nature ecology & evolution

Article

https://doi.org/10.1038/s41559-023-02197-4

Statistically learning the functional landscape of microbial communities

Received: 24 March 2023

Accepted: 11 August 2023

Published online: 2 October 2023

Check for updates

Abigail Skwara[1], Karna Gowda[2,3], Mahmoud Yousef[2,3], Juan Diaz-Colunga[4], Arjun S. Raman[5,6], Alvaro Sanchez[4], Mikhail Tikhonov[7] ✉ & Seppe Kuehn[2,3] ✉

Microbial consortia exhibit complex functional properties in contexts ranging from soils to bioreactors to human hosts. Understanding how community composition determines function is a major goal of microbial ecology. Here we address this challenge using the concept of community-function landscapes—analogues to fitness landscapes—that capture how changes in community composition alter collective function. Using datasets that represent a broad set of community functions, from production/degradation of specific compounds to biomass generation, we show that statistically inferred landscapes quantitatively predict community functions from knowledge of species presence or absence. Crucially, community-function landscapes allow prediction without explicit knowledge of abundance dynamics or interactions between species and can be accurately trained using measurements from a small subset of all possible community compositions. The success of our approach arises from the fact that empirical community-function landscapes appear to be not rugged, meaning that they largely lack high-order epistatic contributions that would be difficult to fit with limited data. Finally, we show that this observation holds across a wide class of ecological models, suggesting community-function landscapes can be efficiently inferred across a broad range of ecological regimes. Our results open the door to the rational design of consortia without detailed knowledge of abundance dynamics or interactions.

Biology is a science of connecting scales of organization. From proteins to ecosystems, we are faced with the question of how variation at a lower scale of organization gives rise to changes at a higher scale. For example, understanding protein evolution requires learning how variation in the primary amino acid sequence determines fold and function. Similarly, at the level of the organism, genetic variation drives changes in phenotype and fitness. In both cases, interactions between constituent parts give rise to functional system properties.

One of the most powerful conceptual frameworks for thinking about how these functional properties emerge from components and their interactions is the notion of a landscape[1], where the height of the landscape encodes a scalar-valued function or fitness and position on the landscape corresponds to a particular configuration of components. Landscape thinking permits us to articulate key properties of the mapping from genotypes to fitness, including the relative fitness of related genotypes[2], the extent and

[1]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA. [2]Center for the Physics of Evolving Systems, University of Chicago, Chicago, IL, USA. [3]Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA. [4]Department of Microbial Biotechnology, National Center for Biotechnology (CNB-CSIC), Madrid, Spain. [5]Department of Pathology, University of Chicago, Chicago, IL, USA. [6]Duchossois Family Institute, University of Chicago, Chicago, IL, USA. [7]Department of Physics, Washington University in St. Louis, St. Louis, MO, USA. ✉e-mail: tikhonov@wustl.edu; seppe.kuehn@gmail.com

nature of interactions between genes[3,4] and the dynamics of evolutionary trajectories[5–7].

Communities of microbes also exhibit functional properties, from degrading complex substrates[8,9] to resisting invasions[10–12], that arise from the constituent parts and their interactions (Fig. 1a). It is natural to ask whether the landscape concept can also be useful in these scenarios. Here we take inspiration from methods for understanding landscapes for proteins and organisms to characterize the functional landscapes of microbial communities[13].

During the past century, it has become routine to infer landscapes in the protein and organismal context statistically[14–17]. Given that directly assaying all possible mutation combinations is typically infeasible, the statistical approach aims to approximate the landscape from a smaller number of measurements via regression[18], with the interactions between mutations quantified via nonlinear epistatic terms. It is important to note that this approach explicitly neglects the details of the complex and often dynamical underlying processes that cause the fitness change. For example, a statistical approach does not explicitly account for the complex physical interactions between residues that alter the function of an enzyme. Similarly, at the organismal level, this approach neglects the details of how mutations impact gene expression or life history traits. Despite this dramatic simplification, regression-based statistical approaches have been highly successful in both these contexts[2,19,20].

Inspired by these successes, here we take a landscape approach to quantitatively predict functions of interest in microbial communities. From this perspective, the presence and absence of species are analogous to mutations in a protein or genome, and a regression can be formulated to predict community function from species presence and absence alone. This is in contrast to most existing approaches to predicting community function, which almost exclusively seek to understand how species presence impacts abundance dynamics and, consequently, function[21–24]. Here we consider the possibility that community function can be understood without the intermediate step of predicting dynamics (Fig. 1b). We note that this approach explicitly ignores priority effects[25], multistability[26] or any other scenario when presence/absence information does not uniquely specify the community state. Nevertheless, as we will show, community-function landscapes prove remarkably predictive across a range of ecological contexts. This does not mean that priority effects are absent, merely that, for the examples considered here, their impact on community function is, on average, weak enough that the predictive power remains high.

Implementing this approach requires measurements of community function for sets of synthetic communities constructed from libraries of taxa. Here we utilize six existing datasets of this type, representing a diverse set of functional properties[21,22,27–29]. For all the datasets we study, we find that the functional landscape is well described by models including only additive and pairwise epistatic terms. Moreover, we find that the ruggedness of these landscapes is surprisingly low, such that the effects of species presence/absence on function are, in fact, dominated by additive terms. We support these observations computationally, showing that a regression approach succeeds in learning the community-function landscape across a large class of ecological models, despite using only additive and pairwise terms and only species presence/absence as input.

Taken together, our results show that, at least in the six examples presented here, learning the properties of communities can be accomplished without a detailed understanding of the interactions between taxa or their abundance dynamics. Our findings enable a powerful conceptual framework for predicting community functional properties, from invasion resistance to biotechnological applications.

Results

Learning community-function landscapes via regression

We first formulated a statistical approach to fitting community-function landscapes using datasets that comprise measured values of microbial community functions for a set of defined species combinations. In each of these experiments, a defined pool of species was used to construct communities combinatorially. Each community was then incubated, typically for a defined period of time, and then a functional property of interest was assayed. For a pool of N total species, there are 2^N-1 possible species combinations, and measuring all possible combinations is frequently intractable. The first goal of our investigation was to ask whether we can predict the function of all 2^N-1 communities by fitting a statistical model to a small subset ($\ll 2^N-1$) of all possible consortia. The resulting model would provide a global picture of the community-function landscape.

We formulated this problem as a linear regression of the following form:

$$y = \beta_0 + \sum_i \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \ldots, \quad (1)$$

where y is the scalar-valued function and x_i represents the presence or absence of species i in that community. The coefficient β_i is the additive effect of including species i in the community, and β_{ij} is analogous to the effect of pairwise epistasis in genetic fitness landscapes, which measures the impact beyond individual additive effects of adding both species i and j. The ellipses denote higher-order epistasis terms, for example, three-way epistatic terms captured by third-order polynomials and so on.

We took the convention that $x_i = 1$ if species i is present in a community and $x_i = -1$ if that species is absent. We denoted absence using −1 instead of 0 to simplify the interpretation of the regression coefficients (discussion in ref. 30). In brief, using $x_i = \pm 1$ allows us to interpret β_i as the average effect of adding species i to community function, where the average is taken over all community compositions. Similarly, the pairwise coefficient β_{ij} captures the average epistatic effect of species i and j together across many consortia. Moreover, this convention (corresponding to the Fourier expansion of the landscape) has some convenient mathematical properties that make it easier to quantify how much variation in the measured function is captured by additive, pairwise and higher-order terms[30,31].

We considered regressions truncated at first, second and third orders. Many of the datasets we utilized sampled a number of community configurations that are comparable to the total number of coefficients to be inferred. To mitigate the risk of overfitting, we employed L_1-regularized regression (LASSO)[32] using a cross-validation procedure to estimate the regularization hyperparameter (Methods). To assess out-of-sample generalization error, we applied an additional leave-one-out cross-validation scheme in which each data point was iteratively left out of sample and the model was fit to all remaining data points, allowing an out-of-sample prediction for each distinct experimental community (Methods).

Community function is predictable from species composition

We compiled six datasets in which synthetic bacterial communities were assembled from a pool of species. These datasets represent a broad spectrum of community functions: Clark et al.[22] measured the production of the short-chain fatty acid butyrate (Fig. 2a); Langenheder et al.[27] measured a combination of biomass and redox activity on the monosaccharide xylose (comment in Methods); Sanchez-Gorostiaga et al.[21] measured the breakdown of the polysaccharide starch (Fig. 2b,c); Diaz-Colunga et al.[28] measured the total production of iron-scavenging siderophores (Fig. 2d). In addition, we considered biomass-related community functions: work by the Sanchez lab (Methods) measured total community biomass (Fig. 2e), and Kehe et al.[29] measured the abundance of a single target species (Fig. 2f). Details about the size and species pools for each dataset are given in Supplementary Table 1.

In each dataset, community function can be defined as a measurable scalar quantity, for example, the concentration of a compound

https://doi.org/10.1038/s41559-023-02197-4

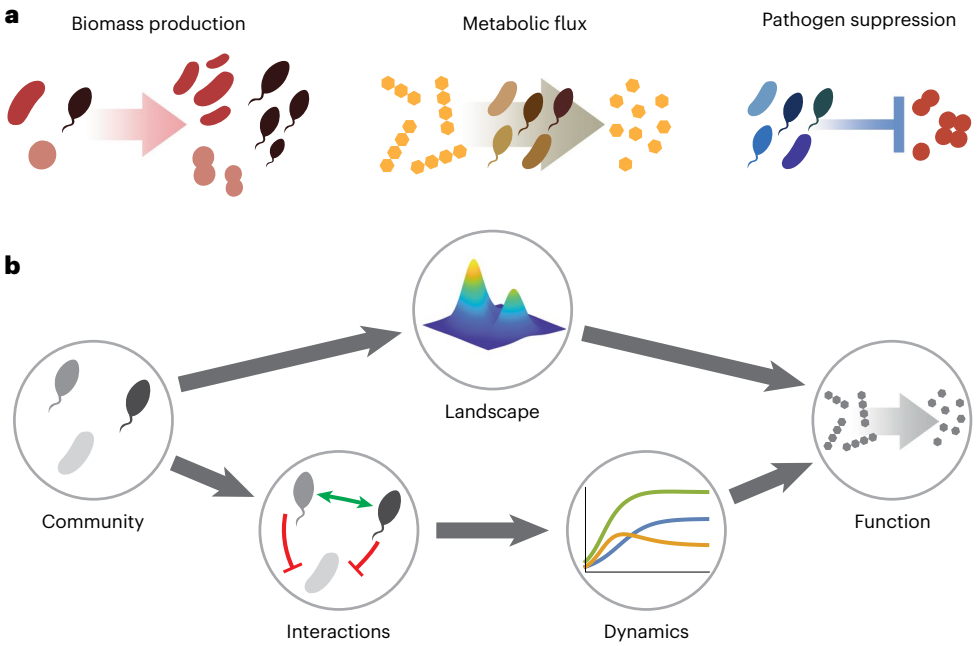


Fig. 1 | Statistically learning community-function landscapes. a, Examples of microbial community functions including (left to right): production of biomass, conversion of substrate to product and suppression of a pathogen. **b**, Contrasting the statistical landscape view (top) of predicting community function with the dynamical view (bottom). In the dynamical view, species abundance dynamics are predicted via an ecological model, which integrates knowledge or measurements of interactions between populations. In contrast, the statistical landscape approach neglects dynamics and measures community function for a set of consortia, allowing functions for all possible community combinations to be inferred statistically.

or a measurement of biomass at a given point in time. Therefore we fit models of the form shown in equation (1). We investigated truncating the model at successively higher orders of epistatic terms to determine what degree of model complexity is needed to accurately predict community function. The quality of the fits are shown in Fig. 2 (bar plots) for increasingly complex models, with shaded bars showing in-sample R^2 and white bars showing out-of-sample R^2 (Methods); scatter plots visualize the quality of second-order out-of-sample model predictions.

Remarkably, across all six datasets studied here, models of first or second order provide high-quality predictions (R^2 ~ 0.8 for second-order models). In most cases, additive models alone (for example, $y = \beta_0 + \sum_i \beta_i x_i$) already have strong predictive power ($R^2 > 0.5$), although the addition of second-order terms yielded an increased quality of fit for all datasets. We note that residuals of observed versus predicted values demonstrate similar patterns of heteroskedasticity across datasets, where communities with higher values of the function relative to the mean tend to be underestimated and those with lower values of the function tend to be overestimated (Fig. 2 and Extended Data Fig. 1). It is possible that these patterns are a consequence of bias induced by regularization[18]. We note that errors in our predictions do not correlate with community richness (Extended Data Fig. 2), indicating that our models generalize well to diverse communities.

For the dataset by Clark et al. (Fig. 2a), the authors of the original study predicted butyrate production using a complex model that parameterized interactions and abundance dynamics. The quality of our statistical predictions using species presence/absence are similar to those obtained using a complex dynamical model[22], suggesting that detailed dissections of community dynamics are not always necessary to make reasonably good predictions of community function.

Empirical community-function landscapes are not rugged

We demonstrated that statistical models based on species presence/absence can predict microbial community functions with surprising accuracy. In particular, simple models containing only additive and/or pairwise epistatic terms explained the vast majority of the variation in the data (Fig. 2). Our statistical approach represents a strategy for approximating the empirical community-function landscapes for these datasets. We wanted to gain intuition for why these regressions appeared to be so successful. To do this, we sought to quantify the ruggedness of community-function landscapes.

In an evolutionary context, the ruggedness of a fitness landscape dictates the number of local fitness optima and has important implications for the predictability of evolution[5]. In a community context, very rugged landscapes are expected to be much harder to approximate globally using regression methods such as those used here, simply because ruggedness arises from a substantial number of strong high-order epistatic terms. High-order terms are challenging to learn statistically due to the explosive combinatorial increase in the number of terms as model order increases. Thus, we expect that rugged landscapes will be difficult to learn statistically, while non-rugged ones will be straightforward to approximate using low-order models.

We quantified ruggedness using two complementary approaches: first, by explicitly quantifying the relative contribution of terms of different orders to the total variance, as explained below; and second, by using an established metric of ruggedness denoted r/s ('roughness over slope'). To pursue the first approach, we started with the combinatorially complete dataset from Langenheder et al.[27] in which growth on xylose was measured for all $2^6 - 1 = 63$ species combinations that can be formed from a six-species pool. This combinatorially complete dataset allowed us to compute the coefficients of the *exact* full-order empirical landscape, which is a model of the form equation (1) that includes epistatic terms of all possible orders (that is, up to sixth order). Using this exactly inferred landscape, we generated a Fourier amplitude spectrum (Methods), which is a decomposition that reflects the total variance of the landscape that is captured by terms of each order[30,31]. This spectrum, shown in Fig. 3a (red line), indicates that ~78% and ~16% of the variance in the landscape is explained by first- and second-order terms, respectively, leaving ~6% of variance remaining for higher-order

https://doi.org/10.1038/s41559-023-02197-4

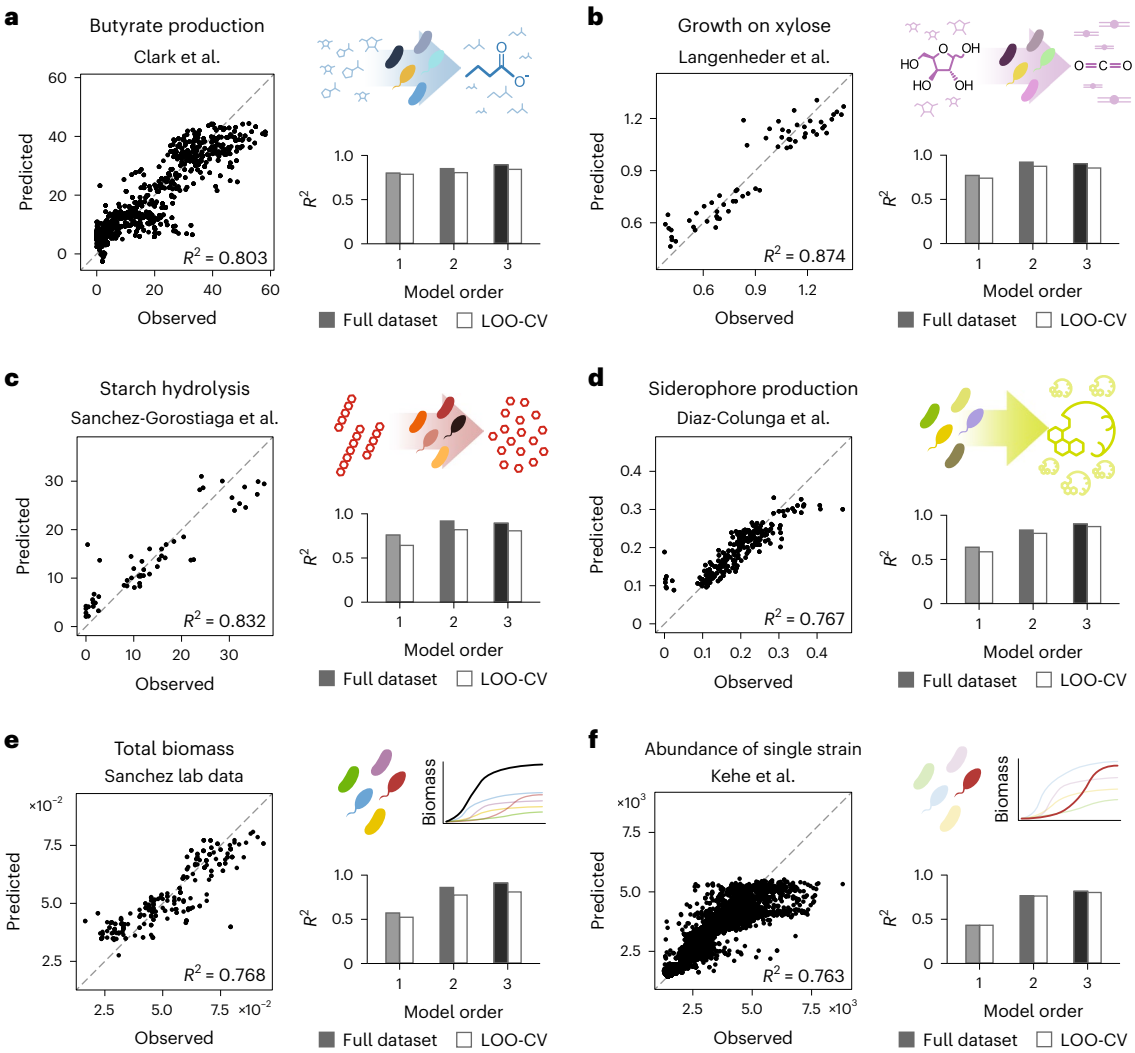


Fig. 2 | Community function is predictable from species presence/absence in empirical datasets. For each dataset, regularized linear regressions were performed using models truncated at the first, second and third order (equation (1)). Bar plots show the quality of fit (R^2) for each of these models, either using all experimental data or using a systematic leave-one-out cross-validation approach (labelled 'LOO-CV'). Scatter plots show out-of-sample prediction values for the second-order regression fits obtained via the leave-one-out cross-validation procedure. **a**–**f**, Analyses are shown for datasets by Clark et al.[22] (**a**), Langenheder et al.[27] (**b**), Sanchez-Gorostiaga et al.[21] (**c**), Diaz-Colunga et al.[28] (**d**), data from the Sanchez lab (**e**) and Kehe et al.[29] (**f**). Extended Data Figs. 1–4 provide more details.

terms. In other words, this exact community-function landscape displays a low degree of ruggedness, as it is dominated by low-order terms and is largely free of consequential higher-order terms.

To assess whether this result reflects a meaningful property of the empirical community-function landscape, we performed a randomization test by computationally shuffling the assignments between function measurements and community compositions. For 100 such randomizations, we inferred the new landscape and computed the Fourier amplitude spectra. The results are plotted in Fig. 3a (black line). For the randomized landscapes, we found that terms of third order were most important and additive terms alone captured only 10% of the variance; the peak at third order arises from the fact that there are combinatorially more terms possible at this order than any other. We concluded that a lack of ruggedness in the true landscape is not spurious but rather a distinctly non-random structural feature of this landscape.

We performed similar analyses for the remaining five datasets to estimate the relative importance of coefficients of different orders. Although these datasets are not combinatorially complete and therefore do not permit the inference of the exact amplitude spectra, we inferred a truncated amplitude spectra via third-order regression. We found that in each case, the dominant coefficients are additive, with pairwise coefficients typically the next most important (Fig. 3b). The large fraction of variance explained by additive and pairwise coefficients across these cases again indicates a low degree of ruggedness.

As a second approach to quantify ruggedness, we computed the roughness/slope ratio (r/s) (refs. 33,34), a commonly used ruggedness metric that quantifies how well a landscape is fit by a purely additive model. Explicitly, roughness r is computed by fitting a model with additive terms only (for example, $y = \beta_0 + \sum_i \beta_i x_i$) and determining a residual to this fit. The roughness (r) is the root-mean-square value of these residuals. Slope s is defined as the mean (absolute) value of the additive coefficients β_i. The ratio r/s represents the typical magnitude of additive model error relative to the typical magnitude of an additive term. Large values of this ratio mean that the approximation afforded by an additive model is poor, indicating a high degree of ruggedness.

To make a well-defined comparison between all datasets, we computed normalized r/s values, defined as the ratio between r/s on the original dataset and r/s for 100 randomized landscapes (Methods). Randomized landscapes served here as a natural high-ruggedness comparison. We found that this ratio was consistently $\ll 1$, indicating

https://doi.org/10.1038/s41559-023-02197-4

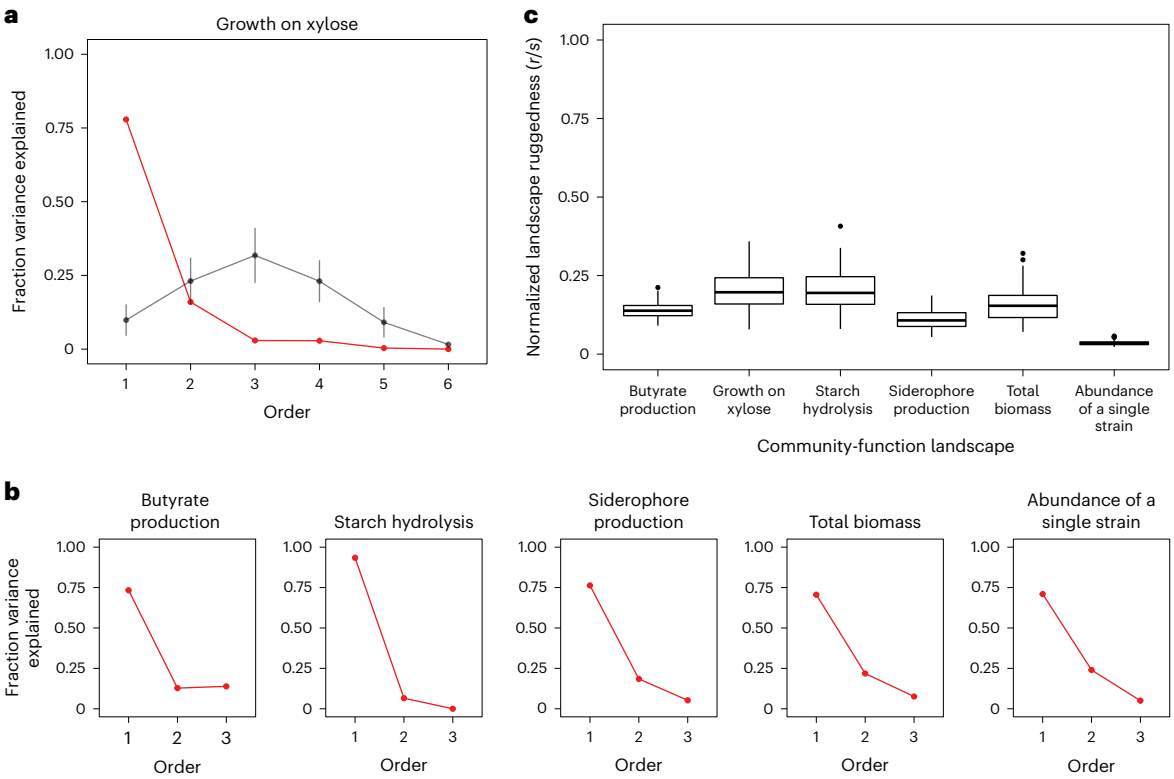


Fig. 3 | Empirical community-function landscapes are not rugged. **a**, Normalized Fourier amplitude spectrum obtained from the combinatorially complete landscape from Langenheder et al.[27]. Normalized amplitude values correspond to the fraction of total landscape variance that is captured by terms at each order. The empirical amplitude spectrum (red) demonstrates that landscape variance is primarily explained by first-order terms, with higher-order terms explaining decreasing fractions of the variance. This denotes a low degree of ruggedness in the empirical landscape. Black traces show spectra obtained from randomized landscapes (points and error bars indicate the mean and standard deviation, respectively, across 100 randomizations). Unlike the empirical landscapes, the randomized versions are rugged: first-order terms explain a relatively small fraction of total variance. **b**, For each additional dataset that is not combinatorially complete, the fitted coefficients of the regularized third-order linear regression are used to infer the normalized amplitude spectrum at first through third order. As in panel **a**, first-order terms explain more variance in the landscape than terms at second and third orders, indicating a lack of ruggedness. **c**, For each dataset, normalized r/s values (a measure of ruggedness; text) are computed by calculating the ratio of empirical r/s values on original landscapes to the r/s value of 100 randomized landscapes. The normalized r/s values for all datasets are notably smaller than 1, again indicating a low degree of ruggedness in empirical landscapes compared to their randomized counterparts. The boxplots show the median as the centre line, with the boxes corresponding to the upper and lower quartiles, whiskers corresponding to values that lie within 1.5× the interquartile range and individual points indicating outliers.

that true landscapes are much less rugged than comparable random landscapes (Fig. 3c).

We concluded that across a diverse range of microbial community functions, empirical landscapes possess a low degree of ruggedness, corresponding to the dominance of low-order (that is, additive and pairwise epistatic) terms, and a notable absence of higher-order epistatic terms. This enables low-order statistical models to faithfully parameterize empirical landscapes and thereby accurately predict community functions.

Ecological models indicate when landscapes can be learned

We have demonstrated multiple empirical examples where the functional landscape of a microbial community proves to be non-rugged, allowing low-order statistical models to predict the function of interest. To understand the expected generality of this observation beyond the six datasets considered here, we turned to ecological models, generating large ensembles of random communities for which the functional landscapes can be evaluated *in silico*. Our goal was not to fit models to the empirical examples above but to probe the conditions under which a lack of ruggedness is expected to be rare or common and to identify the ecological scenarios under which low-order statistical inference is expected to fail.

We first sought to interrogate synthetically generated community-function landscapes using the generalized Lotka–Volterra (gLV) model, with total abundance (biomass) serving as the function of interest. The gLV model has many variants that differ in assumptions regarding the structure of randomly generated species interactions, but as recently argued by Barbier et al., a large class of such variants can be reduced to a four-parameter 'reference model'[35]. We adopted this reference model for our analyses (Methods), focusing on a sweep of the two key parameters describing interaction strength (μ) and interaction variability (σ). Further analysis varying all four parameters is presented in Supplementary Fig. 5.

For each combination of μ and σ, we performed 10 trials of generating a random pool of $N = 10$ species (a pool small enough to evaluate the combinatorially complete set of all possible communities) and computed the exact community-function landscape (mimicking our procedure for the Langenheder et al. data in Fig. 3a). The fraction of variance explained by the first- and second-order terms, equivalent to the predictive power (R^2) of the second-order approximation of the landscape, is shown in Fig. 4a. These values are averaged over the ten trials at each point in the μ–σ plane. In the reference model, positive μ corresponds to interactions that are competitive on average, while larger values of σ correspond to a greater degree of variability in the

https://doi.org/10.1038/s41559-023-02197-4

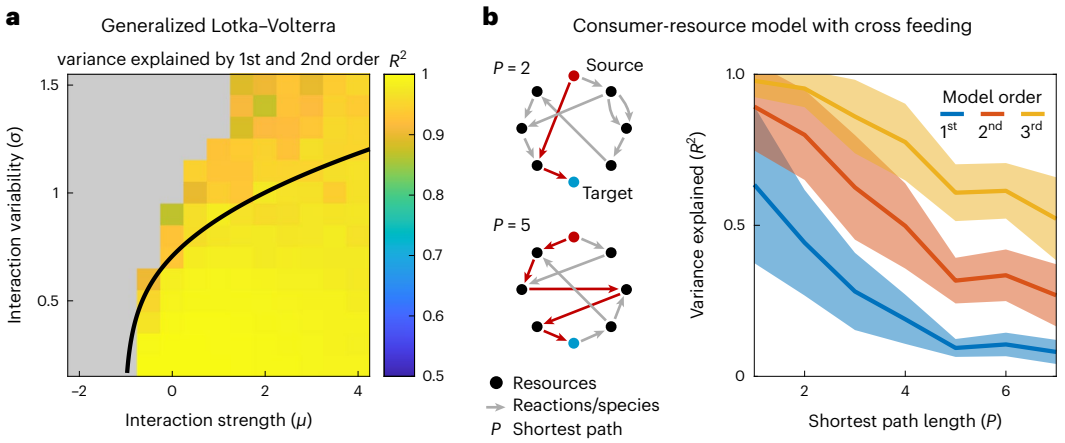


Fig. 4 | Ecological models indicate both optimism and caution for inferring community-function landscapes. a, The gLV model was used to generate combinatorially complete synthetic community-function landscapes, taking total abundance as the function of interest. Landscapes were generated using an $N = 10$ species pool, and two parameters controlling the structure of randomly drawn interactions were varied: an interaction-strength parameter μ and an interaction variability parameter σ, with ten landscape trials for each point in the μ–σ plane. Heat map shows the (trial-averaged) R^2 of the second-order approximation of the exact landscape, that is, the variance explained by first- and second-order terms combined. This value is computed from the exact landscape coefficients inferred as in Fig. 3a (no fitting required). Note that as interaction variability σ becomes large, the model becomes unstable, causing species abundances to diverge. In our simulations, any points in the μ–σ plane that encountered divergences in more than five landscape trials are indicated in grey. The black curve shows an analytical prediction (computed in the $N \to \infty$ limit[36]) for the stability boundary of the gLV model, beyond which species abundances will typically diverge. Supplementary Figs. 5 and 6 detail more. **b**, A consumer resource model (CRM) was used to generate synthetic community-function landscapes with randomly generated cross-feeding networks (schematic). One resource was supplied externally ('source'), while the function of the community was defined to be the concentration of a different resource ('target'). Networks comprising $N = 10$ species and $L = 8$ resources were generated so as to vary the length of the shortest path (P) connecting the source resource to the target. Plot shows the variance explained (R^2) by models truncated at first, second and third order, computed using the exact landscape coefficients and shown as a function of P. The R^2 value declines with P, as expected. Solid lines correspond to mean values obtained across trials, with error bands indicating ±1 standard deviation.

strength of interactions. Note that as interaction variability σ becomes large, the Lotka–Volterra model becomes unstable, causing species abundances to diverge[36]. In our simulations, any points in μ–σ plane that encountered divergences in more than five trials are indicated in grey (Fig. 4a). Additional details about these simulations are described in Methods.

We found that across the entire range of parameters for which the dynamics of the model are stable, second-order regression provides excellent fits ($R^2 \gtrsim 0.9$) to gLV community-function landscapes (Fig. 4a; the non-grey region is all yellow). Further analysis shows that R^2 remains similarly high even as we change other model parameters, including interaction asymmetry and the variability of species' carrying capacity (Supplementary Fig. 5). These results indicate that non-rugged functional landscapes can be observed across a wide range of ecological scenarios.

It is important to stress that the lack of ruggedness is not a trivial property of all gLV models. Although the Lotka–Volterra model studied here contains only second-order interactions, these interactions couple species *abundances*, and therefore we do not expect the functional landscape to be well approximated by a low-order regression on species presence/absence alone. For example, rugged gLV landscapes can and do exist (Supplementary Fig. 6). In the ensemble of models described by Barbier et al., such examples are curiously rare (Supplementary Fig. 6), at least for biomass as the property of interest. However, it is likely that for other ensembles (for example, with a more complex correlation structure of interactions), rugged landscapes could be more common. Identifying the ecological mechanisms dictating whether the landscape of a given functional property will be rugged or smooth is an important question for future work.

As an illustration, we present one mechanism that will generically lead to increased ruggedness. Consider a nutrient X_1 that is broken down through a chain of L reactions following a linear pathway $X_1 \xrightarrow{n_1} X_2 \xrightarrow{n_2} \ldots \xrightarrow{n_L} X_{L+1}$, where each reaction is performed by a specialized community member n_i. If X_1 is the only nutrient supplied to the consortium, and the function of interest is the concentration of the end product X_{L+1}, then the concentration of X_{L+1} will be non-zero if and only if all species $n_1, \ldots n_L$ are present. Mathematically, if presence or absence were denoted as $s_i = 0$ or 1, respectively, such a landscape would be described by a single Lth-order term: $y = \beta_{1\ldots L}\, s_1 s_2 \ldots s_L$. Under the convention used here, where presence or absence is denoted by $x_i = \pm 1$, this corresponds to coefficients at all orders being equally important. (As a simple illustration, for three species $s_1 s_2 s_3 = (x_1 + x_2 + x_3 + x_1x_2 + x_2x_3 + x_1x_3 + x_1x_2x_3)/8$, which is easy to check by a direct substitution $s_i \equiv \frac{x_i+1}{2}$.) In this scenario, regressions exploiting only low-order models will provide poor predictive power and the landscape will be rugged (Fig. 4b).

To build on the intuition from this thought experiment, we used a consumer resource model (CRM) to generate synthetic landscapes with $N = 10$ species competing for $L = 8$ nutrients. Only one resource was supplied externally ('source', Fig. 4b, red dot), while the function of the community was defined to be the concentration of one of the other resources ('target', Fig. 4b, blue dot). In each trial, we constructed a random cross-feeding network for a pool of $N = 10$ species, with each species capable of converting one randomly chosen resource into another. We then computed the complete functional landscape, as above for the gLV, by simulating all possible subsets of this ten-species pool. Simulation details are described in Methods.

On the basis of the intuition from the thought experiment above, we expected the predictive power of low-order approximations to correlate with the length of the shortest path (P) connecting the source resource to the target (schematic, Fig. 4b). The landscapes corresponding to cross-feeding networks with long paths from source to target (that is, large P) are expected to be more rugged. To confirm this, Fig. 4b plots the fraction of variance (R^2) explained by low-order approximations of the exact synthetic CRM landscapes, shown as a

function of the shortest chain length P. We observe that while successful in cross-feeding networks with small P, low-order models were increasingly challenged as P increases, with R^2 dropping below 0.5 for second-order models beyond $P = 4$. Increasing model order from additive only to third order substantially improved predictive power, particularly at large P, consistent with the idea that increasing P increases the prevalence of higher-order terms. These results illustrate that low-order approximations of community-function landscapes may fail to make accurate predictions in ecological scenarios with long chains of trophic dependencies or other situations when function is strongly contingent on the simultaneous presence of multiple species.

Discussion

The key result of our study is the demonstration that functional properties of microbial communities can be predicted by simple statistical models knowing only which species are present or absent. Remarkably, in analogy to fitness landscapes describing proteins and organisms, we showed that regressions can quantitatively describe empirical landscapes for a wide range of community functions. We found that the success of these regressions derives from the fact that the underlying landscapes are not rugged, allowing the majority of variation in function to be captured by additive and pairwise terms in the regression. The predictions did not require specialized knowledge or measurement of microscopic system properties, only a dataset comprising quantitative measurements of a function taken from a collection of defined communities drawn from a fixed pool of species.

Our simulations of generalized Lotka–Volterra and consumer resource models demonstrated that we can expect low-order approximations of landscapes to work well across a range of ecological contexts. We identified a clear exception, however, in functions that are strongly contingent on the simultaneous presence of multiple highly specialized community members. One might therefore expect that the low-order landscape approximations might encounter challenges, for example, in systems with long linear chains of reactions, such as those present in anaerobic digesters[37] and Winogradsky columns[38]. A critical direction to be addressed in future work is a systematic analysis of the ecological mechanisms that either enable or impede the performance of approximations of community-function landscapes. A better understanding of these mechanisms would provide a more principled view of community functional properties amenable to our statistical approach.

Because our approach predicts community function from species presence/absence, it explicitly assumes that replicate communities with the same composition will have the same function. This assumption could fail if communities exhibit alternative stable states[39,40] with distinct final abundances and functions despite identical presence/absence compositions. While alternative stable states have been documented in synthetic consortia[26], they do not appear to be widespread or have large impacts on function in the communities studied here. Extended Data Fig. 3 shows predictions for replicate communities with identical compositions. We find that most replicates in all six datasets studied are well predicted by our model. We conclude that alternative stable states do not drive substantial functional variation across the consortia studied here. However, none of the datasets studied here systematically varied the initial relative abundances for communities of fixed composition, and this may result in communities not reaching alternative stable states despite their existence. In ecological contexts where alternative stable states are pervasive and drive large functional variation, our predictions would begin to degrade in quality.

Another question is how our results could be extended to contexts with extensive functional redundancy between taxa. Our regression approach would probably struggle in these scenarios. This is easiest to see in the extreme limit of taxa that are perfectly interchangeable: a high-dimensional OR function, where OR is defined as the Boolean logical operator, has Fourier terms of all orders, and any low-order model would be a poor approximation. Thus, the excellent performance of our method on the available datasets is probably aided by the fact that the labour-intensive nature of the combinatorial experiments favours synthetic communities with relatively low redundancy. The experimental cost associated with increasing the number of species discourages the inclusion of taxa highly similar to those already in the pool. This is the opposite regime of the natural communities, which are often phylogenetically under-dispersed[41,42].

Whereas the method as presented would probably struggle in communities with extensive redundancy, the argument above indicates how this limitation could be remedied, namely by grouping redundant taxa before performing regressions. In fact, our results suggest that the high predictive power of a simple regression on the variables describing the presence or absence of any member in a group could be taken as the criterion indicating that the grouping was chosen appropriately. This offers a path towards quantitative prediction of complex community functions in the high-diversity regime and echoes recently proposed ideas from multiple groups[43–45].

Our results complement previous studies demonstrating that community-function landscapes follow patterns of global epistasis[28]. These patterns were first discovered in the context of organismal fitness landscapes[46–48]. Global epistasis refers not to the impacts of individual epistatic terms but rather to properties (for example, diminishing returns) that emerge from the collective impact of many epistatic contributions. In the context of microbial communities, global epistasis arises as a linear relationship between the impact of adding a specific species on community function and the function of the 'background' community to which the species was added[28]. It is worth noting that the regression and global epistasis approaches represent distinct strategies for predicting community function. The regression approach presented here attempts to predict the function of an arbitrary community via an ansatz assuming that only low-order epistatic contributions are important, whereas the global epistasis approach attempts to estimate the effect of adding a species to a given background community harnessing the predictability of patterns arising from combined epistatic contributions up to arbitrary order[47]. Connecting the concept of global epistasis with the regression approach to learning functional landscapes remains an important exercise, as the two approaches may provide complementary insights in different ecological scenarios.

Perhaps the greatest downstream impact of our study is the possibility of using statistically inferred landscapes to rationally design communities with desired functional properties. Because community-function landscapes can be approximated by sampling only a subset of all possible species combinations, it is conceivable that even large synthetic consortia with predefined functions can be designed and optimized computationally given only a small number of measurements. Determining the optimal sampling strategy to accurately infer landscapes remains an important avenue for future work. For example, it is unclear whether the sampling should include a mix of simple communities of <3 taxa or whether high-diversity communities are more informative given a restricted number of measurements. However, even without extensive optimization, our models were able to identify the communities with the highest functional output, even when these communities were left out of sample. The simplicity of the approach makes it readily portable across contexts and functions and its performance could offer an appealing advantage relative to alternative design strategies[22,49].

Methods

Collection and preprocessing of datasets

Datasets were compiled from six experimental efforts to measure various community functions in defined synthetic microbial consortia. Details about these datasets and references are listed in Supplementary Table 1.

The dataset by Langenheder et al.[27] was generated by measuring the activity of synthetic communities with xylose provided as the sole carbon source. Metabolic activity was measured via the absorbance of

a redox dye (tetrazolium violet) at 600 nm over multiple time points. It should be noted that because cell scattering at 600 nm probably also contributed to absorbance, the functional values collected in this study reflect a combination of both redox activity and biomass growth. While this detail complicates the mechanistic interpretation of the data, it does not present a problem for our analysis, as the data still represent a complex functional property for the regression approach to predict. Here the community-function landscape we sought to approximate was constructed from the functional values collected at the endpoint of the experiment (48 h).

The dataset by Sanchez-Gorostiaga et al.[21] included 53 community configurations out of the possible 63. We note that as we confirmed with the authors of the original study, the potentially ambiguous phrasing in the original manuscript ('every combination of six amylolytic soil bacteria') referred to every *pairwise* combination being included in the list, not that the dataset was combinatorially complete.

Total community biomass data collected by Diaz-Colunga et al. were not included in the original manuscript[28]. Detailed experimental protocols for isolation, culturing and community assembly are given in refs. 28,50, with the biomass dataset differing from these methodological details in the following ways: (1) the eight isolates used in the biomass dataset were distinct from the siderophore production isolates (though isolated using the same methodology), (2) the medium used in biomass growth experiments lacks a trace mineral supplement, (3) biomass was quantified by measuring the optical density (OD) at 600 nm using 100 µl of endpoint cultures in an AccuSkan FC plate reader (Fisher Scientific) and (4) monoculture measurements for four strains were omitted, resulting in a dataset of 160 unique configurations (instead of 164 as should otherwise result from the methodology in ref. 28). Though isolates used in the biomass dataset were not sequenced or taxonomically identified, colonies either possessed distinct morphologies, possessed distinct colour profiles when grown on chromogenic agar plates (CHROMagar Mastitis GN) and/or were isolated from separate environmental samples and were therefore probably genomically distinct.

Data from Kehe et al.[29] were generated using a microwell array approach in which each community was assembled by randomly grouping nanolitre droplets of defined species composition into 2–19 droplet combinations. Due to this stochastic assembly, initial species abundances varied beyond binary presence/absence: for example, a three-droplet combination containing two droplets of species A and one droplet of species B will have different initial abundances than the combination of one droplet of species A and two droplets of species B. Because the formulation of community-function landscapes in equation (1) operates on binary species presence/absence, this variation in initial abundances was ignored and a species was considered to be present in a community if it had positive initial abundance.

In cases where datasets contained experimental replicates, the mean over replicates was taken to obtain a single community-function value for each unique species combination.

Statistical inference of landscapes via regression

Community-function landscapes were approximated for empirical datasets by fitting equations of the form equation (1) truncated at first, second and third order via LASSO regularization[32]. Tenfold cross validation was used to estimate the regularization hyperparameter. All models were fit using the package glmnet in R version 4.1. 2. For all datasets considered here, the computational demands of fitting our models were modest, and all model fitting and analysis was performed using a personal computer.

To fit the data, two strategies were employed. First, all available data points were used to fit landscape models. The coefficients of determination (R^2) for these fits are shown in grey bars in Fig. 2, and predicted versus observed values are shown in Extended Data Figs. 3 and 4. Second, to obtain an estimate of out-of-sample model accuracy, a leave-one-out procedure was employed. Individual data points corresponding to each distinct experimental community were systematically left out of sample, and models were then fit to all remaining data points using tenfold cross validation. The observed versus predicted values of left-out points estimated via this approach for a second-order model are shown in the scatter plots of Fig. 2. The prediction quality (R^2) for left-out points are shown in white bars in Fig. 2.

Calculation of Fourier amplitude spectrum

Because equation (1) with $x_i \in \{-1, 1\}$ corresponds to the Fourier expansion of a fitness landscape[30], the Fourier amplitude equation[31] at order p can be written simply as

$$A_p = \sum_{i \in p} \beta_i^2. \tag{2}$$

It can be shown that sum of amplitudes across all orders is equal to the total variance of the landscape. This permits the calculation of the fraction of total variance explained by terms at order p as $A_p / \sum_{p'} A_{p'}$.

Calculation of *r*/*s* ruggedness metric

The roughness-slope ratio, r/s, is a measure of landscape ruggedness that quantifies how well a landscape is fit by a purely additive model. This quantity was computed by first fitting a linear model of the form:

$$y = \beta_0 + \sum_i \beta_i x_i, \tag{3}$$

where y is the measured community function, and the coefficients β_0 and β_i are obtained ordinary least-squares regression.

The roughness, r, is defined as the root-mean-squared-error of the resulting fit, or

$$r = \sqrt{\frac{1}{L} \sum_i (y_i - \hat{y}_i)^2}, \tag{4}$$

where L is the number of data points in the landscape and $\hat{y}$ is the value of y fitted by linear regression. The slope, s, is defined as the mean of the absolute value of coefficients β_i

$$y = \frac{1}{n} \sum_i |\beta_i|. \tag{5}$$

Larger values of r/s indicate a greater deviation from linearity, therefore a more rugged landscape. In contrast, an r/s value of 0 would correspond to a perfectly additive landscape.

The values of r/s depend on landscape size, and sensible comparisons of this quantity between datasets require an appropriate normalization. Because randomized landscapes represent a natural high-ruggedness comparison, r/s values computed on randomized landscapes were used as scaling factors, for example

$$r/s_{\text{normalized}} = \frac{r/s_{\text{original}}}{r/s_{\text{randomized}}}, \tag{6}$$

was computed. Normalized r/s values close to 1 indicate that the empirical landscape is as rugged as the randomized landscape, while values close to zero indicate a low degree of ruggedness in the empirical landscape. One hundred randomizations were performed for each dataset, and values of $r/s_{\text{normalized}}$ were computed for each; the distributions of these values are shown in Fig. 3c.

Simulations of the generalized Lotka–Volterra model

Community-function landscapes were synthetically generated using the generalized Lotka–Volterra (gLV) model, with total abundance

serving as the community function of interest. A four-parameter 'reference model' formulated by Barbier et al.[35] was used to explore important dimensions of the gLV parameter space. The model can be written as follows:

$$\dot{N}_i = \frac{r_i}{K_i} N_i \left(K_i - N_i - \sum_j \alpha_{ij} N_j \right), \tag{7}$$

where for species i in a pool of N total species, N_i is the abundance, r_i is the intrinsic growth rate and K_i is the carrying capacity. The parameters α_{ij} are interaction coefficients between species i and j.

The equilibria of equation (7) are determined by the values of K_i and α_{ij}. For each synthetic landscape, these parameters were randomly generated. The carrying capacities K_i were independently drawn from a gamma distribution with mean 1 and variance ζ^2, while the interaction coefficients were drawn from a normal distribution with mean μ / N and variance σ^2 / N. These definitions ensure that the full range of distinct qualitative behaviours is spanned by parameter values of order 1 (ref. 51 for details). A broad, ecologically relevant regime of the interaction parameter space was explored in Fig. 4a ($\mu \in [-1, 4]$ and $\sigma \in [0, 1.5]$). The standard deviation of carrying capacities was fixed at $\zeta = 0.3$. The fourth and final parameter is interaction asymmetry, defined as $\gamma = \mathrm{corr}(\alpha_{ij}, \alpha_{ji})$. In Fig. 4a, interactions were set to be symmetric by taking $\gamma = 1$. These parameter values and ranges are identical to those used in Fig. S2 of ref. 35, allowing direct comparison with Fig. 4a. Supplementary Fig. 5 provides a more thorough parameter sweep varying both γ and ζ.

Landscapes were generated over a grid of points in μ–σ space. Each landscape was generated through the following steps:

1. A set of carrying capacity parameters K_i and interaction parameters α_{ij} were drawn for a pool of $N = 10$ species as described above. All r_i were fixed to 1.
2. Equation (7) was simulated to equilibrium for all species combinations.
3. Total endpoint 'biomass' (sum of abundances $\sum_i N_i(\infty)$) was computed for each simulation.
4. Exact, full-order landscapes (equation (1)) were computed, using biomass as the community function.

Initial species abundances were drawn from an exponential distribution with mean 0.1. Numerical integration was performed using ode15s (MATLAB). At each point in the μ–σ space, ten trials of landscapes were generated. Note that for larger values of σ, species abundances in equation (7) are more likely to diverge. Parameter combinations for which divergences were encountered in more than half of trials are indicated by grey values in Fig. 4a.

Simulations of the consumer resource model with random cross-feeding networks

Community-function landscapes were synthetically generated using a consumer resource model (CRM), taking the equilibrium concentration of a terminal waste product as the function of interest. The CRM is given as follows:

$$\begin{aligned} \dot{N}_i &= N_i (r_i - m_i), \\ r_i &= \sum_\alpha C_{i\alpha} R_\alpha, \\ \dot{R}_\beta &= \frac{K_\beta - R_\beta}{\tau_\beta} - \sum_i C_{i\beta} R_\beta N_i + \sum_i \gamma r_i N_i D_{i\beta}. \end{aligned} \tag{8}$$

Here N_i, r_i and m_i are the abundance, total resource uptake and maintenance costs, respectively, of species i in a pool of N total species. R_α is the concentration of resource α. The matrices $C_{i\alpha}$ and $D_{i\beta}$ describe which resources a species consumes and secretes, respectively, with secretions assumed proportional to the metabolic uptake r_i. An efficiency factor $\gamma < 1$ ensures that energy cannot be gained but only lost. Here γ was set to 0.5. The decay rates τ_β and species maintenance costs m_i are all set to 1 for simplicity. The resource carrying capacity for the single externally supplied resource, K_1, was set to 10^5, and all remaining resource carrying capacities, K_β, were set to 0. The total number of resources was fixed to $L + 1$, with resource $L + 1$ representing a terminal waste product that no species can consume.

Random cross-feeding networks were generated to explore how chains of trophic dependencies impact the ruggedness of community-function landscapes. To do this, it was assumed that each species i can consume and secrete exactly one resource, denoted in_i and out_i, respectively. These were selected through the following steps, performed independently for each species:

1. The identity of the consumed resource in_i was chosen randomly between 1 and L with equal probability.
2. The secreted resource out_i can range from 1 to $L + 1$, where resource $L + 1$ is the terminal waste product; however it must be distinct from in_i, leaving L possible choices. With probability p, we set $\mathrm{out}_i = \mathrm{in}_i + 1$; otherwise out_i was drawn from any of the remaining values at random.

The parameter p thus allows the exploration of a range of network topologies, from long linear pathways with a high degree of trophic dependency (at $p \approx 1$) to random graphs (at $p \approx 1 / L$).

After selecting in_i and out_i for all species, we verified that resource $L + 1$ was 'reachable' through the network from resource 1, that is, whether there existed at least one path from resource 1 to resource $L + 1$. If this was not the case, the functional landscape (the concentration of resource $L + 1$ for a given set of species) would be identically zero; such networks were discarded as invalid and the steps above were repeated until a valid circuit was obtained.

For Fig. 4b, random cross-feeding networks were generated by carrying out the steps above across a range of values of p, fixing $N = 10$ total species and $L = 8$ total resources. Twenty valid random trials over 15 values of $p \in (1 / L, 1)$ were generated to create 300 total random cross-feeding networks. The community-function landscapes for these networks were then computed by setting initial species abundances to 0.1, simulating equation (8) to equilibrium for all species combinations (ode15s, MATLAB), taking the equilibrium concentration of resource $L + 1$ as the function of interest and using this complete combinatorial landscape to determine the exact coefficients of its Fourier decomposition (equation (1)).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data analysed here are either available from the original studies or in the following repository: https://github.com/abbyskwara2/regression_on_landscapes.

Code availability

Code to run all analyses presented in this paper is available in the following repository: https://github.com/abbyskwara2/regression_on_landscapes.

References

1. Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. XI Int. Congr. Genet.* **8**, 209–222 (1932).
2. Ferguson, A. L. et al. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* **38**, 606–617 (2013).
3. Chou, H. H., Chiu, H. C., Delaney, N. F., Segrè, D. & Marx, C. J. Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science* **332**, 1190–1192 (2011).

4. Khan, A. I., Dinh, D. M., Schneider, D., Lenski, R. E. & Cooper, T. F. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* **332**, 1193–1196 (2011).
5. Kauffman, S. & Levin, S. Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.* **128**, 11–45 (1987).
6. Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445**, 383–386 (2007).
7. Kryazhimskiy, S., Tkăcik, G. & Plotkin, J. B. The dynamics of adaptation on correlated fitness landscapes. *Proc. Natl Acad. Sci. U.S.A.* **106**, 18638 (2009).
8. Datta, M. S., Sliwerska, E., Gore, J., Polz, M. F. & Cordero, O. X. Microbial interactions lead to rapid micro-scale successions on model marine particles. *Nat. Commun.* **7**, 11965 (2016).
9. Solden, L. M. et al. Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. *Nat. Microbiol.* **3**, 1274–1284 (2018).
10. Jones, M. L., Rivett, D. W., Pascual-Garria, A. & Bell, T. Relationships between community composition, productivity and invasion resistance in semi-natural bacterial microcosms. *eLife* **10**, e71811 (2021).
11. Wagner, A. Competition for nutrients increases invasion resistance during assembly of microbial communities. *Mol. Ecol.* **31**, 4188–4203 (2022).
12. Cheng, A. G. et al. Design, construction, and in vivo augmentation of a complex gut microbiome. *Cell* **185**, 3617–3636 (2022).
13. Sanchez, A. et al. The community-function landscape of microbial consortia. *Cell Syst.* **14**, 122–134 (2023).
14. Fisher, R. A. *The Genetical Theory of Natural Selection* (Clarendon Press, 1930).
15. Price, G. R. Selection and covariance. *Nature* **227**, 520–521 (1970).
16. Lande, R. & Arnold, S. J. The measurement of selection on correlated characters. *Evolution* **37**, 1210–1226 (1983).
17. Uffelmann, E. et al. Genome-wide association studies. *Nat. Rev. Methods Primers* **1**, 59 (2021).
18. Otwinowski, J. & Plotkin, J. B. Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proc. Natl Acad. Sci. U.S.A.* **111**, E2301 (2014).
19. Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O. & Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* **5**, e16965 (2016).
20. Poelwijk, F. J., Socolich, M. & Ranganathan, R. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nat. Commun.* **10**, 4213 (2019).
21. Sanchez-Gorostiaga, A., Bajić, D., Osborne, M. L., Poyatos, J. F. & Sanchez, A. High-order interactions distort the functional landscape of microbial consortia. *PLoS Biology* **17**, e3000550 (2019).
22. Clark, R. L. et al. Design of synthetic human gut microbiome assembly and butyrate production. *Nature Communications* **12**, 3254 (2021).
23. Gowda, K., Ping, D., Mani, M. & Kuehn, S. Genomic structure predicts metabolite dynamics in microbial communities. *Cell* **185**, 530–546 (2022).
24. van den Berg, N. I. et al. Ecological modelling approaches for predicting emergent properties in microbial communities. *Nat. Ecol. Evol.* **6**, 855–865 (2022).
25. Debray, R. et al. Priority effects in microbiome assembly. *Nat. Rev. Microbiol.* **20**, 109–121 (2022).
26. Amor, D. R., Ratzke, C. & Gore, J. Transient invaders can induce shifts between alternative stable states of microbial communities. *Sci. Adv.* **6**, eaay8676 (2020).
27. Langenheder, S., Bulling, M. T., Solan, M. & Prosser, J. I. Bacterial biodiversity-ecosystem functioning relations are modified by environmental complexity. *PLoS ONE* **5**, e10834 (2010).
28. Diaz-Colunga, J., Skwara, A., Vila, J. C. C. & Bajic, D. Global epistasis and the emergence of ecological function. *bioRxiv* https://doi.org/10.1101/2022.06.21.496987 (2023).
29. Kehe, J. et al. Massively parallel screening of synthetic microbial communities. *Proc. Natl Acad. Sci. U.S.A.* **116**, 12804–12809 (2019).
30. Poelwijk, F. J., Krishna, V. & Ranganathan, R. The context-dependence of mutations: a linkage of formalisms. *PLoS Comput. Biol.* **12**, e1004771 (2016).
31. Hordijk, W. & Stadler, P. F. Amplitude spectra of fitness landscapes. *Adv. Complex Syst.* **01**, 39–66 (1998).
32. Hastie, T., Tibshirani, R. J., & Friedman, J. *The Elements of Statistical Learning* 2nd edn (Springer, 2008).
33. Szendro, I. G., Schenk, M. F., Franke, J., Krug, J. & De Visser, J. A. G. Quantitative analyses of empirical fitness landscapes. *J. Stat. Mech.: Theory Exp., 1:* P01005 (2013).
34. de Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* **15**, 480–490 (2014).
35. Barbier, M., Arnoldi, J.-F., Bunin, G. & Loreau, M. Generic assembly patterns in complex ecological communities. *Proc. Natl Acad. Sci. U.S.A.* **115**, 2156–2161 (2018).
36. Bunin, G. Ecological communities with Lotka–Volterra dynamics. *Phys. Rev. E* **95**, 042414 (2017).
37. Vanwonterghem, I. et al. Deterministic processes guide long-term synchronised population dynamics in replicate anaerobic digesters. *ISME J.* **8**, 2015 (2014).
38. Esteban, D. J., Hysa, B. & Bartow-McKenney, C. Temporal and spatial distribution of the microbial community of Winogradsky columns. *PLoS ONE* **10**, e0134588 (2015).
39. Goyal, A., Dubinkina, V. & Maslov, S. Multiple stable states in microbial communities explained by the stable marriage problem. *ISME J.* **12**, 2823–2834 (2018).
40. Dubinkina, V., Fridman, Y., Pandey, P. P. & Maslov, S. Multistability and regime shifts in microbial communities explained by competition for essential nutrients. *eLife* **8**, e49720 (2019).
41. Darcy, J. L. et al. A phylogenetic model for the recruitment of species into microbial communities and application to studies of the human microbiome. *ISME J.* **14**, 1359–1368 (2020).
42. O'Dwyer, J. P., Kembel, S. W. & Green, J. L. Phylogenetic diversity theory sheds light on the structure of microbial communities. *PLoS Comput. Biol.* **8**, e1002832 (2012).
43. Louca, S. et al. Function and functional redundancy in microbial systems. *Nat. Ecol. Evol.* **2**, 936–943 (2018).
44. Shan, X., Goyal, A., Gregor, R. & Cordero, O. X. Annotation-free discovery of functional groups in microbial communities. *Nat. Ecol. Evol.* **7**, 716–724 (2023).
45. Moran, M. A. et al. Microbial metabolites in the marine carbon cycle. *Nat. Microbiol.* **7**, 508–523 (2022).
46. Lyons, D. M., Zou, Z., Xu, H. & Zhang, J. Idiosyncratic epistasis creates universals in mutational effects and evolutionary trajectories. *Nat. Ecol. Evol.* **4**, 1685–1693 (2020).
47. Reddy, G. & Desai, M. M. Global epistasis emerges from a generic model of a complex trait. *eLife* **10**, e64740 (2021).
48. Diaz-Colunga, J. et al. Global epistasis on fitness landscapes. *Philos. Trans. R. Soc. B* **378**, 20220053 (2023).
49. Harcombe, W. et al. Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Rep.* **7**, 1104–1115 (2014).
50. Diaz-Colunga, J. et al. Top-down and bottom-up cohesiveness in microbial community coalescence. *Proc. Natl Acad. Sci. U.S.A.* **119**, e2111261119 (2022).
51. Barbier, M., Arnoldi, J. F., Bunin, G. & Loreau, M. Generic assembly patterns in complex ecological communities. *Proc. Natl Acad. Sci. U.S.A.* **115**, 2156–2161 (2018).

Acknowledgements
We thank S. Allesina and the members of the Center for the Physics of Evolving Systems at the University of Chicago for useful discussions. We thank J. Softcheck for assistance with the experiments. S.K., K.G., M.Y., and M.T. acknowledge funding from the National Science Foundation (EF-2025293, MCB-2117477, PHY-2310746). S.K. acknowledges funding from the National Institutes of Health (NIH R01GM151538). A. Sanchez acknowledges support from the Spanish Ministry of Science and Innovation under project PID2021-125478NA-100.

Author contributions
S.K., M.T., K.G., M.Y., A. Skwara, A. Sanchez and A.S.R. conceptualized the study. A. Skwara, K.G., M.Y., M.T. and S.K. developed the methodology. A. Skwara, K.G., M.Y., M.T. and S.K. designed and conducted formal analysis. J.D.-C. and A. Sanchez designed experiments and collected experimental data, and J.D.-C. assisted with data analysis and visualization. K.G., S.K., M.T. and A. Skwara wrote the paper. S.K. and M.T. supervised the project and acquired funding.

Competing interests
The authors declare no competing interests.

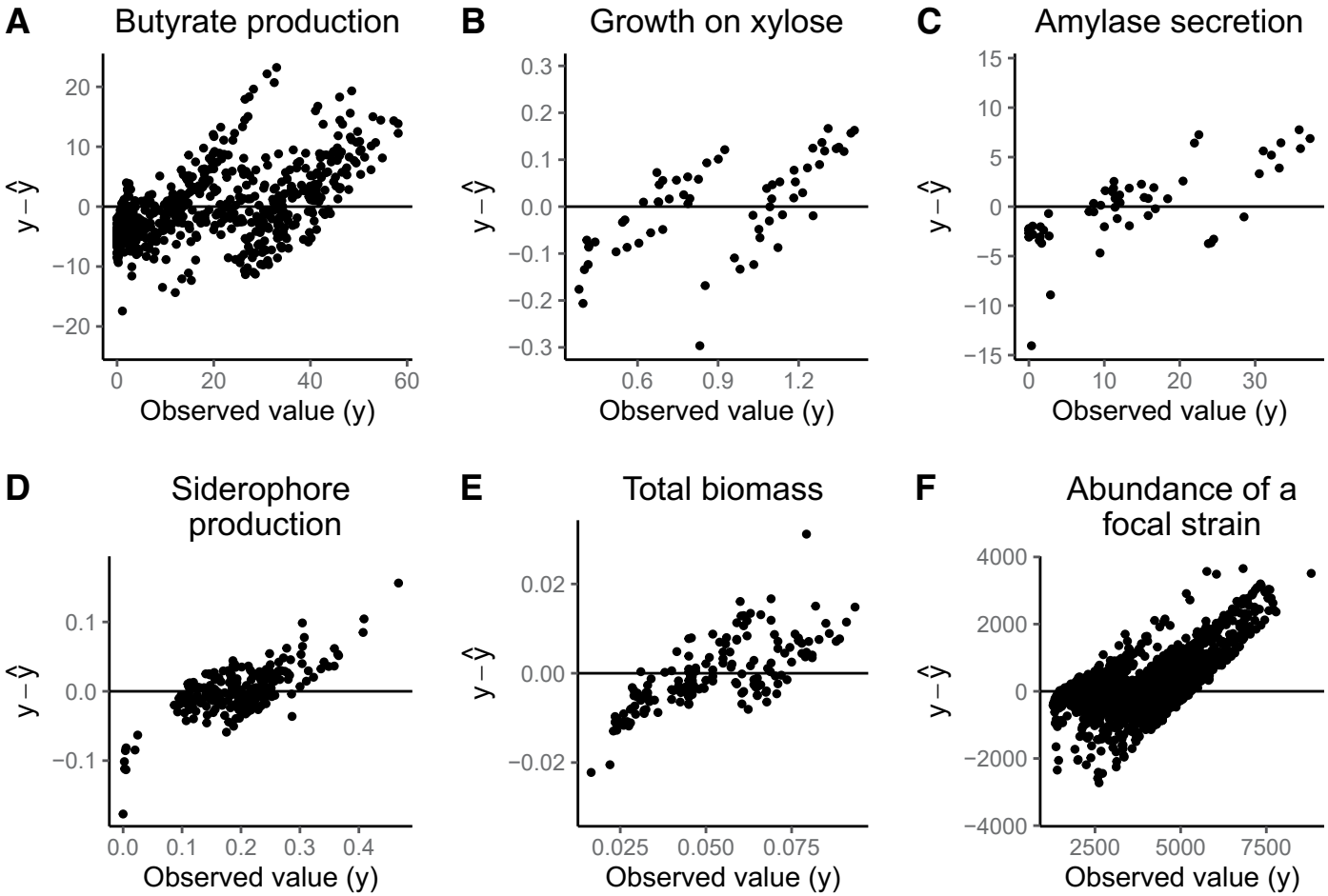
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41559-023-02197-4.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-023-02197-4.

Correspondence and requests for materials should be addressed to Mikhail Tikhonov or Seppe Kuehn.

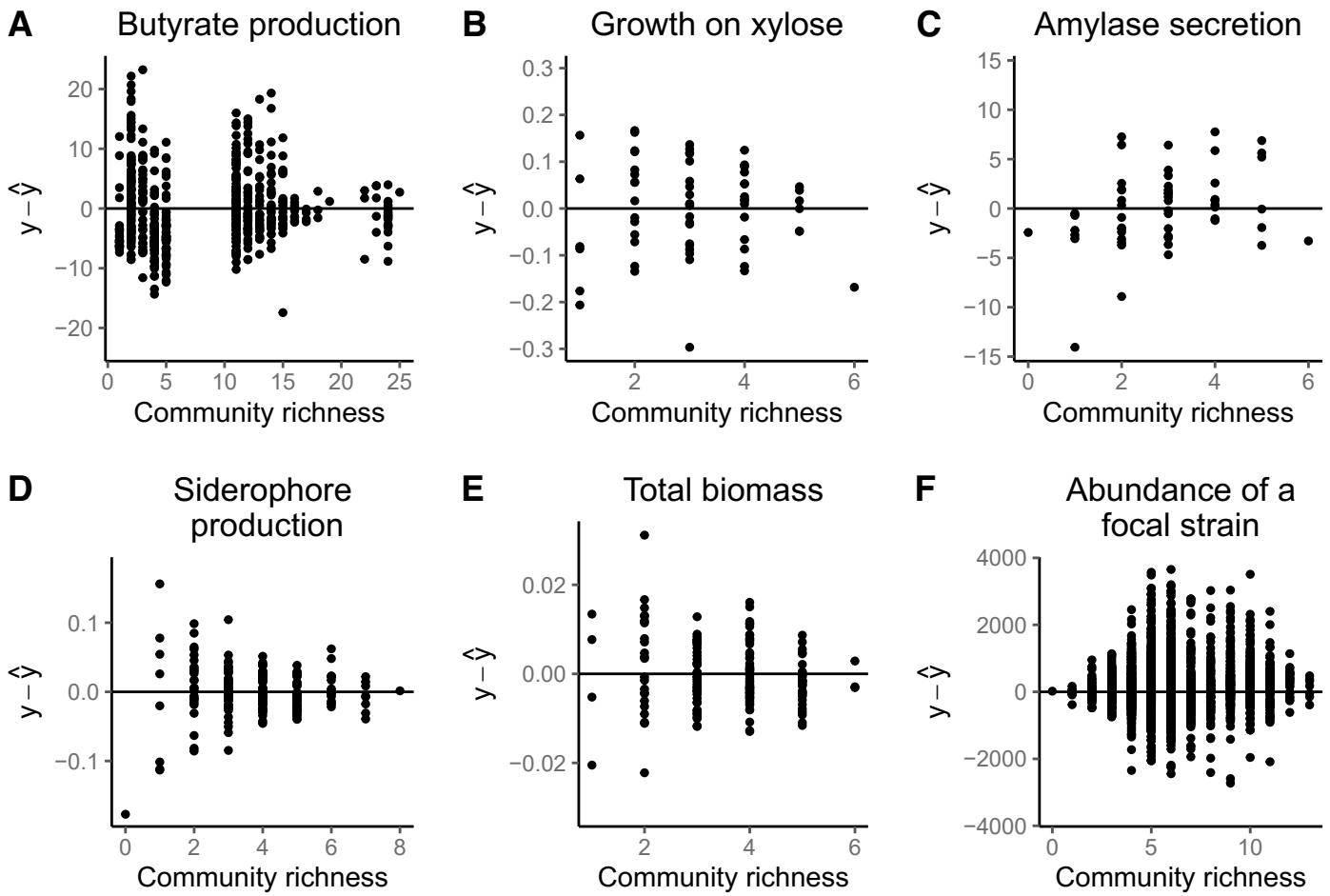
Peer review information *Nature Ecology & Evolution* thanks Elle Barnes, Daniel Amor and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41559-023-02197-4



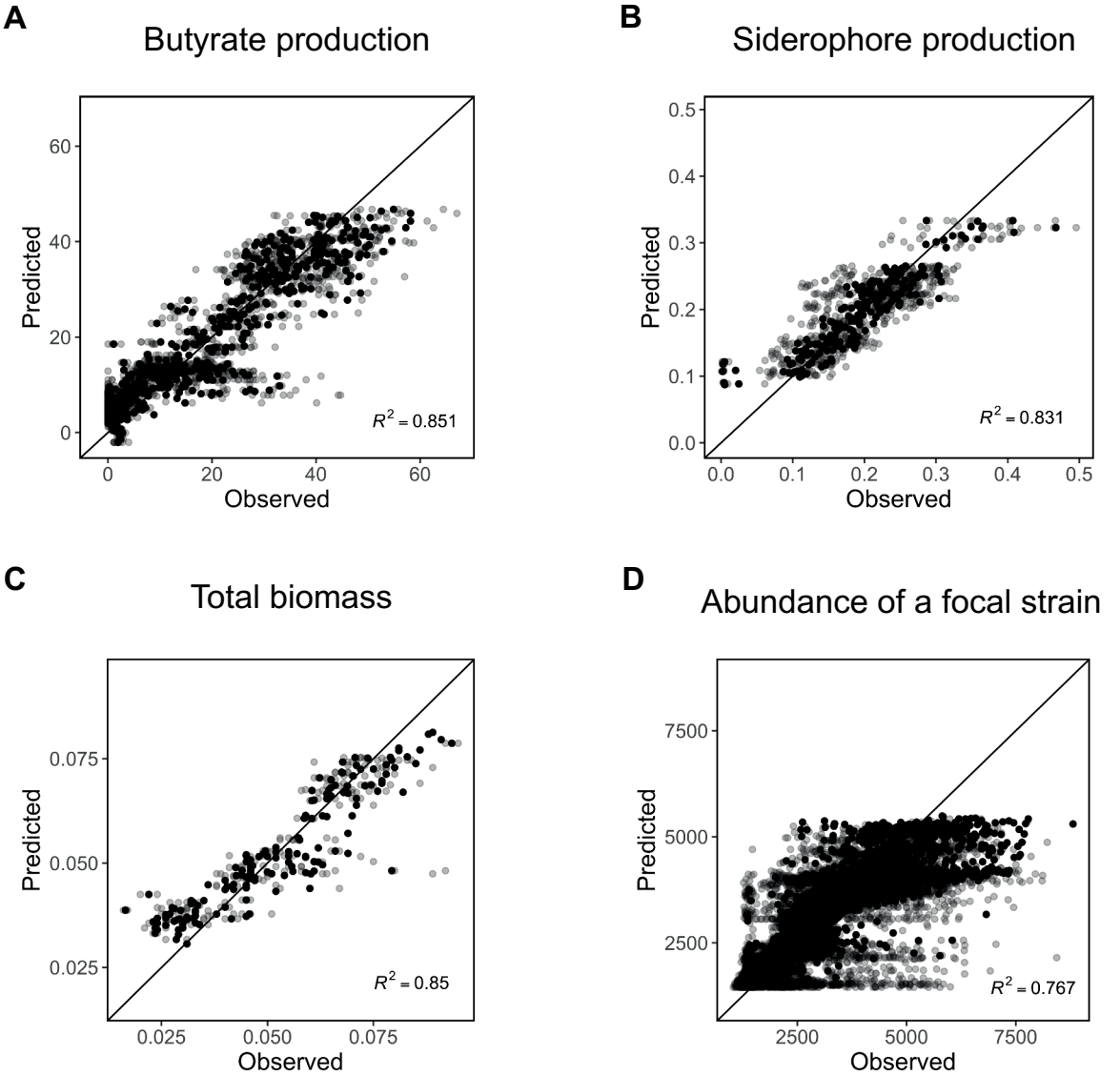
Extended Data Fig. 1 | Residuals obtained from second-order regressions. Data from (A) Clark et al.[22], (B) Langenheder et al.[27], (C) Sanchez-Gorostiaga et al.[21], (D) Diaz-Colunga et al.[28], (E) data from the Sanchez lab, and (F) Kehe et al.[29].

https://doi.org/10.1038/s41559-023-02197-4



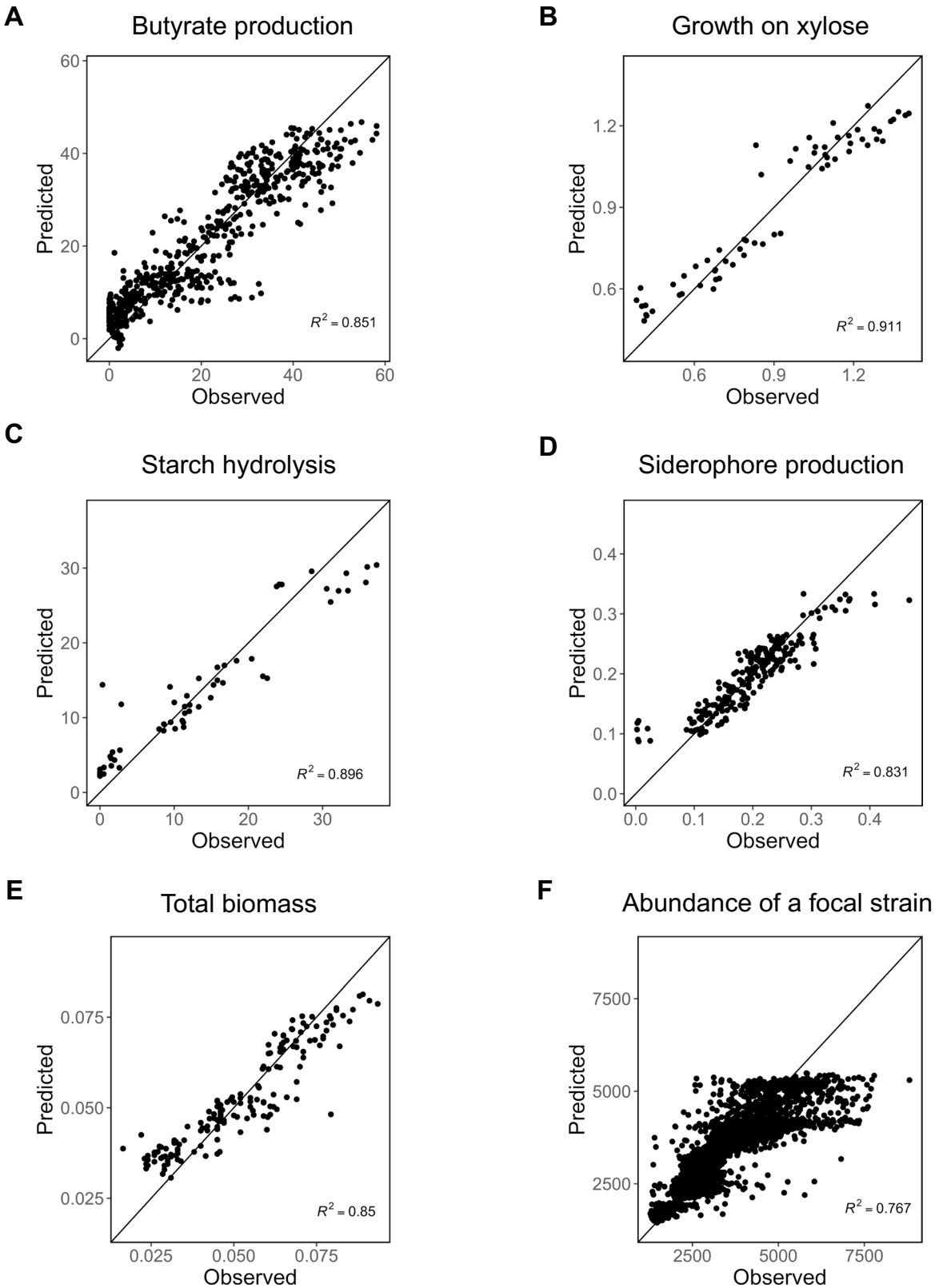
Extended Data Fig. 2 | Residuals obtained from second-order regressions as a function of community richness. Data from (A) Clark et al.[22], (B) Langenheder et al.[27], (C) Sanchez-Gorostiaga et al.[21], (D) Diaz-Colunga et al.[28], (E) data from the Sanchez lab, and (F) Kehe et al.[29].

https://doi.org/10.1038/s41559-023-02197-4



Extended Data Fig. 3 | Model predictions from second-order regression fits with all data in-sample and experimental replicates included. Data from (A) Clark et al.[22], (B) Diaz-Colunga et al.[28], (C) data from the Sanchez lab, and (D) Kehe et al.[29]. Black points correspond to the mean observed value for each distinct experimental community, and are identical to the points shown in Extended Data Fig. S1. Gray points correspond to values for distinct experimental replicates. The datasets here include only datasets for which experimental replicates were measured.

https://doi.org/10.1038/s41559-023-02197-4



Extended Data Fig. 4 | Model predictions from second-order regression fits with all data in-sample. Data from (A) Clark et al.[22], (B) Langenheder et al.[27], (C) Sanchez-Gorostiaga et al.[21], (D) Diaz-Colunga et al.[28], (E) data from the Sanchez lab, and (F) Kehe et al.[29].

nature portfolio

Corresponding author(s): Mikhail Tikhonov, Seppe Kuehn

Last updated by author(s): Jul 6, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For the dataset collected for this study, optical density measurements with an AccuSkan FC plate reader (Fisher Scientific) were made with the standard SkanIt software v5.0. All simulation data was generated using MATLAB R2022b, and all associated scripts are available at https://github.com/abbyskwara2/regression_on_landscapes.
Data analysis	All dataset analyses were performed using R v4.2.1, R package tidyverse v2.0.0, and R package glmnet v4.1-7. All code and scripts necessary to reproduce these analyses are available at https://github.com/abbyskwara2/regression_on_landscapes.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data analyzed in this study is either available from the original studies or in the following repository: https://github.com/abbyskwara2/regression_on_landscapes.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	na
Reporting on race, ethnicity, or other socially relevant groupings	na
Population characteristics	na
Recruitment	na
Ethics oversight	na na

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This is a method development study which made use of 6 datasets of combinatorial microbial communities, as well as simulated datasets of combinatorial microbial communities.
Research sample	These datasets considered here were comprised of microbial communities for which a scalar-valued community function was measured for each distinct experimental community. These datasets included one dataset which was collected for this study, and 5 existing datasets from Langenheder et al. 2010, Clark et al. 2021, Sanchez-Gorostiaga et al. 2019, Diaz-Colunga et al. 2023,and Kehe et al. 2019. This study additionally generated simulated datasets of combinatorial microbial communities.
Sampling strategy	No sample size calculations were performed for the dataset collected in this study. For this dataset, 160 distinct experimental communities were generated, a comparable number to similar studies.
Data collection	For the dataset collected in this study, all experiments were performed in the Sanchez lab by Juan Diaz-Colunga.
Timing and spatial scale	For the dataset collected in this study, all experiments were performed in January 2022.
Data exclusions	No data were excluded from the analyses.
Reproducibility	For the dataset collected in this study, the majority of experimental communities contained biological replicates. All replicates were included in the analysis, and all replicates are plotted in Supplementary Figure 2C.
Randomization	The experimental communities for the dataset collected in this study were not randomized, as the experimental setup was designed to ensure that many different community compositions with a range of community richness would be tested.
Blinding	For the dataset collected in this study, blinding during data collection was not possible as the researcher performing the experiments was required to know the identities of the microbes added to each experimental community.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging