nature ecology & evolution

Supplementary information

https://doi.org/10.1038/s41559-024-02440-6

Metabolic complexity drives divergence in microbial communities

In the format provided by the authors and unedited

1 Measuring divergence with the Aitchison distance metric

In our study, we use the Aitchison distance[1] to measure divergence (i.e. the beta-diversity between communities within the same condition) because of its sensitivity to non-overlapping species between communities and its statistical properties with respect to handling compositional data. To illustrate why we chose the Aitchison distance, we will define it alongside other commonly used beta-diversity metrics - Bray-Curtis[2] and Jensen-Shannon Distance (JSD)[3] - and consider the effects of each across different hypothetical scenarios.

1.1 Metric definitions

$$\text{Aitchison}(\mathbf{x}, \mathbf{y}) = E(clr(\mathbf{x}), clr(\mathbf{y})) = \sqrt{\sum \left(clr(\mathbf{x}) - clr(\mathbf{y})\right)^2} \tag{1}$$

Where E is the Euclidean distance, clr is the center log ratio transformation $clr(\mathbf{x}) = \log \mathbf{x}/\mathrm{G}(\mathbf{x})$, and $\mathrm{G}(\mathbf{x})$ is the geometric mean. A pseudocount of 1 is added to each vector $(\mathbf{x} + 1)$ to allow for a meaningful calculation of log transformations and $\mathrm{G}(\mathbf{x})$. See Section 1.3 for more information about the effect of pseduocounts.

$$\text{Bray-Curtis}(\mathbf{x}, \mathbf{y}) = \frac{\sum |\mathbf{x} - \mathbf{y}|}{\sum |\mathbf{x} + \mathbf{y}|} \tag{2}$$

$$\text{JSD}(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{D(\mathbf{x}||m) + D(\mathbf{y}||m)}{2}} \tag{3}$$

Where D is the Kullback-Liebler divergence: For distributions $\mathbf{a}$(i) and $\mathbf{b}$(i), $D(\mathbf{a}(i)||\mathbf{b}(i)) = \sum \mathbf{a}(i) \log \frac{\mathbf{a}(i)}{\mathbf{b}(i)}$. And where $m = (\mathbf{x} + \mathbf{y})/2$.

1.2 Simple hypothetical scenarios

To compare these three metrics, we can interpret their results for the following four scenarios. For simplicity's sake we can assume pairs of communities where all taxa that are found in each community are sequenced to equal amounts:

Scenario 1: Entirely non-overlapping communities

$C1:$	100.	0.	0.	0.	0.	0.	0.	0.	0.	0.
$C2:$	0.	100.	0.	0.	0.	0.	0.	0.	0.	0.

Scenario 2: Partially overlapping low-diversity communities

$C1:$	100.	0.	0.	0.	0.	0.	0.	0.	0.	0.
$C2:$	100.	100.	0.	0.	0.	0.	0.	0.	0.	0.

Scenario 3: Entirely non-overlapping diverse communities

$C3:$	100.	100.	100.	100.	100.	0.	0.	0.	0.	0.
$C4:$	0.	0.	0.	0.	0.	100.	100.	100.	100.	100.

Scenario 4: Partially overlapping diverse communities

$C3:$	100.	100.	100.	100.	100.	0.	0.	0.	0.	0.
$C4:$	100.	0.	0.	0.	0.	100.	100.	100.	100.	100.

	Aitchison	Bray-Curtis	JSD
Scenario 1: Non-overlapping, low diversity	6.526766	1.000000	0.832555
Scenario 2: Partially overlapping, low diversity	3.263383	0.333333	0.464501
Scenario 3: Non-overlapping, high diversity	14.594293	1.000000	0.832555
Scenario 4: Partially overlapping, high diversity	13.768228	0.818182	0.752880

Table 1: **Supplementary Table 1 — Divergence between hypothetical communities using different beta-diversity metrics.** The distances computed in scenarios 1 (non-overlapping taxa) and 2 (partially overlapping taxa) are with low diversity communities (C1, C2, and C3) while distances computed in scenarios 3 (non-overlapping) and 4 (partially overlapping) are with high diversity communities (C4, C5, and C6).

These scenarios reveal how each metric is sensitive to overlapping taxa and diversity, but how only the Aitchison distance detects differences in the special case where no taxa overlap between communities. The sensitivity to overlapping taxa is clear when comparing scenarios 1 and 2 or when comparing scenarios 3 and 4 - in both cases, all metrics decrease in value when a taxon becomes shared between communities (Scenario 1 > Scenario 2 and Scenario 3 > Scenario 4). The sensitivity to diversity is apparent by comparing scenarios 2 and 4, where all metrics increase in value when more unshared taxa are present in each community (Scenario 4 > Scenario 2). Comparing scenarios 1 and 3 reveals the special case where the diversity of both communities increases without any overlapping taxa. Here, the value of Bray-Curtis and JSD is unchanged, while Aitchison interprets this increase in differences between communities as an increase in distance (Scenario 3 > Scenario 1 for Aitchison and Scenario 1 = Scenario 3 for Bray-Curtis and JSD).

The Aitchison distance may be best for the scenario in our study, where we are comparing communities cultured under the same conditions (i.e. divergence). We are precisely interested in the situation in which communities that may start off similar to each other undergo different assembly processes (including the selection of entirely non-overlapping taxa) despite growing on the same conditions. With the Aitchison, we can interpret non-overlapping communities of few species as more similar than non-overlapping communities of many species, whereas the other distance metrics make no distinction (Scenario 1 versus Scenario 3). We further discuss the relationship between diversity and the Aitchison distance in the following section, "Divergence is sensitive to richness and evenness" (Section 2).

An additional and critical justification for using the Aitchison distance relates to how it fundamentally handles "compositionality", which other dissimilarity metrics (e.g. Bray-Curtis and JSD) cannot address. As described in "Microbiome Datasets Are Compositional: And This Is Not Optional" by Gloor et al.[4], since sequencing data is constrained by the number of "slots" in the sequencer, sequencing reads are non-independent random samples from a population and all that is effectively being captured during sequencing is the proportion (or composition) of these reads from the larger population. As a consequence, it is critical to treat sequencing data with compositional techniques, such as the Aitchison distance, which is the Euclidean distance between two compositions following centered log-ratio (clr) transformation (Equation 1).

As outlined in Gloor et al., ratio transformations between proportions capture the same relationship between counts of the same data and log transformations of these ratios result in symmetrically distributed and linearly related data[4]. While information about the true absolute abundance of taxa is lost during sequencing, ratio transforms provide a framework to compare all taxa to the same reference within a sample. The centered log-ratio in particular is scale-invariant, meaning that, in principle, this ratio will be the same regardless of the sample read depth. These properties allow for the proper use of standard statistical methods on clr-transformed data, ultimately making the Aitchison distance a more appropriate method for compositional data than other common alternatives.

1.3 Effect of pseudocount

As mentioned in Equation 1, a pseudocount is required to compute the Aitchison distance when non-overlapping taxa (0 counts) are present to allow for log-based calculations[4]. We added a pseudocount of 1 to all taxa in each sample, so that 0-count entries for a given sample, $\mathbf{x}$ became $\log[1/G(\mathbf{x})]$ following the clr-transformation. Since in real sequencing samples, $G(\mathbf{x})$ and most read counts $>> 1$, this transformed pseudocount value is much smaller than most of the real clr-transformed counts that it is being compared

to, and therefore does not introduce any systematic noise into our calculations. When computing divergence on our communities, we used the union of all taxa detected in our study for each calculation. As a result, some calculations included instances of computing differences between entries that were each 0-count before clr-transformation (0-0 pairs). Leaving 0-0 pairs in principle could alter results compared to removing them; however, we repeated all of our calculations in the manuscript with the alternative approach of removing 0-0 pairs, and we found that all results were nearly identical.

2 Divergence is sensitive to richness and evenness

As shown in **Figure 3**, we observed that communities with higher diversity diverge more from each other. We were interested in understanding whether this correlation should be expected merely based on (uniform) random chance or if this results from how species are distributed in natural communities. Here we demonstrate that the answer is both: our divergence metric (the Aitchison distance) does correlate with richness (i.e. the number of taxa) and, in addition, skewness (unevenly distributed taxa) causes communities of the same richness to diverge further. This notion supports the logic we depict in **Figure 4d** where the endemic (uneven) distribution of specialists contributes to the increased divergence experienced by communities in more complex conditions.

In order to test the relationships between richness, evenness, and divergence, we simulated the divergence between communities sampled from multinomial distributions[5]. With a multinomial distribution, we can generate simplified simulated "communities" with n total counts (akin to sequencing depth) over k taxa where each taxon i has probability p_i. When we tested the effect of richness on divergence, by sampling communities with increasing k under the null assumption that all taxa are equally likely ($p_i = 1/k$), we found that communities diverge more with increasing richness (**Supplementary Figure S1a**). However, we know that real communities have a substantially skewed distribution that deviates from the uniform one used under this null assumption[6]. To best capture the distribution of taxa in nature, we used the distribution of abundances from one of our source experimental communities with 300 observed taxa (**Supplementary Figure S1b**). Since our post-inoculation communities (from day 3 and onwards) have a richness an order of magnitude less than our source communities, the following analysis provides a conservative assessment of the effects of richness on divergence for our post-inoculation communities.

To understand how divergence is affected by the distribution of taxa, we generated fifty communities each from a multinomial distribution parameterized by the real taxonomic abundance distribution of our representative experimental community (**Supplementary Figure S1b**) and then calculated the divergence between all pairs of these generated communities. We repeated these two steps to generate another fifty communities from a uniform distribution with the same richness as our representative community, and compared the skewed and uniform divergence outcomes (**Supplementary Figure S1c**). We find that even though these calculations were performed on communities of the same size, the divergence of the skewed communities is significantly greater than those of the uniformly distributed ones, showing that skewness contributes to the signal we see in our experiment (**Figure 3-4**). For all of the previous analyses, we fixed $n = 8,564$, the "sequencing depth" of our representative experimental community.

Even upon taking into account the natural distribution of taxa, the divergence observed for randomly sampled communities falls short of the divergence experimentally measured across communities in our experiment and simulations. In our experiments, divergence reaches values of ∼20-30 for our post-inoculation communities (**Figure 2**), which is far greater than the mean of the distribution of the skewed distribution cases (∼8). Again, our experimental post-inoculation communities diverge this much even though they are an order of magnitude smaller than our source communities, suggesting that skewness provides a substantial contribution to divergence in these experimental communities compared to richness. These analyses reveal how both increased richness (which has shown to be correlated with metabolic complexity[7]) and the uneven distribution of taxa (as we show) contribute to divergence and how further differences between natural communities shape their divergence in our experiments.

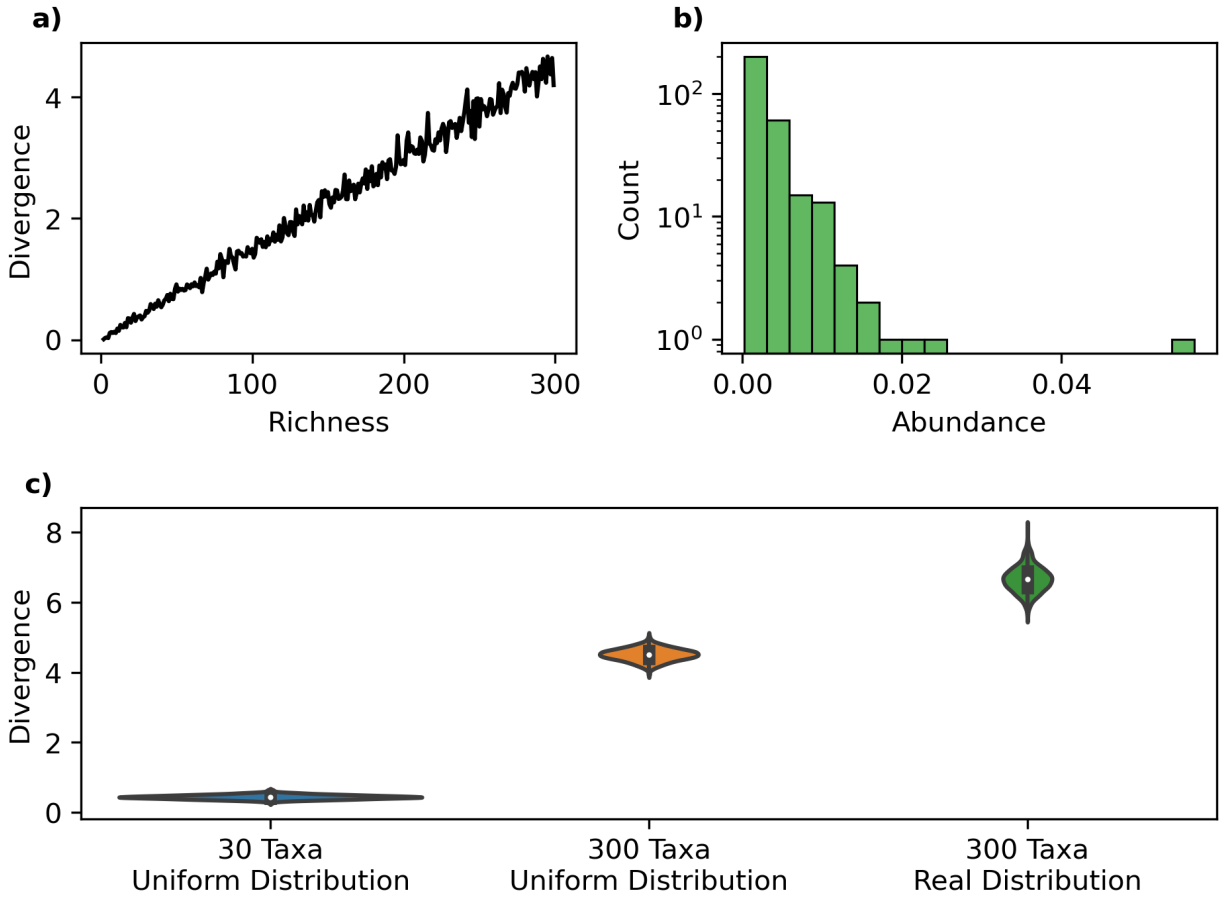

Figure 1: **Supplementary Figure S1 — Divergence is sensitive to richness and evenness. a)** The divergence (measured by the Aitchison distance) between two randomly sampled communities of uniformly likely taxa of increasing size (richness). Communities with higher richness diverge more than communities with less. **b)** The (uneven) distribution of taxa abundances from a real community in our study. **c)** The distribution of divergence between fifty simulated communities with (i) 30 uniformly distributed taxa, (ii) 300 uniformly distributed taxa, and (iii) 300 unevenly distributed taxa (following the distribution in **b**). Each violin describes the distribution of pairwise distances between all fifty simulated communities (N=1,255 per violin). Each violin outlines the kernel density estimate and contains a box which is bound by the interquartile range with an open circle at the media and whiskers that extend up to 1.5 times the interquartile range. The increase in divergence between (i) and (ii) shows how divergence is sensitive to richness and the increase in divergence between (ii) and (iii) shows how divergence is then further sensitive to skewness. Note that divergence between our real communities ranges from 20-30, which is far greater than in these simulated communities (**Figure 2**).

3 Generality of cross-feeding mechanism in our consumer-resource model

3.1 Consumer-resource model definition

The microbial consumer-resource model (CRM) is a valuable method[8–12] for simulating the dynamics of complex microbial communities. In this model, consumers uptake resources for growth and in the process leak some of the resulting transformed byproducts into the environment. In the text below we use the terms "leakage" to encompass different ways that microbially-transformed metabolites are made available to the

environment, including exudation, active and passive secretion, and extracellular degradation. Note also that the standard CRMs do not differentiate between costly (ATP-dependent) transport and free diffusion. Cross-feeding, where the byproducts of one population's activity become available to others for consumption[1]3, emerges from these dynamics. As described in our Methods, our CRM is directly adapted from previous work8 and is defined with the following system of equations,

$$\frac{dN_i}{dt} = N_i \left(\sum_{\alpha} (1-l) c_{i,\alpha} R_\alpha - m \right) \tag{4}$$

$$\frac{dR_\alpha}{dt} = \left(R_\alpha^0 - R_\alpha\right) - \sum_j N_j c_{j,\alpha} R_\alpha + \sum_{j,\beta} N_j c_{j,\beta} R_\beta D_{\alpha,\beta} l \tag{5}$$

Where N_i is the abundance of species i, R is the concentration of resource α, R_α^0 is the resource supply concentration, l is the leakage fraction i.e. how much each resource is "leaked" (how much of α is converted into β, where the rest is converted into biomass), m is the consumer maintenance cost, c_i, α,is the consumer preference matrix, and D_α, β, is the resource transformation matrix describing the rate that β turns into α following consumption. In our simulations, we parameterize the c and D matrices to represent the trophic structure of microbial communities, where groups of taxa consume types of resources (defined by a structured c matrix) and complex resources hierarchically transform into simpler resources following consumption (defined by a structured D matrix).

3.2 Representations of cross-feeding in CRMs

As mentioned in the Results and Discussion sections, an important limitation of this model is the simplified formulation of cross-feeding defined by a linear coupling of resource transformation, leakage, and growth.This is meant to capture in an idealized way the complexity of metabolism. While this model has been successful in reproducing various ecological phenomena[9,10], it does so at the expense of a detailed and accurate description of the many possible ways microbes may interact. The simplified view portrayed by a typical CRM encodes (through the D matrix) a series of possible transformations, implicitly assumed to be unimolecular intracellular reactions that convert a given metabolite into different ones, with a rate directly proportional to the uptake rate of the substrate[9,11]. We argue here that, despite the simplicity of the original interpretation, the CRM can be viewed as capturing, in an approximated way, a much broader set of alternative cross-feeding mechanisms. While these mechanisms can be very complex and dependent on environmental variables that are absent from the model definition, they may be expected to display on average a behavior that is consistent with the CRM formalism.

Cross-feeding can be mediated by a large assortment of molecules that can be produced, secreted, exchanged, imported, or extracellularly modified through a variety of mechanisms[13]. We focus below on three specific processes that are known to mediate metabolic cross-feeding: extracellular degradation, fermentation, and stress-induced cross-feeding. While the biochemical mechanisms underlying these processes are not explicitly described with the CRM formalism, their overall phenomenological outcome (i.e. consumption of an incoming metabolite and production of an outgoing resource) is reasonably captured by the standard metabolite transformation term ($N_j c_{j,\beta} R_\beta D_{\alpha,\beta} l$) of the CRM, where the net outcome can be approximately described with an outgoing byproduct production that is proportional to biomass and substrate amount. Therefore for the purpose of interpreting our data, we posit that the microbial consumer resource model with trophic structure can sufficiently approximate cross-feeding dynamics as measured in real complex communities.

Below, we examine how the metabolite transformation term, $N_j c_{j,\beta} R_\beta D_{\alpha,\beta} l$, relates to the availability of the leaked metabolites (dR_α/dt) in fermentation, extracellular degradation, and stress-induced cross-feeding and then summarize these details in **Supplementary Table 2**.

3.2.1 Fermentation

With fermentation, sugars imported by the cell for ATP production are not fully metabolized to CO_2 (as is the case with respiration), but rather result in the production and secretion of ("simpler") organic acid byproducts ($D_{\alpha,\beta}$) which can then be secreted and made available to other community members. Different

organisms are capable of fermenting different sugars ($c_{j,\beta}$)[14] resulting in a diversity of byproducts. The net amount of byproducts secreted during fermentation (l, R_α)[15,16] and the amount of microbial biomass (N_i)[17] generally correlate with the amount of sugar present in the environment (R_β). Therefore we expect that secretions due to fermentation and the subsequent cross-feeding that emerges are phenomenologically captured as a potential avenue for metabolic interdependence in our model.

We should note that the extent of fermentation vs. respiration often depends on the availability of oxygen, a metabolic detail that is missing in CRMs. For the purpose of the present work, it is reasonable to assume that oxygen is somehow limiting both in our experimental setup and in the natural soil from which the microbial samples are extracted, and that a certain degree of fermentation is pervasive in the communities[18]. We don't expect the details of which microbes secrete what specific byproduct under a certain level of oxygen to be relevant for the conclusions of our work, but this is still a fundamental limitation of CRM that may be addressed in future models.

3.2.2 Extracellular degradation

Bacteria produce a diverse set of extracellular enzymes[19,20] to degrade polymers (e.g. cellulose) that are too large to be directly imported. The extracellular activity of these enzymes, which are often attached to, or found in the vicinity of the producing cell, lead to the environmental release (l) of simpler products (e.g. monosaccharides), which constitute common goods, available for import and utilization by surrounding microbes ($c_{j,\beta}$)[13]. Significant portions of the liberated byproducts can be utilized by individual taxa that do not participate in the production of relevant extracellular enzymes (including "cheaters", which consume byproducts without contributing to enzyme production). The effective massive loss of substrate (i.e. the fact that cellulose is not directly used by the enzyme producer) can be encoded in a large l parameter of our model. One can view this process as an overall transformation of a given compound (e.g., cellulose) by a given organism into a simpler product (e.g. glucose) usable by other organisms, hence giving rise to a particular case of cross-feeding.

In our model, for this specific subset of metabolic processes, $c_{j,beta}$ can be interpreted as representing the extent to which organism j can produce extracellular enzymes that degrade molecule β, $D_{\alpha,\beta}$ represents the extent to which complex polymers β is transformed into simpler molecule α, l represents the fraction of degraded byproducts that are available to other community members, m represents the cost of producing extracellular enzymes, growth is proportional to the amount of consumed resource, and cross-feeding is proportional to the size of the degrader population and the leakage fraction.

Note that in our implementation of the CRM we do not explicitly distinguish between intracellular and extracellular catabolism, but rather encodes an overall hierarchy of molecular structures. What matters for the purpose of our analysis is that the diversity of cross-feeding processes (intracellular and extracellular) likely occurring in our experimental communities can be captured simultaneously in an approximated way by our CRM.

3.2.3 Stress-induced

While cross-feeding is typically thought of as a process associated with metabolism during active (typically exponential) growth (ex. organic acid secretion in fermentation and the availability of extracellular degradation byproducts), it can also emerge as a means for non-growing cells to modify the environment. One example is acid stress-induced cross-feeding, where a growth-arrested population experiencing acid stress consumes the acid in its environment and then secretes many of the resulting central carbon metabolic intermediates without growing[21]. In particular, one study found that one organism facing acid stress converted acetate into other simple metabolites such as pyruvate, lactate, and glutamate as secreted byproducts[21]. In this way, the stressed population detoxifies its environment and a significant amount of cross-feeding emerges without the consumer harnessing energy or performing biosynthesis.

Despite the complex and subtle mechanisms involved in this stress-induced cross-feeding phenomenon, it is possible to think of our model as encoding in an approximate way the overall net transformation of environmental metabolites. Our model represents the consumption of specific acids with $c_{j,\beta}$, the transformation into byproducts with $D_{\alpha,\beta}$ with leakage l (which can approach 100% of the incoming substrate if cells are not growing at all) that is proportional to population size N_i even if these populations are not actively growing. We have constructed our $D_{\alpha,\beta}$ matrix such that most of a resource is transformed into a simpler

molecule, but a significant transformation among molecules of comparable complexity (e.g. acetate to lactate, as seen in the above example) is also allowed (see off-diagonal components of D matrix in **Extended Data Fig. 7a**). While the detoxifying population does not grow while secreting byproducts, over time larger populations will secrete more byproducts than smaller ones, so we would expect that our model still captures the proportional relationship between resource transformation, leakage, and population size (N_i) under stress-induced cross-feeding conditions. An important difference between stress-induced cross-feeding and the other mechanisms we describe is that in the stress-induced case, secretions may specifically occur only once the cells stop growing - a feature that is not currently captured by the CRM.

As with the other categories of cross-feeding mechanisms outlined above, our CRM does not explicitly distinguish between specific mechanisms, but rather serves as a generalized model that broadly captures each of these possibilities.

Variable	**Fermentation**	**Extracellular**	**Stress-induced**
c	Sugar preference	Polymer preference	Acid preference
D	Feasible transformations into organic acids	Possible degradation byproducts	Producible metabolic intermediates
l	Degree of fermentation byproduct leakage	Amount of extracellular degradation products made available	Amount of byproduct secretion

Table 2: **Supplementary Table 2 — Representations of cross-feeding mechanisms in the microbial consumer resource model.**

4 References

1. Aitchison, J., Barcelo-Vidal, C., Martın-Fernandez, J. A. & Pawlowsky-Glahn, V. Logratio Analysis and Compositional Distance. Math. Geol. (2000).

2. Bray, J. R. & Curtis, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. Ecol. Monogr. 27, 325–349 (1957).

3. Lin, J. Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory 37, 145–151 (1991).

4. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. Front. Microbiol. 8, 2224 (2017). 5. Egozcue, J. J., Graffelman, J., Ortego, M. I. & Pawlowsky-Glahn, V. Some thoughts on counts in sequencing studies. NAR Genomics Bioinforma. 2, (2020).

6. Grilli, J. Macroecological laws describe variation and diversity in microbial communities. Nat. Commun. 11, 4743 (2020).

7. Dal Bello, M., Lee, H., Goyal, A. & Gore, J. Resource–diversity relationships in bacterial communities reflect the network structure of microbial metabolism. Nat. Ecol. Evol. 5, 1424–1434 (2021).

8. Marsland, R. et al. Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities. PLOS Comput. Biol. 15, e1006793 (2019).

9. Marsland, R., Cui, W. & Mehta, P. A minimal model for microbial biodiversity can reproduce experimentally observed ecological patterns. Sci. Rep. 10, 3308 (2020).

10. Cui, W., Marsland, R. & Mehta, P. Diverse communities behave like typical random ecosystems. Phys. Rev. E 104, 034416 (2021).

11. Goldford, J. E. et al. Emergent simplicity in microbial community assembly. 7 (2018). 12. Diaz-Colunga, J. et al. Top-down and bottom-up cohesiveness in microbial community coalescence. Proc. Natl. Acad. Sci. 119, e2111261119 (2022).

13. Fritts, R. K., McCully, A. L. & McKinlay, J. B. Extracellular Metabolism Sets the Table for Microbial Cross-Feeding. Microbiol. Mol. Biol. Rev. 85, (2021).

14. Palframan, R. J., Gibson, G. R., Rastall, R. A. & Vriers, D. Carbohydrate Preferences of Bifidobacterium Species-Isolated from the Human Gut.

15. Zentou, H. et al. A New Model of Alcoholic Fermentation under a Byproduct Inhibitory Effect. ACS Omega 6, 4137–4146 (2021).

16. Grau, F. H. End Products of Glucose Fermentation by Brochothrix thermosphacta. Appl. Environ. Microbiol. 45, 84–90 (1983).

17. Reischke, S., Rousk, J. & Bååth, E. The effects of glucose loading rates on bacterial and fungal growth in soil. Soil Biol. Biochem. 70, 88–95 (2014). 18. Tiedje, J., Sexstone, A., Parkin, T. & Revsbech, N. Anaerobic processes in soil. Plant Soil 76, 197–212 (1984).

19. Zimmerman, A. E., Martiny, A. C. & Allison, S. D. Microdiversity of extracellular enzyme genes among sequenced prokaryotic genomes. ISME J. 7, 1187–1199 (2013).

20. Berlemont, R. & Martiny, A. C. Phylogenetic Distribution of Potential Cellulases in Bacteria. Appl. Environ. Microbiol. 79, 1545–1554 (2013).

21. Amarnath, K. et al. Stress-induced metabolic exchanges between complementary bacterial types underly a dynamic mechanism of inter-species stress resistance. Nat. Commun. 14, 3165 (2023).