nature chemical biology

Supplementary information

https://doi.org/10.1038/s41589-026-02272-4

# Designing fiber–gut microbiome interactions with active learning

In the format provided by the authors and unedited

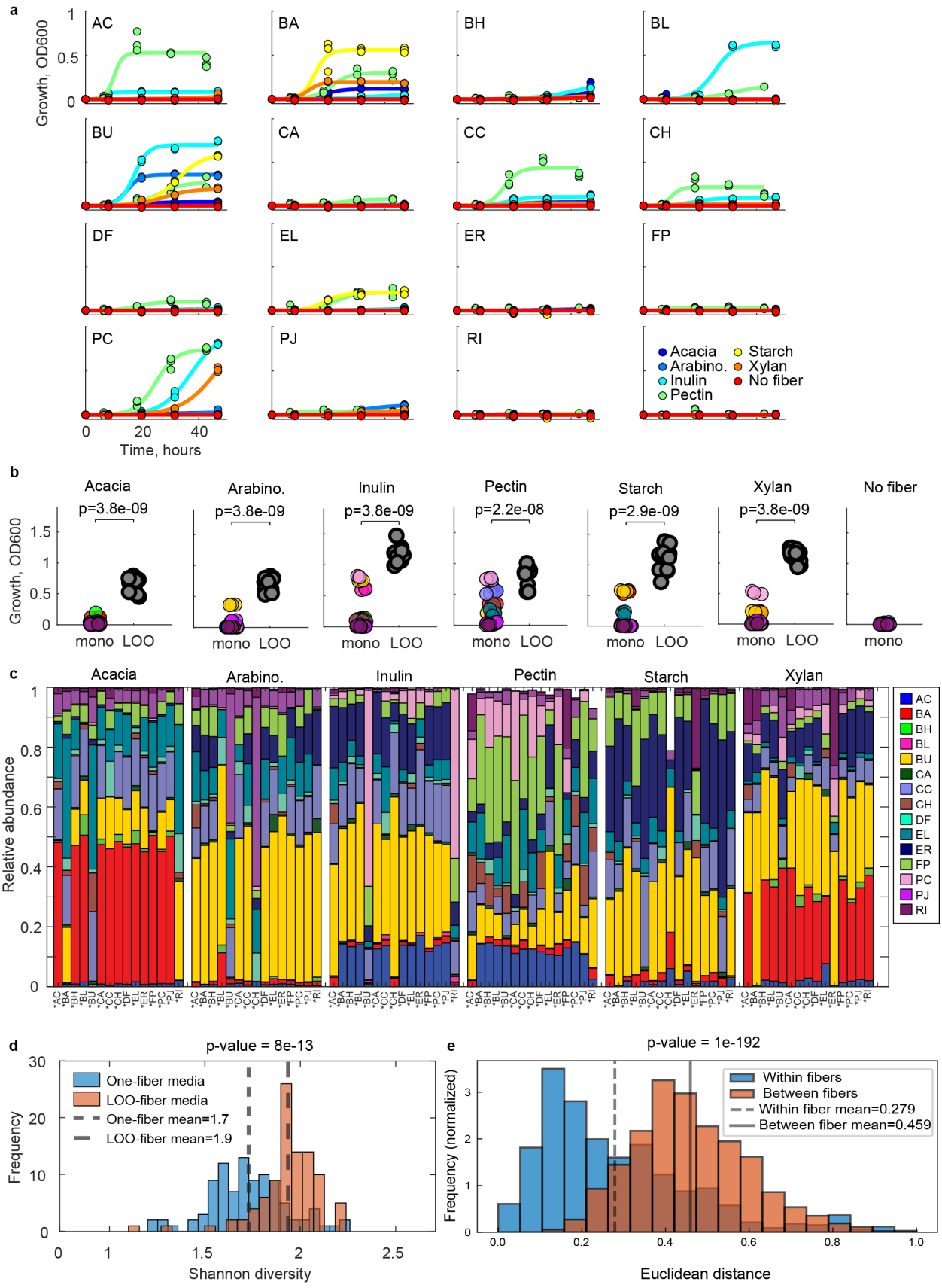
**Supplementary information contents**

| Model | Training data | Validation Step 1: cross validation on corresponding DTL data | Validation step 2: retroactive validation on additional DTL data |
|---|---|---|---|
| MR0 | DTL0 | Training – training sets for 20-fold cross validation on DTL0; monoculture 48hr timepoint<br><br>Test – test sets for 20-fold cross validation with DTL0 data | Training – All DTL0 data; monoculture 48hr timepoint<br><br>Test – All DTL1, 2, 3, & VAL data |
| MR1 | DTL0 & 1 | Training – training sets for 20-fold cross validation on DTL0 & 1; monoculture 48hr timepoint<br>Test – test sets for 20-fold cross validation with DTL0 & 1 data | Training – All DTL0 & 1 data; monoculture 48hr timepoint<br><br>Test – All DTL1, 2, 3, & VAL data |
| MR2 | DTL0, 1, & 2 | Training – training sets for 20-fold cross validation on DTL0, 1, & 2; monoculture 48hr timepoint<br><br>Test – test sets for 20-fold cross validation with DTL0, 1, & 2 | Training – DTL0, 1, & 2; monoculture 48hr timepoint<br><br>Test – All DTL2, 3, & VAL data |
| MR3 | DTL0, 1, 2, & 3 | Training – training sets for 20-fold cross validation on DTL0, 1, 2, & 3; monoculture 48hr timepoint<br><br>Test – test sets for 20-fold cross validation with DTL0, 1, 2, & 3 | n/a |

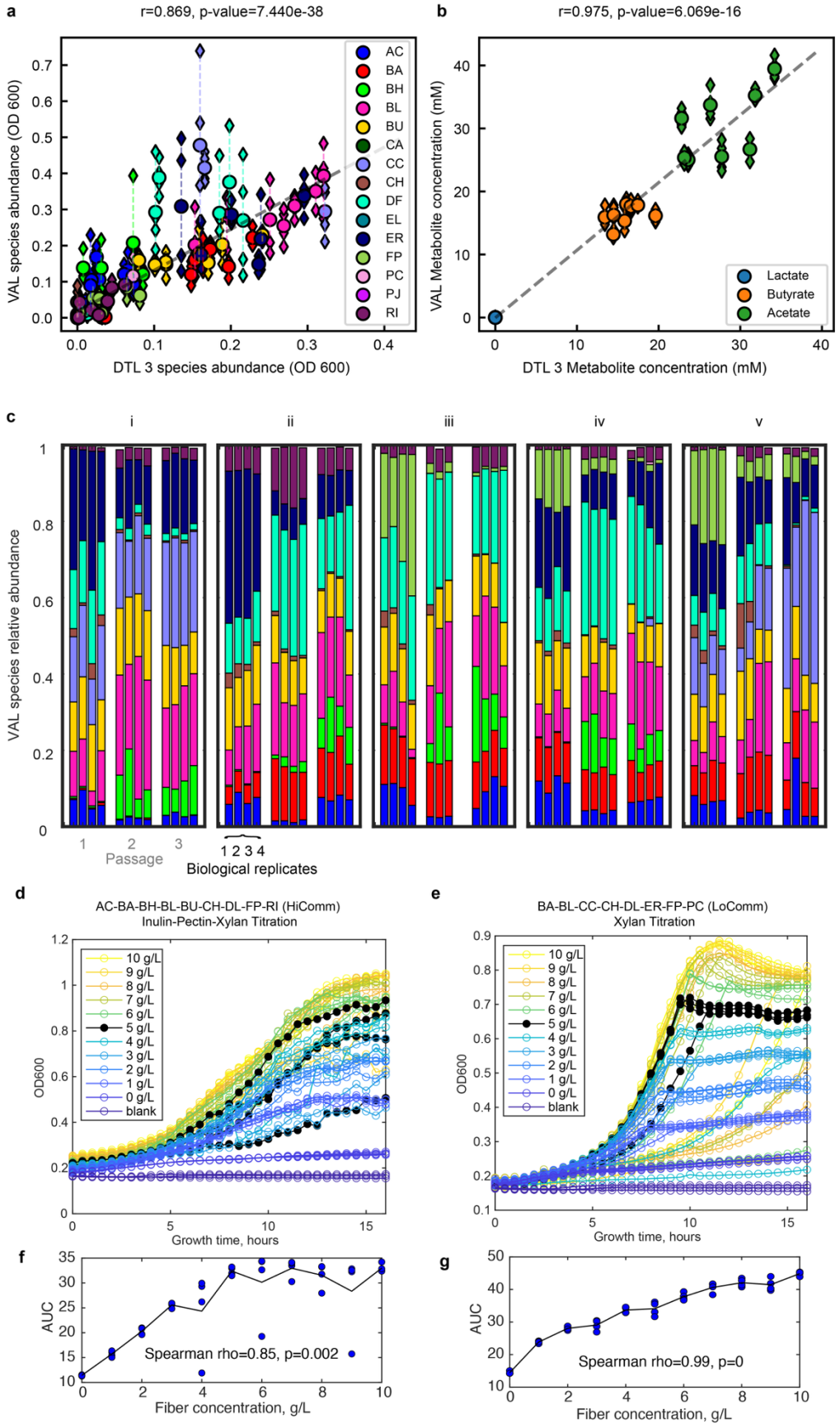
**Supplementary Table 1** Model iterations and training/test data set descriptions corresponding to **Fig 2**. Models were trained on community abundance and pH data for all passages and datasets. Models were trained on organic acid data for passages one and three for DTL1, 2, and 3. DTL1-3 consist of 2-15 species, 1-6 fiber combinations designed using Bayesian experimental design and/or optimization. DTL0 represents all 15 single-species dropout communities cultured on all 6 single-fiber media and all six single-fiber dropout media (192 conditions); organic acids were not measured for DTL0.

| DTL cycle | Design approach | Exploration | Exploitation | Number of conditions |
|---|---|---|---|---|
| 0 | N/A | N/A | N/A | 387 |
| 1 | Expected information gain | Yes | No | 288 |
| 2 | Expected information gain + objective | Yes | Yes | 288 |
| 3 | Thompson sampling | Yes | Yes | 288 |
| 4 | Objective | No | Yes | 8 |

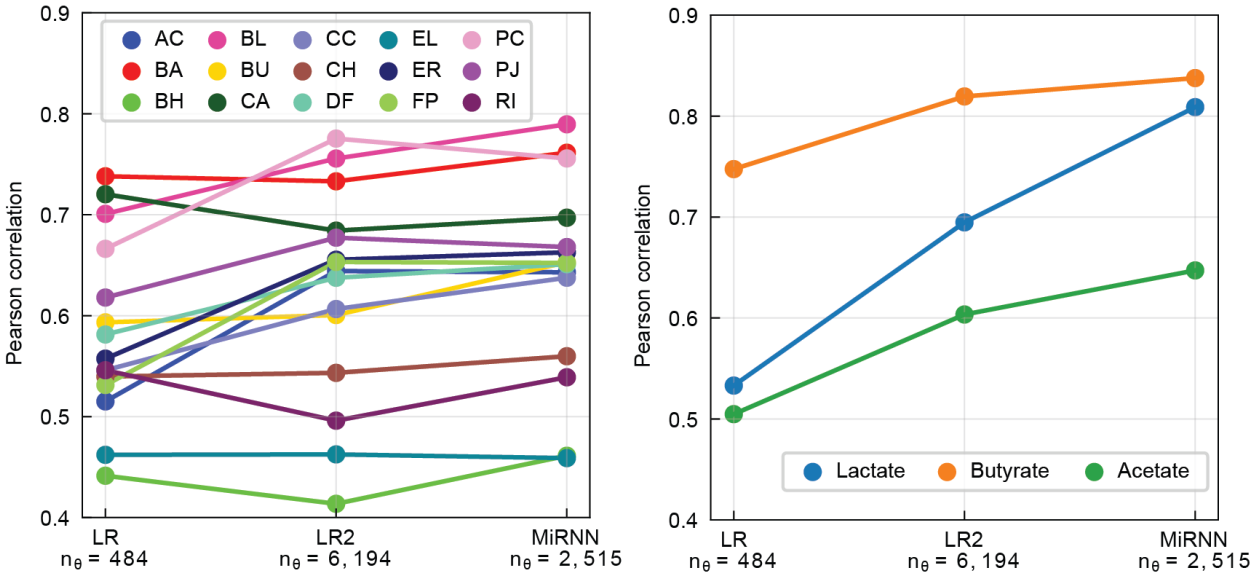
**Supplementary Table 2** Experimental design algorithm used in each DTL cycle. In DTL 0, a preliminary experimental design consisted of the full community and communities of all species except one (leave-one-out) communities cultured in the presence of individual fibers or single-fiber dropout media. DTL 0 also included single species in the presence of individual fibers, single-fiber dropouts, and no-fiber media. In total, this initial data set comprised 387 community-fiber combinations. In DTL 1, a pure exploration strategy was used where 288 conditions were selected such that the expected information gain of the set of conditions was maximized. In DTL 2, this strategy was extended to account for the design objective that quantified model predicted butyrate production, diversity, and stability. In DTL 3, we expanded the design space to include non-equal ratios of fibers and used a Thompson sampling strategy to select 288 conditions. Exploration in Thompson sampling is achieved by randomly sampling parameter values from the posterior distribution every time a condition is selected. Finally, we selected 8 conditions using pure exploitation in DTL 4 to confirm convergence of the design objective.


a
Growth, OD600
AC
BA
BH
BL
BU
CA
CC
CH
DF
EL
ER
FP
PC
PJ
RI
Time, hours
Acacia
Arabino.
Inulin
Pectin
Starch
Xylan
No fiber
b
Growth, OD600
p=3.8e-09
p=3.8e-09
p=3.8e-09
p=2.2e-08
p=2.9e-09
p=3.8e-09
mono
LOO
c
Relative abundance
d
p-value = 8e-13
One-fiber media
LOO-fiber media
One-fiber mean=1.7
LOO-fiber mean=1.9
Frequency
Shannon diversity
e
p-value = 1e-192
Within fibers
Between fibers
Within fiber mean=0.279
Between fiber mean=0.459
Frequency (normalized)
Euclidean distance

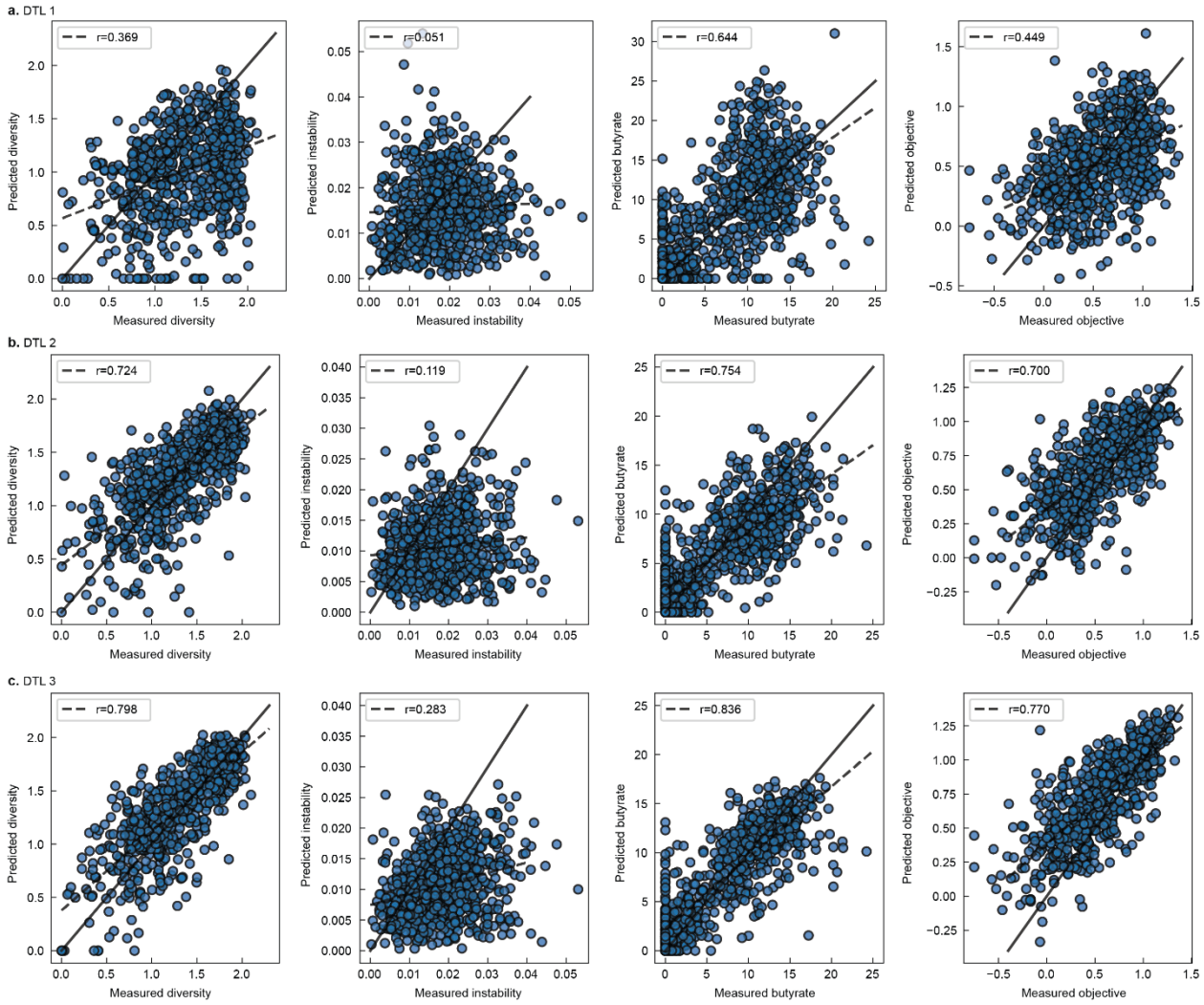
**Supplementary Fig. 1.** Monoculture and leave-one-out communities cultured on single-fiber media **a** Line plots indicating growth curves for all single-species, single-fiber conditions. Colored markers indicate average of n=3 biological replicates and colored lines indicate logistic model fits. Colors indicate species with the color legend shown in the bottom right panel. The bottom right panel shows timeseries sampling of blank media. **b** Categorical scatter plot of highest growth monoculture timepoint (panel a) vs. total leave-one-out (LOO) community growth in each single-fiber media. A two sided Mann-Whitney U test is used to determine statistical significance with the p-value labeled above each plot alongside the corresponding fiber. “Acacia” indicates acacia gum and “Arabino.” indicates arabinogalactan. **c** Stacked bar plots indicating endpoint (48 hour) relative abundances of all leave-one-out communities cultured on all single-fiber media. The species not included in each community is labeled with an asterisk on the x-axis. **d** Histograms indicating distributions of Shannon diversities for single-fiber media (“one-fiber”) and leave-one-fiber-out media (five fibers) for all leave-out-out species communities. The p-value for a two-sided Mann-Whitney U test is shown above the plot. **e** Histograms of Euclidean distances between communities of leave-one-out species cultured in the same fiber (blue) and in different fibers (red-orange). The Euclidean distance between communities when cultured in the same fiber was significantly less than distances between fibers (two-tailed independent t-test p-value shown above plot)


a
r=0.869, p-value=7.440e-38
VAL species abundance (OD 600)
DTL 3 species abundance (OD 600)
AC
BA
BH
BL
BU
CA
CC
CH
DF
EL
ER
FP
PC
PJ
RI
b
r=0.975, p-value=6.069e-16
VAL Metabolite concentration (mM)
DTL 3 Metabolite concentration (mM)
Lactate
Butyrate
Acetate
c
i
ii
iii
iv
v
VAL species relative abundance
1
2
3
Passage
1 2 3 4
Biological replicates
d
AC-BA-BH-BL-BU-CH-DL-FP-RI (HiComm)
Inulin-Pectin-Xylan Titration
OD600
Growth time, hours
10 g/L
9 g/L
8 g/L
7 g/L
6 g/L
5 g/L
4 g/L
3 g/L
2 g/L
1 g/L
0 g/L
blank
e
BA-BL-CC-CH-DL-ER-FP-PC (LoComm)
Xylan Titration
OD600
Growth time, hours
f
AUC
Spearman rho=0.85, p=0.002
Fiber concentration, g/L
g
AUC
Spearman rho=0.99, p=0
Fiber concentration, g/L

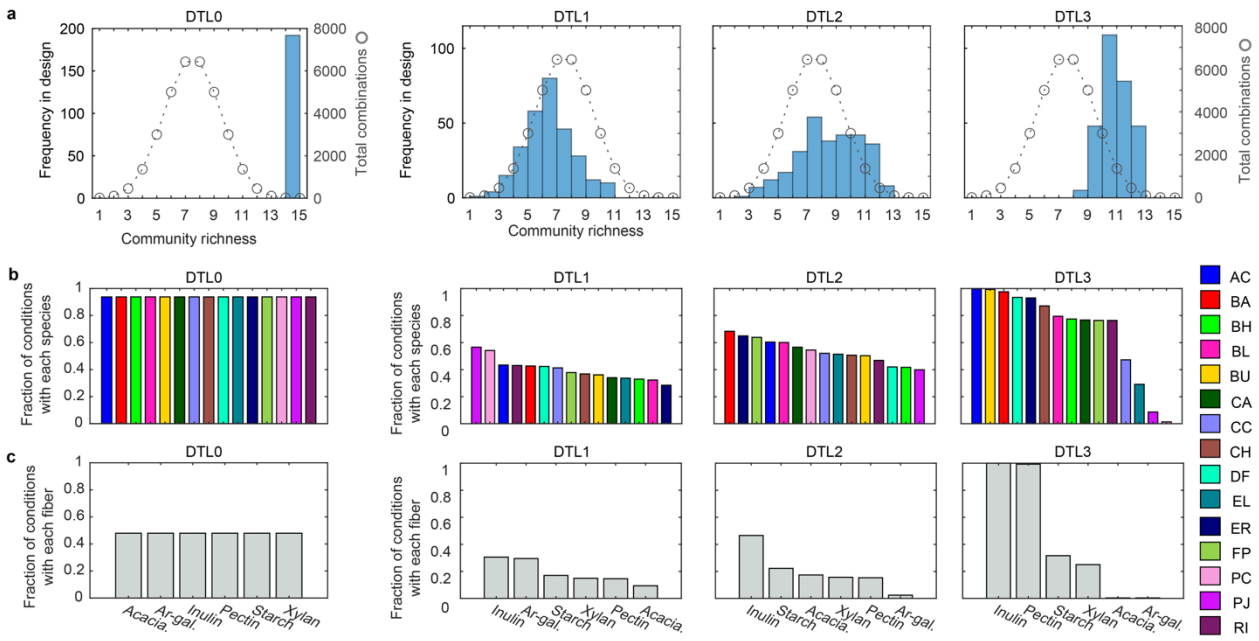
**Supplementary Fig. 2.** Design objective validation data and frequency of feature selection. **a** Scatter plot of absolute abundances for the DTL3 validation experiment (n=4 biological replicates, individual replicates shown as small diamond markers and average value of replicates shown as circles) vs. abundances of (n=1 biological replicate) for corresponding communities cultured in DTL3. Dashed line indicates regression line. Pearson correlation coefficient corresponding two-sided p-value are indicated above the plot. **b** Same as panel **a** but showing metabolite concentrations. Formatting and statistics are as described in panel a legend. **c** Stacked bar indicating relative abundances of all biological replicates for DTL3 validation conditions of which the mean and standard deviation are shown in Fig. 4c. **d** Line plot depicting growth curves of increasing total fiber concentrations with a representative high objective community (AC-BA-BH-BL-BU-CH-FP-RI, 4Inulin-1pectin-1xylan ratio, "HiComm/IPX"). Colors indicate total fiber concentration in the media, n=4 biological replicates are shown. The 5 g/L condition is shown in black as this was the concentration used throughout the DTL cycles and all other *in vitro* experiments. **e** Line plot depicting growth curves of increasing fiber concentrations with a representative low objective community (BA-BL-CC-CH-DL-ER-FP-PC, xylan) 4Inulin-1pectin-1xylan ratio, "HiComm/IPX"). Colors indicate total fiber concentration in the media, n=4 biological replicates are shown. The 5 g/L condition is shown in black as this was the concentration used throughout the DTL cycles and all other *in vitro* experiments. **f,g** Scatter plot with blue circles indicating area under the curve (AUC) vs. total fiber concentration (n=4 biological replicates), and where line indicates mean AUC vs. total fiber concentration. (f) BA-BL-CC-CH-DL-ER-FP-PC, xylan corresponding to panel **d** and (g) AC-BA-BH-BL-BU-CH-FP-RI, 4Inulin-1pectin-1xylan corresponding to panel **e**. Area under the curve is calculated using MATLAB's "trapz" numerical integration function. Spearman correlation coefficient and corresponding two-sided p-value are indicated in the plot.



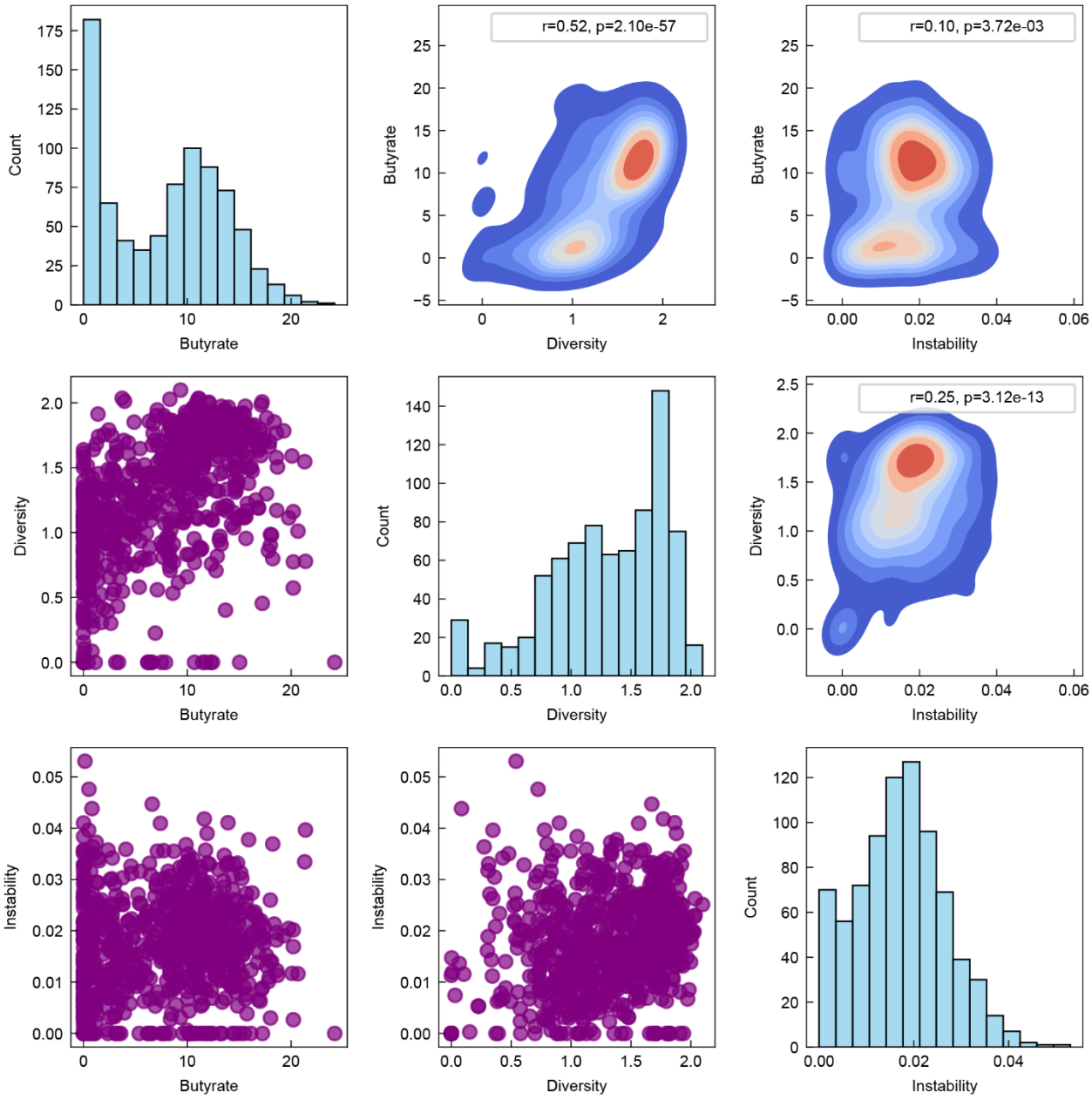

**Supplementary Fig. 3 Comparison of prediction performance between a linear state-state model** with only main effects (LR), a linear state-space model with main and pairwise-effects (LR2), and the MiRNN with a 32-dimensional hidden state. The LR2 model outperformed LR1 for 11 of the 15 species and for all metabolites, and the MiRNN outperformed the LR2 model for 10 of the 15 species and all metabolites. The number of parameters of each model is shown in the x-axis, indicating that the inclusion of pairwise-terms results in more than 3 times the number of parameters of the MiRNN.



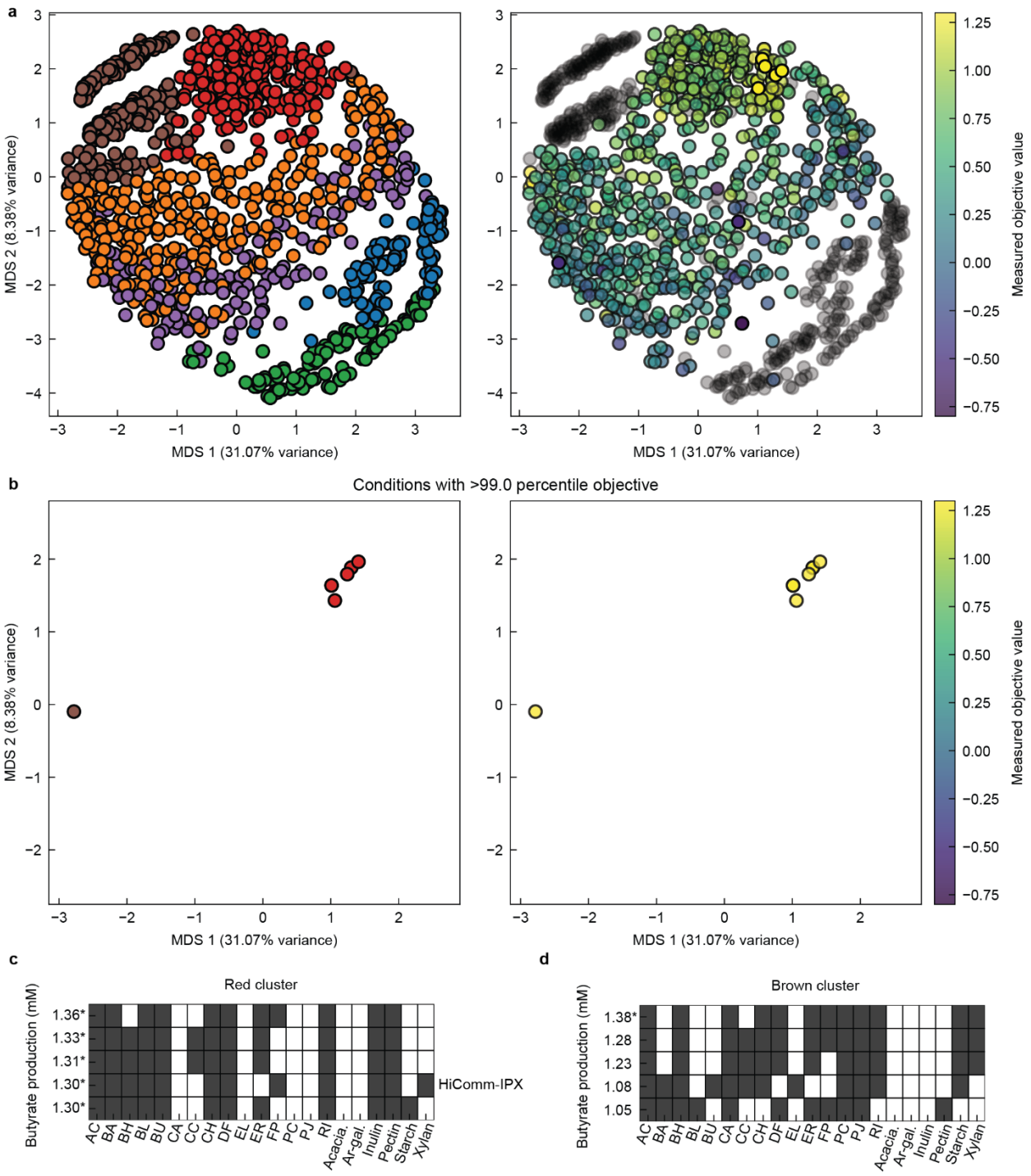

**Supplementary Fig. 4** Scatter plot of model predicted diversity, instability, butyrate, and overall objective vs. measured values. Dashed line indicates regression line. Pearson correlation corresponding to regression line are shown in top left. **a** Prediction performance on held-out data using only data from DTL 0-1 as training data. **b** Prediction performance on held-out data using only data from DTL 0-2 as training data. **c** Prediction performance on held-out data using data from DTL 0-3 as training data.




**Supplementary Fig. 5. Exploration and exploitation of species-fiber design across DTL cycles** (**a)** Histograms depicting frequency of community richness (number of species in each condition) for experimental designs for DLT0, DTL1, DTL2, and DTL3 (lefthand axis). Circular markers connected by dashed lines indicate the total possible number of combinations for each richness level (righthand axis). Combinations are calculated using the “n choose k” formula with n=15 and k=1 to 15. (**b),(c)** Bar plots showing fraction of conditions in each experimental design cycle containing each species (f) and fiber (g). Species and fibers are sorted left to right along the x-axis (1) alphabetically for DTL0 and by rank order of frequency for DTL1, DTL2, and DTL3. Species bar colors correspond to the legend at bottom right.

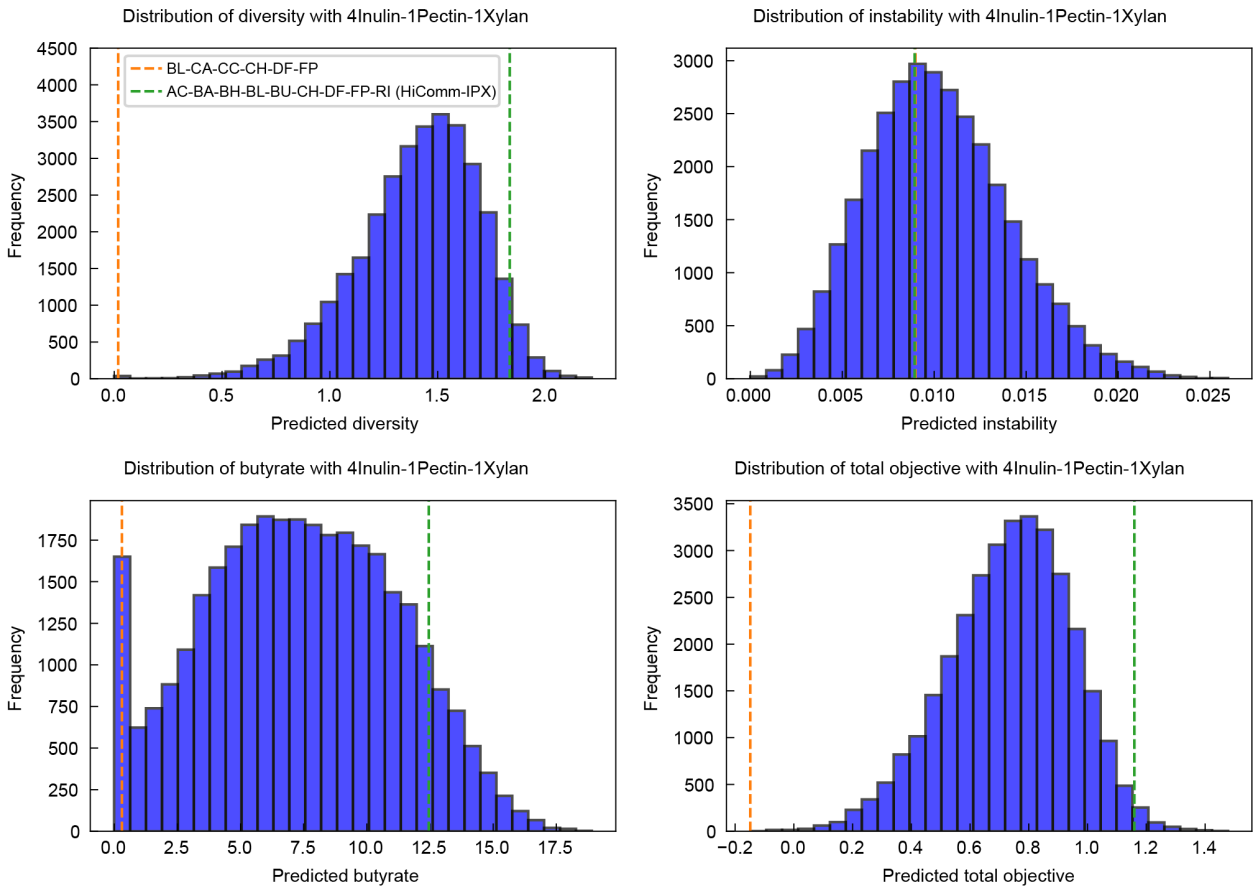



**Supplementary Fig. 6** Histograms (diagonal), scatter plots (lower triangle), and kernel density estimate plots (upper triangle) of experimentally measured butyrate, diversity, and instability.

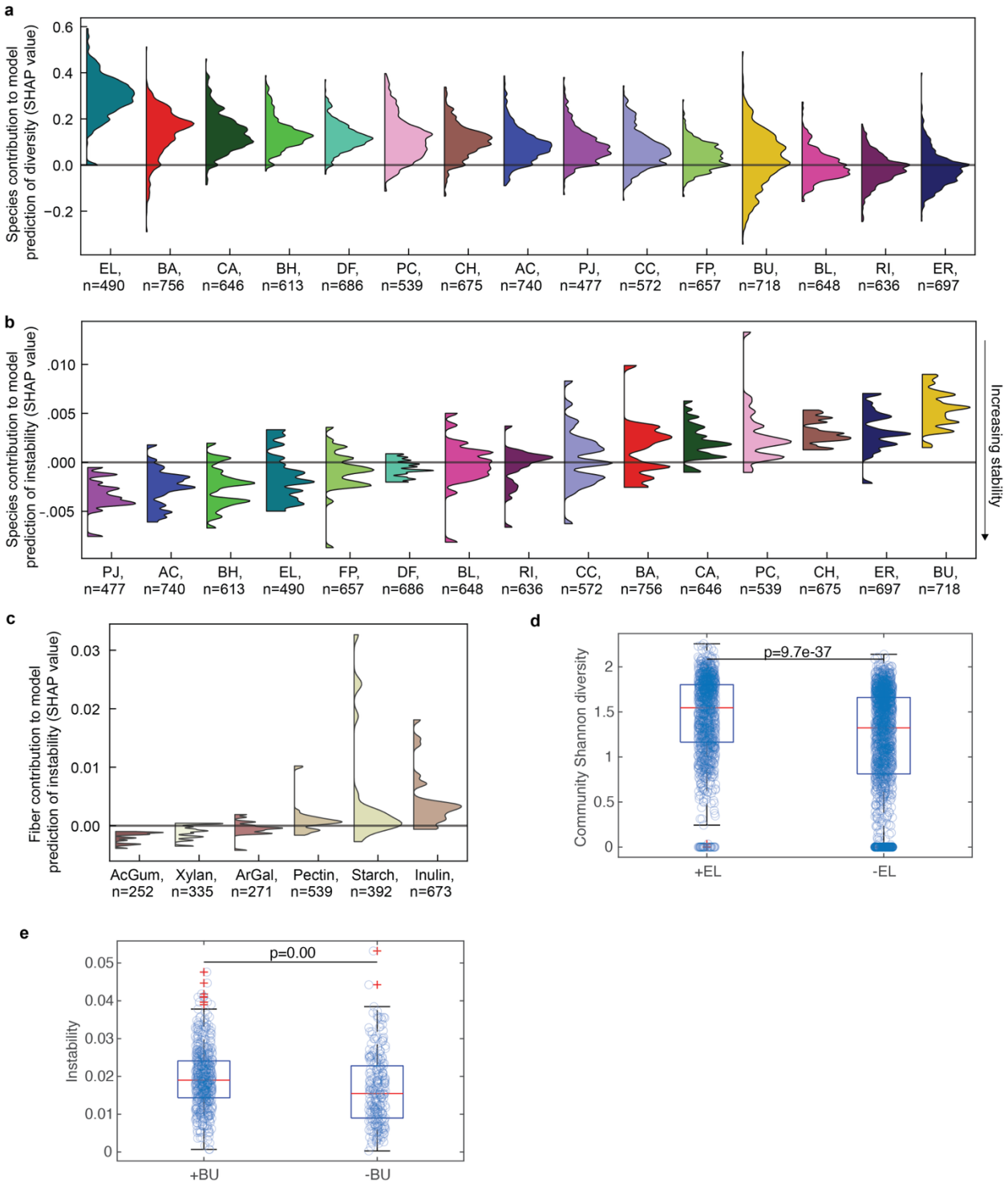



**Supplementary Fig. 7 K-means clustering of the design space. a** Left panel shows K-means clustering with K=6 of the design space where each condition is colored according to its cluster assignment. Right panel shows each condition where colors represent the measured objective value. **b** Both panels show remaining conditions and

cluster assignments after thresholding only conditions in the 99$^{th}$ percentile of measured objective values. **c** Presence (dark squares) or absence (blank squares) of species or fibers in the set of conditions in the red cluster with the 5 highest measured objective values. Rows with asterisks indicate that the condition was in the 99$^{th}$ percentile of objective values. **d** Same as panel c but showing conditions from the brown cluster.

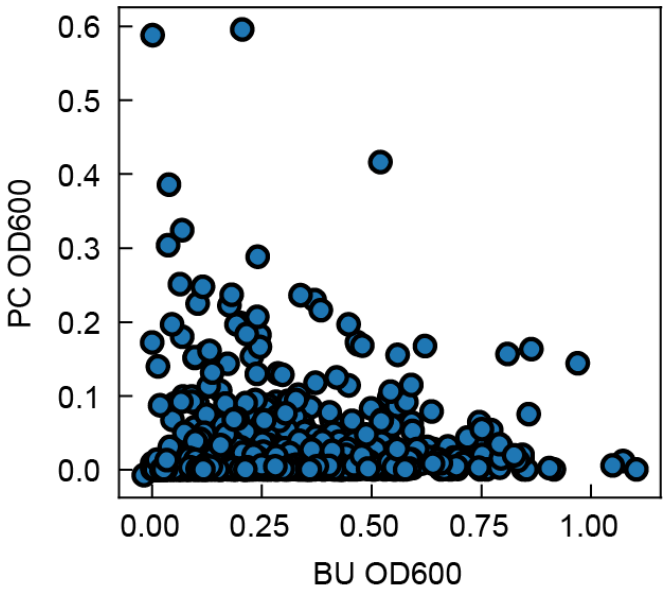



**Supplementary Fig. 8 The effect of species combinations on model predicted objectives keeping the fiber ratio fixed.** With the fiber ratio fixed to 4 inulin : 1 pectin : 1 xylan, the MiRNN trained on the full data set was used to predict each component of the design objective for each possible community combination. The distributions show that the fiber ratio is not sufficient to guarantee high butyrate, high diversity, and low instability. The community combination *Bifidobacterium longum* (BL), *Collinsella aerofaciens* (CA), *Coprococcus comes* (CC), *Clostridium hiranonis* (CH), *Dorea formicigenerans* (DF), and *Faecalibacterium prausnitzii* (FP) resulted in the worst overall objective (plotted as orange dotted lines) with nearly zero predicted butyrate and diversity with moderate instability. In comparison, HiComm-IPX (plotted as green dotted lines) yielded high diversity and high butyrate with similar stability.

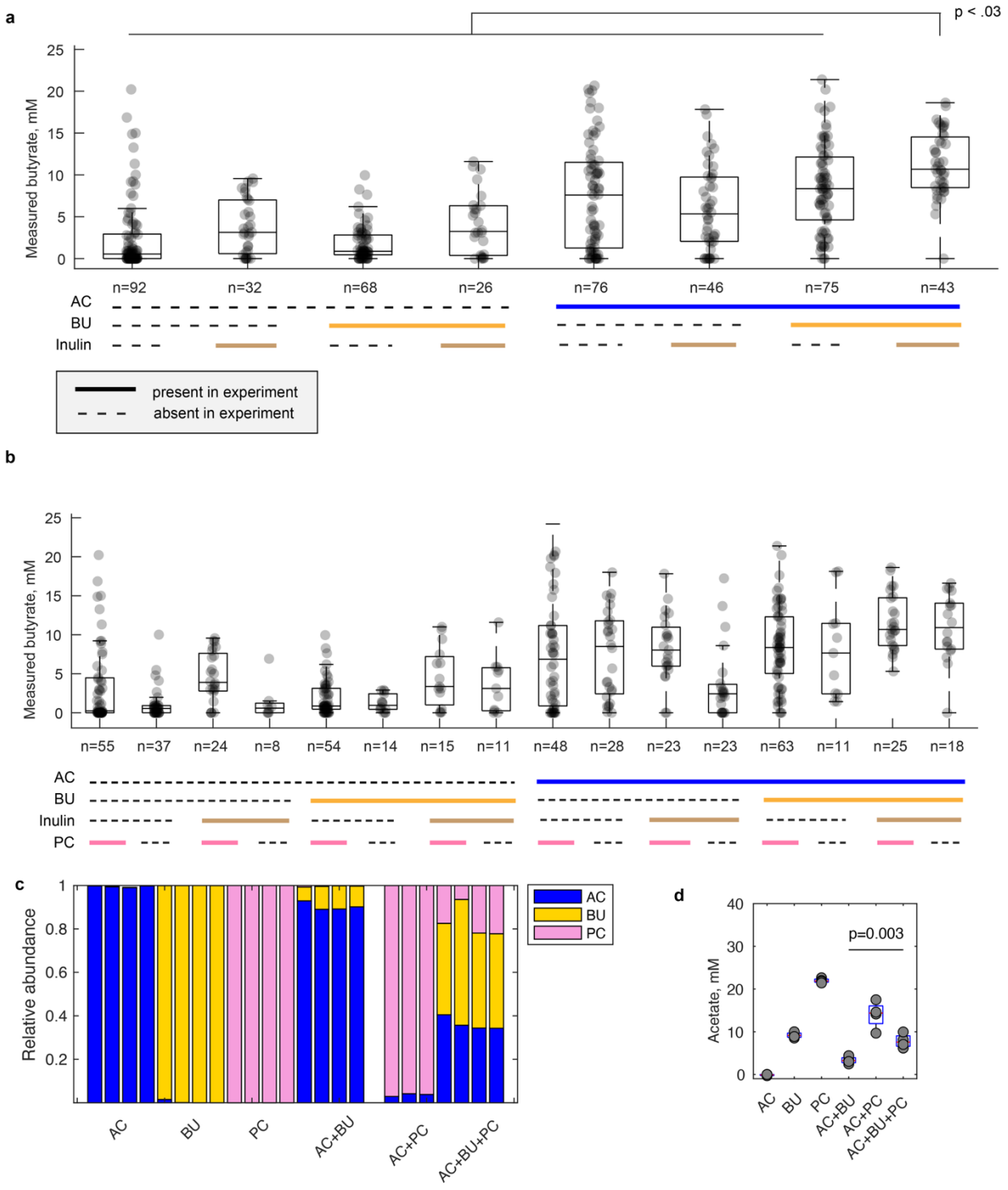



**Supplementary Fig. 9 Understanding drivers of diversity using explainable machine learning. a** Distributions of SHAP values (Shapley additive explanation) of species contribution to model predicted diversity. SHAP values in the plotted distributions correspond to conditions with the species present, with the number of points in each distribution indicated in the x-axis tick labels. The horizontal line indicates a value of zero.

The distributions are sorted by highest to lowest median value. **b** Distributions of SHAP values of species contribution to model predicted instability. The distributions are sorted by highest to lowest median value. **c** Distributions of SHAP values of fiber contribution to model predicted instability for conditions containing respective fiber. **d** Boxplot with overlayed markers indicating community Shannon diversity and **e** instability measurements for experimental communities containing or not containing EL and BU, respectively, n=1 biological replicate per condition. Center line, box edges, and whiskers indicating median, quartiles, and quartile plus 1.5 times interquartile range, respectively. Red markers indicate outliers falling beyond whisker range. The p-values correspond to a two-sided, unpaired t-test.

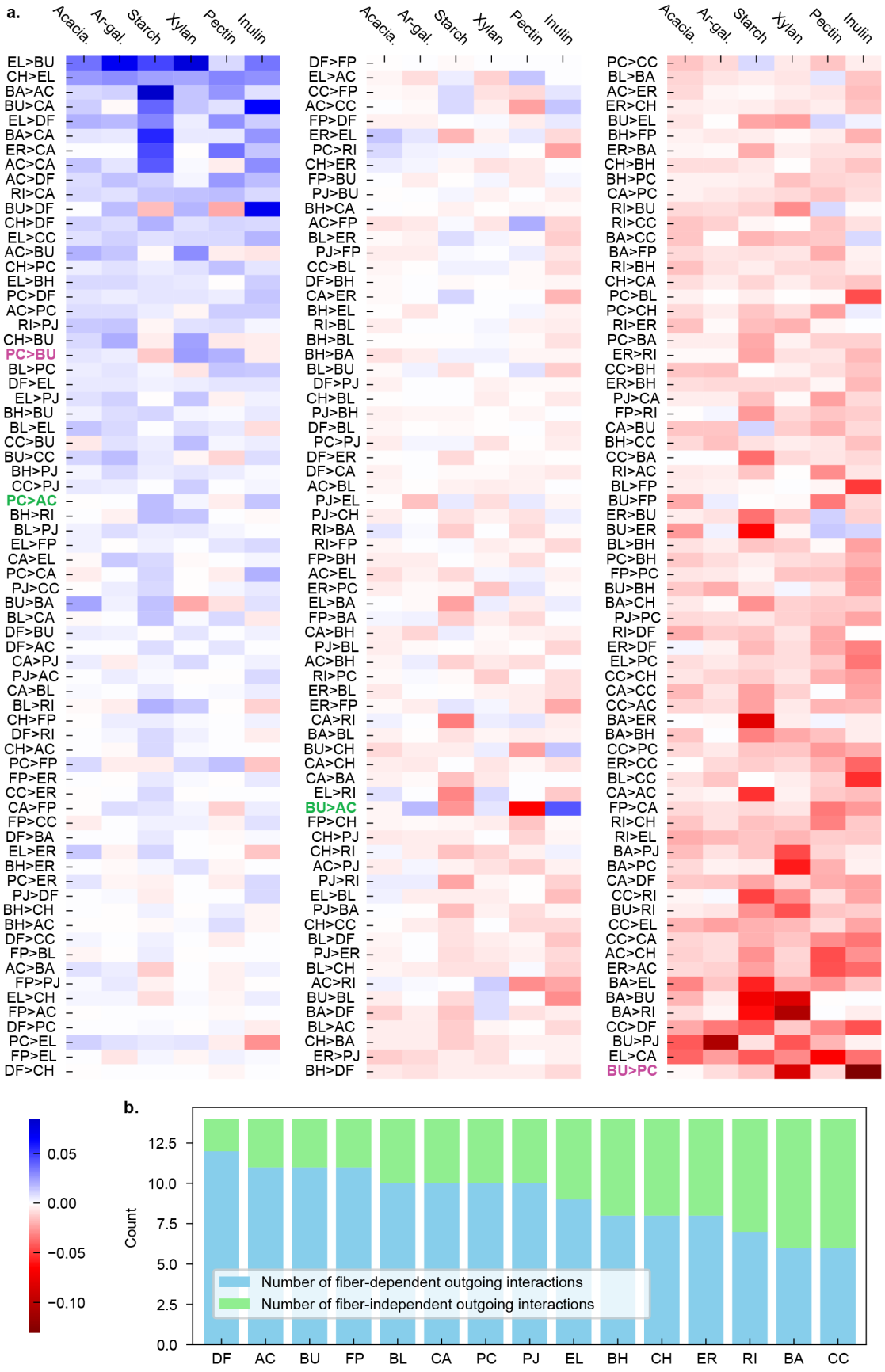



**Supplementary Fig. 10 Scatter plot of *P. copri* (PC) versus *B. uniformis* (BU)** A scatter plot of experimentally measured absolute abundance (OD600) of *P. copri* versus *B. uniformis* from all experimental conditions that contained both species suggests competitive exclusion, where if one species has a high absolute abundance, the other has a low absolute abundance.

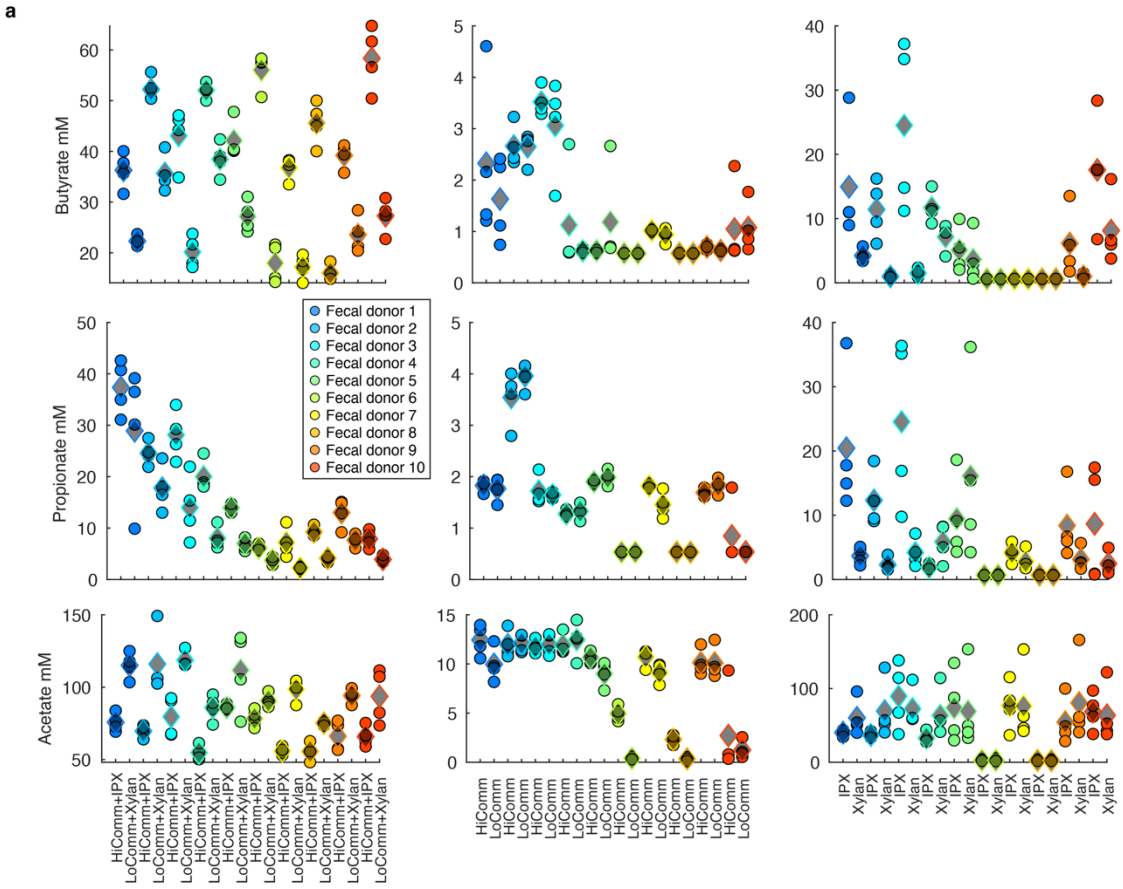


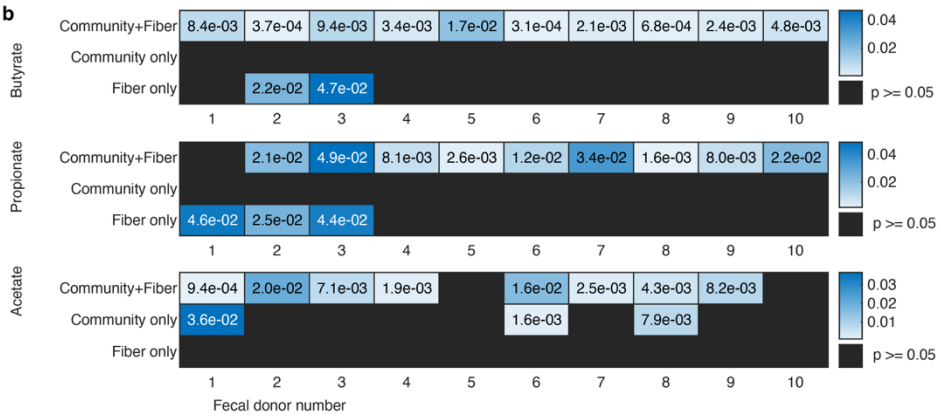

**Supplementary Fig. 11 Analysis of passage three experimental butyrate measurements for higher-order interactions. a** Categorical scatter plot of experimentally measured butyrate concentrations for communities arrange by every combination of presence and absence of *A. caccae* (AC), *B. uniformis* (BU), and inulin for passage three. Inulin presence indicates that inulin was the only fiber present in the condition. Presence is indicated by the solid line of which the color corresponds to the

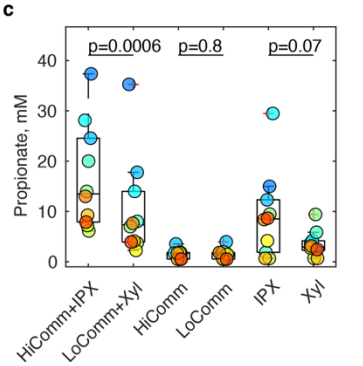
variable labeled on the left, and absence is indicated by a dashed black line. Box plots indicate summary statistics of the distribution with center line, box edges, and whiskers indicating median, quartiles, and quartile plus 1.5 times interquartile range, respectively. The number of experimental conditions in each distribution is indicated below the box plot by “n”, of which each is a single biological replicate. The statistical comparison at the top of the panel indicates group eight (copresence of AC, BU, and inulin) is significantly higher than all other individual groups using pairwise comparisons (p-values < .03). All pairwise group comparisons are made using one-way analysis of variance (ANOVA) followed by Tukey’s honest significant difference (HSD) procedure using MATLAB’s “anova1” and “multcompare” functions. All 28 pairwise comparisons are reported in **Supplemental Data 4**. **b** Categorical scatter plot of experimentally measured butyrate concentrations for every combination of presence and absence of *A. caccae* (AC), *B. uniformis* (BU), inulin, and *P. copri* (PC) for passage one. Statistics, boxplot formatting, and replicates are as described in panel a legend. All 120 pairwise group comparisons are reported in **Supplemental Data 4. c** Stacked bar plots where bar height indicates relative abundance (16S fraction) of AC, BU, and PC from cultures on inulin corresponding to butyrate and propionate measurements shown in **Fig. 5c,d)**. Each group of four bars indicates n=4 biological replicates. The first replicate of AC+PC is omitted due to cross-contamination. **d** Box and categorical scatter plots of acetate concentrations for corresponding to **Fig. S9c**, **Fig. 5c,d**. Markers indicate concentration of n=4 biological replicates measured after passage 3. Measurements correspond to butyrate and propionate measurements shown in **Fig. 5c,d**. The p-value corresponds to a two-sided, unpaired t-test.

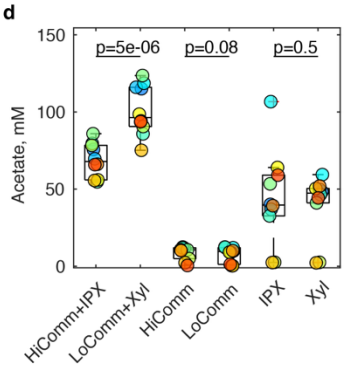

a.
Acacia
Ar-gal.
Starch
Xylan
Pectin
Inulin
EL>BU
CH>EL
BA>AC
BU>CA
EL>DF
BA>CA
ER>CA
AC>CA
AC>DF
RI>CA
BU>DF
CH>DF
EL>CC
AC>BU
CH>PC
EL>BH
PC>DF
AC>PC
RI>PJ
CH>BU
PC>BU
BL>PC
DF>EL
EL>PJ
BH>BU
BL>EL
CC>BU
BU>CC
BH>PJ
CC>PJ
PC>AC
BH>RI
BL>PJ
EL>FP
CA>EL
PC>CA
PJ>CC
BU>BA
BL>CA
DF>BU
DF>AC
CA>PJ
PJ>AC
CA>BL
BL>RI
CH>FP
DF>RI
CH>AC
PC>FP
FP>ER
CC>ER
CA>FP
FP>CC
DF>BA
EL>ER
BH>ER
PC>ER
PJ>DF
BH>CH
BH>AC
DF>CC
FP>BL
AC>BA
FP>PJ
EL>CH
FP>AC
DF>PC
PC>EL
FP>EL
DF>CH
DF>FP
EL>AC
CC>FP
AC>CC
FP>DF
ER>EL
PC>RI
CH>ER
FP>BU
PJ>BU
BH>CA
AC>FP
BL>ER
PJ>FP
CC>BL
DF>BH
CA>ER
BH>EL
RI>BL
BH>BL
BH>BA
BL>BU
DF>PJ
CH>BL
PJ>BH
DF>BL
PC>PJ
DF>ER
DF>CA
AC>BL
PJ>EL
PJ>CH
RI>BA
RI>FP
FP>BH
AC>EL
ER>PC
EL>BA
FP>BA
CA>BH
PJ>BL
AC>BH
RI>PC
ER>BL
ER>FP
CA>RI
BA>BL
BU>CH
CA>CH
CA>BA
EL>RI
BU>AC
FP>CH
CH>PJ
CH>RI
AC>PJ
PJ>RI
EL>BL
PJ>BA
CH>CC
BL>DF
PJ>ER
BL>CH
AC>RI
BU>BL
BA>DF
BL>AC
CH>BA
ER>PJ
BH>DF
PC>CC
BL>BA
AC>ER
ER>CH
BU>EL
BH>FP
ER>BA
CH>BH
BH>PC
CA>PC
RI>BU
RI>CC
BA>CC
BA>FP
RI>BH
CH>CA
PC>BL
PC>CH
RI>ER
PC>BA
ER>RI
CC>BH
ER>BH
PJ>CA
FP>RI
CA>BU
BH>CC
CC>BA
RI>AC
BL>FP
BU>FP
ER>BU
BU>ER
BL>BH
PC>BH
FP>PC
BU>BH
BA>CH
PJ>PC
RI>DF
ER>DF
EL>PC
CC>CH
CA>CC
CC>AC
BA>ER
BA>BH
CC>PC
ER>CC
BL>CC
CA>AC
FP>CA
RI>CH
RI>EL
BA>PJ
BA>PC
CA>DF
CC>RI
BU>RI
CC>EL
CC>CA
AC>CH
ER>AC
BA>EL
BA>BU
BA>RI
CC>DF
BU>PJ
EL>CA
BU>PC
0.05
0.00
−0.05
−0.10
b.
Count
12.5
10.0
7.5
5.0
2.5
0.0
DF
AC
BU
FP
BL
CA
PC
PJ
EL
BH
CH
ER
RI
BA
CC
Number of fiber-dependent outgoing interactions
Number of fiber-independent outgoing interactions

**Supplementary Fig. 12 Species interactions determined using explainable machine learning. a** Interspecies interactions were determined by computing the average SHAP contribution of each inoculated species to the prediction of species in the third passage. For each fiber, the average was computed over all experimentally measured conditions that included only that fiber. Interactions hypothesized to promote butyrate production are highlighted in green and interactions hypothesized to inhibit butyrate production are highlighted purple. **b** Each of the 15 species has 14 outgoing interactions with other species. Of the 14 outgoing interactions, the stacked bar plot shows how the interactions partitioned into those that remained either positive or negative in all fiber conditions (green) or changed in sign depending on the fiber (blue).

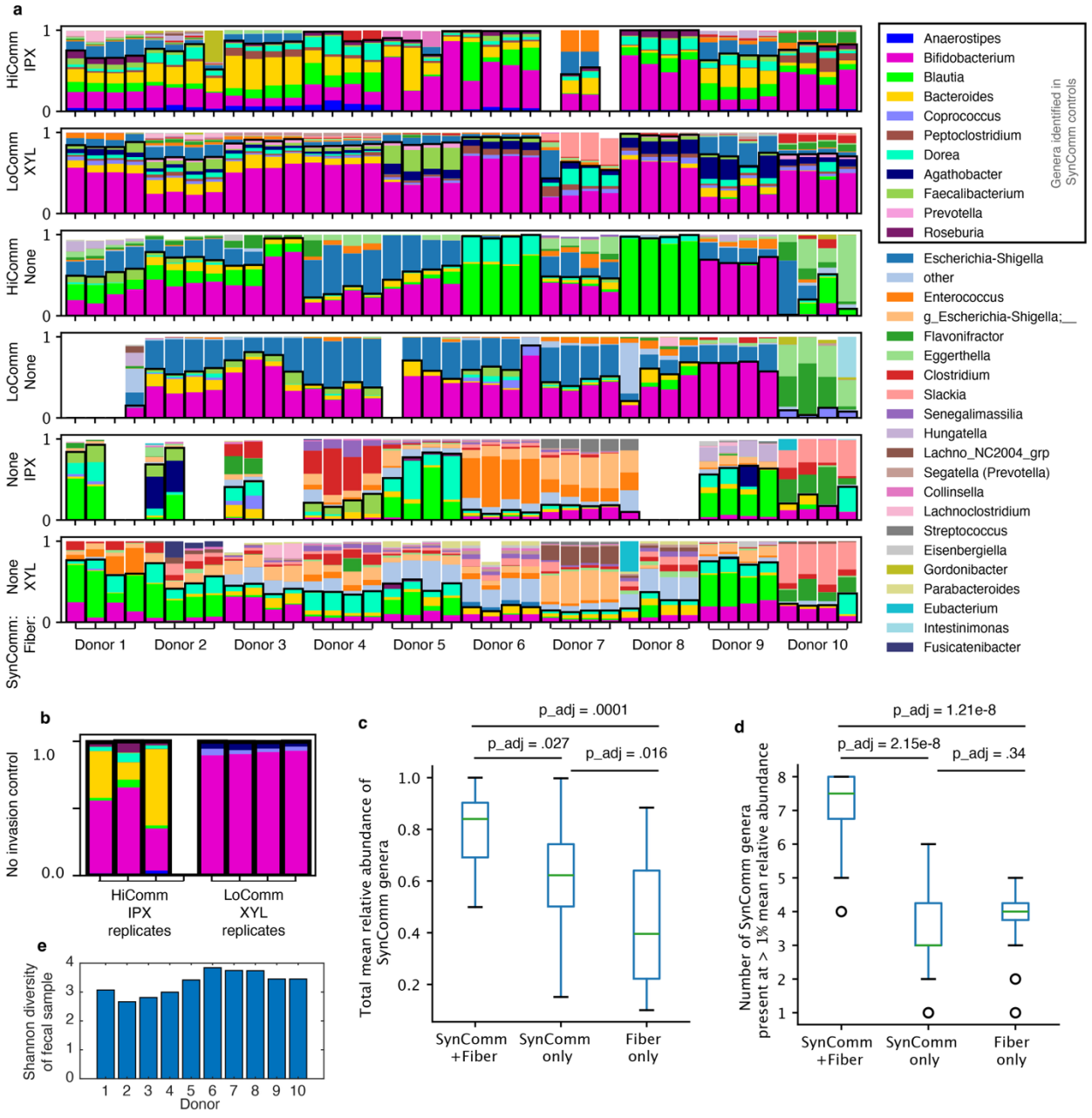

a
Butyrate mM
Propionate mM
Acetate mM
Fecal donor 1
Fecal donor 2
Fecal donor 3
Fecal donor 4
Fecal donor 5
Fecal donor 6
Fecal donor 7
Fecal donor 8
Fecal donor 9
Fecal donor 10
HiComm+IPX
LoComm+Xylan
HiComm
LoComm
IPX
Xylan


b
Butyrate
Propionate
Acetate
Community+Fiber
Community only
Fiber only
8.4e-03 3.7e-04 9.4e-03 3.4e-03 1.7e-02 3.1e-04 2.1e-03 6.8e-04 2.4e-03 4.8e-03
2.2e-02 4.7e-02
2.1e-02 4.9e-02 8.1e-03 2.6e-03 1.2e-02 3.4e-02 1.6e-03 8.0e-03 2.2e-02
4.6e-02 2.5e-02 4.4e-02
9.4e-04 2.0e-02 7.1e-03 1.9e-03 1.6e-02 2.5e-03 4.3e-03 8.2e-03
3.6e-02 1.6e-03 7.9e-03
0.04
0.02
0.03
0.01
p >= 0.05
1 2 3 4 5 6 7 8 9 10
Fecal donor number


c
Propionate, mM
p=0.0006
p=0.8
p=0.07
HiComm+IPX
LoComm+Xyl
HiComm
LoComm
IPX
Xyl


d
Acetate, mM
p=5e-06
p=0.08
p=0.5
HiComm+IPX
LoComm+Xyl
HiComm
LoComm
IPX
Xyl

**Supplementary Fig. 13. Extended metabolite data and analysis from fecal invasion experiment (a)** Categorical scatter plot corresponding to **Fig. 5** indicating butyrate, propionate, and acetate concentrations at the endpoint of fecal invasion experiment for all biological replicates (n=44 per treatment) where color indicates fecal donor number and median is indicated with gray diamond. **(b)** Heatmap indicating p-value for two-sided, unpaired t-test comparing High objective vs. Low objective community+fiber, community, or fiber for each treatment, donor. (**c,d**) Categorical scatter plot indicating propionate and acetate concentration at the endpoint of fecal invasion experiment where color indicates fecal donor number and markers represent average of n=4 biological replicates. All replicates and statistics for each donor are shown in panels a,b. Center line, box edges, whiskers, and red crosses indicate median, quartiles, quartile plus 1.5 times interquartile range, and outliers (beyond 1.5 times interquartile range), respectively. The p-values correspond to a two-sided, paired (by donor) t-test.




**Supplementary Fig. 14. Compositional analysis of fecal invasion experiment (a)** Stacked bar plots depicting n=4 biological replicates of invasion cultures, genus-level composition, at endpoint of fecal invasion experiments using AmpliSeq analysis pipeline. Blank spaces indicate biological replicates with fewer than 3000 reads that were omitted from analyses; n=4 biological replicates are present for most conditions. The stacked bar plots and legend show the 31 most abundant genera (11 HiComm+LoComm genera + 20 fecal origin genera). Missing bars correspond to low-read replicates with fewer than 2750 reads (average number of reads was >80,000). (**b**) Non-invasion (synthetic community only) conditions were included as bioinformatic controls and analyzed in parallel, with genera classified in the non-invasion control outlined by a black box in both the legend and individual stacked bar plots. **(c)** Boxplot indicating distribution of total mean relative

abundances of genera identified in syncom controls for each treatment category (i.e. height of black rectangle outlines in **Fig. 5d**). The distribution contains n=20 means of different donor/community conditions, each taken across n=4 biological replicates. Treatment category refers to SynComm+Fiber (HiComm/IPX and LoComm/XYL), SynComm (HiComm and LoComm), and Fiber (IPX and XYL). P-values are computed using a two-sided, paired (by donor) t-test and adjusted using Benjamini-Hochberg procedure for multiple comparisons. Box plots indicate summary statistics of the distribution with center line, box edges, whiskers, and circles indicating median, quartiles, quartile plus 1.5 times interquartile range, and outliers beyond 1.5 times interquartile range, respectively. **(d)** Boxplot indicating distribution of number of genera identified in SynComm controls present at > 1% average relative abundance for each treatment category in fecal invasion experiment (n=20 per boxplot). Treatment category refers to SynComm+Fiber (HiComm/IPX and LoComm/XYL), SynComm (HiComm and LoComm), and Fiber (IPX and XYL). Statistics and boxplot formatting are as described in legend c. (**e**) Bar plot indicating Shannon diversity of fecal material for 10 donors used in invasion experiment.