Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Network propagation in bipartite metabolite–reaction graphs for metabolomic data exploration

Julia Kuligowski
Instituto de Investigación Sanitaria La Fe
Marta Moreno-Torres
Universitat de València
David Pérez-Guaita
Universitat de València
Francesc Albert Esteve-Turrillas
Universitat de València
Guillermo Quintas
gquintas@leitat.org

Leitat Technological Center

Research Article

Keywords: metabolomics, multivariate analysis, data exploration, network propagation, Partial Least Squares – Discriminant Analysis (PLS-DA)

Posted Date: July 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10129122/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Metabolomics on September 16th, 2026. See the published version at https://doi.org/10.1007/s11306-026-02529-y.

Abstract

Introduction:

Interpretation of metabolomic data is frequently limited by incomplete metabolite coverage and the predefined representation of biochemical organization provided by pathway-based approaches.

Objectives

This study aims to develop and evaluate a network propagation framework to facilitate metabolomic data exploration by diffusing statistical relevance scores across a metabolic topology.

Methods

A directed bipartite metabolite–reaction network was reconstructed using the KEGG database, comprising 8236 nodes with 1601 metabolites and 6635 reactions. Simulated metabolomic datasets with 75% sparsity mimicking two metabolic perturbations (oxidative stress and mitochondrial dysfunction) and two no-effect control groups, were generated to test the strategy. Node-level statistical signal scores were distributed across the network topology using a random walk-based diffusion operator combined with a semi-supervised clamping procedure to preserve initially observed measurements.

Results

Network propagation redistributed statistical signals across connected biochemical regions, allowing unmeasured metabolites and reactions to accumulate topology-aware significance scores. For both OS and MD scenarios, supervised univariate analyses revealed that significantly altered nodes formed highly interconnected, non-random topological modules. Furthermore, PLS-DA demonstrated that network propagation significantly enhanced multivariate predictive performance, yielding statistically validated performance gains via permutation testing, while effectively managing background noise in null comparisons.

Conclusions

The proposed bipartite metabolite–reaction network propagation framework provides an exploratory tool for metabolomics. By integrating topological context with statistical relevance, this approach complements established pathway analyses for hypothesis generation in partially observed metabolic systems.

1. Introduction

Multivariate modeling has become a central tool in metabolomics for capturing coordinated biochemical variation across large sets of metabolites associated with physiological states or external interventions. However, the interpretability of these models remains challenging. While machine learning approaches often provide strong predictive performance, they typically offer limited mechanistic insight into the biological processes underlying model outputs. This has driven increasing interest in explainable artificial intelligence (XAI) methods[1, 2], including perturbation-based strategies, which aim to identify relevant variables or groups of variables by assessing the impact of controlled input modifications on model predictions.

Interpretation of metabolomic data is commonly facilitated by grouping metabolites into biologically meaningful sets, most often defined by metabolic pathways. This approach has proven useful for linking multivariate model behavior to known biochemical functions[3]. However, pathway-based representations present several limitations. Pathways are predefined, often inconsistently curated across databases, and do not fully capture the interconnection and overlap of metabolic processes[4, 5]. Moreover, metabolites frequently participate in multiple pathways, complicating the attribution of functional relevance, and metabolomic coverage is incomplete, thus limiting the interpretability of pathway analyses[6].

To address these limitations, recent work has proposed grouping metabolites based on topological properties of metabolic networks, such as community structures derived from graph-based analyses[7]. This network-driven approach enables the identification of modules that reflect biochemical connectivity and might be complementary to the use of predefined pathways. Current approaches typically rely on metabolite-metabolite (unipartite) networks, where edges represent inferred relationships between metabolites. However, metabolism is fundamentally organized around enzyme-catalyzed reactions, which define the transformation of substrates into products and encode the directionality and mechanism of biochemical processes. Unipartite graphs collapse these relationships, not including the directionality and constraints of multi-substrate reaction dependencies. Representing metabolic systems as bipartite networks, where metabolites and reactions are explicitly modeled as distinct node types, preserves the relationships between metabolites and reactions and reaction dependencies.

As mentioned before, technical limitations in metabolomics result in the incomplete coverage of the metabolic networks. Due to technical constraints in detection and annotation, only a subset of metabolites present in the biological system can be detected and a fraction of them annotated in untargeted studies, resulting in partially observed network states. This sparsity can bias downstream analyses, particularly when estimating the contribution of specific network structures, whose relevance may be underestimated.

Network propagation refers to a class of graph-based methods that diffuse node-associated information across a network according to its connectivity structure[8]. Signals assigned to individual nodes are distributed to neighboring nodes, enabling the integration of local measurements with global network

topology. Network propagation enables the integration of biological signals across interaction networks, facilitating the characterization of biological mechanisms and providing insight into biological processes[8]. Over the last years, a number of bioinformatic techniques have been proposed for genetic, proteomic and other omic analysis using networks, based on random walks, information diffusion, and electrical resistance among others. These strategies have been successfully applied to identify disease genes, genetic modules and drug targets[9, 10]. Although network propagation methods are frequently used in e.g., genomics, transcriptomics, and proteomics, their application in metabolomics is still limited, particularly in the context of bipartite metabolite–reaction networks.

In this work, we describe an adaptation of network propagation to bipartite metabolite-reaction networks for metabolomic data exploration[11, 12]. In this approach, the signal associated with each node does not represent metabolite concentration or abundance, but rather a measure of statistical relevance, a so called statistical score (SC), derived from univariate or multivariate analyses (e.g., –log p-values from differential testing). The strategy utilizes a random walk-based smoothing approach combined with a semi-supervised clamping strategy to propagate SCs while preserving initially observed values. This enables the estimation of metabolite and reaction-level propagated scores, derived from both direct contributions of neighboring metabolites and indirect contributions mediated through the broader network topology. Under this formulation, node values describe the strength of association with a given condition, and network propagation can be interpreted as the diffusion of that relevance through the network topology.

By leveraging connectivity patterns, this approach allows nodes that are topologically close to statistically significant metabolites to accumulate signal, thereby identifying metabolites and reactions that, while not experimentally detected or annotated, may be functionally relevant due to their network context. This strategy does not aim to reconstruct missing concentrations, but to enable the exploration of statistical results beyond the directly observed metabolome, guiding the analysis with well-known information about metabolite and reactions.

Using simulated bipartite metabolite–reaction networks describing different perturbation scenarios, results show that the proposed approach might support the interpretation of metabolomic data, providing a tool for hypothesis generation and data exploration.

2. Materials and methods

2.1 Bipartite Reaction-Metabolite network

A bipartite reaction-metabolite network reconstruction was performed using the KEGG REST API (https://www.kegg.jp/kegg/rest/) to retrieve compound–reaction associations and reaction-level annotations. KEGG compound identifiers (C#####) were queried in batch mode, and associated reactions (R#####) were extracted from compound entries, followed by retrieval of full reaction equations. Reaction reversibility was stored as an edge attribute but was not used for data analysis. To

improve biological relevance and interpretability, metabolites were filtered to retain 1813 participating in at least one of a curated set of 57 metabolic pathways summarized in **Table SM1**, while excluding others such as drug metabolism, xenobiotic degradation, and other peripheral pathway classes. Additional filtering steps were applied to remove currency metabolites (e.g., ATP, NADH, H_2O) known to introduce non-specific connectivity, as well as compounds with low molecular weight (< 50 Da), thereby reducing topological bias and improving modular structure, leaving a total of 1601 metabolites.

A directed bipartite graph was then constructed, in which metabolites and reactions were represented as distinct node classes. For each reaction, directed edges were defined from substrate metabolites to the reaction node and from the reaction node to product metabolites. To ensure consistency between node and edge sets, only metabolites retained after filtering were allowed to form edges, while all associated reaction nodes were preserved. The resulting network was exported as node and edge tables for topological analysis in Gephi 0.10.1. The network resulted in a graph with 8236 nodes (1601 metabolites and 6635 reactions) and 11868 edges and it was visualized using the Force Atlas 2 layout with a scaling factor = 2, and the 'dissuade hubs' and 'prevent overlap' options enabled.

2.2 Simulated metabolomic data

To evaluate the performance of the proposed method, synthetic datasets were generated. While experimental metabolomic datasets offer real-world complexity, they lack a verifiable ground truth regarding which specific metabolites, reactions, or pathways are truly altered. Network propagation processes redistribute signal across the topology, and so evaluating downstream classification improvements without a known baseline risks confounding true biological signal with topological smoothing artifacts. Simulated data provides a kown ground truth (e.g., altered vs. non altered nodes) enabling the statistical properties of the algorithm to be evaluated independently of experimental confounding factors and allowing accurate estimation of sensitivity and specificity.

Simulated metabolomic datasets were generated to represent two simulated metabolic perturbation scenarios, i.e. oxidative stress (OS) and mitochondrial dysfunction (MD), and two no-effect scenarios (NE1 and NE2). The NE1 and NE2 groups were simulated using identical procedures but independent random seeds, ensuring that they share the same statistical properties while containing no systematic differences to enable the evaluation of null comparisons (NE1 vs NE2). OS represents a multi-pathway perturbation involving high-degree core hubs such as L-glutamate and glutathione that link peripheral amino acid and redox subsystems. On the other hand, MD models a highly concentrated, localized perturbation focused on the central energy-metabolism backbone and highly connected nodes (e.g., Acetyl-CoA and 2-oxoglutarate).

Each scenario (OS, MD, NE1, and NE2) consisted of 15 independent simulated studies comparing a control population with a perturbed population. For example, in the MD scenario, each study could represent an experiment assessing the metabolomic effects of a different drug known to induce mitochondrial dysfunction in a biological system.

The simulated datasets did not consist of metabolite abundance measurements. Instead, they consisted of metabolite-level statistical scores (SCs describing the results of each metabolomic study. This is, for each study a SC per metabolite, analogous to a − log(p-value) obtained from a differential analysis comparing case and control groups, was simulated (see Fig. 1). SCs were simulated under controlled sparsity (25%), such that 75% of the metabolic nodes of the KEGG metabolic network were consistently set to zero. The subset of 400 detected (non-zero) metabolic nodes was shared across all conditions and replicates, ensuring a common feature space and identical missingness structure. This shared set included all metabolites included in Table 1.

Table 1
Sets of metabolites simulating perturbations associated with each biological mechanism selected.

Effect	**Systematically altered metabolites**
oxidative stress	C00025 (Glutamic), C00037 (Glycine), C00051 (GSH), C00127 (GSSG), C00212 (Adenosine), C00242 (Guanine), C00294 (Inosine), C00387 (Guanosine), C01762 (Xanthosine), C01879 (Pyroglutamic)
mitochondrial dysfunction	C00022 (Pyruvic), C00024 (Acetyl-CoA), C00026 (Oxoglutaric), C00042 (Succinic), C00122 (Fumaric), C00149 (Malic), C00158 (Citric), C00186 (Lactic), C00424 (Lactaldehyde)

Baseline values for the ‘detected’ (i.e. non-zero) metabolites were randomly sampled from a uniform distribution on [0, 1] and then -log transformed. Condition-specific metabolic profiles were then introduced by selectively perturbing the metabolites associated with each mechanism summarized in Table 1. For these metabolites, SCs were reassigned by randomly sampling from a lower value uniform distribution ([0.015, 0.2]) and then -log transformed, generating subtle but structured differences relative to baseline. The sets of perturbed metabolites differed between OS and MD conditions, and no systematic perturbation was applied to any metabolite in the NE1 and NE2 vs control metabolic profiles. This design ensured consistent metabolite detection across all samples while including biological differences through condition-specific perturbation of subsets of metabolites participating in selected mechanisms. As a result, inter-condition variation was driven by structured signal differences, and intra-condition variability arose from random intensity sampling. Reaction nodes were initialized to zero prior to propagation.

Figure 2 shows the distribution of signals for boosted metabolites across conditions. Results depicted show that the induced change in the metabolic profiles was subtle, and the statistically significant difference relative to control (NE2) (two-sample t-test, $p < 0.05$) was found only in a subset of the metabolites in each condition.

SCs obtained from 15 independent ‘case vs control’ studies per condition are shown for oxidative stress (OS, blue), mitochondrial dysfunction (MD, green), and ‘no effect’ (NE1 (red) and NE2 (orange)). For each condition, vertical lines represent the 5th–95th percentile range, and thick bars indicate the interquartile range (25th–75th percentiles). Circles denote the mean SC value. Filled black circles indicate a

statistically significant difference relative to NE2 (two-sample t-test, $p < 0.05$), while open circles indicate non-significant differences (two-sample t-test, $p \geq 0.05$).

2.3 Network-aware signal propagation and metabolite-reaction signal inference

To incorporate biochemical network structure into the analysis, node-level metabolite signals were propagated across the bipartite metabolite–reaction network[13] using a graph diffusion framework[14]. Let $X \in \mathbb{R}^{n \times m}$ denote the node-level signal matrix, where n is the number of nodes and m the number of samples or experimental conditions. The adjacency matrix of the network, A, was row-normalized to obtain a transition matrix $W = D^{-1}A$, where D is the diagonal matrix of node out-degrees. Signal propagation was then performed using the closed-form diffusion operator $F = (I - \alpha W)^{-1}X$, where F is the propagated signal matrix, I is the identity matrix, and $\alpha \in [0,1])$ is the diffusion parameter that adjust the balance between local and global information, where lower values emphasize immediate neighbors and higher values allow broader diffusion across the network. Following propagation, a semi-supervised clamping procedure was applied to preserve initially observed values. Specifically, metabolite nodes with originally non-zero signal were reset to their original values, while nodes with zero initial signal retained their network-propagated estimates. Thus, reaction nodes acquire values that reflect topology-aware aggregations of propagated statistical signals from connected metabolites.

An example of a small bipartite network consisting of 8 metabolite nodes and 10 reaction nodes is shown in Fig. 3. The figure shows the impact of different α values on the trade-off between local and global information flow, and how reaction node values are estimated as topology-aware aggregates of signals from both directly connected and indirectly related metabolites. Initially, the network contains only two non-zero metabolite nodes (C1 and C7), with SC = 10, respectively (Fig. 3). Signal propagation was then performed using a diffusion parameter α = 0.25, 0.50 and 0.75. After propagation, a semi-supervised clamping step restored the original values of C1 and C7. As α increased, the signal progressively diffuses from the initially non-zero nodes to neighboring reaction nodes and subsequently to more distant metabolites through shared connectivity. At low α (0.25), propagation remained localized, with signal change largely limited to immediate neighbors, and nodes C3, C4 and C5 remained unaffected. At intermediate α (0.5), signal spread more broadly while still preserving local structure, enabling nearby reactions to accumulate moderate propagated signal. At higher α (0.75), diffusion extends across larger portions of the network, leading to increased signal in distal nodes. In this work, a moderate value α = 0.5 was selected to integrate network structure without excessive smoothing of the signal.

2.4 Discriminant analysis

Discriminant analysis was performed for each network node using Welch's two-sample t-test (unequal variances) and Partial Least Squared-Discriminant Analysis (PLS-DA) across three pairwise

comparisons: (i) oxidative stress vs no effect (OS vs NE2), (ii) mitochondrial dysfunction vs no effect (MD vs NE2), and (iii) a null comparison between two no-effect groups (NE1 vs NE2).

For a set of networks obtained from N independent case vs control studies testing these pairwise comparisons, the data matrix contained the node SC expressed as -log(p-values) (see Fig. 1). For each node, Welch's t-test was used to compare group means across the three pairwise comparisons, yielding a nominal p-value in each case. In parallel, the fold change (FC) between groups was calculated as the ratio of the mean node value in the treatment group to the mean node value in the reference group. A small constant (10^{-12}) was added to both means to avoid division-by-zero errors. To account for multiple testing across network nodes, p-values were adjusted using the Benjamini–Hochberg false discovery rate (FDR) procedure, producing q-values. Results were visualized using volcano plots, where the x-axis represented $\log_2$(FC) and the y-axis represented $-\log_{10}$(q). Nodes were considered significantly altered when they satisfied both an FDR-adjusted significance threshold ($q < 0.05$) and an effect-size threshold ($|\log_2 FC| > \log_2(1.5)$).

Selection and assessment of the statistical significance of PLS-DA models was carried out using random repeated stratified k-fold cross validation (k-fold CV) (5 repetitions, k = 5). At each repetition, fold assignments were randomly selected while keeping class balance within folds. Identical CV partitions were used for the initial and propagated models within each repetition, to ensure that differences in predictive performance between results obtained in the initial and propagated data were attributable to network propagation and not to variation in fold assignment. For each CV iteration, the optimal number of PLS latent variables (LVs) was selected by minimizing the CV classification error. CV predicted y values were then used to calculate the discriminant Q^2 statistic (dQ^2) [15], and a mean dQ^2 values was calculated after the 5 CV splits. A propagation-induced gain in predictive performance was defined as $\Delta \mathrm{dQ}^2 = \mathrm{dQ}^2_{\mathrm{propagated}} - \mathrm{dQ}^2_{\mathrm{initial}}$, where $\mathrm{dQ}^2_{\mathrm{propagated}}$ and $\mathrm{dQ}^2_{\mathrm{initial}}$ were estimated using the propagated and initial networks, respectively.

Permutation testing (250 iterations) was used to assess the significance of the observed predictive performance (dQ^2) and propagation-induced gain ($\Delta \mathrm{dQ}^2$). In each permutation, class labels were randomly shuffled while preserving the original network structure and feature relationships. The entire modeling pipeline, including fold generation, LV optimization, PLS-DA model selection, and CV, was repeated for each permuted dataset. This procedure generated empirical null distributions for dQ^2 and $\Delta\, dQ^2$under the null hypothesis of no association between data structure and class labels. Empirical one-sided p-values were estimated by comparing the observed statistics obtained using the true labels against their corresponding null distributions from permutation testing. Results obtained were used to estimate the significance of the PLS-DA model in the initial data and whether the increase in predictive performance observed after propagation was greater than expected from random class assignments alone.

Discriminant analysis using the initial network included only metabolite nodes with non-zero values, resulting in a reduced feature space (400 nodes). In the analysis of the data sets obtained after network

propagation, all 1601 non-zero metabolic nodes were included.

2.5 Software and data availability

Data simulation and analysis was carried out in MATLAB 2023b (MathWorks) using in-house written functions and the PLSToolbox 9.5.1 (Eigenvector Research Inc.) for PCA, PLS-DA, and data handling. MATLAB scripts for calculating the standardized within-class sum of squares[16] (SWISS) were obtained from http://cancer.med.unc.edu/nhayes/pubs.html. Network analysis and visualization was carried out in Gephi 0.10.1 (released 202301172018) running on Windows 11 (Microsoft). Over representation analysis (ORA)[17] was carried out in Metaboanalyst[18]. ORA was implemented using the hypergeometric test to evaluate whether a particular metabolite set was represented more than expected by chance within the given compound list using the KEGG library pathway as reference, and one-tailed p-values were provided after adjusting for multiple testing.

3. Results and discussion

3.1 Bipartite Reaction-Metabolite network

The complete filtered network contained 1601 metabolic nodes. However, only 400 metabolites initially contained non-zero simulated signal values. Figure 4 shows the resulting bipartite metabolite–reaction network comprised 8236 nodes and 11868 directed edges. The network exhibited a low average degree (1.44) and a relatively large diameter (43), consistent with a bipartite topology in which connectivity describes biochemical dependencies rather than metabolite co-occurrence. The network was fragmented into multiple weakly connected components, consistent with pathway-level organization and partial disconnection across metabolic subsystems. Community detection identified 109 modules with high modularity (0.831), reflecting strong topological partitioning. The inclusion of reaction nodes facilitates the exploration of network topology by showing multi-substrate dependencies. This representation might also enable downstream assignment of reaction-level scores linked to metabolomic changes.

3.2 PCA of initial and propagated data

PCA provided an overview of the global structure of the dataset (Fig. 5). The analysis was carried out before and after network propagation, considering four scenarios including: (a) only initial non-zero metabolites; and data after signal propagation including (b) all nodes (metabolites and reactions), (c) only metabolite nodes, and (d) only reaction nodes. To quantitatively evaluate the clustering of the differ classes in the PC score, the SWISS parameter was used and its statistical significance was estimated by permutation testing[16]. Before propagation (Fig. 5A), samples from the four different classes (OS, MD, NE1, and NE2) strongly overlap in the PC1 vs PC2 score space (SWISS = 0.6969), indicating that the variance associated with the simulated perturbations was small relative to the overall variability in the data. However, as NE1 and NE2 represent equivalent simulated conditions, the four-class SWISS metric likely overestimates within-group dispersion and underestimates apparent clustering performance.

Nevertheless, the observed overlap is consistent with the simulation design, in which condition-specific effects were introduced as small shifts in a limited subset of metabolites, and reflects a common situation in metabolomics, where class differences are distributed across a subset of correlated features and do not represent the main source of variance. Following network propagation (Figs. 5B-D), the overall distribution of samples in the PCA space remained largely overlapping, and no clear class separation emerged along the leading principal components. However, a slight improvement in sample clustering was observed (SWISS = 0.6530), indicating that propagation redistributed signal across the feature space without introducing dominant variance components. Similar results were observed from the analysis of propagated data including only reaction (SWISS = 0.6591, Figs. 5D) nodes and slightly better clustering was observed including only metabolic nodes (SWISS = 0.5524, Figs. 5C).

3.3 Supervised analysis of initial and propagated data

Univariate supervised analysis (Welch's test) was performed on the network-propagated data matrices for each node across three pairwise comparisons: (i) oxidative stress vs no effect (OS vs NE2), (ii) mitochondrial dysfunction vs no effect (MD vs NE2), and (iii) a null comparison between two no-effect groups (NE1 vs NE2). Propagation was first applied independently to each network using the SCs of each metabolite obtained from each 'case vs control' study as node signal (see Fig. 1). The propagation step provided SC values to metabolic nodes that originally were zero. Then, statistical comparisons were then performed for each node to assess, for each node, the difference in the distributions of propagated SC values between the two situations included (e.g., OS vs NE2). This analysis provided an estimated p-value for each node reflecting differences in the distributions of SCs between the two classes that were FDR corrected for further analysis.

The volcano plots shown in Fig. 6 depict results obtained. Network propagation assigned inferred SC values to metabolites that were not directly measured but were topologically connected to measured regions of the graph (gray circles in Fig. 6). Thus, some metabolites displayed relatively high statistical significance despite having no original signal, indicating that class-associated signal had diffused through the network topology. This behavior is consistent with the propagation framework, which reinforces coordinated topological patterns across connected network regions. It also suggests that the method can incorporate new network context to the interpretation. No metabolic or reaction nodes were found statistically significant in the null comparison between two no-effect groups (NE1 vs NE2).

Figure 7 depicts a section of the network showing the names of the 64 metabolites found significant (q-value < 0.05 and log2(FC) > 0.5) in the OS vs NE2 analysis shown in Fig. 6, including both initially non-zero (green circles) and propagation-inferred (orange circles) metabolic nodes. Figure 7 shows that the set of 64 significant metabolites captured network regions consistent with the simulated oxidative stress perturbation, including metabolites associated with glutathione turnover, redox-related metabolism, and purine-associated processes. The identified metabolites were not randomly distributed across the graph, but instead formed connected topological regions close to the initially perturbed metabolites. In addition to directly perturbed nodes, propagation-associated metabolites linked through

neighboring reactions were also identified. The statistical significance of the clustering of significant metabolites within the network was evaluated using a permutation test (1000 iterations) on the size of the largest connected component (LCC), including both metabolic and reaction nodes. The observed value (LCC = 660) was significantly larger than expected under random selection of 64 nodes ($p = 0.016$).

Figure 8 shows a section of the metabolic network showing some metabolite names among the set of nodes found significant (q-value < 0.05 and $\log_2(FC) > 0.5$) in the MD vs NE2 analysis (see Fig. 6). In this case, the initial perturbation included metabolites associated with central carbon metabolism (pyruvate, citrate, succinate, and malate). The list of 150 significant metabolites included associated with central carbon metabolism, amino acid metabolism, nucleotide metabolism, and cofactor-related processes, forming broad connected regions consistent with the central position of the initially perturbed metabolites in the network.

In this case, the observed largest connected component value (LLC = 1346) was also significantly larger than expected under random node selection ($p = 0.004$), again supporting that the subset of 158 metabolites followed a non-random organization of nodes within the network.

Graph topology influences the distribution of the signal during network propagation. Highly connected or topologically central metabolites tend to accumulate propagated signal due to their multiple connections within the network. In this simulation, the metabolites defining the OS and MD cases showed different topological properties. MD metabolites showed median betweenness and closeness percentiles of 99.9 (range 20–100) and 83.5 (range 64–85), respectively, indicating that these nodes occupy highly central and well-connected positions in the network. In contrast, selected OS metabolites also displayed high betweenness centrality (median 96.1, range 79.4–99.9), but more moderate closeness (median 64.6, range 37.4–84.2), indicating a less central distribution. Several MD metabolites also corresponded to high-degree hubs, including 2-oxoglutarate (degree: 308), Acetyl-CoA (degree: 246), Pyruvate (degree: 211), and Succinate (degree: 200)) (see Fig. 8). OS metabolites showed lower connectivity, with most metabolites having degree < 10 and only Glutamate) with degree > 200. These topological differences were consistent with the broader propagation observed in the MD simulation compared with the more localized propagation pattern observed in OS.

To evaluate the influence of the diffusion parameter α on the propagation process, the number of statistically significant metabolites identified after propagation (q-value < 0.05 and log2FC > 1.5) was assessed across a range of α values for the OS vs NE2 and MD vs NE2 comparisons (Fig. 9). In the OS simulation, increasing α led to a minor but progressive decrease in both the number of significant metabolites and the size of the LCC. In contrast, in the MD simulation, increasing α resulted in a larger number of significant metabolites and LCC. These different behaviors suggest that the effect of the diffusion parameter depends on the topological organization of the initially perturbed metabolites, and evaluation across multiple α values may therefore be necessary.

Pathway analysis using the network-propagated analysis

To evaluate the impact of the simulated perturbations at the pathway level, over-representation analysis (ORA) was carried out using KEGG metabolic pathways. ORA was used to identify metabolic patterns that were significantly enriched within the KEGG pathway library as reference metabolomic dataset. Enrichment-based approaches assess whether functionally related metabolites are overrepresented among discriminant metabolites, enabling pathway-level interpretation of coordinated metabolic responses. Consequently, only selected discriminant metabolites contribute to pathway enrichment.

First, the sets of metabolites initially selected to simulate OS and MD before network propagation were analyzed independently. For the OS simulation, ORA identified pathways associated with glutathione metabolism (Hits 5/28, FDR = 0.0003), and purine metabolism (Hits 5/70, FDR = 0.003) as the most enriched metabolite sets. In the MD simulation, ORA highlighted pathways related to the citrate cycle (TCA cycle) (Hits 7/20, FDR = $6\ 10^{-11}$), pyruvate metabolism (Hits 6/22, FDR = 10^{-8}), alanine, aspartate and glutamate metabolism (Hits 5/28, FDR = $5\ 10^{-6}$), glyoxylate and dicarboxylate metabolism (Hits 4/32, FDR = $3\ 10^{-4}$) as the most enriched metabolite sets. These results indicated that the manually selected metabolite groups were consistent with target mechanisms.

ORA was then applied using the sets of metabolites found statistically significant after network propagation in the OS vs NE2 and MD vs NE2 comparisons. In the OS case, ORA identified purine metabolism (Hits 12/70, FDR = $6\ 10^{-6}$), glutathione metabolism (Hits 8/28, FDR = $9\ 10^{-6}$), and arginine biosynthesis (Hits 4/14, FDR = 0.009) as the most enriched pathways. These results were consistent with the topological distribution of significant metabolites observed in Fig. 7, where metabolites associated with glutathione turnover and purine catabolism formed connected regions surrounding the initially perturbed nodes. Besides, pathway enrichment also included propagation-inferred metabolites linked through neighboring reactions, indicating that signal diffusion extended the altered metabolic context beyond the initially measured metabolites.

ORA of the set of 150 significant metabolites identified in the MD vs NE2 comparison showed the enrichment of 12 pathways including alanine, aspartate and glutamate metabolism (Hits 11/28, FDR = $4\ 10^{-7}$), citrate cycle /TCA cycle) (Hits 8/20, FDR = $3\ 10^{-5}$), butanoate metabolism (Hits 7/15, FDR = $3\ 10^{-5}$), arginine biosynthesis (Hits 7/15, FDR = $3\ 10^{-5}$), pyruvate metabolism (Hits 7/22, FDR = 0.003), phenylalanine metabolism (Hits 4/8, FDR = 0.002), and tyrosine metabolism (Hits 7/22, FDR = 0.003). Compared with the OS simulation, the MD perturbation produced a broader and more diffuse enrichment profile, consistent with the higher centrality and connectivity of the metabolic nodes initially perturbed in this condition.

Taken together, the ORA results and the spatial organization of significant metabolites in the propagated networks suggest that as expected, network propagation does not generate isolated significant nodes distributed across the graph, but instead amplifies coordinated signal within connected biochemical regions. Nevertheless, these pathway-level enrichments after signal propagation should be interpreted as exploratory and hypothesis-generating rather than as direct evidence of pathway regulation.

Multivariate predictive models

Predictive modeling was performed using PLS-DA applied to both the non-propagated and propagated data networks, across the same three pairwise comparisons described above: (i) oxidative stress vs no effect (OS vs NE2), (ii) mitochondrial dysfunction vs no effect (MD vs NE2), and (iii) a null comparison between two no-effect groups (NE1 vs NE2).

In the OS vs NE2 comparison, the initial data set including 400 metabolic nodes showed statistically significant predictive performance ($dQ^2 = 0.280$, permutation test p-value = 0.004), indicating the presence of discriminant biological signal in the data. After network propagation, performance increased significantly ($dQ^2 = 0.515$, permutation test p-value = 0.004). The observed propagation gain was statistically significant ($\Delta dQ^2 = 0.235$, permutation test p-value = 0.016), indicating that propagation improved predictive performance beyond what could be expected under random class assignment.

In the MD vs NE2 comparison, the initial dataset showed limited predictive performance ($dQ^2 = 0.042$, $p = 0.13$). However, after propagation, model performance increased ($dQ^2 = 0.553$, $p = 0.004$) leading to a significant gain ($\Delta dQ^2 = 0.511$, $p = 0.012$). The observed gain was statistically significant, indicating again that the improvement could not be attributed to propagation bias alone.

In the null comparison (NE1 vs NE2), both the initial and propagated models showed no predictive performance (initial: $dQ^2 = -0.229$, $p = 0.9$; propagated: $dQ^2 = -0.08$, $p = 0.3$), consistent with the absence of class-related signal. Despite the lack of predictive signal, a positive propagation gain was observed ($\Delta dQ^2 = 0.152$). Under a label permutation null model, this gain was not statistically significant ($p = 0.06$), indicating that the apparent improvement was compatible with propagation-induced effects and model variability rather than class separability. However, this borderline p-value highlights that propagation acts as a topological smoothing filter that distributes background noise across shared network pathways, and this can artificially inflate a multivariate structure even in the complete absence of a biological effect. This result underscores the necessity of using permutation testing targeting the performance gain rather than evaluating the absolute post-propagation ΔdQ^2, to avoid false-positive results.

Taken together, these results show that, in the tested simulation, the effect of network propagation on predictive performance depended on both the presence of discriminant metabolic nodes and on their topological distribution. In the presence of between-classes differences (OS vs NE2 and MD vs NE2), it increased model performance. In the absence of signal (NE1 vs NE2), propagation enhanced the predictive performance, but this increase was not statistically significant. Although propagation increased the number of non-zero variables included in the models, the use of permutation testing for ΔdQ^2 allowed the assessment of whether the increase in performance was larger than that expected from the effect of increasing the number of features and the propagation.

Some cautions should be considered for the development of network-informed propagated models. The proposed strategy has been evaluated exclusively using simulated data under controlled conditions, which do not fully capture the complexity, noise structure, and heterogeneity of real metabolomic

datasets. In particular, the simulation design assumes a shared missingness structure and controlled signal perturbations, which may favor the effectiveness of network propagation compared to more heterogeneous scenarios. The magnitude of improvement gained after propagation depends on several factors such as the network topology and metabolic coverage of the network and may vary across datasets. Also, network propagation is sensitive to the selected diffusion coefficient (α), which determines the balance between local and global signal integration. Finally, the propagated values represent topology-informed statistical signals and not experimental measurements and should therefore be interpreted for hypothesis generation.

Additionally, network propagation methods are inherently influenced by network topology, and nodes with higher degrees or central positions may accumulate more signal during propagation, irrespective of their initial values. As a result, propagated scores may partially reflect topological properties such as connectivity or centrality. However, the proposed strategy minimizes the risk of false-positive topological inflation. Instead of propagating signal values from a singular profile, the input matrix consists of independent SCs ($-\log_{10}$ p-values) derived from independent experiments. Consequently, for a metabolic hub and its immediate topological neighbors to exhibit elevated post-propagation scores, the hub must show a reproducible SC across the experimental conditions. Finally, to confidently distinguish significant biological signals from topological smoothing effects, the routine implementation of label permutation tests is recommended when applying this approach. Ultimately, experimental validation will be required to assess the accuracy and biological relevance of results obtained. As the bipartite framework assigns significance scores to reactions, these hypotheses using multi-omics approaches such as proteomics or transcriptomics to monitor corresponding enzyme expression or activity could support validation.

4. Conclusions

We present a network-based strategy for metabolomics analysis that integrates metabolite-level SCs with network topology through propagation. Under controlled simulation conditions with known ground truth, network propagation extended statistical signal to connected network regions, and improved discriminant performance while maintaining specificity in null comparisons. These results provide a proof-of-concept that network propagation can support the exploration of partially observed metabolic systems by integrating local experimental data with a reference network structure. While the inferred SCs should be interpreted as hypothesis-generating and require independent experimental validation, this approach can be valuable in metabolomics, where incomplete coverage and annotation limit direct biological interpretation, and exploratory strategies are important in guiding downstream validation. Future work will focus on applying this approach to experimental datasets and assessing its accuracy and robustness.

Declarations

Competing interests

The authors declare no competing interests.

Author Contribution

JK: resources, analyzed data; MMT: resources, analyzed data; DPG: analyzed data; FAET: analyzed data; GQ: conceptualization, methodology, analyzed data. All authors wrote and reviewed the manuscript.

Acknowledgement

This work was supported by the Instituto de Salud Carlos III (ISC III), Spain, and co-funded by the European Union [grant number PI23/00202]. MMT acknowledges the grant RYC2021–031346-I, funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, the project "PI24/00215" from Instituto de Salud Carlos III (ISCIII) and co-funded by the European Union, and the Grant CNS2024–154299 funded by MICIU/AEI /10.13039/501100011033.

Data Availability

The data, the Gephi Network file, and the simulation results that support the findings of this study are available in Zenodo (10.5281/zenodo.20808246).

References

1. Ivanovs, M., Kadikis, R., & Ozols, K. (2021). Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, *150*, 228–234. https://doi.org/10.1016/j.patrec.2021.06.030
2. Sidak, D., Schwarzerová, J., Weckwerth, W., & Waldherr, S. (2022). Interpretable machine learning methods for predictions in systems biology from omics data. *Front Mol Biosci*, *9*, 926623. https://doi.org/10.3389/fmolb.2022.926623
3. Kuligowski, J., Moreno-Torres, M., & Quintás, G. (2024). Improving insights from metabolomic functional analysis combining multivariate tools. *Analytica Chimica Acta*, *1323*, 343062. https://doi.org/10.1016/j.aca.2024.343062
4. Green, M. L., & Karp, P. D. (2006). The outcomes of pathway database computations depend on pathway ontology. *Nucleic Acids Research*, *34*, 3687–3697. https://doi.org/10.1093/nar/gkl438
5. Radrich, K., Tsuruoka, Y., Dobson, P., Gevorgyan, A., Swainston, N., Baart, G., & Schwartz, J. M. (2010). Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *Bmc Systems Biology*, *4*, 114. https://doi.org/10.1186/1752-0509-4-114

6. Lee, K. S., Su, X., & Huan, T. (2025). Metabolites are not genes — avoiding the misuse of pathway analysis in metabolomics. *Nat Metab*, 1–4. https://doi.org/10.1038/s42255-025-01283-0
7. Kuligowski, J., Albiach-Delgado, A., Pérez-Guaita, D., & Quintás, G. (2025). Interpretation of multivariate metabolomic models through network-guided perturbation-based explanations. *Metabolomics*, *21*, 143. https://doi.org/10.1007/s11306-025-02347-8
8. Cowen, L., Ideker, T., Raphael, B. J., & Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, *18*, 551–562. https://doi.org/10.1038/nrg.2017.38
9. Cowen, L., Ideker, T., Raphael, B. J., & Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, *18*, 551–562. https://doi.org/10.1038/nrg.2017.38
10. Visonà, G., Bouzigon, E., Demenais, F., & Schweikert, G. (2024). Network propagation for GWAS analysis: a practical guide to leveraging molecular networks for disease gene discovery. *Briefings In Bioinformatics*, *25*, bbae014. https://doi.org/10.1093/bib/bbae014
11. Pirhaji, L., Milani, P., Leidl, M., Curran, T., Avila-Pacheco, J., Clish, C. B., White, F. M., Saghatelian, A., & Fraenkel, E. (2016). Revealing disease-associated pathways by network integration of untargeted metabolomics. *Nature Methods*, *13*, 770–776. https://doi.org/10.1038/nmeth.3940
12. Hui, T. X., Kasim, S., Md. Fudzee, M. F., Sutikno, T., Hassan, R., Aziz, I. A., Hasan, M. H., Jaafar, J., Alharbi, M., & Sen, S. C. (2023). A Review of Random Walk-Based Method for the Identification of Disease Genes and Disease Modules. *Ieee Access : Practical Innovations, Open Solutions*, *11*, 116366–116383. https://doi.org/10.1109/ACCESS.2023.3324985
13. Frainay, C., & Jourdan, F. (2017). Computational methods to identify metabolic sub-networks based on metabolomic profiles. *Briefings In Bioinformatics*, *18*, 43–56. https://doi.org/10.1093/bib/bbv115
14. Dengyong Zhou, O., Bousquet, T. N., Lal, J., & Weston Bernhard Sch¨olkopf, Learning with Local and Global Consistency, Adv. Neural Inf. Process. Syst. 16 (n.d.) 321–328.
15. Westerhuis, J. A., van Velzen, E. J. J., Hoefsloot, H. C. J., & Smilde, A. K. (2008). Discriminant Q2 (DQ2) for improved discrimination in PLSDA models. *Metabolomics*, *4*, 293–296. https://doi.org/10.1007/s11306-008-0126-2
16. Cabanski, C. R., Qi, Y., Yin, X., Bair, E., Hayward, M. C., Fan, C., Li, J., Wilkerson, M. D., Marron, J. S., Perou, C. M., & Hayes, D. N. (2010). Standardized WithIn Class Sum of Squares to Evaluate Methodologies and Dataset Elements. *PLOS ONE*, *5*, e9905. https://doi.org/10.1371/journal.pone.0009905
17. Wieder, C., Frainay, C., Poupin, N., Rodríguez-Mier, P., Vinson, F., Cooke, J., Lai, R. P., Bundy, J. G., Jourdan, F., & Ebbels, T. (2021). Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. *Plos Computational Biology*, *17*, e1009105. https://doi.org/10.1371/journal.pcbi.1009105
18. Pang, Z., Lu, Y., Zhou, G., Hui, F., Xu, L., Viau, C., Spigelman, A. F., MacDonald, P. E., Wishart, D. S., Li, S., & Xia, J. (2024). MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. *Nucleic Acids Research*, gkae253. https://doi.org/10.1093/nar/gkae253

Figures

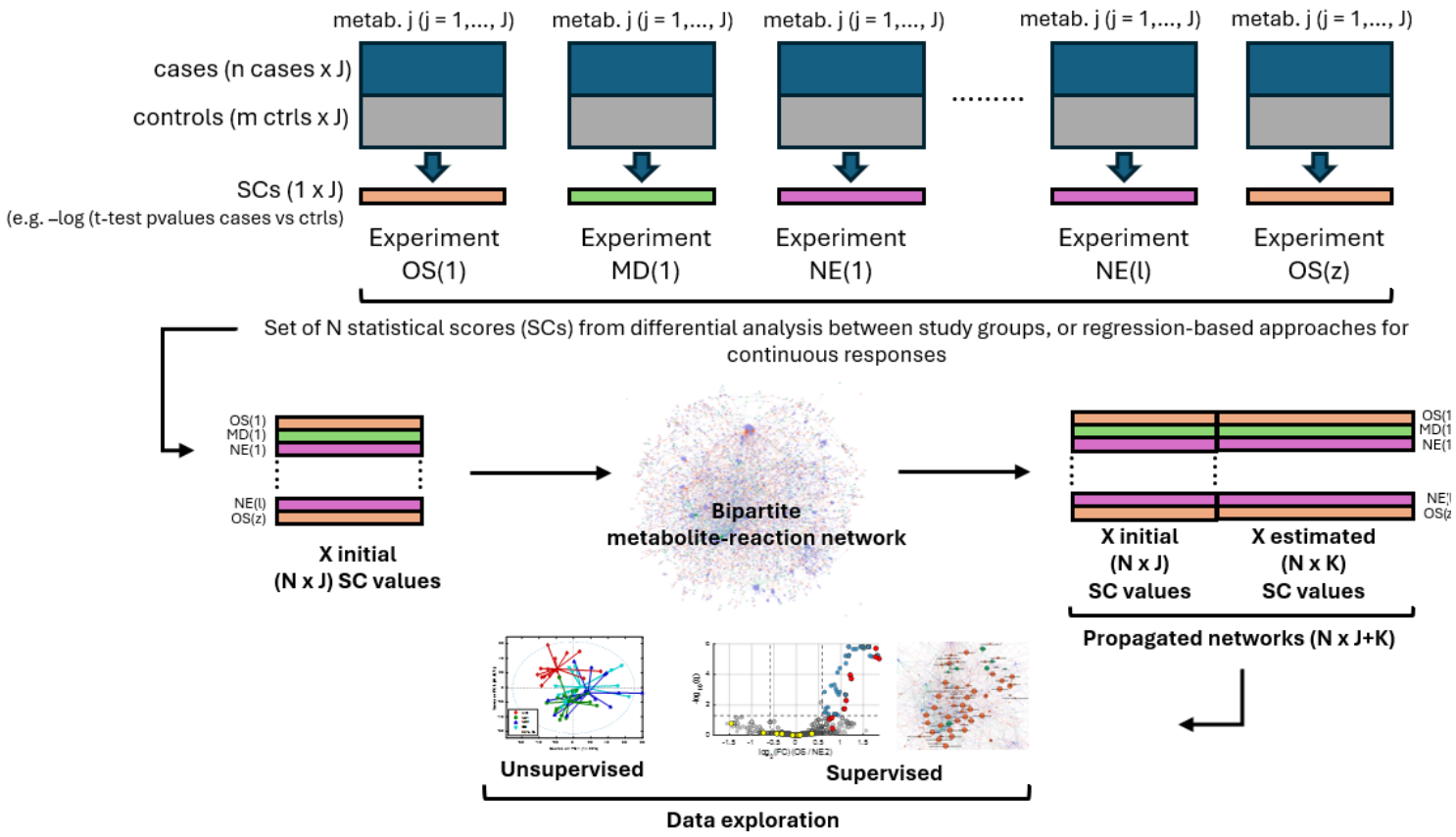


Figure 1

Scheme of data analysis pipeline. Results from a set of N independent studies comparing concentrations or levels of J metabolites between two study groups, or regression-based approaches for continuous responses, are collected. For each study, a (1 x J) *statistical scores* (SCs) vector, representing the statistical significance of each metabolite in the study (e.g., t-tests -log(p-value)) are obtained. Then, the set of SC vectors is rearranged into a matrix $\mathbf{X}_{\mathbf{initial}}$ (N x J). The J metabolites are identified in the bipartite metabolite-reaction network, and network propagation is carried out using the SCs obtained from each independent study included in $\mathbf{X}_{\mathbf{initial}}$ (N x J). The propagated networks, $\mathbf{X}_{\mathbf{propagated}}$ (N x J+K), provide an extended data matrix containing J original and K non-zero propagated values which is analyzed for data exploration.

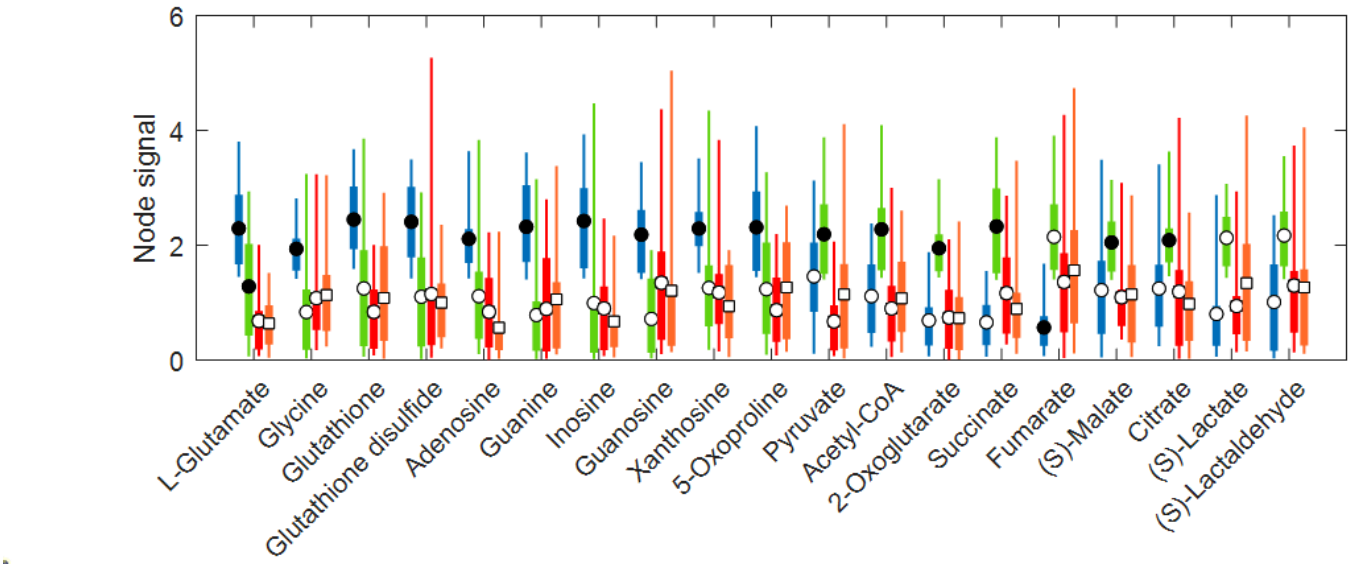


Figure 2

Distribution of SCs for boosted metabolites across conditions.
SCs obtained from 15 independent 'case vs control' studies per condition are shown for oxidative stress (OS, blue), mitochondrial dysfunction (MD, green), and 'no effect' (NE1 (red) and NE2 (orange)). For each condition, vertical lines represent the 5th–95th percentile range, and thick bars indicate the interquartile range (25th–75th percentiles). Circles denote the mean SC value. Filled black circles indicate a statistically significant difference relative to NE2 (two-sample t-test, $p < 0.05$), while open circles indicate non-significant differences (two-sample t-test, $p \geq 0.05$).

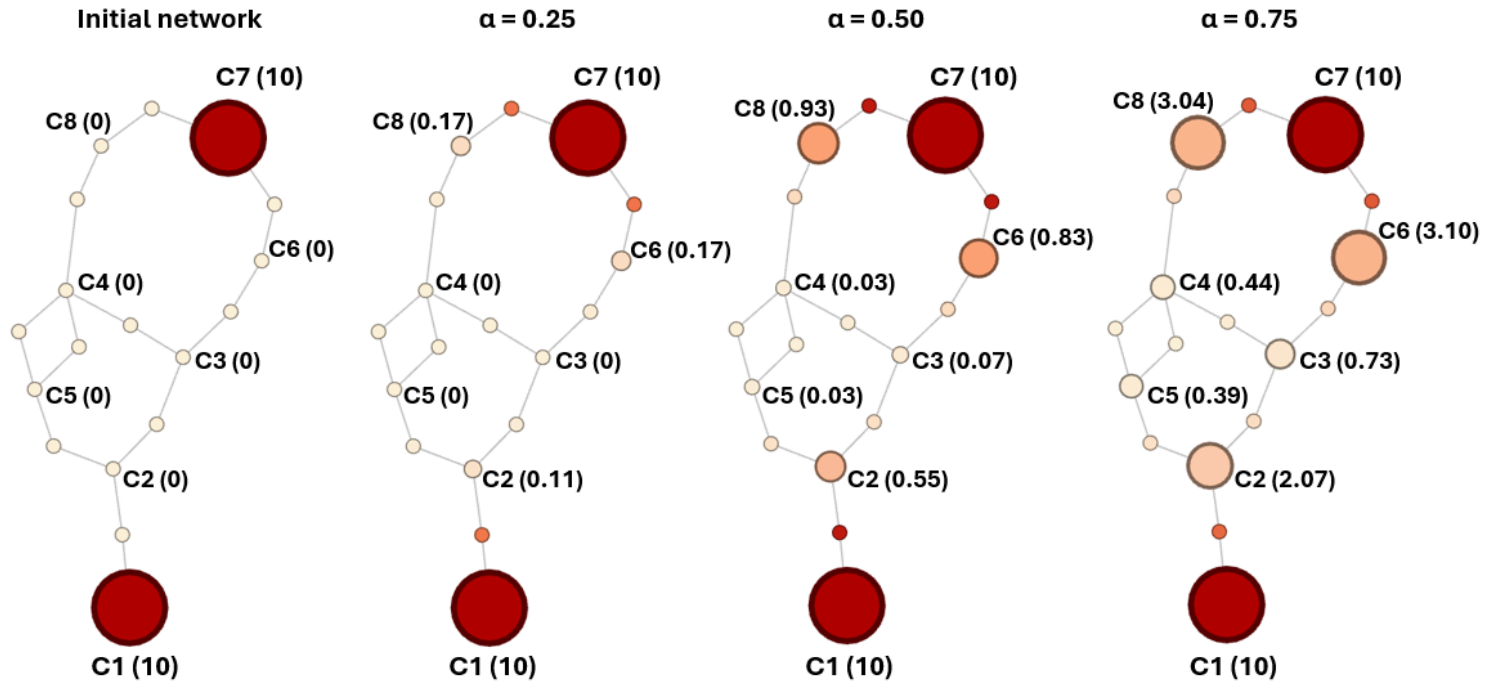


Figure 3

Effect of the parameter α on the network propagation of a small bipartite metabolite–reaction network. Circles connecting metabolites (C labeled nodes) represent reactions. The color and size of metabolic nodes represent the node SC. The size of reaction nodes was kept constant for better visualization.

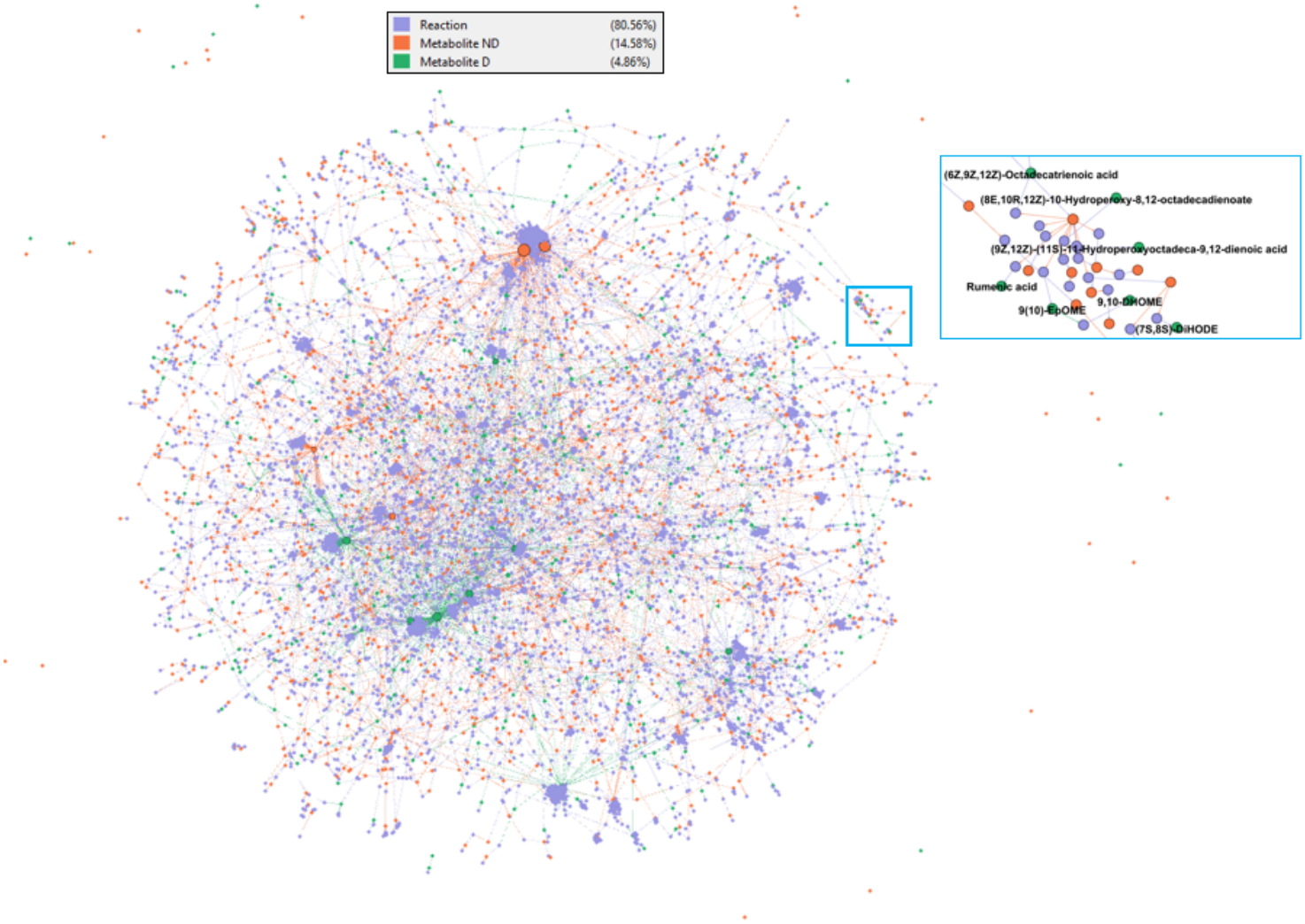


Figure 4

Structure of the metabolic network. Nodes represent metabolites detected (D, green) or not detected in the simulated experimental dataset (ND, orange), and reactions (purple). Edges represent directed associations through a chemical reaction. Node sizes are proportional to the node degree.

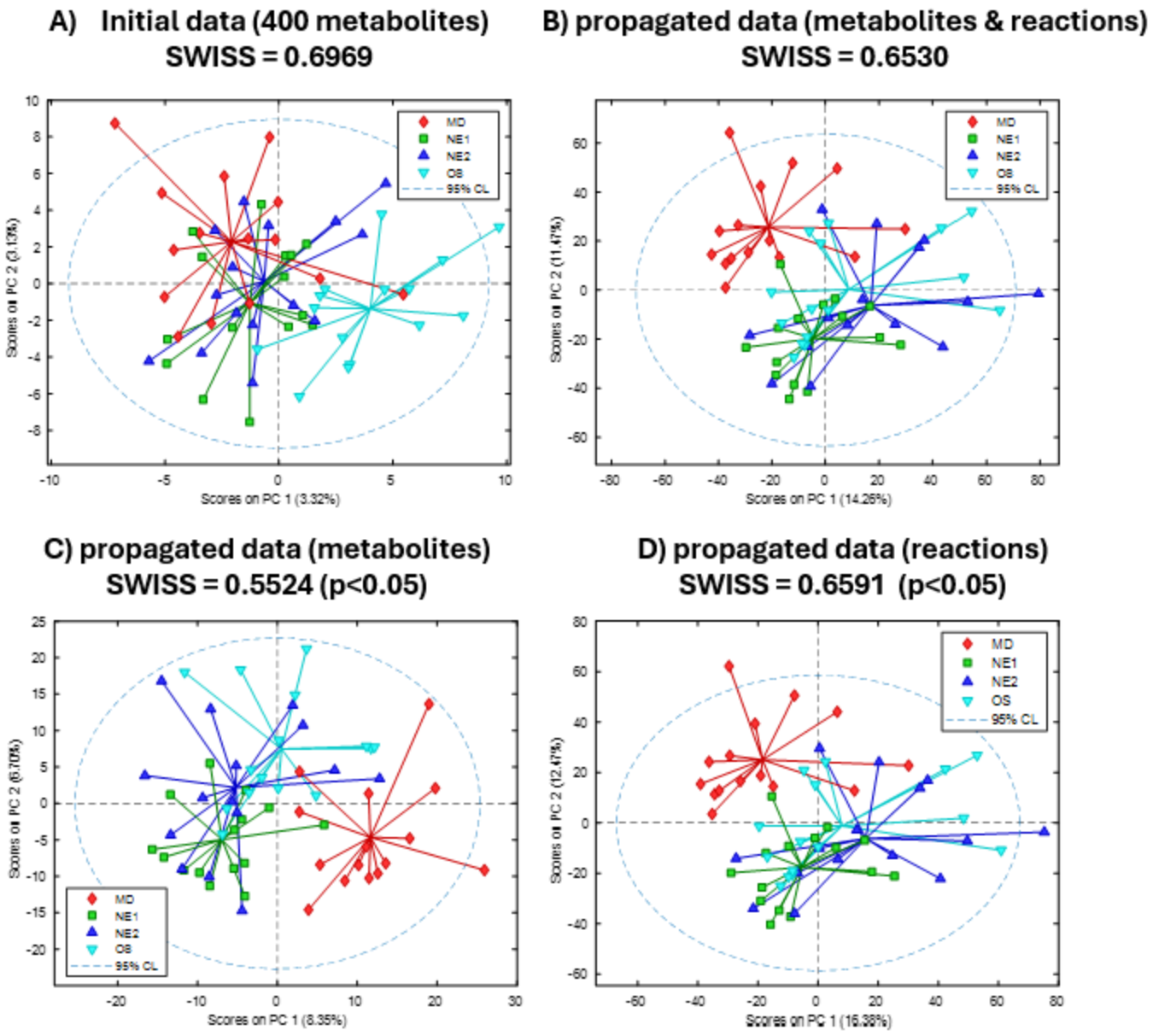


Figure 5

PCA scores plots of simulated network profiles. A) original data (400 metabolite nodes); B) propagated data including metabolite (1182) and reaction (4259) nodes; C) propagated data including only metabolite (1182) nodes; D) propagated data including only reaction (4259) nodes.

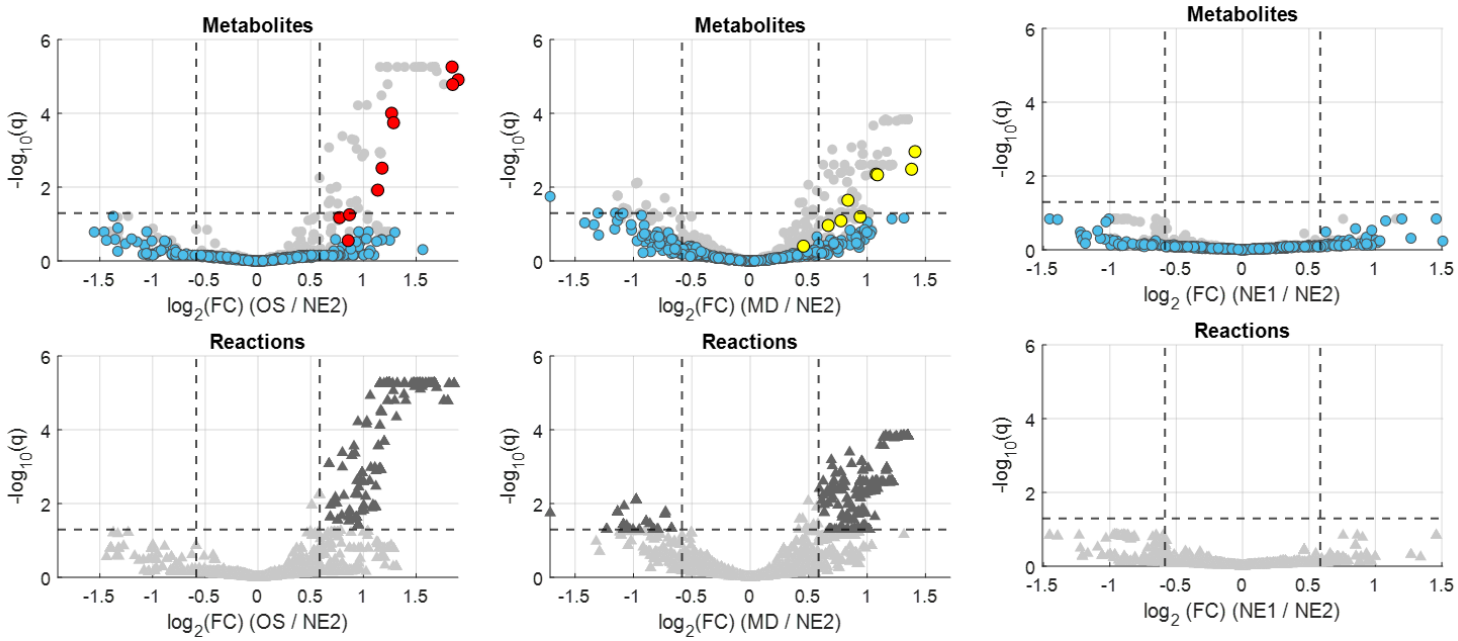


Figure 6

Volcano plots showing the log2 fold change (ratio of mean values) as a function of statistical significance ($-\log_{10}$ of False Discovery Rate-adjusted q-values) for each node, based on propagated networks, in the comparisons OS vs NE2 (left), MD vs NE2 (center), and NE1 vs NE2 (right)(top: metabolites, bottom: reactions). Statistical significance was assessed using Welch's t-test. Metabolites are represented as circles (gray: not initially detected; blue: initially detected; red and yellow: initially detected and included in the sets of metabolites simulating perturbations associated with OS (red) or MD (yellow). Dashed lines indicate the thresholds for statistical significance ($q < 0.05$) and effect size ($|\log_2$ fold change$|$) above the specified cutoff.

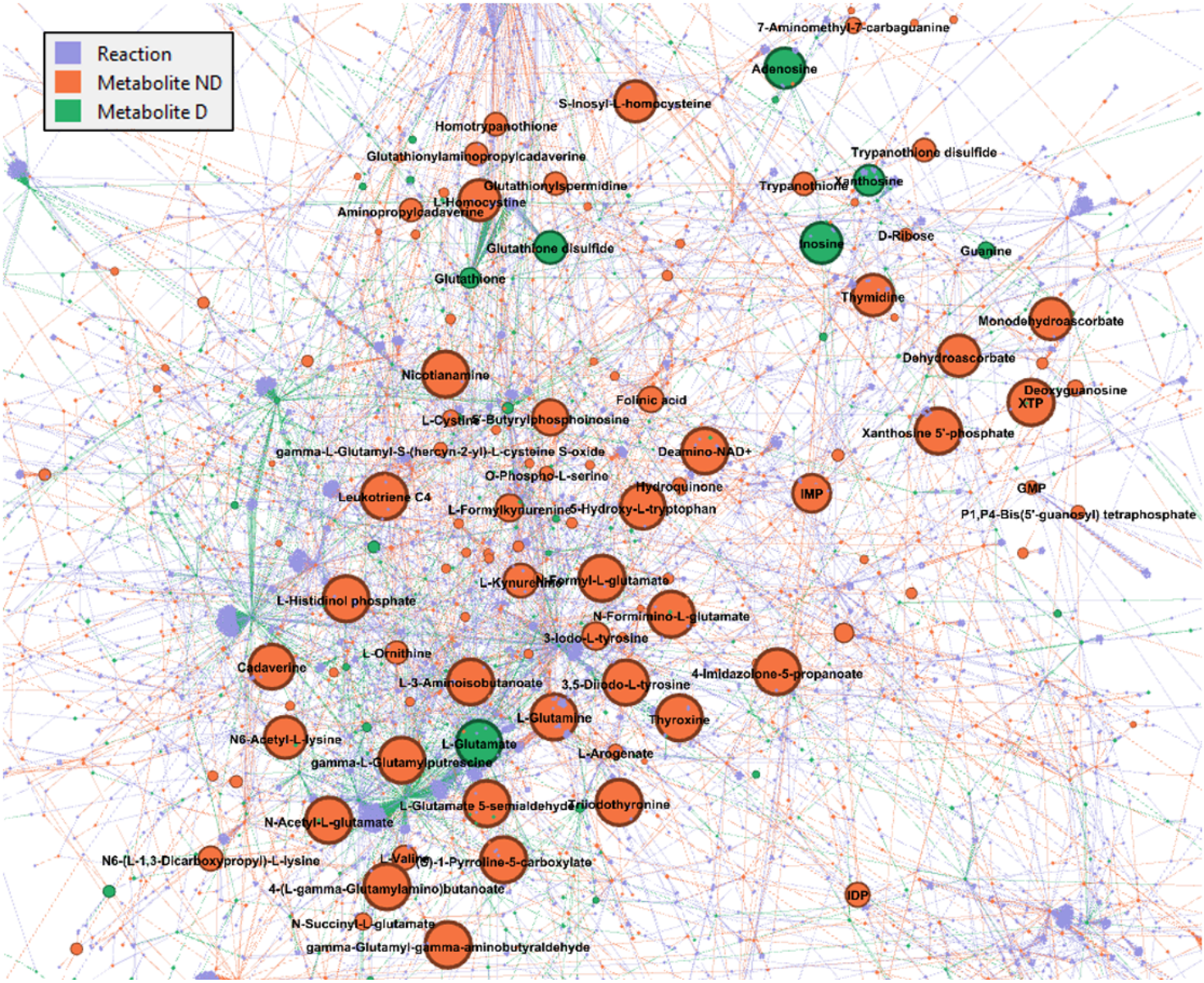


Figure 7

Section of the network showing the names of the metabolites found significant (q-value<0.05 and $\log_2$(FC)>0.5) in the OS vs NE2 analysis. Green nodes: initial non-zero metabolites, orange nodes: propagation-inferred metabolites, purple nodes: reactions. Metabolic node size is proportional to the $-\log_{10}$ of False Discovery Rate-adjusted q-values.

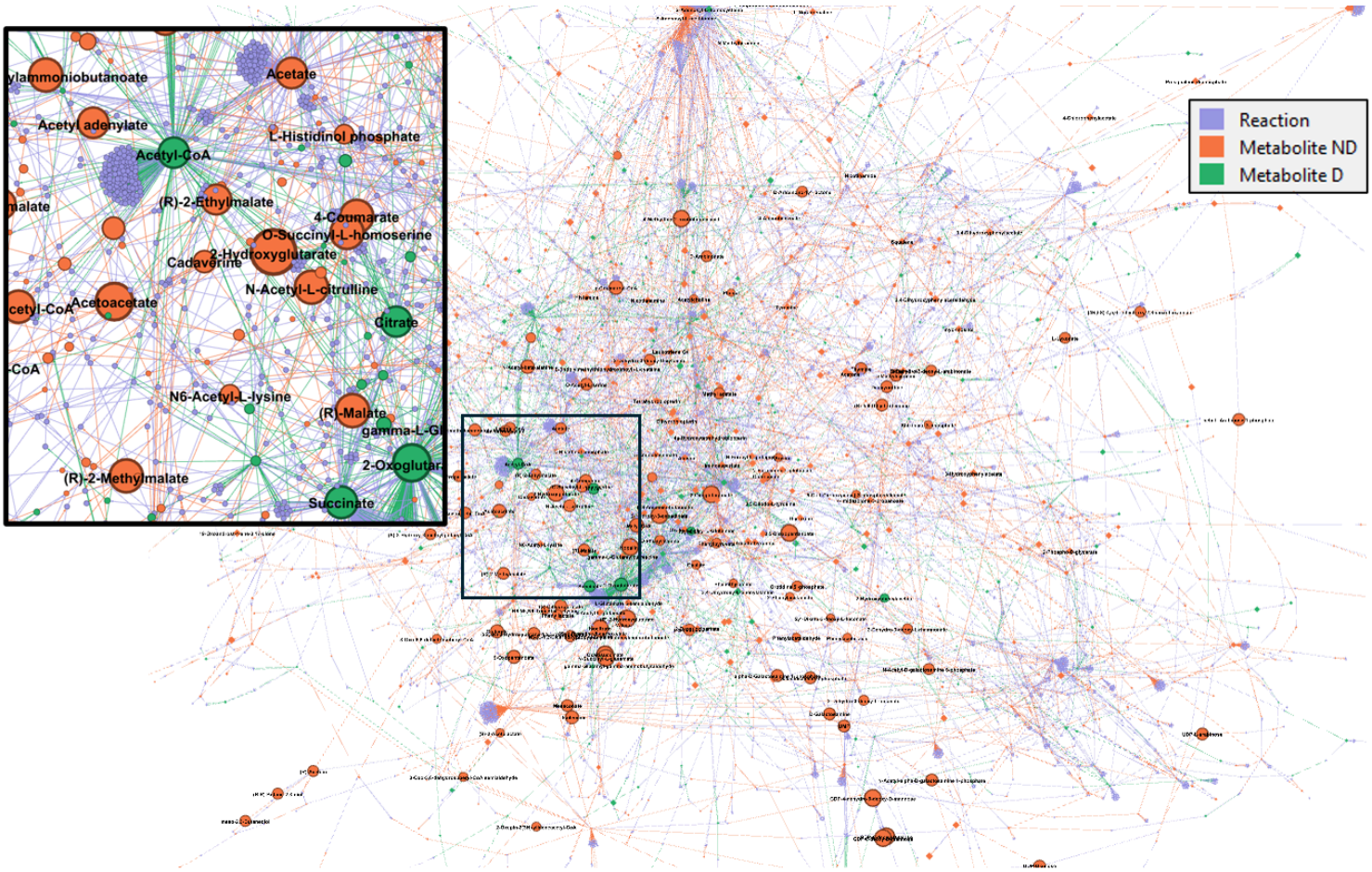


Figure 8

Section of the network showing the names of the metabolites found significant (q-value<0.05 and $\log_2$(FC)>0.5) in the MD vs NE2 analysis. Green nodes: initial non-zero metabolites, orange nodes: propagation-inferred metabolites, purple nodes: reactions. Metabolic node size is proportional to the $-\log_{10}$ of False Discovery Rate-adjusted q-values.

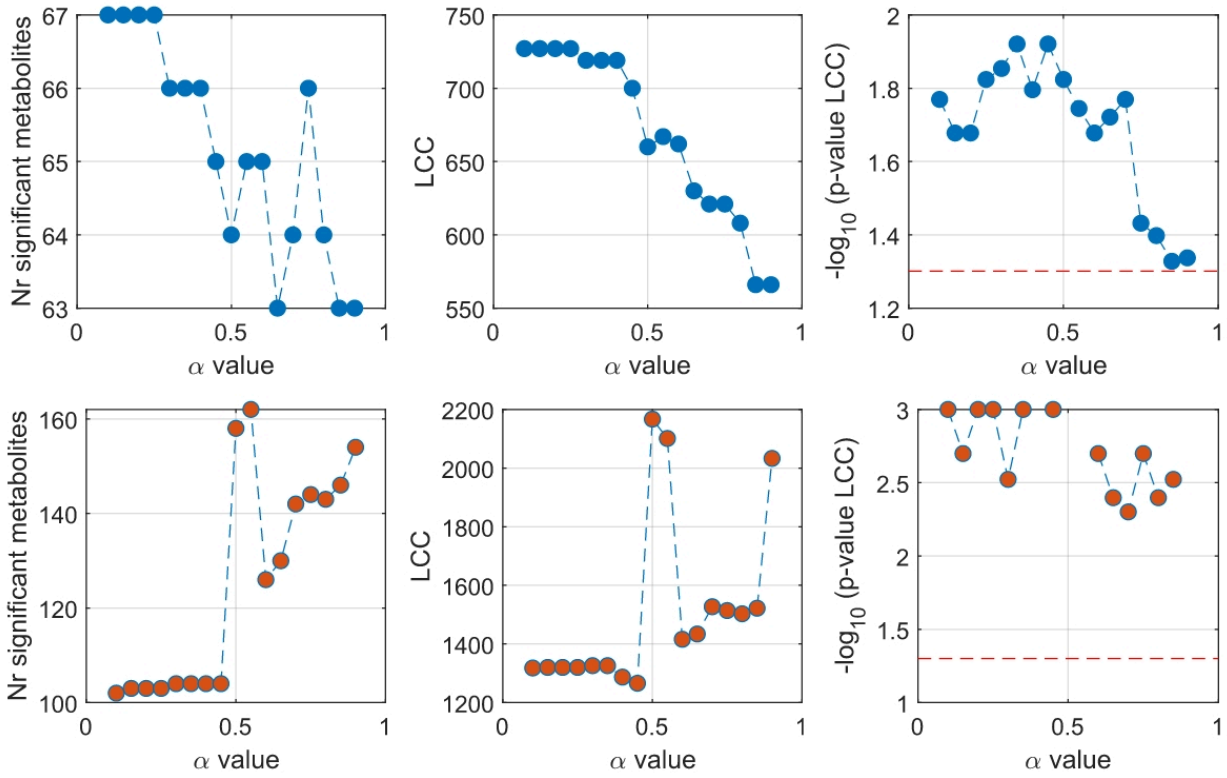


Figure 9

Effect of the diffusion parameter α on the number of statistically significant metabolites identified, LCC, and LCC permutation test p-value, after network propagation in the OS vs NE2 (top) and MD vs NE2 (bottom) comparisons. Significant metabolites were defined using q-value < 0.05 and $\log_2$FC > 1.5 thresholds.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterialTableSM1.xlsx