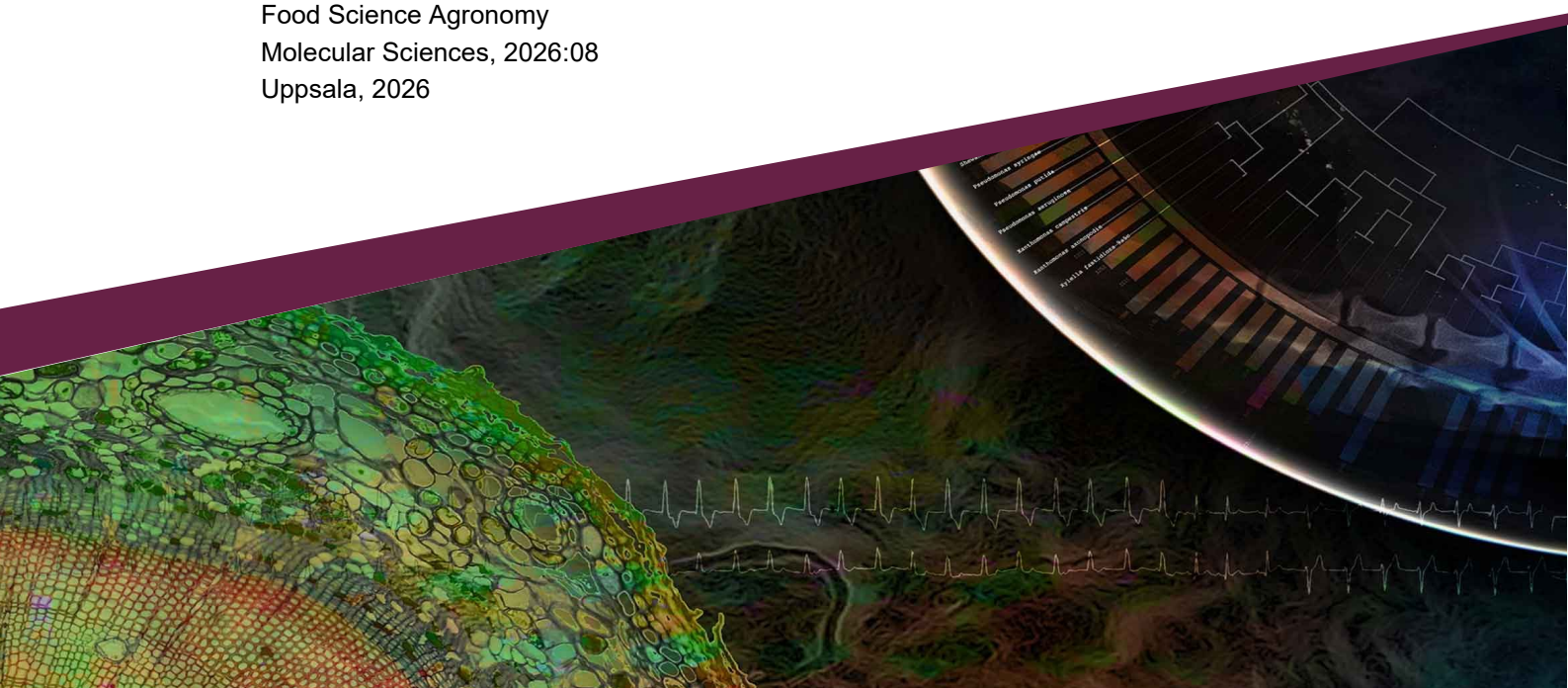
# Metabolomics-Based Biomarker Discovery: Challenges and Opportunities with Machine Learning

Maryam Chaid

Degree project/Independent project • 15 hp
Swedish University of Agricultural Sciences, SLU
NJ Faculty
Food Science Agronomy
Molecular Sciences, 2026:08
Uppsala, 2026



# Metabolomics-Based Biomarker Discovery: Challenges and Opportunities with Machine Learning

**Maryam Chaid**

**Supervisor:** Ali Moazzami, department of Molecular Sciences

**Examiner**: Jan Eriksson, department of Molecular Sciences

| | |
|---|---|
| **Credits:** | 15 |
| **Level:** | G2E |
| **Course title:** | Independent project in Food Science |
| **Course Code:** | EX0876 |
| **Course coordinating dept:** | Department of Molecular Sciences |
| **Place of publication:** | Uppsala |
| **Title of series:** | Molecular Sciences |
| **Part number:** | 2026:08 |
| **Keywords:** | analytical chemistry, biomarker discovery, deep learning, machine learning, metabolomics, systems biology, supervised learning, unsupervised learning |

**Swedish University of Agricultural Sciences**
Faculty of Natural Resources and Agriculture Science (NJ)
Department of Molecular Sciences

# Abstract

This literature study investigates the challenges and opportunities of using machine learning for biomarker discovery in metabolomics, with emphasis on how data characteristics influence model performance. In general, simple machine learning methods are well suited for small datasets and offer higher interpretability, but they may fail to capture complex biological relationships between metabolites, while more advanced methods can model such complexity but require larger datasets, careful tuning and often provide limited interpretability. Importantly, good statistical performance does not necessarily reflect true biological relevance, as evaluation metrics may indicate strong predictive ability even when underlying biological relationships are not properly captured, particularly in high-dimensional and imbalanced datasets. For this reason, model results must be interpreted with caution and supported by robust validation strategies, although commonly used validation approaches can still produce overly optimistic results if not correctly applied and more rigorous methods are often constrained by data availability and computational demands. At the same time, metabolomics data itself presents major challenges due to high dimensionality, missing values and technical variation, which makes data preprocessing a critical step in ensuring reliable results. Overall, the findings suggest that in metabolomics-based biomarker discovery, data quality, preprocessing and study design have a greater impact on outcomes than the choice of machine learning model alone.



*Keywords:* analytical chemistry, biomarker discovery, deep learning, machine learning, metabolomics, systems biology, supervised learning, unsupervised learning

# Table of contents

# List of figures

# Abbreviations

| | |
|---|---|
| AUC | Area under the curve |
| DL | Deep learning |
| GC-MS | Gas chromatography-mass spectrometry |
| kNN | k-nearest neighbours |
| KS | Kolmogorov-Smirnov |
| LASSO | Least absolute shrinkage and selection operator |
| LC-MS | Liquid chromatography-mass spectrometry |
| LR | Logistic regression |
| MCC | Matthew's correlation coefficient |
| ML | Machine learning |
| MS | Mass spectrometry |
| NMR | Nuclear magnetic resonance |
| OPLS-DA | Orthogonal partial least squares discriminant analysis |
| PCA | Principal component analysis |
| PLS-DA | Partial least squares discriminant analysis |
| RF | Random forest |
| ROC | Receiver operating characteristic |
| SHAP | Shapely additive explanations |
| SVM | Support vector machine |
| XGBoost | eXtreme Gradient Boosting |

# 1. Introduction

## 1.1 Metabolomics

Metabolomics is a rapidly evolving field within systems biology that focuses on the comprehensive analysis of small-molecule metabolites in biological systems. Metabolites, which are the end products of cellular processes, reflect the integrated outcome of gene expression, protein activity and environmental influences, and are therefore closely linked to the phenotype of an organism (Worley & Powers, 2013). Unlike genomics or proteomics, which provide information on potential or upstream biological activity, metabolomics captures the actual biochemical state of a system at a given time point. This makes it a particularly powerful approach to studying physiological and pathological processes.

The concept of metabolomics dates back much earlier than the field itself, as people historically used observable traits, such as the sweetness of urine, to identify elevated glucose levels in diabetes (Ren et al., 2015). However, the development of advanced chemical analytical platforms has been central to the growth of metabolomics. Techniques such as mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy enable the detection and quantification of metabolites in complex biological matrices such as blood, urine and tissue samples (Tuerxunyiming et al., 2025). MS, often coupled with separation techniques such as liquid or gas chromatography, offers high sensitivity and the ability to detect low-abundance metabolites, while NMR provides highly reproducible and non-destructive analysis, albeit with lower sensitivity (Ren et al., 2015). These complementary technologies allow for both targeted and untargeted metabolomic approaches, facilitating broad metabolic profiling as well as the precise quantification of specific compounds.

The metabolomics workflow typically involves several interconnected steps, including experimental design, sample collection and preparation, data acquisition, preprocessing and statistical analysis (Anwardeen et al., 2023). Each of these stages can introduce both systematic and random sources of variation, which may influence downstream analyses. In particular, preprocessing steps such as peak detection, alignment, normalization and scaling are critical for transforming raw spectral data into structured data matrix suitable for statistical evaluation (Anwardeen et al., 2023). The quality of these steps has a direct impact on the

reliability of the results, highlighting the importance of rigorous methodological control throughout the analytical pipeline.

One of the primary motivations for applying metabolomics in biomedical research is its potential for biomarker discovery. Biomarkers are measurable biological indicators that can be used to detect or predict disease, monitor disease progression or evaluate responses to treatment. The identification of reliable biomarkers is essential for improving diagnostic accuracy, enabling early disease detection and advanced personalized medicine. Because metabolites are closely linked to biochemical pathways and physiological states, metabolomics offers a unique opportunity to identify biomarkers that are directly associated with disease mechanisms. Also, in many diseases, metabolic alterations occur before the onset of clinical symptoms, making metabolite-based biomarkers particularly valuable for early diagnosis (Shomorony et al., 2020; Anwardeen et al., 2023; Zhang et al., 2023). However, despite numerous reported candidate biomarkers, the translation of metabolomic findings into clinically validated biomarkers remains limited.

A major factor contributing to this limitation is the inherent complexity of metabolomics datasets. A defining feature of such data is high dimensionality, where the number of measured metabolites far exceeds the number of samples (Sun et al., 2024). This imbalance presents significant challenges for statistical analysis, as it increases the likelihood of identifying random correlations that do not reflect true biological relationships. This is also called overfitting, where the model captures noise instead of true biological signals (Zhao et al., 2025). The problem is further compounded by the presence of highly correlated variables, as many metabolites are interconnected within biochemical pathways, leading to redundancy in the data (Shomorony et al., 2020; Wang et al., 2022).

In addition to high dimensionality, metabolomics datasets are often characterized by substantial levels of noise and variability. Technical variation can arise from differences in sample collection, storage, and preparation, as well as from analytical factors such as instrument performance and batch effects, meaning unwanted technical variations in the data caused by non-biological factors (Deng et al., 2024). At the same time, biological variability between individuals, driven by factors such as age, diet, lifestyle, and environmental exposure, can introduce additional complexity. These sources of variation can obscure true biological signals and complicate the identification of robust biomarkers.

Another important challenge is the prevalence of missing data and non-normal data distributions in metabolomics datasets. Missing values frequently occur due to metabolites being present below the detection limit or as a result of technical issues

during data acquisition (Idkowiak et al., 2025). Furthermore, metabolite concentrations often exhibit skewed distributions and heteroscedasticity, violating the assumptions of many classical statistical methods. This necessitates the use of appropriate data transformation, normalization, and imputation strategies, each of which can influence the outcome of the analysis.

Finally, metabolomics studies are particularly susceptible to false discoveries. The simultaneous testing of a large number of metabolites increases the risk of identifying statistically significant results by chance alone. Without appropriate correction for multiple testing and careful validation, this can lead to the reporting of biomarkers that lack reproducibility and biological relevance. Additionally, limited sample sizes, which are common in metabolomics due to the cost and complexity of data acquisition, reduce statistical power and further increase the risk of unreliable findings (Mendez et al., 2019).

Taken together, these challenges highlight that while metabolomics holds great promise for biomarker discovery, its successful application depends critically on careful data handling and rigorous statistical analysis. Understanding the limitations and complexities of metabolomics datasets is therefore essential for ensuring that identified biomarkers are both statistically robust and biologically meaningful.

## 1.2 Machine Learning in Metabolomics

Machine learning (ML) has become an increasingly important tool for the analysis of complex biological data, including metabolomics. In essence, ML refers to a class of computational algorithms designed to identify patterns in data and make predictions or decisions without being explicitly programmed for a specific task (Zhao et al., 2025). By learning from existing datasets, ML models can uncover relationships between variables and outcomes, making them particularly useful in data-rich fields such as systems biology.

In metabolomics, ML methods are commonly applied to tasks such as classification, regression and feature selection. Classification models aim to distinguish between biological groups, for example diseased and healthy individuals, based on their metabolic profiles, while regression models are used to predict continuous outcomes such as metabolite concentrations. These approaches rely on “features,” which represent measurable characteristics derived from the data, such as mass-to-charge ratios or metabolite intensities (Zhao et al., 2025). The ability of ML algorithms to handle large numbers of variables simultaneously makes them well suited for metabolomics datasets, where metabolites are often interconnected through complex biochemical pathways.

The development of ML models involves several key steps, including data preprocessing, model training, and performance evaluation. Preprocessing techniques such as normalization, scaling, and transformation are commonly applied to ensure that the data is suitable for analysis and comparable across samples (Ren et al., 2015; Zhao et al., 2025). The dataset is typically divided into subsets for training and evaluation, allowing the model to learn patterns from one portion of the data and assess its predictive performance on another. Model performance is commonly evaluated using evaluation metrics, which provide quantitative measures of how well the model distinguishes between different classes (Sun et al., 2024).

Deep learning (DL) is a subset of ML that utilizes artificial neural networks with multiple layers to model complex relationships in data (Sha et al., 2021; Zhao et al., 2025). These models are designed to automatically learn hierarchical representations of input data, allowing them to capture both simple and highly complex patterns. In contrast to traditional MLapproaches, DL can reduce the need for feature engineering, as relevant features are learned directly from the data during training.

The application of DL has expanded rapidly in recent years, including within metabolomics and other omics disciplines. Its ability to model nonlinear relationships and process large-scale datasets has contributed to its growing use in biological data analysis. Furthermore, DL approaches can integrate multiple types of data and capture interactions between variables, offering a flexible framework for studying complex biological systems.

Overall, ML and DL provide powerful and versatile approaches for analysing metabolomics data. Their capacity to identify patterns, model relationships, and support predictive analysis has made them valuable tools in modern biomedical research, particularly in studies aimed at understanding disease mechanisms and identifying potential biomarkers.

## 1.3 Aim

The aim of this literature study was to investigate the challenges and opportunities associated with metabolomics-based biomarker discovery using ML, with a particular focus on the statistical and biological characteristics of metabolomics datasets. Although metabolomics combined with ML has shown considerable potential for identifying disease-associated biomarkers, the high complexity, variability and dimensionality of metabolomics data raise important concerns

regarding the reliability, reproducibility and biological relevance of reported findings. This study therefore aimed to critically evaluate how the inherent properties of metabolomics datasets, together with commonly used statistical and ML approaches, influence the outcomes and interpretation of biomarker discovery studies.

# 2. Method

## 2.1 Methodology

This study was conducted as a literature-based review with the aim of identifying and critically evaluating challenges associated with the use of metabolomics in biomarker discovery, with a particular focus on ML-based statistical analysis. Relevant scientific literature was collected, reviewed and synthesized to provide an overview of current knowledge in the field.

## 2.2 Literature Search Strategy

A structured literature search was performed to identify relevant peer-reviewed articles. Google Scholar was used as a scientific database. It was chosen because it provides broad interdisciplinary coverage and articles from multiple scientific publishers, which was suitable for identifying studies related to ML in metabolomics. Since the study was a literature review with a specific focus rather than a systematic review, one comprehensive database was considered sufficient for the scope and timeframe of the project. The search focused on studies related to metabolomics, biomarker discovery, and statistical or computational analysis methods.

The following keywords and combinations of keywords were used during the search process:

- “metabolomics”
- “biomarker discovery”
- “metabolomics data analysis”
- “statistical challenges metabolomics”
- “high dimensional data metabolomics”
- “overfitting metabolomics”
- “false discoveries metabolomics”
- “machine learning metabolomics"
- “PLS-DA”
- “random forest”
- “support vector machine”
- “deep learning metabolomics”
- “XGBoost”
- “SHAP metabolomics"
- “cross-validation bias metabolomics"

- “analytical techniques metabolomics”

## 2.3 Selection Criteria

Articles were selected based on their relevance to the research aim. Priority was given to peer-reviewed studies that addressed statistical methods, data characteristics, and challenges in metabolomics-based biomarker discovery. Both methodological papers and applied studies were included to provide a comprehensive understanding of the topic.

Studies were excluded if they lacked relevance to metabolomics, did not address biomarker discovery, or did not include a statistical or analytical perspective. Initially, around 52 articles were identified, but 30 articles remained after applying the inclusion and exclusion criteria.

## 2.4 Use of AI Tools

ChatGPT (OpenAI) was used as a supportive tool during the literature search process to assist in identifying relevant keywords, refining search queries, and suggesting potentially relevant research topics and articles. However, all sources included in this study were independently verified by consulting original peer-reviewed publications, and no references were included without manual evaluation.

ChatGPT was also used as a supporting tool for creating the figures included in the text.

## 2.5 Data Analysis

The selected articles were systematically reviewed and key findings related to statistical challenges in metabolomics were extracted. Particular attention was given to recurring themes such as high dimensionality, small sample sizes, data variability, model validation, and risk of false discoveries.

# 3. Results

## 3.1 Metabolomics Data and Preprocessing

### 3.1.1 Data Acquisition and Characteristics

Metabolites are quantified using advanced chemical analytical techniques. Among the most commonly used techniques are liquid chromatography-mass spectrometry (LC-MS), gas chromatography-mass spectrometry (GC-MS) and nuclear magnetic resonance (NMR) spectroscopy. These techniques differ in their analytical capabilities, even though they are complementary. MS offers sensitive and unbiased detection of many metabolites, meaning that the method can detect a wide range of metabolites without being specifically targeted toward only certain metabolite classes (Zhao et al., 2025). LC-MS and GC-MS are widely used due to their high sensitivity and ability to detect low-abundance metabolites (Xiao et al., 2012).

Metabolites are first separated using chromatographic techniques and subsequently detected based on their mass-to-charge ratios (Xiao et al., 2012; Zhao et al., 2025). High-resolution mass spectrometry instruments are commonly used for untargeted metabolomics, such as quadrupole time-of-flight and Orbitrap systems, in which as many metabolites as possible are measured in a biological sample (Zhao et al., 2025).

NMR spectroscopy is also used for the detection of metabolites. In this method, nuclei with a magnetic spin are placed inside a magnetic field, and the different energy levels generated can be observed using radiofrequency waves (Griffin et al., 2011). NMR spectroscopy is a non-destructive and highly reproducible method for metabolite identification and quantification, because the sample is not in contact with the detector and requires no chemical derivation (Ren et al., 2015). However, it offers a lower sensitivity compared to MS since only metabolites in medium or high abundance can be detected.

Metabolomics data is characterized by high dimensionality, where the number of measured variables far exceeds the number of samples, since modern analytical techniques can detect thousands of metabolites (Worley & Powers, 2013; Labory et al., 2024). This imbalance leads to what is commonly referred to as the “curse of dimensionality”, where the presence of many variables increases the complexity of the data. This is a major drawback of metabolomics, since the cost of data collection is also financially high and the number of participants is limited (Labory et al.,

2024). Furthermore, many metabolites are correlated and, in the context of biomarker discovery, not directly related to the disease in question, which leads to noisy, redundant information (Wang et al., 2022; Labory et al., 2024). This poses a great challenge for statistical analysis, as many methods assume independence between variables.

As a result, metabolomics data often require preprocessing steps such as dimensionality reduction, feature selection or transformation to make the data more suitable for statistical analysis. Deconvolution, library-based identification and alignment is then used to convert the spectral data to presence of metabolites in each sample, which links the raw data to statistical analysis (Anwardeen et al., 2023). As Anwardeen et al. (2023) put it, "statistical machine learning-based methods are geared towards the identification of unknowns based on feature similarity with the knowns".

In addition to high dimensionality, missing values is a common feature of metabolomics datasets (Anwardeen et al., 2023; Idkowiak et al., 2025). This is due to metabolite concentrations falling below the limit of detection, technical issues during peak detection or alignment, and variability in sample preparation or instrument performance. Although metabolomics data has been subjected to general statistical techniques for multiple imputation, the non-random nature has necessitated the development of more tailored approaches. One such approach is MetabImpute, an R package that recognizes if the missingness is random or not random (Anwardeen et al., 2023).

### 3.1.2 Data Preprocessing

Data preprocessing is a critical step in metabolomics workflows, as it also influences the validity, interpretability and reproducibility of downstream statistical analyses. The preprocessing steps are normalization, transformation, imputation and scaling, and they are not merely technical adjustments. Rather, they fundamentally shape the statistical structure of the data and, in turn, the outcomes of biomarker discovery efforts.

Because preprocessing aims to reduce unwanted technical variation while preserving biologically meaningful differences, normalization is commonly applied to correct for systematic variation between samples. These differences could be due to differences in sample concentration, dilution effects or analytical variability (Ren et al., 2015; Zhao et al., 2025). By scaling each sample (row) by a sample specific constant factor, the intensities of metabolites become more comparable across

observations (Ren et al., 2015). Also, to standardize the dataset, z-score normalization can be applied to scale the data to a standard normal distribution (Sun et al., 2024). The researchers distinguish between two different normalization methods; pre-acquisition normalization strategies where standardization is based on sample volume, mass or metabolite concentration, and post-acquisition statistical approaches, including sum, median and probabilistic quotient normalization. (Idkowiak et al., 2025).

Although normalization can mitigate or even correct datasets with analytical variation, pre-analytical variation, such as sample collection or handling, is harder to account for (Idkowiak et al., 2015). Also, the distinction between technical and biological variation is not always clear, and inappropriate normalization may either fail to remove confounding variability or inadvertently eliminate relevant biological differences. This should be accounted for before comparing metabolomes.

Due to missing values and outliers, metabolomics data deviate from a symmetric Gaussian distribution and exhibit right- or left-skewed patterns (Anwardeen et al., 2023). Because the variance depends on the mean, this deviation violate assumptions of many parametric statistical methods. Log transformation is commonly used to stabilize the variance and reduce skewness (Idkowiak et al., 2025). More advanced approaches, such as rank-based inverse normal transformation, have been applied to address the non-Gaussian distributions (Shomorony et al., 2020). Also, the order in which preprocessing steps are applied can influence statistical bias, since performing transformation prior to covariate adjustment has been shown to reduce bias compared to alternative sequences (Pain et al., 2018).

More advanced methods to deal with missing values are k-nearest neighbors (kNN) and random forest-based imputation. kNN is widely used to estimate missing values based on the similarity between samples (Frölich et al., 2024; Idkowiak et al., 2025). In this approach, the “k” most similar observations (neighbors) are identified using a distance metric, such as Euclidean distance, calculated from the available features. The missing value is then imputed using a function of the corresponding values in the neighboring samples, commonly the mean or a weighted average. This method assumes that similar samples will exhibit similar metabolite profiles and can therefore provide biologically plausible estimates.

Random forest-based imputation is based on an ensemble of decision trees to predict missing values (Idkowiak et al., 2025). Each variable with missing values is treated as a response variable and predicted using the remaining variables as predictors (Labory et al., 2024). By iteratively refining these predictions, the

algorithm constructs multiple decision trees and aggregate their outputs, which captures complex interactions between features/metabolites (Anwardeen et al., 2023). It can handle both continuous and categorical data and is robust to outliers and normalization techniques.

Finally, scaling is another essential preprocessing step, in which a uniform scale is found for all variables (Idkowiak et al., 2025). However, a major disadvantage of data scaling is its ability to amplify instrumental noise (Worley & Powers, 2013). This is particularly challenging for multivariate analysis, since principal component analysis (PCA) and partial least squares (PLS) are sensitive to this kind of data characteristic.

## 3.2 Feature Engineering in Metabolomics

### 3.2.1 Feature Selection

Since metabolomics data is high-dimensional, dimensionality reduction is essential for the data to be analyzed correctly. Dimensionality reduction is categorized by feature selection and feature extraction. Feature selection methods aim to identify a subset of the original variables that maximize the performance of a predictive model while minimizing redundancy and noise (Labory et al., 2024). It results in improved learning performance, which is recognized by increased classification accuracy, reduced computational power and enhanced model interpretability. The feature selection approach is classified into three groups: supervised, unsupervised and semi-supervised.

Filter, wrapper and embedded methods are commonly used for feature selection (Labory et al., 2024). Filter methods evaluate the relevance of individual features independently of any specific ML algorithm, using statistical criteria such as variance, correlation and hypothesis testing. For example, Kolmogorov-Smirnov (KS) is a supervised filter model that assesses whether the distribution of a metabolite differs significantly between groups. The main advantage of filter methods is their computational efficiency and independence from model-specific bias. However, traditional feature selection methods fail to capture the complex interactions and correlations between metabolites (Sun et al., 2024). Researchers have therefore proposed the Mutual Information-Random Forest (MI-RF) method, in which mutual information metrics are combined with the feature importance scored from random forest.

Wrapper methods evaluate subsets of features based on the predictive performance of a specific ML model (Labory et al., 2024). An example is the Boruta algorithm, which builds upon random forest classification to iteratively tested and assessed using model performance metrics. By accounting for interactions between features and model behavior, wrapper methods have the ability to achieve higher predictive accuracy compared to filter methods.

Embedded methods integrate feature selection directly into the model training process (Labory et al., 2024). Techniques such as least absolute shrinkage and selection operator (LASSO) regression impose a penalty on model coefficients, shrinking some coefficients to zero and selecting a subset of relevant features (Ge et al., 2025). LASSO performed better in predicting model selection and identifying predictors, compared to classical statistical methods. However, linear approaches such as LASSO may struggle to capture nonlinear relationships, which can even exclude biologically relevant metabolites (Sun et al., 2024). Additionally, the stability of selected features can be sensitive to small changes in the data, raising concerns about reproducibility in biomarker discovery.

### 3.2.2 Feature Extraction

Feature extraction aims to lower the dimensionality of high-dimensional features, thereby constructing a new feature space with linear or nonlinear combinations of the original features (Labory et al., 2024). Among linear feature extraction methods, PCA remains one of the most widely used unsupervised approaches (Idkowiak et al., 2025). PCA performs a linear transformation of the data into orthogonal components that capture the maximum variance, thereby enabling visualization of the global structure and detection of clustering patterns. It is a computationally efficient method that scales well with large datasets (Idkowiak et al., 2025), and it can be used as checkpoint during quality control to screen for outliers (Anwardeen et al., 2023). However, because it maximizes total variance rather than class separation, it only reveals group structure when between-group variation is much higher than within-group variation (Worley & Powers, 2013). Therefore, class separation in PCA scores plot could be misleading if it is not a function of the algorithm but occurs due to sample preparation problems, inappropriate data preprocessing or experimental bias. Consequently, while PCA is valuable for exploratory analysis, its direct utility for identifying discriminative biomarkers is limited.

To address these limitations, supervised linear methods such as partial least squares discriminant analysis (PLS-DA) and its extension orthogonal PLS-DA (OPLS-DA)

are commonly employed. While PLS-DA uses predictor variable X to optimally explain the response variable Y, OPLS-DA divides the variance into parts that predict the experimental groups and parts that arises because of noise (Anwardeen et al., 2023). But as with other models, both PLS-DA and OPLS-DA are associated with substantial statistical challenges. A major concern is their tendency to overfit (Gromski et al., 2015; Ge et al., 2025). Because these models are optimized to maximize class separation, they can produce apparent discrimination even in random or noisy data, leading to misleading separation, which stresses the fact that "visual separation alone is not evidence of a valid model" (Gromski et al., 2015). Also, inadequate validation practices can result in overly optimistic estimates of predictive performance, so metrics such as $R^2$ (explained variance) and $Q^2$ (predictive ability) may not reliably reflect true model generalizability. Even variable importance in projection (VIP) scores can be unstable and sensitive to small changes in data.

Non-linear extraction methods have also gained increasing attention in metabolomics. Techniques such as kernel-based methods (Labory et al., 2024; Mendez et al., 2019) and manifold learning approaches (Idkowiak et al., 2025) are capable of capturing complex, non-linear relationships between metabolites that linear methods cannot represent, especially given that metabolite interactions are often non-linear and highly independent. However, this increased complexity comes at the cost of reduced interpretability, making it difficult to link extracted features to specific metabolites or biological pathways.

## 3.3 Statistical Learning in Metabolomics

### 3.3.1 Univariate and Multivariate Analysis

Univariate and multivariate analyses are fundamental statistical approaches in metabolomics and are widely used for biomarker discovery, pattern recognition and interpretation of metabolic alterations associated with disease states (Idkowiak et al., 2025). Univariate analysis examines one metabolite at a time to determine whether its abundance differs significantly between experimental groups (figure 1). Commonly applied methods include the Student's t-test, analysis of variance (ANOVA), Mann-Whitney U test and fold change analysis. These methods are frequently used as an initial screening step to identify potentially discriminative metabolites between healthy and diseased cohorts. The main advantage of univariate analysis lies in its simplicity and interpretability, since each metabolite is independently evaluated for statistical significance (Saccenti et al., 2014). Furthermore, univariate methods are computationally efficient and relatively

straightforward to implement, making them useful for preliminary biomarker discovery workflows.

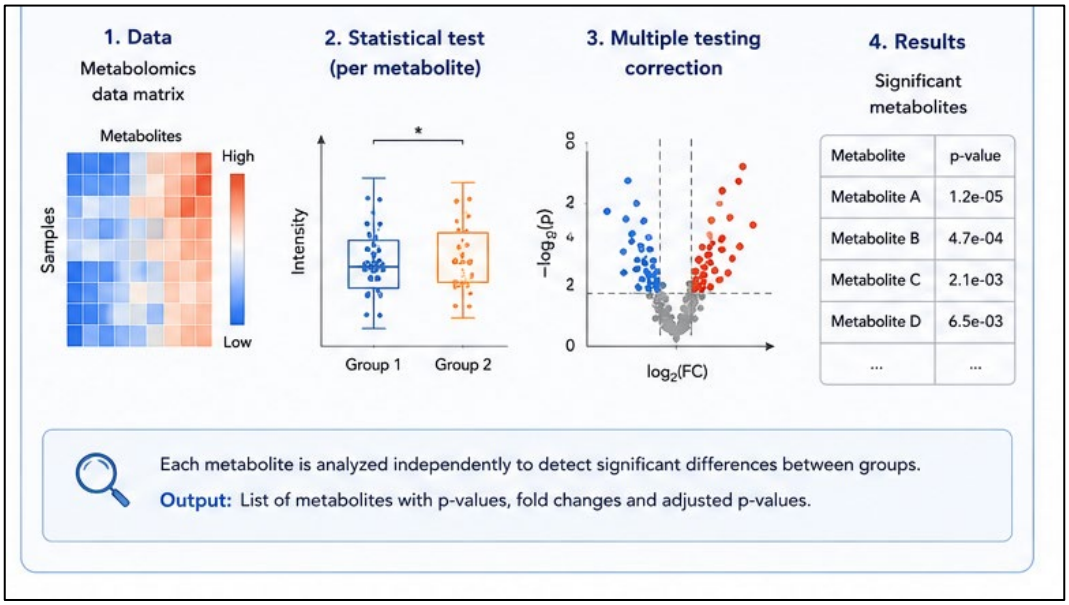



*Figure 1. Workflow of univariate analysis of metabolomics. The analysis starts with data generation and proceeds with statistical testing of each metabolite. To decrease the risk of false-positive finding, multiple testing correction is applied before analyzing the significant metabolites. Picture created with assistance of ChatGPT.*

Despite these advantages, univariate analysis has important limitations when applied to metabolomics datasets. Since metabolites are evaluated independently, biological relationships and correlations between metabolites are ignored. This represents a significant drawback because, as mentioned before, metabolites function within highly interconnected metabolic pathways rather than as isolated variables. In addition, metabolomics datasets often contain thousands of metabolites, increasing the risk of false-positive findings due to multiple hypothesis testing, which occurs when thousands of statistical tests are conducted simultaneously (Broadhurst & Kell, 2006). Also, testing each metabolite independently leads to inflated type I error rates, and with a typical threshold of $p < 0.05$, hundreds of false positives are expected. Therefore, it is recommended to apply multiple testing corrections, such as Bonferroni correction, and control the false discovery rate (FDR) rather than individual error rates (Broadhurst & Kell, 2006).

In contrast, multivariate analysis evaluates several metabolites simultaneously and is therefore better suited for capturing the complexity of metabolic systems (figure 2). Multivariate methods are particularly advantageous in metabolomics because

they can account for metabolite correlations, reduce dimensionality and identify hidden structures within high-dimensional datasets (Idkowiak et al., 2025). Commonly applied multivariate approaches include (PCA), (PLS-DA) and (OPLS-DA).

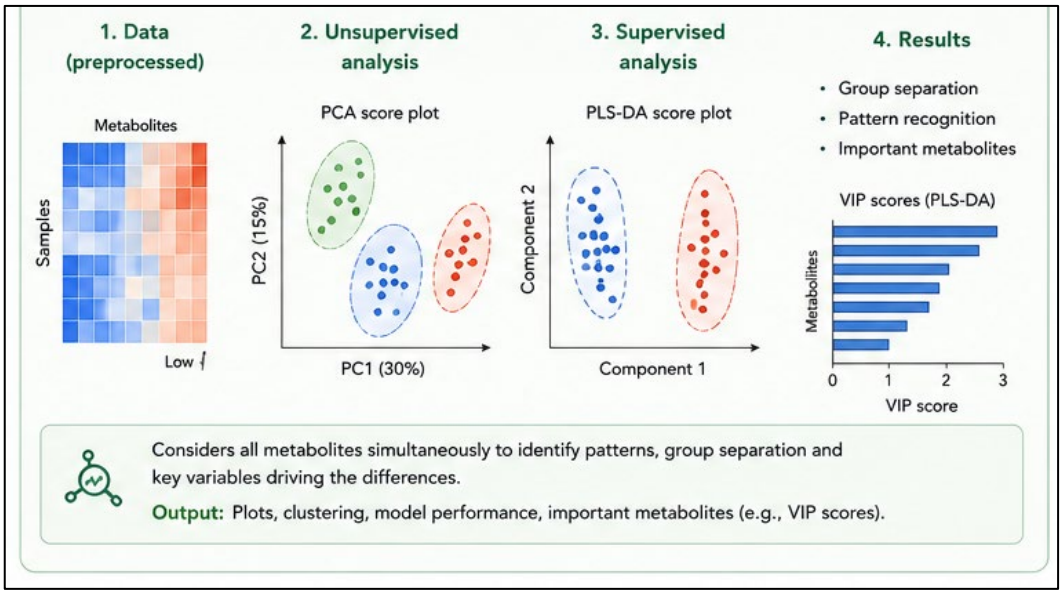



*Figure 2. Workflow of multivariate analysis, in which several metabolites are analyzed simultaneously. After preprocessing of the data, unsupervised analysis such as PCA is applied, followed by supervised analysis such as PLS-DA. The results generated show group separation, pattern recognition and significant metabolites. Picture created with assistance of ChatGPT.*

Multivariate and ML overlap conceptually because both analyze variables simultaneously to identify patterns in complex datasets. ML methods build upon the principles of methods such as PCA, PLS-DA and OPLS-DA by applying computational algorithms capable of learning complex relationships within high-dimensional data for prediction and classification purposes (Zhao et al., 2025). The distinction between the fields is therefore not absolute, as both aim to extract meaningful biological information from multidimensional datasets. Because of this, the abovementioned methods are described in more detail in the coming section.

## 3.3.2 Machine Learning Methods

### *3.3.2.1 Unsupervised Machine Learning Methods*

Unsupervised ML methods are used in metabolomics for exploratory data analysis, particularly when the underlying structure of the data is unknown (Idkowiak et al., 2025). Methods such as PCA, hierarchial clustering and k-means clustering are

commonly applied to reduce dimensionality and identify natural groupings in high-dimensional metabolite profiles.

As mentioned before, PCA is one of the most established techniques and allows for intuitive visualization of metabolic differences between experimental groups, making the technique useful as an initial quality control step to assess whether biological separation exists before any supervised modelling is applied (Wang et al., 2022). However, PCA is inherently linear and therefore struggles to capture non-linear metabolic interactions, which are common in biological systems. As a result, subtle but biologically relevant patterns may remain hidden.

k-means clustering is another unsupervised method that assigns samples to clusters based on distance to centroids (Ren et al., 2015). While computationally efficient, it assumes spherical and equally sized clusters, an assumption that rarely holds in metabolomics datasets that are characterized by heterogeneity, skewed distributions and outliers (Idkowiak et al., 2025). Consequently, cluster assignments may become unstable or biologically non-interpretable. Hierarchial clustering, although more flexible, is highly sensitive to distance metrics and linkage criteria, which can significantly alter clustering outcomes.

More advanced unsupervised approaches such as variational autoencoders (VAEs) address some of these limitations by learning non-linear latent representations using neural networks (Wang et al., 2022). VAEs can capture more complex structures than PCA, but they require large datasets, careful tuning and often lack interpretability, which limits their use in biomarker discovery.

#### *3.3.2.2 Simple Supervised Machine Learning Methods*

Simple supervised ML methods incorporate outcome variables and are therefore more directly suited for biomarker discovery in metabolomics. Common examples include PLS-DA, OPLS-DA and logistic regression (LR).

PLS-DA is particularly popular in metabolomics because it can handle high-dimensional and highly collinear data by projecting predictors into latent components that maximize covariance with class labels (Gromski et al., 2015; Anwardeen et al., 2023). This makes it well suited for datasets where the number of variables is much higher than the number of samples. For instance, PLS-DA is frequently used in disease classification studies to separate control and patient groups based on metabolite profiles (Wang et al., 2022). A major limitation is its tendency to overfit, especially in small sample sizes typical of metabolomics studies. Without proper cross-validation or permutation testing, PLS-DA models

may appear to perform well even when the predictive power is limited (Gromski et al., 2015).

OPLS-DA improves interpretability by separating predictive variation from orthogonal (non-predictive) variation, which can help identify discriminative metabolites more clearly (Anwardeen et al., 2023). Nevertheless, it still inherits many of the same overfitting risks and linear assumptions as PLS-DA.

LR, another simple supervised method, provides a more interpretable and probabilistic framework for classification tasks (Ranganathan et al., 2017). It predicts the probability that a sample belongs to a specific class by modeling the relationship between metabolite variables (X) and a binary outcome (Y). While it is highly interpretable and provides probabilities that are useful in clinical decision-making, it again assumes a linear relationship between predictors and outcome.

#### *3.3.2.3 Complex Supervised Machine Learning Methods*

Complex supervised ML methods such as random forest (RF), support vector machine (SVM) and gradient boosting algorithms (e.g. XGBoost) offer greater flexibility in modeling non-linear relationships in metabolomics data.

RF is an ensemble learning method that constructs multiple decision trees and aggregates their predictions (Sun et al., 2024). This allows RF to capture non-linear interactions and handle noisy, high-dimensional data effectively. In metabolomics, RF has demonstrated strong performance in disease classification tasks (Ge et al., 2025). Additionally, RF is relatively robust to missing values and does not require strict feature scaling (Anwardeen et al., 2023). However, interpretability remains limited, and model performance can degrade without careful tuning of tree depth, number of tress and feature selection strategies. Furthermore, in some metabolomics applications such as prostate cancer metabolite prediction, RF and XGBoost have shown suboptimal performance, suggesting that their effectiveness is highly dataset-dependent (Sun et al., 2024).

SVMs are powerful classifiers that can model complex decision boundaries using kernel functions (Mendez et al., 2019). This makes them particularly useful in metabolomics where class separation is not linearly separable. Yet, SVMs are highly sensitive to parameter selection and often require extensive preprocessing and feature scaling (Sun et al., 2024). They also do not inherently provide probabilistic outputs, which can limit interpretability in biomarker studies.

XGBoost further improves predictive performance by sequentially correcting error from previous models (Labory et al., 2024). Although several studies have shown its strong performance in structured biological datasets, it is highly prone to overfitting if not carefully regularized (Wang et al., 2022; Sun et al., 2024).

Overall, complex supervised methods improve predictive power by capturing non-linearities, but this comes at the cost of interpretability, computational complexity and increased risk of overfitting.

A simplified workflow of ML is shown in figure 3.

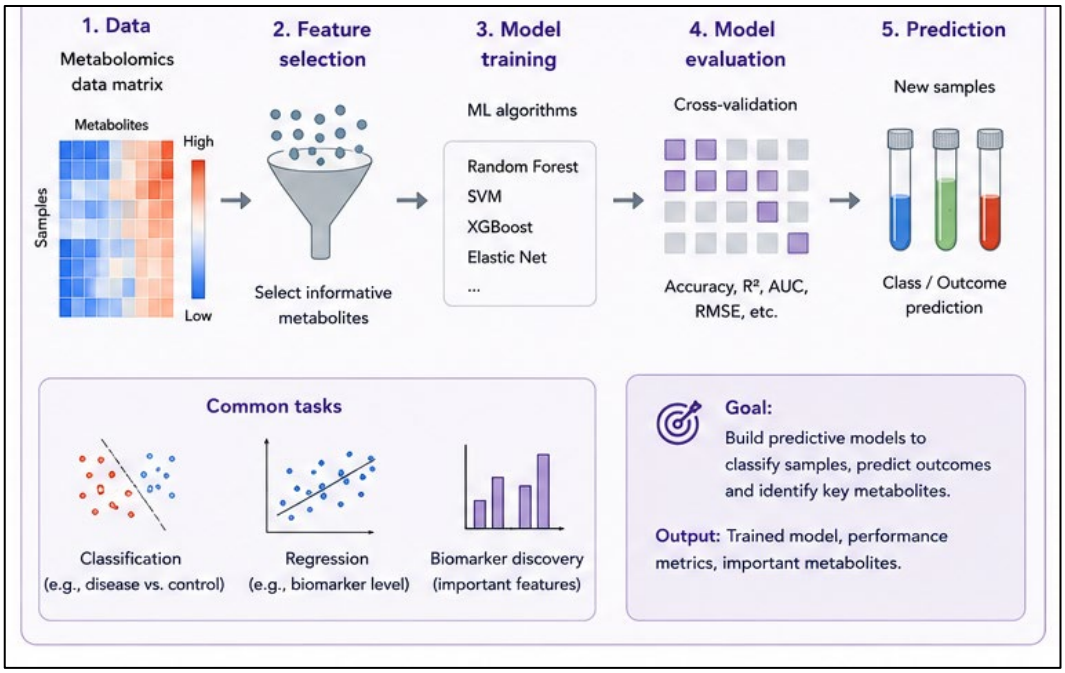



*Figure 3. Workflow of ML in metabolomics. First, data is generated and features are selected. The ML model is then trained and evaluated using cross-validation (see section 3.4), before prediction and eventual clinical application can be made. Picture generated with assistance of ChatGPT.*

### *3.3.2.4 The Bias-Variance Tradeoff: Simple Versus Complex Machine Learning Methods*

Because of differences in variance and bias, there is always a tradeoff between simple and complex ML methods. Variance describes how sensitive a model is to changes in the training data (James et al., 2013). When a model is trained, a dataset is used to estimate a function $\hat{f}$, which is the model's approximation of the true relationship f. If different training sets are collected, the model would not be exactly the same, and variance measures how much the model changes between these different datasets. Low variance indicates that the model stays relatively stable even

if the training data changes a little, and high variance indicates that small changes in the training data lead to very different models.

More flexible or complex models, such as RF and SVM, can fit the training data very closely (Mendez et al., 2019; Sun et al., 2024), which means they capture many patterns in the data. This ability makes them learn random noise and irrelevant variation, making the models highly dependent on the specific training dataset. In other words, these complex models have high variance. High variance is therefore closely related to overfitting, since an overfitted model performs extremely well on the training data but poorly on new unseen data (Zhao et al., 2025).

In contrast, bias is the error that occurs when a model is too simple to capture the true relationship in the data (James et al., 2013). Even though biological systems are complex, statistical models simplify reality in order to make predictions. When a model makes assumptions that are too simple, it cannot fully represent the true underlying patterns, and this difference between the true relationship and the simplified model is called bias. LR, for example, has high bias since the model is overly simple and important patterns are missed (Ranganathan et al., 2017). High bias leads to underfitting, in which the model performs poorly on both training and test data and fails to learn the important structure of the data (Zhao et al., 2025) (figure 4).

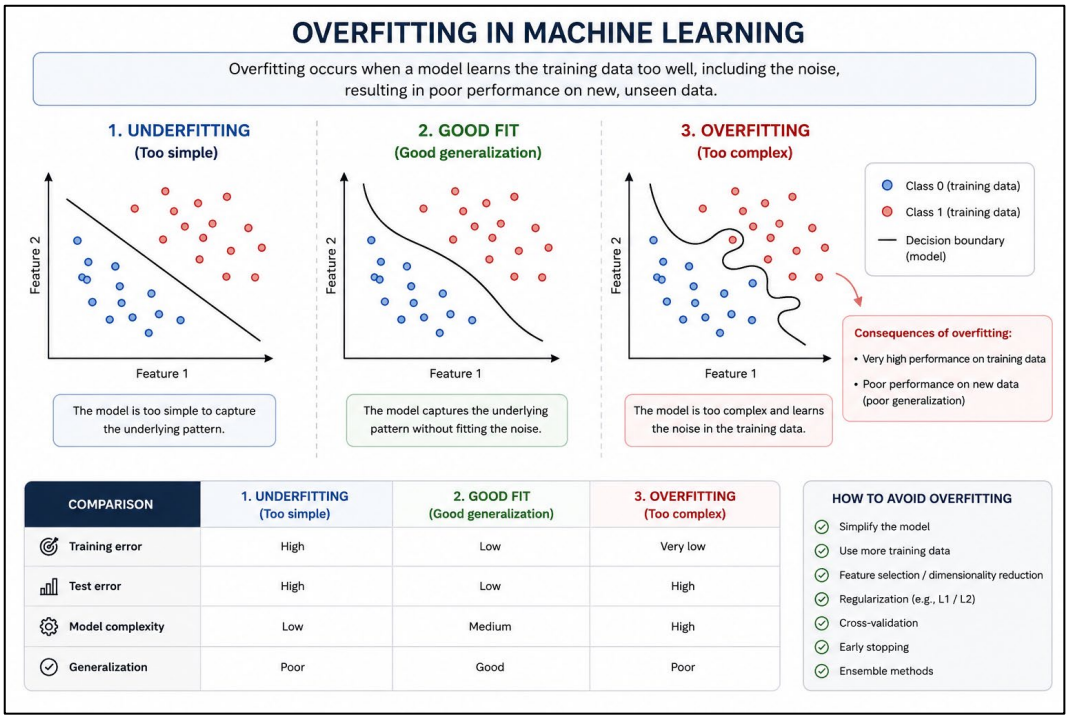



*Figure 4. Graphical representation of underfitting (when the model is too simple), good fit (when the model has good generalization) and overfitting (when the model is too complex). Underfitting makes the model unable to capture the complex relationships between metabolites, while overfitting*

*leads to results based on noise rather than true biological signals. Picture generated with assistance of ChatGPT.*

### 3.3.3 Deep Learning Approaches

Deep learning (DL) approaches have emerged as powerful alternatives to classical ML methods in metabolomics, which is due to their capacity to model complex, nonlinear relationships and to process high-dimensional data without extensive feature engineering (Zhao et al., 2025). These models solve problems such as image recognition, biochemical pattern discovery and natural language processing (figure 5). However, a fundamental issue in DL models, such as neural networks (NN), is that the results are considered as uninterpretable "black boxes", rendering the information less meaningful since it cannot be utilized in a realistic way (Rudin et al., 2019; Watson et al., 2019). To solve this problem, researchers have developed several DL models with striking features.

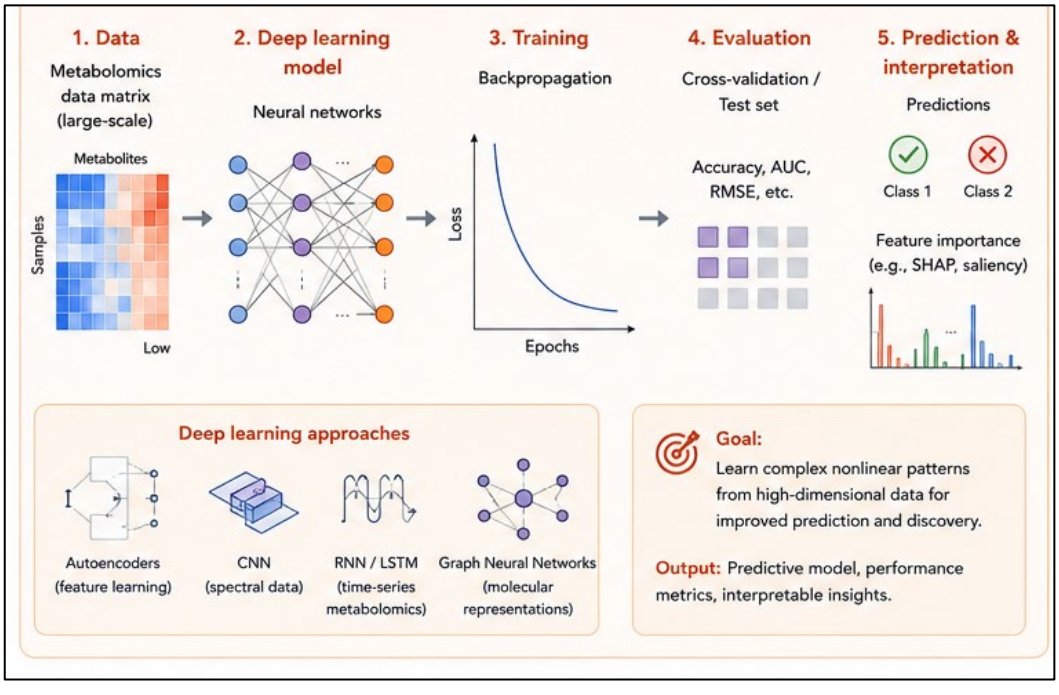



*Figure 5. Workflow of DL. DL models are constructed using artificial neural networks, which are able to capture the complex relationships between metabolites and process high-dimensional data. After model training, the model is evaluated using cross-validation (see section 3.4) and predictions and interpretations can be made. However, as is often the case, the results of DL models are uninterpretable "black boxes". Picture generated with assistance of ChatGPT.*

In order to reveal biomarkers for Parkinson's disease, Zhang et al. (2023) developed a model called "Classification and Ranking Analysis using Neural networks generates Knowledge from Mass Spectrometry" (CRANK-MS), which has several

built-in features. The integrated model parameters avoid the need for preselecting chemical features while analyzing the high-dimensional metabolomics data. The model also incorporates Shapely Additive exPlanations (SHAP) that retrospectively extract chemical features that contribute the most to an accurate model prediction. Finally, CRANK-MS was benchmarked against five ML methods to compare their diagnostic performance and further verify if the selected chemical features were significant.

The CRANK-MS exemplified the ability of DL approaches to process high-dimensional data by being highly tolerant of more than 1500 chemical features that did not contribute to disease prediction (Zhang et al., 2023). Also, across all metrics investigated, which took not more than 10 minutes to obtain, the NN model had the highest diagnostic performance.

The need for numerous parameters to describe DL models have also been circumvented by the development of an evolutionary algorithm (Sha et al., 2021). SMILE, a DL model, uses this algorithm to interpret predictive models, provide explanations for its own decision making and identify key metabolites and their interactions. The algorithm maintains some candidate solutions to a problem, often generated randomly. With each new generation, better solutions are produced probabilistically, and small changes using biologically inspired operators, such as mutation and crossover, are introduced. Using only few metabolites, SMILE was able to find predictive models with high accuracy (Sha et al., 2021).

DL models have also been developed in which the transformer and convolutional neural networks were combined, where sequential and grid-like data are incorporated, respectively (Sun et al., 2024). This hybrid model, TransConvNet, was used for the classification of prostate cancer metabolomics data, and it was based on an attention mechanism. By focusing on critical aspects of the input data, the mechanism allows the model to capture key features and complex relationships. The ability of the model to capture nonlinear relationships was attributed to the interaction between global and local correlations. As with CRANK-MS, TransConvNet outperformed other algorithms in key evaluation metrics, through five-fold cross-validation experiments.

The general notion is that training DL models from scratch requires significantly larger datasets, which limit its use in datasets with small sample sizes or simpler underlying mechanisms (Zhao et al., 2025). However, a DL method called “Metabolite Response Predictor using Coupled Multilayer Perceptrons” (McMLP) was developed to predict endpoint metabolite concentrations based on baseline microbial compositions (Wang et al., 2025). Instead of fine-tuning

hyperparameters, MLP was overparametrized using a large and fixed number of layers and a hidden layer dimension, yielding better performance and, remarkably, decreased tendency to overfit. More importantly, the performance of McMLP was better than RF at small training sample sizes, and closer to RF at large training sample sizes. This contradicts the traditional notion that DL methods are prone to overfit at small sample sizes.

Although DL methods provide a great alternative to traditional ML methods, it is important to note that they are still being developed and not much information about them is obtained. The limitation to DL approaches is the computational cost of running the algorithm and the collection of a large number of independent runs (Sha et al., 2021). Without these algorithms, the DL models tend to overfit if the sample size is small (Zhao et al., 2025), and the results generated are of less employability in the biomedical context (Labory et al., 2024).

## 3.4 Model Evaluation and Validation

### 3.4.1 Evaluation Metrics

The evaluation of ML and DL models in metabolomics is critical for assessing the predictive performance of the models and serves as a guidance for biomarker discovery. However, the selection and interpretation of the evaluation metrics can significantly influence how model performance is perceived and may introduce bias.

The receiver operating characteristic (ROC) curve is a plot of the true positive rate (sensitivity) versus the false positive rate (1 − specificity) (Zhang et al., 2023; Sun et al., 2024). ROC curves are widely used to evaluate the classification performance of predictive models. A perfect classifier has a sensitivity of 1 and a false positive rate of 0, indicating perfect discrimination between classes. The area under the ROC curve (AUC) provides a summary measure of model performance and reflects the ability of a model or biomarker to distinguish between experimental groups (Sha et al., 2021; Anwardeen et al., 2023). Higher AUC values indicate better classification performance.

Matthew's correlation coefficient (MCC) has been proposed as a more informative and reliable metric for the evaluation of binary classification accuracy, as it considers all values in the confusion matrix (true negative, false negative, true positive, and false positive) (Chicco et al. i Zhang et al., 2023). Therefore, MCC is less biased in the presence of class imbalance, providing a more robust evaluation compared to accuracy alone (Zhang et al., 2023).

Accuracy, sensitivity and specificity are common evaluation metrics. Accuracy reflects the proportion of correct predictions, while sensitivity and specificity measure the ability to correctly identify positive and negative cases, respectively (Sun et al., 2024). However, these metrics can be misleading in imbalanced datasets, where a model may achieve high accuracy by predominantly predicting the majority class (Zhao et al., 2025). It is therefore important to include several performance metrics, such as precision, recall or the F1 score, which is the weighted average of precision and recall (Sha et al., 2021; Zhao et al., 2025). F1 score of 1 indicates the best model performance, while 0 indicates the worst performance (Sha et al., 2021).

A major limitation of performance metrics, especially high-performance metrics, is that they do not necessarily indicate biological validity. Near-perfect classification performance, including AUC values of 1, even in relatively small datasets, raises concerns about overfitting and limited generalizability (Ge et al., 2025; Deng et al., 2024). This suggests that strong numerical performance may reflect dataset-specific patterns rather than true underlying biological differences. Finally, evaluation metrics may fail to capture model stability, reproducibility and interpretability, and instead only provide a limited, quantitative summary of model performance, especially when over-relying on a small set of metrics (Zhao et al., 2025).

### 3.4.2 Validation Strategies

Validation strategies are essential for assessing the generalizability and reliability of ML models. Among these strategies, cross-validation is one of the most commonly used validation techniques in metabolomics studies. To verify the robustness of the models and avoid overfitting, k-fold cross-validation and several permutations are performed (Wang et al., 2022; Ge et al., 2025; Idkowiak et al., 2025). In this type of technique, the dataset is randomly split into k subsets, where k-1 subsets are used for training and the remaining subset is used for testing, with the process repeated across all folds (Idkowiak et al., 2025).

Despite its widespread use, cross-validation has important limitations. The more complex the ML model is, the higher the risk of overfitting, despite standard k-fold cross-validation for optimization, because the stability of the model depends on the number of samples that are available for training (Mendez et al., 2019). Even the quality assessment statistics (Q2), which evaluates the predictive ability of a model and is calculated through cross-validation, can be misleading, since an invalid or irrelevant model can still produce a large Q2 value (Worley & Powers, 2013; Ge et

al., 2025). This is due to the fact that cross-validation requires a systematic deletion of large parts of the dataset during training. These limitations are especially common when the underlying model parameters vary due to heterogeneity of the training sets (Mendez et al., 2019). Therefore, it is essential that the complete dataset is representative of the biological question (Mendez et al., 2019; Zhao et al., 2025).

A critical issue arises when cross-validation is used for both model selection and performance evaluation (Sudhir & Richard, 2006). The resulting error estimate is then biased downward, because when multiple models are evaluated on the same data, the best performance is chosen due to random chance (overfitting) and a selection bias is introduced. The bias becomes more severe as the number of tested models increases and as sample size decreases, which are conditions that are typical in metabolomics studies. Consequently, models may appear to be highly accurate during validation while failing to generalize to independent datasets (Sudhir & Richard, 2006).

Nested cross-validation has been proposed as a more robust validation strategy (Sudhir & Richard, 2006). This approach separates model selection and performance evaluation into two independent loops: an inner loop for selecting the optimal model and an outer loop for estimating its predictive performance. The approach provides a more accurate estimate of prediction error. However, it is less frequently applied in metabolomics due to increased computational complexity and resource requirements (Sudhir & Richard, 2006).

Bootstrapping represents another validation approach, particularly in studies with limited sample sizes (Zhang et al., 2023). This method involves repeatedly resampling the dataset with replacement to generate multiple training and testing sets. Researchers have concluded that bootstrapping, when used on models with more replicates, can lower absolute error compared to cross-validation. Also, the approach could be highly time efficient. For example, disease classification using neural networks with 100 bootstraps and 1430 metabolites took less than a minute on a consumer laptop computer, with 1.2 GHz processor (Zhang et al., 2023).
Finally, the absence of independent external validation datasets remains as another major limitation in metabolomics studies. Internal validation methods cannot fully replicate real-world conditions, so independent test sets are considered as the most reliable approach for evaluating model generalizability (Sudhir & Richard, 2006). Sadly, they are often not available due to the cost and complexity of metabolomics data.

# 4. Discussion

This literature study aimed to explore the challenges and opportunities in metabolomics-based biomarker discovery using ML, with emphasis on the statistical and biological properties of metabolomics datasets. ML models have different complexities and therefore different methodological challenges and opportunities. PCA is one of the simpler and most widely used unsupervised method, which aims to uncover intrinsic structures within the data (Idkowiak et al., 2025) and is primarily used for clustering. Supervised methods such as LR, PLS-DA and OPLS-DA incorporate class labels and are more applicable to biomarker identification tasks. These models require fewer parameters and are based on linear assumptions, which means that they are relatively simple, leading to improved interpretability and better suitability for small datasets. This is particularly useful in metabolomics since metabolomics studies often have limited sample numbers (Mendez et al., 2019; Sun et al., 2024). Because they require fewer parameters, these ML models have a reduced risk of overfitting.

On the other hand, methods such as LR and PLS-DA cannot capture the complex relationships between metabolites. Metabolites are interconnected through complex biochemical pathways (Shomorony et al., 2020; Wang et al., 2022) and simple models can therefore lead to oversimplification of biological systems, also called underfitting (Zhao et al., 2025). They are also sensitive to feature selection and preprocessing (Worley & Powers, 2013).

To deal with the abovementioned challenges, more advanced ML models are applied, for example RF, SVM and XGBoost. These models are capable of capturing complex nonlinear relationships and handle high-dimensional data, which leads to improved biomarker discovery. However, these models are highly prone for overfitting since they can model noise as well as biologically relevant signals, and they require extensive hyperparameter tuning (Sun et al., 2024). They also have limited reproducibility across studies and require larger datasets, which is limited in metabolomics.

Overall, there is always a tradeoff between bias and variance when dealing with ML models. While simple linear models such as LR are not too sensitive to changes in the training data, meaning they have low variance and reduced risk of overfitting, more flexible models such as RF and SVM can capture noise as well as biologically relevant patterns in the data, indicating they have high variance and increased risk of overfitting (James et al., 2013). On the other hand, simple models have a reduced ability to capture true relationships in the data (high bias), while more advanced models have low bias and reduced risk of underfitting.

DL techniques have gained attention as more advanced approaches compared to classical ML methods in metabolomics. They are able to model complex and non-linear relationships and process high-dimensional data without extensive feature engineering (Zhao et al., 2025). They have also high computational efficiency in which metrics take only minutes to obtain (Zhang et al., 2023), and although some DL models require large datasets for training, researchers have developed models that performed well even when trained on small sample sizes (Wang et al., 2025). The main challenge with DL models is their limited interpretability, often referred to as "black-box problems" (Rudin et al., 2019; Watson et al., 2019). With other words, they are difficult to interpret biologically, making it challenging to understand how specific metabolites contribute to predictions. Due to lack of interpretability and transparency, DL models have limited clinical applicability. There are also challenges with class imbalance, in which the models tend to bias towards the majority classes in biomedical datasets that contain unequal group sizes (Tasci et al., 2022).

In order to assess the predictive performance of ML and DL models, evaluation metrics are used to serve as guidance for biomarker discovery. Common evaluation metrics are ROC (plot of the true positive rate versus false positive rate), AUC (summary measure of model performance), MCC (evaluates binary classification accuracy), accuracy (reflects the proportion of correct predictions), sensitivity (measure correct identification of positive cases) and specificity (measure correct identification of negative cases). Nevertheless, although evaluation metrics are essential for assessing a model, they should be interpreted with caution. For example, accuracy, sensitivity and specificity can give a distorted picture when classes are imbalanced, since a model may appear highly accurate simply by predicting the dominant class (Zhao et al., 2025). Furthermore, evaluation metrics do not include biological validity (Ge et al., 2025; Deng et al., 2024). This means that good model performance (e.g. high accuracy, AUC, sensitivity etc.) does not necessarily mean that the model has captured true biological meaning or real biological relationships. In other words, a model can look statistically strong but still be biologically questionable. In the high-dimensional metabolomics datasets, a model can achieve excellent metrics by exploiting dataset-specific noise and easily overfit to small sample sizes, capturing correlations that are not biologically meaningful. Therefore, the evaluation metrics do not account for how stable, reproducible or interpretable the model is.

To assess the generalizability and reliability of ML models, validation strategies are essential. Cross-validation is one of the most widely used validation approaches, such as k-fold cross-validation where the dataset is randomly split into k subsets

(Idkowiak et al., 2025). However, the stability of a model depends on the number of available training samples, which makes complex ML models particularly prone to overfitting even when cross-validation is applied. A major issue arises when cross-validation is used simultaneously for both model selection and performance assessment (Sudhir & Richard, 2006). In such cases, the estimated error becomes optimistically biased, and the model that performs best due to random variation is often selected, leading to overfitting. This bias becomes more pronounced as the number of evaluated models increases and the sample size decreases, which is frequently the case in metabolomics studies.

Nested cross-validation, in which model selection and performance evaluation are carried out in separate and independent loops, provides a more robust validation framework. Yet, this approach is computationally demanding and requires substantial resources (Sudhir & Richard, 2006). To yield a more stable and often lower error estimate compared to cross-validation, bootstrapping, where the dataset is repeatedly resampled with replacement to generate multiple training and test sets, is applied (Zhang et al., 2023). In addition, external validation datasets are highly valuable as they better reflect real-world conditions (Sudhir & Richard, 2006). Unfortunately, such datasets are often unavailable due to the high cost and complexity of metabolomics research.

Figure 6 shows a graphic representation of the entire workflow of ML-based metabolomics and the challenges and opportunities associated with it.

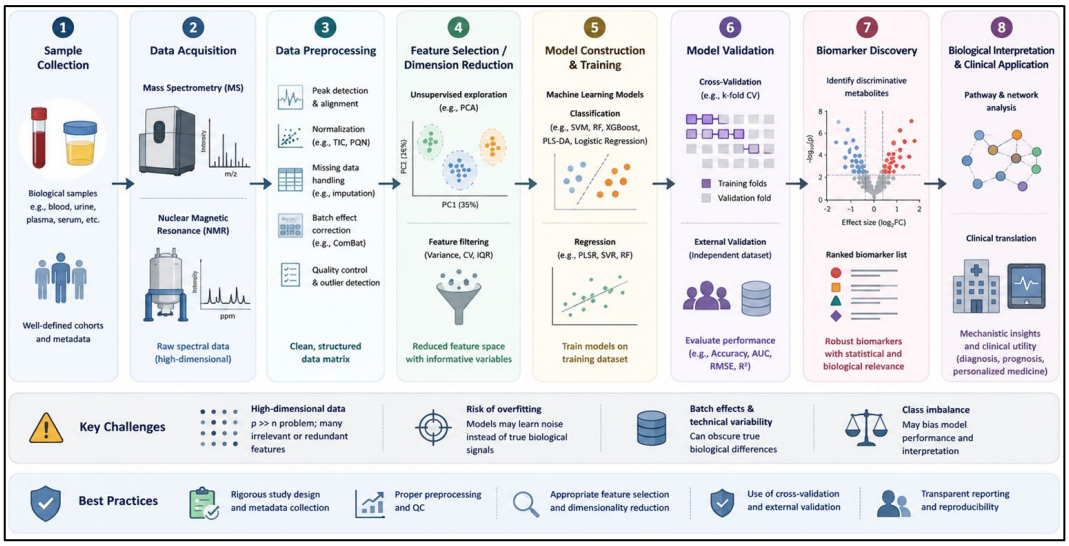



*Figure 6. Workflow of ML-based metabolomics. The key challenges are the high-dimensionality of metabolomics datasets that lead to model overfitting, and the technical variations and class imbalances that lead to statistical bias. Rigorous study design, proper preprocessing, appropriate feature selection and validation strategies are essential to get reliable and interpretable results. Picture generated with assistance of ChatGPT.*

Beyond the challenges associated with ML models and their validation strategies, the complex nature of the metabolomics dataset itself poses obstacles. Since metabolomics data contain more variables than samples, the complexity of the data increases and it becomes high-dimensional, which is also referred to as the "curse of dimensionality" (Worley & Powers, 2013; Labory et al., 2024). The correlation between metabolites, which is based on interconnected biochemical pathways, also lead to noisy and redundant information (Wang et al., 2022; Labory et al., 2024). Furthermore, metabolomics datasets contain missing values (Anwardeen et al., 2023; Idkowiak et al., 2025). These occur when metabolite concentrations are below the detection limit, as well as from technical problems during peak detection or alignment, or inconsistencies in sample preparation and instrument performance. Therefore, data preprocessing is a critical step in metabolomics-based ML, because without proper preprocessing, these issues can obscure biologically meaningful patterns and negatively affect model performance, robustness and reproducibility. Preprocessing techniques such as normalization, transformation, scaling and imputation can help reduce unwanted technical variation while preserving relevant biological information. Preprocessing also includes data cleaning, which identifies and corrects or removes inaccurate, corrupted, duplicated, improperly formatted, noisy or missing data (Zhao et al., 2025). This step is essential because low-quality raw data can reduce model performance and make the training process more complicated. Furthermore, data balance is a crucial component of ML data preparation, because maintaining a similar number of training samples for each class helps prevent the model from becoming biased toward majority classes (Zhao et al., 2025).

The literature suggests that equally important as selecting an ML method is determining how the model is optimized and how its ability to generalize is assessed. Achieving robust predictive models largely depends on having sufficient statistical power (Mendez et al., 2019). According to the curse of dimensionality, the generalization error increases with model complexity and decreases with the number of training samples. As a result, complex models trained on small, high-dimensional datasets often show poor classification performance on unseen data. In simple terms, larger and well-curated datasets are generally more suitable for nonlinear ML approaches.

A common assumption in metabolomics is that the choice of ML model is the primary determinant of analytical performance. However, the literature increasingly suggests that data characteristics play a far more decisive role than the specific model used. In high-dimensional settings with limited sample sizes, many models ranging from traditional statistical methods to advanced ML algorithms are

sufficient to fit the available data. As a result, differences in model performance are often marginal compared to the impact of how the data is generated, processed and structured. In other words, preprocessing decisions and feature selection strategies can have a greater influence on which biomarkers are identified rather than the choice of algorithm itself. Overall, these findings suggest that in metabolomics-based biomarker discovery, priority should be placed on high-quality data generation, rigorous preprocessing and robust study design, rather than optimizing the choice of ML model alone.

# 5. Conclusion

This literature study highlights that the primary challenges in metabolomics-based biomarker discovery are not solely rooted in the choice of ML or DL algorithms, but rather in the intrinsic characteristics of the data and the statistical framework applied to analyze them. High dimensionality, small sample sizes, multicollinearity, missing values and substantial technical and biological variability collectively create a complex analytical landscape in which the risk of overfitting, false discoveries and limited reproducibility is inherently high. Since many commonly used ML approaches can produce seemingly strong predictive performance while failing to capture true biological signals, especially when validation strategies are insufficient or when preprocessing steps introduce bias, high model accuracy does not necessarily equate to biological relevance or generalizability. The study further demonstrates that preprocessing and feature engineering play a decisive role in shaping analytical outcomes. Choices related to normalization, imputation and feature selection can substantially influence which biomarkers are identified, often to a greater extent than the selection of the ML model itself. To improve the robustness and reproducibility of metabolomics-based biomarker discovery, the literature emphasizes the importance of rigorous study design, transparent and appropriate preprocessing strategies and the use of robust validation frameworks, including bootstrapping and independent external datasets. Additionally, greater focus should be placed on statistical rigor, including critical evaluation of model performance metrics. Future studies should seek to shift the focus from model optimization to data-centric approaches, ensuring high-quality data generation, careful preprocessing and statistically sound validation practices.

# References


Anwardeen, N. R., Diboun, I., Mokrab, Y., Althani, A. A., & Elrayess, M. A. (2023). Statistical methods and resources for biomarker discovery using metabolomics. BMC bioinformatics, 24(1), 250.

Broadhurst, D. I., & Kell, D. B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics, 2(4), 171-196.

Camacho, D., De La Fuente, A., & Mendes, P. (2005). The origin of correlations in metabolomics data. Metabolomics, 1(1), 53-63.

Deng, Y., Yao, Y., Wang, Y., Yu, T., Cai, W., Zhou, D., ... & Li, W. (2024). An end-to-end deep learning method for mass spectrometry data analysis to reveal disease-specific metabolic profiles. Nature Communications, 15(1), 7136.

Frölich, N., Klose, C., Widén, E., Ripatti, S., & Gerl, M. J. (2024). Imputation of missing values in lipidomic datasets. Proteomics, 24(15), 2300606.

Ge, C., Lu, Y., Shen, Z., Lu, Y., Liu, X., Zhang, M., ... & Zhu, L. (2025). Machine learning and metabolomics identify biomarkers associated with the disease extent of ulcerative colitis. Journal of Crohn's and Colitis, 19(2), jjaf020.

Gonzalez-Riano, C., Saiz, J., Barbas, C., Bergareche, A., Huerta, J. M., Ardanaz, E., ... & Amiano, P. (2021). Prognostic biomarkers of Parkinson's disease in the Spanish EPIC cohort: A multiplatform metabolomics approach. npj Parkinson's Disease, 7(1), 73.

Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L., & Goodacre, R. (2015). A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding. Analytica chimica acta, 879, 10-23.

Idkowiak, J., Dehairs, J., Schwarzerová, J., Olešová, D., Truong, J. X., Kvasnička, A., ... & Holčapek, M. (2025). Best practices and tools in R and Python for statistical processing and visualization of lipidomics and metabolomics data. Nature Communications, 16(1), 8714.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R* (Vol. 103). New York: springer.

Labory, J., Njomgue-Fotso, E., & Bottini, S. (2024). Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data. Computational and Structural Biotechnology Journal, 23, 1274-1287.

Mendez, K. M., Reinke, S. N., & Broadhurst, D. I. (2019). A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. Metabolomics, 15(12), 150.

Pain, O., Dudbridge, F., & Ronald, A. (2018). Are your covariates under control? How normalization can re-introduce covariate effects. European Journal of Human Genetics, 26(8), 1194-1201.

Ranganathan, P., Pramesh, C. S., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: logistic regression. *Perspectives in clinical research*, *8*(3), 148-151.

Ren, S., Hinzman, A. A., Kang, E. L., Szczesniak, R. D., & Lu, L. J. (2015). Computational and statistical analysis of metabolomics data. Metabolomics, 11(6), 1492-1513.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5), 206-215.

Sha, C., Cuperlovic-Culf, M., & Hu, T. (2021). SMILE: systems metabolomics using interpretable learning and evolution. BMC bioinformatics, 22(1), 284.

Shomorony, I., Cirulli, E. T., Huang, L., Napier, L. A., Heister, R. R., Hicks, M., ... & Shah, N. (2020). An unsupervised learning approach to identify novel signatures of health and disease from multimodal data. Genome medicine, 12(1), 7.

Sun, L., Fan, X., Zhao, Y., Zhang, Q., & Jiang, M. (2024). Deep learning-based metabolomics data study of prostate cancer. BMC bioinformatics, 25(1), 391.

Tasci, E., Zhuge, Y., Camphausen, K., & Krauze, A. V. (2022). Bias and class imbalance in oncologic data—towards inclusive and transferrable AI in large scale oncology data sets. *Cancers*, *14*(12), 2897.

Tuerxunyiming, M., Zhao, Q., Hu, Q., Zhu, P., & Zhu, S. (2025). LC-MS-based conventional metabolomics combined with machine learning models to identify metabolic markers for the diagnosis of type I diabetes. Frontiers in Endocrinology, 16, 1588718.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC bioinformatics, 7(1), 91.

Wang, H., Wang, Y., Li, X., Deng, X., Kong, Y., Wang, W., & Zhou, Y. (2022). Machine learning of plasma metabolome identifies biomarker panels for metabolic syndrome: findings from the China Suboptimal Health Cohort. Cardiovascular Diabetology, 21(1), 288.

Wang, K., Li, J., Meng, D., Zhang, Z., & Liu, S. (2022). Machine learning based on metabolomics reveals potential targets and biomarkers for primary Sjogren's syndrome. Frontiers in Molecular Biosciences, 9, 913325.

Wang, T., Holscher, H. D., Maslov, S., Hu, F. B., Weiss, S. T., & Liu, Y. Y. (2025). Predicting metabolite response to dietary intervention using deep learning. Nature communications, 16(1), 815.

Watson, D. S., Krutzinna, J., Bruce, I. N., Griffiths, C. E., McInnes, I. B., Barnes, M. R., & Floridi, L. (2019). Clinical applications of machine learning algorithms: beyond the black box. Bmj, 364.

Worley, B., & Powers, R. (2013). Multivariate analysis in metabolomics. Current metabolomics, 1(1), 92-107.

Xiao, J. F., Zhou, B., & Ressom, H. W. (2012). Metabolite identification and quantitation in LC-MS/MS-based metabolomics. TrAC Trends in Analytical Chemistry, 32, 1-14.

Zhang, J. D., Xue, C., Kolachalama, V. B., & Donald, W. A. (2023). Interpretable machine learning on metabolomics data reveals biomarkers for Parkinson's disease. ACS Central Science, 9(5), 1035-1045.

Zhao, C., Zhao, T., Shen, Q., Tang, Z., Li, X., & Huan, T. (2025). Practical Guidance for Training Machine Learning Models in Metabolomics and Mass Spectrometry Research. Analytical Chemistry, 97(42), 23005-23013.

**Publishing and archiving**

Approved students' theses at SLU can be published online. As a student you own the copyright to your work and in such cases, you need to approve the publication. In connection with your approval of publication, SLU will process your personal data (name) to make the work searchable on the internet. You can revoke your consent at any time by contacting the library.

Even if you choose not to publish the work or if you revoke your approval, the thesis will be archived digitally according to archive legislation.

You will find links to SLU's publication agreement and SLU's processing of personal data and your rights on this page:

- https://libanswers.slu.se/en/faq/228318

☒ YES, I, Maryam Chaid, have read and agree to the agreement for publication and the personal data processing that takes place in connection with this

☐ NO, I/we do not give my/our permission to publish the full text of this work. However, the work will be uploaded for archiving and the metadata and summary will be visible and searchable.