Metabolomics Pipeline Overview

Table of contents

1 Introduction

This document describes a basic pipeline for analyzing untargeted LC–MS/MS metabolomics data in R.

The example data consist of panels of untargeted LC–MS features from positive and negative ionization mode runs, similar to metabolites measured from tissue, stool samples, or microbial cultures. The files in data_raw/metabolomics-pipeline-example.xlsx and the corresponding positive- and negative-mode .csv exports come from a Thermo Fisher Compound Discoverer v3 workflow and have already undergone feature detection, basic quality control, and signal curation. The values are relative peak intensities for deconvoluted LC–MS features.

These data have been restructured into a hypothetical experimental design with samples assigned to a control group, one of two single-treatment groups (tr1 or tr2), or a dual-treatment group (tr1+tr2) to demonstrate a clear treatment-response pattern.

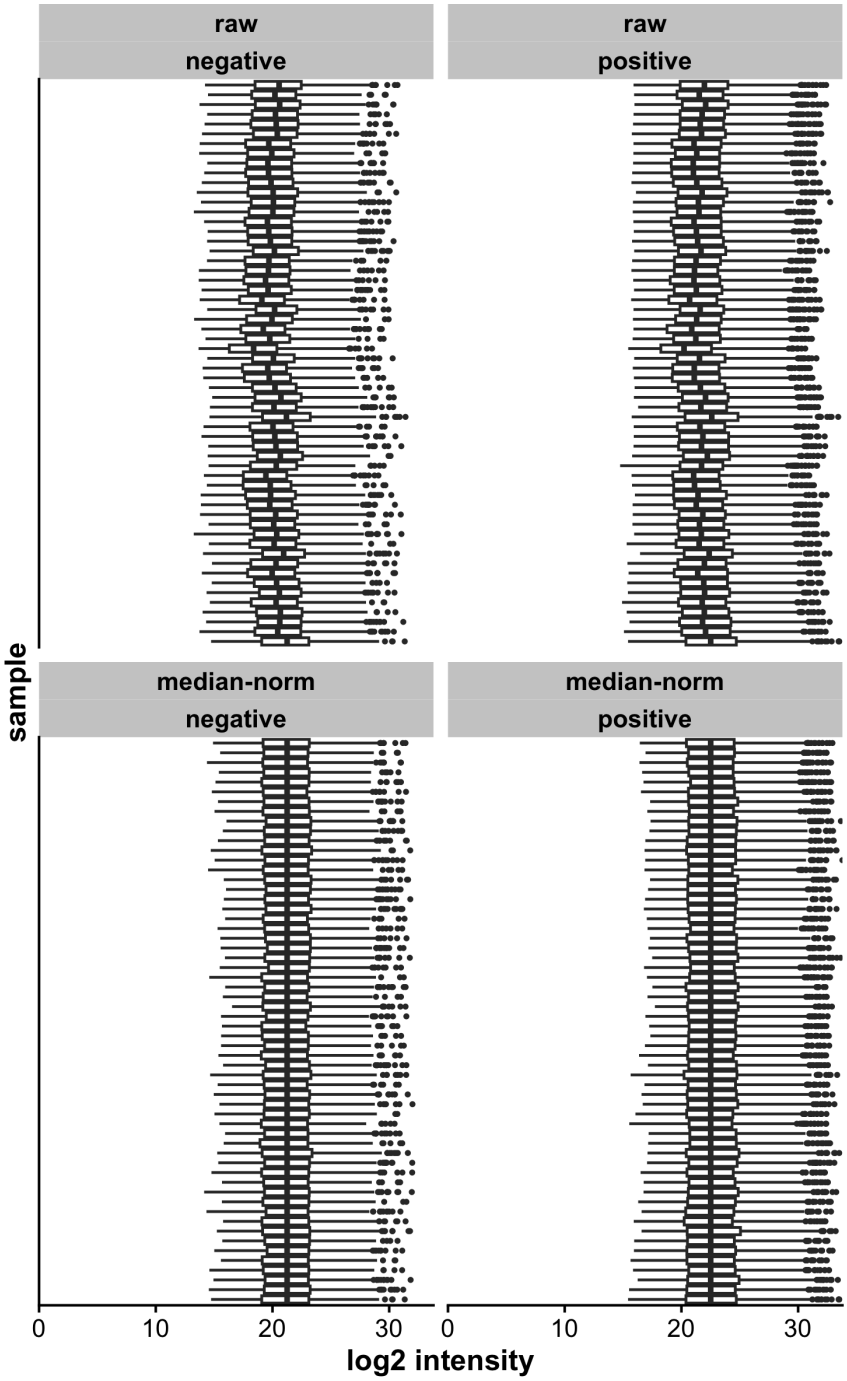
The pipeline illustrates a standard untargeted metabolomics workflow: data cleaning and normalization, unsupervised and supervised multivariate analysis, feature-level differential testing, and putative metabolite annotation using KEGG-based compound information.

All requisite R packages are managed with the renv package (see the renv/ folder for dependencies). R/00_setup.r initializes the environment, loads packages, formats sample metadata, and sets global plotting aesthetics for the figures in this document.

2 Data Distribution

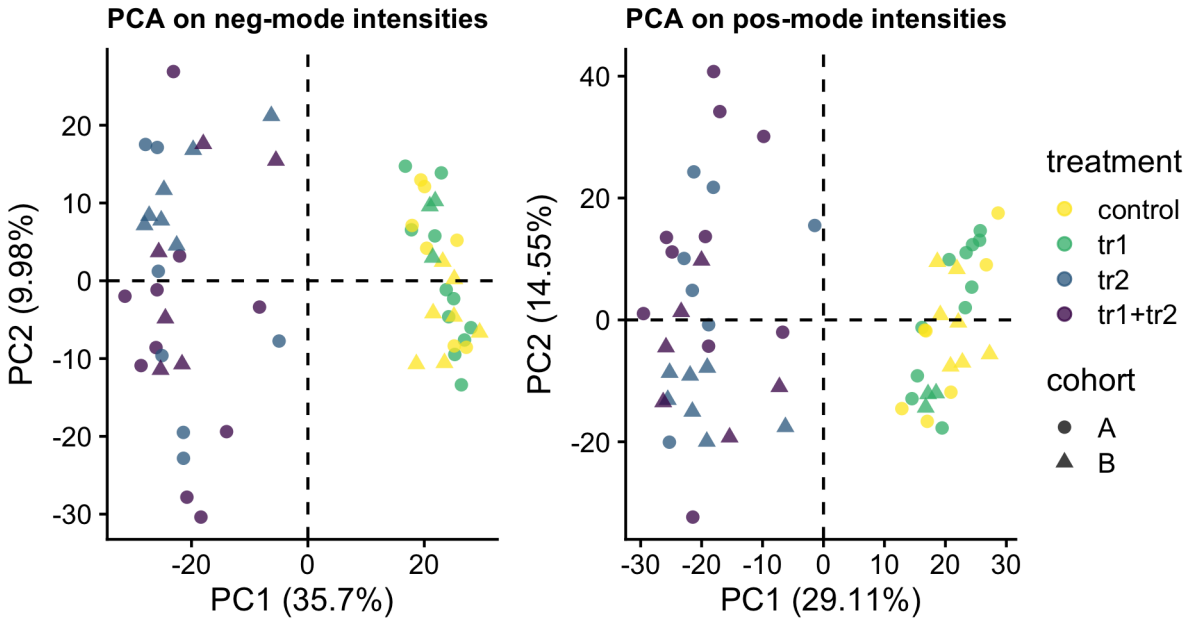
To further clean up the data, 01_preprocess.r merges LC-MS features with identical mass-to-charge ratios and retention times. It then assigns each feature a unique ID by merging its mass-to-charge ratio and retention time ("mz_rt_min").

As LC-MS intensities of metabolomics data typically skew to the right, data normalization is required to improve comparability across samples and help stabilize variance before actual analysis can begin. The figure demonstrates how median normalization reduces between-sample differences in overall intensity while preserving the underlying variability that's of interest in the data. The data are also transformed into a log2 scale to reduce domination by extreme values and bring the data closer to normality.



3 Sample Analysis — PCA

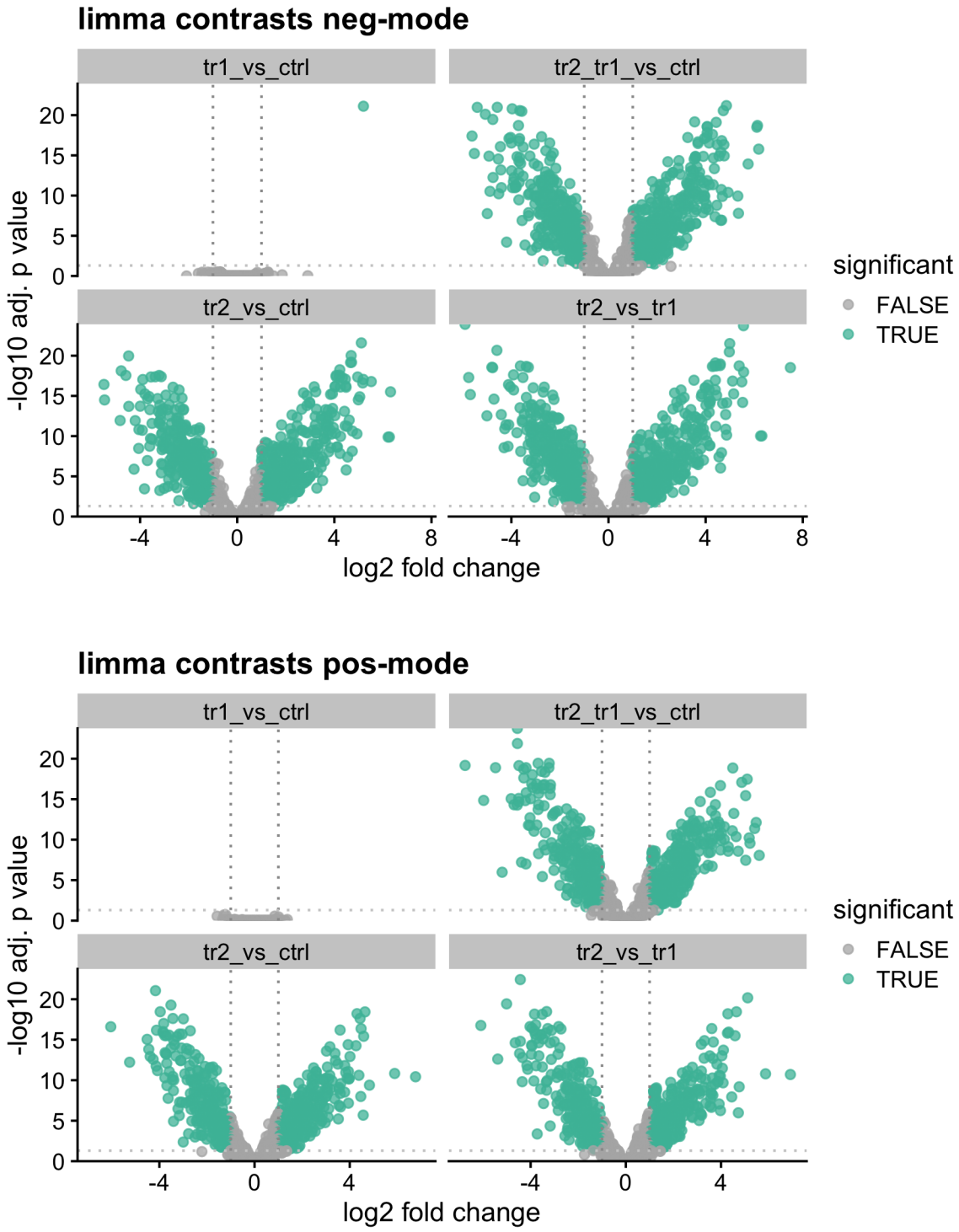
Principal Component Analysis is used to gauge the overall structure of the metabolomic profiles (samples/cases). The median-normalized, log2-transformed intensities are next centered and scaled, then decomposed into a set of orthogonal principal components that capture the dominant sources of variance in the dataset.



Principal component (PC) 1 illustrates a divide amongst the samples based on the treatment group, with controls and tr1 contrasted from tr2 and the dual-treatment (tr1+tr2). PC 1 captures 35.7% of the variance in the data for the negative-mode intensities and 29.11% for the positive-mode intensities, meaning the treatment variable has substantial influence on the overall data structure. PC 2 does not appear to highlight any meaningful secondary structure in the data.

4 Feature Analysis — limma

To evaluate the underlying feature fluctuations responsible for the observed sample differences, linear models for microarrays (limma) were applied to the median-normalized, log2-transformed intensities. This approach tests each LC-MS feature for differential abundance across treatment groups while controlling the false discovery rate using multiple-testing correction.



The figure depicts the differentially abundant features for each pairwise comparison, highlighting metabolites that most strongly distinguish the control, single-treatment (tr1, tr2), and dual-treatment (tr1+tr2) conditions.

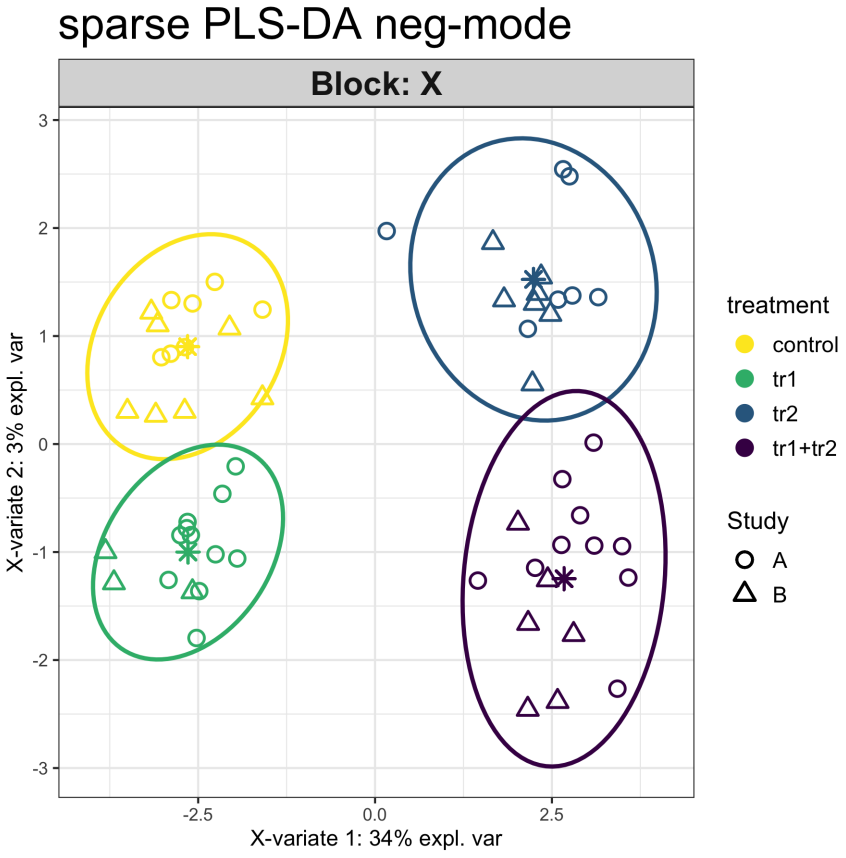
Table 1: limma differential features

Mode	tr1 vs ctrl	tr2 vs ctrl	tr1+tr2 vs ctrl	tr2 vs tr1	total features
negative	1	845	820	825	1499
positive	0	642	643	625	1498

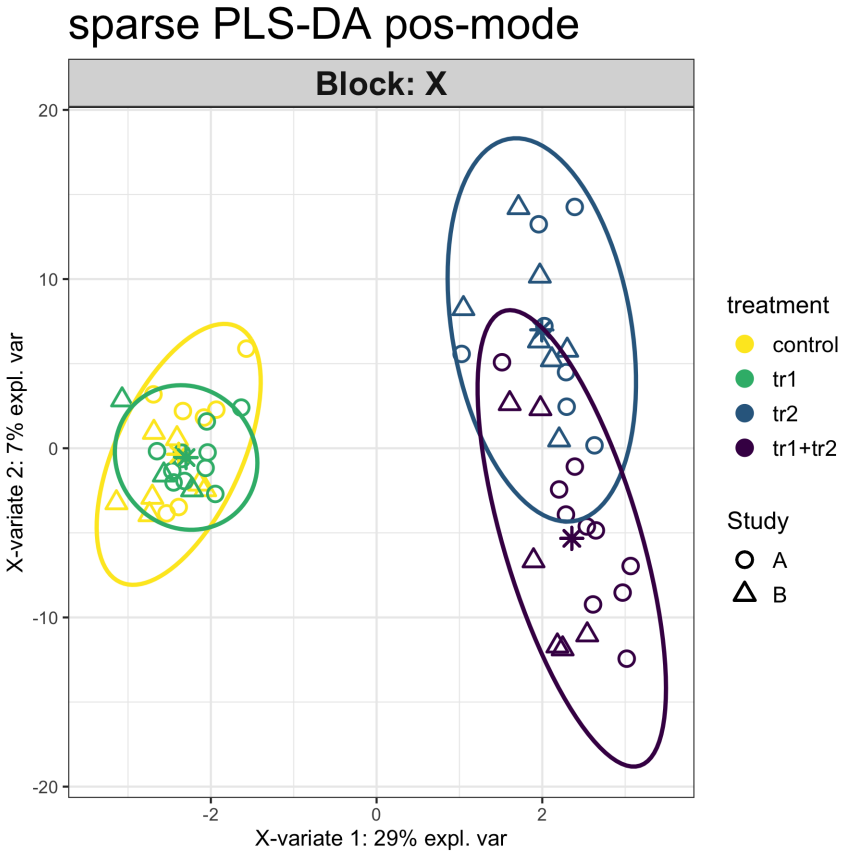
The figure and table results show that treatment 2 induces a pronounced effect in the samples' metabolome with over half of the total features perturbed from the control baseline. Treatment 1 alone has effectively no impact on the metabolome and thus all the differential features between the dual-treatment and controls are likely a consequence of treatment 2.

5 Supervised Analysis — sparse PLS-DA

Partial least squares discriminant analysis (PLS-DA) is a supervised multivariate method that models the relationship between a set of predictors (the LC-MS features) and a categorical outcome (the treatment groups). In this implementation, a sparse variant (sPLS-DA) is used so that only a subset of variables contributes non-zero loadings to the components.

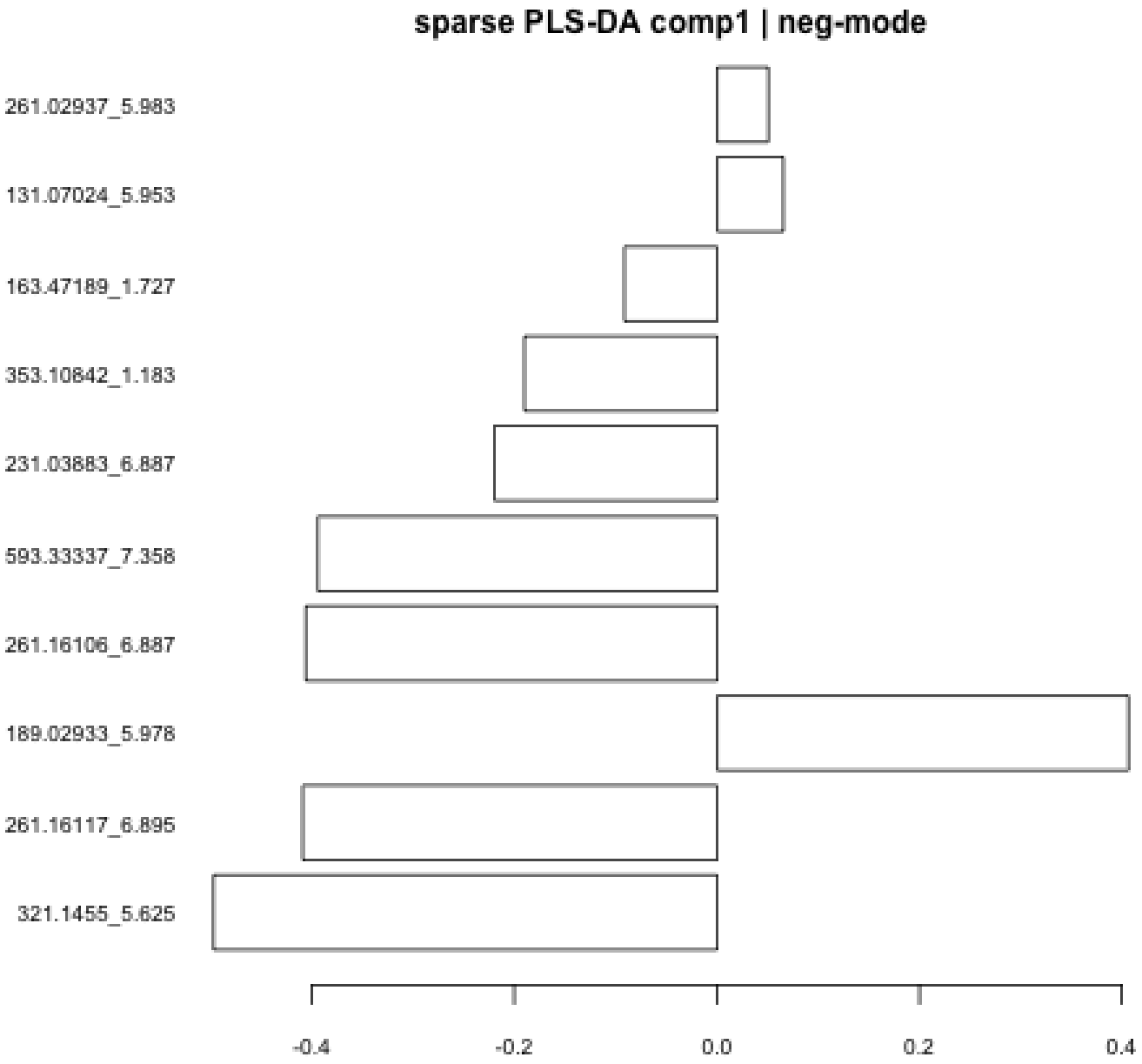


sPLS-DA is applied here to identify a minimal set of discriminative features that best separates the control, single-treatment, and dual-treatment groups in the reduced latent space. The resulting components summarize multivariate treatment effects, while the sparsity constraint highlights candidate metabolites that contribute most strongly to class separation and can be prioritized for downstream annotation.

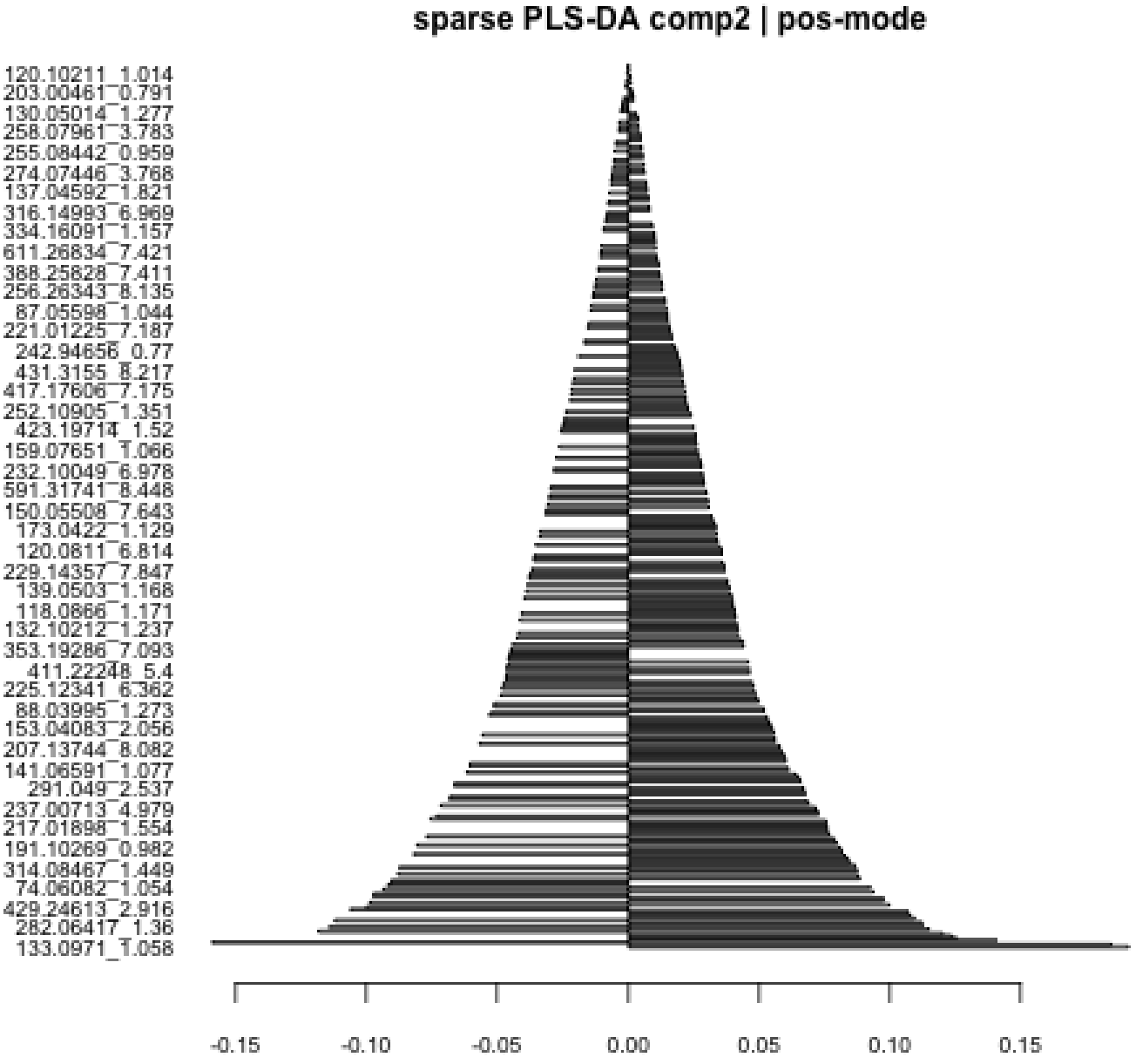


In both ionization modes, the first component emphasizes the contrasts induced by treatment 2, consistent with the unsupervised PCA results. In the negative-mode data, the second component further discriminates the groups but accounts for only 3% of the total variance, so its contribution should be interpreted cautiously.

The LC-MS features driving these separations can be examined via their loadings on each sparse component. For the negative-mode data, both components retain only 10 features out of 1,499, indicating that the variation in this small subset provides an efficient summary of the treatment-related differences.



In the positive-mode data, the first component similarly selects 10 features to distinguish treatment groups. The second component (shown below), by contrast, retains 100 features and shows only weak additional separation between the treatment 2 and dual-treatment samples, consistent with the idea that a larger number of variables is required to represent a relatively subtle effect.



6 Conclusion