Decoding the emergent properties of microbial community functions through sub-community observations and interpretable machine learning

Hidehiro Ishizawa[1*], Sunao Noguchi[1], Miku Kito[1], Yui Nomura[2], Kodai Kimura[2], Masahiro Takeo[1]

Short title: Dissecting emergent microbiome functions

[1]Department of Applied Chemistry, Graduate School of Engineering, University of Hyogo, Himeji, Hyogo, 671-2280, Japan
[2]Department of Applied Chemistry and Chemical Engineering, School of Engineering, University of Hyogo, Himeji, Hyogo, 671-2280, Japan

*Corresponding author
Hidehiro Ishizawa
Address: Shosha 2167, Himeji, Hyogo, 671-2280 Japan
Phone: (+81)79-267-4969
Mail to: ishizawahidehiro@gmail.com

ABSTRACT

The functions of microbial communities, including substrate conversion and pathogen suppression, arise not as a simple sum of individual species' capabilities but through complex interspecies interactions. Understanding how such functions arise from individual species and their interactions remains a major challenge, limiting efforts to rationally design and control microbial communities. Because current holistic (meta-omics) and reductionist (isolation- or single-cell-based) approaches struggle to capture these emergent microbial community functions, this study explores an intermediate strategy: analyzing simple sub-community combinations to enable a bottom-up understanding of community-level functions. To examine the validity of this approach, we used a nine-member synthetic microbial community capable of degrading the environmental pollutant aniline, and systematically generated a dataset of 256 sub-community combinations and their associated functions. Analyses using random forest models revealed that the sub-community combinations of just three to four species enabled the quantitative prediction of functions in larger communities (5–9 member; Pearson's r = 0.780–0.801). Prediction performance remained robust even with limited sub-community data, suggesting applicability to more diverse microbial communities where exhaustive sub-community observation is infeasible. Moreover, interpreting models trained on these simple sub-community combinations enabled the identification of key species and interspecies interactions that strongly influence the overall community function. These findings provide a methodological framework for mechanistically dissecting complex microbial community functions through sub-community-based analysis.

INTRODUCTION

Microbial communities constitute a significant proportion of Earth's total biomass (approximately 90 Gt-C) (Bar-On et al., 2018), encompassing diverse, interacting species. Far from being mere background organisms, they are key drivers of climate regulation, biogeochemical cycling, and phenotypic expression in multicellular hosts (Bardgett et al., 2008; Nicholson et al., 2012; Hassani et al., 2018). Their diverse metabolic capabilities are increasingly being applied in agriculture, bioproduction, and environmental remediation (McCarty and Ledesma-Amaro, 2019; Compant et al., 2025). However, despite decades of research, our ability to rationally manipulate microbial community functions remains limited.

This limitation largely stems from the emergent nature of microbial community functions, which are shaped by complex interspecies interactions rather than the additive effects of individual species' capabilities (Diaz-Colunga et al., 2024). For instance, in pollutant-degrading microbial communities, species lacking genetic capacity for direct degradation still influence overall system performance by supporting downstream metabolism, modulating population dynamics, or providing essential cofactors (Hu et al., 2020; D'souza et al., 2023; Peng et al., 2025). However, existing analytical frameworks often struggle to capture such indirect contributions via interspecies interactions. Holistic meta-omics approaches provide comprehensive data but cannot easily resolve causal relationships between specific species and functional outcomes (Noecker et al., 2019; Pinto et al, 2022). Conversely, reductionist approaches using isolates or single-cell data often miss the interspecies interactions that drive emergent properties. Consequently, understanding interspecies interactions and the true functional contributions of individual species remains challenging, hindering our ability to predict, design, and control microbial community functions.

A promising approach to address this challenge is to focus on sub-community combinations as an intermediate unit of observation and analysis. This approach involves sampling multiple combinations of a few species, typically three to five, from a larger microbial community and evaluating their functional outputs. Unlike statistical inference from whole-community data, this approach facilitates the examination of species' functional roles in simpler and more interpretable contexts, while still capturing emergent

properties not detectable in isolates or single-cell assays. The conceptual basis for this approach derives from the theoretical food web ecology, where ecosystem-level properties have been modeled by decomposing complex systems into small ecological modules of three to four species (Holt, 1977; Kondoh et al., 2008; Stouffer and Bascompte, 2010). Empirical studies have also shown that interactions among as few as three species can recapitulate emergent behaviors relevant to population dynamics and community assembly in both plant and microbial systems (Mayfield and Stouffer, 2017; Friedman et al., 2017; Ishizawa et al., 2024).

However, existing studies have primarily focused on predicting species abundance or community composition, leaving open the question of how sub-community data can be leveraged to dissect functional outputs, such as substrate conversion or host phenotypic modulation. To address this gap, this study aimed to establish an analytical framework that uses sub-community information to achieve bottom-up, mechanistic understanding of microbial community functions. Specifically, a machine learning approach was applied to explain community-level functional performance based on measurements from sub-community combinations. While previous studies have trained machine learning models on whole community data to predict functional outputs (Thompson et al., 2019; Dutta et al., 2022), we hypothesized that models trained on sub-community data would further highlight influential species and interspecies interactions, providing rational targets for microbial community engineering.

To evaluate the proposed framework in a controlled setting, we constructed a nine-member synthetic microbial community capable of degrading the environmental pollutant aniline as a model system. Aniline, a representative aromatic amine widely used to produce pesticides and dyes, is considered a high-priority bioremediation target due to its toxicity and environmental persistence (Kulkarni and Chaudhari, 2007). Using this system, we systematically assessed all possible sub-community combinations to test two hypotheses: (1) functional outputs of more complex communities can be accurately predicted from simpler sub-community data, and (2) sub-community observation allows identification of key species and interspecies interactions driving emergent community functions. Although derived from a controlled synthetic system, our findings provide a methodological blueprint for leveraging sub-community data to gain mechanistic insight

106 into the functions of more complex natural microbial communities.

107

108 **MATERIALS AND METHODS**

109 **Establishment of a reproducible aniline-degrading community**

110 To establish a model aniline-degrading microbial community, 2.5 g of soil collected from
111 a household vegetable garden in Kakogawa City, Hyogo, Japan (34.77 N 134.85 E), was
112 suspended in 30 mL of MSM medium (Coleman et al., 2002), mixed vigorously, and
113 centrifuged (1,000 × *g*, 1 min). Then, 5 mL of the supernatant containing microbial cells
114 was transferred to 50 mL of MSM medium supplemented with 0.5 mM aniline (hereafter
115 referred to as MSM aniline medium) and incubated at 30 °C with shaking at 150 rpm for
116 48 h. Subsequently, 1 mL of this culture was inoculated into 50 mL of fresh MSM aniline
117 medium and incubated for four days. During cultivation, 1 mL samples were periodically
118 collected to measure residual aniline concentrations using HPLC (Table S1). The
119 resulting enriched culture was preserved at −80 °C in cryotubes containing 16.7%
120 glycerol.

121 For community restoration, 1.2 mL of thawed glycerol stock was added to 10
122 mL of MSM medium and microbial cells were collected via centrifugation (10,000 × *g*,
123 4 °C, 10 min), washed twice, and resuspended in 5 mL of MSM. Then, 1 mL of the cell
124 suspension was inoculated into 50 mL of MSM aniline medium and incubated at 30 °C
125 with shaking at 150 rpm for 48 h. The resulting culture was used as model aniline-
126 degrading community for subsequent analyses. To assess the community structure and
127 aniline degradation ability, 1 mL of the model community was inoculated into 50 mL of
128 MSM aniline medium and incubated in four replicate flasks at 30 °C with shaking at 150
129 rpm for 48 h. Reproducibility was assessed by performing the restoration procedure twice
130 on separate days, followed by the evaluation of residual aniline concentration using
131 HPLC and community structure using 16S rDNA amplicon sequencing.

132 For amplicon sequencing, the 16S rRNA gene library (V4 region; 515F–806R)
133 was prepared using a two-step PCR procedure and sequenced on the Illumina Miseq
134 platform (2×300 bp) at Bioengineering Lab. Co. Ltd. (Sagamihara, Japan). Sequencing
135 reads were processed using DADA2 (Callahan et al., 2016) to generate an ASV table, and
136 taxonomy was assigned using the SILVA v138 database (Quast et al., 2013).

137

138 **Metagenome analysis**

139 DNA was extracted from the restored model aniline-degrading community using
140 NucleoSpin Microbial DNA (Macherey-Nagel, Düren, Germany). For shotgun
141 metagenome sequencing, libraries were prepared with the MGIEasyFS DNA Library Prep
142 Set (MGI Tech, Shenzhen, China), and 150 bp paired-end sequencing was performed on
143 a DNBSEQ-G400RS platform (MGI Tech) at Genome-Lead Inc. (Takamatsu,
144 Japan). Sequencing reads were quality-filtered and adapter-trimmed with fastp v0.23.2
145 and assemblies were generated using metaSPAdes v3.15.5 (Nurk et al, 2017).
146 Subsequently, reads were mapped to assembled scaffolds using BWA v0.7.17 (Li and
147 Durbin, 2009), and metagenome binning was performed using MetaBATt2 v2.15 (Kang
148 et al., 2019). The resulting bins were evaluated using checkM v1.2.2 (Parks et al., 2015),
149 and those with completeness > 50% and contamination < 5% were regarded as valid
150 metagenome-assembled genomes (MAGs). MAGs were annotated using DFAST v1.6.0
151 (Tanizawa et al., 2017) and BlastKOALA v3.0 (Kanehisa et al., 2016). The relative
152 abundances of the MAGs were estimated by mapping the metagenome reads using
153 BBMap v36.01.1. Homology search and gene feature visualizations were performed
154 using the BLAST and CAGECAT (van de Belt, 2023).

155

156 **Isolation, identification, and characterization of synthetic community members**

157 The restored model aniline-degrading community was diluted, spread on R2A and PCAT
158 agar (Draghi et al., 2014), and incubated at 30 °C for up to one week. Morphologically
159 distinct colonies were isolated, purified, and subjected to PCR amplification of the 16S
160 rDNA using the 27F–1392R primer set (Amann et al., 1995). Amplicons were purified
161 using the NucleoSpin Gel and PCR Purification Kit (Macherey-Nagel) and sequenced via
162 Sanger sequencing on a 3730xl DNA analyzer (Applied Biosystems, Foster City, CA,
163 USA) at Macrogen Japan Corp. (Tokyo, Japan). The taxonomic positions of the isolates
164 were identified using similarity searches against the EzBioCloud 16S database. Based on
165 these results, nine bacterial strains were selected as members of the synthetic community:
166 *Achromobacter veterisilvae* MA031 (Ach031), *Pseudomonas monteilii* MA033 (Pse033),
167 *Sphingomonas pseudosanguinis* MA034 (Sph034), *Microbacterium laevaniformans*

168 MA039 (Mic039), *Bosea spartocytisi* (Bos046), *Bacillus mexicanus* MA050 (Bac050),
169 *Pseudomonas* sp. MA056 (Pse056), *Caballeronia cordobensis* MA074 (Cab074), and
170 *Sphingomonas* sp. MA075 (Sph075) (Table S2). Given the absence of rRNA gene
171 sequences in our MAG sequences, we predicted the closest MAG–isolate correspondence
172 by comparing species-level classifications obtained independently for MAGs and isolates.
173 The nine strains were cultivated overnight in 2 mL R2A medium in a 24-well plate (30 °C,
174 150 rpm), washed twice with MSM medium, and used for subsequent experiments.
175 For one isolate (Cab074), identified as the primary aniline degrader, a
176 fluorescence marker (eGFP) was introduced using the pMRE152 plasmid as described by
177 Schlechter et al. (2018). To achieve this, filter mating was performed by placing a 0.45
178 μm filter on solid R2A medium and co-culturing overnight cultures of the wildtype
179 Cab074 and *E. coli* S17λpir carrying pMRE152 at 30 °C for 24 h. After incubation, the
180 cells were collected and plated onto solid MSM aniline medium containing 50 μg/mL
181 kanamycin. The obtained fluorescent transformants (hereafter Cab074F) were purified
182 and used as the member of synthetic communities in place of the wildtype.
183 To evaluate the ability of each synthetic community member to utilize aniline,
184 individual strains were inoculated into 500 μL MSM aniline medium in a 24-well plate at
185 an initial OD_{600} of 0.01. After 48 h of incubation, the residual aniline content in the
186 medium was quantified using HPLC.
187
188 **Full factorial evaluation of the sub-community combinations**
189 A 384-well plate-based assay was conducted to evaluate all possible sub-community
190 combinations of the synthetic community that included Cab074F (i.e., $2^8 = 256$ patterns;
191 $n = 4$). Each well of the black 384-well plate contained 20 μL of MSM aniline medium
192 supplemented with 2 μg/mL of resazurin, a microbial metabolic marker reduced to
193 fluorescent resorufin in proportion to microbial aerobic metabolism (Sarker et al., 2007).
194 Because aniline was the sole carbon and energy source, fluorescence intensity was
195 considered a proxy for aniline utilization. Strains were inoculated into wells at an initial
196 OD_{600} of 0.001 each. To minimize pipetting errors, a robotic pipettor (Assist Plus, Integra
197 Biosciences, Zizers, Switzerland) was used for culture preparation. The plates were
198 incubated in a microplate reader (Synergy HTX, Agilent Technologies, Santa Clara, CA,

199 USA) at 30 °C with orbital shaking at 237 rpm. During the 72 h cultivation period, GFP
200 (excitation/emission: 485/528 nm) and resorufin (528/590 nm) fluorescence were
201 measured every 10 min. Three plates were analyzed to cover all sub-community
202 combinations in four replicates each. Every plate included at least 16 Cab074F-only and
203 16 blank wells without microbial inoculation.

204 The "aniline-utilization activity" and the "Cab074 growth" in each well were
205 defined as the area under the curve (AUC) calculated from the fluorescence time courses
206 of resorufin and GFP during cultivations, respectively. AUC was calculated using the R
207 package "gcplyr" (Blazanin et al., 2024) as follows: (1) baseline correction by subtracting
208 the mean relative fluorescence units (RFU) values of blank wells from all wells on the
209 same plate; (2) alignment of time windows across plates by shifting to match the lag phase
210 of the Cab074F-only wells, as estimated using gcplyr; (3) normalization of the RFU
211 values by setting adjusted time 0 to 1 for each well, correcting subsequent values below
212 1 to 1; and (4) normalization of the AUC values by dividing the mean AUC of the
213 Cab074F-only wells.

214 Based on the obtained data, the functional effect of each non-degrader strain was
215 calculated as the change in aniline-utilization activity between sub-communities that
216 differed only by the presence or absence of the focal strain. The Shapley value,
217 representing the overall contribution of each species to community function, was
218 calculated as the average functional effect across all possible assembly pathways (Shapley,
219 1953). Additionally, the functional interactions between all pairs of non-degrader strains
220 were estimated as the difference in the mean functional effect of a focal strain in the
221 presence versus absence of a partner strain.

222

223 **Validation of the fluorescence-based measurements**

224 To validate that resorufin fluorescence reflected aniline-utilization capacity, cultivation
225 assays for 18 randomly selected combinations were performed using the same procedure
226 scaled up to 300 μL in a 96-well plate to obtain sufficient samples for HPLC analysis. For
227 each combination, cultures were prepared in duplicate wells, and samples were collected
228 after incubating for 48 and 72 h to evaluate aniline degradation via HPLC. These results
229 were compared with the RFU values of resorufin measured at the corresponding time

230 points. Similarly, the correlation between GFP fluorescence and Cab074F abundance was
231 validated by plating 13 Cab074F cultures in a 96-well plate after 24, 48, and 72 h on solid
232 MSM aniline medium and comparing Cab074F abundance in CFU/mL with the
233 corresponding GFP RFU values.
234
235 **Random forest model**
236 Random forest models were employed to examine the bottom-up predictability of aniline-
237 utilization activity. Models were constructed using the "randomForest" R package, with
238 the presence–absence data of eight non-degrader strains as explanatory variables and the
239 measured aniline-utilization activity as the response variable. Training data comprised
240 communities with specific species richness levels (i.e., two to seven strains) using the
241 following optimized parameters: *mtry* = 2, number of trees = 500, and minimum node
242 size = 5. The trained models were then used to predict the aniline-utilization activity in
243 communities containing five to nine members, and Pearson's correlation coefficient
244 between predicted and observed values was calculated to evaluate model performance.
245 When communities with 5–7 members were used as training data, overlapping
246 combinations were excluded from the test set. To assess model performance under limited
247 training data, additional analyses were conducted using randomly sampled subsets
248 comprising 50% or 25% of the original training data. Each model was generated 100
249 times, and the mean and standard deviation of prediction accuracy were calculated to
250 account for training variability. Figures present the model with median performance (51st
251 ranked prediction accuracy among the 100 trials).
252 To identify important species and interspecies interactions contributing to
253 prediction accuracy, permutation importance and Friedman's H statistic were extracted
254 from the random forest models using the "vivid" R package (Inglis et al., 2022).
255
256 **Data availability**
257 The 16S rDNA sequences of the bacterial isolates were deposited in the
258 DDBJ/EMBL/GenBank database under accession numbers LC878797-LC878805.
259 Metagenome sequence reads of the aniline-degrading community were deposited under
260 the accession number DRR707044. All data and analytical codes have been deposited in

261 Zenodo (https://doi.org/10.5281/zenodo.16936110).
262
263 **RESULTS**
264 **Reproducibility of the model aniline-degrading microbial community**
265 A major challenge in systematic analysis of microbial communities is the lack of
266 reproducibility, i.e., the ability to regenerate the same community across experiments. To
267 address this, we first established a reproducible community using a whole-community
268 glycerol stock derived from an aniline-degrading community. Restoring the preserved
269 stock through a standardized procedure confirmed consistent community degradation of
270 0.5 mM aniline within 48 h and a stable community structure dominated by specific
271 genera including *Achromobacter*, *Burkholderia*, and *Pseudomonas*, with only minor
272 variations in relative abundances (Fig. S1). Importantly, the dominant bacteria in the
273 restored communities were consistent at the ASV level, suggesting that the preserved
274 community stock provides a reliable source of reproducible model aniline-degrading
275 communities for consistent investigations.
276
277 **Sequence-based prediction of aniline- and catechol-degrading bacteria**
278 Shotgun metagenome analysis was performed to identify species in the aniline-degrading
279 community harboring genes responsible for aniline degradation. The assembly of
280 approximately 107 million reads enabled the reconstruction of 13 MAGs, encompassing
281 all dominant taxa detected via the 16S rDNA amplicon sequencing (Fig. 1A; Table S3).
282 Reportedly, aniline degradation is initiated by a unique reaction in which a
283 glutamate synthase-like protein (AtdA1) attaches a glutamate molecule to aniline,
284 forming γ-glutamylanilide. Subsequently, dioxygenase proteins (AtdA3–5) oxidize the
285 aromatic ring, yielding catechol, glutamate, and ammonium (Takeo et al., 1998, 2013)
286 (Fig. 1B). Genes encoding these enzymes are often clustered with *atdA2*, encoding the
287 reverse reaction of AtdA1. Catechol, a common intermediate in aromatic compound
288 biodegradation is typically introduced into central anabolic metabolism via either the
289 ortho- or meta-cleavage pathways (Barnsley, 1976).
290 Within the obtained MAGs, only MAG01, assigned to the genus *Caballeronia*,
291 contained a gene with significant similarity to known *atdA1* genes (43.6% similarity to

292 *atdA1* of *Acinetobacter* sp. YAA) (Takeo et al., 2013). Adjacent genes encoded
293 components of an aromatic ring-hydroxylating dioxygenase (a functional analog of
294 *atdA3–5*) and a glutamyl hydrolase (a functional analog of *atdA2*) (Fig. 1C). Although
295 these surrounding genes exhibited <20% similarity to previously reported aniline-
296 degrading gene clusters, we considered this gene cluster responsible for aniline
297 degradation in our community through analogous enzymatic reactions.

298 For catechol degradation, several MAGs dominant in the community—including
299 those affiliated with *Achromobacter*, *Pseudomonas*, and *Pseudorhodopherax*—possessed
300 most genes involved in the ortho-cleavage pathway (Fig. 1D). MAG01, which harbored
301 putative aniline-degrading enzymes, also possessed a complete set of ortho-cleavage
302 pathway genes, suggesting its ability to independently utilize aniline as a sole carbon and
303 energy source. In contrast, no MAG carried the candidate genes for 2-keto-4-pentenoate
304 hydratase (K02554), a key enzyme in the meta-cleavage pathway, indicating that catechol
305 degradation in this community proceeds primarily through the ortho-cleavage pathway.

306 In summary, the metagenomic analysis suggests a potential metabolic relay
307 within the model aniline-degrading community, in which *Caballeronia* first converts
308 aniline into catechol, while several other strains contribute to its subsequent metabolism.
309 Bacteria lacking the ability to degrade neither aniline nor catechol may persist through
310 metabolic cross-feeding or scavenging of cell debris.
311

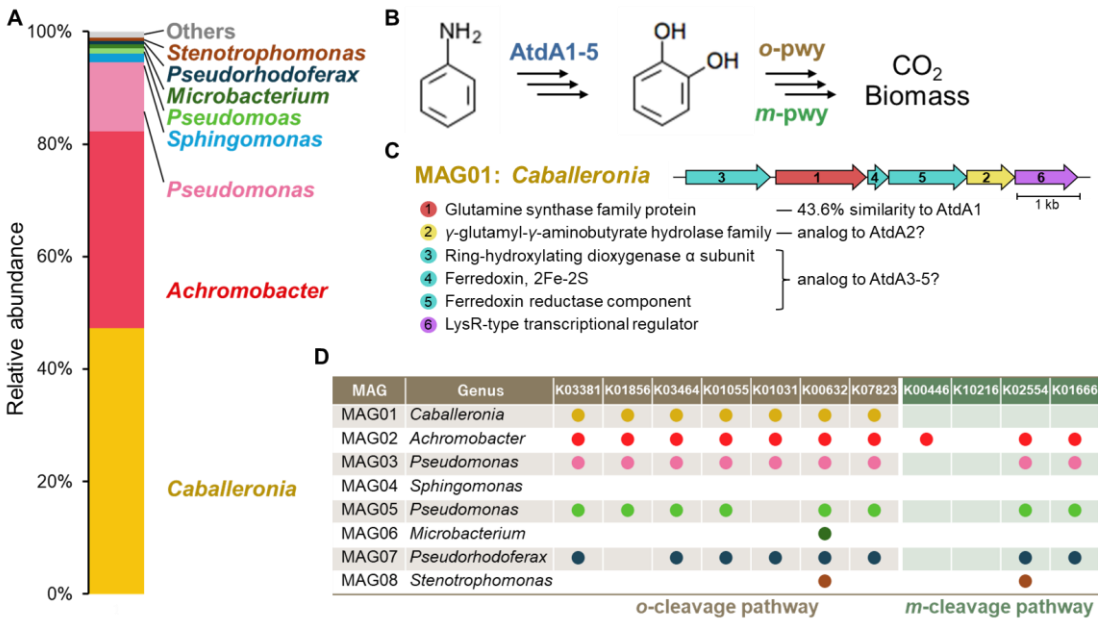


Fig. 1 Metagenomic characterization of the model aniline-degrading community. A) Relative abundance of metagenome-assembled genomes (MAGs). B) Schematic representation of the known aniline degradation pathway. C) The putative aniline oxidation gene cluster identified in MAG01. D) Distribution of catechol-degradation genes among the eight most abundant MAGs in the community.

Characterization of synthetic community members

To construct a tractable synthetic community reflecting the diversity, functionality, and ecological characteristics of the model aniline-degrading community, nine bacterial strains representing a broad range of phylogenetic groups present in the original community were isolated and selected as members of the synthetic community (Table S2). Species-level taxonomic comparisons indicated that Cab074, Ach031, Pse033, Sph075, Pse056, and Mic039 corresponded most closely to the six dominant MAGs (MAG01–MAG06). In contrast, Sph034, Bos046, and Bac050 were considered minor community members without corresponding MAGs. To facilitate detection of the primary aniline degrader, a GFP-labeled strain of Cab074 (Cab074F) was constructed and used in place of the wildtype in subsequent experiments. Degradation assays for individual synthetic community members confirmed that only Cab074F degraded aniline (Fig. S2), consistent with the metagenome characterization of the original community. Overall, the synthetic

community recapitulates key characteristics of the original community: a single *Caballeronia* genotype predominantly performs initial aniline oxidation while other strains would participate in downstream metabolism.

Composition–function landscape

Using the synthetic community, we investigated how community-level functions emerge from interactions between the aniline degrader and non-degrader strains, using a microplate-based assay (Fig. 2A). To increase throughput, we employed a fluorescence-based assay of aniline-utilization activity using resazurin. Preliminary tests showed a positive correlation between RFU and aniline degradation, as evaluated using HPLC (Pearson's $r = 0.774$) (Fig. S3). Although the correlation was moderate, deviations likely arose because resazurin reduction reflects not only aniline degradation but also the metabolism of downstream intermediates. Thus, the resazurin-based assay would capture overall aniline-utilization activity, encompassing both aniline degradation and downstream metabolic processes. Simultaneously, Cab074F abundance was monitored by using GFP fluorescence, which showed strong linear correlation between GFP RFU and MA074F abundance as determined via colony counting (Pearson's $r = 0.892$) (Fig. S3).

By quantifying the aniline-utilization activity of all possible sub-community combinations containing Cab074F (i.e., $2^8 = 256$ patterns), we generated a composition–function landscape in which each data point represents a sub-community combination and each edge denotes the functional effect of a single non-degrader (Fig. 2B). In this study, "functional effect" was defined as the change in aniline-utilization activity between two sub-community combinations differing only in the presence or absence of the focal strain. If the functional effect of each strain were constant regardless of co-existing members, the edge angles would remain consistent. Indeed, Ach031 and Pse033 consistently exhibited negative and positive functional effects, respectively, with few exceptions (Fig. 2C). Sph034 and Mic039 also showed relatively consistent effects, whereas the remaining strains exhibited highly variable effects, ranging from strongly positive to strongly negative depending on the composition of the surrounding community.

To systematically quantify the contribution of each species to community-level

function, we applied Shapley values to the composition–function landscape, calculating the average functional effect of each strain across all possible assembly pathways from the Cab074F-only community to the nine-member community (8! = 40,320 pathways). As shown in Fig. 2D, Pse033 contributed most positively, followed by Bac050 and Sph075, whereas Ach031 and Sph034 contributed negatively.

We then evaluated how the presence or absence of a given non-degrader strain altered the functional effect of another. For example, Sph034 had a near-neutral effect in the absence of Pse033, but became negative when Pse033 was present (Fig. 2E). We quantified such "functional interaction" as changes in the mean functional effect of a focal strain due to a partner strain's presence or absence. As summarized in Fig. 2F, we identified ten significant negative and three positive functional interactions (*t*-test, $p < 0.05$), indicating combinations that lead to either enhanced or diminished community-level functions compared with simple additive assumptions. Notably, despite its high individual contribution to aniline degradation, Pse033 exhibited negative interactions with six other strains, suggesting diminished contributions in co-culture. In contrast, Mic039 showed no significant interactions, indicating that its functional effect was largely independent of other species.

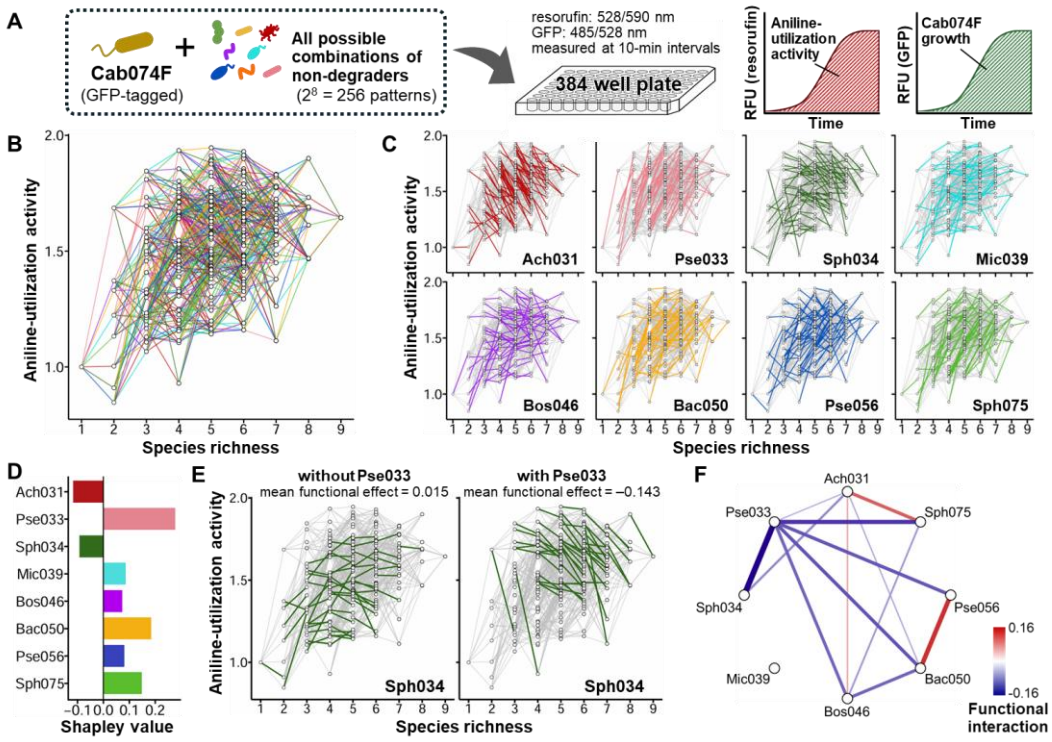
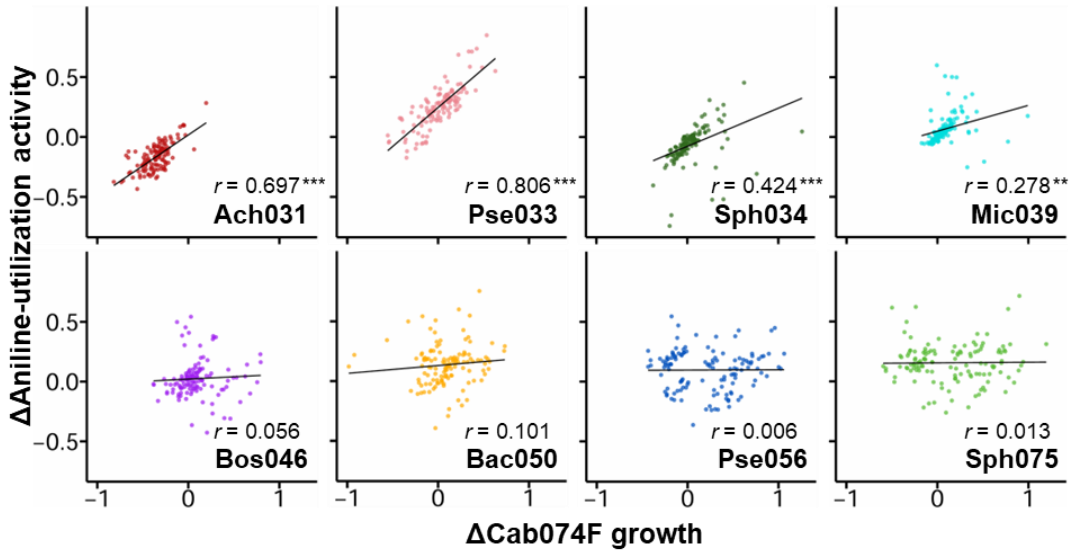


Fig. 2 Full factorial evaluation of the synthetic aniline-degrading community. A) Schematic overview of the experimental workflow. B) Composition–function landscape constructed from the full dataset. Each data point represents a sub-community combination with a given species richness. Edge angles indicate the functional effect of each non-degrader strain in different community contexts (color-coded as in panel C). C) Individual composition–function landscapes for each of the eight non-degrader strains. D) Overall contributions of the eight non-degrader strains to community-level function, quantified as Shapley values. E) Example of a functional interaction: Sph034 exhibited a near-neutral effect in the absence of Pse033 but a negative effect when Pse033 was present. The change in the mean functional effect was defined as the functional interaction between two species. F) Pairwise functional interactions among the eight non-degrader strains. Statistically significant interactions (t-test, $p < 0.05$) are indicated as edges.

Abundance of primary aniline degrader

Given that Cab074F was the sole strain capable converting aniline to catechol, we hypothesized that variations in its growth across sub-community combinations could explain differences in community-level aniline-utilization activity. To test this hypothesis, we generated a composition–abundance landscape for Cab074F using GFP fluorescence data (Fig. S4) and assessed the correlations between the functional effect on aniline-utilization activity and the corresponding effects on Cab074F growth within the same community contexts (Fig. 3). The results varied significantly across non-degrader strains: Ach031, Pse033, Sph034, and Mic039 showed significant positive correlations, whereas the others showed weak or no correlations. These positive correlations suggest that the influence of these non-degrader strains on community-level aniline-utilization activity can be largely explained by their facilitation or inhibition of Cab074F. In contrast, the absence of correlation implies the involvement of additional mechanisms, such as effects on downstream metabolism or interactions with coexisting non-degraders. Interestingly, Ach031 and Pse033, which exhibited notably consistent functional effects regardless of community composition (Fig. 2C), also showed exceptionally strong positive correlations. In contrast, strains that exhibited highly variable functional effects (Bos046, Bac050,

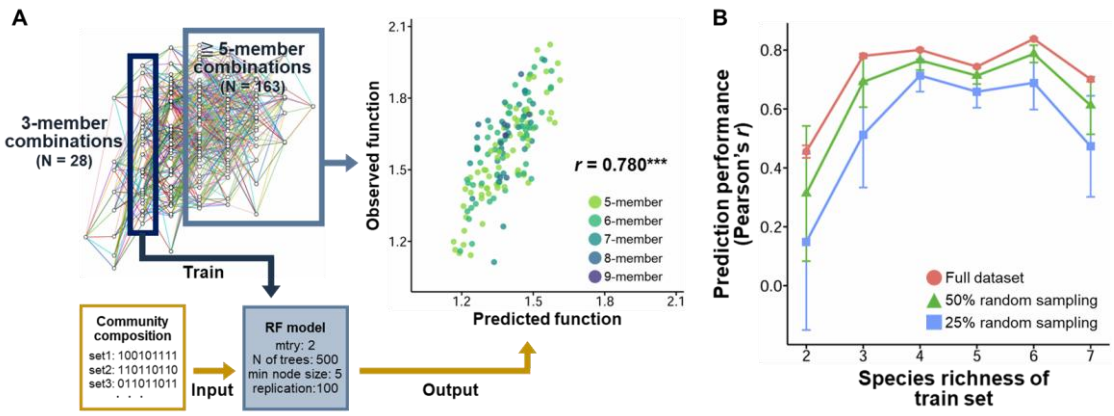
412 Pse056, and Sph075) showed weak or no correlations. These findings suggest that the
413 functional effects mediated through Cab074F abundance tend to be robust to community
414 composition, whereas effects involving other mechanisms are more sensitive to the
415 community context.
416



417 **Fig. 3** Correlation between changes in Cab074F's growth and aniline-utilization activity.
418 Each point represents differences in Cab074F growth (measured using GFP) and aniline-
419 utilization activity (measured using resazurin) between sub-community combinations
420 differing only by the presence or absence of a focal non-degrader strain. Asterisks denote
421 significance levels in Pearson's correlation analysis: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.
422
423
424 **Bottom-up predictability of the composition-abundance landscape**
425 Using the obtained dataset, we tested whether the functions of complex microbial
426 communities can be predicted based on observations of simpler sub-community
427 combinations. To assess how species richness in the training data influences predictive
428 performance, we trained random forest models using sub-communities of defined
429 richness levels and applied them to predict aniline utilization activity in communities
430 containing 5–9 species (Fig. 4A). Notably, models trained solely on three-member
431 combinations (n = 28) achieved high predictive accuracy for more complex communities
432 (r = 0.780) (Fig. 3A). Incorporating more diverse combinations into the train set did not

433 substantially improve the prediction performance (Fig. 3B). However, models trained
434 solely on two-member combinations (i.e., Cab074F plus one other strain) showed
435 markedly lower accuracy (r = 0.487). These results indicate that sub-community
436 combinations of just three to four species can support robust bottom-up prediction of
437 microbial community functions.

438 Because exhaustively sampling all sub-community combinations is infeasible in
439 real microbial communities, we next tested performance with reduced training data (50%
440 and 25%). Reducing the dataset caused only modest decreases in predictive performance.
441 For example, models trained on 25% of the three-member combinations (n = 7) exhibited
442 lower performance (r = 0.512), whereas those trained on 25% of the four-member
443 combinations (N = 14) still achieved high performance (r = 0.714) (Fig. 4B). These
444 findings suggest that random forest models can reliably predict community-level
445 functions without exhaustive data, provided that enough sub-community combinations
446 are included in the train set.



447 **Fig. 4.** Bottom-up prediction of aniline-utilization activity using random forest models.
448 A) Schematic of the bottom-up prediction framework. Random forest models were
449 trained on data from sub-community combinations to predict the aniline-utilization
450 activity of combinations with ≥5 members. Predicted and observed values were compared
451 to evaluate model performance. B) Prediction performance of random forest models
452 trained on sub-community combinations of varying species richness. Models trained on
453 reduced datasets (50% or 25% random sampling) are also shown to assess the impact of
454 data availability on prediction accuracy.
455

Interpretation of the random forest model to identify key species and interactions

Finally, we tested whether interpretating the trained random forest models could reveal key species and interspecies interactions that shape community-level functions. To this end, we evaluated two model-derived metrics: permutation importance, which quantifies each feature's contribution by measuring changes in prediction accuracy after random shuffling of focal features (Altmann et al., 2010), and Friedman's H statistic, which estimates the strength of interactions between features by assessing how the effect of one feature on aniline utilization activity depends on another (Friedman and Popescu, 2008).

We first applied these interpretability techniques to the random forest model trained on three-, four-, and five-member sub-community combinations (Fig. 5A–C). The permutation importance scores consistently identified Ach031 and Pse033 as the most influential species, followed by Sph034, Bac050, and Sph075. This ranking closely matched species contributions estimated from the full dataset using Shapley values (Fig. 2D), indicating that random forest models trained on simple sub-community data can reliably identify the dominant contributors to community-level functions.

Additionaly, Friedman's H statistics derived from models trained on three- and four-member sub-communities correlated significantly with empirical functional interaction strengths shown in Fig. 2E (r = 0.402–0.446, $p < 0.05$), whereas those trained on two-member sub-communities showed no correlation ($r = 0.141$, $p = 0.457$). Notably, the three- and four-member models highlighted different sets of pairwise relationships, and predictive accuracy improved markedly when models were trained on combined dataset of all two- to four-member combinations ($r = 0.712$, $p < 0.0001$) (Fig. 5D). In contrast, Friedman's H statistics derived from models trained on five-member sub-communities were no longer significantly correlated with the observed functional interactions ($r = 0.196$, $p > 0.05$), and a similar loss of correspondence occurred in six-member models ($r = 0.139$, $p > 0.05$). These results suggest that increasing training data complexity reduces the ability to resolve specific interspecies interactions, likely due to statistical difficulties in pinpointing influential feature pairs within high-dimensional observational units.

Together, these findings show that machine-learning models trained on simple

sub-community combinations can identify key species and interspecies interactions underlying emergent community functions. In particular, sub-community combinations of three to four species appear to represent the most informative unit for dissecting the emergent properties of microbial community functions.

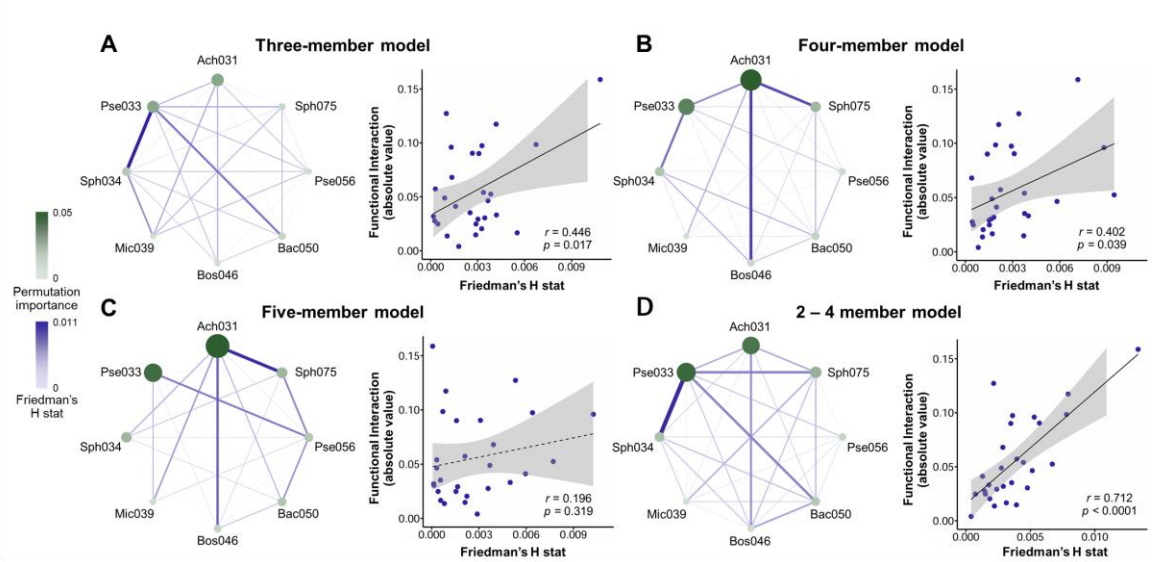


Fig. 5. Key species and interspecies interactions inferred from the random forest models. Permutation importance and Friedman's H statistic were calculated from random forest models trained on sub-community combinations comprising three, four, and five members, as well as from models trained on two- to four-member combinations. Permutation importance was used to quantify the relative contributions of individual species to community-level aniline utilization activity, whereas Friedman's H statistic was used to assess the strength of pairwise functional interactions. To evaluate the reliability of inferred interactions, Friedman's H values were compared with the empirically derived functional interaction strengths shown in Fig. 2.

DISCUSSIONS

Using a synthetic microbial community capable of degrading aniline, this study showed that analyzing simple sub-community combinations enables the identification of key species and interactions shaping community-level functions (Fig. 5). Traditional analytical approaches to microbial community analysis have tended to follow either

508 holistic (meta-omics-based) or reductionist (isolated strain/single-cell-based) strategies.
509 In contrast, this study proposes an intermediate, sub-community-based framework that
510 effectively captures the emergent properties of microbial community functions. Our
511 proposed framework involves three steps: (1) generating multiple sub-community
512 combinations from a target microbial community, (2) evaluating their functions and
513 compositions, and (3) analyzing the resulting dataset to infer the species contributions
514 and interactions, thereby identifying rational targets for microbial community engineering.
515 Recent advances in high-throughput microbial cultivation platforms (Hsu et al., 2019;
516 Jackman et al., 2019) have made it increasingly feasible to generate and characterize
517 thousands of sub-community combinations without isolation. Thus, the findings and
518 analytical framework presented here provide a foundation for applying this approach to
519 more complex natural microbial communities in the near future.

520 Consistent with recent studies investigating composition–function landscapes of
521 synthetic microbial communities (Sanchez-Gorostiaga et al., 2019; Ruiz et al., 2023), our
522 results demonstrate that the functional effects of individual species are highly context-
523 dependent (Fig. 2). To explain such context dependency, Diaz-Colunga et al. (2024)
524 proposed that a species' functional effect can be approximated by simple linear regression
525 against the functional level of the background community—tending to be positive when
526 the baseline function is low, and negative when it is high. This pattern was evident in our
527 dataset (Fig. S5) and appears to reflect both functional redundancy among species and
528 complex interaction networks that obscure individual contributions. Beyond this
529 statistical explanation, our results suggest that mechanistic differences in species' modes
530 of action also contribute: strains that affected aniline degradation through direct
531 promotion of the degrader Cab074F's growth showed relatively stable functional effects,
532 whereas those acting through other mechanisms exhibited greater context sensitivity (Fig.
533 3). To the best of our knowledge, no studies have explicitly integrated mechanistic
534 insights with statistical patterns in composition–function landscapes; doing so may
535 represent a promising direction for future research.

536 Another key finding of this study is that microbial community functions can be
537 predicted in a bottom-up manner by combining simple sub-community observations with
538 machine learning. Similar bottom-up approaches have recently gained traction in studies

using synthetic microbial communities, with several strategies showing promise. For example, genome-scale metabolic models have been applied to predict the metabolic output of complex microbial consortia (Clark et al., 2021; Gowda et al., 2022; Ruan et al., 2024). In parallel, mathematical models that infer community function from the geometry of composition–function landscapes have also been developed (Skwara et al., 2023; Ruiz et al., 2023; Diaz-Colunga et al., 2024), enbaling prediction without detailed knowledge of metabolic pathways. Although many of these modeling studies relied on full datasets in composition–function landscapes and leave-one-out validation strategies (Skwara et al., 2023; Ruiz et al., 2023; Diaz-Colunga et al., 2024), our framework demonstrated that reasonable prediction can be achieved using only a limited set of simple sub-community combinations. This flexibility makes our approach suitable not only for the rational design of synthetic microbial consortia but also for investigating natural communities, where exhaustive sub-community observation is infeasible.

This study also introduced a method for identifying key species and interspecies interactions by interpreting variable-importance measures derived from machine-learning models. Notably, increasing species richness in observed sub-community combinations does not necessarily improve analytical insights (Fig. 5), likely due to statistical challenges in resolving interspecies interactions from more complex combinations. These findings suggest that sub-community combinations consisting of three to four species may represent an analytical sweet spot, balancing predictive power and interpretability. This observation aligns with previous reports where three- to four-member sub-communities provided robust models for predicting species abundance or survival (Friedman et al., 2017; Ishizawa et al., 2024), highlighting their promise as a unit for dissecting emergent properties in microbial communities.

In summary, this study demonstrates that simple sub-community combinations contain valuable information for bottom-up, mechanistic understanding of microbial community functions. The machine learning–based framework proved robust to data sparsity, making it a promising tool for analyzing diverse and noisy microbial systems. While further validation is needed to assess its applicability to highly diverse and dynamic microbial ecosystems, integrating this sub-community-based strategy with existing meta-omics and cultivation-based methods offer a promising path towards predicting and

rationally controlling emergent properties of microbial community functions.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

ACKNOWLEDGEMENTS

This work was supported by JST, ACT-X (JPMJAX22B1), Research Grants by Kurita Water and Environment Foundation (22E053), and Gene-grant 2023 by Nippon Genetics Co., Ltd. to Hidehiro Ishizawa.

AUTHOR CONTRIBUTIONS

Hidehiro Ishizawa: Conceptualization, Formal analysis, Funding acquisition, Project administration, Software, Visualization, Writing – Original Draft. Sunao Noguchi: Formal analysis, Investigation, Visualization, Writing – Review & Editing. Miku Kito: Investigation, Writing – Review & Editing. Yui Nomura: Investigation, Writing – Review & Editing. Kodai Kimura: Investigation, Writing – Review & Editing. Masahiro Takeo: Project Administration, Writing – Review & Editing, Supervision.

References

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010) Permutation importance: a corrected feature importance measure. *Bioinformatics* **26**: 1340–1347.

Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Review* **59**: 143–169.

Bardgett, R.D., Freeman, C., and Ostle, N.J. Microbial contributions to climate change through carbon cycle feedbacks. *The ISME Journal* **2**: 805–814.

Barnsley, E.A. (1976) Role and regulation of the ortho and meta pathways of catechol metabolism in pseudomonads metabolizing naphthalene and salicylate. *Journal of Bacteriology* **125**: 404–408.

Bar-On, Y.M., Phillips, R., and Milo, R. (2018) The biomass distribution on Earth. *Proceedings of the National Academy of Sciences* **115**: 6506–6511.

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

Clark, R.L., Connors, B.M., Stevenson, D.M., Hromada, S.E., Hamilton, J.J., Amador-Noguez, D., and Venturelli, O.S. (2021) Design of synthetic human gut microbiome assembly and butyrate production. *Nature Communications* **12**: 3254.

Coleman, N.V., Mattes, T.E., Gossett, J.M., and Spain, J.C. (2002) Biodegradation of cis-Dichloroethene as the Sole Carbon Source by a β-Proteobacterium. *Applied and Environmental Microbiology* **68**: 2726–2730.

Compant, S., Cassan, F., Kostić, T., Johnson, L., Brader, G., Trognitz, F., and Sessitsch, A. (2025) Harnessing the plant microbiome for sustainable crop production. *Nature Reviews Microbiology* **23**: 9–23.

Diaz-Colunga, J., Skwara, A., Vila, J.C.C., Bajic, D., and Sanchez, A. (2024) Global epistasis and the emergence of function in microbial consortia. *Cell* **187**: 3108–3119.

Draghi, W.O., Peeters, C., Cnockaert, M., Snauwaert, C., Wall, L.G., Zorreguieta, A., and Vandamme, P. (2014) *Burkholderia cordobensis* sp. nov., from agricultural soils. *International Journal of Systematic and Evolutionary Microbiology* **64**: 2003–2008.

D'Souza, G., Schwartzman, J., Keegstra, J., Schreier, J.E., Daniels, M., Cordero, O.X., et al. (2023) Interspecies interactions determine growth dynamics of biopolymer-degrading populations in microbial communities. *Proceedings of the National Academy of Sciences* **120**: e2305198120.

Dutta, A., Goldman, T., Keating, J., Burke, E., Williamson, N., Dirmeier, R., and Bowman, J. S. (2022) Machine learning predictions biogeochemistry from microbial community structure in a complex model system. *Microbiology Spectrum* **10**, e01909-21.

Friedman, J.H., and Popescu, B.E. (2008) Predictive learning via rule ensembles. *The annals of Applied Statistics* **2**, 916–954.

Friedman, J., Higgins, L.M., and Gore, J. (2017) Community structure follows simple assembly rules in microbial microcosms. *Nature Ecology and Evolution* **1**: 0109.

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

Hassani, M.A., Durán, P., and Hacquard, S. (2018) Microbial interactions within the plant holobiont. *Microbiome* **6**: 58.

Holt, R.D. (1977) Predation, apparent competition, and the structure of prey communities. *Theoretical Population Biology* **12**: 197–229.

Hsu, R.H., Clark, R.L., Tan, J.W., Ahn, J.C., Gupta, S., Romero, P.A., and Venturelli, O.S. (2019) Microbial Interaction Network Inference in Microfluidic Droplets. *Cell Systems* **9**: 229–242.

Hu, B., Wang, M., Geng, S., Wen, L., Wu, M., Nie, Y., et al. (2020) Metabolic exchange with non-alkane-consuming *Pseudomonas stutzeri* SLG510A3-8 improves n-alkane biodegradation by the alkane degrader *Dietzia* sp. Strain DQ12-45-1b. *Applied and Environmental Microbiology* **86**: e02931-19.

Inglis, A., Parnell, A., and Hurley, C.B. (2022) Visualizing variable importance and variable interaction effects in machine learning models. *Journal of Computational and Graphical Statistics* **31**: 766–778.

Ishizawa, H., Tashiro, Y., Inoue, D., Ike, M., and Futamata, H. (2024) Learning beyond-pairwise interactions enables the bottom–up prediction of microbial community structure. *Proceedings of the National Academy of Sciences* **121**: e2312396121.

Jackman, C.M., Deans, K.W., Forney, L.J., and Lin, X.N. (2019) Microdroplet co-cultivation and interaction characterization of human vaginal bacteria. *Integrative Biology* **11**: 69–78.

Kanehisa, M., Sato, Y., and Morishima, K. (2016) BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology* **428**: 726–731.

Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**: e7359.

Kondoh, M. (2008) Building trophic modules into a persistent food web. *Proceedings of the National Academy of Sciences* **105**: 16631–16635.

Kulkarni, M. and Chaudhari, A. (2007) Microbial remediation of nitro-aromatic compounds: An overview. *Journal of Environmental Management* **85**: 496–512.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-

Wheeler Transform. *Bioinformatics* **25**: 1754–1760.

Mayfield, M.M. and Stouffer, D.B. (2017) Higher-order interactions capture unexplained complexity in diverse communities. *Nature Ecology and Evolution* **1**: 0062.

McCarty, N.S. and Ledesma-Amaro, R. (2019) Synthetic biology tools to engineer microbial communities for biotechnology. *Trends in Biotechnology* **37**: 181–197.

Nicholson, J.K., Holmes, E., Kinross, J., Burcelin, R., Gibson, G., Jia, W., and Pettersson, S. (2012) Host-gut microbiota metabolic interactions. *Science* **336**: 1262–1267.

Noecker, C., Chiu, H.-C., McNally, C.P., and Borenstein, E. (2019) Defining and evaluating microbial contributions to metabolite variation in microbiome-metabolome association studies. *mSystems* **4**: 10.1128/msystems.00579-19.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Research* **27**: 824–834.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**: 1043–1055.

Peng, Q., Zhao, C., Wang, X., Cheng, K., Wang, C., Xu, X., and Lin, L. (2025) Modeling bacterial interactions uncovers the importance of outliers in the coastal lignin-degrading consortium. *Nature Communications* **16**: 639.

Pinto, S., Benincà, E., Nes, E.H. van, Scheffer, M., and Bogaards, J.A. (2022) Species abundance correlations carry limited information about microbial network interactions. *PLOS Computational Biology* **18**: e1010491.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**: D590–D596.

Ruan, Z., Chen, K., Cao, W., Meng, L., Yang, B., Xu, M., et al. (2024) Engineering natural microbiomes toward enhanced bioremediation by microbiome modeling. *Nature Communications* **15**: 4694.

Ruiz, J., De Celis, M., Diaz-Colunga, J., Vila, J.C., Benitez-Dominguez, B., Vicente, J., et al. (2023) Predictability of the community-function landscape in wine yeast ecosystems. *Molecular Systems Biology* **19**: e11613.

Sanchez-Gorostiaga, A., Bajić, D., Osborne, M.L., Poyatos, J.F., and Sanchez, A. (2019)

High-order interactions distort the functional landscape of microbial consortia. *PLoS Biology* **17**: e3000550.

Sarker, S.D., Nahar, L., and Kumarasamy, Y. (2007) Microtitre plate-based antibacterial assay incorporating resazurin as an indicator of cell growth, and its application in the in vitro antibacterial screening of phytochemicals. *Methods* **42**: 321–324.

Schlechter, R.O., Jun, H., Bernach, M., Oso, S., Boyd, E., Muñoz-Lintz, D.A., et al. (2018) Chromatic bacteria – a broad host-range plasmid and chromosomal insertion toolbox for fluorescent protein expression in bacteria. *Frontiers in Microbiology* **9**: 3052.

Shapley, L. (1953) A value for n-person games. In: Kuhn, H. and Tucker, A., Eds., Contributions to the Theory of Games II, Princeton University Press, Princeton, 307–317.

Skwara, A., Gowda, K., Yousef, M., Diaz-Colunga, J., Raman, A.S., Sanchez, A., et al. (2023) Statistically learning the functional landscape of microbial communities. *Nature Ecology and Evolution* **7**: 1823–1833.

Stouffer, D.B. and Bascompte, J. (2010) Understanding food-web persistence from local to global scales. *Ecology Letters* **13**: 154–161.

Takeo, M., Fujii, T., and Maeda, Y. (1998) Sequence analysis of the genes encoding a multicomponent dioxygenase involved in oxidation of aniline and o-toluidine in *Acinetobacter* sp. strain YAA. *Journal of Fermentation and Bioengineering* **85**: 17–24.

Takeo, M., Ohara, A., Sakae, S., Okamoto, Y., Kitamura, C., Kato, D., and Negoro, S. (2013) Function of a glutamine synthetase-like protein in bacterial aniline oxidation via γ-glutamylanilide. *Journal of Bacteriology* **195**: 4406–4414.

Tanizawa, Y., Fujisawa, T., and Nakamura, Y. (2018) DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. *Bioinformatics* **34**: 1037–1039.

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

van den Belt, M., Gilchrist, C., Booth, T. J., Chooi, Y, Medema, M. H., and Alanjary, M.

725 (2023) CAGECAT: the comparative gene cluster analysis toolbox for rapid search
726 and visualization of homologous gene clusters. *BMC Bioinformatics* **24**: 181.