Emergent ecosystem functions follow simple quantitative rules

Juan Diaz-Colunga[1,2] *, Abigail Skwara[1,2] *, Jean C. C. Vila[1,2], Djordje Bajic[1,2] ✉, and Álvaro Sánchez[1,2,3] ✉

[1]Department of Ecology & Evolutionary Biology, Yale University, New Haven, CT 06511, USA
[2]Microbial Sciences Institute, Yale University, West Haven, CT 06516, USA
[3]Current address: Department of Microbial Biotechnology, National Center for Biotechnology (CNB-CSIC), 28049 Madrid, Spain
✉D.B.: djordje.bajic@yale.edu, A.Sa.: alvaro.sanchez@cnb.csic.es
*These authors contributed equally to this work

Abstract | The functions and services provided by ecosystems emerge from myriad interactions between organisms and their environment. The difficulty of incorporating this complexity into quantitative models has hindered our ability to predictively link species-level composition with ecosystem function. This represents a major obstacle towards engineering ecological systems for environmental and biotechnological purposes. Inspired by similar findings in evolutionary genetics, here we show that the function of ecological communities often follows simple equations that allow us to accurately predict and optimize ecological function. This predictability is facilitated by emergent "species-by-ecosystem" interactions that mirror the patterns of global epistasis observed in many genetic systems. Our results illuminate an unexplored path to quantitatively linking the composition and function of ecological communities, bringing the tasks of predicting biological function at the genetic, organismal, and ecological scales under the same quantitative formalism.

The Earth's ecosystems carry out countless functions of technological importance, from food production in farms and crop fields to biofuel production in sugarcane biorefineries (*1*,*2*). Learning how we may engineer and optimize ecological functions is a major aspiration of modern science, with the potential to resolve a wide range of currently open technological challenges across research fields and sectors of the economy. Addressing this challenge requires us to find a general answer to a simple question: Given a list of candidate species, which ones should one choose to form a community that maximizes a target function? This question has been posed in a wide range of contexts, from which crop mixtures should be used to maximize yield or improve soil health (*1*,*3*) to which phage cocktails are most effective at clearing bacterial infections (*4*,*5*), but a general strategy to solve it is still lacking. Purely empirical approaches are generally unfeasible given the astronomic dimensionality of the problem: with as few as 25 candidate species, one could form over 30 million possible combinations, and testing them all is unpractical. Theoretical approaches have not yet delivered a general solution either. Ecological function emerges from complex webs of molecular, physiological, and organismal interactions. Incorporating all of this complexity into predictive models has only been achieved in a small number of case studies, and those required extensive parametrization (*6*–*9*).

The challenging nature of predicting biological function is not exclusive to ecology. At the organismal, genetic, and molecular scales, biological function is also highly complex, emerging from physiological, biophysical and biochemical interactions between components. For instance, the growth rate of a cell emerges from interactions between its metabolic pathways, while the catalytic activity of an enzyme arises from biophysical and biochemical interactions between its amino acids. Given this complexity, predicting the effect of a mutation on the fitness of an organism or on the stability of an enzyme might also appear to be a formidable task, not that different in scope from predicting the change in ecosystem function after adding a new species to a community. Encouragingly, quantitative genetics research has consistently found that the phenotypic and fitness effects of a mutation are often well-estimated by simple linear equations, which can be empirically determined for each mutation from a small number of measurements and do not require extensive parameterization nor fine-grain modeling. For instance, the effect of a particular mutation on the relative

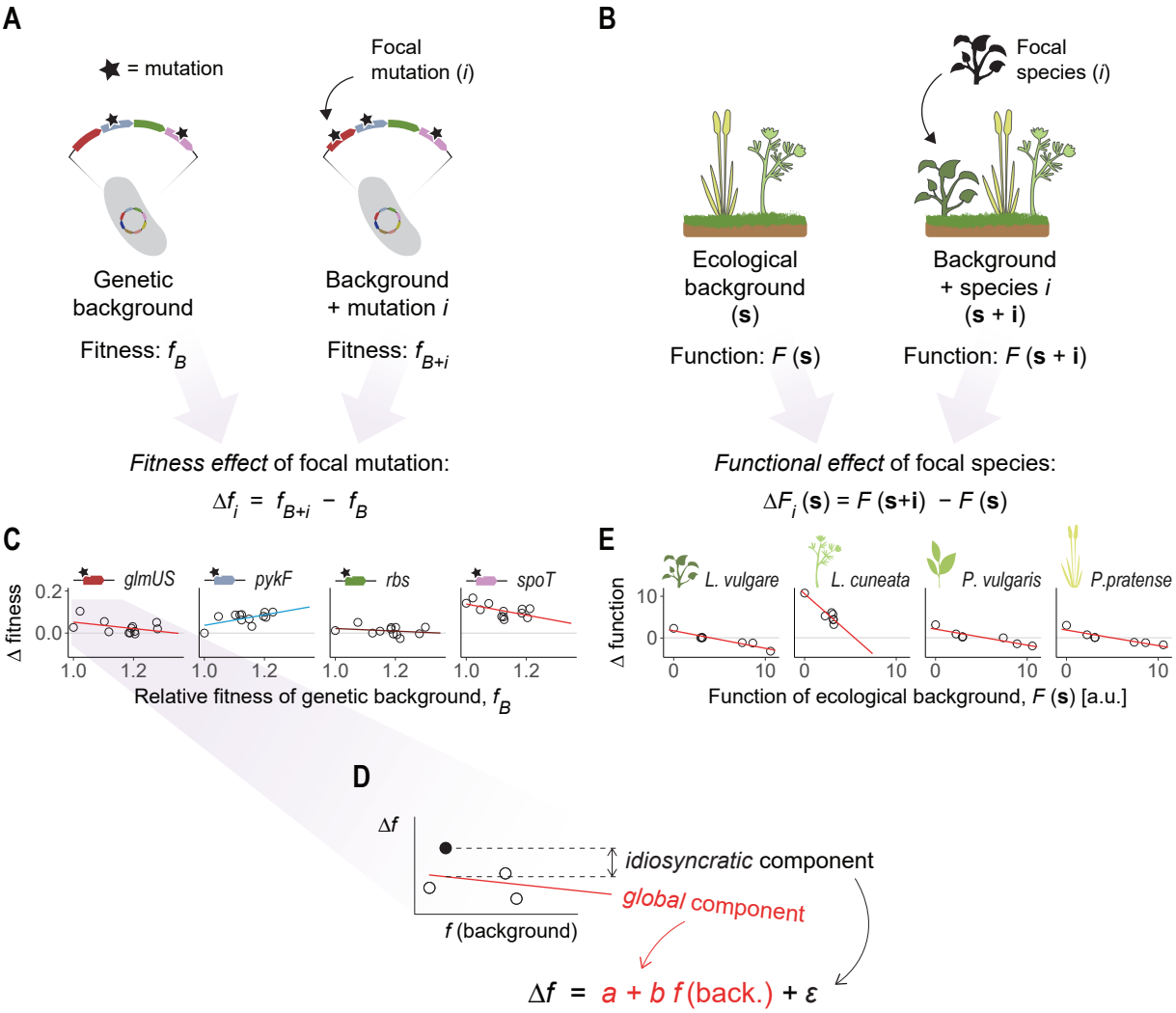

Fig. 1. An ecological parallel to global epistasis. (**A**) Recent work in quantitative genetics has found that the fitness effect of a mutation is often dependent on the fitness of the genetic background where it arises — a phenomenon that has been termed *global epistasis*. (**B**) We hypothesize that an ecological parallel to global epistasis might exist, where the addition of a species to a community might induce a change in ecosystem function that depends on the function of the community to which it is added. (**C**) The fitness effect of a mutation exhibits a global scaling with the background fitness that is often well estimated by a linear fit. The slope and intercept of the fit vary across different mutations. Data from Khan et al. (*11*) (**D**) The fitness effect of a mutation can be broken down into two contributions: first, a *global* contribution that scales with the background fitness and is approximated by a linear equation, and, second, an *idiosyncratic* contribution that is not predictable from the background fitness and is represented by the residuals of the fits. (**E**) Analogously, the functional effect of a species on an ecological background scales linearly with the background function (here, the above-ground biomass of a plant consortium). Data from Kuebbing et al. (*33*), non-native species.

fitness of a microorganism is often predictably linked to the fitness of its genetic background (Fig. 1A). The existence of these quantitative patterns is a manifestation of *global epistasis* (*10*–*19*), a phenomenon which includes (but is not limited to) the common observation that beneficial mutations have smaller positive effects in fitter backgrounds (*diminishing returns epistasis*). The existence of global epistasis makes it possible to build predictive effective models of biological function that do not require the integration of fine-grain mechanisms (even though those are of course valuable for interpretation purposes). Recent studies have exploited global epistasis to develop highly promising methodologies that are successful at inferring the full map between genotype and phenotype in large combinatorial spaces from just a subset of measured genotype-phenotype pairs (*20*–*24*).

Inspired by this idea, we hypothesized that an ecological analogue to global epistasis might exist, where the functional effect of adding a species to an ecosystem (an effective *ecological background*) could be well estimated from simple, linear relationships linking it to the function of the communities to which it is added (Fig. 1B). If this hypothesis were correct, we could then predict how adding a species to a community should change its function. This would pave the way to predictively connecting species-level composition to quantitative function. To test this hypothesis, we set out to examine previously published data of plant, bacterial, and algal ecosystems, under distinct environmental conditions and for a variety of collective functions. We found that a parallel concept to global epistasis can indeed be formulated for ecological systems. By conducting new experiments, we show that, as we had hoped, this allows us to build accurate quantitative models that predict and optimize ecological function. Our findings argue that the same quantitative formalism can be applied to predict biological function across widely different scales and levels of biological organization, from molecules and organisms to ecological communities.

Species-by-ecosystem effects across different ecological contexts

Genetic interactions capture how the fitness effect of a mutation changes in different genetic contexts. Historically, the study of genetic interactions (*epistasis*) has broken them down as the sum of pairwise interactions (G×G), third-order interactions (G×G×G), fourth-order, and so on (*25*). This has paralleled the similar partitioning of ecological interactions as the sum of pairwise species-by-species (S×S) and higher-order (e.g., S×S×S) effects (*26*–*32*). Recent work in genetics has proposed that epistasis can be instead partitioned into a global epistasis component, described by a linear regression between the fitness effect of a mutation and the fitness of the background, and an idiosyncratic component described by the residuals of this fit (Fig. 1C-D). Based on the success of recently found parallelisms between genetic and functional ecological interactions (*28*,*29*,*31*), we reasoned that the latter can be partitioned in the same manner, as the sum of (i) a global, species-by-community (S×C) interaction described by how the functional effect of a species scales with the function of the community to which it is added, and (ii) an idiosyncratic interaction captured by the residuals.

To assess the possible merits of this hypothesis, we first re-examined published data from a recent experiment that combinatorially assembled (almost) all possible combinations of four different plants (*33*). Each species assemblage can be described by a unique combination of species presence/absence ($\mathbf{s}$). The function of each assemblage ($F(\mathbf{s})$), which in this case was the above-ground biomass, was measured at harvest time. From such data, one can determine the functional effect of adding each species (i) to various background communities formed by different plant combinations (Fig. 1B) as, i.e., $\Delta F_i(\mathbf{s}) = F(\mathbf{s}+\mathbf{i}) - F(\mathbf{s})$, where we have called $\mathbf{s}+\mathbf{i}$ the assemblage resulting from the addition of species i to the background $\mathbf{s}$ (Fig. 1B). In Fig. 1E we plot the functional effects of each species — $\Delta F_i(\mathbf{s})$ for species i — against the function of its ecological backgrounds, $F(\mathbf{s})$. As a comparison, in Fig. 1C we show data from ref. (*11*), which measured the fitness effects of various different beneficial mutations in *E. coli* placed in several combinatorial backgrounds made up by the other mutations (Fig. 1C). The functional effect of species additions exhibits a strong parallel with the patterns of global epistasis observed in genetic systems, scaling linearly with the function of the background community. As is the case for mutations, the particular linear equation that estimates the functional effects is unique for each species.

Global epistasis has been seen in a wide range of other genetic contexts, including yeast (*14*,*17*) and bacteria (*12*). To determine how general this parallel to global epistasis may be in ecological systems, we analyzed a collection of published data sets from our own laboratory and others. Each community in these data sets is made up by different organism types: terrestrial plants (*33*), phytoplankton (*34*), and both Gram-negative and Gram-positive bacteria (*8*,*29*,*35*). The ecological conditions of these communities vary widely, including the number of organismal generations, the type and frequency of resource addition, and the form of propagation. The functions themselves are very different too: from the production of biomass or the net metabolic activity to the secre-

Organisms type	Number of species	Ecosystem function	Source of data set
Terrestrial plants	Two sets of 4 each	Above-ground biomass	Kuebbing et al. (*33*)
Phytoplankton	5	Biomass production	Ghedini et al. (*34*)
Bacteria	6	Xylose oxidation rate	Langenheder et al. (*35*)
Bacteria	6	Starch hydrolysis rate	Sanchez-Gorostiaga et al. (*29*)
Bacteria	25	Butyrate secretion	Clark et al. (*8*)

Table 1. Data sets of combinatorial ecosystem function used in this study.

tion of specific metabolites or the degradation of environmental polymers. Table 1 summarizes the data sets we considered, all of which include multiple combinatorial assemblages of species from candidate pools of between 4 and 25 taxa.

As shown in Fig. 2, we found that the functional effect of a species was in general well described by simple linear relationships of the form $\Delta F_i(\mathbf{s}) = a_i + b_i F(\mathbf{s}) + \epsilon_i(\mathbf{s})$. We generically call this expression the *functional effect equation* (FEE) of species i. The intercepts (a_i) and slopes (b_i) of the fitting lines differ across taxa, suggesting that they are determined by the interplay between each individual species and the rest of the community — and thus can be interpreted as emergent species-by-ecosystem interactions as we expected. The terms $\epsilon_i(\mathbf{s})$ (i.e., the residuals of the fits) capture the idiosyncratic component of said interactions. Global S×C interactions were present and strong across species and data sets (average $R^2 = 0.42$, fig. S1).

Many species (∼50%) across all datasets in Table 1 display negatively sloped FEEs (red lines in Fig. 2). This trend is also commonly observed in population genetics: the fitness effect of a genomic mutation most often becomes either less beneficial or more deleterious as the fitness of the genetic background increases (*10*–*12*,*15*,*17*,*18*). These two situations are typically referred to as *diminishing returns* and *increasing costs*, respectively. Often, diminishing returns and increasing costs are exhibited by the same species, which can be beneficial or deleterious depending on the function of the background community in which where they are introduced: they can increase the community function when added to low performing ecological backgrounds, but decrease it when added to high performing ones. A second major fraction of all species (∼45%) have effects on ecosystem function that are dominated by idiosyncrasies in the species-by-community interactions, making it so the functional effect displays no global relationship with $F(\mathbf{s})$ and instead depends on the particular composition of each ecological background (black lines in Fig. 2). As we shall see in what follows, these flat patterns are also informative and useful for predictive purposes. Finally, a smaller number of species (∼5%) exhibit positively sloped FEEs (blue lines in Fig. 2), becoming more beneficial (or less deleterious) in backgrounds with higher functions. We refer to these patterns as *accelerating returns* (or *decreasing costs*).

Notably, in one of the data sets we examined (Sanchez-Gorostiaga et al. (*29*)) one bacterial species (*P. polymyxa*, Fig. 2D, rightmost panel) displays a functional effect on the amylolytic rate of the consortia that can be described by two distinct FEEs, i.e., its FEE appears split into two "branches". Closer examination of this case indicates that the two branches are determined by the presence or absence of a second species (*B. thuringiensis*) in the ecological background (fig. S2). This suggests that some specific species-by-species pairwise interactions may not be well captured by a global species-by-ecosystem trend, and instead can induce major shifts in the FEEs. Comparable patterns have been observed in population genetics, where strong idiosyncratic mutation-by-mutation interactions have been found that modify the global mutation-by-genotype fitness effects (*19*).

Together, our analyses suggest that global species-by-ecosystem interactions can be observed across a wide range of ecological contexts and functions. The specific molecular mechanisms through which species interact with one another and contribute to collective functions are often complex, context-dependent and difficult to characterize. However, the emergence of FEEs suggests that these complex microscopic details may be absorbed into an emergent species-by-community functional trend, which can in principle be fit from a small number of observed communities. This

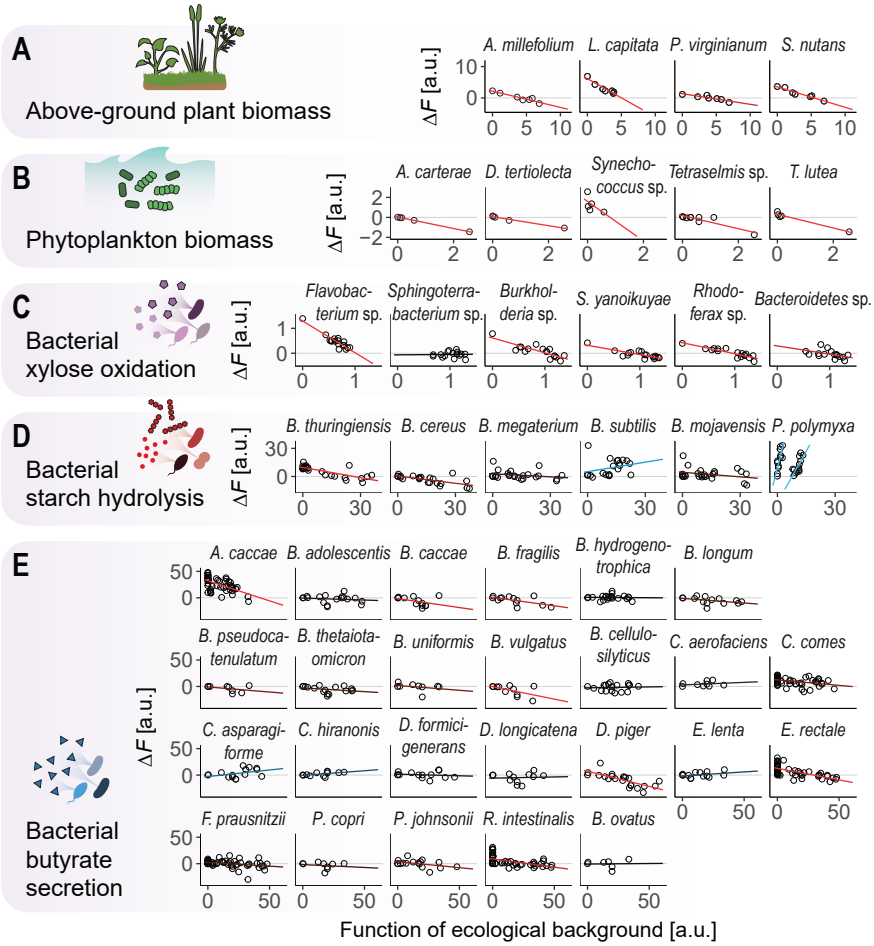

Fig. 2. Functional effects across species and ecosystems. The functional effect of a species often scales with the function of the community to which it is added. This phenomenon is observed across very different organism types, ecological conditions, and collective functions (Table 1). The scaling is frequently well described by a linear relationship (red lines: negative slopes, blue lines: positive slopes, black lines: flat slopes). (**A**) Data from Kuebbing et al. (*33*), native species. (**B**) Data from Ghedini et al. (*34*) (**C**) Data from Langenheder et al. (*35*) (**D**) Data from Sanchez-Gorostiaga et al. (*29*) (**E**) Data from Clark et al. (*8*)

indicates that the functional effect of a taxon on a given ecological background may be predictable with no prior information on the traits of that taxon or its interactions with all its ecological partners. Thus, we hypothesized that FEEs could be exploited to predict community function without the need for fine-grained mechanistic ecological models.

Global functional effects for the design of optimal consortia

Our starting hypothesis is simple: if we have a set of species and, for all of them, we know how adding them to a community would change its function, then we should be able to predict the function of any combinatorial assemblage from that set. Knowledge of the FEEs of a set of species should thus find a solution to the question we posed at the outset of this paper: Given a list of species, which ones should one choose to form a community that maximizes a given function? To test this hypothesis, we built a small library consisting of eight bacterial species that were isolated from soil samples (Materials and Methods). Five of these species were Pseudomonas strains that produce pyoverdines in monoculture, while the remaining three were non-producing Enterobacteriaceae (Fig. 3A, Materials and Methods). The cumulative production of pyoverdines is a good candidate for a community function: first, it can be quantified using simple readings of optical density (Materials and Methods) and, second, the production of pyoverdines responds to intra-species signaling (*36*) and

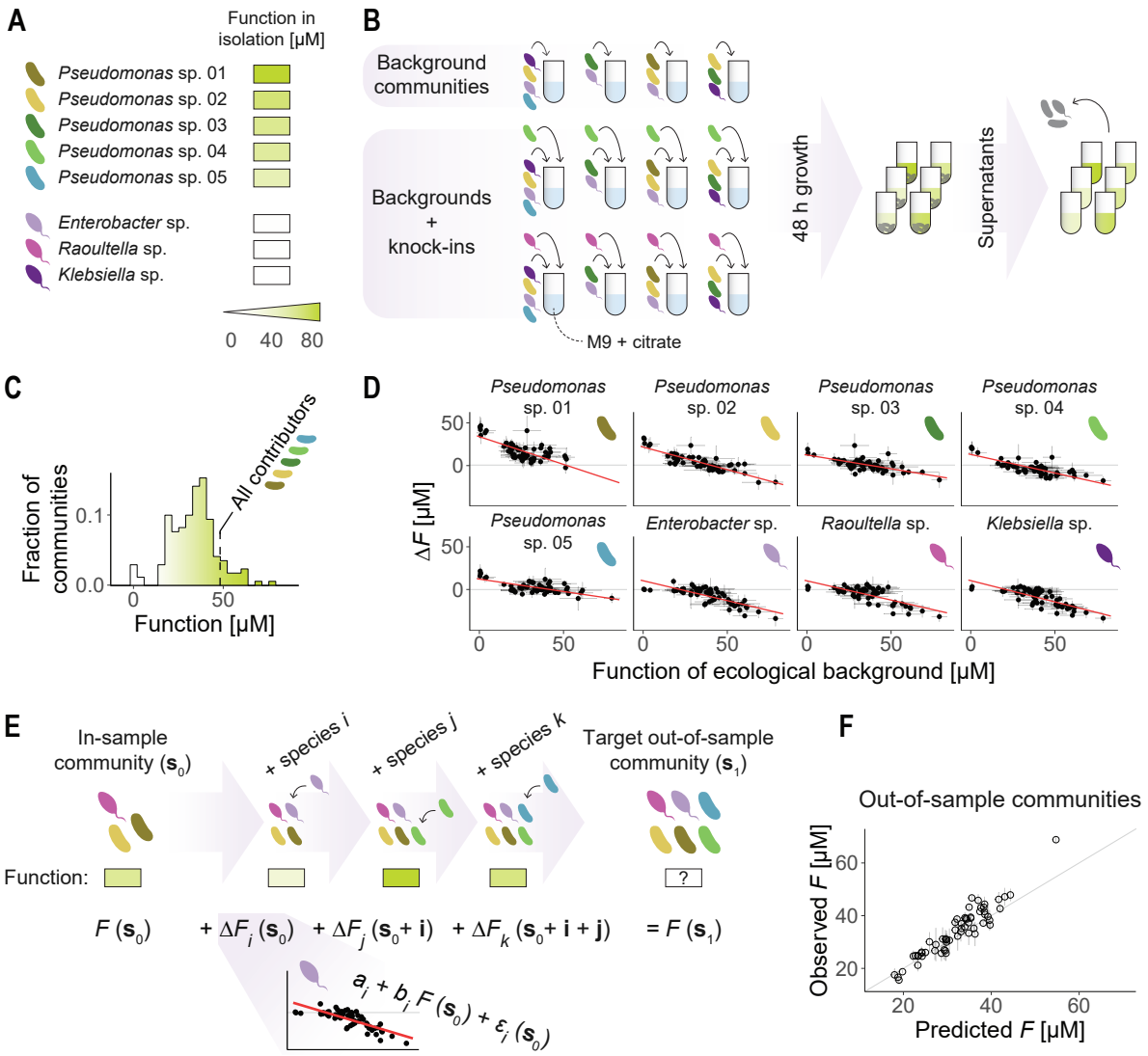

Fig. 3. Global functional effects can be exploited to predict community function. (**A**) We isolated and eight bacterial species from environmental samples and identified them at the genus level (Materials and Methods). Five of them exhibited secretion of pyoverdines when grown in monoculture in minimal M9 citrate medium (Materials and Methods). (**B**) We assembled 164 consortia by inoculating combinations of these eight species into minimal M9 citrate medium, and incubated them still for 48 h. We then collected the spent media and quantified the concentration of pyoverdines (Materials and Methods). (**C**) We found variable levels of pyoverdines secretion, with the concentrations in the supernatants ranging from 0 to roughly 70 μM. About 20% of the assemblages exhibited higher activity than the consortium formed by all five pyoverdines secretors. (**D**) Global species-by-community interactions emerged in our experiment, as evidenced by the correlations between the functional effects of the species and the functions of their ecological backgrounds ($R^2 \sim 0.5$ for all species). Dots and error bars represent means and standard deviations across three biological replicates. (**E**) We hypothesize that sequentially adding functional effects could serve to predict the function of an out-of-sample community ($\mathbf{s}_1$) from that of an in-sample community ($\mathbf{s}_0$) as described in the main text. (**F**) We evaluated the viability of our prediction method by assembling 61 new consortia (which served as the out-of-sample test set of communities) and comparing their predicted and measured levels of pyoverdines secretion. We found a good agreement ($R^2 = 0.8$) between the observations and the predictions. Dots and error bars represent means and standard deviations across two biological replicates.

is often controlled by population size via quorum sensing (*37*). Due to the potential for interactions in our system, it is not immediately obvious which of the 255 potential consortia one could assemble would produce the most pyoverdines under our conditions.

Using this function and species set as our case study, we combinatorially assembled a set of background consortia by inoculating unique combinations of those species in minimal media at fixed inoculum sizes (Materials and Methods). We then allowed each assemblage to grow for 48 h, and measured the concentration of pyoverdines in the spent media at harvest time (Fig. 3B, Materials and Methods). In parallel experimental lines, we added each of the eight isolates to each of the background consortia — giving a total of 164 unique assemblages with variable levels of pyoverdines secretion (between 0 and 70 µM concentration in the spent media, Fig. 3C). We thus quantified the functional effects of each isolate in every background, and fit a linear regression for each species obtaining its functional effect equation. Consistent with what we found in the other data sets, clear linear FEE patterns were observed, indicating the presence of global species-by-community interactions (Fig. 3D).

A simple visual inspection of the FEEs can be useful from the perspective of ecosystem design. Species whose functional effects remain below or close to zero can be expected to have a deleterious (or at best insignificant) impact on function regardless of their ecological context, and thus it is reasonable to exclude them from a prospective optimal community. This straightforward observation can serve to narrow down the list of potentially desirable species. In our experiment, the functional effects of all three non-producers (*Enterobacter* sp., *Raoultella* sp. and *Klebsiella* sp.) were almost always negative or very small ($\Delta F \lesssim 0$) (Fig. 3D), as we had expected. The five pyoverdines producers, on the other hand, had positive functional effects ($\Delta F > 0$) in at least some ecological contexts. If there were no interactions, we should expect that the best community would include all five producers. However, we found that roughly 20% of the assemblages in our experiment had higher function than this naive assemblage of all contributing species (Fig. 3C). Out of the communities tested in our experiment, the highest functional output was achieved by a single species in monoculture (*Pseudomonas* sp. 01). While this is the case for this particular experiment, it is worth noting that the best consortium is not necessarily a monoculture. In other experimental data sets, the best performing community contained multiple taxa (fig. S3), even including some that had no activity in isolation — such as *P. polymyxa* in the Sanchez-Gorostiaga et al. data set (Fig. 2D), or *C. aerofaciens* in the Clark et. al data set (Fig. 2E). Together, these experiments and analyses indicate that the combination of species that optimizes a particular function is not trivial to know a priori or to predict relying on intuition alone. We reasoned that, once the FEEs are known, they could be leveraged to predict community functions based in composition, and thus to find optimal consortia.

To test this hypothesis, we developed a simple method based on concatenating species functional effects (Fig. 3E). Suppose that we have measured the function of a consortium (i.e., one of the 164 assemblages used to produce the ΔF-vs-F plots in Fig. 3D; henceforth an *in-sample* community), and we are interested in predicting the function of an assemblage that has not been tested (an *out-of-sample* community). We call the in-sample and out-of-sample communities $\mathbf{s}_0$ and $\mathbf{s}_1$, respectively, and their functions $F(\mathbf{s}_0)$ and $F(\mathbf{s}_1)$ respectively. In the example shown in Fig. 3E, $\mathbf{s}_1$ has three more species (i, j and k) than $\mathbf{s}_0$. Because we know the FEEs for each of those species, we hypothesized that sequentially adding their functional effects to the starting in-sample function $F(\mathbf{s}_0)$ could serve to predict the function of the out-of-sample community $F(\mathbf{s}_1)$. For instance, the first addition of species i to the in-sample community $\mathbf{s}_0$ would have an effect in function that we can estimate from the linear FEE for species i: $\Delta F_i(\mathbf{s}_0) = a_i + b_i F(\mathbf{s}_0) + \epsilon_i(\mathbf{s}_0)$. This procedure can be iterated for species j and k, ultimately giving a prediction for the function of the out-of-sample-community $F(\mathbf{s}_1)$. Predictions can be further refined by estimating the residuals of the FEEs using maximum likelihood, as discussed in the Supplementary Text.

To test the viability of this idea, we built a set of 61 new consortia that had not been assembled in our first experiment. These served as our out-of-sample test set of communities. We used the method described above to predict their functions, and then assembled them experimentally (under identical conditions to those in the first round of experiments) to quantify their empirical levels of pyoverdines secretion (Fig. 3B, Materials and Methods). As shown in Fig. 3F, we found a good agreement between the predictions and the observations ($R^2 = 0.8$). Notably, reducing the number

of in-sample communities used to fit the FEEs only moderately affected the ability of our method to predict out-of-sample functions. Even when FEEs were fit to a very small number of points (as few as ∼4), the signal was still strong ($R^2 \sim 0.5$) and the method was able to successfully identify optimal consortia (fig. S4). This suggests that our approach could be scalable to much larger combinatorial spaces: while the number of potential assemblages scales exponentially with the number of candidate species, our results indicate that only a few measurements per species could suffice to provide quantitative predictions of community function.

To test whether this simple method could be robust across ecological conditions, organism types, and ecosystem functions, we turned to the five data sets described in Table 1. For each of them, we applied the method described above (Fig. 3E) to predict the functions of a subset of randomly chosen out-of-sample communities. We repeated this process 500 times, each of them with a different set of out-of-sample assemblages, and quantified the R^2 between predictions and observations. We generally found our method to be reliable (average R^2 between 0.5 and 0.8 depending on the data set, Fig. 4), even when the number of data points used to fit the FEEs was further reduced (fig. S5). Interestingly, the Clark et al. data set (*8*) yielded the smallest R^2 between predictions and observations. This is not entirely surprising: besides having the smallest fraction of communities in the training set (as the total number of potential communities exceeds 33 million) this data set contains the most species with flat FEEs (Fig. 2E, black lines), that is, whose functional effects are dominated by an idiosyncratic rather than a global component (Fig. 1E). Note, however, that flat FEEs are informative. The magnitude of the deviations from the FEE (even if flat) are useful to discern between those species whose contribution to ecosystem function is relatively independent of their ecological background (i.e., those for which the residuals are small) and those whose contribution depends on their ecological context in a highly idiosyncratic manner (i.e., those with large residuals). While the former case might be well captured by our predictive method, the latter could suggest the presence of highly specific species-by-species interactions — not absorbed into a global species-by-community trend — for which fine-grained ecological models might be more appropriate.

Given the apparent ubiquity and usefulness of global species-by-ecosystem functional effects, we asked how generally they can be expected to emerge. Can any arbitrary mapping between community composition and function lead to ΔF-vs-F correlations? Intuitively, one might expect that a negative slope should be seen if the association between composition and function were random. In this scenario, the functions of any two communities differing in the presence of a single taxon would be completely uncorrelated, and they can be seen as independent "draws" from a generic distribution of functions. If the first draw gives a large value for the function, the second is likely to give a smaller one and vice-versa. Thus, the subtraction of the two random functions (namely $F_2 - F_1$) would be likely to be positive if F_1 was small and negative if F_1 was large, leading to a negative correlation between $F_2 - F_1$ and F_1.

To test this intuition, we randomized the pairing between communities and functions in our data 500 times. Consistent with our reasoning, we found that the functional effects and the background functions exhibited a negative correlation in the randomized data sets (fig. S6). Interestingly, though, the FEEs we fit to our empirical data were significantly different to those in the randomized control (fig. S6). Negative slopes around -1 are generically observed when the association between community composition and function is random, but significantly different slopes commonly emerge in many real ecological contexts (e.g., Fig. 2 and Fig. 3D). Despite the existence of negatively sloped ΔF-vs-F correlations, randomizing the association between composition and function should eliminate, or at the very least severely diminish, the ability of FEEs to predict community function out of sample. Application of our predictive method to the randomized data set yielded unsurprisingly poor results (fig. S6). Together, these realizations suggest that the observed FEEs in empirical data sets across ecosystems and functions are not a trivial consequence of having a bounded set of functional values. This randomization control provides a benchmark against which we can determine whether the empirical FEEs do indeed capture ecologically meaningful information on how species contribute to ecosystem function.

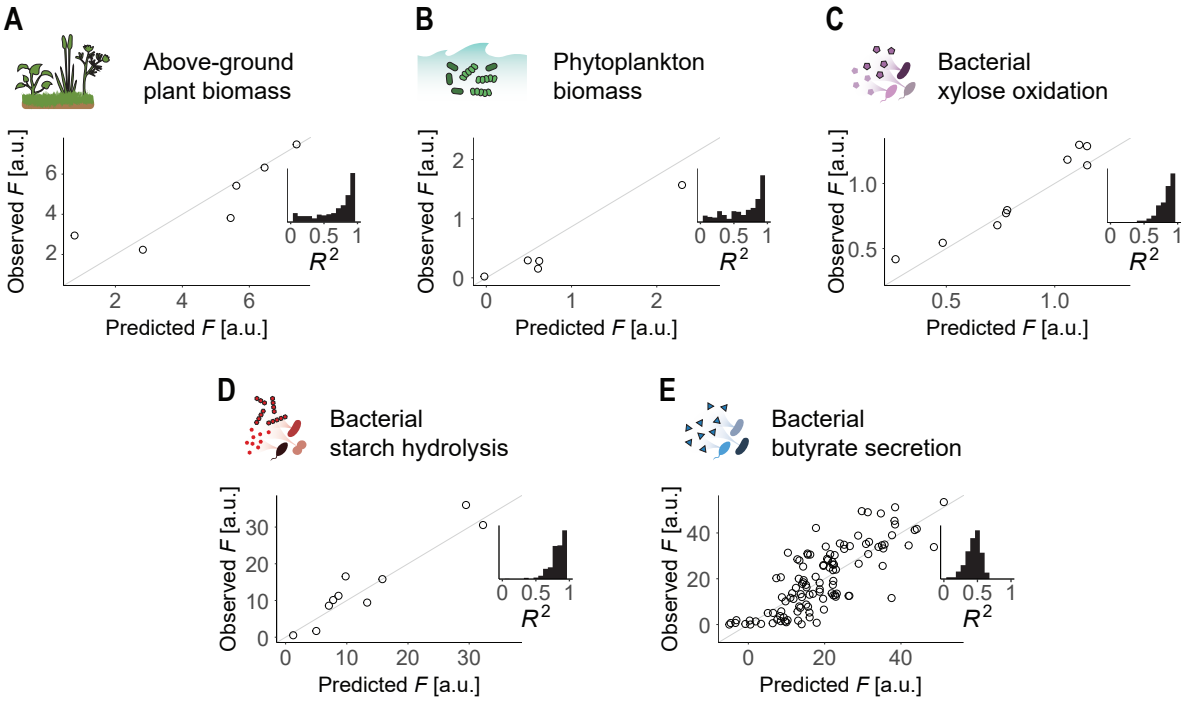

Fig. 4. Predicting community function across data sets. We evaluated the ability of the method described in the main text (and Fig. 3E) to predict community functions in all data sets in Table 1. For that, we left 20% of the communities in the data sets out of the sample, used the remaining 80% to fit FEEs, and applied our method to predict the function of the out-of-samples. We quantified the accuracy of the method as the R^2 between the predictions and the observations. We repeated the same process 500 times, each leaving a different subset of communities out of sample (randomly chosen). Main plots show an example of predicted against observed functions for one of the runs. Insets show histograms of the R^2 between predictions and observations across the 500 runs. (**A**) Data from Kuebbing et al. (*33*). (**B**) Data from Ghedini et al. (*34*) (**C**) Data from Langenheder et al. (*35*) (**D**) Data from Sanchez-Gorostiaga et al. (*29*) (**E**) Data from Clark et al. (*8*)

Discussion

Our experiments and analyses demonstrate that, despite their enormous microscopic complexity, emergent community functions are determined by simple quantitative rules. The core finding of this paper is that the change in community function caused by adding a new species to a community is often well estimated by simple linear equations. These *functional effect equations* represent an ecological parallel to the phenomenon known as *global epistasis* in quantitative genetics, where the fitness effect of a mutation scales with the fitness of the genetic background to which it is added. We propose that these linear trends may be interpreted as emergent species-by-ecosystem interactions, which approximate the functional effect of a species without having to specify every pairwise and higher-order interaction individually. The existence of these patterns reveals a tractable structure in the mapping between community compositions and functions, which we have shown can be exploited to identify optimal consortia from a very limited amount of empirical observations.

Building fine-grained predictive models that integrate the complex web of molecular and organismal interactions that take place in ecological communities has been and remains extremely challenging. Even in those studies that have reported success (*6–9*), parameterization required exhaustive empirical work, which is highly specific to the taxa, environmental conditions, and functions being studied. Machine learning strategies are more scalable (*38*,*39*), but extracting relevant, interpretable biological information from them is generally difficult. If we abandon fine-grained models and opt instead for coarse-graining the description of our communities, we find a more generalizable strategy to explain ecosystem function that consists of condensing community structure through a metric of

its biodiversity (*40*,*41*). When averaged across communities, biodiversity is indeed often related to ecosystem function, but the variation is generally high. By compressing the compositional state of a multi-species community (a high dimensional vector) to a scalar metric of biodiversity, we lose the level of granularity that is needed for rational ecosystem design.

Overcoming these limitations, our results point to a general, scalable, and interpretable solution to the problem of optimizing ecosystem function. Most importantly, they show that the problem of connecting structure to function in biology can be approached from the same modeling framework at all biological scales — from the molecular to the ecological. At the organismal level and below, recent studies have been successful at inferring the map between genotypes and phenotypes from partial observations and without the need for fine-grained, molecular-level description of biological function (*20*,*22*–*24*). These methods rely on the existence of regularities in genotype-phenotype maps, which are revealed by the emergence of global epistasis. Our work demonstrates that analogous regularities may exist in the mapping between ecosystem composition and function. This suggests that the increasingly large assortment of predictive and analytical tools from evolutionary genetics could be adapted and imported to ecology, exposing an unexplored path to predictively linking structure and function in ecosystems, and opening opportunities for cross-pollination across fields.

References

1. Wuest SE, Peter R and Niklaus PA (2021). Ecological and evolutionary approaches to improving crop variety mixtures. *Nature Ecology & Evolution* **5(8)**:1068–1077

2. Lino F, Bajic D, Vila JC, Sanchez A and Sommer M (2021). Complex yeast–bacteria interactions affect the yield of industrial ethanol fermentation. *Nature Communications* **12**:1498

3. Liu CLC, Kuchma O and Krutovsky KV (2018). Mixed-species versus monocultures in plantation forestry: Development, benefits, ecosystem services and perspectives for the future. *Global Ecology and Conservation* **15**:e00419

4. Chan BK, Abedon ST and Loc-Carrillo C (2013). Phage cocktails and the future of phage therapy. *Future Microbiology* **8(6)**:769–783

5. Wright RC, Friman VP, Smith MC and Brockhurst MA (2021). Functional diversity increases the efficacy of phage combinations. *Microbiology* **167(12)**:001110

6. Chen Y, Lin CJ, Jones G, Fu S and Zhan H (2009). Enhancing biodegradation of wastewater by microbial consortia with fractional factorial design. *Journal of Hazardous Materials* **171(1-3)**:948–953

7. Eng A and Borenstein E (2019). Microbial community design: methods, applications, and opportunities. *Current Opinion in Biotechnology* **58**:117–128

8. Clark RL, Connors BM, Stevenson DM, Hromada SE, Hamilton JJ, Amador-Noguez D and Venturelli OS (2021). Design of synthetic human gut microbiome assembly and butyrate production. *Nature Communications* **12(1)**:1–16

9. Gowda K, Ping D, Mani M and Kuehn S (2022). Genomic structure predicts metabolite dynamics in microbial communities. *Cell* **185(3)**:530–546.e25

10. MacLean R, Perron G and Gardner A (2010). Diminishing returns from beneficial mutations and pervasive epistasis shape the fitness landscape for rifampicin resistance in Pseudomonas aeruginosa. *Genetics* **186(4)**:1345–1354

11. Khan AI, Dinh DM, Schneider D, Lenski RE and Cooper TF (2011). Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* **332(6034)**:1193–1196

12. Chou HH, Chiu HC, Delaney NF, Segrè D and Marx CJ (2011). Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science* **332(6034)**:1190–1192

13. Perfeito L, Sousa A, Bataillon T and Gordo I (2014). Rates of fitness decline and rebound suggest pervasive epistasis. *Evolution* **68(1)**:150–162

14. Kryazhimskiy S, Rice DP, Jerison ER and Desai MM (2014). Global epistasis makes adaptation predictable despite sequence-level stochasticity. *Science* **344(6191)**:1519–1522

15. Schoustra S, Hwang S, Krug J and de Visser JAG (2016). Diminishing-returns epistasis among random beneficial mutations in a multicellular fungus. *Proceedings of the Royal Society B: Biological Sciences* **283(1837)**:20161376

16. Otwinowski J, McCandlish DM and Plotkin JB (2018). Inferring the shape of global epistasis. *Proceedings of the National Academy of Sciences* **115(32)**:E7550–E7558

17. Johnson MS, Martsul A, Kryazhimskiy S and Desai MM (2019). Higher-fitness yeast genotypes are less robust to deleterious mutations. *Science* **366(6464)**:490–493

18. Wei X and Zhang J (2019). Patterns and mechanisms of diminishing returns from beneficial mutations. *Molecular Biology and Evolution* **36(5)**:1008–1021

19. Bakerlee CW, Nguyen Ba AN, Shulgina Y, Rojas Echenique JI and Desai MM (2022). Idiosyncratic epistasis leads to global fitness–correlated trends. *Science* **376(6593)**:630–635

20. Romero PA, Krause A and Arnold FH (2013). Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences* **110(3)**:E193–E201

21. Tareen A, Kooshkbaghi M, Posfai A, Ireland WT, McCandlish DM and Kinney JB (2022). MAVE-NN: learning genotype-phenotype maps from multiplex assays of variant effect. *Genome Biology* **23**:98

22. Otwinowski J (2018). Biophysical Inference of Epistasis and the Effects of Mutations on Protein Stability and Function. *Molecular Biology and Evolution* **35(10)**:2345–2354

23. Sailer ZR, Shafik SH, Summers RL, Joule A, Patterson-Robert A, Martin RE and Harms MJ (2020). Inferring a complete genotype-phenotype map from a small number of measured phenotypes. *PLoS Computational Biology* **16(9)**:e1008243

24. Tonner PD, Pressman A and Ross D (2021). Interpretable modeling of genotype-phenotype landscapes with state-of-the-art predictive power. *bioRxiv*

25. Taylor MB and Ehrenreich IM (2015). Higher-order genetic interactions and their contribution to complex traits. *Trends in Genetics* **31(1)**:34–40

26. Guo X and Boedicker JQ (2016). The contribution of high-order metabolic interactions to the global activity of a four-species microbial community. *PLoS Computational Biology* **12(9)**:e1005079

27. Guo X and Boedicker J (2016). High-order interactions between species strongly influence the activity of microbial communities. *Biophysical Journal* **110(3)**:143a

28. Gould AL, Zhang V, Lamberti L, Jones EW, Obadia B, Korasidis N, Gavryushkin A, Carlson JM, Beerenwinkel N et al. (2018). Microbiome interactions shape host fitness. *Proceedings of the National Academy of Sciences* **115(51)**:E11951–E11960

29. Sanchez-Gorostiaga A, Bajić D, Osborne ML, Poyatos JF and Sanchez A (2019). High-order interactions distort the functional landscape of microbial consortia. *PLoS Biology* **17(12)**:e3000550

30. Mickalide H and Kuehn S (2019). Higher-order interaction between species inhibits bacterial invasion of a phototroph-predator microbial community. *Cell Systems* **9(6)**:521–533

31. Eble H, Joswig M, Lamberti L and Ludington WB (2021). High dimensional geometry of fitness landscapes identifies master regulators of evolution and the microbiome. *bioRxiv*

32. Korkmazhan E and Dunn AR (2022). High-order correlations in species interactions lead to complex diversity-stability relationships for ecosystems. *Physical Review E* **105**:014406

33. Kuebbing SE, Classen AT, Sanders NJ and Simberloff D (2015). Above-and below-ground effects of plant diversity depend on species origin: an experimental test with multiple invaders. *New Phytologist* **208(3)**:727–735

34. Ghedini G, Marshall DJ and Loreau M (2022). Phytoplankton diversity affects biomass and energy production differently during community development. *Functional Ecology* **36(2)**:446–457

35. Langenheder S, Bulling MT, Solan M and Prosser JI (2010). Bacterial biodiversity-ecosystem functioning relations are modified by environmental complexity. *PLoS ONE* **5(5)**:e10834

36. Mould DL, Botelho NJ and Hogan DA (2020). Intraspecies signaling between common variants of Pseudomonas aeruginosa increases production of quorum-sensing-controlled virulence factors. *mBio* **11(4)**:e01865–20

37. Stintzi A, Evans K, Meyer Jm and Poole K (1998). Quorum-sensing and siderophore biosynthesis in Pseudomonas aeruginosa: lasRllasI mutants exhibit reduced pyoverdine biosynthesis. *FEMS Microbiology Letters* **166(2)**:341–345

38. Thompson J, Johansen R, Dunbar J and Munsky B (2019). Machine learning to predict microbial community functions: an analysis of dissolved organic carbon from litter decomposition. *PLoS One* **14(7)**:e0215502

39. Qu K, Guo F, Liu X, Lin Y and Zou Q (2019). Application of machine learning in microbiology. *Frontiers in Microbiology* **10**:827

40. Midgley GF (2012). Biodiversity and ecosystem function. *Science* **335(6065)**:174–175

41. Shade A (2017). Diversity is the question, not the answer. *The ISME journal* **11(1)**:1–6

Acknowledgements

We thank M. Tikhonov, B. Ogbunugafor, M. Rebolleda-Gómez and all members of the Sanchez lab for helpful discussions. **Funding:** This work was supported by a Packard Foundation Fellowship to A.Sa. and by the National Institutes of Health through grant 1R35 GM133467-01 to A.Sa. **Author contributions:** D.B. and A.Sa. conceived the study. J.D.-C. and A.Sk. analyzed data. J.D.-C. and A.Sa. designed the experiments. J.D.-C. performed experiments. J.D.-C. and A.Sk. processed and analyzed experimental data. J.D.-C., A.Sk., J.C.C.V, D.B. and A.Sa. discussed and interpreted results. J.D.-C. and A.Sa. wrote the paper, with input from A.Sk, J.C.C.V and D.B. **Competing Interests:** The authors declare that no competing interests exist in relation to this manuscript. **Data and materials availability:** All data and code is available at https://github.com/jdiazc9/eco_global_epist.

Supplementary Materials

Materials and Methods
Supplementary Text
Figs. S1 to S6
Table S1
References (*42–43*)