iScience

CellPress
OPEN ACCESS

Perspective

An ensemble approach to the structure-function problem in microbial communities

Chandana Gopalakrishnappa,[1,4] Karna Gowda,[2,3,4] Kaumudi H. Prabhakara,[2,3,4] and Seppe Kuehn[2,3,*]

SUMMARY

The metabolic activity of microbial communities plays a primary role in the flow of essential nutrients throughout the biosphere. Molecular genetics has revealed the metabolic pathways that model organisms utilize to generate energy and biomass, but we understand little about how the metabolism of diverse, natural communities emerges from the collective action of its constituents. We propose that quantifying and mapping metabolic fluxes to sequencing measurements of genomic, taxonomic, or transcriptional variation across an ensemble of diverse communities, either in the laboratory or in the wild, can reveal low-dimensional descriptions of community structure that can explain or predict their emergent metabolic activity. We survey the types of communities for which this approach might be best suited, review the analytical techniques available for quantifying metabolite fluxes in communities, and discuss what types of data analysis approaches might be lucrative for learning the structure-function mapping in communities from these data.

INTRODUCTION

The structure-function problem in microbial communities

The evolutionary history of the biosphere is inextricably linked to the metabolic activities of microbes. Since life arose on this planet, microbes have lived in consortia that saturated nearly every biochemical niche on the planet, driving global transformations in the chemical composition of the biosphere via metabolic processes from fermentation to photosynthesis to respiration (Falkowski et al., 2000; Canfield et al., 2010; Nelson et al., 2016; Sunagawa et al., 2015; Zakem et al., 2020). As such, microbes and the communities in which they reside are the result of an ongoing eco-evolutionary process that couples the transformation of metabolites to the complex dynamics of interacting ecological systems across many spatial and tempo-ral scales.

Given the importance of the metabolic activity of microbial communities, we argue that a major goal for the field should be to predict, design, and control the metabolism of microbial communities in complex, natural, and engineered settings. Accomplishing this goal requires understanding how the structure of a community, in terms of the taxa present and its genomic composition, determines its metabolic activity in a given environmental context. The sequencing revolution has revealed the structure of microbial communities at the level of the taxa present, the genes they possess, and the dynamics of gene expression. This means that we now have a detailed and dynamic "parts list" for microbial communities in terms of taxo-nomic and genomic composition across a range of environments, from anaerobic digesters (Bocher et al., 2015; Toerien and Hattingh, 1969; Vanwonterghem et al., 2014) to the human gut (Blanton et al., 2016; Raman et al., 2019), soils (Bahram et al., 2018), and the ocean (Sunagawa et al., 2015). For some metabolic processes, we can interpret gene content and taxa in terms of the specific metabolic processes that they are capable of. For example, we know the dominant taxa that perform processes such as nitrification (Bock and Wagner, 2013) or polysaccharide degradation (Sanchez-Gorostiaga et al., 2019). Further, by anno-tating metagenomic data, we can assign specific functional roles for many (but not all) of the genes present in a given community. As a result, we can measure the prevalence of enzymes that perform the reactions necessary for specific metabolic processes. Despite the remarkable scale and breadth of these sequencing

[1]Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[2]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637, USA

[3]Center for the Physics of Evolving Systems, University of Chicago, Chicago, IL 60637, USA

[4]These authors contributed equally

*Correspondence: seppe.kuehn@gmail.com

https://doi.org/10.1016/j.isci.2022.103761

data, we still do not have a predictive, quantitative framework for using these data to understand, predict, and design the metabolism of the communities in complex environments.

In this perspective, we explore what makes this problem both challenging and important. We propose a specific approach to begin to address this question, and examine what types of communities and associated metabolic processes might be amenable to this approach. We review the techniques that are relevant to implementing the approach with a focus on methods for quantifying metabolites.

The significance of finding a solution

Microbial communities play an outsized role in driving fluxes of nutrients through the biosphere. Photosynthetic microbes are responsible for nearly half of the carbon fixation on the planet (Falkowski, 1994). These phototrophs work in concert with heterotrophic bacteria that enable primary productivity in terrestrial, marine, and freshwater environments (Kirchman, 2012; Madigan et al., 2018). We are only beginning to glimpse the role of the collective in this nearly 100 gigaton annual carbon flux. Bacteria and archaea in anaerobic environments degrade complex carbon sources to methane, playing an important role in carbon recycling and climate change (Madigan et al., 2018).

In the nitrogen cycle, microbes play a key role in nitrogen fixation (dinitrogen gas to ammonia), nitrification (ammonia to nitrate), and denitrification (nitrate to dinitrogen gas) (Stein and Klotz, 2016). These processes are key for wastewater treatment (Cydzik-Kwiatkowska and Zielinska, 2016) and human health (Turnbaugh et al., 2007). A critical challenge is to form a quantitative and predictive understanding of how microbial communities drive these fluxes. To give a concrete example, the process of denitrification, performed by bacterial communities in soils, reduces nitrate to dinitrogen gas. An intermediate in the conversion of nitrate to dinitrogen is the potent greenhouse gas and ozone depleting compound nitrous oxide (Tian et al., 2020). Denitrifying communities in some cases (especially in agricultural soils) can leak nitrous oxide, but in other cases fully convert nitrous oxide to harmless dinitrogen gas. The question then becomes: What controls the production of nitrous oxide from denitrifying communities in soils? Can we manipulate these microbial communities to limit nitrous oxide production? To address this, we need to understand how the structure of the community and the environmental context determine the flux of metabolites through the system.

Similarly, the essential importance of resident microbiota in host health is now clear (Turnbaugh et al., 2007), but as yet, it is unclear how to rationally manipulate these communities to benefit the host. There exists tantalizing evidence that this can be done, for example, by altering metabolic phenotypes (Turnbaugh et al., 2006) or treating persistent infections (Lawley et al., 2012), but we lack general approaches for developing such strategies. Here we focus on environmental microbiomes, but we emphasize that the strategy proposed here could, and in a few cases has been (Raman et al., 2019), applied to host-asso-ciated communities.

Defining structure and function

Before moving forward, we take a moment to define community structure and function. We define the structure of the community as the taxa present as well as the genomic structure of each taxon, which may include everything from the detailed knowledge of the regulatory architecture of each gene, to the syntenic organization of the genome (Junier et al., 2018), to even the presence of phage. The structure of the community may, if necessary, include transcriptional or proteomic information at the metagenomic or single-taxon level as well.

We define the function of a community as the collective metabolic activity of all constituent organisms, which, therefore, operates in the space of metabolites. The dynamic or steady-state flux of metabolites through the consortium defines its metabolic function. Depending on the context, the most important metabolic fluxes may include electron donors (e.g., organic carbon), electron acceptors (e.g., oxygen, nitrate), secondary metabolites, biomass, overall catabolic activity, or byproducts. A note about usage: Some readers may find the term "function" teleological, implying some sort of purpose on microbial communities. We use the term function to mean the activity or action of a community without any implication of purpose. Despite this potential confusion, we find that the term function is a useful shorthand. In particular, we would like to invoke a certain symmetry between the ideas presented here and the problem of sequence, structure, and function at the level of proteins. The structure-function problem for microbial

communities is therefore to deduce the mapping from the space of genes, transcripts, proteins, and taxonomic organization to metabolite fluxes, and to understand the environmental dependence and context in which this mapping is relevant.

How should we approach the problem?

Understanding how the metabolic activity of a microbial community emerges from the taxa present and their metabolic capabilities is a problem of connecting hierarchical scales of biological organization, from genes to phenotypes and interactions in specific environmental contexts. What makes this challenging is the fact that processes at different scales feedback on one another. For example, compounds that mediate interactions between strains can do so by modulating gene expression (Beliaev et al., 2014). Similarly, phenotypic variation in individuals, determined by gene expression and interactions, can modify community interactions (Mickalide and Kuehn, 2019) and the chemical environment, with widespread impacts on other members of the consortium (Ratzke et al., 2018).

One way to proceed is via the reductionist mode that has motivated biology over the last century (Woese, 2004). In the context of communities, this would mean dissecting the mechanistic and physiological metabolic properties of each member of the community and understanding how metabolite dynamics emerge at the level of the collective. The challenge is the complexity of these systems, which makes a detailed, mechanistic understanding of the collective metabolism a massive undertaking. For synthetic communities comprising model organisms, where detailed information is available, this type of approach has had some success (Orth et al., 2010; Harcombe et al., 2014). For comparatively simple communities of a few strains, such as bacteria cross-feeding amino acids (Wintermute and Silver, 2010) or ciliates consuming bacteria (Mickalide and Kuehn, 2019), detailed models have been constructed and validated. However, in environmental or host-associated contexts, where the number of strains present are enormous and many organisms present are poorly studied or challenging to work with in the laboratory, this approach faces huge challenges.

In this scenario, what are we to do? On the one hand, building detailed models as described above is ill-advised. Even when such detailed models can be built, the challenge of distilling simple principles from these models increases with their complexity (see Borges' ''On Exactitude in Science'' (Borges and Hurley, 1998)). On the other hand, we know from many examples that a huge number of processes from antibiotic warfare (Vetsigian et al., 2011) to competition (Friedman et al., 2017), mutualism (Hom and Murray, 2014), and stress responses (Amarnath et al., 2021) influence interactions and metabolism. So, how can we justify not building models that include these details?

In the spirit of Philip Anderson's influential essay (Anderson, 1972), it may be that understanding communities requires explaining entirely new and potentially simpler properties that are emergent at the level of the collective. In this case, due to the scale and complexity of the community, new and distinct phenomena emerge from the individual parts, and discovering the organizing principles requires an approach that goes beyond dissecting the detailed metabolic phenotypes of each individual member of the community. To be clear, we are not advocating the idea that reductionist approaches are not useful. Their utility is clear from many examples (Jacob and Monod, 1961; Zumft, 1997; Alon et al., 1999; Basan et al., 2020; Amarnath et al., 2021). Instead, we are asking whether making headway on the ''structure-function problem'', as we have defined it, might require a complementary approach that focuses not on the detailed mechanisms and physiology of each organism, but on patterns that are evident at the level of the collective. Here we discuss such an approach.

Learning the right variables: the power of statistics across ensembles

We find it useful to cast the structure-function question in terms of a prediction problem. In this framing, the goal is to predict metabolite dynamics or fluxes from community structure for a given environment or set of environmental conditions. The key question then becomes: What structural elements must we quantify to predict function? Equivalently, we could ask, how predictive of metabolic function is community gene content, taxonomy, transcriptional or proteomic data?

One approach to this problem, which has found success in both physics and biology, is to look for statistical regularities across replicate systems and allow these patterns to naturally define variables that can be used

to make predictions. In many cases, this approach can reveal the salient, emergent features of a system and provide deep insight into their function.

Specifically, from proteins (Halabi et al., 2009) to multicellular organisms (Alba et al., 2021), examining statistical variation across many replicates of a system, or an ensemble, has proven powerful. For example, covariation across ensembles of homologous proteins have been used to reveal which amino acids are in contact in the folded structure (Russ et al., 2005), and co-evolving groups of residues that correlate with enzyme function (Halabi et al., 2009). Careful analysis of behavioral variation in large numbers of mi-crobes (Jordan et al., 2013) and flies (Berman et al., 2014) suggests that apparently very complex behaviors can, in fact, be described by a relatively small number of elementary behavioral features (Berman et al., 2014; Katsov et al., 2017). Statistical analysis of morphological variation across higher organisms suggests that morphologies adhere to constraints that some believe are associated with specific functional capabil-ities (Raup and Michelson, 1965; Shoval et al., 2012). A recent study of variation in patterning in the fly wing shows that a single mode of variation describes the response of the wing developmental program to ge-netic and environmental perturbations (Alba et al., 2021). For a recent piece discussing why low-dimension-ality might be an inherent property of evolved systems, see Eckmann and Tlusty (2021).

The common feature of all of these examples is that, by judiciously studying the variation across a carefully chosen ensemble of systems, one can often discover simple, relatively low-dimensional features that enable the prediction of the functional properties of the system. Here we advocate a similar approach to communities.

An ensemble approach to the structure-function problem

In light of these considerations, we propose an ensemble approach to the structure-function problem in microbial communities. We motivate this approach by analogy to a similar approach taken at the level of proteins (Figure 1). For proteins, the structure-function problem is to predict the fold and function of a protein from its amino acid sequence. One way to approach this problem is by performing detailed phys-icochemical simulations of a polypeptide chain via molecular dynamics (Figure 1A). While much progress has been made in this approach, it has proven a huge technical challenge. However, a statistical approach, ignoring the mechanistic details and instead considering only statistics of a multiple sequence alignment does a remarkably good job of predicting protein folds (Morcos et al., 2011). Similar approaches reveal low-dimensional structure in proteins which is predictive of function (Halabi et al., 2009; Russ et al., 2020). These studies suggest that much can be learned by carefully considering variation across a suitably chosen ensemble of systems.

In the context of microbial communities, flux balance models of community metabolism are analogous to molecular dynamics simulations in proteins because they attempt a detailed mechanistic accounting of all of the phenomena within a community (Figure 1B). However, in communities, there is comparatively little work pursuing a statistical approach analogous to the one taken at the level of proteins.

Taking such an approach is precisely what we are advocating here. We propose quantifying the structure of a collection, or ensemble, of communities using sequencing while simultaneously measuring metabolite dynamics. Given such data, one can then approach the structure-function problem by asking whether variation in community structure (e.g., across metagenomes or metatranscriptomes) permits quantitative insights into the functional properties of these communities. The proposal is then to leverage these insights to design, predict, and control community function. Several studies in the past few years have begun to explore statistical approaches (Gowda et al., 2022; Raman et al., 2019). However, we suggest that this approach is under-explored and that there is a pressing need to collect new datasets that are explicitly designed to pursue a statistical approach to the structure-function problem in communities. Moreover, while we focus on community metabolic function, the approach we propose could just as easily be applied to other salient features of ecosystems from spatial structure to resilience to stability.

Overview of the paper

We will focus on three main challenges that must be surmounted to apply the ensemble approach to the structure-function problem: (1) Choosing an ensemble across which one should make comparisons and look for patterns, (2) measuring metabolite dynamics, which requires analytical chemistry techniques

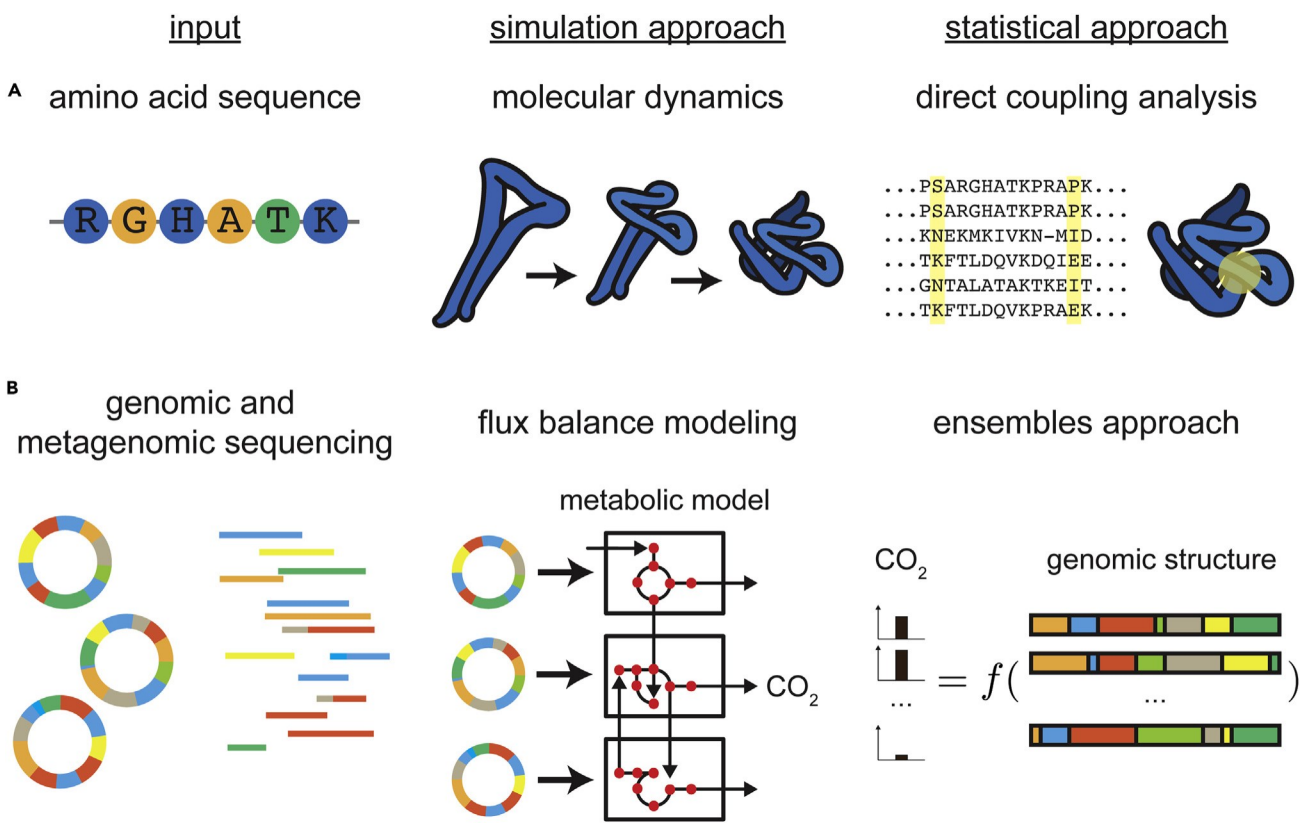

Figure 1. Sequence, structure and function in proteins and microbial communities

We propose that there exist analogous solutions to the sequence-structure problem in protein folding and the structure-function problem in microbial communities.

(A) The mapping from amino acid sequence to 3D protein structure can be accomplished either by a simulation approach (e.g., molecular dynamics) or by a statistical approach (e.g., direct coupling analysis). The former is a computationally intensive strategy to simulate 3D protein structure based on first-principles modeling of atomic interactions. The latter leverages information about residue coevolution from an ensemble of amino acid sequences to infer which residues are in contact, allowing for an elegant and interpretable statistical inference of 3D structure.

(B) The mapping from genomic and metagenomic sequences to community metabolic activity can be achieved through community flux balance modeling or, as we propose, a statistical ensembles approach. The former requires genome-level metabolic models of each organism to be built, a labor-intensive iterative process that so far has been successful primarily in a handful of model organisms. The latter leverages the diversity and variation in an ensemble of communities to learn an effective mapping between community sequence content metabolic activity

that are often not standard practice in microbial ecology labs, and (3) using these data to distill the mapping from structure to function.

Our intention is to provide a roadmap for how such an approach might be applied across communities of interest. We recognize that this roadmap is far from complete and that many pitfalls exist that may render this approach challenging in various circumstances.

We will not review sequencing technologies, which have been widely and capably recapped elsewhere (Knight et al., 2018). We will focus largely on microbial communities in environmental contexts rather than health-related (human microbiome) contexts, in part because environmental microbiology is where our expertise lies, but also because of the abundance of existing literature on the latter topic. Finally, we will neglect the many recent advances in theoretical ecology, in particular the renaissance in con-sumer-resource models applied to communities, in service of focusing on questions that can be settled empirically.

MODEL MICROBIAL COMMUNITIES

The first challenge is choosing a community and an associated metabolic process to interrogate. The choice is far from trivial and there is no simple prescription. Instead we appeal to the intuition of microbial ecologists, experts in physiology and the quantitative considerations of applied mathematics and physics. The goal should be to circumscribe a well-defined community and, if possible, an associated metabolic process where quantitative measurements in many replicates can be made.

Choosing a metabolic process, and therefore specific metabolite measurements to be made, is a significant challenge, and compromises are inevitable. Here we reiterate the point that microbial communities are not "functional" in the sense that they are designed with some purpose. Despite this fact, as we will discuss below, the importance of metabolic activity of communities in specific niches is clear, as is the relevance of these processes for the biosphere more broadly. To make this point clear, we begin by discussing specific communities and their associated environments and metabolic processes, with an eye toward how one might apply the ensemble approach to dissecting their function.

Structure-function in the wild

Soils

Perhaps no microbial community on Earth is more important than that which inhabits soils. The soil microbiome plays a key role in plant growth and physiology (Saleem et al., 2019), in particular through nitrogen fixing bacteria that provide reduced nitrogen for plant hosts. The storage of carbon and production of CO_2 via respiration of reduced organic compounds in soils are key components of the global carbon cycle (Lal, 2004). As the climate warms, the rate of microbial respiration of CO_2 from soils increases (Kirschbaum, 1995; Allison et al., 2010), potentially driving a positive feedback loop with dire consequences for the global climate. Similarly, denitrification in agricultural soils is responsible for roughly 80% of the anthropogenic release of nitrous oxide (N_2O) (see https://www.epa.gov/ghgemissions/overview-greenhouse-gases and (Tian et al., 2020)). Nitrous oxide is 300-fold more potent than CO_2 as a greenhouse gas and responsible for approximately 10% of the global warming potential from human activity. For these reasons, there is keen interest in associating soil microbiome structure to process rates such as CO_2 or N_2O production and N_2 fixation. However, most attempts to find a relationship between soil microbiome structure and the rates of key metabolic processes in soils have found only marginal success (Graham et al., 2016; Rillig et al., 2019; Rocca et al., 2015; Fierer, 2017).

Despite these difficulties, we see reason for optimism. It is known that a relatively small number of environmental factors are the dominant drivers of variation in soil community structure: pH, moisture, carbon and nitrogen availability, temperature, and redox potential (Fierer, 2017). Moreover, while soils are routinely cited as very complex microbial communities, much like the human gut, they are typically dominated by a handful of taxa (e.g., Acidobacteria, Verrucomicrobia (Fierer, 2017; Crits-Christoph et al., 2018)), with most other strains present in relatively low abundances. Moreover, there are clear patterns in the abundances of bacteria and fungi in soils, with high biomass turnover environments such as grasslands dominated by bacteria and low biomass turnover forests dominated by fungi (Fierer, 2017).

As acknowledged in recent meta-studies (Graham et al., 2016), one challenge in associating soil community structure to metabolic function is a lack of high quality datasets where process rates (e.g., CO_2 production) are measured in a large ensemble of soil communities. Exceptions to this include a recent survey of global topsoil microbiomes (Bahram et al., 2018), and microcosm studies documenting the role of multiple environmental perturbations applied to soils (Rillig et al., 2019). Despite these advances, the consensus remains that predicting metabolic processes in soils from microbial community structure is challenging (Fierer, 2017). However, we note that in some cases this conclusion is derived from meta-studies that aggregate data from different experiments or labs. Such comparisons can be challenging given systematic errors in sequencing measurements between protocols (McLaren et al., 2019), and the strong dependence of measurements such as soil pH on the technique employed (Miller and Kissel, 2010).

Given these considerations, we propose that one route forward is the judicious collection of data from large ensembles of soil communities followed by careful quantification of process rates and community structure in a consistent manner. We acknowledge that even in the presence of such data, relating community composition at the taxonomic, genomic or transcriptomic levels to metabolite fluxes may remain a challenge. It may be that soil taxa are not "good variables" for understanding function, and that chemical

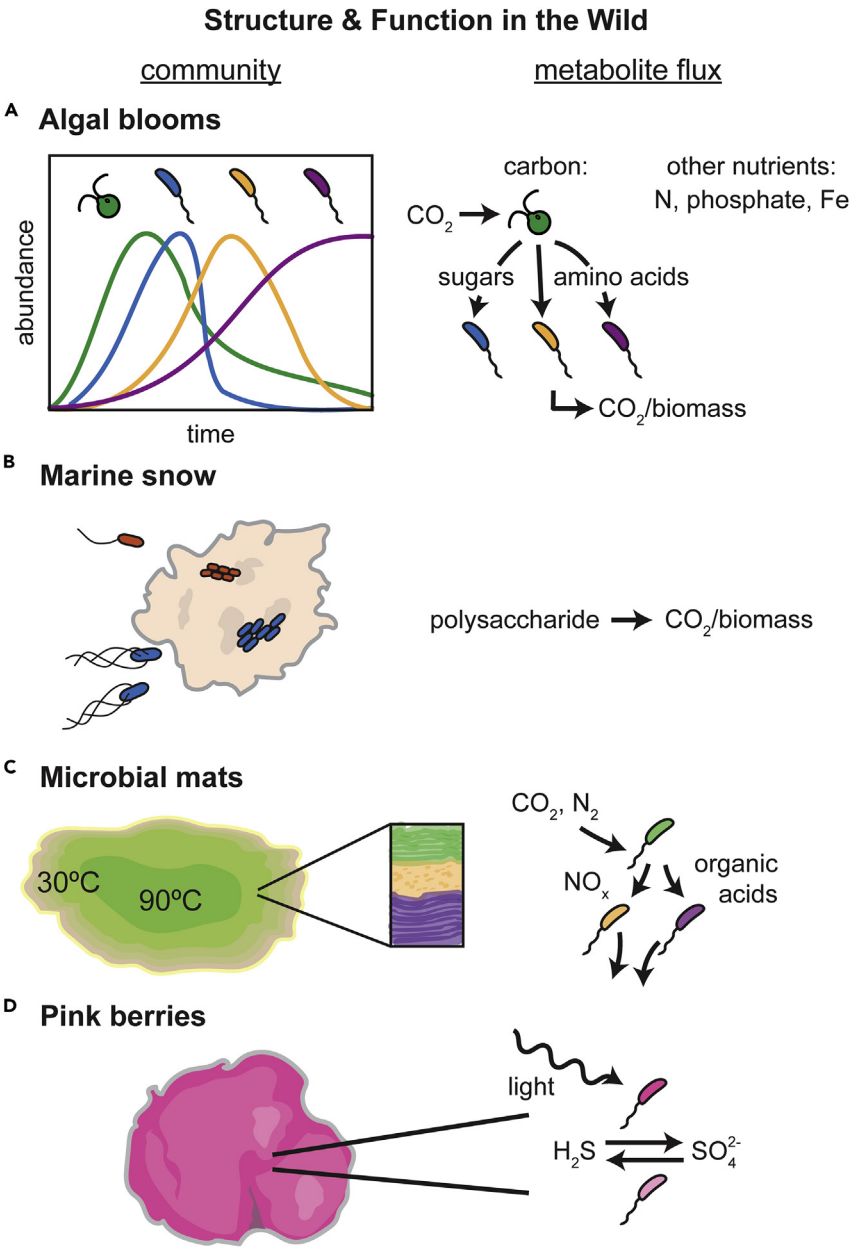

Figure 2. Community structure and function in the wild

(A) Algal blooms are microbial successional processes that follow from the input of exogenous nutrients to aquatic environments. Reduced carbon fixed from CO_2 by algae is consumed along with other nutrients by heterotrophic bacteria in reproducible successional dynamics.

(B) Marine snow particles are aggregates of organic carbon that are formed near the ocean's surface and subsequently sink to the ocean floor. Microbial communities can degrade these particles, and the amount of carbon that is mineralized to CO_2 versus the amount that is sequestered on the ocean floor depends strongly on the structure of the microbial community.

(C) Microbial mats are layered communities that occur at air-water interfaces, often in extreme thermal environments such as hot springs. The spatial structure of these communities follows from exchanges of nutrients governed by redox gradients.

(D) Pink berries are microbial aggregates that cryptically (internally) cycle sulfur between photosynthetic purple sulfur bacteria and anaerobic sulfate reducing bacteria

or abiotic properties such as the redox state of available organic carbon are necessary (Keiluweit et al., 2017). Regardless, the importance of understanding the metabolic activities of soil communities cannot be overstated.

Algal blooms

In aquatic systems, exogenous inputs of nitrogen and phosphorous, often driven by human activity, result in the dramatic successional process of the algal bloom (Teeling et al., 2012) (Figure 2A), wherein photosynthetic microbes explode in abundance. The rise of phototrophic microbes brings with it a complex community of non-photosynthetic (heterotrophic) bacteria that form tight symbioses with the phototrophs. These phototroph-heterotroph communities cycle large quantities of carbon, with phototrophs fixing CO_2 to reduced organic carbon, which is in turn consumed by the associated heterotrophic bacteria.

In this system, a natural framing of the structure-function problem would be to ask how the taxonomic composition of the community impacts the fixation of carbon and eventual bacterial biomass production. Teeling et al. (2012) have found that specific classes of bacteria grow at different phases of the bloom, e.g., Alphaproteobacteria dominated the pre-bloom; as blooming commenced, Bacteroidetes increased more rapidly than others, and then Gammaproteobacteria grew much later. Concurrent metagenomic studies showed that the abundances of enzymes capable of degrading carbohydrates and sulfatases required to degrade sulfated algal polysaccharides increased in abundance during the bacterial succession. The degradation of the larger polysaccharides leads to the production of shorter organic compounds, which was revealed by the expression of the relevant transporters. These results, and those of other studies (Kimbrel et al., 2019; Louati et al., 2015; McFeters et al., 1978; Parulekar et al., 2017; Ramanan et al., 2015), indicate a link between taxonomy of bacteria associated with certain photosynthetic strains during blooming events.

These results suggest that these associations are at least in part determined by the carbon catabolic activity of the bacteria and the fixed carbon that phototrophs excrete (see also (Buchan et al., 2014)). Taking an ensemble approach to this problem could be accomplished by recapitulating bloom dynamics in a laboratory context (Riemann et al., 2000), where the identity of the phototroph and the composition of the bacterial community could be manipulated and the resulting total carbon flux quantified (de Jesus´Astacio et al., 2021). Such studies might shed insights on how to control blooms in natural settings.

Marine snow

Marine snow is a term used to describe micrometer- to millimeter-scale aggregates in the ocean, which are typically made of detritus containing carbon and nitrogen in the form of polysaccharides, microbes, and inorganic substances. These particles form at the upper levels of the ocean and sink to the ocean floor, transporting carbon in a "pump" that removes carbon from the atmosphere over geologic timescales (Alldredge and Silver, 1988; Gralka et al., 2020). Communities of bacteria and other microbes consume these aggregates as they sink (Figure 2B). Understanding how the structure of marine snow communities impacts the degradation of carbon and the production of metabolic byproducts (e.g., CO_2) is a critical question for understanding global carbon fluxes.

Studying these particles in situ is challenging (Kiørboe, 2007). To circumvent this difficulty, some studies use synthetic particles to study community assembly and function. One study used agar beads (Kiørboe et al., 2003) to study the colonization by microbes in raw seawater, revealing successional dynamics underlying particle degradation. The particles were initially colonized by bacteria, and later by flagellates. Cordero et al. have taken a similar approach to understand colonization dynamics (Datta et al., 2016; Ebrahimi et al., 2019; Enke et al., 2019; Pontrelli et al., 2021). These studies shed light on the metabolic roles played by different players during particle colonization dynamics: primary degraders excrete enzymes to break down polysaccharide chains, which creates a niche for secondary degraders capable of consuming mono-mers and oligomers of the polysaccharide, which in turn makes way for scavengers that take up metabolic byproducts of the primary and secondary degraders (ammonia, amino acids).

In terms of the approach proposed here for mapping structure to function, particle degradation offers a powerful model system because so much is known about how the communities collectively degrade the particle. For example, it would be interesting to see whether one could take a statistical approach to inferring the key metabolic traits of primary and secondary degraders and scavengers from sequencing data

alone. In this case the function of the community to predict would be the fraction of carbon degraded or the total CO_2 respired. The power of this system is the ability to manipulate structure, but the particles make it challenging to measure carbon degradation directly (see the next section for further discussion).

Plastisphere

Closely related to the degradation of polysaccharide particles in the ocean is the recent rise of plastic debris in freshwater and marine environments, and the microbial communities associated with its degradation (Amaral-Zettler et al., 2020). Given the fact that a few million tons of plastic enter the ocean per year (Jambeck et al., 2015), this is an important ecosystem to study, not only to understand how these pollutants affect the ecosystem, but also to find potentially efficient plastic degrading communities. Further, plastic is a comparatively new environment on evolutionary timescales, making it interesting to study from the perspective of evolution. It has been shown that the community composition on plastic is different from the composition on other substances in the same conditions, and that certain taxa are commonly found on plastics (Dussud et al., 2018; Kirstein et al., 2019). Diatoms have been observed in high numbers, although strong succession dynamics were observed. Other photoheterotrophs, heterotrophs, ciliates, fungi, and pathogens have also been observed. Although the functional role of these microbes is unclear, it is speculated that chemotaxis, interactions with metals and degradation of low molecular weight polymers are important factors that determine community composition (Amaral-Zettler et al., 2020). In this system the structure-function problem is similar to that outlined for marine snow: How does community structure determine the rate of plastic degradation?

Microbial mats

Microbial mats are stratified communities that often form at air-water interfaces in extreme environments such as hot springs (Klatt et al., 2013). Mat communities harbor a top layer of photoautotrophic bacterium (typically Synechococcus sp.) that use light to fix carbon during the day and often fix nitrogen at night. Below the top layer are strata of various heterotrophs and anaerobic autotrophs (Ward et al., 1998) (Figure 2C). These communities have been studied for decades, and much is known about the metabolic roles each strain plays in the community (Bateson and Ward, 1988; Anderson et al., 1987) and metagenomic data-sets are available (Lee et al., 2018). It is remarkable that similar mat structures form in many different contexts across the globe. In these mats, much of the nutrients are fixed from CO_2 and N_2 (Steunou et al., 2006), heterotrophic community members then consume reduced organic carbon excreted by the autotrophs. In this context, the question of relating structure to function falls to mapping the flux of C, N and other metabolites through the mat to the taxonomic and metagenomic structure of the system. For example, what is the simplest community that will stably form a mat? What pathways, for example in the primary autotroph, are essential to mat formation and which are dispensable? Further, given that mats support remarkable allelic diversity driven by extensive recombination (Rosen et al., 2015), how is the functional genetic repertoire of these communities maintained? Preliminary work on mats in California (Lee et al., 2018) showed the presence of a core genome across samples, and various other genes thought to be useful for specialized functions, but further metagenomic analysis is likely to be useful in addressing the structure-function question in mats. The two main challenges in these communities are obtaining axenic isolates from the mats and making high quality metabolite dynamics measurements in situ. Petroff et al. have recently made exquisite quantitative measurements of oxygen dynamics in mats (Petroff et al., 2017; Tejera et al., 2018), addressing the latter challenge, which opens the door to making quantitative links between community composition and metabolic activity.

Cryptic sulfur cycling in microbial aggregates

Microbial metabolism drives the global cycling of sulfur through energy-generating oxidation and reduction reactions. It has become increasingly evident that much of this cycling occurs cryptically (Canfield et al., 2010; Callbeck et al., 2018) (i.e., with low steady state metabolites levels but substantial fluxes), often in the context of cellular aggregates where oxidation and reduction reactions occur in physical proximity (Wilbanks et al., 2014; Callbeck et al., 2018). Inferring the presence and rate of cryptic sulfur cycling in microbial aggregates has important implications for our understanding of other elemental cycles, since sulfur cycling is often tightly coupled with the carbon and iron cycles.

Remarkable examples of cryptic sulfur cycling phenomenon are the so-called “pink berry” consortia (Seitz et al., 1993; Wilbanks et al., 2014), which are discrete macroscopic (1 cm) aggregates that occur on the surface of submerged sediments in the intertidal pools of Sippewissett Salt Marsh (Massachusetts, USA)

(Figure 2D). The bright pink coloration of these aggregates is attributable to the purple sulfur bacteria (PSB) that make up the majority of cellular biomass. These PSB oxidize sulfide to sulfate via the process of anoxygenic photosynthesis. Accompanying these PSB are sulfate-reducing bacteria (SRB), which derive energy from anaerobic respiration by catalyzing the reverse process, reducing sulfate to sulfide. Together these PSB and SRB are capable of locally and cryptically cycling sulfur via the syntrophic exchange of oxidized and reduced sulfur compounds (Wilbanks et al., 2014).

While culture-based approaches to characterizing and reconstituting the pink berry consortia in the lab remain a challenge, the discrete nature of the pink berries presents an opportunity to statistically characterize the structure-function relationship at the level of individual aggregates harvested from the wild. Bulk metagenomic sequencing of pink berry consortia indicates that typically two to three phylotypes make up the majority of biomass in the aggregates (Wilbanks et al., 2014). Laboratory measurements of sulfide oxidation or sulfate reduction rates using spatially explicit microprobe measurements followed by 16S metagenomic sequencing on harvested pink berries could provide insight into how the abundance of these phylotypes quantitatively relates to sulfur cycling. More generally, such an approach could be applied to characterizing cryptic sulfur cycling in other aggregated contexts, such as in marine particles (Callbeck et al., 2018). Predicting sulfur cycling is important to understanding other critical elemental cycles: Sulfide oxidation by PSB can contribute substantially to carbon fixation (Dyksma et al., 2016) and sulfate reduction by SRB can be an important sink for reduced carbon (Liamleam and Annachhatre, 2007) and iron (Enning and Garrelfs, 2014) in environments where electron acceptors are otherwise scarce.

Community structure and function under domestication

Many key insights in the On the Origin of Species came from studying trait variation under domestication. In the same way, we propose that learning the rules for mapping structure to function in communities should leverage the many instances in which communities have been domesticated. Here we explore some of these opportunities.

Microbial communities in the dairy industry

The production of yogurt, cheese and other dairy products relies heavily on microbes. The work of Dutton et al. on cheese rinds (Wolfe et al., 2014) is an important example in this context. This study compared and characterized more than 100 cheese rinds from across 10 countries, and found that less than 15 bacterial taxa and 10 fungal taxa were present in abundances of more than 1%. Further, most of these taxa are not present in the starting cultures, and their function is unknown. They found that the community compo-sition is strongly correlated to the aging process and moisture rather than geography or milk source. The functional profile of the communities, found using shotgun metagenomics, was correlated to pH as well, and pathways were shown to correlate as expected with the cheese types. These results point to the idea that the chemical environment is perhaps the strongest determinant of community structure.

Remarkably, when the dominant taxa were reconstituted in vitro and cultured as a community in media to mimic treatments in different types of cheese rinds, divergent communities were formed depending on the treatments, which showed some properties similar to the original cheese rinds including their abundance dynamics. This remarkable result shows that these communities can be reconstituted in the laboratory to recapitulate some of the basic functional features of the domesticated cheese communities. It would be interesting to extend these studies further by looking quantitatively at the metabolite dynamics. The fact that these communities can be so readily manipulated means that learning the relationship between composition and metabolic activity is now accessible. What remains unclear to us is what the salient metabolic features of these communities are that should be explained. One way to approach this question would be to chemically characterize the transformations that occur in specific rinds and to then ask whether subsets of the full community can or cannot recapitulate these processes.

Another promising microbiome in the dairy sector is the kefir community (Figure 3A). Kefir grains are used to make a fermented drink like yogurt. They have about 50 bacterial and yeast taxa, which are resilient to stresses, and most of which perform lactic and acetic acid fermentation. A recent study (Blasche et al., 2021) found that kefir grains, which are polysaccharide matrices synthesized by the bacterial consortium, collected from diverse locations had a very similar core community and differed only in rare species. Kefir grains are sustained much like sourdough starters and added to milk to initiate a fermentation process. When added to milk the composition of the grain community is stable while the community

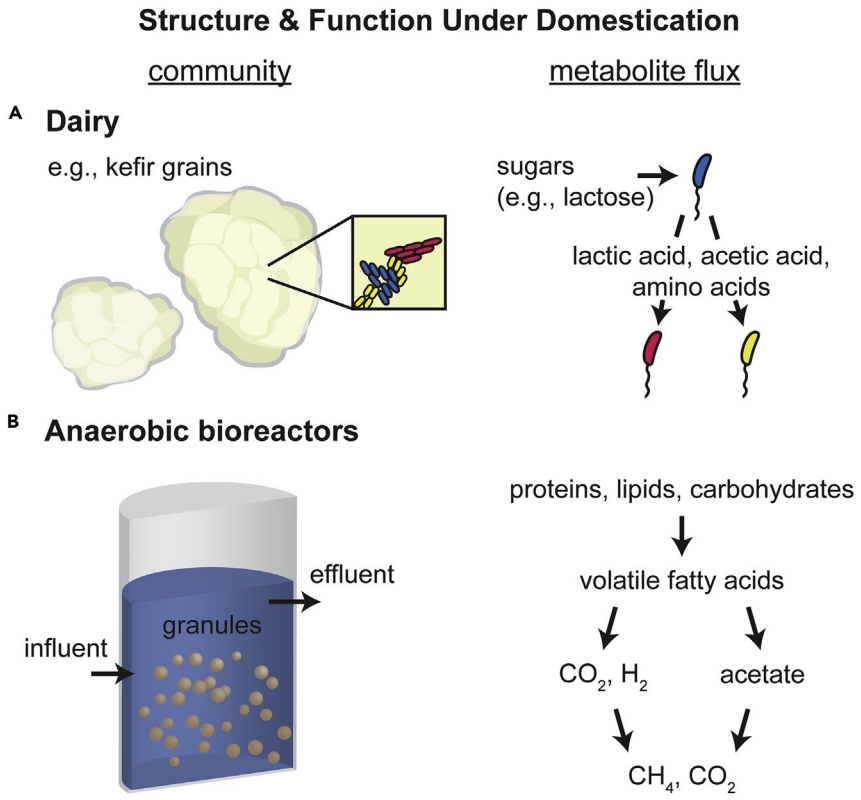

Figure 3. Community structure and function under domestication

(A) Kefir grains extracellular-polymeric aggregates host to microbial communities that inoculate milk for the production of kefir. These communities undergo a reproducible successional process that involves the production and consumption of fermentation byproducts, which ultimately give kefir its desired flavor.

(B) Anaerobic bioreactors often use granulated microbial communities to remove waste products such as reduced carbon, nitrogen, and phosphorous from water. Improving the performance and efficiency of these systems through the ensembles-informed design of communities would increase their viability as alternatives to traditional wastewater treatment approaches that expend significant energy on aeration

present in the milk exhibits a succession. Metabolite changes during the colonization showed similar succession dynamics. The study dissects the interactions between the community on the grains and that in the milk, and demonstrates reproducible metabolite dynamics in this system. As a result, kefir constitutes another powerful platform for manipulating community structure (e.g., composition of the community on the grains) and learning the impact of those changes on the metabolite dynamics. Blasche et al. have already made significant progress in this regard, but it remains to investigate in a high-throughput statistical fashion how the composition of the grain community confers the remarkable robustness they observe.

Anaerobic bioreactors

The treatment of wastewater for reuse and release into the environment requires the removal of large quantities of organic matter, much of which is insoluble or otherwise slow to degrade. Anaerobic bioreactors serve an important role in this industrial process, harnessing microbial metabolism to degrade such organic matter into CO_2 and CH_4 gases (Figure 3B). Because the bioreactors are fed a range of inputs and are operated under a variety of conditions, the microbial communities that populate bioreactors are highly functionally and taxonomically diverse (Werner et al., 2011). Organisms that excrete extracellular enzymes degrade insoluble polymers into soluble monomers, while fermenters consume these compounds and excrete products including acetate and H_2. Methanogenic archaea can then ferment these products to produce CH_4(Toerien and Hattingh, 1969). Many functional attributes are used in practice to characterize the performance of anaerobic bioreactors, including removal of chemical oxygen demand (COD, a proxy for

aerobically metabolizable matter in a water sample) and methanogenic activity. The resilience of a bioreactor community to input fluctuations has also been of interest (Fernandez et al., 2000; Hashsham et al., 2000; Werner et al., 2011).

Several studies have explored the statistical relationship between community taxonomic structure and methanogenic bioreactor performance. In an important early study, Tiedje et al. found that, under constant conditions, COD removal in a laboratory bioreactor was stable while community composition varied substantially over a two-year period (Fernández et al., 1999). This work suggested the role of functional redundancy (Louca et al., 2018) in maintaining community function, and implied a degeneracy in the relationship between community taxonomy and function. However, a more recent study of several industrial-scale bioreactors observed relative stability in both community taxonomy and reactor performance over a year-long period (Werner et al., 2011). Notably, variations in reactor performance were found to be related to community composition. This indicated that taxonomy is predictive of community function, although the authors argued that taxonomy is simply a proxy for functional genomic content due to a close correspondence between phylogeny and metabolic function for organisms in anaerobic bioreactor systems. The conflict between these two results indicates that much is still unknown about the structure-function relationship in anaerobic bioreactors.

Recent work has advanced our understanding of structure and function in methanogenic bioreactors, leveraging a statistical ensembles approach to discover a predictive relationship between gene content and methanogenic activity (Bocher et al., 2015). The authors generated 49 diverse enrichment cultures by seeding laboratory bioreactors with inocula taken from a large and eclectic collection of industrial-scale bioreactors. This ensemble enabled a linear regression approach to mapping methanogen activity (as measured by methane production rate per unit biomass) to variation in genomic content, specifically the abundance of sequence variants of a gene important to methanogenesis (mcrA). Remarkably, their approach produced a predictive model with relatively few variables, suggesting that only a few key mcrA variants, or strains possessing these variants, are important for methanogenic activity. A powerful consequence of this approach is a prediction for which gene variants would improve the performance of an underperforming bioreactor.

Given the importance and widespread use of bioreactors to process organic matter, these constitute important model systems. The field faces two important challenges. First, cultivating many of the slow growing taxa in these communities is difficult and this means that only equipped and experienced labs can readily work with these organisms. Second, the complexity of these communities makes carefully controlled and reproducible experiments a challenge, and as a result, comparisons from one study to the next can be difficult. Standardizing conditions and starting inocula would therefore be a major advance.

Bringing wild communities into the lab: enrichment cultures

Another route to studying communities that is similar to domestication is to bring complex communities from nature into the laboratory and culture them in defined conditions. While these experiments allow the experimentalist to control the growth and incubation conditions, they often result in a drastic loss of diversity. As a result, this approach is likely poorly suited to understanding the structure of natural systems, but nonetheless should enable important insights into engineering and controlling communities.

Winogradsky columns

One of the most well known methods for enrichment culture is a Winogradsky column (Figure 4A). The method was developed by Sergei Nikolaievich Winogradsky to study (and discover) chemosynthesis, a process where energy is derived from inorganic compounds in the absence of light (Zavarzin, 2006). Winogradsky columns are glass cylinders (or bottles) loaded with a sediment, supplemented with an organic carbon source (typically paper) and a sulfur source, which are then sealed off and illuminated. With time, different microbes occupy different levels of the column based on their metabolic capabilities. Each layer is characterized by distinct redox reactions (electron donor/acceptor pairs) that support metabolism locally (much like the mats discussed above), and via diffusion to other strata impact the metabolism of the entire column. For example, the top layer supports phototrophs that produce carbon and oxygen, which drives a second layer that aerobically respires carbon and oxygen, resulting in a third layer that is anaerobic and typically uses alternate electron acceptors such as nitrate or sulfate. Recent work has begun to show that these complex communities are amenable to quantitative interrogation

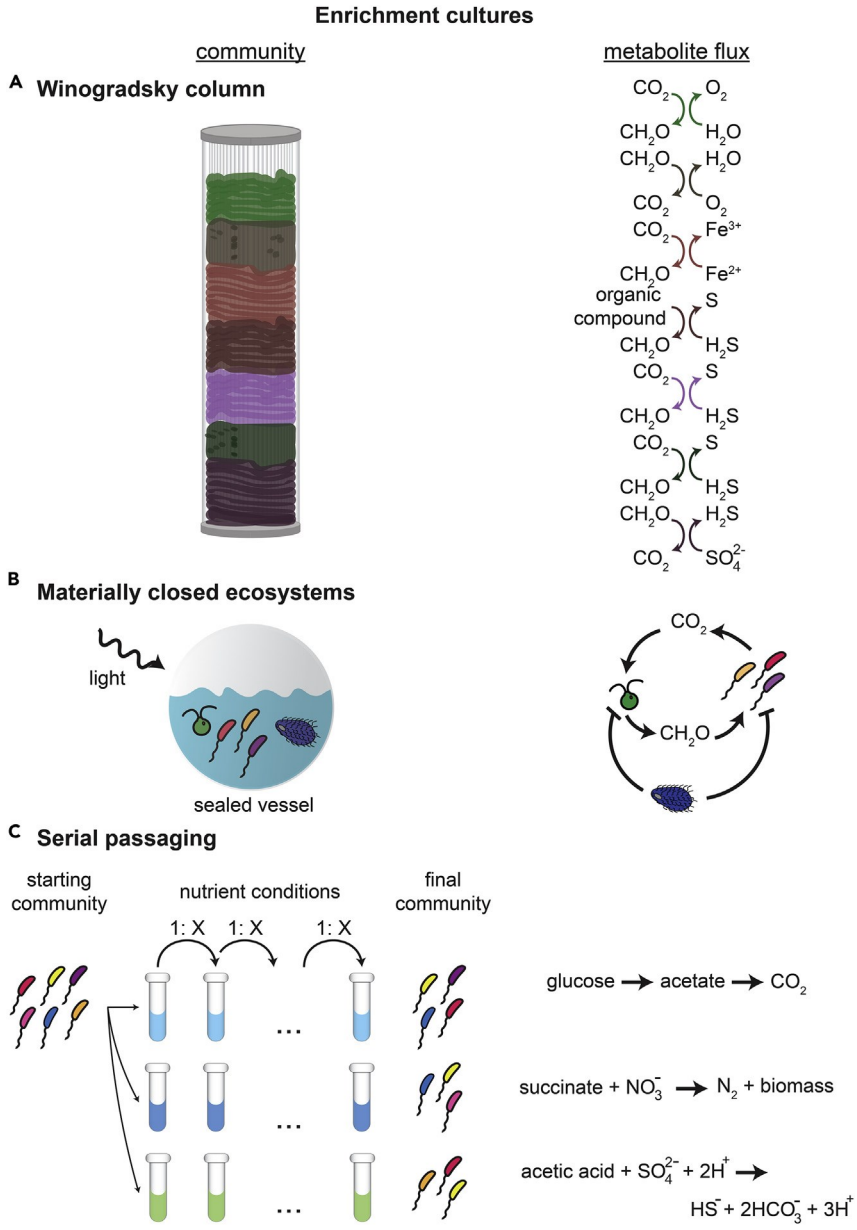

Figure 4. Bringing wild communities into the lab

(A) Winogradsky columns are laboratory-assembled communities with distinctive and reproducible spatially-stratified metabolite fluxes. These fluxes, and consequently community structure, arise from emergent redox gradients.

(B) Materially closed ecosystems are communities grown in sealed vessels whose only energy input is light. Nutrient cycling in closed ecosystems arises from phototrophic organisms generating reduced carbon by fixing CO_2, which can then be consumed by heterotrophic organisms. Predators such as ciliates can consume whole cells, facilitating the recycling of macromolecular biomass.

(C) Serially passaged communities enrich a complex environmentally-derived community on laboratory-controlled nutrient conditions (e.g., a fixed carbon source). The resulting communities are typically low-complexity, and demonstrate reproducible trophic roles and patterns of nutrient exchange

in the laboratory (Pagaling et al., 2014, 2017). Of particular note is a study that used centimeter long glass capillaries to assemble stratified communities cultured in the presence of dyes to report pH and redox (Quinn et al., 2015). A simple imaging setup then permitted the acquisition of quantitative spatiotemporal data on community assembly in many replicates. These systems were used to simulate

communities in the lung of a cystic fibrosis patient, but an opportunity to extend this work remains and the platform is well-suited to the ensemble approach outlined here.

Given that metabolic niches are spatially separated in Winogradsky culture systems, they are ideal systems for understanding how this stratification self-organizes, their energetics and how this organization depends on the boundary and initial conditions in the system. For example, if the capillary system of Quinn et al. (2015) could be combined with defined nutrient conditions, controlled illumination, and quantitative imaging, one could ask how the layers (and therefore the metabolite fluxes) depend on community composition. Given the timescale of these experiments (months), working with many replicates in parallel will be key.

Materially closed ecosystems

Closely related to Winogradsky columns are materially closed microbial ecosystems (CES, Figure 4B). Closed ecosystems are hermetically sealed, typically aquatic, microbial communities that have been shown to sustain life with only light as an input for decades in some cases. These 'ecospheres' are available commercially (https://eco-sphere.com/) and have been studied in an academic setting since the 1960s. CES contain photosynthetic microbes, typically algae or bacteria, and either simple or complex consortia of heterotrophic bacteria and predators. When provided with only light these communities self-organize to sustain nutrient cycles and therefore the community itself. CES act as model biospheres because they require nutrient cycling to persist (Rillig and Antonovics, 2019). In the context of the structure-function problem, the salient metabolic property of CES is nutrient cycling, and one question is how the composition of the community determines nutrient cycling rates and persistence.

Work from a number of groups (Obenhuber and Folsome, 1984, 1988; Kearns and Folsome, 1981; Kawabata et al., 1995; Taub, 1974, 2009) showed that CES containing primarily microbes were tractable model systems and methods for quantifying nutrient cycling were developed (Obenhuber and Folsome, 1988; Taub, 2009). More recently, Leibler and Hekstra (Hekstra and Leibler, 2012) and later Leibler, Frentz and Kuehn (Frentz et al., 2015) studied population dynamics in a synthetic closed ecosystem comprised of three species using sophisticated microscopy-based methods. These studies revealed remarkably deterministic population dynamics in CES. However, few studies have quantitatively characterized the nutrient cycling capabilities of CES with the notable exceptions of Obenhuber (Obenhuber and Folsome, 1988) and later Taub (2009). The early work of Obenhuber showed that complex bacterial communities mixed with photoautotrophic algae or bacteria could sustain a carbon cycle for many months (Obenhuber and Folsome, 1988). Inspired by these studies, we recently developed a higher throughput method for quantifying carbon cycling that uses low-cost microelectromechanical sensors (MEMS) made for mobile devices to measure small changes in pressure inside a CES (de Jesús Astacio et al., 2021). As appreciated by Obenhuber and Folsome, changes in pressure reflect carbon cycling, because oxygen has lower solubility in water than CO_2. When photosynthetic microbes fix CO_2, they produce oxygen and the pressure rises. When bacteria respire carbon, the opposite happens and one can quantify carbon cycling by measuring pressure oscillations during light-dark cycles. We constructed CES using bacterial communities derived from soil samples combined with a domesticated strain of the alga Chlamydomonas reinhardtii. We found that taxonomically diverse CES stably cycled carbon for as long as six months. Metabolic profiling of these communities showed that, despite taxonomic variability across replicate CES, each community exhibited similar metabolic capabilities in terms of the carbon compounds they could utilize. It remains to be understood how such taxonomically distinct communities can retain this diversity while exhibiting similar metabolic capabilities. To address this question using the approach outlined here would require studying many synthetic communities with varying composition while measuring carbon cycling.

We propose that CES constitute powerful model communities for understanding nutrient cycling. In the context of the approach outlined here, CES could be used to understand how initial nutrient supply (C, N, P, S, etc.) controls community structure and cycling. Further, since light is the only source of energy, CES can be used to explore how energy availability impacts the structure-function mapping in terms of cycling.

Enrichment in defined media

Recently, Sanchez et al. performed serial enrichment cultures on complex natural communities in simple defined media containing a single carbon source (Goldford et al., 2018; Estrela et al., 2021) (Figure 4C).

In this case, the function of the community is the conversion of glucose to CO_2 and biomass. Over a few tens of generations these communities assembled into relatively simple, and predictable consortia comprised a few cross-feeding strains. This convergent structure was typified by characteristic ratio of strains from the Enterobacteriaceae and Pseudomonadaceae families. To explain this conserved structure, strains were isolated from the endpoint of the enrichment experiment. They were then assayed for growth on glucose, and Enterobacteriaceae strains were found to grow faster on glucose than the Pseudomonadaceae. Metabolomic measurements (mass spectrometry) revealed that Enterobacteriaceae strains excreted intermediates such as acetate, which Pseudomonadaceae were observed to consume rapidly. Therefore, similar to the polysaccharide particle experiments discussed above (Datta et al., 2016), these experiments revealed a primary carbon degrader that rapidly consumed the supplied carbon source but released secondary metabolites (in this case acetate) that supported the growth of other strains. Very recent work suggests that this type of cross-feeding may arise from stress induced release of nutrients that arises due to serial dilution (Amarnath et al., 2021). In these communities the salient metabolic property of the system is carbon degradation. These studies have already revealed much about how the structure of these communities impacts the carbon degradation (e.g., the role of secondary consumers). It would be interesting to more fully elucidate the mechanistic basis of this cross-feeding in the service of understanding how the genotypes of each strain present determine their uptake and release of carbon compounds. Hopefully, these insights would allow us to understand how the cross-feeding depends on available carbon compounds and environmental parameters such as pH. These generalizations could prove insightful in natural contexts such as soils where carbon degradation is a critical phenomenon for climate change.

Bridging structure and function with synthetic communities

In recent years a number of studies have attempted to bridge structure and function of wild communities by performing experiments on synthetically assembled communities of natural isolates. This approach offers an opportunity to capture much of the substantial taxonomic and genomic diversity of natural communities within a setting where environment and composition can be controlled, and function can be measured accurately.

A handful of studies have explored how interactions between strains in a community affect carbon utilization. In part, these studies attempt to determine how prevalent, different types of interaction between strains are (e.g., mutualism, parasitism, etc.), and how these interactions vary based on community composition and the identity of the carbon substrate provided. Using strains isolated from tree-associated environments, Foster and Bell (2012) assembled synthetic communities and measured CO_2 evolution during growth on a complex medium as a metric for community productivity. The measured values of productivity were compared to a "non-interacting" null prediction obtained by summing the productivities of the constituent strains grown in monoculture. What was observed in the vast majority of pair cultures and higher-order communities was consistent with resource competition rather than mutualism, suggesting that competitive interactions for carbon utilization dominate in natural communities. This implies that the carbon utilization of a community should saturate with the diversity of a community, which is precisely what is shown in a more recent study by Yu et al. (2019). In this study, a diverse ensemble of communities with vary-ing strain composition and diversity was generated from a seawater inoculum via serial dilution and dilute-to-extinction approaches. Cell density, protein concentration, and CO_2 evolution were measured along with community diversity (via 16S metagenomic sequencing). These community function measurements show saturating behavior as a function of taxonomic richness. Abundant strains were isolated to disen-tangle the effects driven by individuals and effects driven by interactions. Interactions increase and saturate with diversity, suggesting both competition and complementation increase simultaneously with diversity.

Additional recent studies have added nuance to the picture of what interactions are prevalent in carbon-degrading communities (Kehe et al., 2019, 2021), finding a high prevalence of parasitic and mutualistic interactions within communities of isolates that are consistent with cross-feeding. These studies leverage a high-throughput droplet microfluidics platform to perform combinatorial community assembly and culture in multiple different carbon source conditions, and the growth of community constituents was measured using fluorescent labeling. In the first study (Kehe et al., 2019), a high prevalence of competitive interactions was observed, particularly for communities and carbon sources where the constituent strains all showed strong growth on the carbon source in monoculture. However, positive interactions were frequently observed in cases where one strain grew poorly on a carbon source in monoculture, consistent with that strain growing in a community because of cross-feeding of metabolic intermediates. The second study (Kehe et al., 2021) broadened this observation by focusing on communities comprising strains from two taxonomic orders, Enterobacterales and Pseudomonadales. Again,

it was observed that positive interactions were common in communities where one of the constituent strains grew poorly by itself on a given carbon source. It is likely that the positive interactions were generated by cross-feeding of overflow metabolism intermediates (Amarnath et al., 2021). These positive interactions likely arise from the same mechanism discussed above in the enrichment culture experiments of Sanchez et al. (Goldford et al., 2018; Estrela et al., 2021). Altogether, these results indicate that community carbon degradation can be a relatively simple function of the taxonomic structure of a community. Linking these relationships to genomic structure remains to be accomplished.

Another study by our group (Gowda et al., 2022) set out to explicitly identify the genomic attributes of a community that are predictive of community function. Using a statistical ensemble of isolates that perform denitrification, a process of anaerobic respiration involving the reduction of oxidized nitrogen compounds, we first mapped the genotypes of individual isolates to the precisely quantified kinetics of nitrate and nitrite reduction, which were parameterized using a consumer-resource modeling approach. We used a regularized linear regression approach to predict nitrate and nitrite reduction kinetics from the presence and absence of denitrification-pathway genes possessed by each isolate. We then assembled communities of these isolates and determined that resource-competitive interactions are prevalent and predictable from single-strain kinetics via the consumer-resource model. Thus we inferred that the conserved properties of metabolic genes allow the prediction of community-level function. This study shows that synthetic communities comprised natural isolates, combined with statistical approaches, can yield insights into the mapping from gene content to metabolite dynamics. It remains to be seen if this approach can be applied to more complex communities.

Our work in Gowda et al. (2022) points to the idea that focusing on the genomics and ecology of specific metabolic processes can be a powerful approach. In particular, this study leads us to believe that denitrification offers a remarkable system not only for the quantitative interrogation of structure and function using the statistical approach outlined here, but also for detailed physiological studies of specific metabolic processes. The advantages of denitrification include the facts that the taxa that perform the process are easily isolated and grow well in the laboratory (Lycus et al., 2017), the metabolites can be quantified in high-throughput (Gowda et al., 2022), the molecular genetics of denitrification are well-understood (Zumft, 1997) and the process has been characterized in the wild to some extent (Tiedje et al., 1983). These opportunities have spawned several compelling recent studies (Lilja and Johnson, 2016,2019; Goldschmidt et al., 2018), including a study of the role of carbon source identity in driving denitrification in wild communities (Carlson et al., 2020).

EXPERIMENTAL TECHNIQUES TO ENABLE AN ENSEMBLE APPROACH

Having reviewed a variety of model communities for undertaking an ensemble approach to the structure function problem we now turn our attention to experimental methods to study these systems. We focus here on culturing and isolation methods as well as analytical techniques for measuring metabolites. We do not review sequencing approaches that have been discussed in detail elsewhere (Hugerth and Ander-sson, 2017).

Isolation techniques

A vast majority of microbes that are known to exist in nature remain uncultured in the laboratory. The staggering difference between the number of cells counted from microscopy and those obtained on agar plates was discovered in the early 19th century (Amann, 1911) and was dubbed as the "the great plate count anomaly" in 1985 by Staley and Konopka (1985). Developments in sequencing techniques have further widened the gap between the number of cultured and uncultured bacteria. The causes for microbial uncultivability include requirement of growth factors present in the natural environment, slow growth, need for interspecies interactions and transitions to dormancy. Development of isolation techniques that overcome these drawbacks is important to capture the high microbial diversity that exists in the wild. Such techniques will aid in the construction of synthetic communities with high genotypic and phenotypic diversity and hence benefit the study of structure-function problem.

Recently developed isolation techniques that offer some advantages over the conventional approach of plating on agar include culturomics, microdroplets and diffusion chambers. In culturomics, communities are tested for growth in a multitude of media conditions using high-throughput techniques, followed by subjecting the communities to mass spectrometry and sequencing (Lagier et al., 2012; Seng et al., 2009). By performing MALDI-TOF (see below) mass spectrometry directly on colonies, bacterial taxa can be iden-tified with high fidelity (Seng et al., 2009). In case MALDI-TOF fails to identify taxa, 16S sequencing is used.

The use of mass spectroscopy combined with sequencing facilitates accurate, rapid, and comprehensive strain identification. In a recent study (Lagier et al., 2016), the culturomics approach was shown to be very successful in increasing the number of species isolated from the human gut (by 2-fold).

A higher-throughput method involves encapsulating cells from natural communities in gel microdroplets (GMDs) made of agar. The GMDs are then incubated in growth chambers flushed with low nutrient media. Following this, GMDs with microcolonies are sorted using flow cytometry and individual GMDs are subsequently transferred into microtiter plate wells containing rich organic medium for biomass enrichment and isolation (Zengler et al., 2002, 2005). In this technique, the porous nature of the GMDs facilitates exchange of metabolites between droplets during the incubation. As a result, strains that require metabolites produced from resource-mediated interspecies interactions can be isolated using this technique.

Diffusion chambers work by culturing communities in chambers exposed to their native environments through porous membranes (Kaeberlein et al., 2002; Bollmann et al., 2007; Chaudhary et al., 2019). Thus, the setup allows for the growth of microbes that require growth factors present in their natural environments and/or produced from native community interactions. Significant developments improving the throughput of this isolation method include the isolation chip (Berdy et al., 2017) and the Hollow-Fiber Membrane Chamber (HMFC) (Aoi et al., 2009).

In addition to these non-targeted isolation techniques, targeted isolation techniques have been recently attempted. These involve designing the isolation methods to target desired phenotypes (e.g., antibiotic resistance or sporulation (Browne et al., 2016)). One recent study successfully isolated desired cell types using fluorescently labeled antibodies against predicted cell surface proteins combined with flow cytometry for cell sorting (Cross et al., 2019).

Overall, both the available targeted and stochastic isolation techniques have proven to be useful for isolating previously unculturable bacteria. Hence, these techniques may prove valuable for generating ensembles of bottom-up assembled microbial communities.

High throughput culturing platforms

Our proposed ensemble approach for studying structure-function relationship in microbial communities requires creation of many replicate communities. Hence, high throughput culturing platforms are critical for its implementation.

A majority of high throughput experimental platforms so far have been droplet-based or microfluidic-based devices. One such recently developed device is 'kChip', a microfluidic platform that facilitates combinatorial construction of microbial communities (Blainey et al., 2018). A study involving synthetically constructed microbial communities on kChips successfully identified sets of strains among 19 soil isolates that promote growth of model plant symbiont Herbaspirillum frisingense, by screening 100,000 multispe-cies communities (Kehe et al., 2019). Although kChip is a high throughput platform, it only enables bottom-up construction of microbial communities that requires isolation of microbes prior to the experiments. Additionally, only metabolic functions with optical readouts can be assayed as physical access to the microdroplets at this scale is not feasible.

Another similar microfluidic platform that enables parallel co-cultivation of microbial communities was developed by Park et al. (2011). Their platform was able to successfully detect pairwise symbiotic interactions in communities when the symbionts were in as low an abundance as 3 percent of the total population. Here again, only optically detectable metabolic properties can be measured, but the device enables top-down construction of microbial consortia through random compartmentalization of community members. Though this was not a structure-function study per se, inclusion of automated droplet sorting and characterization of communities in the retrieved droplets can easily enable structure-function studies. In fact, this was achieved in a more recent study by Terekhov et al., where microbes conferring antibiotic resistance in the oral microbiota of Siberian bears were identified (Terekhov et al., 2018). This was done by functional profiling of the encapsulated communities from the oral microbiome that suppressed the growth of the pathogen Staphylococcus aureus.

Table 1. Comparison of methods for measuring metabolites in microbial communities

Method	Sensitivity	Specificity	Range of applicability	Throughput
NMR	Low	High	High	Low
Mass Spectrometry	High	High	High	Low/moderate
Infrared/Raman	Moderate	High	High	High
UV/Vis	Moderate	Low	Low	High
Targeted assays	Moderate	High	Low	High

Sensitivity refers to the minimum detectable concentration. Specificity refers to the ability of the assay to detect a specific metabolite. Range of applicability refers to the diversity of metabolites that can be detected with the technique. Throughout is the number of measurements that can be made in parallel.

From the aforementioned studies, it can be said that the choice of the experimental platform largely depends on the nature of the study. Some existing platforms support a top-down approach whereas others support a bottom-up approach. Further, the methods of determining structure and function can differ across platforms. There is room to improve these methods to incorporate other analytical techniques for measuring metabolites. For example, if large scale culturing platforms could be combined with spectroscopic or automated mass spectrometry methods, this would enable the rapid construction of large quantitative datasets.

Measuring metabolite dynamics

Metabolites, unlike nucleic acids, require distinct analytical techniques depending on the metabolite of interest. Here we review the available methods, their applicability and opportunities for improving these methods for microbial consortia. See Table 1 for the specific strengths and limitations of each technique discussed here.

Nuclear magnetic resonance spectroscopy

A number of high quality textbooks describe the fundamental physics (Slichter, 1990) and chemistry (Levitt, 2008) of nuclear magnetic resonance (NMR) spectroscopy.

Here we give an intuitive explanation of the basis of this technique and go on to the practical applications of measuring metabolites in microbial communities.

NMR spectroscopy exploits the spin magnetic moment of atomic nuclei such as hydrogen, carbon, and nitrogen to characterize chemical structure. In an applied magnetic field, nuclei behave as weak magnets, collectively aligning with the field. The collective alignment of the nuclear magnetic moments can then be manipulated with applied electromagnetic fields in the radio frequency (MHz) and detected as emitted fields in the same spectral region. The small magnetic moments of nuclei cause them to emit very weak ra-diation, meaning that relatively high concentrations of metabolites of interest are necessary for detection. Nuclei experience minuscule changes in the local magnetic field due to their local chemical context, result-ing in what are termed “chemical shifts.” For example, a proton in hydrogen on an alkane (e.g., methane) will emit a distinct radio frequency (its resonance frequency) from one in an aromatic hydrocarbon (e.g., benzene). These small changes in emitted radio frequency fields are of the order of parts per million (ppm). Typical resolution of modern instruments is a fraction of a ppm and depends on the field strength of the spectrometer and technical details of the detection scheme. State-of-the-art spectrometers (oper-ating at 600MHz and above) are widely available at core facilities.

For metabolomics the two most common types of NMR are proton (^{1}H) and carbon NMR, each of which has both advantages and disadvantages. First, the advantages of proton NMR are (1) rapid acquisition due to the relatively high signal-to-noise ratio in proton spectra, (2) the fact that no isotopic labeling is necessary, and (3) the ability of the technique to detect a broad range of relevant organic compounds. One disadvantage is the fact that spectra from mixtures of unknown metabolites are complex, often containing hundreds of peaks corresponding to the many different compounds present. As a result, it can be challenging to detect the presence or absence of specific metabolites via proton NMR. Further, water contributes a broad and strong solvent signal in the middle of the relevant range of chemical shifts. There are two main routes to

removing this signal: (1) drying the sample and replacing the water with D2O and (2) using clever pulse sequences to decouple the water signal from the metabolite signals (Mckay, 2011). The former requires specialized equipment and increases the cost and reduces throughput. Therefore, it is recommended to use decoupling. The fundamental physics of how this decoupling works is beyond the scope of this review, but it is recommended to use 1D NOESY (Nuclear Overhauser Effect Spectroscopy) to isolate metabolite signals from water (Emwas et al., 2019). The approach is robust, widely applied, and requires no sample processing to be done. It should be noted that because of variation in technical specifications between instruments, performing measurements on a single spectrometer across samples is key to maintaining reproducibility of measurements (Mckay, 2011).

Carbon NMR in contrast, detects signals from the magnetic moments of carbon nuclei. As with protons, the chemical context of the carbon nucleus gives rise to chemical shifts in the resonance and this permits the disambiguation of carbon nuclei in different compounds. One advantage of Carbon NMR is that it does not require suppression of water signals. The main drawback to carbon NMR is sensitivity. The dominant isotope of carbon (^{12}C, 99% prevalence) is not NMR active, while ^{13}C is NMR active, but present at about 1% natural abundance. This means that most nuclei do not contribute to the observed signal. Second, ^{13}C has a magnetic moment that is roughly 4-fold lower than ^{1}H reducing the signal-to-noise ratio. These two considerations imply that acquiring carbon spectra requires extensive averaging and can take hours for a single sample. However, the low isotopic abundance of ^{13}C can be overcome by using ^{13}C labeled compounds as nutrients, with order of magnitude increases in signal-to-noise. Unfortunately, these compounds are expensive (hundreds of dollars per gram) increasing costs.

Carbon and proton NMR are the two most commonly applied metabolomic profiling techniques for microbial communities. Typical studies range from targeted detection of a single metabolite in a well-defined community (Andrade-Domínguez et al., 2014) to untargeted profiling in very complex consortia such as anaerobic digesters (Gonzalez-Gil et al., 2015). Here we present a few examples of NMR based measurements of metabolite dynamics in communities as case-studies that might be more broadly applicable.

One compelling approach taken by Date et al. (2010) and Nakanishi et al. (2011) is to combine NMR based metabolite measurements in time with quantification of abundance dynamics. Date et al. use ^{13}C labeled glucose to initiate growth in fecal microbiota (Date et al., 2010) (note that ^{13}C is a stable isotope). The authors then performed time series of abundance dynamics and carbon NMR measurements. Given labeled glucose as the sole carbon source, the authors could track the dynamic production and consumption of ^{13}C labeled compounds as the glucose was converted to organic acids in time. The authors could then (crudely, given the electrophoresis methods at the time of the study) classify the community into primary and secondary degraders. The compelling aspect of this study is the potential to statistically correlate large-scale variation in the community structure with metabolite dynamics. One could imagine a similar experiment with amplicon sequencing-based abundance dynamics measurements. This approach would be especially powerful for looking at the community level response to carbon fixation by autotrophs in systems like mats or CES. In these situations, initiating a community with ^{13}C labeled bicarbonate as the sole carbon source would allow the direct measurement of carbon flux from autotrophs to associated heterotroph communities. Since NMR relies on magnetic fields and emitted RF signals, it is non-invasive and could be applied to ongoing experiments (e.g., CES) without invasive sampling.

We conclude with a brief note on throughput. NMR spectrometers are large, expensive machines that rely on superconducting magnets to apply large magnetic fields. Running parallel experiments on NMR machines is therefore prohibitive. Throughput is achieved by using robotics or fluidic systems to automatically load samples into the spectrometer. Such experiments typically achieve a throughput of order 100 samples per day (Macnaughtan et al., 2003; Soininen et al., 2009). Increasing NMR throughput by a factor of 10–100 would constitute a major advance.

Finally, despite the current limitations, there is a revolution underway in quantum sensing systems from superconducting quantum interference devices (SQUIDs) (McDermott et al., 2002) to spin-based magnetic field sensors, impurities in diamonds (nitrogen-vacancy (NV) centers) (Cujia et al., 2019) or force measurements (Kuehn et al., 2008). SQUIDs enable ultra-low field NMR, obviating the need for large expensive magnets, and NV-centers enable high sensitivity magnetic resonance detection at the single-molecule level. The applications of these technologies to metabolic function in microbial communities await future

discovery, but one can imagine massively parallel NMR measurements or in situ detection of metabolites in complex settings.

Mass spectrometry

Mass spectrometry is the most widely used platform for metabolomics and several good reviews of the methodology are available (Beale et al., 2018; Mastrangelo et al., 2015; Dettmer et al., 2007; Raftery, 2014; Alseekh et al., 2021; Jemal, 2000).

Therefore, our discussion of mass spectrometry will be limited, but we include it as an important point of comparison with the other techniques discussed in this section.

Mass spectrometry ionizes the molecules in a sample, using a variety of different methods, and then accelerates the charged molecules using an electric field. The beam of ions is then passed through a magnetic field that (via the Lorentz force law) results in a force on each ion that depends on its mass to charge ratio. The result is a physical separation of ions in space in proportion to the mass-to-charge ratio. A huge number of variations exist on this basic theme, including measurements of time of flight (TOF) and quadrupole mass filters that apply oscillating fields to the ion beam. The reviews cited above contain detailed discussions of the type of ionization and detection methods that are best suited for metabolomic applications.

In the context of metabolomics, where samples are often of considerable chemical complexity, mass spectrometry is almost always preceded by either gas or liquid chromatography to separate compounds and therefore increase the specificity and sensitivity of downstream mass spectrometry. As a result, gas chromatography-mass spectrometry (GC-MS) and liquid chromatography-mass spectrometry (LC-MS) are the two most widely used forms of mass spectrometry for metabolomics. GC-MS is restricted to volatile compounds typically of molecular weight less than 600 Da (Beale et al., 2018), while LC-MS applies to a broader spectrum of metabolites.

Mass spectrometry has significantly higher sensitivity than NMR, and can achieve single molecule sensitivity (Robertson et al., 2007), although this is not routine. This advantage is especially crucial for measuring metabolite fluxes in cells and communities where concentrations are often low (micromolar and below) (Bennett et al., 2009; Basan et al., 2020). This combination of sensitivity and the ability to distinguish broad classes of compounds has contributed to the widespread usage of MS-based metabolomics methods (Aiyar et al., 2017).

Mass spectrometry is often used in communities where specific nutrients (amino acids, sugars) are labeled with stable heavy isotopes (e.g., ^{13}C or ^{15}N). Using stable isotope labels on nutrients allows measurements of fluxes even when the steady state concentrations of metabolites are low, for example when nutrients are produced/consumed at high rates. These powerful techniques allow for quantifying which pathways are utilized (Zamboni et al., 2009), and to observe dynamic changes in those pathways (Yuan et al., 2008). By integrating these approaches with protein mass spectrometry, flux measurements can be assigned to phylogenetically specified strains in a community (Ghosh et al., 2014; Jehmlich et al., 2010; Rühl et al., 2011) and measurements can even capture spatiotemporal dynamics (Mandy et al., 2014).

The throughput of these techniques is significantly lower than the optical methods discussed below and comparable, at present, to NMR. So it is routine to run 100 samples over the course of a day or two (see (Jemal, 2000) or a review). Some robotic systems have been developed to automate the sampling and analysis process (Molstad et al., 2007). However, making mass spectrometry measurements on thou-sands or tens of thousands of samples, while feasible, is costly and slow. Given the remarkable sensitivity, specificity and broad applicability of the technique, especially for untargeted measurements of metabolite pools, it would represent a major advance if mass spectrometry could be routinely applied to thousands of samples in parallel. Here, we mention a handful of notable studies that leverage mass spectrometry to quantify metabolite dynamics in microbial communities. The goal is not to present an exhaustive list but simply to point the reader toward some representative studies.

Amarnath et al. (2021) used untargeted metabolomics to study the metabolites that are exchanged between two strains of bacteria in a serial dilution experiment. The authors revealed a broad spectrum of metabolites excreted by one strain in response to stress. These excreted compounds facilitated cross-feeding

between the two strains. Shi et al. (2017) use LC-MS to study metabolite exchange in a fungal-bacterial community. A statistical analysis of the LC-MS data shows that the metabolites excreted are distinct for co-cultures and mono-cultures.

Mass spectrometry is widely used to study the degradation of compounds from pharmaceuticals to soil contaminants (Pieper et al., 2010; Thelusmond et al., 2016). For example, common environmental contaminants are polycyclic aromatic hydrocarbons (PAHs), which are routinely degraded by bacterial consortia. The breakdown of these compounds in time is typically interrogated by GC- or LC-MS (Luan et al., 2006).

Mass spectrometry is widely applied to food-related microbial communities. In these cases the untargeted nature of mass spectrometry is important as the compounds of interest (e.g., for flavor) are typically unknown. For example, GC-MS has been used to identify starting components in traditional Cambodian rice wine (Ly et al., 2018). Similarly, it was used in fermentation of red peppers to investigate changes in bacterial and fungal communities and volatile flavor compounds (Xu et al., 2020), and high throughput GCMS has been used to correlate metabolites with taxonomic structure in kimchi fermentation (Park et al., 2019), glutinous rice wine (Huang et al., 2019), the liquor Daqu (Jin et al., 2019) and pickled radishes (Rao et al., 2020).

Infrared spectroscopy

Infrared spectroscopy detects absorption and emission of photons in the infrared region of the electromagnetic spectrum and characterizes molecular vibrations. As with NMR, the resonance frequency of a molecular vibration depends strongly on molecular structure. This dependence affords IR spectroscopy its chemical specificity. Infrared spectroscopy offers perhaps the best combination of sensitivity, specificity, high-dimensional characterization of complex metabolite pools and throughput. There are two commonly used methods for measuring infrared spectra that differ in their fundamental physical mechanisms: (1) IR absorption and (2) Raman spectroscopy.

Infrared absorption involves passing light in the infrared range (2.5 mm–10 mm wavelengths or 1000 cm^1 to 4000 cm^1 wave numbers) through an aqueous or gas phase sample and measuring the absorption. Compounds containing different chemical bonds absorb light at different frequencies and the resulting spectrum can provide extensive information on the chemical composition of the sample. As with proton NMR, a major downside of absorption spectroscopy is the broad absorption of water in the informative region of the spectrum (around 3200 cm^1 and 1600 cm^1), which can limit the information for aqueous samples without cumbersome drying. Simple dispersive spectrometers that shine a narrow band of wavelengths through a sample have limits on sensitivity, spectral resolution and the duration of acquisition. These limitations can be overcome using Fourier Transform infrared spectroscopy (FTIR), which uses broadband excitation and an interferometer to rapidly acquire spectra in specific spectral bands, and this is the most commonly applied technique. Plate readers that perform FTIR measurements en masse on microtiter plates are available and can acquire data from both liquid and solid phase samples.

In contrast to FTIR or dispersive IR spectroscopy, Raman spectroscopy measures molecular vibrations using photons in the visible portion of the spectrum. When visible photons interact with a sample, most are scattered with the same energy as the incident photon (Reyleigh scattering). However, with low probability, the incident light undergoes inelastic scattering and in the process photons are emitted from the sample with either slightly lower (Stokes) or higher (anti-Stokes) energy than the incident radiation. These small changes in the emitted photon wavelength correspond to the molecular vibrations in the sample. For reasons beyond the scope of this review, Raman spectroscopy does not suffer from broad band absorption from water, making it especially attractive for microbial communities in the aqueous phase. However, due to the inefficiency of inelastic photon scattering, Raman spectroscopy requires high laser power and the resulting heating can be a problem for biological samples, a limitation that can be overcome by techniques such as resonance Raman spectroscopy or surface enhanced Raman scattering. However, these methods are not yet routine for metabolomic profiling. Raman spectroscopy can be performed on bulk samples using plate readers, or integrated with a microscope for localized measurements. More recently, Raman spectra can be acquired via flow cytometry (Suzuki et al., 2019; Nitta et al., 2020). These platforms enable much higher throughput than is now standard by NMR or mass spectrometry.

FTIR and Raman spectroscopy have proven to be powerful methods for interrogating cellular physiology at the single-cell level when combined with microscopy (see Hatzenpichler et al., 2020 for a recent review). Remarkably, Raman spectroscopy signals can be used as fingerprints to demarcate cells of one species in different growth states (Escoriza et al., 2006), or different taxa at the strain, species and genus levels (Rosch et al., 2005; Harz et al., 2005). A recent study showed that the global transcriptional profile of yeast and bacteria could be predicted via a linear model from single cell Raman spectra (Kobayashi-Kirschvink et al., 2018). This success owes to the high-dimensional nature of Raman spectra, which are often challenging to interpret in terms of individual peaks but are rich in information that can be decoded statistically.

Despite the power of infrared spectroscopies for chemical characterization, they have seen comparatively little use in the context of communities of microbes. We regard this as a missed opportunity, and suggest that these methods could and should be used more broadly. One of the limitations of the technique is the challenge of assigning specific peaks to specific compounds. As Kobayashi et al. have shown, this limitation can be overcome by using simple statistical methods to map infrared spectra to other cellular properties (Kobayashi-Kirschvink et al., 2018). The approach is to measure Raman or IR spectra on a set of samples and then use a lower throughput technique such as LC-MS to measure the absolute concentration of a metabolite of interest. A combination of dimensionality reduction and regression can then be used to map the LC-MS data to the infrared spectra. This approach has been used to track substrate concentrations in time in monocultures (Paul et al., 2016) and phenol degradation in complex communities (Wharfe et al., 2010).

The advantages of Raman spectroscopy and, to a lesser extent, FTIR over mass spectrometry and NMR have the potential for the rapid acquisition of high-dimensional characterization of metabolite pools. The complexity of the resulting spectra is similar to proton NMR, and therefore is perhaps most useful for statistically characterizing differences between community metabolite profiles. Such complex spectra can then be used either to measure specific metabolites via a calibration approach discussed in the previ-ous paragraph, or to demarcate global metabolic states of consortia without concern for specific metab-olite levels.

UV-visible spectroscopy

We briefly note that simple UV and visible spectroscopy (UV-Vis) can be used to characterize electronic transitions in compounds of interest for metabolic characterization. These methods can be performed with widely available plate readers, particularly those that are equipped with monochromators rather than filters, which permit excitation and emission to be arbitrarily selected by the user. The main limitation of this technique is the fact that electronic transitions in the visible are restricted primarily to chemical spe-cies with delocalized electron density (e.g., conjugated rings such as benzene, tryptophan). As a result, these spectra are low specificity and cannot be used for targeted metabolomics. However, the high throughput of common plate readers facilitates rapid measurements, and the spectra can give coarse characterization of excreted compounds from autotrophs, for example (Tenorio et al., 2017). Moreover, targeted UV-Vis measurements can be integrated with common fluorescence microscopes and therefore offer the possibility of detecting metabolites in massively parallelized platforms such as droplet microfluidics (Kehe et al., 2019, 2021) or in spatially structured communities.

Targeted assays

In situations where the metabolic function of the community under study is known a priori and restricted to a specific chemical compound a targeted metabolite assay can be powerful. Such assays typically utilize chemical reactions to create an optically active compound in proportion to the concentration of a metabolite of interest. For example, starch can be degraded to glucose enzymatically and then glucose concentration can be assayed via standard methods (Holm et al., 1986) or iodine can be used to stain starch directly and detect degradation (Fuwa, 1954). Similarly, nitrate and nitrite can be detected via the Griess assay (Miranda et al., 2001), which utilizes a colorimetric reporter (dye) generated via a reaction with nitrite. Such assays can easily be performed in 96-well plates facilitating high throughput (Gowda et al., 2022). These measurements can be powerful for studying specific metabolic processes in communities. However, typically the chemistry involved is not easily automated nor are the conditions of the reaction biocompatible. So, such measurements are made offline after sampling and are challenging to automate in situ.

QUANTIFYING STRUCTURE

We briefly outline the main ways in which community structure can be quantified. As mentioned above a suite of next generation sequencing technologies are capable of quantifying community structure on multiple levels. For example, amplicon sequencing of the 16S rRNA gene uses PCR to amplify this universally conserved ribosomal subunit and then uses the number of reads mapping to sequence variants (amplicon sequence variants, ASVs) as a proxy for the relative abundance of each taxon. This widely used method has many well documented downsides including variation in the copy number of the 16S gene across taxa, PCR bias and the challenge of associating taxonomy with metabolic capabilities of each strain (Callahan et al., 2016). Despite these shortcomings, amplicon sequencing does permit rapid and high throughput characterization of the community composition, and methods exist for inferring metabolic capabilities of strains from 16S gene sequence alone (Douglas et al., 2020).

In contrast, shotgun metagenomic sequencing amplifies all genomic DNA in a sample. These data enable the characterization of the gene content of an entire community by annotating reads. Metagenomics gives a much more complete picture of the genomic structure of a community but several technical hurdles limit this approach. First, annotating reads mapping to genes remains a challenge and roughly 30–50% of the open reading frames are annotated, leaving much of the genomic content unclassified in terms of function. Second, assigning annotated genes to specific taxa within the community and inferring their relative abundances remains a hard problem. In particular, assembling reads into genomes (metagenome-assembled genomes, MAGs) has been performed but the quality of these assemblies remains hard to assess. Applying cutting edge machine learning methods is likely to improve this process (Nissen et al., 2021). Another method to analyze shotgun metagenomics data is to use a database of reference genomes as templates to recruit reads from a metagenome. The upside of this approach is the ability to reliably detect variation at the level of single nucleotides (Garud et al., 2019), and reliably assemble genomes from metagenomes. The cost of course is that the method misses any diversity that is not present in the reference genome data-base. Despite these challenges, metagenomics perhaps gives clearest picture of the genomic structure of a community as a whole.

Transcriptional profiling of entire communities is also feasible via RNA-sequencing based methods (Antunes et al., 2016; Zhang et al., 2021). Despite the potential predictive power of knowing which genes in a community are transcriptionally active, these measurements have been applied much less widely than taxonomic amplicon or shotgun metagenomic methods. However, as the costs of sequencing continue to fall, it remains a compelling proposition to use metagenomics and transcriptional profiling on the same samples. We propose that such measurements could very well lead to deeper insights into the com-munity structure and function by potentially simplifying the picture. For example, transcriptional profiling could reveal which collective components of the metagenome are inactive and therefore could potentially be left out of a predictive framework.

LEARNING FROM DATA: FUNCTION FROM STRUCTURE

Equipped with measurements of metabolites (either dynamically or at a fixed point in time) and some characterization of the community structure, we are then left to ask what to do with these data. In reality, the answer to this question is an empirical one that depends on the structure the data, the model system under study and the precise question being posed. We offer no pipeline or prescription for how best to proceed, but instead offer a few suggestions and examples and point out some important technical pitfalls. Our intention is to suggest some approaches to learning the structure-function mapping from these data and to leverage the results for predicting community metabolism. As with most data analysis tasks, simple is better. Using the latest methods in machine learning or dimension reduction may be tempting but it is almost always better to explore the data with methods that are simple to interpret and straightforward to implement.

Typically, any sequencing based characterization of community structure will be high dimensional. For example, 16S amplicon sequencing will often yield 10 to 1000 of taxa per sample, similarly metagenomic data can contain 10,000 or more annotated open reading frames depending on the complexity of the community. In contrast, assembling an ensemble of more than 100 or 1000 communities is a huge challenge even for the highest throughput methodologies. As a result, we are almost always in the limit of a small number of data points (communities) and large number of variables (taxa, genes, transcripts, etc.). In this regard, predicting functional properties from these structural data requires reducing the

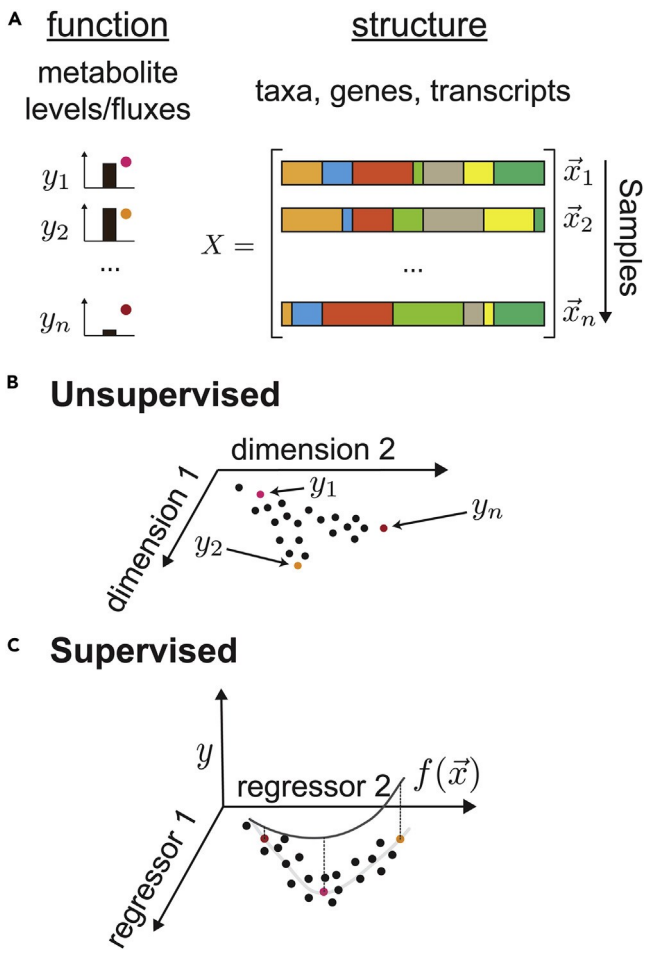

Figure 5. Learning the structure-function map from data

(A) Hypothetical structure-function data from an ensemble of n communities. y denotes a metabolite measurement, either level or rate that could also be dynamic. Colored dots correspond to data points in (B and C). X denotes a matrix of n rows each denoting a single community in an ensemble. The columns denote the relative abundances (colored bars) of taxa, genes in the metagenome or transcripts in the metatranscriptome. (B) An unsupervised approach where dimensionality reduction is applied to X yielding a lower dimensional representation of community structure that is then associated with communities of differing function. (C) A supervised approach where the function fðxÞ is learned for mapping structural variation to functional variation. Regressors denote independent variables in a lower dimensional representation of X that provides good predictive power of y.

dimensionality of the data describing community structure. For the purposes of the discussion below we define the number of features (genes, taxa, transcripts) as p and the number of communities in a given ensemble n (Figure 5A). A sequencing dataset can then be described by a matrix X that is n3p with n p in most cases. The rows of this matrix, x_i correspond to sequencing data for the ith community in the dataset. The entries of this vector are then the number of reads mapping to a specific ASV in a 16S dataset or gene in the metagenome of that community. For each of these communities we assume that the dataset includes some functional measurement, y_i, which may be a dynamic quantity ($y_i$ðtÞ). The goal then is to learn a representation of this functional measurement of the community in terms of the columns of X.

Compositional data and zero counts

All of the standard sequencing methods for quantifying community structure result in compositional data – that is, they do not report the absolute abundances of taxa, genes or transcripts in the sample, but only the relative contribution (e.g., the rows of X are defined only up to an unknown constant). Recently, methods using qPCR or the addition of oligos at known concentrations have been developed to measure absolute

abundances via sequencing. However, as yet these methods are not widespread. Therefore, in any analysis we must contend with the compositional nature of sequencing data. Much has been written about this problem (Gloor et al., 2017), and members of the field are now generally aware of the issues that can arise when the compositional nature of the data are ignored.

Briefly, compositional data can, and should, be log-ratio transformed using log ratios of counts. Log-ratio transformations take compositional data from a simplex and map them to real numbers with the properties of a vector space. This transformation therefore permits the application of conventional statistical approaches to compositional data. Typically, this is done via the center-log transform (CLR) or an additive log-transform (ALR) (Aitchison et al., 2000; Gloor et al., 2017). Computing log-ratio is not compatible with zeros (e.g., zero abundances of an ASV or gene transcript), a problem that has received a significant amount of attention. A host of methods from adding pseudocounts uniformly to all zeros or using Bayesian approaches to replacing zeros (Aitchison, 1982; Love et al., 2014; McMurdie and Holmes, 2014) have been proposed. We urge caution here, as many methods are both ad hoc and can qualitatively impact the results of downstream analyses. We will not review the technical details, but readers should engage carefully with their data rather than blindly applying existing pipelines for log-transforms and handling zeros. For example, variance decompositions applied to log-transformed data can be dominated by large numbers of taxa or transcripts with zero counts. In this scenario, the details of how zeros are handled (e.g., the magni-tude of the pseudocounts added to all taxa) can have huge impacts on the variance decomposition. More recently, phylogenetically aware methods have been developed, which may enable more reliable decom-position of relative abundance data (Silverman et al. (2017)).

Dimension reduction: with or without supervision

The goal is to associate structural components or sets of components with the metabolic function of a community. As discussed above we are almost always in the limit of small numbers of data points. We are therefore forced to consider reducing the dimensionality of the data from p by some form of dimension reduction. Above we suggested that low-dimensionality is a common feature of biological systems from proteins to higher organisms and behavior. Despite the fact that this observation seems to hold quite broadly, it is important that we not take low-dimensionality as given in any analysis of a microbial community. In this sense, we must make a principled search for a simpler description of community structure while judiciously considering the possibility that no such description exists or that we are considering the system at the wrong level of organization. For example, in an ensemble of communities with strong functional redundancy, where very distinct taxa perform similar metabolic functions, it may not be possible to find a low-dimensional description of the ensemble in terms of taxa present across replicates in the ensemble.

There are two ways to go about approaching this problem: supervised and unsupervised dimension reduction (Figures 5B and 5C). We note that for most of the techniques described below, the statistical learning textbook from Hastie, Tibshirani and Friedman (Hastie et al., 2016) provides an excellent and readable reference. For a more recent review article we recommend the paper of Mehta et al. (2019) that covers both supervised and unsupervised methods.

Unsupervised learning

In unsupervised dimension reduction, we seek a lower dimensional representation of the matrix X without any explicit consideration given to y in selecting low-dimensional features (Figure 5B). The idea is to reduce the dimensionality of X and then examine the relationship between this lower dimensional representation of community structure and the function, $\vec{y}$. The hope is that some lower dimensional representation of the community captures all of the useful information in the matrix X. All of these methods change the basis of X in order to recast community structure in terms of groups of genes or taxa. The dimension reduction arises when a small number of such groups of taxa provide a good description of all communities in the dataset. If this is case, each community $\vec{x}$ can be represented not in terms of the abundances of p taxa, but instead as a combination of m p groups or combinations of taxa. Formally, these methods are based on a matrix decomposition (X = AB, with A an n3m matrix and B an m3p matrix) subjected to constraints. Dimension reduction comes about if X can be well approximated for m p. In this case, each community in an ensemble can be represented in terms of just m features. These low-dimensional representations, when they exist, can sometimes dramatically simplify our understanding of complex systems.

The canonical unsupervised methods are variance-based decompositions like principal component analysis (PCA) or the more general version, singular-value decomposition (SVD). These methods find the set of orthogonal directions in the p-dimensional feature space that maximize the variance in the data along each direction (eigenvectors of the covariance matrix of X). The advantage here is the clean interpretability of these decompositions. Each direction (principal component) is a simple p-dimensional vector represent-ing a direction of high data variance. Each data point (community) in the original dataset can be repre-sented in the basis of principal components $x_i, p_1; \ldots; x_i, p_j$. (with p_j the principle components). If a small number of principal components capture a substantial fraction of the data variance, then m is small and each community can be represented as projections on a handful of p_j components.

Many modifications to this basic idea exist. Perhaps the two most notable are independent component analysis (ICA) and non-negative matrix factorization (NMF). ICA finds a lower dimensional representation of X in terms of components that are statistically independent, rather than simply uncorrelated, through an iterative process. The approach is to represent each community x as a sum of statistically independent components (Hyvärinen and Oja, 1997). In this case $x_i = \sum_j a_{i,j} s_j$ where the $a_{i,j}$ are the weights that 'mix' the independent components s to form the observed data x. ICA is most often applied to signal separation problems where multiple independent inputs are combined. One can speculate that this perspective may be useful for communities if they possess a modular organization where independent modules within the community are responsible for functionally distinct processes. NMF is another matrix decomposition method that is applicable whenever the entries of X are constrained to be positive (as is the case in sequencing data). NMF decomposes the data as $X \approx WH$ with all entries of W and H positive (Lee and Seung, 1999). In this case $x_i = \sum_j w_{i,j} h_j$ where the w are weights and the h are vectors of length p. The advantage of this approach is that the columns of H, that act as 'eigen-communities' contain all positive entries and are therefore interpretable. In contrast, the eigenvectors of a PCA decomposition can contain negative entries, which is not interpretable in the context of X that contains only positive values (abundances or relative abundances). NMF has seen limited application in the microbiome context (Cai et al., 2017). We have focused here on well-established methods with simple interpretations. We are aware that over the past two decades many new methods have been developed, especially those that can learn low-dimensional representations of highly non-linear data (e.g., autoencoders (Kramer, 1991) or stochastic neighbor embedding methods (Hinton and Roweis, 2002)). These methods may be useful in the context discussed here, but we advocate starting with the simpler approaches discussed above before moving on to these methods.

Regardless of the method of unsupervised learning applied, the result is a new representation of community structure (X) in a new basis (e.g., principal components, p, s, h). Ideally, m p and communities with many hundreds or thousands of genes or taxa can be represented in a much lower dimensional space (Fig-ure 5B). The task then remains to associate this lower dimensional representation with metabolic function (y_i). A common approach to this problem is simply to ask which basis vectors correlate with specific meta-bolic properties of the community. One common approach is to treat the question as a regression problem to predict y_i from the decomposed x_i for example by using the projections of each x_i along each principal component as independent variables.

Supervised learning

One major shortcoming of the unsupervised approach is that the low-dimensional representation of X that we learn by unsupervised dimension reduction may not be the best representation of the data in terms of predicting y_i. In essence, PCA or NMF may find a low dimensional representation of the data, but there is no reason or guarantee that this representation will be informative of the community metabolic function. Indeed, the unsupervised approach artificially separates the process of finding low-dimensional descriptions of the community and predicting the response variable (metabolic function). A supervised approach overcomes these limitations by performing both dimension reduction and prediction at the same time.

In the supervised approach we seek some prediction of the metabolic function in terms of the structure (Figure 5C). Concretely, we would like to estimate $y_i = f(x_i)$ for our entire dataset. This can be posed as a regression problem where we either posit a functional form for $f(x_i)$ or, using more flexible but less interpretable methods like neural networks, learn a mapping from x_i to y_i without positing a specific functional form.

Before discussing how one can approach this problem we would like to clarify the meaning of $f(\vec{x}_i)$. We are proposing learning a statistical map from $\vec{x}_i$ to y_i. We are not proposing fitting an explicit ecological model such as a consumer-resource or Lotka-Volterra model to the data. We regard fitting such a model as a much harder proposition than learning a statistical mapping from structure to function. Indeed, a statistical approach to learning f necessarily abstracts away these dynamics that relate $\vec{x}_i$ to y_i. We note that two of us recently took a hybrid approach to the problem with synthetic communities that did explicitly model the ecological dynamics (Gowda et al., 2022). In this case, we used a regression to map gene content to consumer-resource model parameters and then used the consumer-resource model to predict metabolite dynamics in consortia. Remarkably, the approach worked, but the downside is that it requires isolating individual taxa and constructing synthetic communities, a more laborious task than studying communities directly.

There are many statistical approaches to learning the function f. The simplest approaches are linear regression methods that simply posit a model of the form $f(\vec{x}_i) = b_0 + \sum_{k=1}^{p} b_k x_{i;k}$, where the b are regression coefficients. There are two major problems with this approach. First, if $n \ll p$ we have many fewer data points than independent variables, which means that an ordinary least-squares regression will almost certainly overfit and yield poor out of sample predictions (as determined by cross validation). The second related issue is that this approach gives equal weight to each entry (gene, taxon) in $\vec{x}$ and does not provide any dimensional reduction. One way to solve this problem is via regularization (Hastie et al., 2016) where the model is optimized with an additional penalty term that seeks to reduce the number of non-zero b coefficients (see LASSO and Ridge Regression). Regularization provides a solution to the problem of selecting which regressors (entries of $\vec{x}_i$) provide the most predictive power while also avoiding overfitting. In situations where the levels of noise are not too high and a sparse solution (small number of non-zero b_k) do allow for good predictions, these regularization methods typically succeed (Fraebel et al., 2020). However, if the noise levels are high or the underlying process is not sparse, then even these methods will fail. Care must be taken in diagnosing when such a regression works and when it does not, see (Fraebel et al., 2020; Gowda et al., 2022).

A major shortcoming of the simple formulation outlined above is that it lacks any interaction terms (e.g., $x_{k;i}x_{l;i}$). Adding these terms to the regression above increases the number of independent variables from p to $p + p^2 = 2$. Given limited data n, it is typically not advised to take this approach. However, including such interaction terms is desirable given the utility of considering pair-wise interactions in complex systems (Bialek and Ranganathan, 2007; Schneidman et al., 2006). One way to proceed is to use linear regression approaches that use groups of independent variables as regressors. For example, principal component regression (Hastie et al., 2016) uses principal components as independent variables in a linear regression. This is qualitatively similar to the unsupervised approach outlined above.

The models above are linear, and this aids in their interpretation. However, explicitly non-linear methods for estimating f are also possible. When and why such approaches are more or less appropriate in the microbial context is not yet clear. However, decision tree based methods such as random forests have proven useful for relating taxonomic variation to host phenotypes in the microbiome (Blanton et al., 2016). Random forests can model complex non-linear relationships between regressors and response variables (Hastie et al., 2016), while retaining some interpretability by assigning 'importance scores' to each independent variable. As a result, these methods can be used to assess the impact of a given taxon or gene on the community function (in a statistical, not causal, sense). Random forest regressions fit many decision trees to bootstrapped replicates of the data and average the result. This averaging procedure, often called 'bagging', reduces the variance of the prediction and as a result, high variance/low bias regression approaches can provide lower variance predictions.

Finally, as the amount of high quality data on microbial communities increases, the applicability of more recently developed neural network supervised prediction methods (Mehta et al., 2019) will become more appropriate. These methods typically contain many millions or even billions of parameters and therefore require reasonably sized datasets to train. Advances in methods such as transfer learning (Weiss et al., 2016), where existing trained networks are trained to solve a new problem given some data, mean that the user need not start from scratch. The challenge will be what we can learn from these networks once they are used to approximate f. In many cases neural networks do not generalize well and are susceptible to small amounts of noise on the input variables. We regard the application and interpretation of these network

approaches to communities a problem at the forefront. It may be that we need to reconsider how networks are trained to properly learn the salient features of the structure function problem in microbial communities (Blazek and Lin, 2021).

WHAT WE DO AND DO NOT LEARN FROM A STATISTICAL APPROACH

What can the ensemble approach coupled with a statistical analysis like the one described above teach us about communities? We contend that by looking at an ensemble of well-chosen communities, and learning the main statistical features of community structure that determine function we can begin to learn what general properties of communities must be present to admit their functional properties. A handful of recent studies have begun to show the power of this approach (Goldford et al., 2018; Blanton et al., 2016; Raman et al., 2019; Gowda et al., 2022). What we recover from these studies is what reproducible features of communities are retained across replicate consortia. These reproducible features can be regarded as 'good variables' for predicting community function from structure. In some cases, understanding these variables lets us control or predict the functional properties of consortia in synthetic systems (Gowda et al., 2022) and in hosts (Blanton et al., 2016). Ultimately, we hope this approach can be used to design, predict and control microbial consortia in engineered and wild contexts to address the existential threat of climate change.

One limitation of the ensemble approach is that it requires variation in structure and function across the ensemble (Figure 5). One can imagine little variation in community structure might occur for metabolic processes that are performed by only a few closely related taxa (e.g., nitrification). In this case, one approach would be to examine allelic variation in the enzymes present or variation in transcriptional regulation of relevant pathways. However, it may also be that more direct mechanistic approaches are more appropriate when the ensemble exhibits small levels of variation.

However, even when these statistical approaches succeed in predicting community structure from function we often do not understand why, at a mechanistic level, the prediction succeeds. For example, in the case of (Gowda et al., 2022) the reason for the success of the regression from gene content to phenotypes is not entirely clear. Nor do we believe that the regression can predict the impact of gene gain or loss mutations. Similarly, in the case of (Blanton et al., 2016) the precise metabolic role of each bacterial taxon in the stunt-ing of the host is unclear. In essence, the statistical approach lets us find good variables for design and con-trol of communities, but it does not, by itself, tell us why these are good variables.

Understanding the mechanistic basis of community function requires a more detailed look under the hood of a community. In certain cases, where we have some knowledge of the metabolic players in a consortium, metabolic modeling such as flux balance analysis can be a powerful tool to achieve this goal. Recent successes include descriptions of syntrophy in methanogenic communities (Embree et al., 2015) and mutualism in spatially structured communities (Harcombe et al., 2014).

FUTURE DIRECTIONS: EVOLUTIONARY RULES OF THE STRUCTURE-FUNCTION MAPPING

So, while the ensemble approach can help us solve the structure-function problem the deeper question of why nature constructs communities the way it does remains. We argue that the answer to this question will require considering the eco-evolutionary basis of the observed structure function mapping. Addressing this question is subject enough for a separate manuscript. However, we hope that through the careful application of quantitative methods, like those discussed here, to some of the model systems discussed above, we can open the door to understanding how nature constructs dynamic functional consortia.

ACKNOWLEDGMENTS

The authors would like to acknowledge National Science Foundation funding from grants: MCB 2117477, EF 2025293, and MCB 1921439.

AUTHOR CONTRIBUTIONS

C.G., K.G., K.P., S.K.: Conceptualization, Writing - original Draft, Writing - review &Editing. S.K.: Supervision.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Aitchison, J. (1982). The statistical analysis of compositional data. J. R. Stat. Soc. Ser. B (Methodological) 44, 139–160.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A., and Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. Math. Geol. 32, 271–275.

Aiyar, P., Schaeme, D., García-Altares, M., Carrasco Flores, D., Dathe, H., Hertweck, C., Sasso, S., and Mittag, M. (2017). Antagonistic bacteria disrupt calcium homeostasis and immobilize algal cells. Nat. Commun. 8, 1756.

Alba, V., Carthew, J.E., Carthew, R.W., and Mani, M. (2021). Global constraints within the developmental program of the Drosophila wing. eLife 10, e66750.

Alldredge, A.L., and Silver, M.W. (1988). Characteristics, dynamics and significance of marine snow. Prog. Oceanogr. 20, 41–82.

Allison, S.D., Wallenstein, M.D., and Bradford, M.A. (2010). Soil-carbon response to warming dependent on microbial physiology. Nat. Geosci. 3, 336–340.

Alon, U., Surette, M.G., Barkai, N., and Leibler, S. (1999). Robustness in bacterial chemotaxis. Nature 397, 168–171.

Alseekh, S., Aharoni, A., Brotman, Y., Contrepois, K., D'Auria, J., Ewald, J., Ewald, J.C., Fraser, P.D., Giavalisco, P., Hall, R.D., et al. (2021). Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. Nat. Methods 18, 747–756.

Amann, J. (1911). Die direkte Zahlung der Wasserbakterien mittels des Ultramikroskops. Centralbl. f. Bakteriol. 29, 381–384.

Amaral-Zettler, L.A., Zettler, E.R., and Mincer, T.J. (2020). Ecology of the plastisphere. Nat. Rev. Microbiol. 18, 139–151.

Amarnath, K., Narla, A.V., Pontrelli, S., Dong, J., Caglar, T., Taylor, B.R., Schwartzman, J., Sauer, U., Cordero, O.X., and Hwa, T. (2021). Stress-induced cross-feeding of internal metabolites provides a dynamic mechanism of microbial cooperation. bioarxiv. https://doi.org/10.1101/2021.06.24.449802.

Anderson, K.L., Tayne, T.A., and Ward, D.M. (1987). Formation and fate of fermentation products in hot spring cyanobacterial mats. Appl. Environ. Microbiol. 53, 2343–2352.

Anderson, P.W. (1972). More is different. Science 177, 393–396.

Andrade-Domínguez, A., Salazar, E., del Carmen Vargas-Lagunas, M., Kolter, R., and Encarnación, S. (2014). Eco-evolutionary feedbacks drive species interactions. ISME J. 8, 1041–1054.

Antunes, L.P., Martins, L.F., Pereira, R.V., Thomas, A.M., Barbosa, D., Lemos, L.N., Silva, G.M.M., Moura, L.M.S., Epamino, G.W.C., Digiampietri, L.A., et al. (2016). Microbial community structure and dynamics in thermophilic composting viewed through metagenomics and metatranscriptomics. Scientific Rep. 6, 38915.

Aoi, Y., Kinoshita, T., Hata, T., Ohta, H., Obokata, H., and Tsuneda, S. (2009). Hollow-fiber membrane chamber as a device for in situ environmental cultivation. Appl. Environ. Microbiol. 75, 3826–3833.

Bahram, M., Hildebrand, F., Forslund, S.K., Anderson, J.L., Soudzilovskaia, N.A., Bodegom, P.M., Bengtsson-Palme, J., Anslan, S., Coelho, L.P., Harend, H., et al. (2018). Structure and function of the global topsoil microbiome. Nature 560, 233–237.

Basan, M., Honda, T., Christodoulou, D., Horl, M., Chang, Y.-F., Leoncini, E., Mukherjee, A., Okano, H., Taylor, B.R., Silverman, J.M., et al. (2020). A universal trade-off between growth and lag in fluctuating environments. Nature 584, 470–474.

Bateson, M.M., and Ward, D.M. (1988). Photoexcretion and fate of glycolate in a hot spring cyanobacterial mat. Appl. Environ. Microbiol. 54, 1738–1743.

Beale, D.J., Pinu, F.R., Kouremenos, K.A., Poojary, M.M., Narayana, V.K., Boughton, B.A., Kanojia, K., Dayalan, S., Jones, O.A.H., and Dias, D.A. (2018). Review of recent developments in GC–MS approaches to metabolomics-based research. Metabolomics 14, 152.

Beliaev, A.S., Romine, M.F., Serres, M., Bernstein, H.C., Linggi, B.E., Markillie, L.M., Isern, N.G., Chrisler, W.B., Kucek, L.A., Hill, E.A., et al. (2014). Inference of interactions in cyanobacterial–heterotrophic co-cultures via transcriptome sequencing. ISME J. 8, 2243–2255.

Bennett, B.D., Kimball, E.H., Gao, M., Osterhout, R., Van Dien, S.J., and Rabinowitz, J.D. (2009). Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. Nat. Chem. Biol. 5, 593–599.

Berdy, B., Spoering, A.L., Ling, L.L., and Epstein, S.S. (2017). In situ cultivation of previously uncultivable microorganisms using the ichip. Nat. Protoc. 12, 2232–2242.

Berman, G.J., Choi, D.M., Bialek, W., and Shaevitz, J.W. (2014). Mapping the stereotyped behaviour of freely moving fruit flies. J. R. Soc. Interf. 11, 20140672.

Bialek, W., and Ranganathan, R. (2007). Rediscovering the power of pairwise interactions. arXiv. https://arxiv.org/abs/0712.4397.

P.Blainey, A.Kulesa, and J.Kehe. (2018), Massively Parallel On-Chip Coalescence of Microemulsions. US Patent US20180071738A1,

Blanton, L.V., Charbonneau, M.R., Salih, T., Barratt, M.J., Venkatesh, S., Ilkaveya, O., Subramanian, S., Manary, M.J., Trehan, I., Jorgensen, J.M., et al. (2016). Gut bacteria that prevent growth impairments transmitted by microbiota from malnourished children. Science 351, aad3311.

Blasche, S., Kim, Y., Mars, R.A., Machado, D., Maansson, M., Kafkia, E., Milanese, A., Zeller, G., Teusink, B., Nielsen, J., et al. (2021). Metabolic cooperation and spatiotemporal niche partitioning in a kefir microbial community. Nat. Microbiol. 6, 196–208.

Blazek, P.J., and Lin, M.M. (2021). Explainable neural networks that simulate reasoning. Nat. Comput. Sci. 1, 607–618.

Bocher, B.T.W., Cherukuri, K., Maki, J.S., Johnson, M., and Zitomer, D.H. (2015). Relating methanogen community structure and anaerobic digester function. Water Res. 70, 425–435.

Bock, E., and Wagner, M. (2013). Oxidation of Inorganic Nitrogen Compounds as an Energy Source (Springer), pp. 83–118.

Bollmann, A., Lewis, K., and Epstein, S.S. (2007). Incubation of environmental samples in a diffusion chamber increases the diversity of recovered isolates. Appl. Environ. Microbiol. 73, 6386–6390.

Borges, J.L., and Hurley, A. (1998). Collected Fictions (Viking).

Browne, H.P., Forster, S.C., Anonye, B.O., Kumar, N., Neville, B.A., Stares, M.D., Goulding, D., and Lawley, T.D. (2016). Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. Nature 533, 543–546.

Buchan, A., LeCleir, G.R., Gulvik, C.A., and Gonzalez, J.M. (2014). Master recyclers: features and functions of bacteria associated with phytoplankton blooms. Nat. Rev. Microbiol. 12, 686–698.

Cai, Y., Gu, H., and Kenney, T. (2017). Learning microbial community structures with supervised and unsupervised non-negative matrix factorization. Microbiome 5, 110.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. Nat. Methods 13, 581–583.

Callbeck, C.M., Lavik, G., Ferdelman, T.G., Fuchs, B., Gruber-Vodicka, H.R., Hach, P.F., Littmann, S., Schoffelen, N.J., Kalvelage, T., Thomsen, S., et al. (2018). Oxygen minimum zone cryptic sulfur cycling sustained by offshore transport of key sulfur oxidizing bacteria. Nat. Commun. 9, 1–11.

Canfield, D.E., Stewart, F.J., Thamdrup, B., De Brabandere, L., Dalsgaard, T., Delong, E.F., Revsbech, N.P., and Ulloa, O. (2010). A cryptic sulfur cycle in oxygen-minimum-zone waters off the Chilean Coast. Science 330, 1375–1378.

Carlson, H.K., Lui, L.M., Price, M.N., Kazakov, A.E., Carr, A.V., Kuehl, J.V., Owens, T.K., Nielsen, T., Arkin, A.P., and Deutschbauer, A.M. (2020). Selective carbon sources influence the end

products of microbial nitrate respiration. ISME J. 14, 2034–2045.

Chaudhary, D.K., Khulan, A., and Kim, J. (2019). Development of a novel cultivation technique for uncultured soil bacteria. Scientific Rep. 9, 6666.

Crits-Christoph, A., Diamond, S., Butterfield, C.N., Thomas, B.C., and Banfield, J.F. (2018). Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis. Nature 558, 440–444.

Cross, K.L., Campbell, J.H., Balachandran, M., Campbell, A.G., Cooper, S.J., Griffen, A., Heaton, M., Joshi, S., Klingeman, D., Leys, E., et al. (2019). Targeted isolation and cultivation of uncultivated bacteria by reverse genomics. Nat. Biotechnol. 37, 1314–1321.

Cujia, K.S., Boss, J.M., Herb, K., Zopes, J., and Degen, C.L. (2019). Tracking the precession of single nuclear spins by weak measurements. Nature 571, 230–233.

Cydzik-Kwiatkowska, A., and Zielinska, M. (2016). Bacterial communities in full-scale wastewater treatment systems. World J. Microbiol. Biotechnol. 32, 66.

Date, Y., Nakanishi, Y., Fukuda, S., Kato, T., Tsuneda, S., Ohno, H., and Kikuchi, J. (2010). New monitoring approach for metabolic dynamics in microbial ecosystems using stable-isotope-labeling technologies. J. Biosci. Bioeng. 110, 87–93.

Datta, M.S., Sliwerska, E., Gore, J., Polz, M.F., and Cordero, O.X. (2016). Microbial interactions lead to rapid micro-scale successions on model marine particles. Nat. Commun. 7, 11965.

de Jesús Astacio, L.M., Prabhakara, K.H., Li, Z., Mickalide, H., and Kuehn, S. (2021). Closed microbial communities self-organize to persistently cycle carbon. Proc. Natl. Acad. Sci. 118, e2013564118.

Dettmer, K., Aronov, P.A., and Hammock, B.D. (2007). Mass spectrometry-based metabolomics. Mass Spectrom. Rev. 26, 51–78.

Douglas, G.M., Maffei, V.J., Zaneveld, J.R., Yurgel, S.N., Brown, J.R., Taylor, C.M., Huttenhower, C., and Langille, M.G.I. (2020). PICRUSt2 for prediction of metagenome functions. Nat. Biotechnol. 38, 685–688.

Dussud, C., Meistertzheim, A.L., Conan, P., Pujo-Pay, M., George, M., Fabre, P., Coudane, J., Higgs, P., Elineau, A., Pedrotti, M.L., et al. (2018). Evidence of niche partitioning among bacteria living on plastics, organic particles and surrounding seawaters. Environ. Pollut. 236, 807–816.

Dyksma, S., Bischof, K., Fuchs, B.M., Hoffmann, K., Meier, D., Meyerdierks, A., Pjevac, P., Probandt, D., Richter, M., Stepanauskas, R., and Mußmann, M. (2016). Ubiquitous Gammaproteobacteria dominate dark carbon fixation in coastal sediments. ISME J. 1–15.

Ebrahimi, A., Schwartzman, J., and Cordero, O.X. (2019). Cooperation and spatial self-organization determine rate and efficiency of particulate organic matter degradation in marine bacteria. Proc. Natl. Acad. Sci. 116, 23309–23316.

Eckmann, J.-P., and Tlusty, T. (2021). Dimensional reduction in complex living systems: where, why, and how. BioEssays 43, 2100062.

Embree, M., Liu, J.K., Al-Bassam, M.M., and Zengler, K. (2015). Networks of energetic and metabolic interactions define dynamics in microbial communities. Proc. Natl. Acad. Sci. 112, 15450–15455.

Emwas, A.-H., Roy, R., McKay, R.T., Tenori, L., Saccenti, E., Gowda, G.A.N., Raftery, D., Alahmari, F., Jaremko, L., Jaremko, M., and Wishart, D.S. (2019). NMR spectroscopy for metabolomics research. Metabolites 9.

Enke, T.N., Datta, M.S., Schwartzman, J., Cermak, N., Schmitz, D., Barrere, J., Pascual-García, A., and Cordero, O.X. (2019). Modular assembly of polysaccharide-degrading marine microbial communities. Curr. Biol. 29, 1528–1535.e6.

Enning, D., and Garrelfs, J. (2014). Corrosion of iron by sulfate-reducing bacteria: new views of an old problem. Appl. Environ. Microbiol. 80, 1226–1236.

Escoriza, M.F., Vanbriesen, J.M., Stewart, S., and Maier, J. (2006). Studying bacterial metabolic states using Raman spectroscopy. Appl. Spectrosc. 60, 971–976.

Estrela, S., Vila, J.C., Lu, N., Bajic, D., Rebolleda-Gómez, M., Chang, C.-Y., Goldford, J.E., Sanchez-Gorostiaga, A., and Sánchez, Álvaro (2021). Functional attractors in microbial community assembly. Cell Syst. https://doi.org/10.1016/j.cels.2021.09.011.

Falkowski, P., Scholes, R.J., Boyle, E., Canadell, J., Canfield, D., Elser, J., Gruber, N., Hibbard, K., Hogberg, P., Linder, S., et al. (2000). The global carbon cycle: a test of our knowledge of Earth as a system. Science 290, 291–296.

Falkowski, P.G. (1994). The role of phytoplankton photosynthesis in global biogeochemical cycles. Photosyn. Res. 39, 235–258.

Fernández, A., Huang, S., Seston, S., Xing, J., Hickey, R., Criddle, C., and Tiedje, J. (1999). How stable is stable? Function versus community composition. Appl. Environ. Microbiol. 65, 3697–3704.

Fernandez, A.S., Hashsham, S.A., Dollhopf, S.L., Raskin, L., Glagoleva, O., Dazzo, F.B., Hickey, R.F., Criddle, C.S., and Tiedje, J.M. (2000). Flexible community structure correlates with stable community function in methanogenic bioreactor communities perturbed by glucose. Appl. Environ. Microbiol. 66, 4058–4067.

Fierer, N. (2017). Embracing the unknown: disentangling the complexities of the soil microbiome. Nat. Rev. Microbiol. 15, 579–590.

Foster, K.R., and Bell, T. (2012). Competition, not cooperation, dominates interactions among culturable microbial species. Curr. Biol. 22, 1845–1850.

Fraebel, D.T., Gowda, K., Mani, M., and Kuehn, S. (2020). Evolution of generalists by phenotypic plasticity. iScience 23, 101678.

Frentz, Z., Kuehn, S., and Leibler, S. (2015). Strongly deterministic population dynamics in closed microbial communities. Phys. Rev. X 5, 041014.

Friedman, J., Higgins, L.M., and Gore, J. (2017). Community structure follows simple assembly rules in microbial microcosms. Nat. Ecol. Evol. 1, 1034.

Fuwa, H. (1954). A new method for microdetermination Cf amylase activity by the use of amylose as the substrate. J. Biochem. 41, 583–603.

Garud, N.R., Good, B.H., Hallatschek, O., and Pollard, K.S. (2019). Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. PLoS Biol. 17, e3000102.

Ghosh, A., Nilmeier, J., Weaver, D., Adams, P.D., Keasling, J.D., Mukhopadhyay, A., Petzold, C.J., and Martín, H.G. (2014). A peptide-based method for 13C metabolic flux analysis in microbial communities. PLoS Comput. Biol. 10, e1003827.

Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V., and Egozcue, J.J. (2017). Microbiome datasets are compositional: and this is not optional. Front. Microbiol. 8, 2224.

Goldford, J.E., Lu, N., Bajic, D., Estrela, S., Tikhonov, M., Sanchez-Gorostiaga, A., Segrè, D., Mehta, P., and Sanchez, A. (2018). Emergent simplicity in microbial community assembly. Science 361, 469–474.

Goldschmidt, F., Regoes, R.R., and Johnson, D.R. (2018). Metabolite toxicity slows local diversity loss during expansion of a microbial cross-feeding community. ISME J. 12, 136–144.

Gonzalez-Gil, G., Thomas, L., Emwas, A.-H., Lens, P.N.L., and Saikaly, P.E. (2015). NMR and MALDI-TOF MS based characterization of exopolysaccharides in anaerobic microbial aggregates from full-scale reactors. Scientific Rep. 5, 14316.

Gowda, K., Ping, D., Mani, M., and Kuehn, S. (2022). Genomic structure predicts metabolite dynamics in microbial communities. Cell 185. https://doi.org/10.1016/j.cell.2021.12.036.

Graham, E.B., Knelman, J.E., Schindlbacher, A., Siciliano, S., Breulmann, M., Yannarell, A., Beman, J.M., Abell, G., Philippot, L., Prosser, J., et al. (2016). Microbes as engines of ecosystem function: when does community structure enhance predictions of ecosystem processes? Front. Microbiol. 7, 214.

Gralka, M., Szabo, R., Stocker, R., and Cordero, O.X. (2020). Trophic interactions and the drivers of microbial community assembly. Curr. Biol. 30, R1176–R1188.

Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. Cell 138, 774–786.

Harcombe, W., Riehl, W., Dukovski, I., Granger, B., Betts, A., Lang, A., Bonilla, G., Kar, A., Leiby, N., Mehta, P., et al. (2014). Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. Cell Rep. 7, 1104–1115.

Harz, M., Rosch, P., Peschke, K.-D., Ronneberger, O., Burkhardt, H., and Popp, J. (2005). Micro-Raman spectroscopic identification of bacterial cells of the genus Staphylococcus and dependence on their cultivation conditions. Analyst 130, 1543–1550.

Hashsham, S.A., Fernandez, A.S., Dollhopf, S.L., Dazzo, F.B., Hickey, R.F., Tiedje, J.M., and Criddle, C.S. (2000). Parallel processing of substrate correlates with greater functional stability in methanogenic bioreactor communities perturbed by glucose. Appl. Environ. Microbiol. 66, 4050–4057.

Hastie, T., Tibshirani, R., and Friedman, J. (2016). Elements of Statistical Learning Data, Second Edition (Springer).

Hatzenpichler, R., Krukenberg, V., Spietz, R.L., and Jay, Z.J. (2020). Next-generation physiology approaches to study microbiome function at single cell level. Nat. Rev. Microbiol. 18, 241–256.

Hekstra, D.R., and Leibler, S. (2012). Contingency and statistical laws in replicate microbial closed ecosystems. Cell 149, 1164–1173.

Hinton, G., and Roweis, S. (2002). Stochastic neighbor embedding. In Proceedings of the 15th International Conference on Neural Information Processing Systems, NIPS'02 (MIT Press), pp. 857–864.

Holm, J., Bjorck, I., Drews, A., and Asp, N.-G. (1986). A rapid method for the analysis of starch. Starch - Starke 38, 224–226.

Hom, E.F.Y., and Murray, A.W. (2014). Niche engineering demonstrates a latent capacity for fungal-algal mutualism. Science 345, 94–98.

Huang, Z.-R., Guo, W.-L., Zhou, W.-B., Li, L., Xu, J.-X., Hong, J.-L., Liu, H.-P., Zeng, F., Bai, W.-D., Liu, B., et al. (2019). Microbial communities and volatile metabolites in different traditional fermentation starters used for hong qu glutinous rice wine. Food Res. Int. 121, 593–603.

Hugerth, L.W., and Andersson, A.F. (2017). Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. Front. Microbiol. 8, 1561.

Hyvarinen, A., and Oja, E. (1997). A fast fixed-point Algorithm for independent component analysis. Neural Comput. 9, 1483–1492.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3, 318–356.

Jambeck, J.R., Geyer, R., Wilcox, C., Siegler, T.R., Perryman, M., Andrady, A., Narayan, R., and Law, K.L. (2015). Plastic waste inputs from land into the ocean. Science 347, 768–771.

Jehmlich, N., Schmidt, F., Taubert, M., Seifert, J., Bastida, F., von Bergen, M., Richnow, H.-H., and Vogt, C. (2010). Protein-based stable isotope probing. Nat. Protoc. 5, 1957–1966.

Jemal, M. (2000). High-throughput quantitative bioanalysis by lc/ms/ms. Biomed. Chromatogr. 14, 422–429.

Jin, Y., Li, D., Ai, M., Tang, Q., Huang, J., Ding, X., Wu, C., and Zhou, R. (2019). Correlation between volatile profiles and microbial communities: a metabonomic approach to study jiang-flavor liquor daqu. Food Res. Int. 121, 422–432.

Jordan, D., Kuehn, S., Katifori, E., and Leibler, S. (2013). Behavioral diversity in microbes and low-dimensional phenotypic spaces. Proc. Natl. Acad. Sci. 110, 14018–14023.

Junier, I., Frémont, P., and Rivoire, O. (2018). Universal and idiosyncratic characteristic lengths in bacterial genomes. Phys. Biol. 15, 035001.

Kaeberlein, T., Lewis, K., and Epstein, S.S. (2002). Isolating "uncultivable" microorganisms in pure culture in a simulated natural environment. Science 296, 1127–1129.

Katsov, A.Y., Freifeld, L., Horowitz, M., Kuehn, S., and Clandinin, T.R. (2017). Dynamic structure of locomotor behavior in walking fruit flies. eLife 6, e26410.

Kawabata, Z., Matsui, K., Okazaki, K., Nasu, M., Nakano, N., and Sugai, T. (1995). Synthesis of a species-defined microcosm with protozoa. J. Protozool. Res. 5, 23–26.

Kearns, E.A., and Folsome, C.E. (1981). Measurement of biological activity in materially closed microbial ecosystems. BioSystems 14, 205–209.

Kehe, J., Kulesa, A., Ortiz, A., Ackerman, C.M., Thakku, S.G., Sellers, D., Kuehn, S., Gore, J., Friedman, J., and Blainey, P.C. (2019). Massively parallel screening of synthetic microbial communities. Proc.Natl. Acad.Sci. 116, 12804–12809.

Kehe, J., Ortiz, A., Kulesa, A., Gore, J., Blainey, P.C., and Friedman, J. (2021). Positive interactions are common among culturable bacteria. Sci. Adv. 7, eabi7159.

Keiluweit, M., Wanzek, T., Kleber, M., Nico, P., and Fendorf, S. (2017). Anaerobic microsites have an unaccounted role in soil carbon stabilization. Nat. Commun. 8, 1771.

Kimbrel, J.A., Samo, T.J., Ward, C., Nilson, D., Thelen, M.P., Siccardi, A., Zimba, P., Lane, T.W., and Mayali, X. (2019). Host selection and stochastic effects influence bacterial community assembly on the microalgal phycosphere. Algal Res. 40, 101489.

Kiørboe, T. (2007). The Sea Core Sampler: a simple water sampler that allows direct observations of undisturbed plankton. J. Plankton Res. 29, 545–552.

Kiørboe, T., Tang, K., Grossart, H.P., and Ploug, H. (2003). Dynamics of microbial communities on marine snow aggregates: colonization, growth, detachment, and grazing mortality of attached bacteria. Appl. Environ. Microbiol. 69, 3036–3047.

Kirchman, D.L. (2012). Processes in Microbial Ecology (Oxford University Press).

Kirschbaum, M.U.F. (1995). The temperature dependence of soil organic matter decomposition, and the effect of global warming on soil organic C storage. Soil Biol. Biochem. 27, 753–760.

Kirstein, I.V., Wichels, A., Gullans, E., Krohne, G., and Gerdts, G. (2019). The Plastisphere –uncovering tightly attached plastic "specific" microorganisms. PLoS One 14, e0215859.

Klatt, C.G., Inskeep, W.P., Herrgard, M.J., Jay, Z.J., Rusch, D.B., Tringe, S., Niki Parenteau, M., Ward, D.M., Boomer, S.M., Bryant, D.A., and Community, S. Miller. (2013). Structure and function of high-temperature chlorophototrophic microbial mats inhabiting diverse geothermal environments. Front. Microbiol. 4, 106.

Knight, R., Vrbanac, A., Taylor, B.C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., et al. (2018). Best practices for analysing microbiomes. Nat. Rev. Microbiol. 16, 410–422.

Kobayashi-Kirschvink, K.J., Nakaoka, H., Oda, A., Kamei, K.-i. F., Nosho, K., Fukushima, H., Kanesaki, Y., Yajima, S., Masaki, H., Ohta, K., and Wakamoto, Y. (2018). Linear regression links transcriptomic data and cellular Raman spectra. Cell Syst. 7, 104–117.e4.

Kramer, M.A. (1991). Nonlinear principal component analysis using autoassociative neural networks. AIChE J. 37, 233–243.

Kuehn, S., Hickman, S.A., and Marohn, J.A. (2008). Advances in mechanical detection of magnetic resonance. J. Chem. Phys. 128, 052208.

Lagier, J.-C., Armougom, F., Million, M., Hugon, P., Pagnier, I., Robert, C., Bittar, F., Fournous, G., Gimenez, G., Maraninchi, M., et al. (2012). Microbial culturomics: paradigm shift in the human gut microbiome study. Clin. Microbiol. Infect. Official Publ. Eur. Soc. Clin. Microbiol. Infect. Dis. 18, 1185–1193.

Lagier, J.-C., Khelaifia, S., Alou, M.T., Ndongo, S., Dione, N., Hugon, P., Caputo, A., Cadoret, F., Traore, S.I., Seck, E.H., et al. (2016). Culture of previously uncultured members of the human gut microbiota by culturomics. Nat. Microbiol. 1, 1–8.

Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. Science 304, 1623–1627.

Lawley, T.D., Clare, S., Walker, A.W., Stares, M.D., Connor, T.R., Raisen, C., Goulding, D., Rad, R., Schreiber, F., Brandt, C., et al. (2012). Targeted restoration of the intestinal microbiota with a simple, defined bacteriotherapy resolves relapsing Clostridium difficile disease in mice. PLoS Pathog. 8, e1002995.

Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791.

Lee, J.Z., Everroad, R.C., Karaoz, U., Detweiler, A.M., Pett-Ridge, J., Weber, P.K., Prufert-Bebout, L., and Bebout, B.M. (2018). Metagenomics reveals niche partitioning within the phototrophic zone of a microbial mat. PLoS One 13, e0202792.

Levitt, M.H. (2008). Spin Dynamics: Basics of Nuclear Magnetic Resonance (Wiley).

Liamleam, W., and Annachhatre, A.P. (2007). Electron donors for biological sulfate reduction. Biotechnol. Adv. 25, 452–463.

Lilja, E.E., and Johnson, D.R. (2016). Segregating metabolic processes into different microbial cells accelerates the consumption of inhibitory substrates. ISME J. 10, 1568–1578.

Lilja, E.E., and Johnson, D.R. (2019). Substrate cross-feeding affects the speed and trajectory of molecular evolution within a synthetic microbial assemblage. BMC Evol. Biol. 19, 129.

Louati, I., Pascault, N., Debroas, D., Bernard, C., Humbert, J.-F., and Leloup, J. (2015). Structural diversity of bacterial communities associated with bloom-forming freshwater cyanobacteria differs according to the cyanobacterial genus. PLoS One 10, e0140614.

Louca, S., Polz, M.F., Mazel, F., Albright, M.B.N., Huber, J.A., O'Connor, M.I., Ackermann, M., Hahn, A.S., Srivastava, D.S., Crowe, S.A., et al. (2018). Function and functional redundancy in microbial systems. Nat. Ecol. Evol. 2, 936–943.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550.

Luan, T.G., Yu, K.S.H., Zhong, Y., Zhou, H.W., Lan, C.Y., and Tam, N.F.Y. (2006). Study of metabolites from the degradation of polycyclic aromatic hydrocarbons (PAHs) by bacterial consortium enriched from mangrove sediments. Chemosphere 65, 2289–2296.

Ly, S., Mith, H., Tarayre, C., Taminiau, B., Daube, G., Fauconnier, M.-L., and Delvigne, F. (2018). Impact of microbial composition of cambodian traditional dried starters (dombea) on flavor compounds of rice wine: Combining amplicon sequencing with hp-spme-gcms. Front. Microbiol. 9, 894.

Lycus, P., Lovise Bøthun, K., Bergaust, L., Peele Shapleigh, J., Reier Bakken, L., and Frostegard, A. (2017). Phenotypic and genotypic richness of denitrifiers revealed by a novel isolation strategy. ISME J. 11, 2219–2232.

Macnaughtan, M.A., Hou, T., Xu, J., and Raftery, D. (2003). High-throughput nuclear magnetic resonance analysis using a multiple coil flow probe. Anal. Chem. 75, 5116–5123.

Madigan, M.T., Kelly, S.B., Daniel, H., Sattley, W.M., and Stahl, D.A. (2018). Brock Biology of Microorganisms (Pearson).

Mandy, D.E., Goldford, J.E., Yang, H., Allen, D.K., and Libourel, I.G. (2014). Metabolic flux analysis using 13C peptide label measurements. Plant J. 77, 476–486.

Mastrangelo, A., Ferrarini, A., Rey-Stolle, F., García, A., and Barbas, C. (2015). From sample treatment to biomarker discovery: a tutorial for untargeted metabolomics based on GC-(EI)-Q-MS. Analytica Chim. Acta 900, 21–35.

McDermott, R., Trabesinger, A.H., Muck, M., Hahn, E.L., Pines, A., and Clarke, J. (2002). Liquid-state NMR and scalar couplings in microtesla magnetic fields. Science 295, 2247–2249.

McFeters, G.A., Stuart, S.A., and Olson, S.B. (1978). Growth of heterotrophic bacteria and algal extracellular products in oligotrophic waters. Appl. Environ. Microbiol. 35, 383–391.

Mckay, R.T. (2011). How the 1D-NOESY suppresses solvent signal in metabonomics NMR spectroscopy: an examination of the pulse sequence components and evolution. Concepts Magn. Reson. A 38A, 197–220.

McLaren, M.R., Willis, A.D., and Callahan, B.J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. eLife 8, e46923.

McMurdie, P.J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput. Biol. 10, e1003531.

Mehta, P., Bukov, M., Wang, C.-H., Day, A.G.R., Richardson, C., Fisher, C.K., and Schwab, D.J. (2019). A high-bias, low-variance introduction to Machine Learning for physicists. Phys. Rep. 810, 1–124.

Mickalide, H., and Kuehn, S. (2019). Higher-order interaction between species inhibits bacterial invasion of a phototroph-predator microbial community. Cell Syst. 9, 521–533.e10.

Miller, R.O., and Kissel, D.E. (2010). Comparison of soil pH methods on soils of North America. Soil Sci. Soc. America J. 74, 310–316.

Miranda, K.M., Espey, M.G., and Wink, D.A. (2001). A rapid, simple spectrophotometric method for simultaneous detection of nitrate and nitrite. Nitric Oxide 5, 62–71.

Molstad, L., Dorsch, P., and Bakken, L.R. (2007). Robotized incubation system for monitoring gases (O2, NO, N2O N2) in denitrifying cultures. J. Microbiol. Methods 71, 202–211.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. 108, E1293–E1301.

Nakanishi, Y., Fukuda, S., Chikayama, E., Kimura, Y., Ohno, H., and Kikuchi, J. (2011). Dynamic omics approach identifies nutrition-mediated microbial interactions. J. Proteome Res. 10, 824–836.

Nelson, M.B., Martiny, A.C., and Martiny, J.B.H. (2016). Global biogeography of microbial nitrogen-cycling traits in soil. Proc. Natl. Acad. Sci. 113, 8033–8040.

Nissen, J.N., Johansen, J., Allesøe, R.L., Sønderby, C.K., Armenteros, J.J.A., Grønbech, C.H., Jensen, L.J., Nielsen, H.B., Petersen, T.N., Winther, O., and Rasmussen, S. (2021). Improved metagenome binning and assembly using deep variational autoencoders. Nat. Biotechnol. 39, 555–560.

Nitta, N., Iino, T., Isozaki, A., Yamagishi, M., Kitahama, Y., Sakuma, S., Suzuki, Y., Tezuka, H., Oikawa, M., Arai, F., et al. (2020). Raman image-activated cell sorting. Nat. Commun. 11, 3452.

Obenhuber, C.K., and Folsome, C.E. (1984). Eucaryote/procaryote ratio as an indicator of stability for closed ecological systems. BioSystems 16, 291–296.

Obenhuber, C.K., and Folsome, C.E. (1988). Carbon recycling in materially closed ecological life support systems. BioSystems 21, 165–173.

Orth, J.D., Thiele, I., and Palsson, B. (2010). What is flux balance analysis? Nat. Biotechnol. 28, 245–248.

Pagaling, E., Strathdee, F., Spears, B.M., Cates, M.E., Allen, R.J., and Free, A. (2014). Community history affects the predictability of microbial ecosystem development. ISME J. 8, 19–30.

Pagaling, E., Vassileva, K., Mills, C.G., Bush, T., Blythe, R.A., Schwarz-Linek, J., Strathdee, F., Allen, R.J., and Free, A. (2017). Assembly of microbial communities in replicate nutrient-cycling model ecosystems follows divergent trajectories, leading to alternate stable states. Environ. Microbiol. 19, 3374–3386.

Park, J., Kerner, A., Burns, M.A., and Lin, X.N. (2011). Microdroplet-enabled highly parallel Co-cultivation of microbial communities. PLoS One 6, e17019.

Park, S.-E., Seo, S.-H., Kim, E.-J., Byun, S., Na, C.-S., and Son, H.-S. (2019). Changes of microbial community and metabolite in kimchi inoculated with different microbial community starters. Food Chem. 274, 558–565.

Parulekar, N.N., Kolekar, P., Jenkins, A., Kleiven, S., Utkilen, H., Johansen, A., Sawant, S., Kulkarni-Kale, U., Kale, M., and Sæbø, M. (2017). Characterization of bacterial community associated with phytoplankton bloom in a eutrophic lake in South Norway using 16S rRNA gene amplicon sequence analysis. PLoS One 12, e0173408.

Paul, A., Carl, P., Westad, F., Voss, J.-P., and Maiwald, M. (2016). Towards process spectroscopy in complex fermentation samples and mixtures. Chem. Ingenieur Technik 88, 756–763.

Petroff, A.P., Tejera, F., and Libchaber, A. (2017). Subsurface microbial ecosystems: a photon flux and a metabolic cascade. J. Stat. Phys. 167, 763–776.

Pieper, C., Risse, D., Schmidt, B., Braun, B., Szewzyk, U., and Rotard, W. (2010). Investigation of the microbial degradation of phenazone-type drugs and their metabolites by natural biofilms derived from river water using liquid chromatography/tandem mass spectrometry (lc-ms/ms). Water Res. 44, 4559–4569.

Pontrelli, S., Szabo, R., Pollak, S., Schwartzman, J., Ledezma-Tejeida, D., Cordero, O.X., and Sauer, U. (2021). Hierarchical control of microbial community assembly. bioarxiv. https://doi.org/10.1101/2021.06.22.449372.

Quinn, R.A., Whiteson, K., Lim, Y.-W., Salamon, P., Bailey, B., Mienardi, S., Sanchez, S.E., Blake, D., Conrad, D., and Rohwer, F. (2015). A Winogradsky-based culture system shows an association between microbial fermentation and cystic fibrosis exacerbation. ISME J. 9, 1024–1038.

D. Raftery, ed. (2014). Mass Spectrometry in Metabolomics: Methods and Protocols (Humana Press), Number 1198 in Methods in molecular biology.

Raman, A.S., Gehrig, J.L., Venkatesh, S., Chang, H.-W., Hibberd, M.C., Subramanian, S., Kang, G., Bessong, P.O., Lima, A.A., Kosek, M.N., et al. (2019). A sparse covarying unit that describes healthy and impaired human gut microbiota development. Science 365, eaau4735.

Ramanan, R., Kang, Z., Kim, B.-H., Cho, D.-H., Jin, L., Oh, H.-M., and Kim, H.-S. (2015). Phycosphere

bacterial diversity in green algae reveals an apparent similarity across habitats. Algal Res. 8, 140–144.

Rao, Y., Tao, Y., Chen, X., She, X., Qian, Y., Li, Y., Du, Y., Xiang, W., Li, H., and Liu, L. (2020). The characteristics and correlation of the microbial communities and flavors in traditionally pickled radishes. LWT 118, 108804.

Ratzke, C., Denk, J., and Gore, J. (2018). Ecological suicide in microbes. Nat. Ecol. Evol. 2, 867–872.

Raup, D.M., and Michelson, A. (1965). Theoretical morphology of the Coiled shell. Sci. New Ser. 147, 1294–1295.

Riemann, L., Steward, G.F., and Azam, F. (2000). Dynamics of bacterial community composition and activity during a mesocosm diatom bloom. Appl. Environ. Microbiol. 66, 578–587.

Rillig, M.C., and Antonovics, J. (2019). Microbial biospherics: the experimental study of ecosystem function and evolution. Proc. Natl. Acad. Sci. 116, 11093–11098.

Rillig, M.C., Ryo, M., Lehmann, A., Aguilar-Trigueros, C.A., Buchert, S., Wulf, A., Iwasaki, A., Roy, J., and Yang, G. (2019). The role of multiple global change factors in driving soil functions and microbial biodiversity. Science 366, 886–890.

Robertson, J.W.F., Rodrigues, C.G., Stanford, V.M., Rubinson, K.A., Krasilnikov, O.V., and Kasianowicz, J.J. (2007). Single-molecule mass spectrometry in solution using a solitary nanopore. Proc. Natl. Acad. Sci. 104, 8207–8211.

Rocca, J.D., Hall, E.K., Lennon, J.T., Evans, S.E., Waldrop, M.P., Cotner, J.B., Nemergut, D.R., Graham, E.B., and Wallenstein, M.D. (2015). Relationships between protein-encoding gene abundance and corresponding process are commonly assumed yet rarely observed. ISME J. 9, 1693–1699.

Rosch, P., Harz, M., Schmitt, M., Peschke, K.-D., Ronneberger, O., Burkhardt, H., Motzkus, H.-W., Lankers, M., Hofer, S., Thiele, H., and Popp, J. (2005). Chemotaxonomic identification of single bacteria by micro-Raman spectroscopy: application to clean-room-relevant biological contaminations. Appl. Environ. Microbiol. 71, 1626–1637.

Rosen, M.J., Davison, M., Bhaya, D., and Fisher, D.S. (2015). Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science 348, 1019–1023.

Russ, W.P., Lowery, D.M., Mishra, P., Yaffe, M.B., and Ranganathan, R. (2005). Natural-like function in artificial WW domains. Nature 437, 579–583.

Russ, W.P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., and Ranganathan, R. (2020). An evolution-based model for designing chorismate mutase enzymes. Science 369, 440–445.

Ruhl, M., Hardt, W.-D., and Sauer, U. (2011). Subpopulation-specific metabolic pathway usage in mixed cultures as revealed by reporter protein-based 13C analysis. Appl. Environ. Microbiol. 77, 1816–1821.

Saleem, M., Hu, J., and Jousset, A. (2019). More than the sum of its parts: microbiome biodiversity as a driver of plant growth and soil health. Annu. Rev. Ecol. Evol. Syst. 50, 145–168.

Sanchez-Gorostiaga, A., Bajic, D., Osborne, M.L., Poyatos, J.F., and Sanchez, A. (2019). High-order interactions distort the functional landscape of microbial consortia. PLoS Biol. 17, e3000550.

Schneidman, E., Berry, M.J., Segev, R., and Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. Nature 440, 1007–1012.

Seitz, A.P., Nielsen, T.H., and Overmann, J. (1993). Physiology of purple sulfur bacteria forming macroscopic aggregates in Great Sippewissett Salt Marsh, Massachusetts. FEMS Microbiol. Ecol. 12, 225–235.

Seng, P., Drancourt, M., Gouriet, F., La Scola, B., Fournier, P.-E., Rolain, J.M., and Raoult, D. (2009). Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clin. Infect. Dis. 49, 543–551.

Shi, Y., Pan, C., Wang, K., Chen, X., Wu, X., Chen, C.-T.A., and Wu, B. (2017). Synthetic multispecies microbial communities reveals shifts in secondary metabolism and facilitates cryptic natural product discovery. Environ. Microbiol. 19, 3606–3618.

Shoval, O., Sheftel, H., Shinar, G., Hart, Y., Ramote, O., Mayo, A., Dekel, E., Kavanagh, K., and Alon, U. (2012). Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space. Science 336, 1157–1160.

Silverman, J.D., Washburne, A.D., Mukherjee, S., and David, L.A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. eLife 6, e21887.

Slichter, C.P. (1990). Principles of Magnetic Resonance. In Springer Series in Solid-State Sciences, Third edition (Springer-Verlag).

Soininen, P., Kangas, A.J., Würtz, P., Tukiainen, T., Tynkkynen, T., Laatikainen, R., Järvelin, M.-R., Kähönen, M., Lehtimäki, T., Viikari, J., et al. (2009). High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. Analyst 134, 1781–1785.

Staley, J.T., and Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu. Rev. Microbiol. 39, 321–346.

Stein, L.Y., and Klotz, M.G. (2016). The nitrogen cycle. Curr. Biol. 26, R94–R98.

Steunou, A.-S., Bhaya, D., Bateson, M.M., Melendrez, M.C., Ward, D.M., Brecht, E., Peters, J.W., Kühl, M., and Grossman, A.R. (2006). In situ analysis of nitrogen fixation and metabolic switching in unicellular thermophilic cyanobacteria inhabiting hot spring microbial mats. Proc. Natl. Acad. Sci. 103, 2398–2403.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al. (2015). Structure and function of the global ocean microbiome. Science 348, 1261359.

Suzuki, Y., Kobayashi, K., Wakisaka, Y., Deng, D., Tanaka, S., Huang, C.-J., Lei, C., Sun, C.-W., Liu, H., Fujiwaki, Y., et al. (2019). Label-free chemical imaging flow cytometry by high-speed multicolor stimulated Raman scattering. Proc. Natl. Acad. Sci. 116, 15842–15848.

Taub, F.B. (1974). Closed ecological systems. Annu. Rev. Ecol. Syst. 5, 139–160.

Taub, F.B. (2009). Community metabolism of aquatic closed ecological systems: effects of nitrogen sources. Adv. Space Res. 44, 949–957.

Teeling, H., Fuchs, B.M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C.M., Kassabgy, M., Huang, S., Mann, A.J., Waldmann, J., et al. (2012). Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. Science 336, 608–611.

Tejera, F., Libchaber, A., and Petroff, A.P. (2018). Oxygen dynamics in a two-dimensional microbial ecosystem. Phys. Rev. E 98, 042409.

Tenorio, R., Fedders, A.C., Strathmann, T.J., and Guest, J.S. (2017). Impact of growth phases on photochemically produced reactive species in the extracellular matrix of algal cultivation systems. Environ. Sci. Water Res. Technol. 3, 1095–1108.

Terekhov, S.S., Smirnov, I.V., Malakhova, M.V., Samoilov, A.E., Manolov, A.I., Nazarov, A.S., Danilov, D.V., Dubiley, S.A., Osterman, I.A., Rubtsova, M.P., et al. (2018). Ultrahigh-throughput functional profiling of microbiota communities. Proc. Natl. Acad. Sci. 115, 9551–9556.

Thelusmond, J.-R., Strathmann, T.J., and Cupples, A.M. (2016). The identification of carbamazepine biodegrading phylotypes and phylotypes sensitive to carbamazepine exposure in two soil microbial communities. Sci. Total Environ. 571, 1241–1252.

Tian, H., Xu, R., Canadell, J.G., Thompson, R.L., Winiwarter, W., Suntharalingam, P., Davidson, E.A., Ciais, P., Jackson, R.B., Janssens-Maenhout, G., et al. (2020). A comprehensive quantification of global nitrous oxide sources and sinks. Nature 586, 248–256.

Tiedje, J.M., Sexstone, A.J., Myrold, D.D., and Robinson, J.A. (1983). Denitrification: ecological niches, competition and survival. Antonie van Leeuwenhoek 48, 569–583.

Toerien, D.F., and Hattingh, W.H.J. (1969). Anaerobic digestion I. The microbiology of anaerobic digestion. Water Res. 3, 385–416.

Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., and Gordon, J.I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444, 1027–1031.

Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R., and Gordon, J.I. (2007). The human microbiome project. Nature 449, 804–810.

Vanwonterghem, I., Jensen, P.D., Dennis, P.G., Hugenholtz, P., Rabaey, K., and Tyson, G.W. (2014). Deterministic processes guide long-term synchronised population dynamics in replicate anaerobic digesters. ISME J. 8, 2015–2028.

Vetsigian, K., Jajoo, R., and Kishony, R. (2011). Structure and evolution of streptomyces interaction networks in soil and in silico. PLoS Biol. 9, e1001184.

Ward, D.M., Ferris, M.J., Nold, S.C., and Bateson, M.M. (1998). A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. Microbiol. Mol. Biol. Rev. 62, 1353–1370.

Weiss, K., Khoshgoftaar, T.M., and Wang, D. (2016). A survey of transfer learning. J. Big Data 3.

Werner, J.J., Knights, D., Garcia, M.L., Scalfone, N.B., Smith, S., Yarasheski, K., Cummings, T.A., Beers, A.R., Knight, R., and Angenent, L.T. (2011). Bacterial community structures are unique and resilient in full-scale bioenergy systems. Proc. Natl. Acad. Sci. U S A. 108, 4158–4163.

Wharfe, E.S., Jarvis, R.M., Winder, C.L., Whiteley, A.S., and Goodacre, R. (2010). Fourier transform infrared spectroscopy as a metabolite fingerprinting tool for monitoring the phenotypic changes in complex bacterial communities capable of degrading phenol. Environ. Microbiol. 12, 3253–3263.

Wilbanks, E.G., Jaekel, U., Salman, V., Humphrey, P.T., Eisen, J.A., Facciotti, M.T., Buckley, D.H., Zinder, S.H., Druschel, G.K., Fike, D.A., and Orphan, V.J. (2014). Microscale sulfur cycling in the phototrophic pink berry consortia of the Sippewissett Salt Marsh. Environ. Microbiol. 16, 3398–3415.

Wintermute, E.H., and Silver, P.A. (2010). Emergent cooperation in microbial metabolism. Mol. Syst. Biol. 6, 407.

Woese, C.R. (2004). A new biology for a new century. Microbiol. Mol. Biol. Rev. 68, 173–186.

Wolfe, B.E., Button, J.E., Santarelli, M., and Dutton, R.J. (2014). Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity. Cell 158, 422–433.

Xu, X., Wu, B., Zhao, W., Pang, X., Lao, F., Liao, X., and Wu, J. (2020). Correlation between autochthonous microbial communities and key odorants during the fermentation of red pepper (capsicum annuum l.). Food Microbiol. 91, 103510.

Yu, X., Polz, M.F., and Alm, E.J. (2019). Interactions in self-assembled microbial communities saturate with diversity. ISME J 13, 1602–1617.

Yuan, J., Bennett, B.D., and Rabinowitz, J.D. (2008). Kinetic flux profiling for quantitation of cellular metabolic fluxes. Nat. Protoc. 3, 1328–1340.

Zakem, E.J., Polz, M.F., and Follows, M.J. (2020). Redox-informed models of global biogeochemical cycles. Nat. Commun. 11, 5680.

Zamboni, N., Fendt, S.-M., Rühl, M., and Sauer, U. (2009). 13C-based metabolic flux analysis. Nat. Protoc. 4, 878–892.

Zavarzin, G.A. (2006). Winogradsky and modern microbiology. Microbiology 75, 501–511.

Zengler, K., Toledo, G., Rappé, M., Elkins, J., Mathur, E.J., Short, J.M., and Keller, M. (2002). Cultivating the uncultured. Proc. Natl. Acad. Sci. 99, 15681–15686.

Zengler, K., Walcher, M., Clark, G., Haller, I., Toledo, G., Holland, T., Mathur, E.J., Woodnutt, G., Short, J.M., and Keller, M. (2005). High-throughput cultivation of microorganisms using microcapsules. In Methods in Enzymology, volume 397 of Environmental Microbiology (Academic Press), pp. 124–130.

Zhang, Y., Thompson, K.N., Branck, T., Yan, Y., Nguyen, L.H., Franzosa, E.A., and Huttenhower, C. (2021). Metatranscriptomics for the human microbiome and microbial community functional profiling. Annu. Rev. Biomed. Data Sci. 4, 279–311.

Zumft, W.G. (1997). Cell biology and molecular basis of denitrification. Microbiol. Mol. Biol. Rev. 61, 533–616.