Text S2: Analytical solutions for the individual terms of the posterior distribution of the latent variables.

We assume that each microbiome sample is a mixture of $K$ metabosystems, with different mixing probabilities for different samples. The relative contribution of each metabosystem to the microbiome associated with the $n^{th}$ sample is modeled through latent variable $\theta_n$, a probability vector of $K$ values summing to one.

Each metabosystem is comprised of a fixed number ($L$) of metabolic subnetworks. The contribution of each subnetwork to the $k^{th}$ metabosystem, $\varphi_{k,i}$ is modeled by a vector of $L$ mixing probabilities that sum to one. The element $\varphi_{k,i}$ in a $K \times L$ matrix $\varphi$ represents the relative contribution of subnetwork $i$ in metabosystem $k$.

Each subnetwork is comprised of a mixture of reactions, which are linked through shared chemical compounds. Thus, each subnetwork has its own substrate ($S$) group and product ($R$) group. Note that compound participating in intermediary reactions can belong to both the substrate and the product group. For subnetwork $l$, we denote the substrate group as $\delta_l$ and the product group as $\gamma_l$, each a vector of $C$ probability values summing to one, where $C$ is the number of compounds. With $L$ subnetworks, there are two $L \times C$ matrices $\delta$ and $\gamma$, one for substrates and one for products respectively. The value in row $l$ and column $c$ of matrix $\delta$ represents the relative contribution of compound $c$ in the substrate group of subnetwork $l$. A similar definition applies for product groups in matrix $\gamma$.

We assume an independent and identical (iid) sparse symmetric Dirichlet prior on $\theta_n$, and the rows of $\varphi$, $\delta$ and $\gamma$:

$$\theta_n \sim \text{Dirichlet}(\alpha_\theta)$$
$$\varphi_{k,i} \sim \text{Dirichlet}(\alpha_\varphi)$$
$$\delta_i \sim \text{Dirichlet}(\alpha_\delta)$$
$$\gamma_i \sim \text{Dirichlet}(\alpha_\gamma)$$

The complete likelihood of the data given the hyper-parameters of the prior distributions is:
where \( N \) is the total microbiome samples, \( I_n \) is the total number of reactions in the metabolic network associated the \( n^{th} \) microbiome sample and \( J_{ni} \) is the total number of substrate-product pairs in the \( i^{th} \) reaction of the \( n^{th} \) sample. \( Z \) and \( Y \) represent the metabosystem and subnetwork assignments for all reactions in all samples, and as \( Z_{ni} \) and \( Y_{ni} \) represent the metabosystem and subnetwork assignments of reaction \( i \) within sample \( n \).

To infer the model framework, we need to sample from the posterior distribution of latent variables given the data:

\[
P(Z, Y, R, S, \theta, \varphi, \delta, \gamma \mid \alpha_{\theta}, \alpha_{\varphi}, \alpha_{\delta}, \alpha_{\gamma}) =
\prod_{n=1}^{N} \prod_{i=1}^{I_n} P(Z_{ni} \mid \theta_n) P(Y_{ni} \mid \theta_{Z_{ni}}) \prod_{n=1}^{N} \prod_{i=1}^{I_n} \prod_{j=1}^{J_{ni}} P(S_{nj} \mid \delta_{Y_{ni}}) P(R_{nij} \mid \gamma_{Y_{ni}}) \prod_{n=1}^{N} P(\theta_n \mid \alpha_{\theta}) \prod_{k=1}^{K} P(\varphi_k \mid \alpha_{\varphi}) \prod_{l=1}^{L} \left[ P(\delta_l \mid \alpha_{\delta}) P(\gamma_l \mid \alpha_{\gamma}) \right]
\]

where \( Z, Y, \theta, \varphi, \delta, \) and \( \gamma \) are latent variables in the model. We use collapsed Gibbs sampling to integrate out the other latent variables \( \theta, \varphi, \delta, \gamma \) and sample from the posterior distributions of the metabosystem \( (Z) \) and subnetwork \( (Y) \) assignments for each reaction conditional on the assignments of all other reactions. We use the following conditional probability to sample the subnetwork and metabosystem assignment of one reaction in a microbiome sample given we know the subnetwork and metabosystem assignments of all other reactions in every microbiome sample:
\[
P(Z_{ni}, Y_{ni} \mid S, R, Z_{-ni}, Y_{-ni}) = \\
\frac{\iint \int P(S, R, Z, Y, \theta, \varphi, \delta, \gamma) d\theta d\varphi d\delta d\gamma}{\iint \int P(S, R, Z_{-ni}, Y_{-ni}, \theta, \varphi, \delta, \gamma) d\theta d\varphi d\delta d\gamma} = \\
\frac{\int P(Z \mid \theta) P(\theta \mid \alpha_{\theta}) d\theta \times \int P(Y \mid \varphi) P(\varphi \mid \alpha_{\varphi}) d\varphi}{\int P(Z_{-ni} \mid \theta) P(\theta \mid \alpha_{\theta}) d\theta \times \int P(Y_{-ni} \mid \varphi) P(\varphi \mid \alpha_{\varphi}) d\varphi} \times \\
\frac{\int P(S \mid Y, \delta) P(\delta \mid \alpha_{\delta}) d\delta}{\int P(S \mid Y_{-ni}, \delta) P(\delta \mid \alpha_{\delta}) d\delta} \times \frac{\int P(R \mid Y, \gamma) P(\gamma \mid \alpha_{\gamma}) d\gamma}{\int P(R \mid Y_{-ni}, \gamma) P(\gamma \mid \alpha_{\gamma}) d\gamma}
\]

where \(Z_{ni}\) and \(Y_{ni}\) denote the metabosystem and subnetwork assignment for all reactions in all samples except only reaction \(i\) in sample \(n\). The analytical solutions for the individual terms of the above equation are:

\[
\int P(Z \mid \theta) P(\theta \mid \alpha_{\theta}) d\theta = \\
= \prod_{n=1}^{N} \prod_{i=1}^{I_n} P(Z_{ni} \mid \theta_n) \prod_{n=1}^{N} P(\theta_n \mid \alpha_{\theta})d\theta
\]

\[
= \prod_{n=1}^{N} \prod_{k=1}^{K} \theta_{nk}^{G_n^k} \prod_{n=1}^{N} \frac{\Gamma(K\alpha_{\theta})}{\Gamma(\alpha_{\theta})} \prod_{k=1}^{K} \theta_{nk}^{\alpha_{\theta}+G_n^k}d\theta
\]

\[
= \prod_{n=1}^{N} \frac{\Gamma(K\alpha_{\theta})}{\Gamma(\alpha_{\theta})} \prod_{k=1}^{K} \Gamma(\alpha_{\theta} + G_n^k) \Gamma\left(\sum_{k=1}^{K} (\alpha_{\theta} + G_n^k)\right)
\]

where \(G_n^k\) is the number of times the \(n\)th sample is assigned to the \(k\)th metabosystem.
\[
\int P(Y \mid Z, \varphi)P(\varphi \mid \alpha_\varphi) d\varphi = \\
= \int \prod_{n=1}^{N} \prod_{i=1}^{I_n} \prod_{l=1}^{L} \frac{\Gamma(L \alpha_\varphi)}{\Gamma(\alpha_\varphi)} \prod_{k=1}^{K} P(\alpha_k | \alpha_\varphi) d\alpha_k \\
= \prod_{k=1}^{K} \frac{\Gamma(L \alpha_\varphi)}{\Gamma(\alpha_\varphi)} \prod_{l=1}^{L} \left( \alpha_\varphi + G_k^l \right) \prod_{l=1}^{L} \left( \alpha_\varphi + G_k^l \right)
\]

where \( G_k^l \) is the number of times a reaction is assigned to the \( k \)th subnetwork for a sample in the \( k \)th metabosystem.

\[
\int P(S \mid \delta, Y)P(\delta \mid \alpha_\delta) d\delta = \\
= \int \prod_{n=1}^{N} \prod_{i=1}^{J_n} \prod_{j=1}^{J_n} \prod_{l=1}^{L} \frac{\Gamma(C \alpha_\delta)}{\Gamma(\alpha_\delta)} \prod_{c=1}^{C} \delta_c^{SG_c^l} d\delta \\
= \prod_{l=1}^{L} \frac{\Gamma(C \alpha_\delta)}{\Gamma(\alpha_\delta)} \prod_{c=1}^{C} \delta_c^{SG_c^l} \prod_{c=1}^{C} \delta_c^{SG_c^l} = \prod_{l=1}^{L} \frac{\Gamma(C \alpha_\delta)}{\Gamma(\alpha_\delta)} \prod_{c=1}^{C} \left( \alpha_\delta + SG_c^l \right) \prod_{c=1}^{C} \left( \alpha_\delta + SG_c^l \right)
\]

where \( SG_c^l \) is the number of times the \( c \)th compound is assigned to the substrate group of the \( k \)th subnetwork.
\[
\int P(R \mid \delta, Y) P(\gamma \mid \alpha_{\gamma}) d\gamma = \\
= \int \prod_{n=1}^{N} \prod_{i=1}^{I_n} \prod_{j=1}^{J_{ni}} P(R_{nij} \mid \gamma_{Yni}) \prod_{l=1}^{L} P(\gamma_l \mid \alpha_{\gamma}) d\gamma \\
= \int \prod_{l=1}^{L} \prod_{c=1}^{C} \frac{\delta_{l}}{\Gamma(a_{l})} \prod_{l=1}^{L} \frac{1}{\Gamma(a_{\gamma})} \prod_{c=1}^{C} \frac{C !}{\delta_{l}} \prod_{c=1}^{C} \frac{1}{\Gamma(a_{\gamma} + \delta_{l})} \\
= \prod_{l=1}^{L} \frac{\Gamma(a_{\gamma})}{\Gamma(a_{\gamma} + \delta_{l})} \prod_{c=1}^{C} \frac{\Gamma(a_{\gamma} + \delta_{l})}{\Gamma(\sum_{c=1}^{C} (a_{\gamma} + \delta_{l}))} \\
\]

where $RG_c^I$ is the number of times the $c^{th}$ compound is assigned to the product group of the $l^{th}$ subnetwork.