Electronic Supplementary Material (ESI) for Molecular BioSystems

Supplementary information for "Uncovering the hidden geometry of metabolic networks"

M. Ángeles Serrano,[1] Marián Boguñá,[2] and Francesc Sagués[1]

[1]*Departament de Química Física, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain*
[2]*Departament de Física Fonamental, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain*

Contents

THE $\mathbb{S}^1$ MODEL AND ITS EXTENSION TO BIPARTITE NETWORKS

The $\mathbb{S}^1$ model [1] is a complex network generator able to generate networks which are, simultaneously, scale-free, small-worlds, and highly clustered, as observed in the majority of real networks [14]. Nodes in this model are distributed in a metric space (in the simplest case a one-dimensional circle) abstracting (di)similarities among the elements of the network. The $\mathbb{S}^1$ model generates networks according to the following steps:

1. Distribute N nodes uniformly over the circle $\mathbb{S}^1$ of radius $R = N/(2\pi)$, so that the node density on the circle is fixed to 1.

2. Assign to all nodes a hidden variable κ representing their expected degrees. To generate scale-free networks, κ is drawn from the power-law distribution

$$\rho(\kappa) = \kappa_0^{\gamma-1}(\gamma-1)\kappa^{-\gamma}, \quad \kappa \in [\kappa_0, \infty), \tag{1}$$

$$\kappa_0 = \langle k \rangle \frac{\gamma-2}{\gamma-1}, \tag{2}$$

   where κ_0 is the minimum expected degree, and $\langle k \rangle$ is the network average degree.

3. Let κ and κ' be the expected degrees of two nodes located at distance $d = N\Delta\theta/(2\pi)$ measured over the circle, where $\Delta\theta$ is the angular distance between the nodes. Connect each pair of nodes with probability $p(x)$, where the *effective* distance is defined as $x \equiv d/(\mu\kappa\kappa')$, and μ is a constant fixing the average degree.

The connection probability $p(x)$ can be any integrable function. Here we chose the distribution

$$p(x) = \frac{1}{1+x^\beta}, \tag{3}$$

where β is a parameter that controls clustering in the network. This distribution is known in the physics literature as the Fermi-Dirac distribution. It is the distribution that maximizes the randomness of the network under the constraints of having a fixed sequence of expected degrees, given angular coordinates for all nodes, and a maximum of one link between any pair of nodes.

Using the formalism developed in [2], we compute the average degree of a node with hidden variable κ (notice that since the angular distribution is homogeneous, this quantity does not depend on the angular coordinate of the node and, therefore, we chose one node located at $\theta = 0$) as

$$\bar{k}(\kappa) = N \int d\kappa \rho(\kappa') \frac{1}{2\pi} \int_{-\pi}^{\pi} d\theta p\left(\frac{|\theta|R}{\mu\kappa\kappa'}\right) = \frac{2\pi\langle k \rangle \mu\kappa}{\beta \sin\left[\frac{\pi}{\beta}\right]}. \tag{4}$$

By choosing parameter μ as

$$\mu = \frac{\beta}{2\pi\langle k \rangle} \sin\left[\frac{\pi}{\beta}\right]. \tag{5}$$

the expected degree of a node with hidden variable κ is simply $\bar{k}(\kappa) = \kappa$ and, therefore, the degree distribution scales as $P(k) \sim k^{-\gamma}$ for large k. Notice that this is the reason why in the main text we use degrees instead of expected degrees.

The $\mathbb{S}^1 \times \mathbb{S}^1$ model

The $\mathbb{S}^1$ model can be extended to bipartite networks as follows:

1. N_m metabolites and N_r reactions are homogeneously distributed on a circle of radius R. The density of metabolites and reactions over the circle are then $\delta_m = N_m/2\pi R$ and $\delta_r = N_r/2\pi R$. These two densities remain constant in the thermodynamic limit so that the radius of the circle is proportional to the number of metabolites or reactions.

2. Each metabolite is assigned a hidden variable κ_m and each reaction a hidden variable κ_r. These random variables follow probability densities $\rho_m(\kappa_m)$ and $\rho_r(\kappa_r)$, respectively.

3. The connection probability between a reaction with hidden variable κ_r and a metabolite with hidden variable κ_r separated by a distance $d_{mr} = R\Delta\theta_{mr}$ ($\Delta\theta_{mr}$ being the angular separation) is given by

$$p(\kappa_m, \theta_m; \kappa_r, \theta_r) = p\left(\frac{d_{mr}}{\mu\kappa_m\kappa_r}\right), \tag{6}$$

which can be any integrable function.

Analogously to the unipartite case, we compute the average degree of a metabolite with hidden variable κ_m as

$$\bar{k}_m(\kappa_m) = N_r \int d\kappa_r \rho_r(\kappa_r) \frac{1}{2\pi} \int_{-\pi}^{\pi} d\theta p\left(\frac{|\theta|R}{\mu\kappa_m\kappa_r}\right). \tag{7}$$

Similarly, the average degree of a reaction with hidden variable κ_r is

$$\bar{k}_r(\kappa_r) = N_m \int d\kappa_m \rho_r(\kappa_m) \frac{1}{2\pi} \int_{-\pi}^{\pi} d\theta p\left(\frac{|\theta|R}{\mu\kappa_m\kappa_r}\right). \tag{8}$$

By doing the change of variables $x = \frac{\theta R}{\mu\kappa_m\kappa_r}$ and taking the thermodynamic limit $R \to \infty$, we can write

$$\bar{k}_m(\kappa_m) = 2\mu\delta_r I\langle\kappa_r\rangle\kappa_m, \tag{9}$$

$$\bar{k}_r(\kappa_r) = 2\mu\delta_m I\langle\kappa_m\rangle\kappa_r, \tag{10}$$

where $I = \int_0^\infty dx p(x)$. By taking the average again

$$\langle k_m\rangle = 2\mu\delta_r I\langle\kappa_r\rangle\langle\kappa_m\rangle, \tag{11}$$

$$\langle k_r\rangle = 2\mu\delta_m I\langle\kappa_m\rangle\langle\kappa_r\rangle. \tag{12}$$

We immediately see that the following relation holds

$$\frac{\langle k_m\rangle}{\langle k_r\rangle} = \frac{\delta_r}{\delta_m} = \frac{N_r}{N_m}. \tag{13}$$

In terms of the average degrees, parameter μ takes the form

$$\mu = \frac{\langle k_m\rangle}{2\delta_r I\langle\kappa_r\rangle\langle\kappa_m\rangle} = \frac{\langle k_r\rangle}{2\delta_m I\langle\kappa_r\rangle\langle\kappa_m\rangle} \tag{14}$$

and, therefore, Eqs. (9) and (10) can be rewritten as

$$\bar{k}_m(\kappa_m) = \frac{\langle k_m\rangle}{\langle\kappa_m\rangle}\kappa_m \tag{15}$$

$$\bar{k}_r(\kappa_r) = \frac{\langle k_r\rangle}{\langle\kappa_r\rangle}\kappa_r \tag{16}$$

We always have the freedom to chose the averages of the hidden variables κ_m and κ_r to coincide with the actual averages of the observable variables k_m and k_r, that is, $\langle k_m\rangle = \langle\kappa_m\rangle$ and $\langle k_r\rangle = \langle\kappa_r\rangle$. In this case we can write

$$\bar{k}_m(\kappa_m) = \kappa_m \qquad \text{and} \qquad \bar{k}_r(\kappa_r) = \kappa_r \tag{17}$$

with parameter μ

$$\mu = \frac{1}{2\delta_r I\langle\kappa_r\rangle} = \frac{1}{2\delta_m I\langle\kappa_m\rangle}. \tag{18}$$

This is the choice that we shall follow in the rest of the text. The degree distributions can now be easily written as

$$P_m(k_m) = \int d\kappa_m \rho_m(\kappa_m) \frac{1}{k_m!}\kappa_m^{k_m} e^{-\kappa_m} \tag{19}$$

$$P_r(k_r) = \int d\kappa_r \rho_r(\kappa_r) \frac{1}{k_r!}\kappa_r^{k_r} e^{-\kappa_r} \tag{20}$$

Specific model for metabolic networks

In the case of metabolic networks, the distribution of metabolites' degrees is a power law with exponent $\gamma \approx 2.6$ and the distribution of reactions' degrees is Poisson-like. We can generate this type of network by chosing

$$\rho_m(\kappa_m) = (\gamma-1)\kappa_{m,0}^{\gamma-1}\kappa_m^{-\gamma} \text{ with } \kappa_m \geq \kappa_{m,0} = \frac{\gamma-2}{\gamma-1}\langle\kappa_m\rangle \tag{21}$$

and

$$\rho_r(\kappa_r) = \delta(\kappa_r - \langle\kappa_r\rangle). \tag{22}$$

Reaction degrees are then Poisson distributed, that is,

$$P_r(k_r) = \frac{1}{k_r!}\langle\kappa_r\rangle^{k_r}e^{-\langle\kappa_r\rangle} \tag{23}$$

whereas the degree distribution of metabolites is

$$P_m(k_m) = (\gamma-1)\kappa_{m,0}^{\gamma-1}\frac{\Gamma(k_m+1-\gamma,\kappa_{m,0})}{k_m!} \tag{24}$$

We also chose the connection probability

$$p(x) = \frac{1}{1+x^\beta} \tag{25}$$

so that the integral $I = \pi/(\beta\sin(\pi/\beta))$. We can also chose $\delta_m = 1$ without loss of generality. Therefore, the number of relevant (free) parameters of the model are $\langle\kappa_r\rangle$, $\langle\kappa_m\rangle$, β, and γ.

Parameters estimation and finite size effects

All results in the previous section are strictly true in the thermodynamic limit. In finite size networks, some of the expressions have to be corrected by size dependent factors as we will show below. Besides, there is an extra complication due to the fact that this model can generate nodes with zero degree, which are never observed in a real network.

Suppose we are given a real network with N_m^{obs} metabolites and N_r^{obs} reactions and average degrees $\langle k_m\rangle^{obs}$ and $\langle k_r\rangle^{obs}$ with exponent γ. We now want to estimate the values of $\langle\kappa_r\rangle$, $\langle\kappa_m\rangle$, N_m and N_r in our model. The first complication arises because in our model, out of the N_m nodes, there is a fraction $P_m(0)N_m$ nodes with zero degree that cannot be observed. Therefore, if we observe N_m^{obs} metabolites, the best estimation of N_m is

$$N_m = \frac{N_m^{obs}}{1-P_m(0)} \tag{26}$$

and, analogously

$$N_r = \frac{N_r^{obs}}{1-P_r(0)} \tag{27}$$

The second complication is due to the fact that the average degree of a power law distribution strongly depends on the maximum degree observed in the sample. For instance, in the case of our $\rho_m(\kappa_m) = (\gamma-1)\kappa_{m,0}^{\gamma-1}\kappa_m^{-\gamma}$, if the sample is finite, the distribution is truncated at a certain value $\kappa_{m,c}$ that, typically, increases with the size of the sample. If we compute the average of $\rho_m(\kappa_m)$ but only up to the maximum κ_m observed, we have

$$\langle\kappa_m(\kappa_{m,c})\rangle = (\gamma-1)\kappa_{m,0}^{\gamma-1}\int_{\kappa_{m,0}}^{\kappa_{m,c}}\kappa_m^{1-\gamma}d\kappa_m \tag{28}$$

and so

$$\langle\kappa_m(\kappa_{m,c})\rangle = \langle\kappa_m\rangle\left(1-\left(\frac{\kappa_{m,0}}{\kappa_{m,c}}\right)^{\gamma-2}\right) \tag{29}$$

Notice that this large parenthesis converges to 1 in the thermodynamic limit but for $\gamma \approx 2$ it can be fairly large even for large systems. Let us call this factor $\alpha(\kappa_{m,c})$, that is,

$$\alpha(\kappa_{m,c}) \equiv \left(1 - \left(\frac{\kappa_{m,0}}{\kappa_{m,c}}\right)^{\gamma-2}\right) \tag{30}$$

Now we need to keep track of the finite size effects from the very beginning. This means that we have to correct Eqs. (9) and (10) as follows

$$\bar{k}_m(\kappa_m; \kappa_{mc}) = \kappa_m \tag{31}$$

$$\bar{k}_r(\kappa_r; \kappa_{mc}) = \alpha(\kappa_{m,c})\kappa_r \tag{32}$$

and taking averages

$$\langle k_m(\kappa_{mc})\rangle = \alpha(\kappa_{m,c})\langle \kappa_m\rangle \tag{33}$$

$$\langle k_r(\kappa_{mc})\rangle = \alpha(\kappa_{m,c})\langle \kappa_r\rangle \tag{34}$$

Notice that, to write these set of equations we have used that variable κ_r is not power law distributed.

Still, this average $\langle k_m(\kappa_{m,c})\rangle$ cannot be directly identified with the measured average degree because it also accounts for nodes of zero degree. To correct for this effect, we write

$$\langle k_m\rangle^{obs} = \frac{\langle k_m(\kappa_{m,c})\rangle}{1 - P_m(0)} \tag{35}$$

and so

$$\langle \kappa_m\rangle = \frac{1 - P_m(0)}{\alpha(\kappa_{m,c})}\langle k_m\rangle^{obs} \tag{36}$$

and analogously

$$\langle \kappa_r\rangle = \frac{1 - P_r(0)}{\alpha(\kappa_{m,c})}\langle k_r\rangle^{obs} \tag{37}$$

with

$$P_m(0) = (\gamma - 1)\kappa_{m,0}^{\gamma-1}\Gamma(1 - \gamma, \kappa_{m,0}) \tag{38}$$

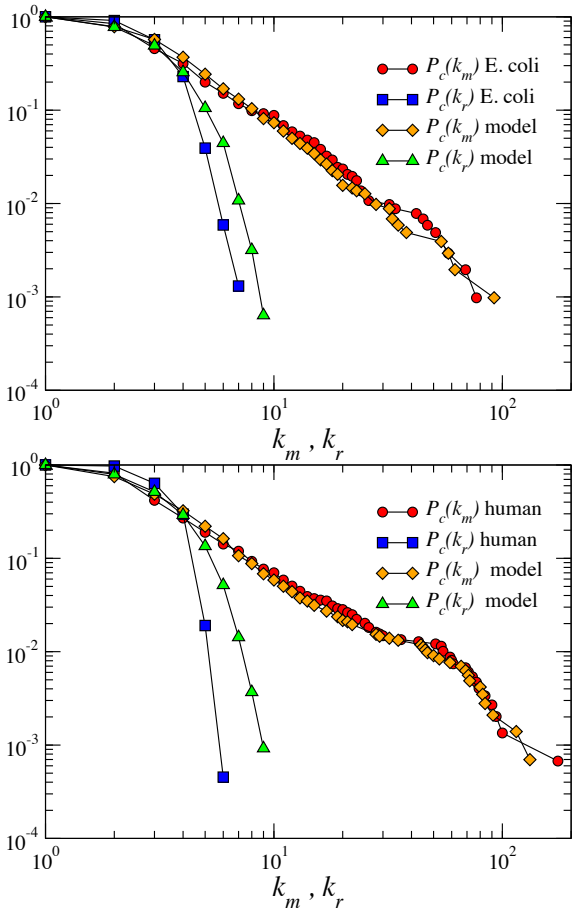
$$P_r(0) = e^{-\alpha(\kappa_{m,c})\langle \kappa_r\rangle} \tag{39}$$

$$\kappa_{m,c} = k_m^{max,obs} \tag{40}$$

Plugging Eqs. (21), (30), (38), and (40) into Eq. (36), we obtain a closed equation for $\langle \kappa_m\rangle$ that can be solved numerically. Once this parameter is known, by inserting it into Eqs. (30) and (38) we obtain the values of $\alpha(\kappa_{m,c})$ and $P_m(0)$. Finally, with the value of $\alpha(\kappa_{m,c})$ and Eqs. (37) and (39) we get the values of $\langle \kappa_r\rangle$ and $P_r(0)$.

Parameters of the real metabolisms

Using information from the BiGG database [3, 4], we build bipartite metabolic network representations of the two analyzed metabolisms, E. coli and human, avoiding reactions that do not involve direct chemical transformations, such as diffusion and exchange reactions. The bipartite representation differentiates two subsets of nodes, metabolites and reactions, mutually interconnected through unweighted and undirected links, without self-loops or dead end reactions. In particular, we analyze the *i*AF1260 version of the K12 MG1655 strain of the metabolism of E. coli [5], and the existing annotated list for human metabolism [6]. For the sake of simplicity and to enhance the resolution of the



Supplementary Figure S1: Empirical vs. model degree distributions Complementary cumulative degree distribution (defined as $P_c(k) = \sum_{k'=k} P(k)$) of metabolites and reactions degrees for the E. coli and human metabolism as compared to two networks generated with the model using the parameters in the text.

applied algorithm, currency metabolites are eliminated, altogether with a few isolated reaction-metabolite pairs and reaction-metabolite-reaction triplets. For E. coli, this leads to a final set of 1512 reactions and 1010 metabolites while human metabolism is nearly 3/2 larger, with 2201 reactions and 1482 metabolites. Characteristic power-law degree distributions for metabolites are readily identified in both organisms, with exponents that are rather similar. Reactions, meanwhile, conform to Poisson-like distributions, whose average values are 2.77 and 2.93 respectively.

There is some controversy on whether metabolic networks are or are not well described by power laws [7–10]. To check this issue, we perform a goodness of fit to test the validity of the null model that the observed empirical metabolite degree distribution has been generated by a power law. We compute the Kolmogorov statistic

$$D = \max_{k \geq k_{min}} \left| P_c(k) - \frac{\sum_{k'=k} k'^{-\gamma}}{\sum_{k'=k_{min}} k'^{-\gamma}} \right|, \tag{41}$$

where k_{min} is the minimum degree beyond which we expect the power law to hold and $P_c(k)$ is the empirical complementary cumulative degree distribution of metabolites. The exponent γ and the minimum degree k_{min} are computed using maximum likelihood methods as described in [7], resulting in $k_{min} = 4$, $\gamma = 2.5(6)$, $D = 0.065$ for the *E. coli* metabolism and $k_{min} = 2$, $\gamma = 2.3(7)$, $D = 0.006$ for the human metabolism. According to the Kolmogorov Smirnov (KS) test, the variable $\sqrt{N}D$ follows the Kolmogorov distribution $P_K(K)$, which 95% confidence level is at $K_{95\%} = 1.35$. Given the size of our samples, we obtain $\sqrt{N}D = 1.17 < 1.35$ for *E. coli* and $\sqrt{N}D = 0.22 << 1.35$ for human. This implies that the null model cannot be ruled out and, consequently, that power laws are a plausible explanation of our data.

- To find the parameters of the E. coli metabolic network, we use a version of the network where different isomers are considered as different metabolites. Further, we remove the following currency metabolites: h-841, h2o-694,

atp-338, pi-308, adp-260, ppi-129, nad-115, nadh-109, amp-85, nadp-83, nadph-81. Ten isolated metabolite-reaction pairs and six isolated reaction-metabolite-reaction triplets have also been removed. For this network, we measure $N_m^{obs} = 1010$, $N_r^{obs} = 1512$, $\langle k_m \rangle^{obs} = 4.15$, and $\langle k_r \rangle^{obs} = 2.77$. Using the formalism described in the previous section, we obtain the following estimation of the parameters: $\langle \kappa_m \rangle = 4.06$, $\langle \kappa_r \rangle = 2.65$, $N_m = 1123$, and $N_r = 1720$, and $R = N_m/2\pi = 178.7$.

- In the case of the Human metabolism, the removed currency metabolites are: h-1250, h2o-916, atp-309, coa-277, pi-240, adp-237, o2-212, nadp-210, nadph-207, nad-202, nadh-195, ppi-114. Three isolated metabolite-reaction pairs have also been removed. We then measure $N_m^{obs} = 1482$, $N_r^{obs} = 2201$, $\langle k_m \rangle^{obs} = 4.34$, and $\langle k_r \rangle^{obs} = 2.93$, which leads to the following estimation of the parameters: $\langle \kappa_m \rangle = 4.22$, $\langle \kappa_r \rangle = 2.73$, $N_m = 1646$, and $N_r = 2326$, and $R = N_m/2\pi = 235.9$.

In Fig. S1, we show the degree distributions for both E. coli and human metabolisms and compare them with those corresponding to networks generated by the $\mathbb{S}^1 \times \mathbb{S}^1$ model. The exponent β takes the value $\beta = 1.3$ in both networks. The agreement between the model and the real metabolic networks is very good for metabolites. However, the model overestimates the probability of reactions involving five or more metabolites. We perform a KS test to determine whether the empirical distribution and the one generated by the model are different. In this case, the Kolmogorov statistic is

$$D = \max_{k \geq k_{min}} \left| P_c(k) - P_c^{model}(k) \right|, \tag{42}$$

where $P_c^{model}(k)$ is the cumulative degree distribution of metabolites generated by our synthetic model with the same parameters as the real one. The parameter to be compared to $K_{95\%}$ is now

$$\sqrt{\frac{N N^{model}}{N + N^{model}}} D. \tag{43}$$

This parameter takes the value 1.03 with $k_{min} = 4$ for the *E. coli* network and 0.95 with $k_{min} = 3$ for the human one. These results indeed justify that networks generated by our model are reproducing well the properties of the real networks. The small discrepancy between our model and the real system is focalized in metabolites of very low degree. However, our embedding method and the results shown in this paper are not affected by these low degree metabolites.

EMBEDDING ALGORITHM AND VALIDATION ON $\mathbb{S}^1 \times \mathbb{S}^1$ SYNTHETIC NETWORKS

Once the parameters $\langle \kappa_r \rangle$, $\langle \kappa_m \rangle$, β, and γ are estimated, we perform the embedding of the bipartite network to infer the angular coordinates of metabolites and reactions. Let $\mathbb{A} \equiv (a_{ij})_{N_m \times N_r}$, $i = 1, \cdots, N_m$, $j = 1, \cdots, N_r$, be the adjacency matrix of the network, defined as $a_{ij} = 1$ if metabolite i participate in reaction j and zero otherwise (in the rest of the text, symbol i is reserved to enumerate metabolites and symbol j to reactions). Our goal is to find the set of coordinates $\{\kappa_{m,i}, \theta_{m,i}, \theta_{r,j}\}$ that best match the $\mathbb{S}^1 \times \mathbb{S}^1$ model in a statistical sense. To this end, we use maximum likelihood estimation (MLE) techniques. Let us compute the posterior probability, or likelihood, that a network given by its adjacency matrix $\mathbb{A}$ is generated by the $\mathbb{S}^1 \times \mathbb{S}^1$ model, $\mathcal{L}(\mathbb{A})$. This probability is

$$\mathcal{L}(\mathbb{A}) = \int \cdots \int \mathcal{L}(\mathbb{A}, \{\kappa_{m,i}, \theta_{m,i}, \theta_{r,j}\}) \prod_{i=1}^{N_m} d\theta_{m,i} d\kappa_{m,i} \prod_{j=1}^{N_r} d\theta_{r,j}, \tag{44}$$

where function $\mathcal{L}(\mathbb{A}, \{\kappa_{m,i}, \theta_{m,i}, \theta_{r,j}\})$ within the integral is the joint probability that the model generates the adjacency matrix $\mathbb{A}$ and the set of hidden variables $\{\kappa_{m,i}, \theta_{m,i}, \theta_{r,j}\}$ simultaneously. Using Bayes' rule, we can compute the likelihood that nodes' coordinates take particular values $\{\kappa_{m,i}, \theta_{m,i}, \theta_{r,j}\}$ given the observed adjacency matrix $\mathbb{A}$. This probability is simply given by

$$\mathcal{L}(\{\kappa_{m,i}, \theta_{m,i}, \theta_{r,j}\}|\mathbb{A}) = \frac{\mathcal{L}(\mathbb{A}, \{\kappa_{m,i}, \theta_{m,i}, \theta_{r,j}\})}{\mathcal{L}(\mathbb{A})} = \frac{\text{Prob}(\{\kappa_{m,i}, \theta_{m,i}, \theta_{r,j}\}) \mathcal{L}(\mathbb{A}|\{\kappa_{m,i}, \theta_{m,i}, \theta_{r,j}\})}{\mathcal{L}(\mathbb{A})}, \tag{45}$$

where

$$\text{Prob}(\{\kappa_{m,i},\theta_{m,i},\theta_{r,j}\}) = \frac{1}{(2\pi)^{N_m+N_r}}\prod_{i=1}^{N_m}\rho_m(\kappa_{m,i}) \tag{46}$$

is the prior probability of the hidden variables given by the model,

$$\mathcal{L}(\mathbb{A}|\{\kappa_{m,i},\theta_{m,i},\theta_{r,j}\}) = \prod_{i=1}^{N_m}\prod_{j=1}^{N_r} p(x_{ij})^{a_{ij}}[1-p(x_{ij})]^{1-a_{ij}} \tag{47}$$

is the likelihood of observing $\mathbb{A}$ if the hidden variables are $\{\kappa_{m,i},\theta_{m,i},\theta_{r,j}\}$,

$$x_{ij} = \frac{N_r\Delta\theta_{ij}}{\beta\sin(\pi/\beta)\kappa_{m,i}}, \tag{48}$$

$$\Delta\theta_{ij} = \pi - |\pi - |\theta_{m,i}-\theta_{r,j}||, \tag{49}$$

and $p(x)$ is given by Eq. (25).

The MLE values of the hidden variables $\{\kappa^*_{m,i},\theta^*_{m,i},\theta^*_{r,j}\}$ are then those that maximize the likelihood in Eq. (45) or, equivalently, its logarithm,

$$\ln\mathcal{L}(\{\kappa_{m,i},\theta_{m,i},\theta_{r,j}\}|\mathbb{A}) = C - \gamma\sum_{i=1}^{N_m}\ln\kappa_{m,i} + \sum_{i=1}^{N_m}\sum_{j=1}^{N_r}\{a_{ij}\ln p(x_{ij}) + (1-a_{ij})\ln[1-p(x_{ij})]\}, \tag{50}$$

where C is independent of the nodes' coordinates $\{\kappa_{m,i},\theta_{m,i},\theta_{r,j}\}$.

MLE for expected metabolites' degrees κ_m

The derivative of Eq. (50) with respect to expected degree $\kappa_{m,l}$ of metabolite l is

$$\frac{\partial}{\partial\kappa_{m,l}}\ln\mathcal{L}(\{\kappa_{m,i},\theta_{m,i},\theta_{r,j}\}|\mathbb{A}) = -\frac{\gamma}{\kappa_{m,l}} - \frac{\beta}{\kappa_{m,l}}\left(\sum_{j=1}^{N_r}p(x_{lj}) - \sum_{j=1}^{N_r}a_{lj}\right). \tag{51}$$

The first term within the parenthesis is the expected degree of metabolite l, while the second term is its actual degree $k_{m,l}$. Therefore, the value $\kappa^*_{m,l}$ that maximizes the likelihood is given by

$$\bar{k}(\kappa^*_{m,l}) = \kappa^*_{m,l} = k_{m,l} - \frac{\gamma}{\beta}. \tag{52}$$

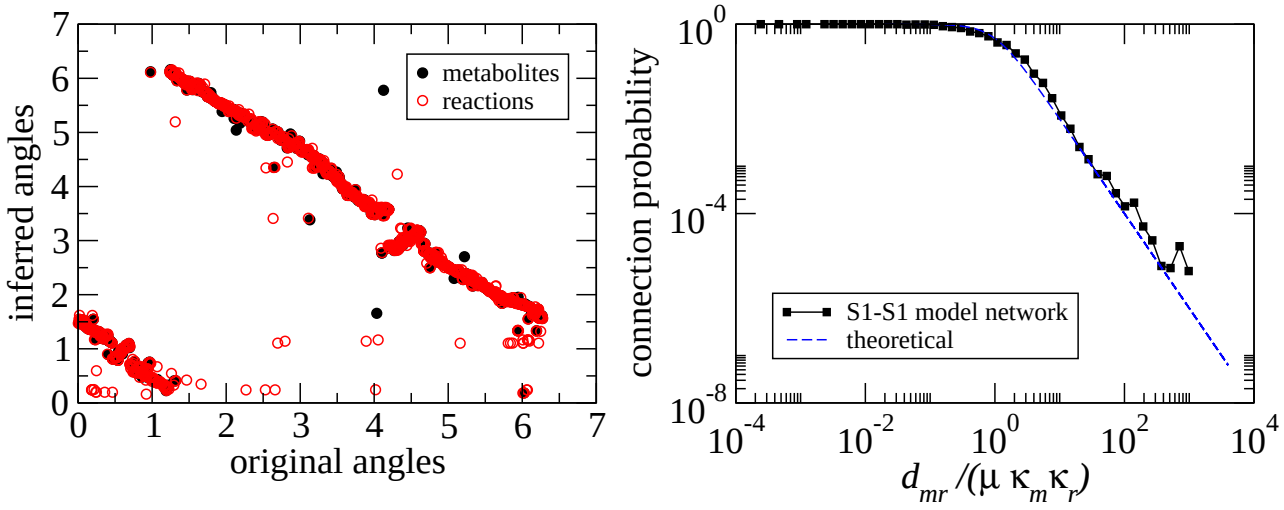
Since κ^*_l can be smaller than κ_0 in the last equation, we set

$$\kappa^*_{m,l} = \max\left(\frac{\gamma-2}{\gamma-1}\langle\kappa_m\rangle, k_{m,l} - \frac{\gamma}{\beta}\right). \tag{53}$$

MLE for angular coordinates θ

Having found the MLE values for expected degrees κ_m, we now have to maximize Eq. (45) with respect to angular coordinates. This task is equivalent to maximizing the partial log-likelihood

$$\ln\mathcal{L}(\mathbb{A}|\{\kappa^*_{m,i},\theta_{m,i},\theta_{r,j}\}) = \sum_{i=1}^{N_m}\sum_{j=1}^{N_r}\{a_{ij}\ln p(x_{ij}) + (1-a_{ij})\ln[1-p(x_{ij})]\}. \tag{54}$$



Supplementary Figure S2: Calibration of the embedding algorithm. The left plot shows the inferred angular coordinates of metabolites and reactions vs. the real ones of a network generated with the $\mathbb{S}^1 \times \mathbb{S}^1$ with the same parameters as the real metabolism. The right plot shows the empirical connection probability obtained from the embedding compared to the theoretical one in Eq. (25)

The maximization of Eq. (54) with respect to the angular coordinates cannot be performed analytically and we have to rely on numerical optimization procedures. Unfortunately, the low degrees of reactions implies that any attempt to maximize Eq. (54) directly is doomed to fail. Indeed, the uncertainty in the position of a low degree reaction is necessary very high. This, in turn, increases the uncertainty in the position of its metabolites' neighbors, which translates into global uncertainty in the localization of nodes and metabolites. We therefore adopt a different strategy. Starting from the original bipartite network, we construct its one mode projection over the space of metabolites, that is, we consider only one type of nodes (metabolites) and declare two metabolites as connected if they participate in the same reaction in the original bipartite net. If metabolites are power-law distributed in the bipartite network, the obtained unipartite network is also power-law distributed with the same exponent. This solves the problem mentioned above because, now, high degree nodes can be located with high accuracy so that we can use afterwards these nodes as a template to find the coordinates of the rest of the nodes.

We find the angular coordinates of metabolites by fitting the one-mode projected network using the $\mathbb{S}^1$ model as described in [11]. Once the angular coordinates $\theta^*_{m,i}$ are known, we find the optimal angular coordinates of reactions by maximizing Eq. (54) but using the already known coordinates of metabolites as fixed inputs. This final maximization is a simple procedure because, being $\theta^*_{m,i}$ fixed, we can maximize the likelihood of each reaction independently.

We first test the described procedure in synthetic networks generated by the $\mathbb{S}^1 \times \mathbb{S}^1$ model with the same parameters as the real E. coli metabolism. Results are shown in Fig. S2. The left plot shows the inferred angles for metabolites and reactions vs. the real ones. As it can be clearly seen, up to minor fluctuations and a global phase shift due to rotational symmetry of the model, the agreement between the real coordinates and those inferred by the algorithm is very good. The right plot shows the connection probability using the inferred coordinates vs. the one used to generate the model Eq. (25). Again, the agreement between the two is excellent.

CLASSIFICATION OF PATHWAYS IN E. COLI DEPENDING ON LOCALIZATION

Supplementary Table SI: Classification of E. coli's pathways. Pathways are classified as "localized" (75% of the pathway localized in a single bin), "bimodal" (75% of the pathway localized in two bins) "multi-peaked" (75% of the pathway localized in three bins or more with at least one peak above 25%), and "transversal" (no bin above 25%) according to the results and bin size of Fig. 3. Pathways in italics indicate that, although they are split in two or three bins, these bins are adjacent and so a change in the bin resolution would lead to their redefinition as more localized pathways.

LOCALIZED	BIMODAL	MULTI-PEAKED	TRANSVERSAL
Glu	His	Ala, Asp	Cofactor and Prosthetic
Folate	Met	Arg, Pro	*Purine and Pirimidine*
Methylglyoxal	Thr, Lys	Cys	Alternate Carbon
Oxidative Phosphorilation	*Anaplerotic*	Gly, Ser	Transport Inner Membrane
Murein B	*Citric Acid Cycle*	*Tyr, Phe, Trp*	
Murein R	Glyoxylate	Val, Leu, Ile	
	Pentose Phosphate	*Nucleotides S*	
	Inorganic Ion Transport	tRNA Charging	
	Membrane Lipid	Glycolisis	
		Pyruvate	
		Nitrogen	
		Lipopolysaccharide	
		Cell Envelope	
		Glycerophospholipid	

DETERMINATION OF ANGULAR SECTORS

Angular sectors shown in Fig. 4 of the main text are computed as follows. Let us assume that all reactions are randomly and homogeneously distributed on the circle. In this case, the probability density of the distance between two consecutive reactions is the exponential distribution

$$\phi(l) = \frac{1}{\bar{l}} e^{l/\bar{l}} \tag{55}$$

where $\bar{l}$ is the average distance between two consecutive reactions. In our case, since concentration of reactions in the circle is fixed to 1, we have $\bar{l} = 1$. The probability to find a gap larger than l is therefore

$$\Phi(l) = \int_l^\infty e^{-l'} dl' = e^{-l}. \tag{56}$$

With this probability, we fix a significance level of 0.1%, leading to a critical distance value $l_{crit} = \ln 1000$ (in angular terms this is equivalent to $\theta_{crit} = 2\pi \ln 1000/1500 \approx 0.03$Rad), that is, the probability to find a gap larger than l_{crit} under the null assumption that reactions are homogeneously distributed is 0.1%. Then, we can consider a gap as significant whenever it is larger than l_{crit}. The significance level 0.1% is chosen such that with the size of our sample (around 1500 reactions) we should expect at most one gap above the critical value. In this case, we can define a sector as the set of points separated by two such significant gaps with the condition that this set has 5 or more reactions.

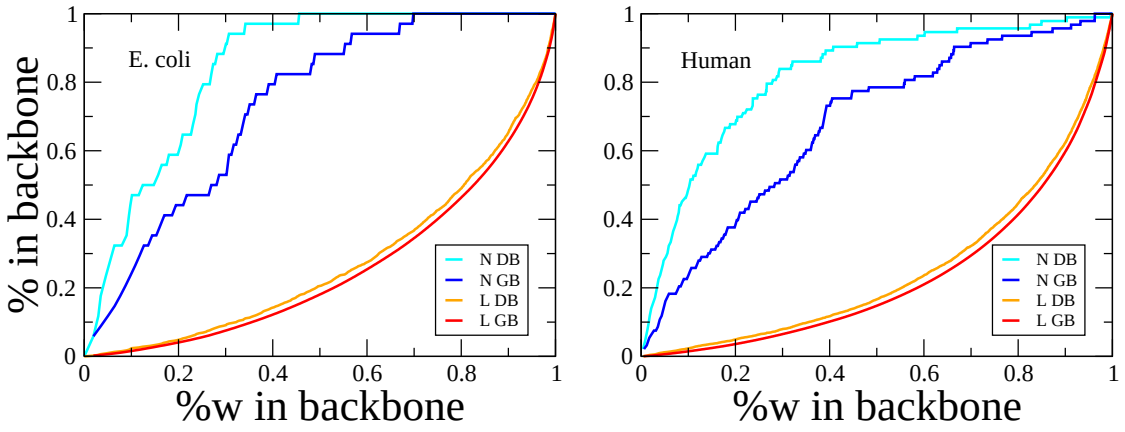
PATHWAYS CROSSTALK AND THE DISPARITY FILTER

We use the following measure of crosstalk between pathways:

$$XT_{PaPb} = \sum_{j \in P_a} \sum_{j' \in P_b} \sum_{i \in \nu} (p(x_{ij}) + p(x_{ij'}))|_{\text{observed links}}, \tag{57}$$

where $\nu \in \mathcal{M}_{ab}$ is the set of metabolites shared by the reactions in the two pathways P_a and P_b, and only probabilities of connections associated to observed links are considered.

Of 561 possible pathway pairs in E. coli, 460 are non-zero crosstalk (82.00%) with a minimum value of 1.80 and a maximum of 159.91. In human cells, of 4278 possible pathway pairs, 1689 are non zero (38.64%) with a minimum crosstalk of 1.19 and a maximum of 131.28. Moreover, there is an isolated pathway (48, Limonene and Pinene Biosynthesys) without crosstalk (no common metabolites with other pathways). So, at this level human cells metabolism seems to be more modular than E. coli's.



Supplementary Figure S3: Disparity backbone vs global threshold backbone.

The obtained pathway crosstalk matrices are filtered to obtain backbones according to the multiscale methodology in [12], which do not belittle small pathways and gives an effective tradeoff between maximum weight and nodes in the backbone with the minimum number of links. A global threshold filter would lose many more nodes for the same number of links and weight in the backbone, see Fig. S3.

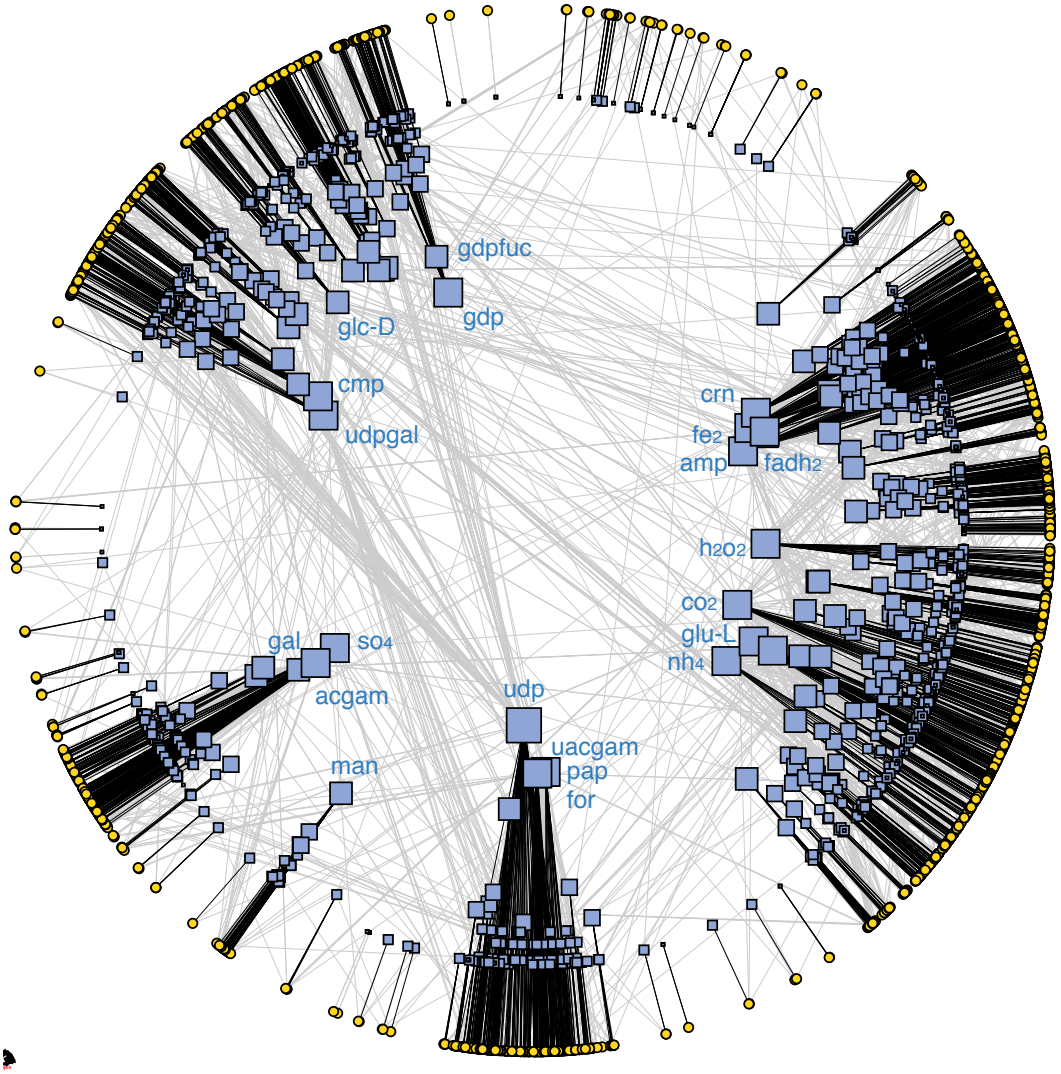
The disparity filter methodology preserves interactions with a statistically significant intensity for at least one of the two nodes the edge is incident to. To decide whether a connection is relevant, the filter compares against a null hypothesis which assumes that the local weights associated to a node are uniformly distributed at random. In this way one discounts intensities that could be explained by random fluctuations. More specifically, a p value –the probability α_{ij} that if the null hypothesis is true one obtains a value for the normalized weight w_{ij}/s_i between nodes i and j larger than or equal to the observed one– is calculated for each edge in the network. By imposing a significance level α, the links that carry weights that can be considered not compatible with a random distribution can be filtered out with a certain statistical significance. Links in the backbone will be then those which satisfy

$$\alpha_{ij} = 1 - (k-1)\int_0^{w_{ij}/s_i} (1-x)^{k-2} dx < \alpha, \tag{58}$$

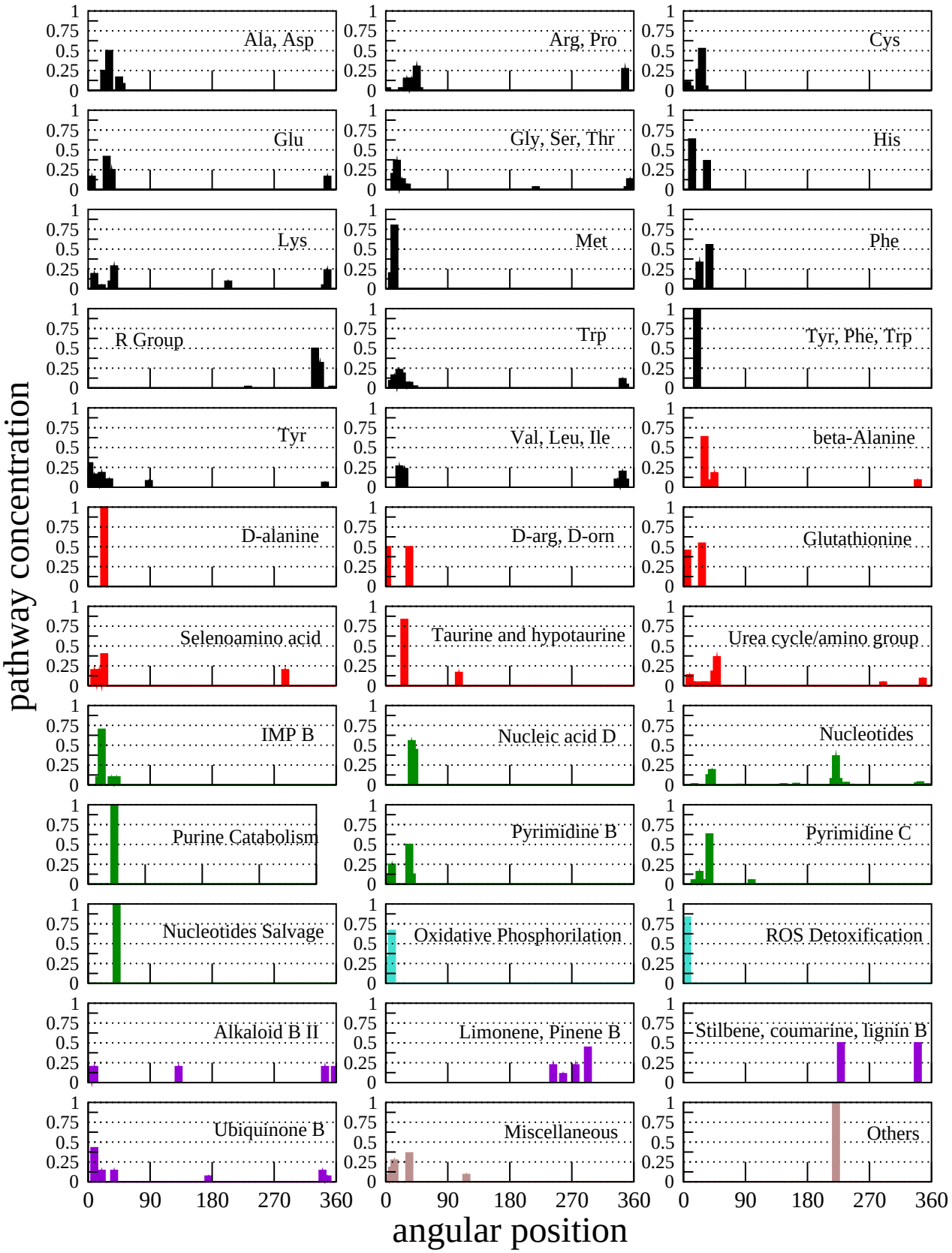
where k is the degree of node i. By changing the significance level, we can filter out the links progressively focusing on more relevant ones. As a result, the disparity filter reduces significantly the number of edges in the original network, while keeping almost a large fraction of the total weight and the total number of nodes. It preserves as well the cutoff of the degree distribution, the form of the weight distribution, and the clustering coefficient.

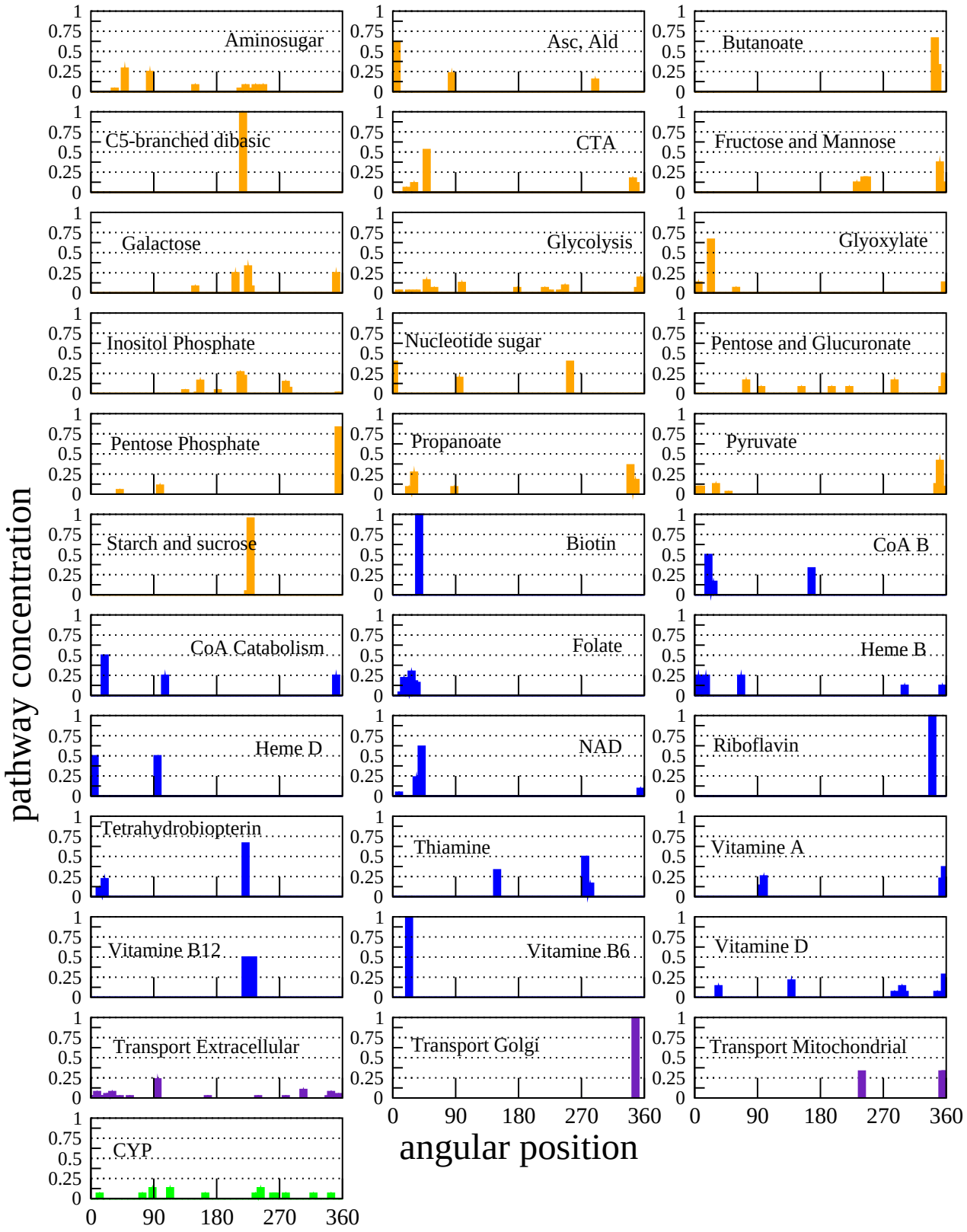
RESULTS FOR HUMAN CELLS METABOLISM

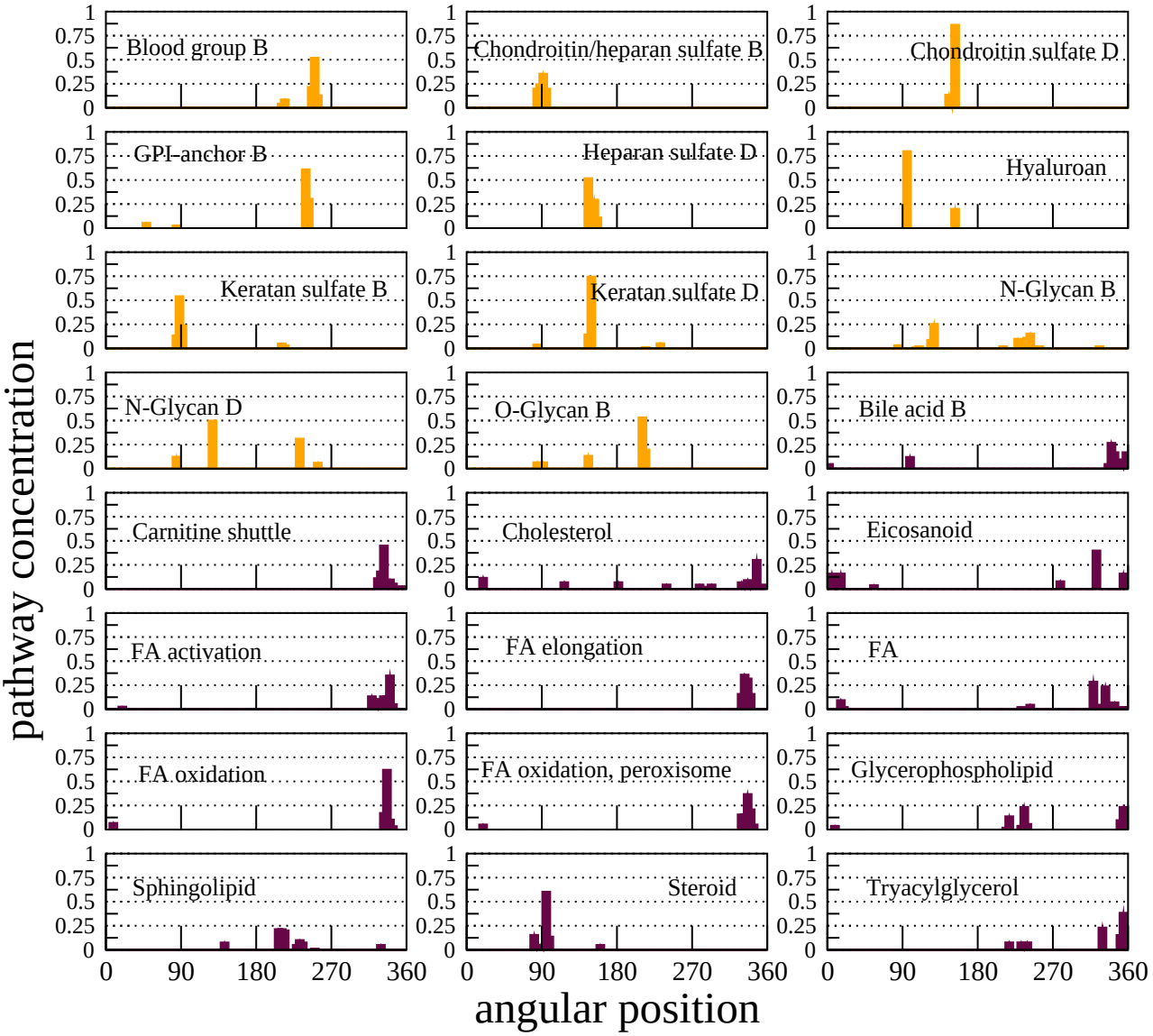
In Fig. S4, we show the embedding representation of human cells metabolism. In Fig. S5, we show the angular distribution on the ring of the whole list of pathways evaluated from the circle-based embedding of the reactions they involve.



Supplementary Figure S 4: Human metabolism map. Yellow circles represent reactions whereas blue squares are metabolites. For each metabolite, the symbol size is proportional to the logarithm of the degree and radially placed according to the expression $r = R - 2 \ln k_m$. Black (grey) connections are those that according to the model have a probability of existence larger (smaller) than 0.5. We used the software "Pajek" to elaborate all network representations in this paper figures'.


pathway concentration
angular position
Ala, Asp
Arg, Pro
Cys
Glu
Gly, Ser, Thr
His
Lys
Met
Phe
R Group
Trp
Tyr, Phe, Trp
Tyr
Val, Leu, Ile
beta-Alanine
D-alanine
D-arg, D-orn
Glutathionine
Selenoamino acid
Taurine and hypotaurine
Urea cycle/amino group
IMP B
Nucleic acid D
Nucleotides
Purine Catabolism
Pyrimidine B
Pyrimidine C
Nucleotides Salvage
Oxidative Phosphorilation
ROS Detoxification
Alkaloid B II
Limonene, Pinene B
Stilbene, coumarine, lignin B
Ubiquinone B
Miscellaneous
Others
0
0.25
0.5
0.75
1
0
90
180
270
360





Supplementary Figure S 5: **Angular distribution of pathways for the human metabolism.** The whole angular domain $[0, 360^o]$ is divided in 50 bins of $7, 2^o$ each and for each bin we compute the fraction of reactions of the pathway in it. Each pathway is shown in a different graph. Different colors indicate different metabolic families. Panel I: black for Amino Acids metabolism (numbering the graphs from left to right and from top to bottom, 1-14), red for metabolism of Other Amino Acids (15-21), dark green for Nucleotide metabolism (22-28), turquoise for Energy metabolism (29,30), purple for biosynthesis of Other Secondary Metabolites (31-34), brown for miscellaneous and others (35,36). Panel II: orange for Carbohydrate metabolism (1-16), blue for metabolism of Cofactors and Vitamins (17-30), violet for Transport pathways (31-33), light green for Xenobiotics Biodegradation (34). Panel III: orange for Glycan metabolism (1-11), and dark brown for Lipid metabolism (12-24). Pathway names have been abbreviated in standard forms whenever possible, see supplementary information in excel file.

[1] M. A. Serrano, D. Krioukov, and M. Boguñá, Phys. Rev. Lett. **100**, 078701 (2008).
[2] M. Boguñá and R. Pastor-Satorras, Phys. Rev. E **68**, 036112 (2003).
[3] J. Schellenberger, J. O. Park, T. C. Conrad, and B. O. Palsson, BMC Bioinformatics **11**, 213 (2010).
[4] *BiGG database*, http://bigg.ucsd.edu/.
[5] A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. O. Palsson, Molecular Systems Biology **3**, 121 (2007).
[6] N. C. Duarte, S. A. Becker, N. Jamshidi, I. Thiele, M. L. Mo, T. D. Vo, R. Srivas, and B. O. Palsson, Proc. Natl. Acad. Sci. USA **104**, 1777 (2007).
[7] A. Clauset, R. Shalizi, and M. Newman, SIAM Review **51**, 661 (2009).
[8] R. Khanin and E. Wit, Journal of Computational Biology **13**, 810 (2006).
[9] G. Lima-Mendez and J. van Helden, Mol. BioSyst. **5**, 1482 (2009).
[10] M. Arita, Proceedings of the National Academy of Sciences of the United States of America **101**, 1543 (2004).
[11] M. Boguñá, F. Papadopoulos, and D. Krioukov, Nat Commun **1** (2010).
[12] M. A. Serrano, M. Boguñá, and A. Vespignani, Proc. Natl. Acad. Sci. USA **106**, 6483 (2009).
[13] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, Science **297**, 1551 (2002).
[14] This type of networks are often called hierarchical in the literature [13]