# Entropy-Based Network Representation of the Individual Metabolic Phenotype

Saccenti, E., Menichetti, G., Ghini, V., Remondini, D., Tenori, L., & Luchinat, C.

# An entropy based network representation of the individual metabolic phenotype

Edoardo Saccenti,[*,†] Giulia Menichetti,[‡] Veronica Ghini,[¶] Daniel Remondini,[‡] Leonardo Tenori,[§,⊥] and Claudio Luchinat[‖,⊥]

*Laboratory of Systems and Synthetic Biology, Wageningen University, Stippeneng 4, 6708 WE, Wageningen, The Netherlands, Department of Physics and Astronomy, University of Bologna, viale Berti Pichat 6/2, 40127, Bologna Italy, Magnetic Resonance Center, University of Florence, via Luigi Sacconi 6, 59100, Sesto Fiorentino, Italy, Department od Experimental and Clinical Medicine, University of Florence, Largo Brambilla 3, 50134, Florence, Italy, and Department of Chemistry, University of Florence, via della Lastruccia 3, 50019, Sesto Fiorentino, Italy*

E-mail: esaccenti@gmail.com

Phone: +31 (0)317 482018

**Abstract**

We approach here the problem of defining and estimating the nature of the metabolite-metabolite association network underlying the human individual metabolic phenotype

[*] To whom correspondence should be addressed

[†] Laboratory of Systems and Synthetic Biology, Wageningen University, Stippeneng 4, 6708 WE, Wageningen, The Netherlands

[‡] Department of Physics and Astronomy, University of Bologna, viale Berti Pichat 6/2, 40127, Bologna Italy

[¶] Magnetic Resonance Center, University of Florence, via Luigi Sacconi 6, 59100, Sesto Fiorentino, Italy

[§] Department od Experimental and Clinical Medicine, University of Florence, Largo Brambilla 3, 50134, Florence, Italy

[‖] Department of Chemistry, University of Florence, via della Lastruccia 3, 50019, Sesto Fiorentino, Italy

[⊥] Magnetic Resonance Center, University of Florence, via Luigi Sacconi 6, 59100, Sesto Fiorentino, Italy

in healthy subjects. We retrieved significant associations using an entropy based approach and a multiplex network formalism. We defined a significantly over-represented network formed by biologically interpretable metabolite modules. The entropy of the individual metabolic phenotype is also introduced and discussed.

Published in *Journal of Proteome Research* 2016 Sep 2;
15(9):3298-307. doi: 10.1021/acs.jproteome.6b00454. Epub 2016 Aug 1.

## Keywords

Metabolite-metabolite association networks, Metabolite mod- ules, Network multiplex.

## Introduction

Humans exhibit great phenotypic diversity. There are multiple regulatory layers underlying the functioning of a living system which organize the response to perturbations and/or modifications.[1,2] These regulatory processes cause biological components, such as metabolites, to change in a coordinated way with respect to these perturbations: the different patterns of metabolite (co)variation and association give birth to the individual metabolic phenotype.

This metabolic phenotype, as defined by NMR spectroscopy of urinary profiles and multivariate modeling, is unique for healthy subjects allowing discrimination with $\approx 100\%$ accuracy.[3] Moreover, it is stable over time[4] and possesses both allostasis and resilience.[5] By quantifying a finite subset of $M$ metabolites unambiguously identified for all subjects, the complexity and the high-dimensionality of NMR profiles arising from the signals of hundreds to thousands of low molecular weight molecules can be reduced of a $n$-fold factor. This reduced $M$-dimensional representation still allows predictive discrimination with high accuracy (see Figure 1A and B), indicating that a subject-specific biological information is thoroughly represented by a limited number of metabolites which constitute the biological components of the human metabolic phenotype. The metabolites excreted in urine represent the output of countless biochemical and biological processes, spanning from kidney functionality to the intricate interplay between the gut microbiome and the human host. For this reason it is of interest to investigate the associations and the interconnections among different metabolites by means of network modeling.

Networks built upon metabolic profiling exhibit a large degree of diversity as shown in Figure 1C and D and this variability reflects the intrinsic diversity observed among the individual metabolic phenotypes. It has recently been shown that such diversity is mainly due to intrinsic factors, such as genetics and epigenetics, being extrinsic factors, such as dietary habits or lifestyle, less important.[6]

Because the biological machinery shaping the urinary metabolic profile of different healthy subjects can be assumed to be the same, it is reasonable to assume the existence of what we termed *population core network* (PCN), that is, a metabolite-metabolite association network representing and underlying the metabolic phenotype observed in the overall population.

The ultimate goal of the present study is to estimate and infer characteristics of such PCN whose direct observation is precluded by the biologically noisy landscape of the sampled metabolic phenotype, which in turn gives origin to slightly different yet subject-specific metabolic networks. Defining a consensus of biological dependencies between metabolites across (healthy) individuals is fundamental since metabolites are read-outs from complex interaction networks and their analysis in a network context can reveal the underlying structure and regulation.[7] On this basis, the resulting metabolite-metabolite association networks can be compared across different subjects or (patho) physiological conditions. For instance, differences in the association patterns of blood metabolites have been observed in subjects with low and high latent cardiovascular risk.[8]

Estimating the PCN requires not only the definition of the most common metabolite associations observed in the population, but also assessing their significance in terms of biological information encoded in the relationship. This problem goes far beyond the results attainable with multivariate and pattern recognition methods and requires a more advanced and non-conventional representation of the individual metabolic phenotype.

We address this problem using the formalism of multiplex network[9–11] where a set of $M = 35$ metabolites (nodes) is connected through a multiplex network consisting of $N = 31$ subject-specific networks (or layers). Subsequently, we evaluate the probability of the $M(M-1)/2$ possible metabolite-metabolite associations by maximizing network entropy, thus using a statistical mechanics approach that takes advantage of network ensembles.

We arrive to an estimation of the PCN by building a significantly over-represented network (SON) that follows from a maximum entropy ensemble null model[12] with uncorrelated layers and fixed average degree sequence in each layer.[11,13] The SON is a weighted network that gathers all the metabolite-metabolite associations that are observed more than expected under the chosen null model and for which biological relevance can be inferred. An overview of the experimental and computational strategy is shown in Figure 2.

By making use of a community detection approach we show that the SON consists of biologically interpretable metabolite modules accounting for the biological mechanisms that shape of the individual metabolic phenotype. Contextually we introduce and discuss the concepts of subject network entropy and single metabolite entropy.

# Materials and Methods

## Entropy calculations

The group of $N$ subject-specific networks $G_\alpha$ can be considered as a multiplex network $\vec{G}$, *i.e.* a set of $M$ metabolites-nodes connected by $N$ networks or layers, each one fully described by its adjacency matrix $\{a_{ij}^{\alpha}\}$.

A proper null model is necessary to quantify how relevant is the abundance of each link, i.e. how far the real population abundance of a given link is from the expected value. Every observed metabolite-metabolite network $G_\alpha$ can be seen as a particular instance of a larger set of networks (an *ensemble*) that are assumed to be biologically equivalent and underlying the individual human metabolic phenotype. These networks share similar features such as the degree sequence (*i.e.* the number of links for each node) that constitute the main observables characterizing our system.[13] It follows that each network has a given probability to be observed, once the main features of the system have been outlined. Similarly, this *ensemble* approach can be applied also to multiplex networks, once the relationships among the layers have been specified.

As all subjects in the study are unrelated (apart two twins and a father-son pair), we considered the subject-specific networks $G_\alpha$ to be independent and we assumed the multiplex ensemble to be uncorrelated; in this case the probability $P(\vec{G})$ of observing a given multiplex can be factorized into the product of the probabilities $P_\alpha(G_\alpha)$ of observing each single network $G_\alpha$:

$$P(\vec{G}) = \prod_{\alpha=1}^{31} P_\alpha(G_\alpha). \tag{1}$$

We considered the particular case of a maximum entropy ensemble null model[12] defined by uncorrelated layers and fixed average degree sequence in each layer.[11,13] The degree sequences $\{k_i^\alpha\}$ are given by the real metabolite-metabolite association networks and are considered as properties to be satisfied on average over the multiplex ensemble. These properties are then likely to be satisfied but they are not matched perfectly by each multiplex network belonging to the ensemble. This multiplex ensemble is then *canonical*, i.e. shaped by *soft constraints.* The entropy value $S$ of the multiplex ensemble is additive in the number of layers and is a function of $p_{ij}^\alpha$, the probability of having a link between metabolite $i$ and $j$ in sample $\alpha$

$$S = \sum_{\alpha=1}^{31} S_\alpha = -\sum_{\alpha=1}^{31}\left[\sum_{i<j}^{35} p_{ij}^\alpha \log(p_{ij}^\alpha) + \sum_{i<j}^{35}(1-p_{ij}^\alpha)\log(1-p_{ij}^\alpha)\right] \tag{2}$$

where $S_\alpha$ is the entropy value of sample $\alpha$. The entropy $S_\alpha$ estimates the logarithm of the number of typical networks in the chosen ensemble, *i.e.* those networks satisfying on average the real degree sequence in layer $\alpha$.

The link probability $p_{ij}^{\alpha}$ is obtained by the constrained maximization of $S$, i.e.

$$\frac{\partial}{\partial p_{ij}^{\alpha}}\left\{S+\sum_{\alpha=1}^{31}\sum_{i=1}^{35}\lambda_i^{\alpha}\left(k_i^{\alpha}-\sum_{j=1}^{35}p_{ij}^{\alpha}\right)\right\}=0 \tag{3}$$

where $\lambda_i^{\alpha}$ are the Lagrange multipliers[14] related to the constraints for the degree sequences $\{k_i^{\alpha}\}$. For each $(i, j, \alpha)$ the resulting marginal probability is

$$p_{ij}^{\alpha}=\frac{e^{-(\lambda_i^{\alpha}+\lambda_j^{\alpha})}}{1+e^{-(\lambda_i^{\alpha}+\lambda_j^{\alpha})}} \tag{4}$$

Using the algorithms developed in,[13] we calculated the probabilities $\{p_{ij}^{\alpha}\}$ and $S_{\alpha}$ for each sample $\alpha$.

## Reconstruction of individual metabolic networks

Individual urine metabolite networks were reconstructed by taking a "wisdom of crowds" approach:[15] three algorithms for networks inference (ARACNE,[16] CLR[17] and PCLRC[8]) were considered and used with default parameters setting. In addition, also a standard correlation map was considered to define a fourth adjacency matrix as commonly used in metabolomics. For each subject $\alpha$ four different weighted adjacency matrices $\{w_{ij}^{\alpha}\}_m$ were built ($m = 1, \ldots, 4$ indicating the $m$-th method)

Binary adjacency matrices $\{a_{ij}^{\alpha}\}_m$ were obtained by imposing a threshold $\tau_m$ on the weighted adjacency matrices

$$\{a_{ij}^{\alpha}\}_m \rightarrow \begin{cases} 1 & \text{if } \left|\{w_{ij}^{\alpha}\}\right| > \tau_m \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The value of the threshold $\tau_m$ depends on the methods considered: values were 0 for ARACNE and CLR methods,[16,17] 0.95 for PCLRC[8] and 0.6 for the correlation map[18] as further detailed in.[19]

For each subject, the four adjacency matrices were superimposed

$$\{a_{ij}^{\alpha}\} = \sum_{m=1}^{4} \{a_{ij}^{\alpha}\}_m \tag{6}$$

The final adjacency matrix representing the metabolite network for every subject was built by retaining only those links inferred by 3 or more methods as suggested in:[19]

$$\{a_{ij}^{\alpha}\} \rightarrow \begin{cases} 1 & \text{if } \{a_{ij}^{\alpha}\} \geq 3. \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The obtained adjacency matrices $\{a_{ij}^{\alpha}\}$ are symmetric, unweighted and they constitute the input for network ensemble modeling. However, the formalism of network ensembles is well developed for both directed and undirected networks[20,21] and directionality can be taken into account as a possible feature of the model, depending on the different techniques of network inference and on the purposes of the analysis. Moreover, the magnitude of the association (e.g. correlation values or mutual information) could be retained to obtain weighted networks, where the binary patterns of link presence/absence is replaced by weights representing the strength of the associations. The formalism of network ensembles has also been extended to weighted multiplex networks.[11,20]

Metabolites are produced from other metabolites and this results in interdependence patterns between their concentrations that do not exist between transcripts or proteins and are constrained by stoichiometry:[18] when available, this information can be used to derive structural biochemical properties of the networks[22] and can be incorporated in the model. However although knowledge of primary metabolism is steadily growing,[23] very little is known about secondary metabolism, and stoichiometry-based analysis of metabolomics data is currently limited and would require extensive experimental work.

## Statistical analysis

### Individual Recognition

Data reduction was carried out by means of projection into a PCA subspace explaining 99.99% of the variance in the data. A test set validation (TSV) approach, which requires that models are constructed without any test set data, was applied to define the multivariate predictive analysis. Data were initially split in test and training data sets. The training set consisted of a random selection of the 90% of data available for all individuals. The test set consisted of the remaining 10% of the data. The training data sets were subjected to canonical analysis (MANOVA) to define a further reduced subspace with optimum group separation in the CA space. The training was then projected into the PCA/CA subspace defined by the training model. A k-NN classification (with k = 7) was applied to each test set for each individual. The procedure was repeated $10^3$ times to derive average recognition accuracy for each subject. Detailed information on the overall procedure is given in the original publications.[3,4]

## Modeling of network similarity

To model and visualize the differences observed in the subject-specific metabolic networks we used COVSCA (COVariance Simultaneous Component Analysis) which is a recently introduced model to analyze communalities and differences across a set of $C_k$ $(k = 1, 2, ..., K)$ covariance matrices simultaneously.[24] Since an adjacency matrix can be considered a particular case of covariance matrix, this method can be used to model adjacency matrices too. In COVSCA the matrices are approximated as a combination of a limited number $(L << K)$ of low dimensional prototypes:

$$C_k \approx \sum_{l=1}^{L} c_{kl} Z_l Z_l^T \tag{8}$$

where $c_{kl} \geq 0$ $(l = 1, 2, \ldots, L)$ are weight coefficients and $Z_l Z_l^T$ are the prototypical covariance matrices; these matrices define a set of loadings $Z$ of size $J \times R_l$ that hold simultaneously for all $C_k$. We have chosen to fit our model with two rank-2 prototype matrices as the best compromise between the goodness-of-fit (70%) of the COVSCA model (100% for perfect fit and 0 for total lack of fit) and the model complexity (rank of the prototypes matrices). This diagnostic measure is defined in Equation 31 of the original publication to which we refer the reader for more details on the model derivation and implementation.

## Metabolite set analysis

Biological interpretation of network modules was carried on by mean of literature mining and with the support of the Human Metabolome Database (www.hmdb.ca).[25] Results were complemented with metabolite set analysis[26] performed using the built in function of MetaboAnalyst 3.0,[27] that employs a library containing 88 metabolite sets based on normal metabolic pathways and using an hypergeometric test. A false discovery rate (FDR) of 0.1 was used as discriminant threshold for statistical significance for multiple testing.

## Experimental methods

Detailed experimental procedures can be retrieved from the original publications.[3–5]

### Sample Collection

Urine samples were collected from 31 healthy volunteers (14 males and 17 females). Each participant provided 37 samples collected on distinct days after an overnight fast, resulting in a total collection of 1147 urine samples. Urine samples were collected into prelabeled sterile collection cups and they were stored at -80°C.

### Sample preparation

Frozen urine samples were thawed at room temperature and shaken before use; 630 $\mu$L aliquot of each urine sample was centrifuged at 14000 g for 5 minutes and 540 $\mu$L of the supernatant were added to 60 $\mu$L of potassium phosphate buffer (1.5 M $K_2HPO_4$, 100%

(v/v) $2H_2O$, 10 mM sodium trimethylsilyl [2,2,3,3-$2H_4$]propionate (TMSP) pH 7.4). 450 $\mu$L of each mixture were transferred into 4.25 mm NMR tubes (Bruker BioSpin srl) for analysis.

### NMR experiments

$^1$H NMR spectra were acquired using a Bruker 600 MHz spectrometer (Bruker BioSpin) operating at 600.13 MHz proton Larmor frequency. Before measurement, samples were kept for at least 3 minutes inside the NMR probehead for temperature equilibration (300 $^o$K). For each urine sample, a monodimensional $^1$H NMR spectrum was acquired with a NOESY-presaturation pulse sequence (Bruker noesygppr1d.comp pulse sequence). 64 scans with 64 K data points were collected, using a spectral width of 12019 Hz, an acquisition time of 2.7s, a relaxation delay of 4s and a mixing time of 100ms.

### NMR spectra processing and Metabolite analysis

Free induction decays were multiplied by an exponential function equivalent to a 1.0 Hz line-broadening factor before applying Fourier transform. Transformed spectra were automatically corrected for phase and baseline distortions and calibrated (TMSP singlet at 0.00 ppm) using TopSpin 3.2 (Bruker Biospin srl). 35 metabolites, whose peaks in the spectra were well defined and resolved, were assigned. Signal identification was performed using a library of NMR spectra of pure organic compounds, public databases (such as HMBD and SDBS, Spectra Database for Organic Compounds, http://sdbs.db.aist.go.jp) storing reference NMR spectra of metabolites, spiking NMR experiments and literature data. The relative concentrations of the various metabolites in the different spectra were calculated by integrating the signal area.

# Results and Discussion

Following recent developments in network theory and systems biology we consider a *multiplex network* representation[9–11] as the more effective framework to model the patterns of metabolite-metabolite associations across the population.

Every subject $\alpha$ (with $\alpha = 1, 2, \ldots, 31$) is represented by a $M \times M$ association network $G_\alpha$, fully described by its adjacency matrix $\{a_{ij}^{\alpha}\}$, in which, for each couple of metabolites $i$ and $j$, the element $a_{ij}^{\alpha}$ can be either 1 or 0, whether the two metabolites are associated or not.

We define as $\phi_{ij}^{obs}$ the number of times that a particular association between metabolites $i$ and $j$ is observed in the population as

$$\phi_{ij}^{obs} = \sum_{\alpha}^{31} a_{ij}^{\alpha} \tag{9}$$

This measure is the first step for the estimation of layer overlap based on experimental observations on our samples. Given the nature of metabolite-metabolite association networks, many inferred links can be due to either sampling effects or biological noise and this is particularly relevant in case of urinary metabolites, where associations can result from metabolites being co-excreted rather than being linked by some biomolecular process. In this light it is of fundamental importance to define an adequate null hypothesis to assess the relevance of metabolite-metabolite associations.

For this task, we considered the analytical tools provided by statistical mechanics of network ensembles, i.e. families of randomized network variants of a given real network, where a set of structural constraints has been specified, while other topological features are completely random.[28] Hence, in this framework, $\{\phi_{ij}^{obs}\}$ is a particular realization on a single network instance of the more general multiplex property $\{\phi_{ij}\}$ whose statistics depends on the specific enforced constraints. A comparison between the empirical values $\{\phi_{ij}^{obs}\}$ and the ensemble statistics allows to quantify metabolite association relevance, providing a link-specific criterion for selection.

## Metabolite-metabolite association probability

The probability $p_{ij}^{\alpha}$ of having an association between metabolite $i$ and $j$ in layer $\alpha$ is obtained by the constrained maximization of the multiplex network entropy $S$. Specifically, we choose a maximum entropy ensemble null model[12] defined by uncorrelated layers and fixed average degree sequence in each layer, based on the observation that the subjects in the study are unrelated. This means that in our null model, for each metabolic network, the number of associations for each metabolite (node) is conserved rather than the observed biological associations between any two metabolites (links).

Figure 3 shows the distributions of $p_{ij}^{\alpha}$, that is the probability of having an association between metabolite $i$ and $j$ in the network of each subject $\alpha$. The distributions are equivalent (Kolmogorov-Smirnov test $p$-value $> 0.05$ for all possible comparisons) and bi-modal, with a peak in proximity of zero, indicating that most part of connections are unlikely, and a peak around 0.95, indicating highly probable metabolite-metabolite association backbone.

The probability $p_{ij}$ of having an association between any two metabolites is directly associated with the likelihood of connectivity of metabolite $i$ and metabolite $j$ given by the constraints imposed on the ensemble.

However, as urine acts as a sink collecting metabolites from different origins not necessarily belonging to the same metabolic pathways or biological processes, it is reasonable that some metabolite associations are spurious. Distinguishing between real and spurious associations is a well known problem in the network inference field and multivariate analysis:[8] the use of a network entropy null model corrects for such spurious association by means of a link-related significance threshold.

We also observed from the null model that $p_{ij}^{\alpha} \approx p_{ij}^{\alpha'}$: the associations between metabolites $i$ and $j$ result similar even if independent. This because under normal physiological conditions the biological processes described by $p_{ij}^{\alpha}$ and $p_{ij}^{\alpha'}$ are likely to occur with similar probability.

## Significantly over-represented network

Starting from the probability of metabolite-metabolite association in the network of each subject, the significance of over-represented links is calculated by comparing $\phi_{ij}^{obs}$ with its expected value over the multiplex ensemble

$$\langle \phi_{ij} \rangle = \sum_{\alpha=1}^{31} p_{ij}^{\alpha} \tag{10}$$

by defining a $z$-score function for each association:[28]

$$z_{ij}[\phi_{ij}] = \frac{\phi_{ij}^{obs} - \langle \phi_{ij} \rangle}{\sigma[\phi_{ij}]} \tag{11}$$

where

$$\sigma^2[\phi_{ij}] = \sum_{\alpha=1}^{31} p_{ij}^{\alpha}(1 - p_{ij}^{\alpha}) \tag{12}$$

is the variance of $\phi_{ij}$.

Since the quantity $z_{ij}$ indicates the deviation of $\phi_{ij}$ from its expected value under the null model, it is a direct measure of the significance of the over-representation of each observed metabolite association: we thus define the SON ( see Figure 4A), the *significantly over-represented network* as a weighted adjacency matrix

$$\{\Omega_{ij}\} = \max(z_{ij}, 0) \tag{13}$$

The expected value of $\phi_{ij}$ is derived under a maximum entropy ensemble null model, defined by uncorrelated networks and fixed average degree sequence in each networks, *i.e.* constraining the average number of associations of each metabolite but not the identity of associating metabolites.

This means that $\phi_{ij}^{obs}$ values for metabolite-metabolite associations in $\{\Omega_{ij}\}$ are caused by some non-random underlying biological process to a different magnitude based on the effective value of the related $z$-score.

The $z$-score $z_{ij}$ has a straightforward probabilistic interpretation if $\phi_{ij}$ follows approximately a Gaussian distribution. We verified the normality of the ensemble distribution for each metabolite-metabolite association by considering the association probabilities $\{p_{ij}^{\alpha}\}$ and generating $10^3$ different multiplex networks. We then calculated the corresponding $\{\phi_{ij}\}$ values for each realization. We found a fair agreement between the theoretical values of $\langle \phi_{ij} \rangle$ and $\sigma[\phi_{ij}]$, obtained analytically from Equations (10) and (12) and the values calculated from the realized distributions ($R^2_{\langle \phi_{ij} \rangle} = 1$ and $R^2_{\sigma[\phi_{ij}]} = 0.99$) as shown in Figure 5. Figure 7 shows the null distribution for 4 selected highly significant metabolite-metabolite associations in a comparison with the corresponding $\phi_{ij}^{obs}$.

## Detection of functional modules in SON

We aimed at defining modules of associated metabolites in the SON with the underlying assumption that metabolites in the same module pertain to the same biological functions. The study of network communities or modules can unravel the essential structure of the system in terms of a clearer functional description. We have chosen to perform a module detection of the SON maximizing the stability of a partition, a measure introduced in[29] that quantifies the quality of a partition in terms of the clustered autocovariance of a dynamic Markov process on the network. In this framework time acts as a resolution parameter, establishing a hierarchy of increasingly coarser partitions but also showing the most stable partitions in terms of time spans. The SON quickly reaches the stable module conformation shown in Figure 4A, where four different metabolite modules appear.

Module I (red) was found to be enriched for the synthesis and degradation of ketone bodies (fdr = 0.04), propanoate (FDR=0.08) and phenylalanine metabolism (fdr = 0.09) pathways. Remarkably, this module is enriched for metabolites that are linked to the activity of the gut microflora (5 out 6 metabolites with the exclusion of arginine; hypergeometric test $p$-value = 0.0004). In normal physiological conditions acetone is typically derived from acetoacetate through the action of microbial acetoacetate decarboxylases found in gut microflora. Indoxyl sulfate originates from bacterial protein fermentation in the large intestine where colonic microbiota degrade tryptophan to indole which is hydroxylated to 3-hydroxy-indole, the majority of which is sulfonated to indoxyl sulfate.[30] Hippurate and Phenylacetylglycine are also urinary gut microbial co-metabolites;[31] hippurate is the product of the conjugation of benzoate with glycine and this conjugation occurs via the formation of an intermediate, benzoyl CoA. This process takes place in the mitochondria of the liver and the kidney but urinary hippurate excretion is modulated by the intestinal microbiome;[32] similarly also phenylacetylglycine levels are modulated by the activity of gut microflora[33] and this metabolite is formed by the conjugation of phenylacetyl coenzyme A (CoA) with glycine.[34] The conjugation relationship between glycine and phenylacetylglycine is apparent in the strong intermodules link connecting these two metabolites in Module II (cyan). Phenylacetylglycine is also strongly connected with m-HPPA (m-hydroxyphenylpropionic acid) being both bacterial-mammal urinary co-metabolites, indicating that links between different modules happen at both the biochemical and functional level.

Module II is also enriched for metabolites linked to gut microflora ($p$-value $= 0.06$). Moreover, Hippurate is also strongly interconnected with 2-hydroxyisobutirate of Module III (green), another metabolite well known to be associated with microbial degradation of dietary proteins that escape digestion in the upper intestinal tract.[35]

Module IV (purple) is significantly enriched for amino acids metabolism and catabolism, namely valine, leucine and isoleucine biosynthesis (adjusted $p$-value $= 0.015$) and glycine, serine and threonine metabolism (adjusted $p$-value $= 0.045$). Interestingly the three branched-chain amino acid are strictly interconnected; valine is linked with leucine of Module II: this module is heterogenous in composition: it is enriched for metabolites involved in glycolate metabolism, pyruvate metabolism, glycine serine and threonine metabolism and aminoacyl t-RNA biosynthesis.

## Percolation of the SON

We further investigated the property of the SON by applying preliminarily a majority rule on the elements of the adjacency matrix $\{\Sigma_{ij}\}$ before thresholding on the z-score:

$$\{\Omega_{ij}^{\beta}\} \to \begin{cases} \Omega_{ij} & \text{if } \phi_{ij}^{obs} \geq \beta N \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

We considered for $\beta$ values 0.5 and 0.75, and applied the community detection algorithm on the resulting networks. Results are shown in Figure 4B and C, respectively. For $\beta = 0.5$ some of the nodes in the modules are lost but the overall structure of the SON is conserved for modules II (cyan), III (green) and IV (purple). Module I is lost and most of its nodes become disconnected. Also for $\beta = 0.75$ the three modules are retained. It is interesting to note that metabolites that get disconnected when the majority rule is applied are those pertaining to the activity of gut microflora. This indicates that although associations between these metabolites are observed in the population more often than expected, they are not uniformly conserved across the population. In contrast, associations pertaining fundamental biological pathways, such as amino acids metabolism and catabolism, are found consistently across the population.

In our opinion this reflects the individuality of relationships between host and gut microflora, since human gut microbiome is highly personalized at both taxonomic and functional levels[36] and thus the patterns of association among metabolites linked to the activity of gut microflora tends to be subject-specific.

This observation is also substantiated by the analysis of the entropy profiles of the single metabolites, as shown in the next Section.

## Single metabolite entropy

The individuality of relationships between metabolites is also substantiated by the entropy profiles of the single metabolites. The single node entropy for metabolite $i$ in sample $\alpha$ is defined as

$$S_i^\alpha = -\sum_{j \neq i} \frac{p_{ij}^\alpha}{k_i^\alpha} \log\left(\frac{p_{ij}^\alpha}{k_i^\alpha}\right) \quad \forall i, \alpha \tag{15}$$

where

$$k_i^\alpha = \sum_{j \neq i} p_{ij}^\alpha. \tag{16}$$

Each metabolite is then characterized by a single-node entropy distribution over the different $N$ samples.

The distribution of the single node entropy is given in Figure 6A. We observed that several metabolites tend to have entropy profiles markedly different from the others (see Figure 6B). Remarkably, different entropy profiles are associated with lower entropy levels as shown in the scatter plot in Figure 8. We found two major clusters that could be related to differential biological processes, namely amino acids metabolism and gut microflora activity. It is interesting to note that most metabolites associated with the activity of gut microflora are characterized by lower entropy values while those associated with amino acids metabolism have larger entropy. From an information theory point of view, lower entropy is associated with higher information content: we speculate that, in the absence of pathophysiological conditions, amino acids bio-synthesis follows the same pathways in all subjects and thus the content of information pertaining the metabolic status of every subject is reduced. On the contrary, metabolites associated with the activity of gut microflora are characterized by higher information level, as already suggested in the discussion of the SON.

## Entropy of the individual metabolic phenotype

The individual metabolic phenotype has been so far investigated by means of classical multivariate and pattern recognition methods. Here we have taken an entropy-based approach, following some recent development in the field. The concept of network entropy has been recently applied in a biological context, as a measure of the âĂIJparameter spaceâĂİ available to the cell (in terms of gene expression profile or clonal diversity) and it allowed successfully characterization of different cell states related to different cancer stages or to physiological ageing.[13]

The examined data set is rather homogeneous, in terms of age and health status, and in previous studies no natural clustering appeared other than between male and females.[3,4] When comparing the single sample entropy estimates $S_\alpha$ for males and females we did not observe any statistically significant difference, indicating no difference in the metabolic phenotypic space accessible to both genders.

We recently showed that the human metabolic phenotype shows both resilience and allostasis properties[5] and an entropy-based representation of both the diffusiveness of metabolic phenotypes and their collective divergence from homeostasis in unperturbed and perturbed system states was suggested.[37] This is consistent with the fact that the macroscopic resilience of a system is correlated to the level of uncertainty or entropy (disorder).

In particular, since profiling of urine and plasma samples can generate biomarkers of many types of organ dysfunctions,[37] this entropy-based approach can find application in the case of external perturbations like exposure to severe toxicity, condition in which metabolite relationship patterns may exhibit substantial changes that could lead to altered entropic profiles.

## Conclusions

Since the cellular function is governed by a complex network of biological interactions it seems natural to explore network properties which may help elucidate some of these features.

Here we addressed the problem of estimating the population core network, which we postulated to represent and underlay the individual metabolic phenotypes observed in healthy subjects. We examined the metabolic profiles of a panel of healthy subjects, who were considered to be representative of a healthy young population, and built a set of subject-specific metabolite-metabolite networks. The variety of network features observed in the cohort reflects the diversity observed in metabolic profiles.

This diversity is commonly attributed to the variation of intrinsic factors (such as genetic variation) and to extrinsic influences (such as diet habits, life-style and environmental conditions). Recent studies have shown that extrinsic factors may play a minor role in shaping the metabolic phenotype and this result has been confirmed in a comparative study involving both humans and *Macaca Mulatta*; the latter study has also shown that the metabolic phenotype accounts for both dynamic (part variable in time) and static (average level) component in the ratio 3/4 and 1/4 respectively. Part of the phenotypical diversity observed, such as sex differences, is encoded in the static part. What is of interest for the estimation of the PCN is the dynamic part, on which the definition of the metabolite-metabolite associations relies. The contribution of extrinsic factors to the shaping of the dynamic part was also found to be negligible, indicating that the metabolic phenotype arises mostly from intrinsic factors, including, to a certain extent, the activity of gut microflora.

The gut microbiome is involved in the regulation of multiple host metabolic pathways, giving rise to interactive host-microbiota metabolic, signaling, and immune-inflammatory axes that physiologically connect the gut, liver, muscle, and brain;[38] its changes and modifications have been associated to several pathophysiological conditions such as obesity,[39] diabetes,[40] autoimmune diseases[41] and neuropsychiatric disorders.[42]

The dynamic patterns of metabolite concentration can be used to define patterns of association (including co-variation and mutual information relationships). The significant occurrence of such associations in the population was estimated by considering a null model, not dissimilarly to what is done in the hypothesis testing framework.

We showed that the model employed is able to describe the diversity of metabolite association found in the metabolic networks related to the individual metabolic phenotype, providing a possibly mechanistic new framework to explore individual metabolic phenotype and its association with pathophysiological conditions.

This is currently investigated through a top-down system biology approach, where multivariate analysis is applied to metabolic phenotyping for detecting key players in the shaping and maintenance of an healthy status. The definition and the characterization of a metabolite-metabolite association network underlying an healthy status can enable a bottom-up system biology approach for detailed modeling of the individual metabolic profile on the basis of its molecular properties. In this a bottom-up approach, molecular networks can be quantitatively studied leading to predictive models[43] that can be applied within system medicine and/or personalized medicine approaches.

# Associated content

**Figure S1** - Single metabolite entropy diversity

# Acknowledgements

This work was partly supported by the European Commission funded FP7 projects *PhenoMeNal* (Contract No. 654241) and *INFECT* (Contract No. 305340). The Fondazione Veronesi granted L.T. through the Post-Doctoral Fellowship-2015. D.R. and G.M. were partially funded by the European Commission funded FP7 *Mimomics* (Contract No. 305280) and IT MIUR CNR Flagship *Interomics* Grant.

## Author Contributions

E.S. and G.M. contributed equally to the paper. E.S. conceived the study. E.S., G.M. and V.G. performed analysis. E.S, G.M. L.T. and D.R. interpreted the results. E.S, G.M. L.T. and V.G. drafted the manuscript. E.S., L.T. and C.L. revised the manuscript. All authors reviewed and approved the final manuscript.

# Additional information

## Competing financial interests

Authors declare no competing financial interests.

# References

(1) Emmert-Streib, F.; Glazko, G. V. Pathway analysis of expression data: deciphering functional building blocks of complex diseases. *PLoS Comput. Biol.* **2011**, *7*, e1002053.

(2) Xia, J.; Wishart, D. S. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.* **2010**, *38*, W71–W77.

(3) Assfalg, M.; Bertini, I.; Colangiuli, D.; Luchinat, C.; SchÃďfer, H.; SchÃijtz, B.; Spraul, M. Evidence of different metabolic phenotypes in humans. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1420–1424.

(4) Bernini, P.; Bertini, I.; Luchinat, C.; Nepi, S.; Saccenti, E.; SchÃďfer, H.; SchÃijtz, B.; Spraul, M.; Tenori, L. Individual Human Phenotypes in Metabolic Space and Time. *J. Proteome Res.* **2009**, *8*, 4264–4271.

(5) Ghini, V.; Saccenti, E.; Tenori, L.; Assfalg, M.; Luchinat, C. Allostasis and resilience of the human individual metabolic phenotype. *J. Proteome Res.* **2015**, *14*, 2951–2962.

(6) Saccenti, E.; Tenori, L.; Verbruggen, P.; Timmerman, M. E.; Bouwman, J.; van der Greef, J.; Luchinat, C.; Smilde, A. K. Of monkeys and men: A metabolomic analysis of static and dynamic urinary metabolic phenotypes in two species. *PloS one* **2014**, *9*, e106077.

(7) Töpfer, N.; Kleessen, S.; Nikoloski, Z. Integration of metabolomics data into metabolic networks. *Front. Plant Sci.* **2015**, *6*.

(8) Saccenti, E.; Suarez-Diez, M.; Luchinat, C.; Santucci, C.; Tenori, L. Probabilistic networks of blood metabolites in healthy subjects as indicators of latent cardiovascular risk. *J. Proteome Res.* **2014**, *14*, 1101–1111.

(9) Boccaletti, S.; Bianconi, G.; Criado, R.; del Genio, C.; Gómez-Gardeñes, J.; Ro-

mance, M.; Sendiña-Nadal, I.; Wang, Z.; Zanin, M. The structure and dynamics of multilayer networks. *Phys. Rep.* **2014**, *544*, 1–122.

(10) Castellani, G. C. et al. Systems medicine of inflammaging. *Brief. Bioinform.* **2015**, bbv062–.

(11) Menichetti, G.; Remondini, D.; Bianconi, G. Correlations between weights and overlap in ensembles of weighted multiplex networks. *Phys. Rev. E* **2014**, *90*, 062817.

(12) Jaynes, E. T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630.

(13) Menichetti, G.; Bianconi, G.; Castellani, G.; Giampieri, E.; Remondini, D. Multiscale characterization of aging and cancer progression by a novel Network Entropy measure. *Mol. BioSyst.* **2015**,

(14) Lagrange, J. L. *Mécanique analytique*; Mallet-Bachelier, 1853.

(15) Marbach, D.; Costello, J. C.; Küffner, R.; Vega, N. M.; Prill, R. J.; Camacho, D. M.; Allison, K. R.; Kellis, M.; Collins, J. J.; Stolovitzky, G. Wisdom of crowds for robust gene network inference. *Nat. Methods* **2012**, *9*, 796–804.

(16) Margolin, A. A.; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Favera, R. D.; Califano, A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics* **2006**, *7*, S7.

(17) Faith, J. J.; Hayete, B.; Thaden, J. T.; Mogno, I.; Wierzbowski, J.; Cottarel, G.; Kasif, S.; Collins, J. J.; Gardner, T. S. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.* **2007**, *5*, e8.

(18) Camacho, D.; de la Fuente, A.; Mendes, P. The origin of correlations in metabolomics data. *Metabolomics* **2005**, *1*, 53–63.

(19) Suarez-Diez, M.; Saccenti, E. Effects of Sample Size and Dimensionality on the Performance of Four Algorithms for Inference of Association Networks in Metabonomics. *J. Proteome Res.* **2015**, *14*, 5119–5130, PMID: 26496246.

(20) Menichetti, G.; Remondini, D.; Panzarasa, P.; Mondragón, R. J.; Bianconi, G. Weighted multiplex networks. *PloS one* **2014**, *9*, e97857.

(21) Menichetti, G.; Remondini, D. Entropy of a network ensemble: Definitions and applications to genomic data. *Theor. Biol. Forum* **2014**, *107*, 77–87.

(22) Schilling, C. H.; Schuster, S.; Palsson, B. O.; Heinrich, R. Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.* **1999**, *15*, 296–303.

(23) Thiele, I.; Swainston, N.; Fleming, R. M.; Hoppe, A.; Sahoo, S.; Aurich, M. K.; Haraldsdottir, H.; Mo, M. L.; Rolfsson, O.; Stobbe, M. D. A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* **2013**, *31*, 419–425.

(24) Smilde, A. K.; Timmerman, M. E.; Saccenti, E.; Jansen, J. J.; Hoefsloot, H. C. J. Covariances Simultaneous Component Analysis: a new method within a framework for modeling covariances. *J. Chemom.* **2015**, *29*, 277–288.

(25) Wishart, D. S. et al. HMDB 3.0âĂŤThe Human Metabolome Database in 2013. *Nucleic Acids Res.* **2013**, *41*, D801–D807.

(26) Xia, J.; Wishart, D. S. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.* **2010**, *38*, W71–W77.

(27) Xia, J.; Sinelnikov, I. V.; Han, B.; Wishart, D. S. MetaboAnalyst 3.0âĂŤmaking metabolomics more meaningful. *Nucleic Acids Res.* **2015**,

(28) Squartini, T.; Garlaschelli, D. Analytical maximum-likelihood method to detect patterns in real networks. *New J. Phys.* **2011**, *13*, 083001.

(29) Delvenne, J.-C.; Yaliraki, S. N.; Barahona, M. Stability of graph communities across time scales. *Proc. Natl. Acad. Sci. U.S.A. of the United States of America* **2010**, *107*, 12755–12760.

(30) Meijers, B. K.; Evenepoel, P. The gutâĂŞkidney axis: indoxyl sulfate, p-cresyl sulfate and CKD progression. *Nephrol. Dial. Transplant* **2011**, *26*, 759–761.

(31) Yap, I. K. S.; Li, J. V.; Saric, J.; Martin, F.-P.; Davies, H.; Wang, Y.; Wilson, I. D.; Nicholson, J. K.; Utzinger, J.; Marchesi, J. R.; Holmes, E. Metabonomic and Microbiological Analysis of the Dynamic Effect of Vancomycin-Induced Gut Microbiota Modification in the Mouse. *J. Proteome Res.* **2008**, *7*, 3718–3728.

(32) Williams, H.; Cox, I. J.; Walker, D.; Cobbold, J.; Taylor-Robinson, S.; Marshall, S.; Orchard, T. Differences in gut microbial metabolism are responsible for reduced hippurate synthesis in Crohn's disease. *BMC Gastroenterology* **2010**, *10*, 108.

(33) Claus, S. P. et al. Colonization-induced host-gut microbial metabolic interaction. *MBio* **2011**, *2*, e00271–10.

(34) Jones, A. R. Some observations on the urinary excretion of glycine conjugates by laboratory animals. *Xenobiotica* **1982**, *12*, 387–395.

(35) Calvani, R.; Miccheli, A.; Capuani, G.; Miccheli, A. T.; Puccetti, C.; Delfini, M.; Iaconelli, A.; Nanni, G.; Mingrone, G. Gut microbiome-derived metabolites characterize a peculiar obese urinary metabotype. *Int. J. Obes.* **2010**, *34*, 1095–1098.

(36) Voigt, A. Y.; Costea, P. I.; Kultima, J. R.; Li, S. S.; Zeller, G.; Sunagawa, S.; Bork, P. Temporal and technical variability of human gut metagenomes. *Genome biol.* **2015**, *16*, 73.

(37) Veselkov, K. A.; Pahomov, V. I.; Lindon, J. C.; Volynkin, V. S.; Crockford, D.; Osipenko, G. S.; Davies, D. B.; Barton, R. H.; Bang, J.-W.; Holmes, E.; Nicholson, J. K.

A Metabolic Entropy Approach for Measurements of Systemic Metabolic Disruptions in Patho-Physiological States. *J. Proteome Res.* **2010**, *9*, 3537–3544.

(38) Nicholson, J. K.; Holmes, E.; Kinross, J.; Burcelin, R.; Gibson, G.; Jia, W.; Pettersson, S. Host-gut microbiota metabolic interactions. *Science* **2012**, *336*, 1262–1267.

(39) Ley, R. E.; Turnbaugh, P. J.; Klein, S.; Gordon, J. I. Microbial ecology: human gut microbes associated with obesity. *Nature* **2006**, *444*, 1022–1023.

(40) Qin, J.; Li, Y.; Cai, Z.; Li, S.; Zhu, J.; Zhang, F.; Liang, S.; Zhang, W.; Guan, Y.; Shen, D. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **2012**, *490*, 55–60.

(41) Cerf-Bensussan, N.; Gaboriau-Routhiau, V. The immune system and the gut microbiota: friends or foes? *Nat. Rev. Immunol.* **2010**, *10*, 735–744.

(42) Cryan, J. F.; Dinan, T. G. Mind-altering microorganisms: the impact of the gut microbiota on brain and behaviour. *Nat. Rev. Neurosci.* **2012**, *13*, 701–712.

(43) Bruggeman, F. J.; Hornberg, J. J.; Boogerd, F. C.; Westerhoff, H. V. *Plant Systems Biology*; Springer, 2007; pp 1–19.

# Figures

### Figure 1

Multivariate representation of the indivdual metabolic phenotype. Projection of metabolite data onto the three dimensional PCA-CA discriminant space. A) samples belonging to same subjects cluster together. B) Natural separation between male and female samples in the PCA-CA subspace. The accuracy for the discrimination in the PCA-CA space is 99% when using the full bucketed NMR spectra and 77% when using $M = 35$ metabolites. Representation in the COVSCA space of the the metabolite-metabolite association networks (see

Methods Section for details): C) color coded by individual and D) color coded by sex. The separation between males and females is lost in the network representation.

**Figure 2**

Overview of the experimental and computational approach to estimate the core population network underlying the individual human metabolic phenotype through the definition of a Significantly Over-represented Network. A) Construction of subject specific metabolite-metabolite association networks. $N = 31$ healthy subjects of both sex are considered and sampled for 37 consecutive days for their urine profiles. $M = 35$ urine metabolites are quantified using NMR spectroscopy. Association networks are inferred considering 4 different methods for network inference. B) Entropy based approach. The 31 individual networks are considered. The group of $N$ subject-specific networks can be considered as a multiplex network $\vec{G}$, *i.e.* a set of $M$ metabolites-nodes differently connected in $N$ networks or layers. By maximizing the entropy of the multiplex the probability of each metabolite-metabolite association can be calculated. C) By defining a null model with uncorrelated networks and fixed average degree sequence in each layer, the significance of metabolite-metabolite association is assessed through a $z$-score $z_{ij}$, measuring the deviation of $\phi_{ij}^{obs}$ from its expected theoretical value $\langle \phi_{ij} \rangle$. D) The Significantly Over-represented Network gathers all metabolite associations satisfying the condition $\max(z_{ij}, 0)$. Contextually the single metabolite and the entropy of the individual metabolic phenotype are introduced.

**Figure 3**

Distribution of the metabolite-metabolite association probability $p_{ij}^{\alpha}$ for each subject $\alpha$ in the study. The probabilities are calculated with Equation (4).

**Figure 4**

A) Graphical representation of the SON. The link weights are the $z$-scores defined in Equation (). The different coloring represent the 4 metabolic modules. A force-based layout is used for network visualization. B) Percolated version of the SON using a majority rule threshold

$\beta = 0.5$. C) Percolated version of the SON with $\beta = 0.75$.

**Figure 5**

A) Observed and theoretical values for for $< \phi_{ij} >$ and B) its standard deviation $\sigma[\phi_{ij}]$

**Figure 6**

A) Overall distribution of the single metabolite entropy across all subjects B) Distribution of the single metabolite entropy

**Figure 7**

Link abundance ensemble distribution for four links, one for each metabolic module. We simulated $10^3$ different multiplex networks in order to generate a distribution for $\{\phi_{ij}\}$. The color code is related to the functional module affiliation consistently with Figure 4. In each panel the observed $\phi_{ij}^{real}$ is significantly higher than the related ensemble average, meaning that the metabolite-metabolite associations are extremely relevant. **A)** the results for the association phenylacetylglycine-hippurate: $\phi_{ij}^{real} = 29$, $\langle\phi_{ij}\rangle = 8.46$ and $z_{ij} = 10.80$. **B)** results for the association choline-creatinine: $\phi_{ij}^{real} = 17$, $\langle\phi_{ij}\rangle = 6.88$ and $z_{ij} = 5.11$. **C)** results for the association trigonelline-glycine: $\phi_{ij}^{real} = 14$, $\langle\phi_{ij}\rangle = 3.75$ and $z_{ij} = 6.90$. **D)** results for the association 2-hydroxyisobutyrate-1-methylnicotinamide: $\phi_{ij}^{real} = 23$, $\langle\phi_{ij}\rangle = 13.30$ and $z_{ij} = 4.46$.

**Figure 8**

Relationship between the median single metabolite (node) entropy and its diversity. For metabolite $i$, diversity is calculated by comparing its entropy values on the 31 samples with the remaining $M-1$ metabolites, using a Kruskal-Wallis test with Tukey's honest significant difference criterion. The percentage of significant tests (over $M-1$) defines the diversity. The entropy diversity for each metabolites is given in Supplementary Figure S1.

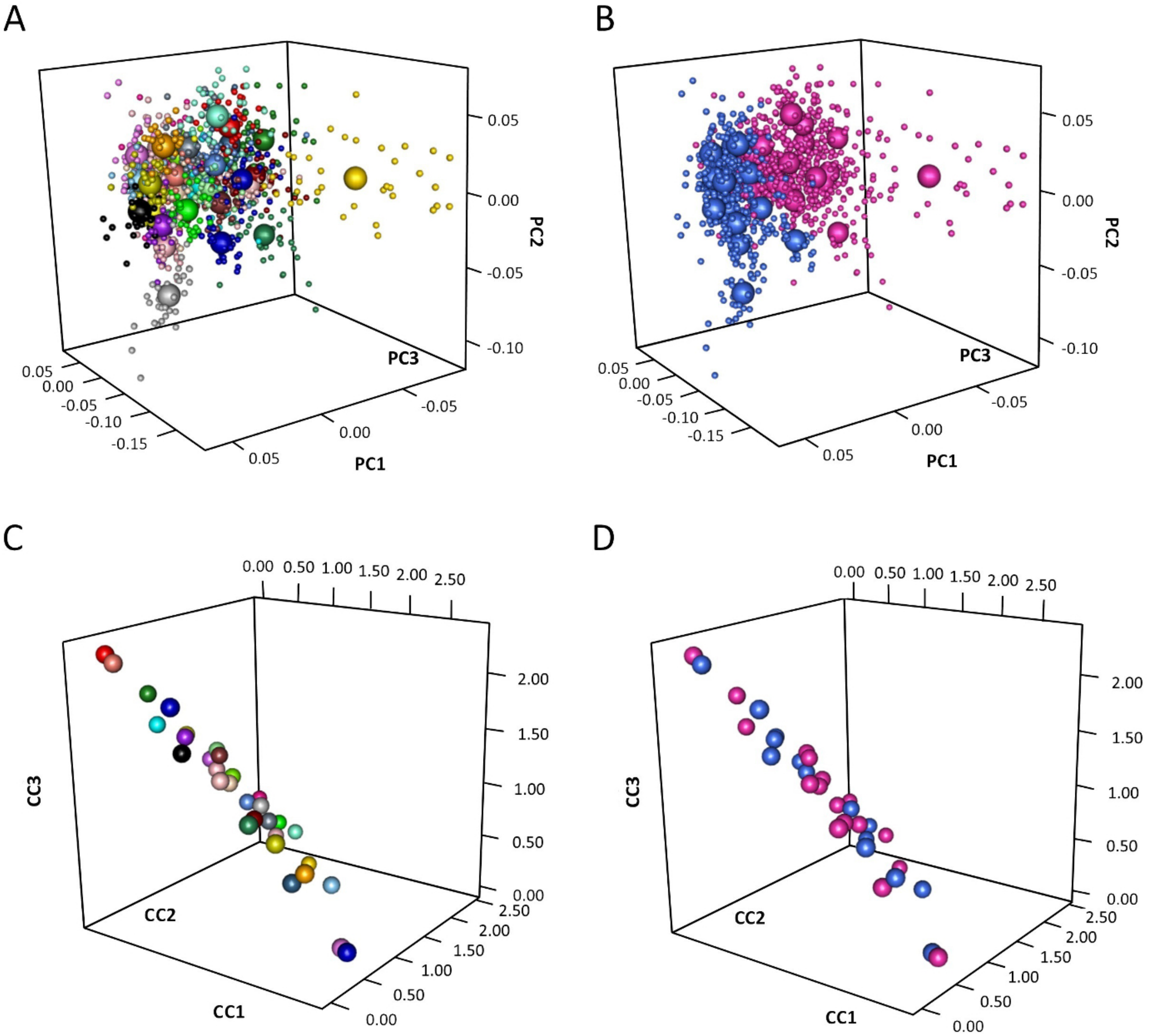
A
0.05
0.00
-0.05
-0.10
PC2
PC3
0.05
0.00
-0.05
-0.10
-0.15
-0.05
0.00
0.05
PC1
B
0.05
0.00
-0.05
-0.10
PC2
PC3
0.05
0.00
-0.05
-0.10
-0.15
-0.05
0.00
0.05
PC1
C
0.00 0.50 1.00 1.50 2.00 2.50
CC3
2.00
1.50
1.00
0.50
0.00
CC2
2.50
2.00
1.50
1.00
0.50
0.00
CC1
D
0.00 0.50 1.00 1.50 2.00 2.50
CC3
2.00
1.50
1.00
0.50
0.00
CC2
2.50
2.00
1.50
1.00
0.50
0.00
CC1

Figure 1

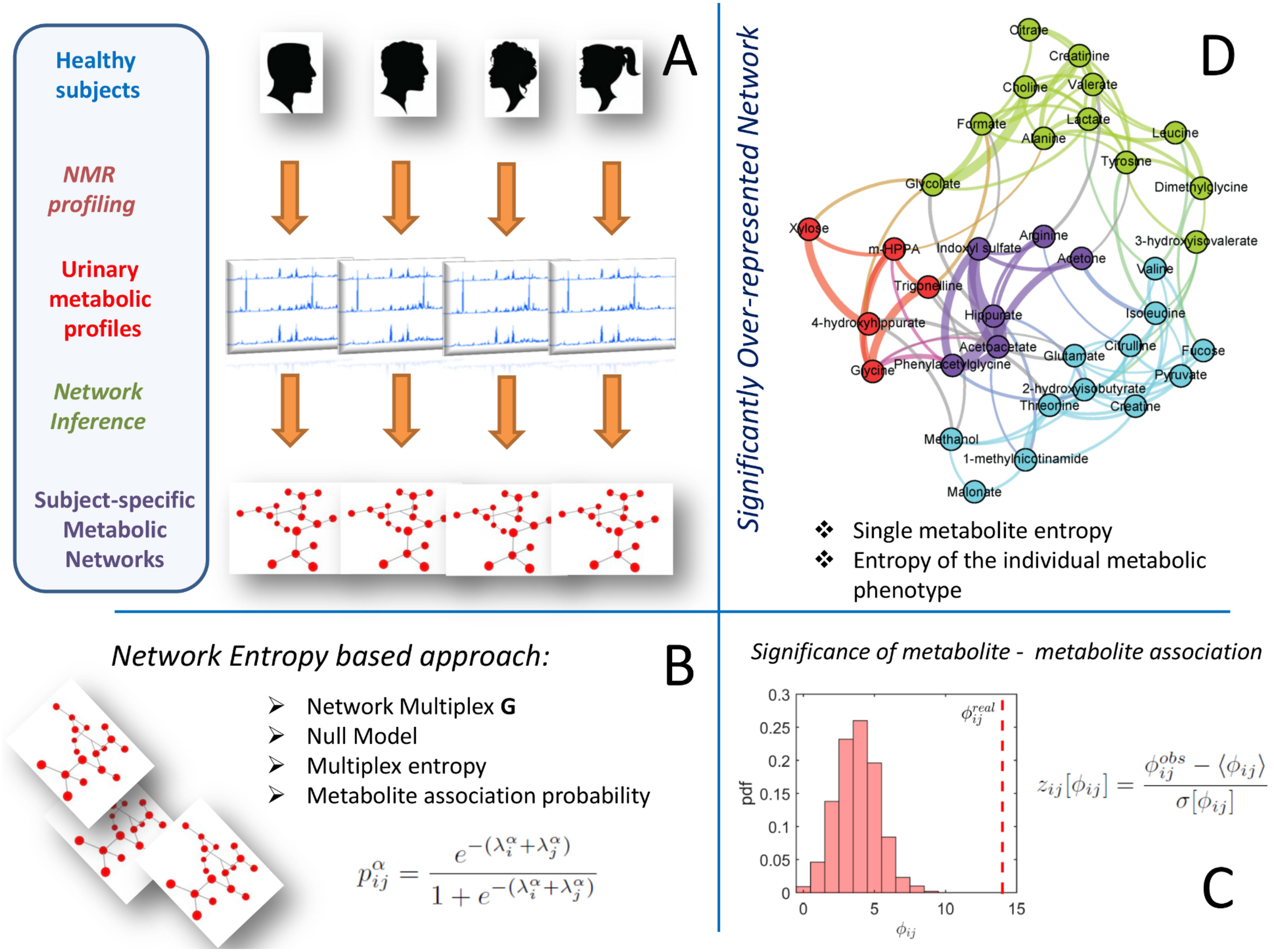
A
Healthy subjects
NMR profiling
Urinary metabolic profiles
Network Inference
Subject-specific Metabolic Networks
B
Network Entropy based approach:
Network Multiplex G
Null Model
Multiplex entropy
Metabolite association probability
$p_{ij}^{\alpha} = \frac{e^{-(\lambda_i^{\alpha}+\lambda_j^{\alpha})}}{1+e^{-(\lambda_i^{\alpha}+\lambda_j^{\alpha})}}$
C
Significance of metabolite - metabolite association
pdf
0.3
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
$\phi_{ij}$
$\phi_{ij}^{real}$
$z_{ij}[\phi_{ij}] = \frac{\phi_{ij}^{obs} - \langle \phi_{ij} \rangle}{\sigma[\phi_{ij}]}$
D
Significantly Over-represented Network
Citrate
Creatinine
Choline
Valerate
Formate
Lactate
Alanine
Leucine
Tyrosine
Glycolate
Dimethylglycine
Xylose
Arginine
m-HPPA
Indoxyl sulfate
Acetone
3-hydroxyisovalerate
Valine
Trigonelline
4-hydroxyhippurate
Hippurate
Isoleucine
Acetoacetate
Citrulline
Fucose
Glutamate
Glycine
Phenylacetylglycine
Pyruvate
2-hydroxyisobutyrate
Threonine
Creatine
Methanol
1-methylnicotinamide
Malonate
Single metabolite entropy
Entropy of the individual metabolic phenotype

Figure 2

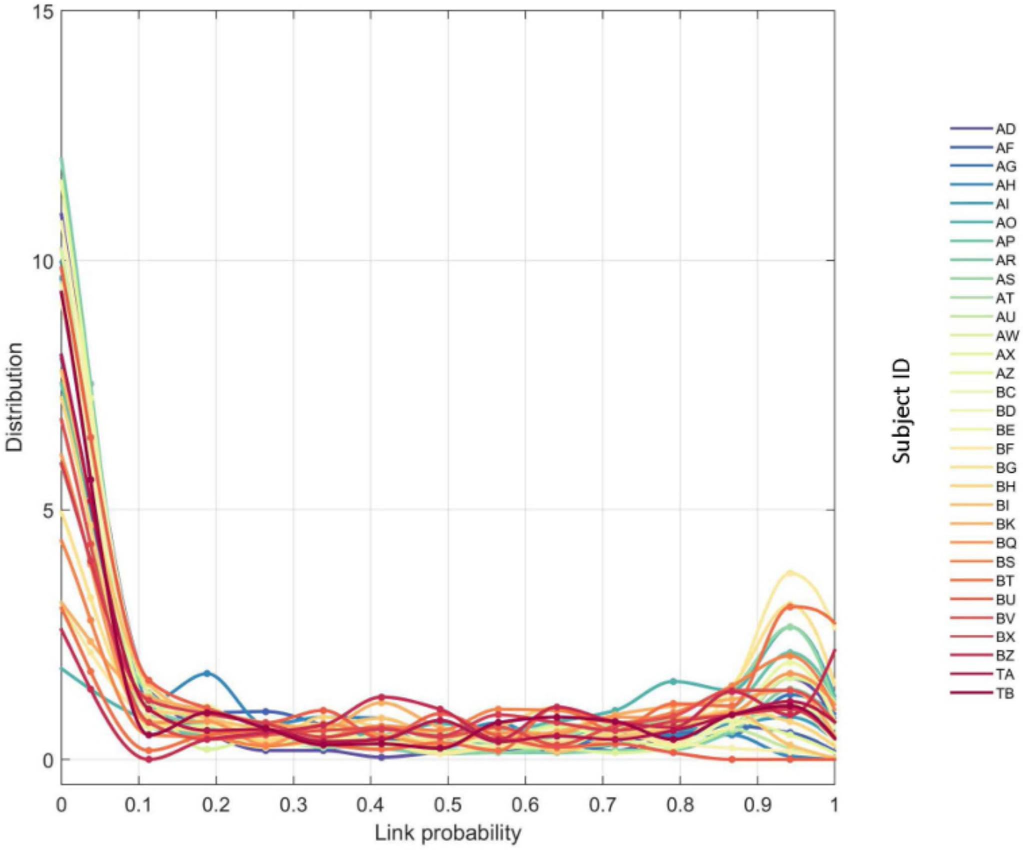
15
10
5
0
Distribution
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Link probability
Subject ID
AD
AF
AG
AH
AI
AO
AP
AR
AS
AT
AU
AW
AX
AZ
BC
BD
BE
BF
BG
BH
BI
BK
BQ
BS
BT
BU
BV
BX
BZ
TA
TB

Figure 3

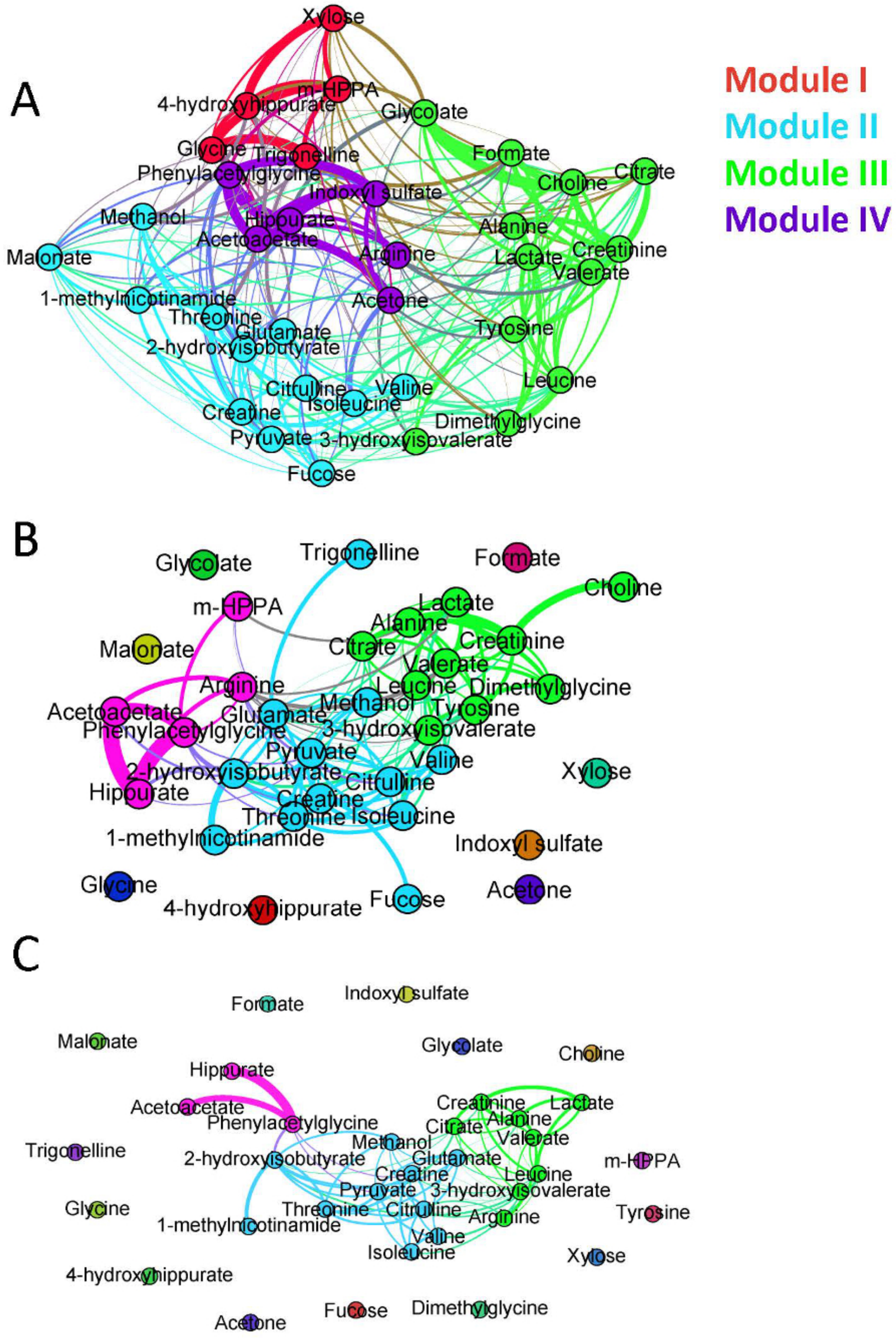
A
Module I
Module II
Module III
Module IV
Xylose
m-HPPA
4-hydroxyhippurate
Glycolate
Glycine
Trigonelline
Formate
Citrate
Choline
Phenylacetylglycine
Indoxyl sulfate
Methanol
Hippurate
Alanine
Acetoacetate
Creatinine
Malonate
Arginine
Lactate
Valerate
1-methylnicotinamide
Acetone
Threonine
Glutamate
Tyrosine
2-hydroxyisobutyrate
Leucine
Citrulline
Valine
Isoleucine
Creatine
Dimethylglycine
Pyruvate
3-hydroxyisovalerate
Fucose
B
Glycolate
Trigonelline
Formate
Choline
m-HPPA
Lactate
Alanine
Malonate
Citrate
Creatinine
Valerate
Arginine
Leucine
Dimethylglycine
Methanol
Acetoacetate
Tyrosine
Glutamate
Phenylacetylglycine
3-hydroxyisovalerate
Pyruvate
Valine
2-hydroxyisobutyrate
Xylose
Hippurate
Citrulline
Creatine
Threonine
Isoleucine
Indoxyl sulfate
1-methylnicotinamide
Glycine
Acetone
4-hydroxyhippurate
Fucose
C
Indoxyl sulfate
Formate
Malonate
Glycolate
Choline
Hippurate
Creatinine
Lactate
Acetoacetate
Alanine
Phenylacetylglycine
Citrate
Valerate
Trigonelline
Methanol
2-hydroxyisobutyrate
Glutamate
m-HPPA
Leucine
Creatine
Pyruvate
3-hydroxyisovalerate
Glycine
Threonine
Citrulline
Tyrosine
Arginine
1-methylnicotinamide
Valine
Isoleucine
Xylose
4-hydroxyhippurate
Fucose
Dimethylglycine
Acetone

Figure 4

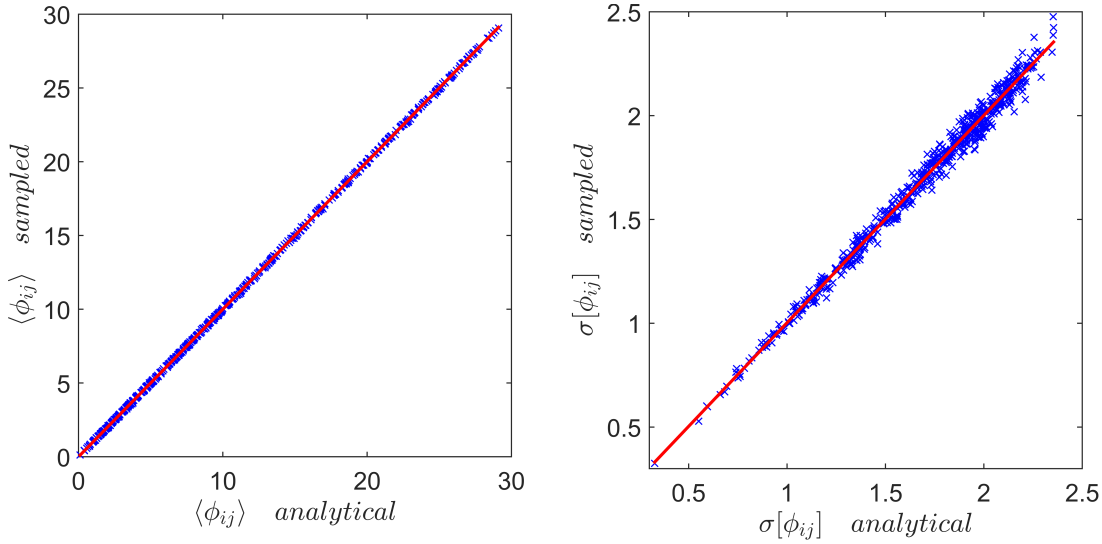
$\langle\phi_{ij}\rangle$ *sampled*
30
25
20
15
10
5
0
0
10
20
30
$\langle\phi_{ij}\rangle$ *analytical*
$\sigma[\phi_{ij}]$ *sampled*
2.5
2
1.5
1
0.5
0.5
1
1.5
2
2.5
$\sigma[\phi_{ij}]$ *analytical*

Figure 5

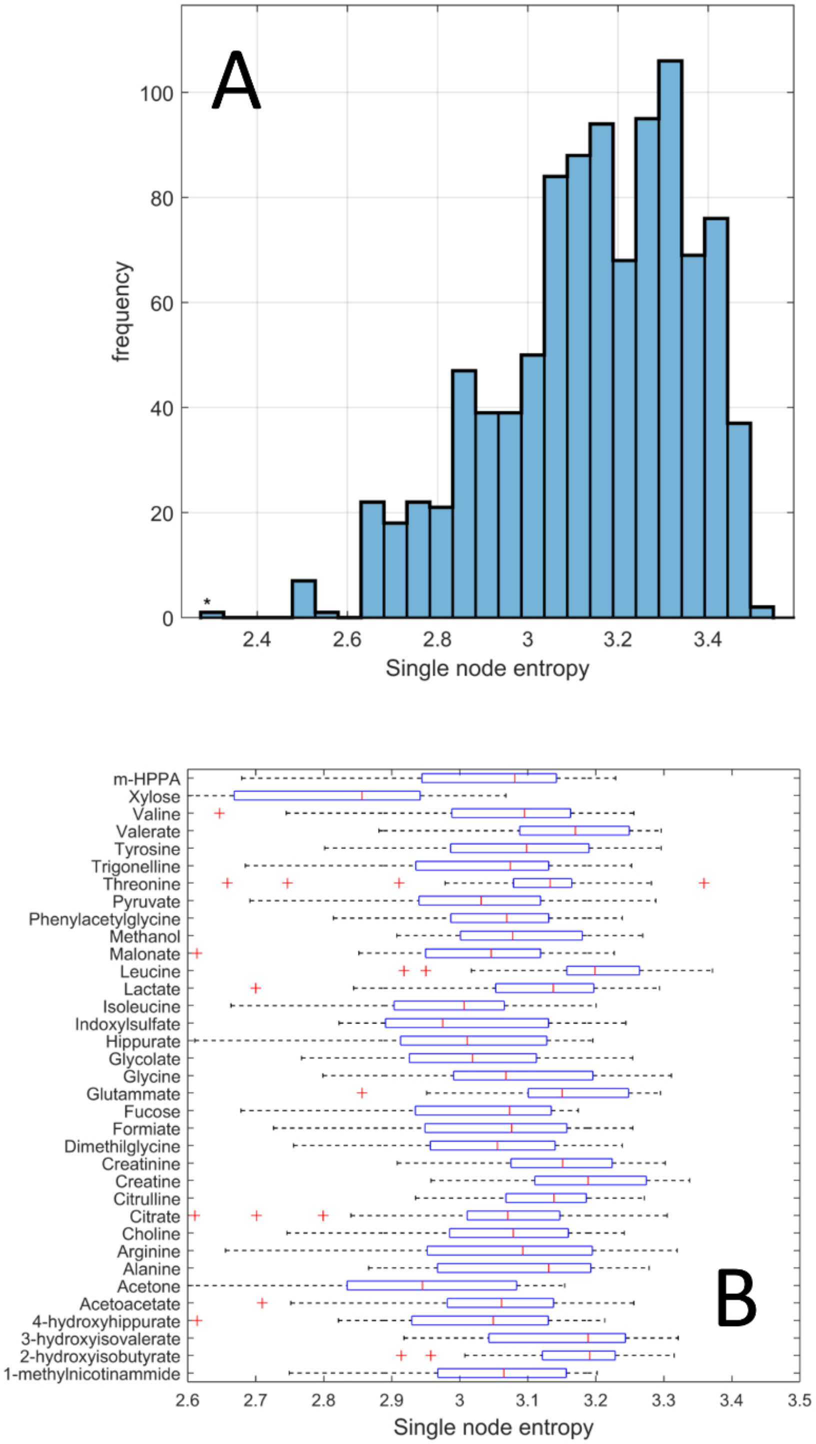
A
frequency
0
20
40
60
80
100
2.4
2.6
2.8
3
3.2
3.4
Single node entropy
*
m-HPPA
Xylose
Valine
Valerate
Tyrosine
Trigonelline
Threonine
Pyruvate
Phenylacetylglycine
Methanol
Malonate
Leucine
Lactate
Isoleucine
Indoxylsulfate
Hippurate
Glycolate
Glycine
Glutammate
Fucose
Formiate
Dimethilglycine
Creatinine
Creatine
Citrulline
Citrate
Choline
Arginine
Alanine
Acetone
Acetoacetate
4-hydroxyhippurate
3-hydroxyisovalerate
2-hydroxyisobutyrate
1-methylnicotinammide
2.6
2.7
2.8
2.9
3
3.1
3.2
3.3
3.4
3.5
Single node entropy
B

Figure 6

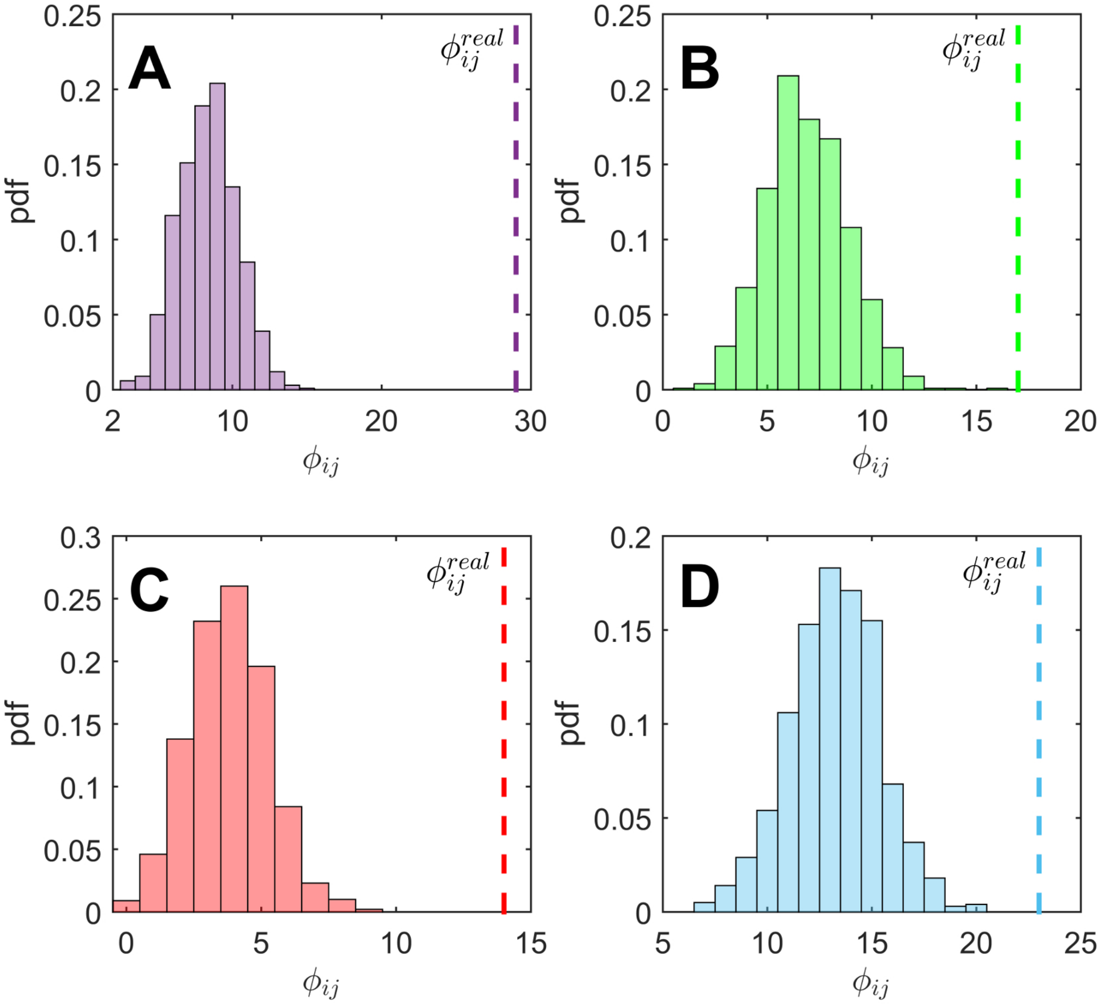
A
B
C
D
pdf
$\phi_{ij}$
$\phi_{ij}^{real}$
0.25
0.2
0.15
0.1
0.05
0
2
10
20
30
0
5
10
15
20
0.3
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
0.2
0.15
0.1
0.05
0
5
10
15
20
25

Figure 7

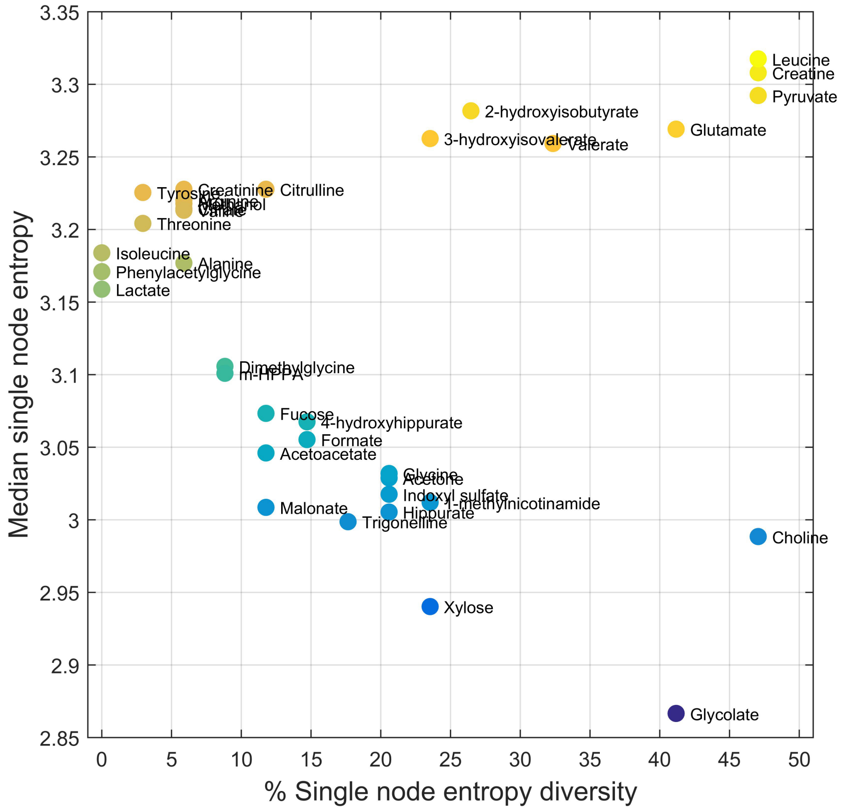
Leucine
Creatine
Pyruvate
2-hydroxyisobutyrate
Glutamate
3-hydroxyisovalerate
Valerate
Tyrosine
Creatinine
Citrulline
Arginine
Methanol
Valine
Threonine
Isoleucine
Alanine
Phenylacetylglycine
Lactate
Dimethylglycine
m-HPPA
Fucose
4-hydroxyhippurate
Formate
Acetoacetate
Glycine
Acetone
Indoxyl sulfate
1-methylnicotinamide
Malonate
Hippurate
Trigonelline
Choline
Xylose
Glycolate
Median single node entropy
% Single node entropy diversity
3.35
3.3
3.25
3.2
3.15
3.1
3.05
3
2.95
2.9
2.85
0
5
10
15
20
25
30
35
40
45
50

Figure 8