Mapping the bacterial ways of life

Ashkaan K Fahimipour & Thilo Gross

Department of Computer Science University of California, Davis, CA 95616, USA

The rise in the availability of bacterial genomes defines a need for synthesis: abstracting from individual taxa, to see the larger patterns of bacterial lifestyles across microbial systems. In community ecology, a central organising theory is the niche concept. A niche is a set of capabilities that enables a population’s persistence, and defines its impact on the environment. The set of possible niches is the niche space, a conceptual space delineating the ways in which persistence in an ecosystem is possible. Understanding the structure of the niche space is perhaps the central question in ecology. Here we use data analysis to map the space of metabolic networks describing thousands of bacterial genera. The results reveal a niche space with continuous branching geometry, whose branches correspond to adaptations to habitats, hosts, and unique resource use strategies. This provides a new perspective on the functional capabilities of known bacteria and lays an ecological foundation for the study of microbiomes. The variables defined here constitute a new way to classify and systematise bacterial populations in ecological terms. We regard this as an important step in the quest to bring methods and results from ecology to bear on microbial communities.

Bacteria perform key ecological functions in virtually all natural systems, engineered environments, and macrobiotic hosts, with rare exception. Although bacterial communities may collectively catalyse thousands of biochemical reactions, single populations are only capable of carrying out a small subset of these reactions at any time. We call the sets of reactions encoded by bacterial genomes metabolic strategies and say that a strategy is feasible if it is capable
of conveying population growth under some conditions. Together the feasible strategies span
the niche space: a space of all strategies that populations may follow to survive.

Mapping the niche space has so far been attempted for individual groups of macrobiotic or-
ganisms (e.g. lizards, fish), where the space is spanned by behavioural or morphological traits.
The studies used principle component analysis (PCA), a linear method, to identify variables rep-
resenting feasible strategies in trait space. Yet, the overwhelming degree of documented bacterial
functional diversity and our own results (below) suggests that metabolic strategies are un-
likely to form a linear space. A powerful nonlinear alternative to PCA is offered by diffusion
maps. This mathematically simple method finds natural variables in which to describe features
of a dataset. While the mathematical procedure does not provide an interpretation of these vari-
ables, our results show that they correspond to meaningful strategies.

Here we use diffusion maps to interpret the shape of the bacterial metabolic niche space. We
generated metabolic networks capturing intracellular metabolic capabilities, for a single represen-
tative genome from all unique bacterial genera in the National Center for Biotechnology In-
formation NCBI RefSeq release 92 (N = 2,634 representatives), analogous to the construction
of the microbial tree of life. Each representative genome was mapped to a point in a 7,438-
dimensional discrete space, where each axis indicates the presence or absence of a unique directed
edge (i.e. substrate-product pairs). This array represents a large fraction of the cytoplasmic bio-
chemical capabilities for most known bacterial genera, and served as input to the diffusion map
algorithm (see Methods).

The diffusion map finds the major (possibly nonlinear) dimensions that span a dataset and re-
turns then in the order of their importance. The most important variable identified by the diffusion
Figure 1: Data analysis reveals some discrete (yes/no) bacterial strategy choices. Variable entries for each genome are visualised as coloured tiles near the tips of a collapsed phylogenetic tree. Large positive or negative values (saturated reds and blues) indicate strong adaptation for the specific strategy (e.g. photosynthesis), whereas whites indicate close to no adaptation. Triangles show collapsed clades with near-zero entries for each of the four example variables. Clades receiving large positive or negative entries in any of the four variables shown are expanded and annotated. The lack of semi-saturated tones indicates that engaging in the strategies represented by these variables is all-or-nothing.

map, variable 1, separates photosynthetic Cyanobacteria from all other bacteria: the 93 cyanobacterial genomes in the dataset are assigned the highest values, while all others have values that are close to zero (Fig. 1A). To confirm that this variable detects photosynthetic activity, we identified the metabolites that are most enriched in the metabolic networks of genera that are located farthest along the variable axis (see Methods). This revealed over-representation of cyanophycin, a nitrogen reserve polymer, ribulose-1,5-bisphosphate (RuBP), used for carbon fixation from Ru-
BisCO during photosynthesis; and bicarbonate, which promotes the accumulation of CO$_2$ near RuBisCO$^{[8]}$, confirming that variable indicates the extent to which organisms follow a photosynthetic strategy (Fig. 1; Supp. Table S1).

The sharp differences in variable 1 show that a photosynthetic lifestyle is a discrete yes-no strategy choice where little middle ground exists. The diffusion map identifies further variables that indicate such discrete clade-level choices including acetic acid production$^{[19]}$ (variable 20), chemoautotrophic and sulfur-oxidization strategies found near deep sea vents or marine sediments (variable 22), and carnitine use for stress tolerance among anaerobic animal associates in the Coriobacterii$^{[20]}$ (variable 23; Fig. 1).

The diffusion map also identifies some continuous strategy axes that are relevant for a wide variety of genera. The most important of these is variable 4 that orders taxa based on metabolic strategies reflecting a spectrum from life in marine environments (e.g. marine genera in the Rhodobacterales and Rhizobiales), to associations with terrestrial plant and animal hosts (Fig. 2). The species that score the lowest (i.e. most negative) values in variable 4 are epipelagic and marine animal-associated bacteria that utilise a broad spectrum of carbon sources. Here the most significant metabolic reactions are all involved in the characteristic production of medium chain-length biopolymers (Fig. 2C; Supp. Table S2). At the opposite end we find host-associated $\gamma$-proteobacteria, Bacilli, and Clostridia (Fig. 2B). Among the top 10 most correlated metabolic activities for species at this extreme are the uptake of the common enteric amino acid L-histidine, the vitamin riboflavin, and the production of the signalling molecule precursor 4,5-dihydroxy-2,3-pentanedione$^{[21]}$ (Fig. 2C; Supp. Table S2). Our interpretation is that this variable traces a broad range of gradual adaptations from a generalist lifestyle in oceans to close interactions with hosts.
**Figure 2:** Broad spectrum of adaptations indicated by variable 4. A Variable entries for each genome are shown as bars near the tips of a phylogenetic tree. Blue and red bars mark positive and negative variable entries. B The ordering of genomes from smallest to largest (left to right) indicated by variable 4, summarised at the taxonomic class-level in 100 equally spaced bins. Taxonomic groups: ■ Actinobacteria, ■ α-proteobacteria, ■ Bacilli, ■ Bacteroidia, ■ β-proteobacteria, ■ Clostridia, ■ Cytophagia, ■ δ-proteobacteria, ■ Erysipelotrichia, ■ Flavobacteriia, ■ γ-proteobacteria, ■ Negativicutes, ■ Tissierellia, and ■ Other (< 0.75%). C The top 10 over/under-produced metabolites in the metabolic networks of taxa with the smallest (red tiles) and largest (blue tiles) variable entries (Supp. Table S2). Black ticks indicate an inability to synthesise. The wide variety of different values of this variable indicate a gradual spectrum of adaptations that reflect a broad scale of possible lifestyles ranging from free life in the oceans (red) to terrestrial host association (blue).

Other variables reflect specific strategies for interactions with a host. Perhaps the most interesting example is variable 3 in which pathogenic γ-proteobacteria score low values. For example Franconibacter pulveris, Escherichia coli O157, Enterobacter cloacae are the top 3 lowest ranked taxa in this variable. Characteristic metabolites of this strategy include production of ferroxamine and ferrioxamine and L-methionine-R-sulfoxide, a molecule related to antioxidant activities and adherence to eukaryotic cells (Supp. Table S3). These indicate that variable 3 reflects strategies to counter attempts by hosts to combat bacteria through iron sequestration and oxidative stress.

Host-microbe interactions also feature in variable 8, which highlights endosymbionts and endoparasites with small genomes. We found the highest values of this variable in animal- and plant-associated Mollicutes, as well as candidate genera like Tremblaya and Sulcia, that have been
Figure 3: Diffusion variables as indicators of functional differentiation and niche convergence. Some variables such as variable 17 (“amino sugar metabolism”, A) highlight functional differences in closely related taxa, corresponding to the appearance of large positive and negative variable values (long red and blue bars, respectively) in close proximity in the phylogenetic tree. Other variables such as variable 16 (“hydrogen as an electron donor”, B) show functional similarities across the tree of life. With respect to these variables, similar adaptations are shared by remotely related taxa (similar bars in distal parts of the tree), providing evidence for niche convergence.

isolated from insect bacteriocytes\textsuperscript{25,26}. Among the top 10 markers of taxa scoring highly in variable 8 are an inability to synthesise key amino acids such as L-histidine, L-phenylalanine, L-leucine, L-isoleucine, L-lysine, and L-valine (Supp. Table \textsuperscript{S4}), signifying organisms with streamlined genomes and strong host dependency.

Some diffusion variables differentiate between closely-related taxa. Variable 17 identifies differences in the metabolic strategies of marine \textit{Rhodobacteraceae}, based on genomic capabilities related to the production of aromatics, and the incorporation of the widespread amino sugar, N-Acetyl-D-glucosamine, into peptidoglycan\textsuperscript{27} (Fig. 3A; Supp. Table \textsuperscript{S5}). Another example is variable 26 which differentiates between different strategies in the \textit{Enterobacteraceae}. Among the top 10 entries on one side of the divide are soil- (\textit{Buttiauxella ferragutiae}, \textit{Kluyvera cryocrescens}), aquatic- (\textit{Enterobacter cloacae}), and insect-associated (\textit{Enterobacillus tribolii}) taxa, with many
acting as opportunistic human pathogens. The metabolic strategies of these taxa center on the use of cysteate and taurine as sulfur sources, and the production of defence molecules (Supp. Table S6), potentially reflecting an ability to obtain compounds from animals to weather stress. The second group included human pathogens (Yersinia pestis, Salmonella enterica, Klebsiella pneumoniae), characterised by an over-representation of metabolites involved in tryptophan metabolism (Supp. Table S6), perhaps indicative of a crosstalk between pathogens and host immunity.

Other diffusion variables show evidence of metabolic niche convergence, wherein similar strategies are seen among distantly-related taxa. For example high values of variable 16 are observed across multiple classes, including Acidobacteria, Planctomycetia, Verrucomicrobiae, Blastocatellia, and Gemmatimonadetes and particularly low values are found in β-proteobacteria, δ-proteobacteria, α-proteobacteria, Bacteroidetes, and Chlamydiae (Fig. 3B). Taxa scoring large values encode for metabolites involved in corrinoid iron sulfur protein production, and aspects of sugar metabolism or peptidoglycan biosynthesis (Supp. Table S7), whereas taxa receiving low values exhibit metabolites in cysteine metabolism, the glucuronate pathway, and coenzyme A products, which enable the use of hydrogen as an electron donor in key biosynthetic processes.

The examples above demonstrate that the diffusion variables provide meaningful coordinates that trace the space of feasible metabolic strategies. Using a procedure proposed by Moon et al. we can combine these variables in a visualisation of the strategy space (Fig. 4). This embedding shows that the bacterial niche space is a filamentous object with multiple quasi one-dimensional branches rising out of a common core. This is in contrast to both Hutchinson’s original idea of the niche space as a solid hypervolume and modern ideas which postulate that feasible strategies form discrete clusters. We conjecture that the filamentous structure reported here has strong
implications for the evolution of bacteria.

In this paper we showed that the metabolic strategies encoded by bacterial genomes can be understood by a combination of metabolic reconstruction and diffusion mapping. The diffusion maps reveal a wealth of biologically salient variables that span the strategy space. Some show evidence of discrete strategies such as photosynthesis in *Cyanobacteria*. Others strategies span a continuous space representing for example different degrees of specialisation or reliance on hosts.
Yet others highlight specific strategies for energy production or stress responses, some of which differentiate closely related species or emerged, likely through convergent evolution, in different branches of the tree of life. The diffusion variables provide an alternative way to organise and systematise the wealth of genomic information that has become available in recent years. Perhaps more importantly, we believe that they provide the right vocabulary for describing bacterial communities, by enabling researchers to discuss to which extent different niches are occupied. From an ecological point of view the present analysis constitutes the most extensive mapping of a niche space so far, and facilitates the application of ecological concepts to bacterial communities.

Our analysis focused largely on the bacteria’s capabilities to catalyse steps in primary metabolism. Even within the realm of primary metabolism the genes reveal only the set of theoretical capabilities, i.e. the fundamental niche\(^2\). Hence our analysis ignores the complexity of biotic interactions, large parts of secondary metabolism, behaviour, and regulation. For any other group of organisms such a limited analysis would be mostly meaningless, however due to the great diversity of metabolic strategies in bacteria it reveals a rich and complex niche space. We envision that with future transcriptomic data, diffusion maps could also map the realised niche, i.e. the metabolic strategies that are employed under a given set of conditions, bringing our understanding of ecology in complex microbial systems closer to the biochemical level.
Methods

Generating metabolic networks. Genomes were obtained from the National Center for Biotechnology Information (NCBI) RefSeq\textsuperscript{16} database (accessed on 2019 March 20, 2019). Namely, we acquired the ‘representative,’ ‘reference,’ ‘complete,’ ‘contig,’ and ‘scaffold’ sets and reduced these to a set of genus-level representatives using the following procedure. We first selected a random representative genome for each unique genus in the combined ‘representative’ and ‘reference’ sets (\(N = 1,849\) genera). Novel genera in the remaining RefSeq categories, that were not already represented in the ‘reference’ and ‘representative’ sets, were then appended to the set (\(N = 801\) genera) in the same way, for a total of \(N = 2,643\) genomes. Metabolic models were automatically constructed for the genome set from protein annotations of genome assemblies using the CarveMe reconstruction algorithm\textsuperscript{32}, that starts with a universal bacterial metabolic model comprising all known biochemical reactions in the BiGG Models\textsuperscript{33} database and generates sets of genome-specific reactions by paring those without proteomic support in the annotated genome.

Models were summarised as metabolic networks — directed graphs in which nodes are chemical metabolite compounds and directed edges link substrates to products\textsuperscript{15}. A single feasible direction was estimated for reversible reactions through flux balance analyses of metabolic networks using the \texttt{fbar} library in the statistical programming environment R. Models that did not exhibit positive biomass growth of the objective function following flux balance analysis were excluded from analyses. The giant component of each metabolic network’s cytoplasmic compartment was retained for diffusion mapping, resulting in a set of 2,621 unique metabolic networks representing major features of cellular metabolism across most currently known bacterial genera.
**Phylogenetic tree generation.** Phylogenetic trees used for visualisation were constructed using the GToTree pipeline\textsuperscript{11} with the “universal” protein set defined by Hug et al\textsuperscript{11}. GToTree identifies target genes using HMMER\textsuperscript{35}, aligns them with MUSCLE\textsuperscript{36}, trims alignments using trimAl\textsuperscript{37}, and then concatenates the output. Trees were generated from the aligned and concatenated gene sets using FastTree\textsuperscript{38}, and visualised using iToL\textsuperscript{39} and the phytools\textsuperscript{40} library in R\textsuperscript{41}.

**Identifying associated metabolites.** We sought to identify metabolites that were overrepresented in the metabolic networks of taxa, that were themselves assigned extreme entries in diffusion map variables. This was accomplished using a permutational analysis which we refer to as a ‘metabolite set enrichment analyses,’ analogous to the gene set enrichment analysis, GSEA\textsuperscript{42}. Genome rankings were provided by the taxonomic orderings specified by each diffusion variable. Analyses were accomplished for each preranked set using the fgsea library in R, with an FDR-adjusted \( P \)-value < 0.05 used as the threshold for retaining metabolites associated with extreme taxa.

**Diffusion map procedure.** Diffusion mapping\textsuperscript{14,43} was performed using the algorithm described by Barter & Gross\textsuperscript{44}. Briefly, the method involves (i) calculating an affinity matrix describing euclidean similarities among the \( k \)-nearest neighbours for samples in a dataset (ii) interpreting this as a weighted adjacency matrix, and (iii) computing the corresponding row-normalised Laplacian matrix. The eigenvectors of the Laplacian represent new diffusion variables describing important variation in the dataset. The first (i.e. most important) variable is given by the eigenvector corresponding to the smallest non-zero eigenvalue, then the second smallest eigenvalue, and so on. An R\textsuperscript{41} implementation of this procedure is provided at [https://github.com/AshkaanF/diffusion_maps](https://github.com/AshkaanF/diffusion_maps).
References

1. Grinnell, J. The niche-relationships of the california thrasher. *Auk* **34**, 427–433 (1917).

2. Hutchinson, G. E. Cold spring harbor symposium on quantitative biology. *Concluding remarks* **22**, 415–427 (1957).

3. Chase, J. M. Ecological niche theory. *The theory of ecology* 93–107 (2011).

4. Winemiller, K. O., Fitzgerald, D. B., Bower, L. M. & Pianka, E. R. Functional traits, convergent evolution, and periodic tables of niches. *Ecology letters* **18**, 737–751 (2015).

5. MacArthur, R. Coexistence of species. in challenging biological problems: Directions toward their solution.(ed. ja behnke.) pp. 253–259 (1972).

6. Gilbert, J. A., Jansson, J. K. & Knight, R. The earth microbiome project: successes and aspirations. *BMC biology* **12**, 69 (2014).

7. Thompson, L. R. *et al.* A communal catalogue reveals earths multiscale microbial diversity. *Nature* **551**, 457 (2017).

8. Hammer, T. J., Sanders, J. G. & Fierer, N. Not all animals need a microbiome. *FEMS microbiology letters* (2019).

9. Mee, M. T., Collins, J. J., Church, G. M. & Wang, H. H. Syntrophic exchange in synthetic microbial communities. *Proceedings of the National Academy of Sciences* 201405641 (2014).

10. Louca, S. *et al.* Function and functional redundancy in microbial systems. *Nature ecology & evolution* **2**, 936 (2018).
11. Chase, J. M. & Leibold, M. A. *Ecological niches: linking classical and contemporary approaches* (University of Chicago Press, 2003).

12. Pianka, E. R., Vitt, L. J., Pelegrin, N., Fitzgerald, D. B. & Winemiller, K. O. Toward a periodic table of niches, or exploring the lizard niche hypervolume. *The American Naturalist* **190**, 601–616 (2017).

13. Hug, L. A. *et al.* A new view of the tree of life. *Nature microbiology* **1**, 16048 (2016).

14. Coifman, R. R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the national academy of sciences* **102**, 7426–7431 (2005).

15. Borenstein, E., Kupiec, M., Feldman, M. W. & Ruppin, E. Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proceedings of the National Academy of Sciences* **105**, 14482–14487 (2008).

16. Pruitt, K. D., Tatusova, T. & Maglott, D. R. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* **35**, D61–D65 (2006).

17. Watzer, B. & Forchhammer, K. Cyanophycin synthesis optimizes nitrogen utilization in the unicellular cyanobacterium synechocystis sp. strain pcc 6803. *Appl. Environ. Microbiol.* **84**, e01298–18 (2018).

18. Espie, G. S. & Kimber, M. S. Carboxysomes: cyanobacterial rubisco comes in small packages. *Photosynthesis Research* **109**, 7–20 (2011).
19. Komagata, K., Iino, T. & Yamada, Y. The family acetobacteraceae. *The prokaryotes: Alphaproteobacteria and betaproteobacteria* 3–78 (2014).

20. Meadows, J. A. & Wargo, M. J. Carnitine in bacterial physiology and metabolism. *Microbiology* **161**, 1161 (2015).

21. De Keersmaecker, S. C. *et al.* Chemical synthesis of (s)-4, 5-dihydroxy-2, 3-pentanedione, a bacterial signal molecule precursor, and validation of its activity in salmonella typhimurium. *Journal of Biological Chemistry* **280**, 19563–19568 (2005).

22. Kingsley, R. A. *et al.* Ferrioxamine-mediated iron (iii) utilization by salmonella enterica. *Appl. Environ. Microbiol.* **65**, 1610–1618 (1999).

23. Levine, R. L., Mosoni, L., Berlett, B. S. & Stadtman, E. R. Methionine residues as endogenous antioxidants in proteins. *Proceedings of the National Academy of Sciences* **93**, 15036–15040 (1996).

24. Zhao, C. *et al.* Role of methionine sulfoxide reductases a and b of enterococcus faecalis in oxidative stress and virulence. *Infection and immunity* **78**, 3889–3897 (2010).

25. Chang, H.-H. *et al.* Complete genome sequence of candidatus sulcia muelleri ml, an obligate nutritional symbiont of maize leafhopper (dalbulus maidis). *Genome Announc.* **3**, e01483–14 (2015).

26. López-Madrigal, S., Latorre, A., Moya, A. & Gil, R. The link between independent acquisition of intracellular gamma-endsymbionts and concerted evolution in tremblaya princeps. *Frontiers in microbiology* **6**, 642 (2015).
27. Riemann, L. & Azam, F. Widespread n-acetyl-d-glucosamine uptake among pelagic marine bacteria and its ecological implications. *Appl. Environ. Microbiol.* **68**, 5554–5562 (2002).

28. Uria-Nickelsen, M. R., Leadbetter, E. R. & Godchaux III, W. Sulphonate utilization by enteric bacteria. *Microbiology* **139**, 203–208 (1993).

29. Olive, A. J. & Sassetti, C. M. Metabolic crosstalk between host and pathogen: sensing, adapting and competing. *Nature reviews Microbiology* **14**, 221 (2016).

30. Braakman, R. & Smith, E. The emergence and early evolution of biological carbon-fixation. *PLoS Computational Biology* **8**, e1002455 (2012).

31. Moon, K. R. *et al.* Phate: a dimensionality reduction method for visualizing trajectory structures in high-dimensional biological data. *bioRxiv* 120378 (2017).

32. Machado, D., Andrejev, S., Tramontano, M. & Patil, K. R. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic acids research* **46**, 7542–7553 (2018).

33. King, Z. A. *et al.* Bigg models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research* **44**, D515–D522 (2015).

34. Lee, M. D. Gtотree: a user-friendly workflow for phylogenomics. *Bioinformatics* **1**, 3 (2019).

35. Eddy, S. R. Accelerated profile hmm searches. *PLoS computational biology* **7**, e1002195 (2011).

36. Edgar, R. C. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792–1797 (2004).
37. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

38. Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one* **5**, e9490 (2010).

39. Letunic, I. & Bork, P. Interactive tree of life (itol) v4: recent updates and new developments. *Nucleic acids research* (2019).

40. Revell, L. J. phytools: an r package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**, 217–223 (2012).

41. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2019). URL [https://www.R-project.org/](https://www.R-project.org/)

42. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).

43. Coifman, R. R. & Lafon, S. Diffusion maps. *Applied and computational harmonic analysis* **21**, 5–30 (2006).

44. Barter, E. & Gross, T. Manifold cities: social variables of urban areas in the uk. *Proceedings of the Royal Society A* **475**, 20180615 (2019).

45. Sergushichev, A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *BioRxiv* 060012 (2016).

**Acknowledgements** We thank Jonathan A. Eisen and James P. O’Dwyer for comments.
Competing Interests  The authors declare that they have no competing financial interests.

Correspondence  Correspondence and requests for materials should be addressed to A.K.F. (email: afahimipour@ucdavis.edu).
Table S1: Top 10 over-represented metabolites in the metabolic networks of the Cyanobacteria identified by the highly localised variable 1. The Synthesised column indicates whether the inferred metabolic network is capable of synthesising metabolites in the rows (i.e., the node has an in degree > 0 in the network). The Norm. Enrich. Score and Adj. P-value columns show the normalised ‘enrichment score’ and FDR-adjusted P-value from the permutational enrichment analysis. Positive enrichment scores indicate that metabolites are over-represented in the networks of taxa that receive the largest diffusion variable entries.
| Metabolite                                                      | Synthesised | Norm. Enrich. Score | FDR adj. $P$-value |
|---------------------------------------------------------------|-------------|---------------------|--------------------|
| 4,5-dihydroxy-2,3-pentanedione                                 | Yes         | 2.990               | 0.0005             |
| Hydroxylamine                                                 | Yes         | 2.855               | 0.0005             |
| Riboflavin                                                    | No          | 2.797               | 0.0005             |
| Menaquinone 8                                                 | No          | 2.756               | 0.0005             |
| O$_2$                                                         | No          | 2.734               | 0.0005             |
| L-Histidine                                                  | No          | 2.716               | 0.0005             |
| Copper                                                        | No          | 2.703               | 0.0005             |
| (2R,4S)-2-methyl-2,4-dihydroxydihydrofuran-3-one             | Yes         | 2.703               | 0.0005             |
| (2R,4S)-2-methyl-2,3,3,4-tetrahydroxytetrahydrofuran        | Yes         | 2.703               | 0.0005             |
| 4-hydroxy-5-methyl-3(2H)-furanone                            | Yes         | 2.671               | 0.0005             |
| $C_{10}:0$-Medium-chain length (MCL) Polyhydroxyalkanoate     | Yes         | −3.400              | 0.0005             |
| $C_{12}:0$ MCL Polyhydroxyalkanoate                          | Yes         | −3.400              | 0.0005             |
| $C_{12}:1$ MCL Polyhydroxyalkanoate                          | Yes         | −3.400              | 0.0005             |
| $C_{14}:0$ MCL Polyhydroxyalkanoate                          | Yes         | −3.400              | 0.0005             |
| $C_{14}:1$ MCL Polyhydroxyalkanoate                          | Yes         | −3.400              | 0.0005             |
| $C_{16}:0$ MCL Polyhydroxyalkanoate                          | Yes         | −3.400              | 0.0005             |
| $C_{18}:0$ MCL Polyhydroxyalkanoate                          | Yes         | −3.400              | 0.0005             |
| MCL Polyhydroxyalkanoate                                     | Yes         | −3.400              | 0.0005             |
| (R)-Hydroxydodecanoyl-5-en-CoA                                 | Yes         | −3.400              | 0.0005             |
| (R)-3-hydroxy-cis-myristol-7-eoyl-CoA                         | Yes         | −3.400              | 0.0005             |

Table S2: Top 10 over-represented metabolites in the metabolic networks of taxa that are farthest along variable
4. The Norm. Enrich. Score and Adj. $P$-value columns show the normalised ‘enrichment score’ and FDR-adjusted
$P$-value from the permutational enrichment analysis$^{[4,5]}$. Positive enrichment scores indicate that metabolites are
over-represented in the networks of taxa that receive the largest diffusion variable entries; negative scores indicate that
metabolites are over-represented for taxa that receive the smallest entries.
| Metabolite                                      | Synthesised | Norm. Enrich. Score | FDR adj. P-value |
|------------------------------------------------|-------------|--------------------|------------------|
| 2-Methylbutanoyl-CoA                           | No          | 2.499              | 0.022            |
| Phenylacetyl-CoA                                | No          | 2.139              | 0.027            |
| Hexanoate                                       | No          | 2.073              | 0.017            |
| 6-Phospho-D-gluconate                          | No          | 1.804              | 0.032            |
| 2-Demethylmenaquinol 8                         | No          | 1.735              | 0.026            |
| N-Succinyl-2-L-amino-6-oxoheptan...            | No          | 1.709              | 0.039            |
| Acetyl-cystine-bimane                           | Yes         | -1.547             | 0.0002           |
| Bimane                                          | No          | -1.547             | 0.0002           |
| Bimane conjugated mycothiol                     | Yes         | -1.547             | 0.0002           |
| Cys-1D-myo-inositol 2-deoxy-D-gl...             | Yes         | -1.547             | 0.0002           |
| 1D-myo-inositol 2-deoxy-D-glucop...             | Yes         | -1.547             | 0.0002           |
| L-Methionine Sulfoxide                          | No          | -1.586             | 0.0003           |
| D-Cysteine                                      | No          | -1.592             | 0.0002           |
| Generic ferrioxamine-Fe-III                    | No          | -1.599             | 0.0002           |
| Ferroxamine minus Fe(3)                         | Yes         | -1.599             | 0.0002           |
| L-methionine-R-sulfoxide                        | No          | -1.619             | 0.0003           |

Table S3: Top over-represented metabolites in the metabolic networks of taxa that are farthest along variable 3. The *Norm. Enrich. Score* and *Adj. P-value* columns show the normalised ‘enrichment score’ and FDR-adjusted *P*-value from the permutational enrichment analysis\(^{42,45}\). Positive enrichment scores indicate that metabolites are over-represented in the networks of taxa that receive the largest diffusion variable entries; negative scores indicate that metabolites are over-represented for taxa that receive the smallest entries. For this variable, only 6 metabolites are significantly associated with positive variable entries.
| Metabolite                          | Synthesised | Norm. Enrich. Score | FDR adj.  | P-value   |
|------------------------------------|-------------|---------------------|-----------|-----------|
| Riboflavin                         | No          | 2.151               | 0.001     |           |
| L-Histidine                        | No          | 2.049               | 0.001     |           |
| L-Phenylalanine                    | No          | 2.040               | 0.001     |           |
| L-Leucine                          | No          | 2.011               | 0.001     |           |
| L-Isoleucine                       | No          | 1.980               | 0.001     |           |
| L-Lysine                           | No          | 1.975               | 0.001     |           |
| (R)-Pantothenate                   | No          | 1.947               | 0.001     |           |
| Chorismate                         | No          | 1.930               | 0.001     |           |
| L-Valine                           | No          | 1.917               | 0.002     |           |
| Adenosine                          | No          | 1.886               | 0.001     |           |
| Cyclic de-hypoxanthine futalosine  | Yes         | −1.887              | 0.001     |           |
| De-hypoxanthine futalosine         | Yes         | −1.887              | 0.001     |           |
| Futalosine                         | Yes         | −1.887              | 0.001     |           |
| Mannobiose                         | Yes         | −1.907              | 0.001     |           |
| Mannotriose                        | Yes         | −1.936              | 0.001     |           |
| Mannotetraose                      | No          | −1.936              | 0.001     |           |
| 5,6-dihydrouracil                  | Yes         | −1.943              | 0.001     |           |
| Carboxyspermidine                  | Yes         | −1.958              | 0.001     |           |
| 4-Hydroxyphenylacetyl-CoA          | Yes         | −1.995              | 0.001     |           |
| Oxidized ferredoxin                | Yes         | −2.061              | 0.001     |           |

Table S4: Top 10 over-represented metabolites in the metabolic networks of taxa that are farthest along variable 8. The Norm. Enrich. Score and Adj. P-value columns show the normalised ‘enrichment score’ and FDR-adjusted P-value from the permutational enrichment analysis. Positive enrichment scores indicate that metabolites are over-represented in the networks of taxa that receive the largest diffusion variable entries; negative scores indicate that metabolites are over-represented for taxa that receive the smallest entries.
| Metabolite                               | Synthesised | Norm. Enrich. Score | FDR adj. $P$-value |
|-----------------------------------------|-------------|--------------------|-------------------|
| 2-Hydroxy-cis-hex-2,4-dienoate          | Yes         | 2.132              | 0.002             |
| 4-Hydroxy-2-oxohexanoic acid            | Yes         | 2.132              | 0.002             |
| 4-methylbenzyl alcohol                  | Yes         | 2.132              | 0.002             |
| P-methylbenzaldehyde                    | Yes         | 2.132              | 0.002             |
| 4-Methylcatechol                        | Yes         | 2.132              | 0.002             |
| Cis-1,2-Dihydroxy-4-methylcyclohexan... | Yes         | 2.132              | 0.002             |
| 2-Hydroxy-5-methyl-cis,cis-muconate...  | Yes         | 2.132              | 0.002             |
| P-toluene                               | Yes         | 2.132              | 0.002             |
| P-methyltoluene                         | No          | 2.132              | 0.002             |
| Decanoyl CoA                            | No          | 2.067              | 0.002             |
| Octanoyl-CoA                            | No          | −1.523             | 0.004             |
| Guanosine                               | No          | −1.531             | 0.011             |
| Decanoyl-CoA                            | No          | −1.548             | 0.003             |
| Thymidine                               | No          | −1.561             | 0.007             |
| Deoxyadenosine                          | No          | −1.652             | 0.003             |
| Xanthosine                              | No          | −1.655             | 0.003             |
| Dodecanoate                             | No          | −1.677             | 0.002             |
| Glycerophosphoglycerol                  | No          | −1.693             | 0.002             |
| Trans-Tetradec-2-enoyl-CoA              | No          | −1.728             | 0.002             |
| N-Acetyl-D-glucosamine                  | Yes         | −1.753             | 0.002             |

Table S5: Top 10 over-represented metabolites in the metabolic networks of taxa that are farthest along variable 17. The Norm. Enrich. Score and Adj. $P$-value columns show the normalised ‘enrichment score’ and FDR-adjusted $P$-value from the permutational enrichment analysis. Positive enrichment scores indicate that metabolites are over-represented in the networks of taxa that receive the largest diffusion variable entries; negative scores indicate that metabolites are over-represented for taxa that receive the smallest entries.
| Metabolite                        | Synthesised | Norm. Enrich. Score | FDR adj. P-value |
|----------------------------------|-------------|---------------------|------------------|
| L-Cysteate                       | No          | 1.798               | 0.011            |
| Taurine                          | Yes         | 1.798               | 0.011            |
| L methionine R oxide             | No          | 1.771               | 0.011            |
| (S)-Methylmalonate semialdehyde  | Yes         | 1.754               | 0.011            |
| L Methionine S oxide             | No          | 1.718               | 0.011            |
| 2-Hydroxy-cis-hex-2,4-dienoate   | Yes         | 1.718               | 0.011            |
| 4-Hydroxy-2-oxohexanoic acid     | Yes         | 1.718               | 0.011            |
| 4-methylbenzyl alcohol           | Yes         | 1.718               | 0.011            |
| P-methylbenzaldehyde             | Yes         | 1.718               | 0.011            |
| 4-Methylcatechol                 | Yes         | 1.718               | 0.011            |
| Glucosyl-heptosyl-heptosyl-kdo2-...| Yes       | −1.507              | 0.011            |
| Heptosyl-glucosyl-heptosyl-hepto...| Yes       | −1.507              | 0.011            |
| (R)-Hydroxyhexanoyl-CoA           | Yes         | −1.537              | 0.019            |
| Indole 3 acetaldehyde            | No          | −1.540              | 0.015            |
| (R)-Hydroxyoctanoyl-CoA           | Yes         | −1.542              | 0.017            |
| Benzoate                         | No          | −1.543              | 0.015            |
| (R)-3-Hydroxydodecanoyl-CoA       | Yes         | −1.557              | 0.011            |
| Folate                           | Yes         | −1.558              | 0.011            |
| 3’,5’-Cyclic GMP                  | Yes         | −1.571              | 0.011            |
| Tryptophanyl-beta-D-glucuronide   | Yes         | −1.646              | 0.011            |

Table S6: Top 10 over-represented metabolites in the metabolic networks of taxa that are farthest along variable 26. The Norm. Enrich. Score and Adj. P-value columns show the normalised ‘enrichment score’ and FDR-adjusted P-value from the permutational enrichment analysis. Positive enrichment scores indicate that metabolites are over-represented in the networks of taxa that receive the largest diffusion variable entries; negative scores indicate that metabolites are over-represented for taxa that receive the smallest entries.
| Metabolite                                           | Synthesised | Norm. Enrich. Score | FDR adj. $P$-value |
|----------------------------------------------------|-------------|---------------------|-------------------|
| Corrinoid Iron sulfur protein                      | Yes         | 1.996               | 0.003             |
| Methylcorrinoid iron sulfur prote...               | Yes         | 1.996               | 0.003             |
| DTDP-4-acetamido-4,6-dideoxy-D-galac...             | Yes         | 1.896               | 0.004             |
| DTDP-4-amino-4,6-dideoxy-D-galac...                | Yes         | 1.896               | 0.004             |
| UDP-N-acetyl-D-mannosamine                          | Yes         | 1.896               | 0.004             |
| UDP-N-acetyl-D-mannosaminouronat...                | Yes         | 1.896               | 0.004             |
| Undecaprenyl-dipospho-N-acetylg...                 | Yes         | 1.896               | 0.004             |
| Undecaprenyl-dipospho N-acetylg...                 | Yes         | 1.896               | 0.004             |
| Undecaprenyl diphospho N-acetyl...                 | Yes         | 1.817               | 0.005             |
| 3′,5′-cyclic adenosine monophosp...                 | Yes         | 1.663               | 0.007             |
| 4-methylbenzyl alcohol                              | Yes         | $-1.740$            | 0.001             |
| P-methylbenzaldehyde                                | Yes         | $-1.740$            | 0.001             |
| 4-Methylcatechol                                    | Yes         | $-1.740$            | 0.001             |
| Cis-1,2-Dihydroxy-4-methylcyclohex...               | Yes         | $-1.740$            | 0.001             |
| 2-Hydroxy-5-methyl-cis,cis-mucon...                 | Yes         | $-1.740$            | 0.001             |
| P-toluate                                           | Yes         | $-1.740$            | 0.001             |
| P-methyltoluene                                    | No          | $-1.740$            | 0.001             |
| Dodecanoyl CoA n C<sub>12</sub>:0 CoA               | No          | $-1.792$            | 0.001             |
| Decanoyl CoA                                       | No          | $-1.854$            | 0.001             |
| D-Cysteine                                         | No          | $-1.899$            | 0.002             |

Table S7: Top 10 over-represented metabolites in the metabolic networks of taxa that are farthest along variable 16. The Norm. Enrich. Score and Adj. $P$-value columns show the normalised ‘enrichment score’ and FDR-adjusted $P$-value from the permutational enrichment analysis[^42[^45]. Positive enrichment scores indicate that metabolites are over-represented in the networks of taxa that receive the largest diffusion variable entries; negative scores indicate that metabolites are over-represented for taxa that receive the smallest entries.