Chemometrics and Intelligent Laboratory Systems 232 (2023) 104720

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Analysis of longitudinal metabolomic data using multivariate curve resolution-alternating least squares and pathway analysis

Isabel Ten-Doménech [a], Marta Moreno-Torres [b,c,d], Juan Daniel Sanjuan-Herráez [e], David Pérez-Guaita [f], Guillermo Quintás [e,g,*], Julia Kuligowski [a]

[a] Neonatal Research Group, Health Research Institute Hospital La Fe, Valencia, Spain
[b] Unidad de Hepatología Experimental y Trasplante Hepático, Health Research Institute Hospital La Fe, Valencia, Spain
[c] Departamento de Bioquímica y Biología Molecular, Facultad de Medicina, Universidad de Valencia, Valencia, Spain
[d] Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Instituto de Salud Carlos III, Madrid, Spain
[e] Health and Biomedicine, LEITAT Technological Center, Terrassa, Spain
[f] Departamento de Química Analítica, Universidad de Valencia, Burjassot, Spain
[g] Analytical Unit, Health Research Institute Hospital La Fe, Valencia, Spain

ARTICLE INFO

Keywords:
Metabolomics
Pathway analysis
MCR-ALS
Longitudinal analysis
Variance reduction

ABSTRACT

Extraction of meaningful biological information from longitudinal metabolomic studies is a major challenge and typically involves multivariate analysis and dimensional reduction methods for data visualization such as Principal Component Analysis or Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS). Besides, a variety of computational tools have been developed to identify changes in metabolic pathways including functional analysis and pathway analysis. In this work, the joint analysis of results from MCR-ALS and metabolic pathway analysis is proposed to facilitate the interpretation of dynamic changes in longitudinal metabolomic data. The strategy is based on the use of MCR-ALS to remove unstructured random variation in the raw data, thus facilitating the interpretation of dynamic changes observed by metabolic pathway analysis over time. A simulated data set representing dynamic longitudinal changes in the intensities of a subset of metabolites from three metabolic pathways was initially used to test the applicability of MCR-ALS to support pathway analysis for detecting pathway perturbations. Then, the strategy is applied to real data acquired for the analysis of changes during $CD8^+$ T cell activation. Results obtained show that MCR-ALS facilitates the interpretation of longitudinal metabolomic profiles in multivariate data sets by identifying metabolic pathways associated with each detected dynamic component.

1. Introduction

Longitudinal metabolomics involves the analysis of biological samples over time, where the same matrix of measurements is repeated at different times to monitor, for example, the effect of an external intervention such as drug, exercise, or the evolution of a disease or a given treatment on a given population. Longitudinal studies are critical for the understanding of the evolution of biological processes and provide major advantages described elsewhere [1]. However, the extraction of useful information from metabolomic data to unravel the biochemical events in longitudinal studies is challenging. Data acquired over time is often arranged as a two-dimensional matrix in which each row corresponds to a given time point and each column to a metabolic feature. The number of detected variables in metabolomic studies largely exceeds the number of samples, variables are usually correlated, and the biological information is typically contained in a fraction of the variables. Furthermore, data acquired from the analysis of samples collected in longitudinal studies typically show autocorrelations. Nevertheless, data analysis is frequently initiated by using exploratory methods such as Principal Component Analysis (PCA) for dimensional reduction and data visualization, or Partial Least Squares (PLS) for discriminant analysis [2].

PCA is arguably the most widely used method in metabolomics for the unsupervised analysis of multivariate data sets. PCA reduces the dimensionality of the experimental data matrix **D** (n x j), where n is the number of metabolic profiles and j the number of metabolic features, to

* Corresponding author. Health and Biomedicine, LEITAT Technological Center, Terrassa, Spain.
E-mail address: gquintas@leitat.org (G. Quintás).

https://doi.org/10.1016/j.chemolab.2022.104720
Received 15 September 2022; Received in revised form 10 November 2022; Accepted 26 November 2022
Available online 30 November 2022

a lower number of independent variables that take the interdependence of the original variables into account through linear functions which successively maximize variance and are uncorrelated with each other [3–5]. Dimensionality reduction facilitates the analysis of the correlations among variables and data visualization to identify the underlying structure. In the case of longitudinal studies, the scores of the PCs can be correlated with the "longitudinal variable" (e.g., time) to identify patterns of evolution of the metabolome. However, there is no guarantee that the directions of maximum variance in the original multivariate space contain biologically relevant information. Furthermore, a single PC does not always capture biological variation of interest and the information is contained in the combination of several PCs, making difficult the interpretation. Low variance components might contain biological variation of interest, while components with large variances might describe non-biological sources of variance [6]. The PCA loadings describe how much each original variable contributes to a particular principal component, and whether the variable and the principal component are positively or negatively correlated. Another algorithm used in the analysis of dynamic systems is Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS). MCR-ALS models multivariate data by considering a set of components whose concentrations evolve at different rates and have different characteristic metabolic features. MCR-ALS iteratively solves the bilinear equation $\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathbf{T}} + \mathbf{E}$, where the data matrix **D** (n x j) is decomposed in a pure variable matrix **S** (j x i) and a concentration matrix **C** (n x i), and **E** (n x j) corresponds to unexplained data variance, where i is the selected number of MCR-ALS components. Matrix **C** describes the evolution of the concentration of metabolites as a function of time, and **S** represents metabolic profiles that are jointly regulated in the biological samples. The iterative ALS algorithm used to solve the bilinear equation calculated at each iteration provides estimates of **C** and **S** by minimizing the residual unexplained variance **E**, until a convergence criterion is met. The optimization requires the initial selection of the number of components (i) and the use of initial estimates of **C** or **S**, that can be obtained using different methods such as SIMPLe-to-use Interactive Self-modeling Mixture Analysis (SIMPLISMA) [7], evolving factor analysis (EFA) [8], or pure components. MCR methods suffer from non-uniqueness of their results, and there are three types of ambiguities in MCR methods: permutation, intensity and rotation ambiguities [9]. The application of constrains that are both mathematically tractable and physically explicable such as unimodality in the concentration profiles, or non-negativity in **C** and/or **S**, improves the accuracy and interpretability of the retrieved spectra and concentration profiles and can partly overcome the ambiguities inherent to the factor analysis decomposition [10]. Furthermore, selectivity or local rank constraints can be used when some species are not present in certain samples to improve the estimation of the **C** profiles during the iterations [11]. As MCR-ALS was initially developed and has been widely applied to resolve optical spectra of chemical mixtures, matrices **S** and **C** are commonly known as the "spectral" and "concentration" matrices. Although in this study the matrix **S** refers to a metabolic profile instead of a characteristic spectrum, we maintained the use of the standard expression 'spectral matrix' to facilitate the interpretation of the manuscript.

MCR-ALS has been used for the analysis of metabolomic data in the compression and pre-processing of multivariate datasets generated by spectral and hyphenated techniques such as nuclear magnetic resonance (NMR) spectroscopy [12] liquid chromatography-mass spectrometry (LC-MS), gas chromatography-mass spectrometry (GC-MS) or MS-imaging (MSI) [13–15], as well as for modeling temporally designed NMR-based metabolomics data [2]. MCR-ALS has also been proposed for the identification of reliable components in metabolomics based on their reproducibility in repeated MCR-ALS calculations with the number of components changing for each iteration [16].

In parallel, pathway analysis is widely used to facilitate the biological interpretation of changes observed over time in metabolomic data [17,18]. Over-representation analysis (ORA) [19] tests if a group of metabolites is represented more than expected by chance within a list of altered metabolites identified using e.g., univariate testing and a pre-defined cutoff based on p-values, to determine whether metabolites involved in a particular pathway are enriched compared to random hits. The *mummichog* algorithm [20] is also widely used for functional metabolic analysis to infer pathway activities from a ranked list of MS peaks identified by untargeted metabolomics. It starts with the putative annotation of LC-MS peaks identified by their mass-to-charge ratio (m/z) or m/z and retention time, considering different adducts and polarities. Then, the annotated features are mapped onto pathway libraries (e.g., Kyoto Encyclopedia of Genes and Genomes, KEGG) for pathway activity prediction using a user-selected m/z accuracy [18]. The algorithm implements an ORA method to evaluate pathway-level enrichment based on significant features selected using a pre-defined cutoff based on p-values [21]. In summary, pathway analysis typically relies on the identification of significant features using a univariate test (e.g., t-test) and a pre-defined p-value cutoff, or on their ranking.

$CD8^+$ T cells detect and kill infected or cancerous cells. When activated from their naïve state, T cells undergo a complex transition, including major metabolic reprogramming. The time-dependent transition from naïve to effector T cells has been recently studied *in vitro* using a combination of two flow injection analysis (FIA) and three LC methods in combination with positive and negative high-resolution MS modes [22]. Results from univariate analysis, PCA, and supervised PLS indicated that, depending on the method, between 54% and 98% of measured metabolic features change in a time-dependent way. Moreover, results showed that the impact of the $CD8^+$ T cell activation process on the metabolome was not constant. For example, results obtained by PLS analysis of the polar metabolites measured by FIA ± using the time (h) as independent variable indicated that the major metabolic differences occurred in the first 48 h of $CD8^+$ T cell activation. Likewise, time profiles of key metabolites for fatty acid oxidation (carnitine), glycolysis (lactic acid) and polyamine biosynthesis (arginine, spermidine, ornithine and spermine) were markedly different. Carnitine concentrations showed a gradual decline over the first 60 h. Lactic acid concentrations increased during the first 48 h, reaching a plateau and showed a decline at 84 h. Arginine and ornithine concentrations decreased over the first 24 h, and spermidine and spermine increased at different relative rates over the first 36 h and 48 h, respectively. Besides, transient increases in cofactors S-adenosylmethionine (SAM) and methyl-thioadenosine (MTA) with maximum values at 36 and 48 h, respectively, were found. PCA scores plots of lipid data sets indicated again that the major changes in the lipid composition occurred in the first 48 h of the experiment, and that the concentrations of membrane lipids such as phosphatidylethanolamines (PE), lysophosphatidylcholines (LPC), and lysophosphatidylethanolamines (LPE) increased transiently between 24 and 72 h post-activation before returning to values similar to pre-activation. These results indicate the presence of simultaneous asynchronous metabolic programming processes that impact the concentrations of the metabolites.

The aim of this work is to demonstrate the utility of using MCR-ALS for the analysis of metabolomic data derived from longitudinal studies to facilitate the interpretation of observed changes through metabolic pathway analysis. A simulated data set representing longitudinal changes in the concentrations of a set of metabolites from three metabolic pathways was initially used to test the applicability of MCR-ALS to support pathway analysis in the detection of pathway perturbations. Then, the strategy was used in real data acquired during an *in vitro* study focused on $CD8^+$ T cell activation.

Results obtained in simulated and real data show that the joint analysis of results from metabolic pathway analysis and MCR-ALS facilitate the interpretation of changes in dynamic metabolomic profiles, and the identification of metabolic pathways associated.

2. Materials and methods

2.1. Simulated data set

A data set was initially built as model example simulating three replicate longitudinal experiments. Accordingly, three $\mathbf{X_i}$ (20 x 469) matrices (i = [1,2,3]) of normally distributed random numbers with a relative standard deviation of 0.15% were built, where each row represented a data point (t = [1, 2, …, 20]) and each column a metabolite included in the following nine KEGG pathways: Tryptophan metabolism (map00380, 84 metabolites), Phenylalanine metabolism (map00360, 47 metabolites), Steroid hormone biosynthesis (map00140, 99 metabolites), Glutathione metabolism (map00480, 29 metabolites), Vitamin B6 metabolism (map00750, 29 metabolites), Cyanoamino acid metabolism (map00460, 41 metabolites), Thiamine metabolism (map00730, 20 metabolites), Glyoxylate and dicarboxylate metabolism (map00630, 46 metabolites) and Arachidonic acid metabolism (map00590, 74 metabolites). Then, the intensities of 30 metabolites included in the Tryptophan, the Phenylalanine, and the Glyoxate and dicarboxylate metabolisms were modified to show a dynamic profile described by the functions shown in Fig. 1A in each simulated longitudinal experiment $\mathbf{X_i}$. The 10 metabolites affected by each simulated effect were: Formyl-N-acetyl-5-methoxykynurenamine, Formyl-5-hydroxykynurenamine, 5-Hydroxy-N-formylkynurenine, 5-Hydroxykynurenine, 5-Hydroxykynurenamine, 4,6-Dihydroxyquinoline, 4-(2-Amino-5-hydroxyphenyl)-2,4-dioxobutanoate, 6-Hydroxykynurenic acid, Anthranilic acid, and Formylanthranilate in the Tryptophan metabolism (effect 1); L-Glutamate, Glycine, Glutathione, Ascorbate, L-Ornithine, L-Cysteine, Glutathione disulfide, Putrescine, Spermidine, and γ-L-Glutamyl-L-cysteine in the Glutathione metabolism (effect 2); and 2-Oxoglutarate, Acetate, Oxaloacetate, Glyoxylate, Formate, Propanoyl-CoA, Butanoyl-CoA, (S)-Malate, Citrate, and Glycolate in the Glyoxate and dicarboxylate metabolism (effect 3).

An augmented data matrix $\mathbf{X_s}$ (60 x 469) = [$\mathbf{X_1}$, $\mathbf{X_2}$, $\mathbf{X_3}$] was then created by row-wise matrix concatenation. In this simulated set of longitudinal experiments, using as reference the metabolic concentrations at t = 1, six metabolic pathways were labeled as non-perturbed (map00140, map00480, map00750, map00460, map00730, map00590), and three metabolic pathways (map00380, map00360, and map00630) were identified as perturbed following specific time profiles shown in Fig. 1B. Then, a second matrix $\mathbf{X_{sn}}$ (60 x 469) was built by adding random normally distributed noise to $\mathbf{X_s}$ with a standard deviation of 0.3 (see Fig. 1C). Thinking of the Student's t-test in terms of signal to noise ratio, in this simulation the signal would be the difference between the mean of two samples (i.e., the difference of mean values of a given metabolite at $t = 1$ and $t = i$ (i = [2, 3, …, 20]). Thus, increasing the width of the distributions by adding noise in $\mathbf{X_{sn}}$ makes the difference in means less likely to be significant and more sensitive to chance variation, affecting the power of pathway analysis to detect perturbed pathways.

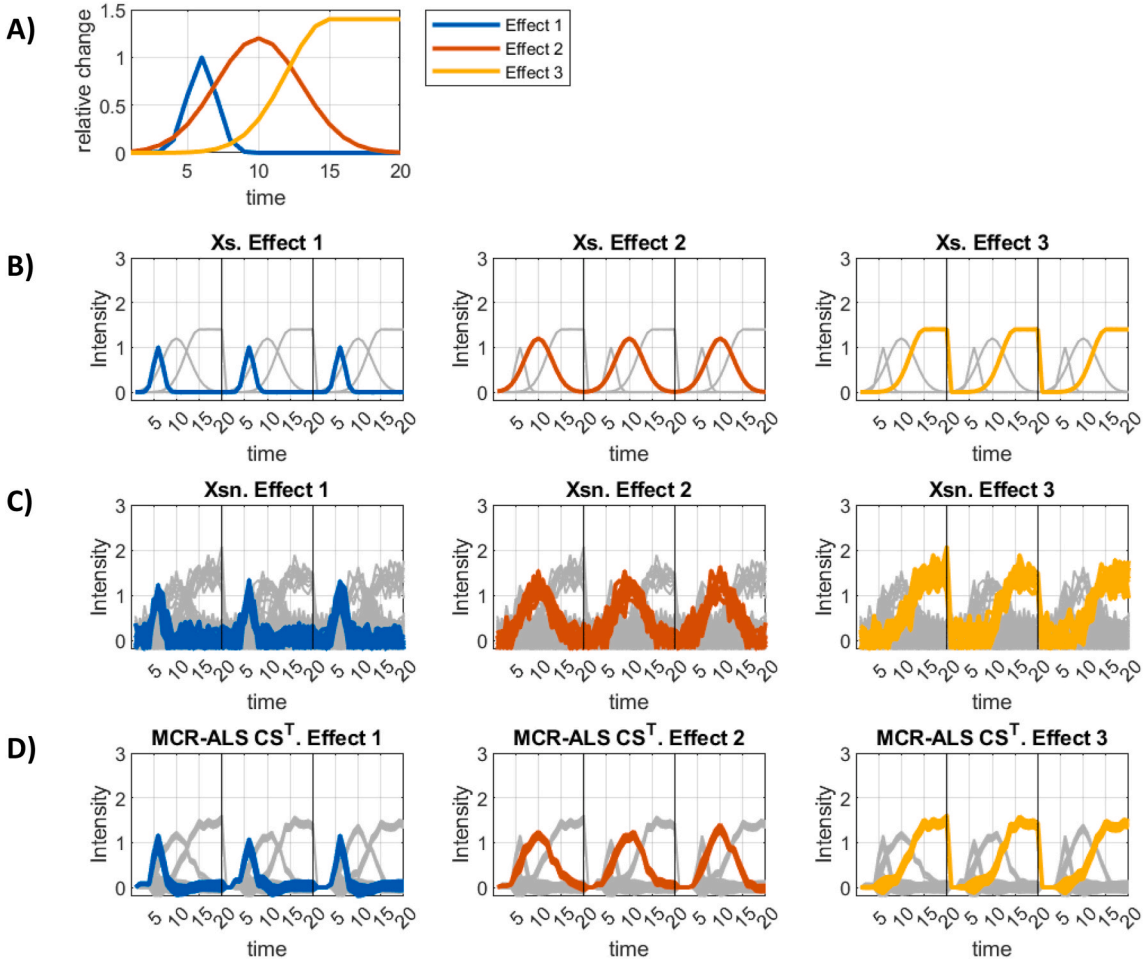

Fig. 1. Simulated data set. Intensities of the three artificial effects (A). Intensities of the set of variables as a function of time in the simulated data with low ($\mathbf{X_s}$) (B) and high ($\mathbf{X_{sn}}$) noise levels (C), and in the MCR-ALS reconstructed $\mathbf{CS^T}$ matrix (D). Note: In B, C, and D, variables modified by effects 1 (left), 2 (center), or 3 (right) were highlighted in separate plots for a better visualization.

2.2. Metabolomic data set and clean-up

A complete description of the data set can be found elsewhere [22]. Briefly, $CD8^+$ T cells were activated *in vitro* and intra- and extracellular metabolites were extracted every 12 h for 4 days in total. The study involved sample collection at 9 time points (at 0, 12, 24, 36, 48, 60, 72, 84, and 96 h) performed by triplicate and their analysis using complementary analytical methodologies: FIA – MS using positive and negative electrospray ionization (ESI+/−), hydrophilic interaction liquid chromatography-quadrupole time of flight MS (HILIC-QTOF-MS) using ESI-, reversed phase LC-QTOF-MS using ESI+, and LC-QTOF-MS using ESI+/− for lipid analysis. This strategy enabled the detection of a large number of features supporting a broad metabolite coverage: FIA-ESI (+)-MS: 1887 features; FIA-ESI(−)-MS: 2416 features; HILIC-ESI (−)-QTOF-MS: 1671 features; LC-ESI(+)-QTOF-MS: 1549 features; lipidomics LC-ESI(+)-QTOF-MS: 1819 features; and lipidomics LC-ESI (−)-QTOF-MS: 1745 features. An initial data clean-up was applied to identify and remove uninformative features. Accordingly, for each data sub-set, univariate *t*-tests were applied to compare the distributions of the intensities of each feature at each time point with respect to blank samples. Features for which the null hypothesis of equal means could be rejected (Student's *t*-test, p-value<0.01) in, at least, three consecutive time points, were retained for further analysis. Thus, after clean-up data analysis included 2730 features distributed as FIA-ESI(+)-MS: 144; FIA-ESI(−)-MS: 201; HILIC-ESI(−)-QTOF-MS: 414; LC-ESI (+)-QTOF-MS: 309; lipidomics LC-ESI(+)-QTOF-MS: 943; and lipidomics LC-ESI(−)-QTOF-MS: 719. A mid-level data fusion strategy was used merging the six data sets to identify the most relevant metabolites and pathways altered during $CD8^+$ T cell activation.

2.3. Pathway and network analysis

Pathway Analysis combined results from enrichment analysis with topology analysis to support the identification of relevant pathways in the simulated data sets [18]. For the analysis of metabolomic data collected during $CD8^+$ T cell activation, functional metabolic analysis was used to extract biological information within relevant networks using the *mummichog* algorithm [20] with a common *m/z* accuracy of 10 ppm for all data sets. *p*-values used as input for functional analysis were estimated from univariate *t*-tests from the comparison of the distribution of relative intensities between 0 h and each time point between 12 and 96 h. Results from functional analysis were summarized using the *p*-values calculated from the enrichment analysis and the pathway impact value calculated from pathway topology analysis estimated using the *mummichog* algorithm.

2.4. Software and statistics

Functional analysis was carried out in MetaboAnalyst 5.0 [18] (http://www.metaboanalyst.ca) using the human KEGG pathway database [23]. Student's *t*-tests assessed the null hypothesis that the data of two groups (e.g., 0 h vs 96 h) came from independent, random samples with equal means with unknown and unequal variances at the 5% significance level. PCA was carried out using autoscaled data, and MCR-ALS using the unscaled values. Data analysis was carried out in MATLAB 2021a (Mathworks Inc., Natick, MA, USA) using in-house written scripts, the PLS Toolbox 8.9 (Eigenvector Research Inc., Wenatchee, USA), and the MCR-ALS toolbox (https://mcrals.wordpress.com/download/mcr-als-2-0-toolbox/, accessed on October 1, 2021) [24].

The raw MS and pre-processed data [22] was downloaded from www.ebi.ac.uk/metabolights/MTBLS2145.

3. Results and discussion

3.1. Simulated data set

To test the applicability of MCR-ALS and pathway analysis for identifying dynamic components and infer metabolic perturbations in longitudinal data, we first applied it to a simulated data set. Fig. 1A shows the simulated effects modifying the concentrations of three subsets of ten metabolites included in the Tryptophan (effect 1), Phenylalanine (effect 2), and Glyoxate and dicarboxylate (effect 3) metabolisms. The intensities of the subsets of altered metabolites as a function of time in the simulated data with very low ($\mathbf{X_s}$) and high noise levels ($\mathbf{X_{sn}}$) are depicted in Fig. 1B and C, respectively. Pathway Analysis was carried out using as input a list of altered metabolites identified using univariate *t*-tests and pre-defined thresholds for fold change and *p*-values. In this analysis, metabolites were classified as altered at a given time point *i* if two conditions were observed: i) p-value<0.05 for a *t*-test comparing the distribution of intensities at $t = i$ with that observed at $t = 1$ as reference, and ii) an absolute shift in the mean value > 0.2. Reference results from the analysis of $\mathbf{X_s}$ depicted in Fig. 2A, showed that, as expected, the simulated effects on subsets of metabolites of the three selected metabolic pathways lead to statistically significant perturbations in those pathways, with minimum *p*-values and maximum impacts matching the position of the maxima of the effects. The Tryptophan pathway was only found altered between $t = 4$ and 6. In the case of Glyoxate and dicarboxylate metabolism, a slightly larger perturbation (*i.e.*, larger $-\log_{10}(p\text{-values})$) was observed between $t = 10$ and 16, due to the overlap with the Phenylalanine metabolism modified by the effect 2, and the Phenylalanine metabolism also showed larger $-\log_{10}(p\text{-values})$ between $t = 4$ and 8. No effect was observed (p-values>0.05, impact $= 0$) in the Glutathione and Steroid hormone biosynthesis pathways that were not modified by the effects. In the presence of a higher noise level ($\mathbf{X_{sn}}$) the power to accurately infer pathways activities was reduced (see Fig. 2C). In the case of the Tryptophan pathway, where the simulated effect was less intense than in the case of the Phenylalanine (effect 2) and Glyoxylate and dicarboxylate (effect 3), it lead to the identification of a 'false positive'. At $t = 2$ results from pathway analysis indicated a significant alteration (p-values<0.05) not observed in $\mathbf{X_s}$ (see Fig. 2B). Besides, the evolution of the impact was noisy and did not follow the trend observed in the data set used as reference ($\mathbf{X_s}$). Results for the Phenylalanine metabolism in $\mathbf{X_{sn}}$ (Fig. 2C) showed 'false negatives' (p-value>0.05) between $t = 14$ and 17, in contrast to the statistically significant p-values<0.05 observed in $\mathbf{X_s}$. Phenylalanine pathway impact values did not follow the trend observed in $\mathbf{X_s}$, and impact values > 0 were found at $t =$ 2 and 17. Results for the analysis of the Glyoxylate metabolism in $\mathbf{X_{sn}}$, showed some differences compared to those found in $\mathbf{X_s}$ including unstable *p*-values at $t > 10$, and a large impact and low *p*-value at $t = 9$. Furthermore, the Steroid hormone biosynthesis pathway was found altered at $t = 5$.

Then, the $\mathbf{X_{sn}}$ data set was analyzed by MCR-ALS. During MCR-ALS optimization, the model is fitted with a predefined number of components using initial estimates of either the **C** or the **S** matrix. In this case, three components were selected from $\log_{10}$(eigenvalues) obtained from singular value decomposition (SVD) (see Fig. 3). The MCR-ALS initial concentration estimates (**C**) were obtained using the *pure* algorithm. This algorithm assumes that the data set includes variables related to a single spectral component. Alternative strategies such as the use of Independent Component Analysis (ICA), or EFA could be used [25]. Fig. 4A shows the resolved concentration (i.e., **C**) profiles estimated by MCR-ALS after 51 iterations (96% variance explained (R^2) at the optimum) using unimodality (10% tolerance) in the concentration profiles of the three components as constraint. Besides, during ALS optimization, an equality constrain was applied in the concentration profile to fix the initial (*i.e.*, basal state) concentration of the three components to zero. Data depicted showed that the three components had concentration peaks at $t \sim 6$, 10, and 15, respectively, matching the theoretical peak

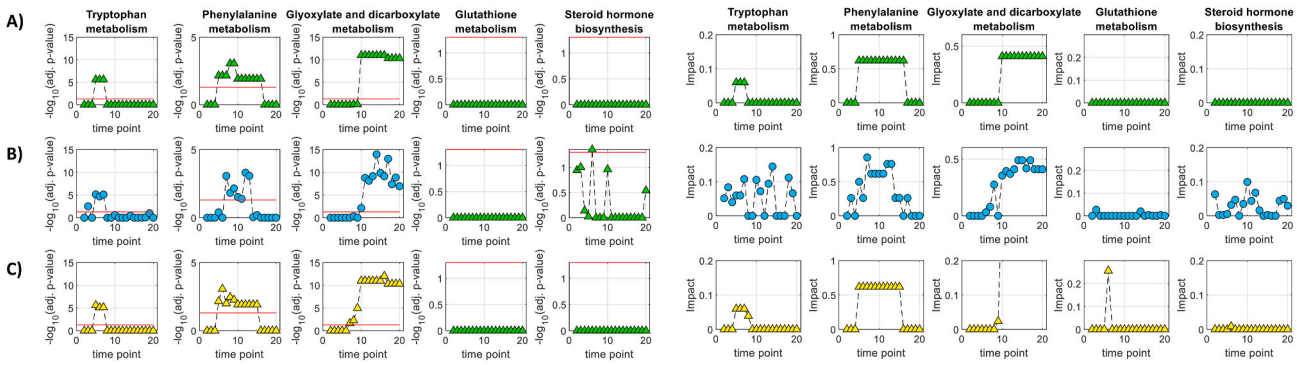

Fig. 2. Results from pathway analysis using data t = 1 as reference and *t*-test for the identification of differential metabolites (*p*-values<0.05) in $\mathbf{X_s}$ (A), $\mathbf{X_{sn}}$ (B) and MCR-ALS reconstructed $\mathbf{CS^T}$ matrix (C). Results show the evolution of the significance (left) and impact (right) of five metabolic pathways. Tryptophan, Phenylalanine and Glyoxylate and dicarboxylate pathways were altered by effects 1, 2, and 3 respectively. The Glutathione and Steroid hormone biosynthesis pathways were randomly simulated, and no alteration was expected at any time point.

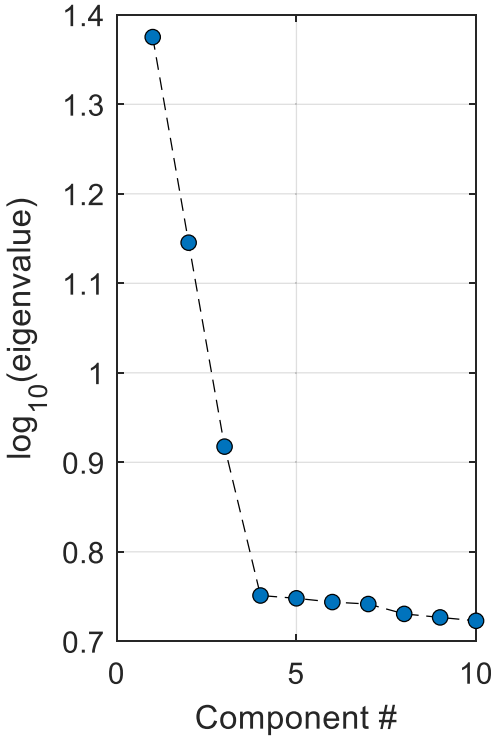

Fig. 3. $\log_{10}$(eigenvalues) obtained from the singular value decomposition (SVD) analysis of $\mathbf{X_{sn}}$.

maxima of the perturbations introduced, respectively. Fig. 4B shows the spectral profiles of the MCR-ALS components. Metabolites of Tryptophan, Phenylalanine, and Glyoxate and dicarboxylate modified by effects 1, 2, and 3 respectively (see Fig. 1A), showed values close to 1 in **S**, in agreement with the artificially introduced effects. Fig. 1D shows data reconstructed using the three low-order MCR-ALS components (i.e., $\mathbf{D} = \mathbf{CS^T}$). Data reconstruction led to a reduction in the noise levels measured as the root mean square value of the intensities of the set of non-altered metabolites, from 0.20 in $\mathbf{X_{sn}}$ to 0.04 in **D**. Furthermore, results depicted in Fig. 2C showed that in this case, the use of reconstructed data **D** also provided an improve in the accuracy of pathway analysis. Estimated *p*-values and impacts matched more closely to those estimated in the reference data ($\mathbf{X_s}$) in the three altered pathways, as well as in the unaltered Glutathione metabolism and Steroid hormone biosynthesis pathways, with no false positive or false negative perturbed time points. These results show that MCR-ALS can be used to identify and isolate dynamic sources of variation in the data set, thus facilitating downstream pathway analysis. MCR methods suffer from non-uniqueness of their results and instrumental noise propagates uncertainty to the bilinear solutions. However, the analysis of the set of feasible profiles in the bilinear decomposition of the metabolic data as a function of instrumental noise can be analyzed using *e.g.*, MCR-BANDS or N-BANDS as described elsewhere [26].

3.2. *Metabolic dynamics of in vitro CD8$^+$ T cell activation*

PCA was employed to identify dynamic components in the data set. Fig. 5A shows the PCA scores as a function of time of the first three PCs explaining 69% of the variance. PC1 and PC2 showed dynamic trends with maxima at 48 and 36 h, respectively. Then, functional metabolic network analysis was used to retrieve biological information within relevant networks using the *mummichog* algorithm, the set of 2730 metabolic features with an *m/z* accuracy of 10 ppm, and the KEGG library of pathways [20]. Metabolic profiles from samples collected at $t =$ 0 h were selected as reference and they were compared to those collected during T-cell activation over the following 8 time points. Fig. 6A shows the list of enriched pathways at each time point. In spite of the different strategies employed, results from PCA and functional analysis agreed with previously reported results showing that metabolite concentrations changed in a time-dependent way. For example, in the original study [22], the separate analysis of FIA-ESI(±)-MS data indicated that the major metabolic differences occurred in the first 48 h of CD8$^+$ T cell activation, in agreement with the large shift in the PC1 scores in that time window (see Fig. 5A). Also, results from functional analysis in Fig. 6A show a larger number of altered pathways within the first 24 h of CD8$^+$ T cell activation, and different patterns of altered pathways along the activation process. However, results obtained from functional analysis did not correlate with the scores obtained by PCA of the longitudinal data set. In this case, the analysis of the correlation between the outcome from functional analysis for each metabolic pathway and the PCA scores did not show any statistically significant (*p*-value<0.01) linear association (data not shown).

Then, MCR-ALS was applied to the $\mathbf{T_{cell}}$ data set. The MCR-ALS model was fitted using concentration profiles obtained using the *pure* algorithm as initial concentration estimates. Fig. 5B shows the resolved concentration profiles estimated by MCR-ALS for three components after 17 iterations (16% variance explained (R^2) at the optimum) using unimodality (10% tolerance) in the concentration profiles as unique constraint. Data depicted show that the three components had concentration peaks at $t = 12$, 48, and 48 h, respectively, with overlapping and markedly different profiles to those observed by PCA (see Fig. 5A). The

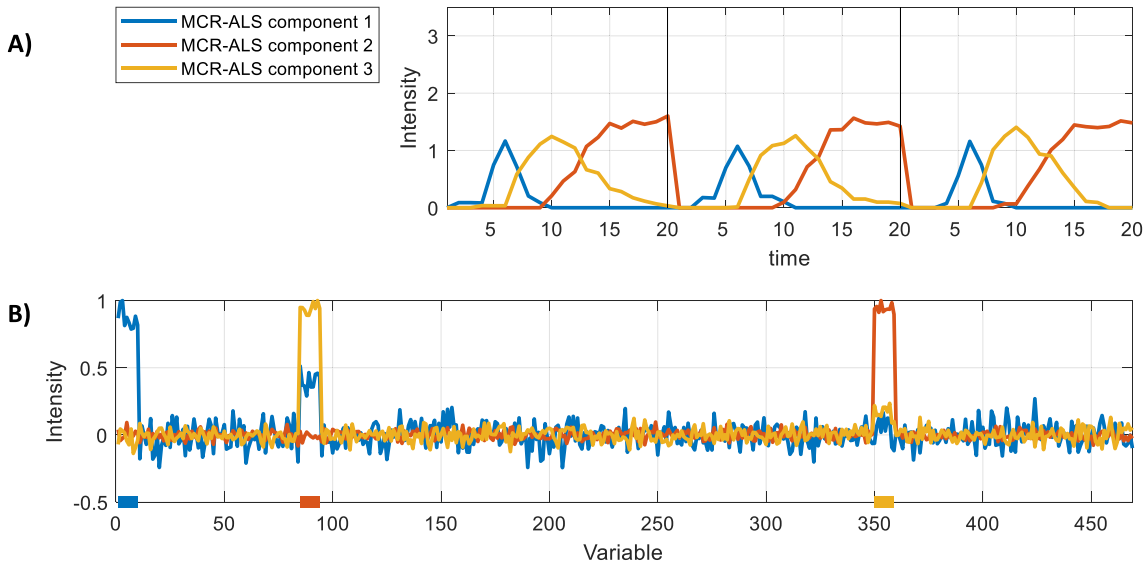

Fig. 4. MCR-ALS analysis of the simulated data set $\mathbf{X}_{sn}$. Concentration (**C**, top) and spectral (**S**, bottom) profiles estimated by MCR-ALS after 51 iterations (96% variance explained (R^2) at the optimum). Note: blue, orange and yellow horizontal bars in the bottom figure at intensity = −0.5 indicate the variables affected by the first, second, and third simulated effects. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

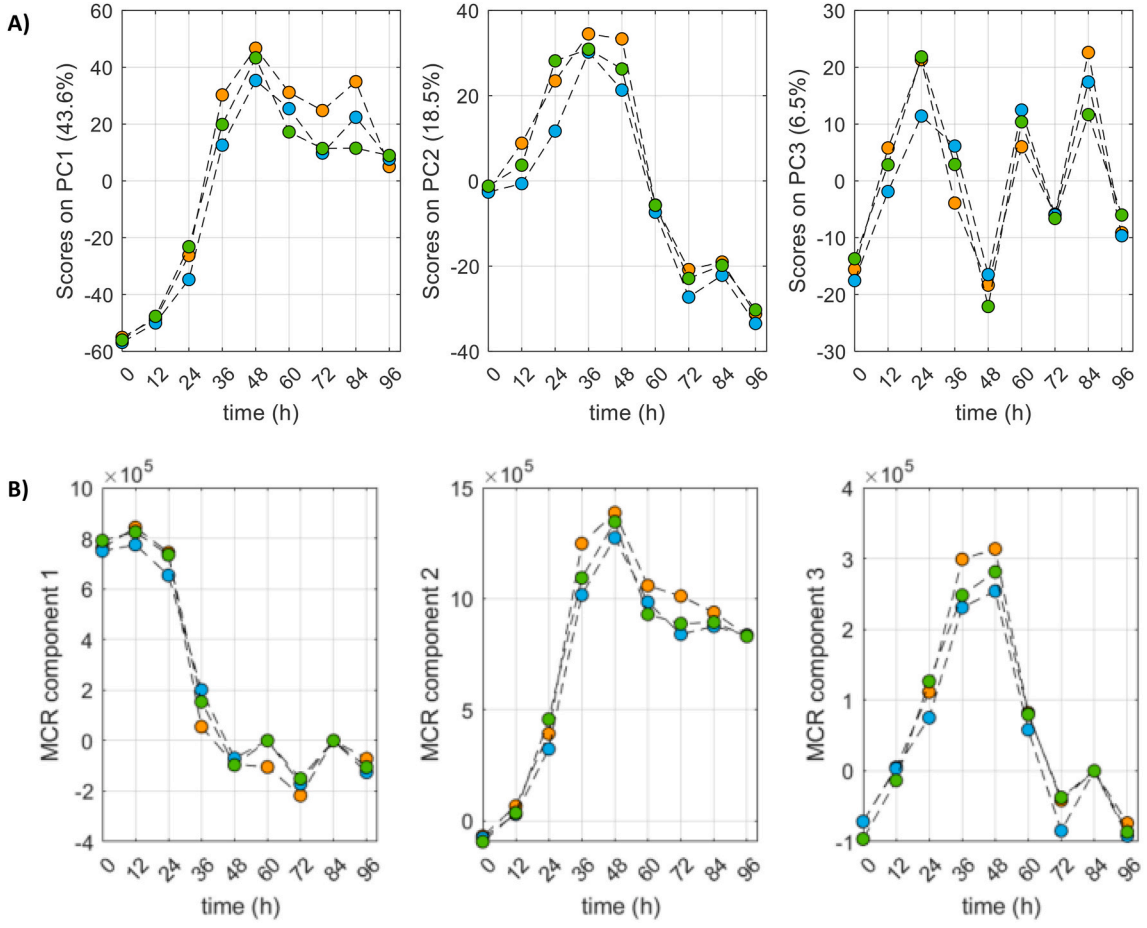

Fig. 5. A) PCA scores plot for the first three principal components, from the analysis of pre-processed data obtained during $CD8^+$ T cell activation using six complementary analytical platforms. B) MCR-ALS concentration profiles obtained after 17 iterations.

intensity of the first component decreased during the first 36 h, and then remained approximately stable until the end of the study. The second component increased over the first 48 h and then decreased, reaching a plateau at 60 h. The third component gradually increased its intensity the first 48 h and then decrease down to approximately the initial values after 72 h of T-cell differentiation.

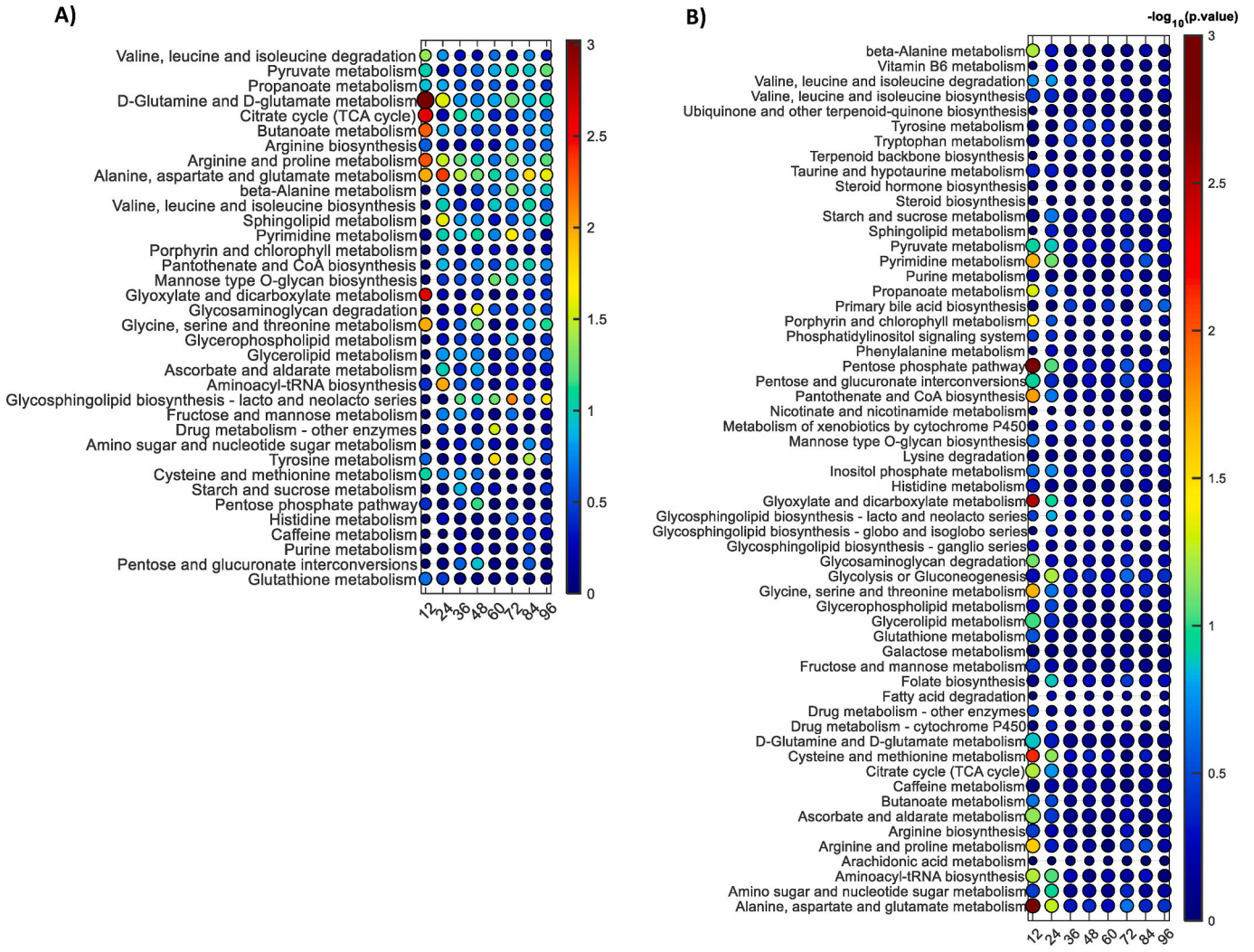

Fig. 6. Pathway analysis results from the analysis of $CD8^+$ T cell activation with six complementary analytical platforms using pre-processed data (A) and the MCR-ALS reconstructed $\mathbf{CS}^{\mathbf{T}}$ matrix (B).

Then, functional metabolic analysis was used to extract biological information as described above. Using the MCR-ALS reconstructed metabolic profiles $\mathbf{D} = \mathbf{CS}^{\mathbf{T}}$, samples collected at $t = 0$ h were compared to those collected during $CD8^+$ T-cell activation over the following eight time points. Results in this case show a larger number of altered pathways within the first 24 h of $CD8^+$ T cell activation than later, and different patterns of altered pathways along the activation process (see Fig. 6B), in agreement with results obtained from the analysis of the initial data matrix $\mathbf{T}_{\mathbf{cell}}$ shown in Fig. 6A]. Moreover, results obtained from the analysis of the MCR-ALS reconstructed data **D** showed pathways displaying statistically significant (p-value<0.01) correlation with the concentration profiles obtained by MCR-ALS. In particular, five pathways (Aminoacyl-tRNA biosynthesis; Glycine, serine and threonine metabolism; Citrate cycle (TCA cycle); Cysteine and methionine metabolism; and Pantothenate and CoA biosynthesis) showed positive significant correlations with the median intensities of the first MCR-ALS component, and two pathways (Metabolism of xenobiotics by cytochrome P450; and Tryptophan metabolism) were correlated with the third MCR-ALS component (see Fig. 7).

4. Conclusions

Results obtained using MCR-ALS in simulated and real data showed the utility of this approach for the interpretation of metabolic changes MS-based longitudinal data sets. Furthermore, it is a non-parametrical approach that only assumes a linear relationship between the concentration and spectral profiles of a limited number of components. However, the type of initial estimates might influence the MCR-ALS factorization depending on the data structure. The identification of metabolic pathways associated with each detected dynamic component can support the understanding of overlapping asynchronous metabolic programming processes in a biological context.

Author statement

Isabel Ten-Domenech: Formal analysis, Investigation, Writing – review & editing. Marta Moreno-Torres: Investigation, Writing – review & editing. Juan Daniel Sanjuan-Herráez: Investigation, Writing – review & editing. David Pérez-Guaita: Conceptualization, Data curation, Writing – review & editing. Guillermo Quintás: Conceptualization, Software, Formal analysis, Investigation, Writing – review & editing. Julia Kuligowski: Conceptualization, Investigation, Data curation, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

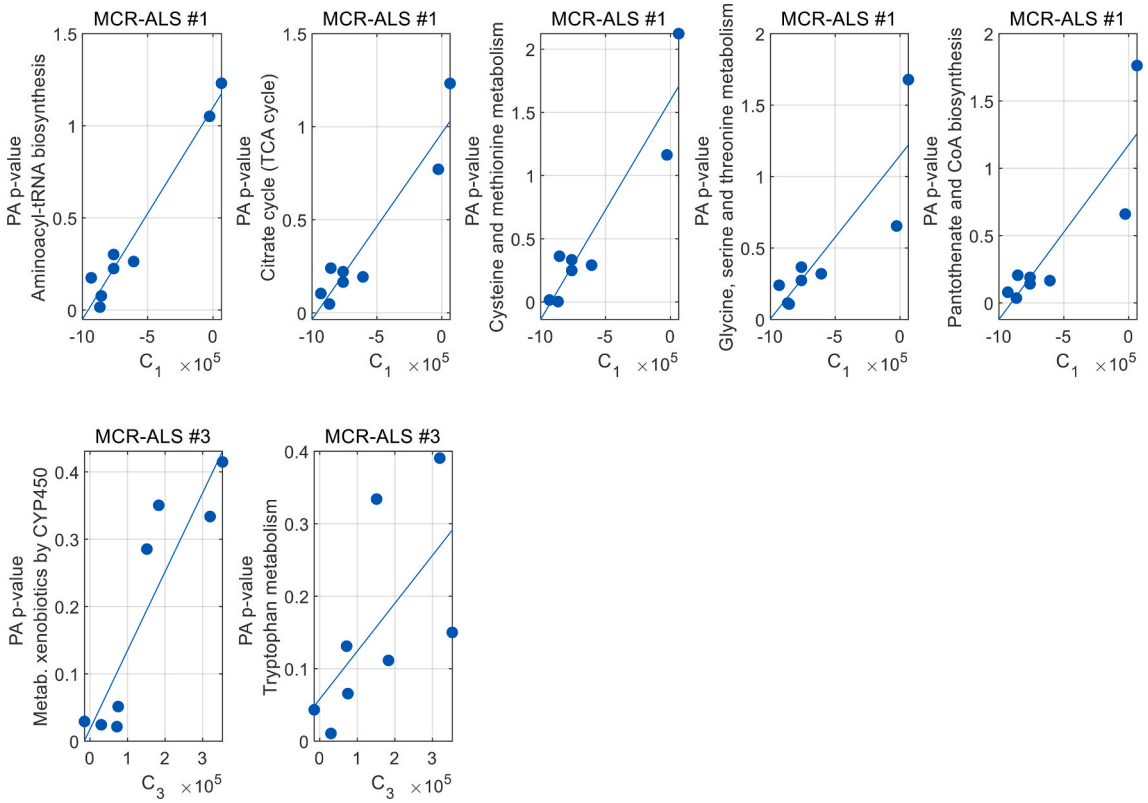

Fig. 7. Linear correlation between *p*-values obtained from pathway analysis using the MCR-ALS reconstructed data **D** and the MCR-ALS concentration profiles of the first (top-row) and the third (bottom row) MCR-ALS component.

the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

GQ acknowledges grant PID2021-125573OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF A way of making Europe. MMT acknowledges support from grant IJC2018-036209-I funded by MCIN/AEI/10.13039/501100011033. JK and ITD acknowledge support received from *Instituto de Salud Carlos III* (Spain) and co-funded by European Regional Development Fund "A way to make Europe" with grant numbers CPII21/00003, and PI20/00964 and CD19/00176. ITD acknowledges financial support from the *Generalitat Valenciana* (GV/2021/186). DPG and MMT acknowledge the financial support from RYC2019-026556-I and RYC2021-031346-I.

References

[1] P. Sperisen, O. Cominetti, F.-P.J. Martin, Longitudinal omics modeling and integration in clinical metabonomics research: challenges in childhood metabolic health research, Front. Mol. Biosci. 2 (2015) 44, https://doi.org/10.3389/fmolb.2015.00044.

[2] T.K. Karakach, R. Knight, E.M. Lenz, M.R. Viant, J.A. Walter, Analysis of time course 1H NMR metabolomics data by multivariate curve resolution, Magn. Reson. Chem. MRC. 47 (Suppl 1) (2009) S105–S117, https://doi.org/10.1002/mrc.2535.

[3] M. Björklund, Be careful with your principal components, Evolution 73 (2019) 2151–2158, https://doi.org/10.1111/evo.13835.

[4] Principal component analysis. https://link.springer.com/book/10.1007/b98835. (Accessed 13 July 2022).

[5] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philos. Trans. R. Soc. Math. Phys. Eng. Sci. 374 (2016), 20150202, https://doi.org/10.1098/rsta.2015.0202.

[6] Á. Sánchez-Illana, J.D. Piñeiro-Ramos, J.D. Sanjuan-Herráez, M. Vento, G. Quintás, J. Kuligowski, Evaluation of batch effect elimination using quality control replicates in LC-MS metabolite profiling, Anal. Chim. Acta 1019 (2018) 38–48.

[7] Willem Windig, Jean Guilment, Interactive self-modeling mixture analysis, Anal. Chem. 63 (1991) 1425–1432, https://doi.org/10.1021/ac00014a016.

[8] H. Gampp, M. Maeder, C.J. Meyer, A.D. Zuberbuehler, Evolving factor analysis of spectrophotometric titrations: forget about the law of mass action, Chim. Switz. 39 (1985) 10. https://www.osti.gov/etdeweb/biblio/7150586. (Accessed 7 November 2022).

[9] H. Abdollahi, R. Tauler, Uniqueness and rotation ambiguities in multivariate curve resolution methods, Chemometr. Intell. Lab. Syst. 108 (2011) 100–111, https://doi.org/10.1016/j.chemolab.2011.05.009.

[10] R. Tauler, A. Smilde, B. Kowalski, Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution, J. Chemom. 9 (1995) 31–58, https://doi.org/10.1002/cem.1180090105.

[11] R.R. de Oliveira, K.M.G. de Lima, R. Tauler, A. de Juan, Application of correlation constrained multivariate curve resolution alternating least-squares methods for determination of compounds of interest in biodiesel blends using NIR and UV–visible spectroscopic data, Talanta 125 (2014) 233–241, https://doi.org/10.1016/j.talanta.2014.02.073.

[12] F. Puig-Castellví, I. Alfonso, R. Tauler, Untargeted assignment and automatic integration of 1H NMR metabolomic datasets using a multivariate curve resolution approach, Anal. Chim. Acta 964 (2017) 55–66, https://doi.org/10.1016/j.aca.2017.02.010.

[13] E. Ortiz-Villanueva, F. Benavente, B. Piña, V. Sanz-Nebot, R. Tauler, J. Jaumot, Knowledge integration strategies for untargeted metabolomics based on MCR-ALS analysis of CE-MS and LC-MS data, Anal. Chim. Acta 978 (2017) 10–23, https://doi.org/10.1016/j.aca.2017.04.049.

[14] E. Gorrochategui, J. Jaumot, R. Tauler, ROIMCR: a powerful analysis strategy for LC-MS metabolomic datasets, BMC Bioinf. 20 (2019) 256, https://doi.org/10.1186/s12859-019-2848-8.

[15] M. Pérez-Cova, C. Bedia, D.R. Stoll, R. Tauler, J. Jaumot, MSroi: a pre-processing tool for mass spectrometry-based studies, Chemometr. Intell. Lab. Syst. 215 (2021), 104333, https://doi.org/10.1016/j.chemolab.2021.104333.

[16] H. Motegi, Y. Tsuboi, A. Saga, T. Kagami, M. Inoue, H. Toki, O. Minowa, T. Noda, J. Kikuchi, Identification of reliable components in multivariate curve resolution-alternating least squares (MCR-ALS): a data-driven approach across metabolic processes, Sci. Rep. 5 (2015), 15710, https://doi.org/10.1038/srep15710.

[17] I. Ten-Doménech, M. Moreno-Torres, J.V. Castell, G. Quintás, J. Kuligowski, Extracting consistent biological information from functional results of metabolomic pathway analysis using the Mantel's test, Anal. Chim. Acta 1187 (2021), 339173, https://doi.org/10.1016/j.aca.2021.339173.

[18] Z. Pang, J. Chong, G. Zhou, D.A. de Lima Morais, L. Chang, M. Barrette, C. Gauthier, P.-É. Jacques, S. Li, J. Xia, MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights, Nucleic Acids Res. (2021), https://doi.org/10.1093/nar/gkab382.

[19] C. Wieder, C. Frainay, N. Poupin, P. Rodríguez-Mier, F. Vinson, J. Cooke, R.P. Lai, J.G. Bundy, F. Jourdan, T. Ebbels, Pathway analysis in metabolomics: recommendations for the use of over-representation analysis, PLoS Comput. Biol. 17 (2021), e1009105, https://doi.org/10.1371/journal.pcbi.1009105.

[20] S. Li, Y. Park, S. Duraisingham, F.H. Strobel, N. Khan, Q.A. Soltow, D.P. Jones, B. Pulendran, Predicting network activity from high throughput metabolomics, PLoS Comput. Biol. 9 (2013), https://doi.org/10.1371/journal.pcbi.1003123.

[21] J. Xia, D.S. Wishart, MetPA: a web-based metabolomics tool for pathway analysis and visualization, Bioinforma. Oxf. Engl. 26 (2010) 2342–2344, https://doi.org/10.1093/bioinformatics/btq418.

[22] J. Edwards-Hicks, M. Mitterer, E.L. Pearce, J.M. Buescher, Metabolic dynamics of in vitro CD8+ T cell activation, Metabolites 11 (2020) 12, https://doi.org/10.3390/metabo11010012.

[23] M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Res. 28 (2000) 27–30, https://doi.org/10.1093/nar/28.1.27.

[24] J. Jaumot, A. de Juan, R. Tauler, MCR-ALS GUI 2.0: new features and applications, Chemometr. Intell. Lab. Syst. 140 (2015) 1–12, https://doi.org/10.1016/j.chemolab.2014.10.003.

[25] L. Valderrama, R.P. Gonçalves, P.H. Março, D.N. Rutledge, P. Valderrama, Independent components analysis as a means to have initial estimates for multivariate curve resolution-alternating least squares, J. Adv. Res. 7 (2016) 795–802, https://doi.org/10.1016/j.jare.2015.12.001.

[26] A.C. Olivieri, K. Neymeyr, M. Sawall, R. Tauler, How noise affects the band boundaries in multivariate curve resolution, Chemometr. Intell. Lab. Syst. 220 (2022), 104472, https://doi.org/10.1016/j.chemolab.2021.104472.