Machine Learning-Assisted Identification of Bioindicators Predicts Medium-Chain Carboxylate Production Performance of An Anaerobic Mixed Culture

Bin Liu
Helmholtz-Zentrum fur Umweltforschung UFZ
Heike Sträuber
Helmholtz-Zentrum fur Umweltforschung UFZ
Joao Pedro Saraiva
Helmholtz-Zentrum fur Umweltforschung UFZ
Hauke Harms
Helmholtz-Zentrum fur Umweltforschung UFZ
Sandra Godinho Silva
Universidade de Lisboa Instituto Superior Técnico: Universidade de Lisboa Instituto Superior Tecnico
Jonas Coelho Kasmanas
Helmholtz-Zentrum fur Umweltforschung UFZ
Sabine Kleinsteuber
Helmholtz-Zentrum fur Umweltforschung UFZ
Ulisses Rocha (ulisses.rocha@ufz.de)
Helmholtz-Zentrum fur Umweltforschung UFZ Abteilung Umweltmikrobiologie https://orcid.org/0000-0001-6972-6692

Research

Keywords: Predictive biology, carboxylate platform, model ecosystems, reactor microbiota, microbial chain elongation

Posted Date: July 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-78714/v2

Version of Record: A version of this preprint was published at International Chain Elongation Conference 2020 on November 2nd, 2020. See the published version at https://doi.org/10.18174/icec2020.18013.

Machine learning-assisted identification of bioindicators predicts medium-chain carboxylate production performance of an anaerobic mixed culture

Bin Liu[1], Heike Sträuber[1], João Saraiva[1], Hauke Harms[1], Sandra Godinho Silva[2], Jonas Coelho Kasmanas[1,3,4], Sabine Kleinsteuber[1]* and Ulisses Nunes da Rocha[1]*

* Authors followed by an asterisk contributed equally to this work

[1]Department of Environmental Microbiology, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany

[2]Institute for Bioengineering and Biosciences, Department of Bioengineering, Instituto Superior Técnico Universidade de Lisboa, Lisbon, Portugal

[3]Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, Brazil

[4]Department of Computer Science and Interdisciplinary Center of Bioinformatics, University of Leipzig, Leipzig, Germany

Corresponding Authors:

sabine.kleinsteuber@ufz.de / ulisses.rocha@ufz.de (ordered alphabetically according to last name).

Authors email addresses:

Bin Liu: liu.bin@ufz.de; Heike Sträuber: heike.straeuber@ufz.de; João Saraiva: joao.saraiva@ufz.de; Hauke Harms: hauke.harms@ufz.de; Sandra Godinho Silva: sandragodinhosilva@gmail.com; Jonas Coelho Kasmanas: jonas.kasmanas@usp.br; Sabine Kleinsteuber: sabine.kleinsteuber@ufz.de; Ulisses Nunes da Rocha: ulisses.rocha@ufz.de

Abstract

Background: The ability to quantitatively predict ecophysiological functions of microbial communities provides an important step to engineer microbiota for desired functions related to specific biochemical conversions. Here, we present the quantitative prediction of medium-chain carboxylate production in two continuous anaerobic bioreactors from 16S rRNA gene dynamics in enriched communities.

Results: By progressively shortening the hydraulic retention time (HRT) from 8 days to 2 days with different temporal schemes in two bioreactors operated for 211 days, we achieved higher productivities and yields of the target products *n*-caproate and *n*-caprylate. The datasets generated from each bioreactor were applied independently for training and testing machine learning algorithms using 16S rRNA genes to predict *n*-caproate and *n*-caprylate productivities. Our dataset consisted of 14 and 40 samples from HRT of 8 and 2 days, respectively. Because of the size and balance of our dataset, we compared linear regression, support vector machine and random forest regression algorithms using the original and balanced datasets generated using synthetic minority oversampling. Further, we performed cross validation to estimate model stability. The random forest regression was the best algorithm producing more consistent results with median of error rates below 8%. More than 90% accuracy in the prediction of *n*-caproate and *n*-caprylate productivities was achieved. Four inferred bioindicators belonging to the genera *Olsenella*, *Lactobacillus*, *Syntrophococcus* and *Clostridium* IV suggest their relevance to the higher carboxylate productivity at shorter HRT. The recovery of metagenome-assembled genomes of these bioindicators confirmed their genetic potential to perform key steps of medium-chain carboxylate production.

Conclusions: Shortening the hydraulic retention time of the continuous bioreactor systems allows to shape the communities with desired chain elongation functions. Using machine learning, we demonstrated that 16S rRNA amplicon sequencing data can be used to predict bioreactor process performance quantitatively and accurately. Characterising and harnessing bioindicators holds promise to manage reactor microbiota towards selection of the target processes. Our mathematical framework is transferrable to other ecosystem processes and microbial systems where community dynamics is linked to key functions. The general methodology used here can be adapted to data types of other functional categories such as genes, transcripts, proteins or metabolites.

Keywords: Predictive biology, carboxylate platform, model ecosystems, reactor microbiota, microbial chain elongation

Background

Microbes form complex communities that play essential roles in ecosystem functioning. Identifying bioindicators derived from community analysis and using them to predict process performance may delineate potential cause-effect relationships with ecosystem functioning [1,2]. The knowledge gained from prediction can be used to generate hypotheses on the role of key species. At ecosystem level, designing effective control strategies for key species holds promise to manage the community towards selection of the target processes, which is crucial for microbiota-based biotechnologies [3–5].

Our goals were to investigate how environmental manipulations affect ecosystem functioning

and to predict performance metrics of the quantifiable biological processes by following microbial community dynamics. Model ecosystems offer the opportunity to link microbial diversity and ecosystem functioning in a quantifiable and predictable way [6–8]. Such simplified ecosystems can be still complex regarding microbial interactions and involved metabolic processes [6]. Here, we used anaerobic fermentation reactors as model ecosystems and considered microbial chain elongation (CE) as the quantifiable model ecosystem process. CE is a microbial process that produces medium-chain carboxylates (6 to 8 carbon atoms) through reverse β-oxidation [9]. Recently we enriched a mixed culture that produces *n*-butyrate (C4), *n*-caproate (C6) and *n*-caprylate (C8) from xylan and lactate in a daily-fed reactor system [10], to simulate the feedstock conditions of anaerobic fermentation of ensiled plant biomass [11]. For this bioprocess to be viable, it needs to include diverse functions such as xylan hydrolysis, xylose fermentation and CE with lactate as electron donor. Mixed culture fermentation is characterised by different trophic groups that may cooperate or compete with each other to metabolise complex substrates [9]. Species involved in these interactions can drive shifts in community structure and function [1]. During the long-term stable reactor operation, the community developed towards predominating C4 and biomass production at the cost of C6/C8 production [10]. The current study was conducted on the enriched chain-elongating microbiota in two parallel bioreactors, in order to explore how process parameter changes shape the existing microbiota to optimise the process towards the target products C6 and C8. To promote C6 and C8 production and enrich the functional groups relevant to process performance, we reduced the hydraulic retention time (HRT). HRT refers to the average time soluble compounds reside in the bioreactor. Shortening the HRT is a common operation-based strategy for increasing C6/C8 production [12–16] and a key factor influencing microbial diversity [17]. It is relevant to the

92 microbial growth rate in reactors without biomass retention, and it affects biomass concentration
93 and community composition [18]. Following variations in diversity induced by HRT reduction,
94 we tested if productivity and yield of the target products (C6 and C8) could be predicted by using
95 machine learning. To provide insight into the dynamics of community structure and function, we
96 measured process performance and collected samples for community analysis using high-
97 throughput sequencing of the 16S rRNA gene. Community analysis using 16S rRNA amplicon
98 sequencing data combined with environmental variables can reveal relationships between
99 microbial communities and ecosystem functioning. For example, Werner et al. demonstrated
100 strong relationships between the phylogenetic community structure, reflected by time-resolved
101 16S rRNA amplicon data, and the methanogenic activity in full-scale anaerobic digesters, by
102 applying constrained ordination [19].

103 Predictive analytics using machine learning has shown promise in microbiota-based
104 biotechnologies [6,20,21]. Identification of bioindicators based on microbial community data is
105 an important application of machine learning predictive models [22]. Different machine learning
106 algorithms, such as linear regression [23], support vector machine [24] and random forest
107 regression [25] have been used in microbiome studies. Our machine learning analysis consisted
108 of the identification of the Amplicon Sequence Variants (ASVs) that were relevant to community
109 dynamics caused by HRT reduction and the prediction of C6/C8 production based on the
110 selected ASVs (hereafter, HRT bioindicators). To determine the HRT bioindicators heuristically
111 we used the ASVs as features to predict the target HRT. We first used the microbiome automated
112 machine learning pipeline (hereafter, mAML) [26] to test several different algorithms on our
113 dataset for microbiome-based classification tasks. Once we had prediction accuracies from the
114 different tested algorithms, we selected the algorithm with the highest prediction accuracy that

115 can rank feature relevance. Since we want to gain insight into our data via the learned
116 relationship between feature and target variable, it is crucial that the selected algorithm for
117 suggesting bioindicators not only demonstrates high prediction accuracies but also is
118 interpretable and can rank feature relevance. After determining the HRT bioindicators, we
119 created C6/C8 production regression models using the selected ASVs. It is important to mention
120 that our dataset is imbalanced regarding the number of samples from the different HRT. The
121 dataset consists of 54 samples: 14 from HRT 8 days and 40 from HRT 2 days. Imbalanced
122 datasets can create a bias to the learning task, prioritising the prediction of the majority target.
123 Consequently, for the creation of the C6/C8 production regression model, we also determined the
124 differences in the predictive performance of the original (unbalanced) datasets and of datasets
125 that were balanced by oversampling to verify if our models can handle the imbalance found in
126 our data. Finally, we used *k*-fold cross-validation to estimate the stability of the model.
127
128 **Results**
129
130 **Effects of HRT decrease on process performance and microbial diversity.** The progressive
131 HRT decrease from 8 to 2 days increased the C6 and C8 productivities and yields in two
132 independent bioreactors (Figure 1). We first shortened the HRT to 6 days and then to 4 days in
133 bioreactor A, which allowed the reactor microbiota to adapt to the new conditions and improved
134 productivities of C4, C6 and C8 (Figure 1a). Further HRT decrease to 2 days confirmed the
135 increasing trend in productivity. At the end of the 2-day HRT period in bioreactor A, we
136 achieved the highest productivities (mmol C L^{-1} d^{-1}) of C4, C6 and C8 up to 115.0, 64.1 and 5.9,
137 respectively. To confirm the observed effects of HRT shortening on the CE process and reactor

138 microbiota, we executed a fast transition mode in bioreactor B and generated a different dataset
139 from the parallel system. Comparable increases in productivity were observed (Figure 1b). We
140 obtained maximum productivities (mmol C L^{-1} d^{-1}) of C4 up to 102.4, C6 up to 62.9 and C8 up to
141 7.0. The C6 and C8 yields (in terms of C mole product to consumed substrate ratio) increased
142 along with decreasing HRT at the cost of C4 yield (Figure 1 and Additional file 1: Table S1).
143 Our results suggest that the shorter HRT favoured lactate-based CE producing C6 and C8 over
144 C4 production. C4 can be produced by CE of acetate but also from sugars by butyric acid
145 fermentation [27]. In both bioreactors at 2-day HRT, a temporary accumulation of lactate was
146 observed that coincided with fluctuations of the C4, C6 and C8 production (Figure 1a). Lactate
147 concentrations were negatively correlated with C4 concentrations (Spearman Rho = -0.90, $P <$
148 0.05) and C6 concentrations (Rho = -0.89, $P < 0.05$), which reflects how lactate was produced
149 and converted by the reactor microbiota. The HRT reduction resulted in higher gas production
150 and hydrogen content (Additional file 1: Figure S1). Besides, an increase in cell mass production
151 (Additional file 1: Figure S2) suggests a facilitating effect of short HRT on the growth of
152 enriched populations with desirable activities, i.e., more biocatalysts were available in the high
153 C6/C8 production phase.

154 Decreasing the HRT affected the composition and diversity of the reactor microbiota. Changes in
155 the relative abundance of ASVs categorised from phylum to genus between the HRT of 8 days
156 and 2 days are shown in Additional file 1 (Figure S3). Alpha diversity metrics showed
157 significantly lower observed ASV counts (pairwise *t*-test, $P < 0.05$) and higher Shannon index
158 values (pairwise *t*-test, $P < 0.05$) for HRT of 8 days compared with 2 days (Additional file 1:
159 Figure S4). Beta diversity analysis revealed a significant difference between the communities at
160 different HRTs (PERMANOVA; Pseudo-$F = 103.1$, $P < 0.001$) but no significant difference

between the communities in both reactors at the same HRT (Pseudo-F = 3.3, P > 0.05) (Figure 2).

Selection of HRT bioindicators. To determine HRT bioindicators, we used HRT of 8 days and 2 days as classes and relative abundances of ASVs as features. To choose the most fitting machine learning algorithm for our dataset, different algorithms integrated into the mAML automated machine learning pipeline [26] were tested heuristically. We selected random forest since it can rank feature relevance and it showed the highest prediction accuracies during the 5-fold cross-validation process (Additional file 2). We measured the prediction strength of our models in two folds. First, we trained the models using the data from bioreactor A and then tested them using bioreactor B. After we trained the models using the data from bioreactor B and tested them using bioreactor A. We selected the 15 top-ranked ASVs that gave the best discrimination between the HRT phases, based on higher than 1% of the mean decrease in Gini scores for both reactors in the prediction accuracy of HRT. The 15 most relevant ASVs to identify HRT changes were defined as "A- or B-HRT bioindicators", potentially reflecting the key species correlating with HRT changes in either bioreactor (feature importance in Figure 3). The two bioreactors shared 11 HRT bioindicators assigned to nine different genera.

Prediction of process performance. To answer the question whether HRT bioindicators can be used to predict process performance in terms of C6 and C8 productivity, we performed a regression analysis. We created regression models using the dataset with the original distribution of samples, i.e., 14 samples from HRT 8 days and 40 samples from HRT 2 days, equally divided among the two different bioreactors. We also created regression models using artificially

184 balanced datasets. We used the Synthetic Minority Oversampling Technique (SMOTE) method
185 to oversample the training datasets to have 100 samples with a balanced distribution of the two
186 HRT classes. The datasets from bioreactors A and B were trained and tested independently.
187 Consequently, we had the following experimental configuration: models were trained with the
188 original dataset from bioreactor A/B and tested with the samples from bioreactor B/A; models
189 were trained with the oversampled dataset from bioreactor A/B and tested with the samples from
190 bioreactor B/A. Finally, all created models were evaluated with 5-fold cross-validation.

191 HRT bioindicators were first chosen as features to train the models. Considering that community
192 assembly is affected by time, we then determined the 15 ASVs most relevant to each non-HRT
193 process parameter (i.e., concentrations of lactate, C4, C6 and C8; productivities and yields of C4,
194 C6 and C8; hereafter, non-HRT bioindicators). Initially, we trained regression models using
195 three different machine learning algorithms: linear regression algorithm, support vector machine
196 with radial kernel, and random forest for regression. We used root mean squared errors (RMSE)
197 as the evaluation metric, and the results are visualised as boxplots in Additional file 1 (Figure S5
198 for the HRT bioindicators and Figure S6 for the non-HRT bioindicators). The random forest
199 regression algorithm performed overall better than linear regression and support vector machine
200 with radial kernel. When using the HRT bioindicators as features for the regression, the random
201 forest algorithm had the lowest RMSE median in 7 out of the 8 tested configurations, as shown
202 in Additional file 1 (Figure S5). In addition, the model trained with random forest showed
203 consistency when comparing its performance in the original and the balanced datasets, which
204 indicates that this algorithm is able to handle the imbalance present in our dataset. Therefore, the
205 random forest for regression algorithm was selected as the best algorithm to determine HRT

bioindicators. In our case, random forest could explain more than 80% of the variance in C6 and C8 productivities (Additional file 1: Tables S2-S3).

Using the selected random forest for regression algorithm, we evaluated its prediction performance by comparing the predicted and measured values of process parameters. The average relative root mean square error (RRMSE) for the predictions made using the HRT bioindicators was 4.6% (Figure 4), and the average RRMSE for the predictions made using the non-HRT bioindicators was 5.8% (Additional file 1: Figure S7). We further tested samples in all HRT phases with HRT and non-HRT bioindicators. The C6 and C8 predicted productivities in all cases showed RRMSE below 7.2% (Additional file 1: Figures S8 and S9). Therefore, we considered HRT bioindicators irrespective of time as the ASVs presented in HRT bioindicators and not in non-HRT bioindicators (feature importance in Additional file 1: Figures S10 and S11). Interestingly, the same four ASVs assigned to the genera *Olsenella*, *Lactobacillus*, *Syntrophococcus* and *Clostridium* IV were identified for C6 and C8 productivity (Figure 5). We thus hypothesise that species represented by these four ASVs determined the increased C6/C8 productivities in the CE process manipulated by changing operational conditions, i.e. shortening the HRT.

Functional role of HRT bioindicators. Genomic information on the species of HRT bioindicators indicated their roles in driving the catabolism of xylan and lactate to C6/C8 (Figure 6). Among 108 metagenome-assembled genomes (MAGs; dereplicated into 29 species; Figure 7 and Additional file 3), we recovered 12 species with similar phylogenies as the four genera representing the HRT bioindicators (Table 1). In view of the fermentation process, we annotated the genetic potential for xylan hydrolysis, xylose fermentation and CE with lactate (Additional

file 1: Figure S12 and Additional files 4-7). Specifically, *Clostridium* IV species were reported as lactate-based chain-elongating bacteria [28]. Our results suggest that four *Clostridium* IV species (*Acutalibacteraceae* spp. according to GTDB-Tk) can convert lactate to C6/C8. Two *Syntrophococcus* species (*Eubacterium*_H spp. according to EZBioCloud [29]) are potential C6/C8-producers as they hold complete gene sets encoding enzyme complexes that catalyse CE reactions. This genetic potential was also found in genomes of closely related *Syntrophococcus* species (*Eubacterium cellulosolvens* according to EZBioCloud; Additional file 7), which was not described before. Lactate formation from xylose by lactic acid bacteria can enhance CE by providing additional electron donors [30–34]. A recent study reported an enriched community dominated by *Lactobacillus* and chain-elongating species, and their co-occurrence suggested lactate produced by *Lactobacillus* to be a key intermediate for C6/C8 production [35]. Network analysis of our previous study [10] revealed the co-occurrence of *Olsenella* with potential chain-elongating species. Species of *Lactobacillus* and *Olsenella* are potential xylose-consuming lactate producers (Figure 6b). Genes encoding xylanases were not found in *Lactobacillus* MAGs but in those assigned to other bioindicators (Figure 6a). Taken together, the delineated synergy effects between these bioindicator species suggest a division of labour with mutual benefits, converting xylan and lactate to C6/C8. A correlation network shows HRT, C6 and C8 productivity being the most highly connected nodes (Additional file 1: Figure S13). Their co-occurrence with ASVs assigned to *Clostridium* IV, *Olsenella* and *Syntrophococcus* indicates strong associations among these taxa, the changed environment and corresponding functions. The predictability of C6 and C8 productivities was relatively poor when using only the four HRT bioindicators irrespective of time (Additional file 1: Figure S14). Besides, we found redundancy in the main functions of catabolising xylan and lactate to C4, C6 and C8 (Figure 6), with the

relevant HRT bioindicators increasing in relative abundances (Additional file 1: Figure S15). Thus, the involved metabolic pathways seem to be strongly coupled to HRT decrease. The genetic potential overlaps with that of other distinct taxa of the reactor microbiota, suggesting that HRT bioindicators might be key species of the process, but ecological interactions with other species are critical to ensure the C6/C8 production (functional annotations of xylose fermentation and chain elongation in Additional files 6-7).

Discussion

Bioreactor performance and community dynamics. Continuous reactor systems maintain cultures in a specific growth rate and physiological state [36]. Therefore, these systems are perfect for the exploration of CE as a biotechnological platform for continuous production of medium-chain carboxylates [9]. In this study, we used continuous anaerobic bioreactors with the enriched chain-elongating microbiota [10] as model ecosystems. Two reactors were operated in parallel starting from one inoculum, thus representing biological replicates, and with frequent sampling over 211 days. We demonstrated that shortening the HRT from 8 to 2 days improved C6/C8 productivity and caused specific shifts in the microbial community in both reactors independently of the temporal scheme applied for HRT reduction (i.e. gradual decrease vs. fast transition mode). As we had stable biomass concentrations and detected certain species at all times, we can make sure that these species were growing in each bioreactor since otherwise they would have been washed out from the reactor microbiome. Using multivariate analysis, we demonstrated that the microbial communities established at 8 days HRT were different from those at 2 days HRT. These analyses also showed that the microbial communities sampled from

the two reactors at the corresponding HRT regime were not significantly different (PERMANOVA, $P < 0.05$). Commonly only two lab-scale reactors are run in parallel for such long-term experiments with complex reactor microbiomes [35,37–40]. In contrast to natural ecosystems with their spatial and temporal heterogeneities and uncontrollable environmental factors, bioreactors represent highly controlled model ecosystems that can be sampled at high frequency over long experimental periods, thereby accounting for stochastic effects despite the comparably low number of biological replicates. The obtained time series data are robust and have been used, for instance, to explore pH effects on the CE process [41] and to unravel long-term successional patterns of community assembly in anaerobic processes [42].

Evaluation of the machine learning approach. Machine learning methods can simultaneously incorporate the relative abundances of multiple ASVs and their context-dependency, surpassing traditional statistical approaches that consider each ASV in isolation (e.g., the empirical Bayes moderated t-statistics) [43]. Multivariate analysis has been shown to enable superior performance compared to individual analysis in the context of sensitivity, specificity, and robustness, as it considers potential synergies between the features [44]. Therefore, we used a machine learning approach based on the retrieved 16S rRNA ASVs in two steps of the study: to identify potential bioindicators of HRT, and to create predictive models of *n*-caproate and *n*-caprylate productivities.

For the identification of potential bioindicators, it is necessary to assess the value of the features from the microbiome in an unbiased way – identifying not only their statistical significance but also their prediction accuracy on independent samples [45]. Consequently, to increase the generality of our approach and to reduce any potential bias present in the samples, we

systematically used samples from one bioreactor for training the machine learning models while using the samples from the other bioreactor for testing the model. On the other hand, deploying a machine learning solution is not trivial. To avoid over-optimistic results, it is important to consider the distribution and format of the training data as well as the intrinsic differences of the algorithms themselves [46].

When searching for the optimal manner of dealing with our data, we faced two potential problems: our dataset class distribution is imbalanced concerning the HRT classes (40 samples from 2 days HRT and 14 samples from 8 days HRT), and the total number of samples we have, which is 54, may be limiting to train a robust model. Most machine learning algorithms evaluate themselves during the learning process by comparing the predicted target with the original labelled sample. This creates a bias in the algorithms towards the majority target [47]. In addition, training models with small datasets may create overfitted models that are overly sensitive to outliers and noise. In this work, we tackled the imbalance and the limited number of samples in our data by pre-processing our dataset to generate new samples using the SMOTE method.

Finally, a validation strategy was also integrated into our machine learning pipeline. Validation is one of the most important techniques when creating a generalised model since it estimates the stability of the model when dealing with new data. The validation approach we used is the *k*-fold cross-validation. The general idea of using *k*-fold cross-validation was to train our model with a selected group of samples from our data and validate it with the remaining samples, rotate the training and validation groups *k* times until we used all samples to train a model, and all samples to validate a trained model. By letting us use all the data to train different models, this approach provides much more confidence in the results [48].

Initially, we wanted to determine potential HRT bioindicators. Therefore, the initial step of our machine learning pipeline was to heuristically try several different classification algorithms to determine which of them can better differentiate 2 days HRT and 8 days HRT. To do so, we used the mAML pipeline to systematically create classification models using several tree-based and non-tree-based classifiers. Most of the algorithms had more than 90% classification accuracy. This indicates that the microbiome composition of 2 days HRT and 8 days HRT should be considerably different, and thus directly divisible. To select an algorithm, however, we also considered the ability of the algorithm to rank feature relevance, since we wanted to select the most important ASVs to differentiate the target HRT. Random forest has been shown to run efficiently and accurately on high-dimensional datasets with multi-features by constructing an ensemble of decision trees [49]. Further, it avoids overfitting through the integration of out-of-bag estimates [49]. Finally, other studies that used 16S rRNA sequencing data in machine learning solutions also reported random forest to show good prediction performance [43,50,51]. For these reasons, we selected the random forest algorithm to extract HRT bioindicators.

Once we selected the potential HRT bioindicators, we wanted to develop regression models to predict *n*-caproate and *n*-caprylate productivities. Our machine learning solution for creating the regression models attempts to take into consideration all the potential problems mentioned (i.e., selecting an adequate algorithm, dealing with an imbalanced dataset and potentially insufficient number of samples, avoiding overfitting, and increasing the generality of the model). We evaluated three different regression algorithms with different biases: linear regression, support vector machine, and random forest regression algorithm. In all cases, we balanced our dataset and increased the number of samples using the SMOTE technique. Boxplots were created to visually interpret the results of the 5-fold cross-validation.

For comparing the results from the models created with the original and balanced datasets, for instance comparing Figure S5a and S5b, we can see that the RMSEs of the models trained with the balanced datasets are lower than the ones trained with the original dataset. This indicates that balancing and increasing the number of samples in our dataset generated a more precise and stable prediction model, being more suitable for predictions of new data. In addition, the random forest regression produced more consistent results with lower error rates when compared with the other algorithms. Thus, we selected random forest regression to generate our predictive models, as it has shown to better handle imbalance issues and being more stable across the tests. However, random forest is not the only machine learning algorithm used for predictive analytics in microbiome studies. For example, with an integration of the phylogenetic tree information into the predictive framework, the recently proposed phylogeny-regularised sparse generalised linear model [52] and regression model [53] showed superior prediction power in real microbiome dataset applications. By using human gut microbiome data for continuous age prediction, the so-called glmmTree model achieved the best performance as indicated by the highest R^2 of 70% and the lowest predicted mean square error of a median value 1.3, with a 5-fold cross validation being applied [52]. The random forest algorithm used in this study achieved results comparable to the glmmTree model with R^2 over 80%.

Function of bioindicator species in chain elongation. Mining the functional potential of MAGs affiliated to bioindicators may indicate key functions of these species in the CE process. In particular the MAGs of the lactate-based CE species such as *Clostridium* IV revealed all genes necessary for lactate oxidation and CE by reverse β-oxidation. To validate this hypothesis, we also annotated the genome of the chain-elongating *Ruminococcaceae* bacterium CPB6 affiliated

to *Clostridium* IV [28], which contained complete gene sets encoding enzyme complexes for converting lactate to C6. Interestingly, our results revealed novel species with the genetic potential for chain elongation. Our results may guide other researchers studying CE to characterise novel chain-elongating bacteria in previously reported CE microbiomes.

Here we used metagenomics to unravel the function of key species in CE that were inferred from 16S rRNA sequencing data. This is more reasonable than inferring the function of species only based on the 16S rRNA sequencing data, but the genetic potential alone does not guarantee that the respective metabolic process is actually performed [54]. Therefore, follow-up studies involving multi-omics are necessary to verify if the genetic potential found in the MAGs corresponds with active pathways. Beside multi-omics experiments, the novel genetic information related to the CE process could be validated in wet-lab experiments using defined mixed cultures of isolated strains representing the bioindicator species [55]. By constructing synthetic microbial consortia with different combinations of those representative bioindicator species and monitoring their growth and metabolic behaviour under controlled conditions, mechanistic and metabolic modelling could be used to verify the ability of our machine learning framework to predict ecophysiological functions from 16S rRNA sequencing data.

Engineering microbial communities for bioprocesses with distributed pathways. In engineered and natural ecosystems, phylogenetic diversity can be linked to ecosystem processes in which microbial communities perform key functions [56]. The machine learning approach used in the current study enabled the quantitative prediction of community functioning (i.e., CE) in the anaerobic bioreactor system (Figure 8). Converting xylan and lactate to medium-chain carboxylates is a complex metabolic process consisting of mainly four functions, i.e. xylan

hydrolysis, xylose fermentation, C4 formation from lactate and acetate, and CE with lactate producing C4, C6 and C8, with more than 30 enzymes being involved. We showed that alternative pathways can be used for this complex conversion (Additional file 1: Figure S12). Because of this complexity, it is likely that the observed increase in C6/C8 productivity after shortening HRT from 8 to 2 days was not driven by a single microorganism but by the joint effort of multiple species within our bioreactors. However, not all species in the bioreactor were directly involved in CE. Our feature selection approach helped us identify the species linked to metabolic pathways potentially involved in CE. This was possible because we included quantitative metadata such as time series data of substrate and product concentrations, which facilitated to filter species linked to the CE process. A similar analysis identified key species that could predict the overall quality of soils [25]. In the latter study, the authors showed that using the indicators of soil bacterial community associated to metadata of soil physicochemical variables facilitated to predict the soil quality with 50–95% accuracy [25].

We also provided new biological insights into the reactor microbiomes of lactate-based CE. The importance of *in situ* lactate formation in the lactate-based CE process has been emphasised by several studies [30–35]. Our results indicated that species of the genera *Lactobacillus* and *Olsenella* were potential xylose degraders but *Lactobacillus* cannot utilise the polysaccharide xylan due to the lack of genes encoding xylanases. This result indicates different functional roles of lactic acid bacteria in the degradation of biomass residues containing hemicellulose, which was reported to be more degradable than cellulose during acidogenic fermentation of maize silage [57]. These new insights into the microbial ecology of the CE process may open doors for further valorisation of carbohydrate-rich waste streams. For example, bioaugmentation of xylan-hydrolysing lactic acid bacteria such as *Olsenella* species in CE communities may optimise the

413 breakdown of hemicellulosic compounds. In addition, we demonstrated that C4 is not only
414 produced by CE of acetate but also from xylose by butyric acid fermentation [27], which
415 competes with CE in the recovery of carbon from sugars. This xylose fermentation to C4 was
416 also described as competing process in other CE studies [10,58]. Currently it is still a challenge
417 to steer the CE community functioning to only medium-chain carboxylates in the mixed culture
418 fermentation, but the direction of creating synthetic microbial consortia with modularity (e.g.,
419 spatial niches) could be a wise option to mediate a multi-step bioprocess and to utilise metabolic
420 diversity in any single reactor system [59].

421 In our engineered ecosystems with well-controlled conditions (temperature, pH and no
422 immigration of other microbes; Figure 8a), HRT was the most influencing factor controlling
423 community assembly (Figure 8b). However, we cannot exclude the impact of other deterministic
424 factors like microbial interactions within temporal patterns, particularly for such a long-term
425 reactor experiment. When the random forest regression models took time instead of HRT into
426 account, the results indicated that the non-HRT bioindicators might result from the intrinsic
427 community dynamics alone. Thus, prediction results of the HRT bioindicators can be biased by
428 these autoregressive data present in time series. Even though the HRT bioindicators irrespective
429 of time seem to be key species for the increase in C6/C8 productivity caused by HRT decrease,
430 we cannot ignore the contribution of the non-HRT bioindicators to community assembly and
431 functioning, particularly with functional redundancy shown in the main functions of the CE
432 process. Therefore, effects of compositional stochasticity on community assembly also need to
433 be considered [60,61]. Further studies on these ecological principles will help manage reactor
434 microbiota towards beneficial traits, such as high specificities for C6/C8 production.
435

Conclusions

The continuous reactor systems with enriched communities facilitated the selection of reactor microbiomes with desired CE functions (i.e., high C6 and C8 productivities). We demonstrated that 16S rRNA amplicon sequencing data can be used to predict CE process performance quantitatively (> 90% accuracy). The described machine learning framework (Figure 8c) may be suitable for other ecosystem processes and more complex communities. For that, it would be necessary to design experiments with (i) sufficient temporal and/or spatial resolution, (ii) parallel sampling for amplicon sequencing data and metadata from desired ecosystem processes, and (iii) correlation of phylogenetic diversity with the ecosystem processes. Our approach was based on phylogenetic diversity (relative ASV abundances) that in some ecosystems may correlate with ecosystem processes where microbiota perform key functions. Due to the use of unbalanced datasets, the high dimensionality and more direct link with different ecosystem processes found in omics data, our general methodology can be adapted to other data types, including functional genes, transcripts, proteins or metabolites. Our approach opens new doors for prediction and hypothesis testing in microbiome research. Further studies are needed to reveal which data types reflect different ecosystem processes and communities with different levels of complexity.

Methods

Reactor operation and monitoring of process parameters and community composition. The inoculum was initially taken from a continuous lab-scale bioreactor that produced C6 and C8 by anaerobic fermentation of lactate-rich corn silage [11]. Enrichment was performed in a reactor

459 that was daily fed with mineral medium (pH 5.5; Additional file 1: Table S4) containing water-
460 soluble xylan (more than 95% xylooligosaccharides, from corncob; Roth, Karlsruhe, Germany)
461 and lactic acid (85%, FCC grade; Sigma Aldrich, St. Louis, USA) as defined carbon sources and
462 produced C4, C6 and C8 over 150 days [10]. For the present study, two 1-L bioreactors (A and
463 B; BIOSTAT® A plus, Sartorius AG, Göttingen, Germany) were filled up with 0.5 L of the
464 enriched culture. Both bioreactors were daily fed with 0.125 L medium containing 1.47 g lactic
465 acid and 1.25 g xylan, without withdrawing effluent. After four days the contents of both
466 bioreactors were mixed by pumping them three times from bioreactor A to B and back while
467 keeping anoxic conditions. Eventually, they were equally distributed to both bioreactors, which
468 is considered the starting point (day 0) of the experiment.

469 We employed semi-continuous stirred tank reactors for anaerobic fermentation, which were
470 operated at 38 ± 1°C and constantly stirred at 150 rpm. The pH of the reactor broth was
471 automatically controlled at 5.5 by addition of 1 M NaOH. For each bioreactor, the produced gas
472 was collected in a coated aluminium foil bag that also served for compensating underpressure in
473 the reactor system. The bag was connected after a MilliGascounter® (MGC-1; Ritter, Bochum,
474 Germany) that measured on-line the volume of the produced gas. A gas-sample septum was
475 placed in the gas pipe of each bioreactor.

476 In the beginning, both bioreactors were operated as duplicates with an equal HRT of 8 days. For
477 daily feeding, 1.47 g lactic acid and 1.25 g xylan were supplied in mineral medium. After 51
478 days, we gradually decreased the HRT of bioreactor A from 8 days to 6 days, and further to 4
479 days and 2 days while operation of reactor B was continued at HRT of 8 days as a control as
480 shown in Additional file 1: Table S5. Next, we shortened the HRT of bioreactor B from 8 days to
481 2 days in a fast transition mode and with the same substrate load as in bioreactor A, in order to

reproduce the HRT transition in the second reactor. Considering the effect of time on community assembly, we conducted unequal HRT changes in the two bioreactors and aimed to delineate the model prediction strength with the two different datasets. Finally, both bioreactors were operated in parallel at an HRT of 2 days until day 211.

Gas samples were taken through the septum twice per week. Samples for measuring optical density (OD) and for DNA extraction were collected twice per week from the reactor effluent. Concentrations of xylan, carboxylates and alcohols were measured in the effluent supernatants [10]. In total, effluent samples were collected on 59 time points for each bioreactor. At the beginning and the end of the experiment, pelleted biomass from the effluent was used to determine the cell dry mass as previously described [10]. For microbial community analysis, pelleted cells from 2 mL effluent were washed with 100 mM Tris-HCl pH 8.5 and stored at -20˚C until DNA extraction.

Analytical methods. Daily produced gas volume was monitored with the MGC-1 and normalised to standard pressure and temperature [30]. Gas composition (H_2, CO_2, N_2, O_2 and CH_4) was determined by gas chromatography in triplicate [62]. Concentrations of carboxylates and alcohols were analysed in triplicate by gas chromatography [10]. Concentration of xylan was measured by a modified dinitrosalicylic acid reagent method [10]. Cell mass concentration was calculated from OD values that were correlated with the cell dry mass [10]. The calculated mean correlation coefficients were 1 OD_{600} = 0.548 g L^{-1} for bioreactor A and 1 OD_{600} = 0.537 g L^{-1} for bioreactor B.

Microbial community analysis. Total DNA was isolated from frozen cell pellets sampled twice

505 per week using the NucleoSpin® Microbial DNA Kit (Macherey-Nagel, Düren, Germany).
506 Methods for DNA quantification and quality control were as described previously [63]. For high-
507 throughput amplicon sequencing, V3-V4 regions of the 16S rRNA genes were PCR-amplified
508 using primers 341f and 785r [64]. Sequencing was performed on the Illumina Miseq platform
509 (Miseq Reagent Kit v3; 2 × 300 bp). A total of 12,168,404 sequences ranging from 57,612 to
510 389,963 pairs of reads per sample (mean: 135,205; median: 122,367) were obtained.

511 The demultiplexed sequence data were processed with the QIIME 2 v2019.7 pipeline [65] using
512 the DADA2 plugin [66]. The DADA2 parameters were set as follows: trim-left-f 0, trim-left-r 0,
513 trunc-len-f 270, trunc-len-r 230, max-ee 2 and chimera-method consensus. A total of 4,194,700
514 sequences ranging from 13,518 to 138,498 reads per sample were retained, with a mean of
515 46,608 reads per sample. The generated feature table indicates the frequency of each ASV
516 clustered at 100% identity. Taxonomic assignment was done with a naïve Bayes classifier trained
517 on 16S rRNA gene sequences of the database MiDAS 2.1 [67], and curated using the RDP
518 Classifier 2.2 with a confidence threshold of 80% [68]. For downstream analyses, ASVs of all
519 samples were rarefied to a sequencing depth of 13,518 reads (rarefaction curve reached the
520 plateau, Additional file 1: Figure S16). We obtained a total of 71 unique ASVs in 90 samples.

521 Alpha diversity based on rarefied ASV data was evaluated by the observed ASV counts and the
522 Shannon index [69], which were determined using the R package phyloseq v1.30.0 [70].
523 Dissimilarities in bacterial community composition (beta-diversity) were calculated using Bray-
524 Curtis distance [71] based on rarefied ASV abundances and visualised as nonmetric
525 multidimensional scaling (NMDS) plots. Statistical analyses of beta-diversity results were
526 performed using permutational multivariate analysis of variance (PERMANOVA) [72] in the R
527 package "vegan" (v2.5.6, "adonis" function, Monto-Carlo test with 1000 permutations); *P* values

were adjusted for multiple comparisons using the false discovery rate (FDR) method [73].

Network analysis. The co-occurrence network analysis was performed using the method described by Ju et al. [74]. Briefly, we constructed a correlation matrix by computing possible pairwise Spearman's rank correlations using the rarefied ASV abundances and abiotic parameters (HRT; concentrations of C4, C6, C8 and lactate; productivities and yields of C4, C6 and C8). Correlation coefficients below -0.7 or above 0.7 and adjusted *P*-values (FDR method) lower than 0.05 were considered statistically robust. Network visualisation and topological features analysis were conducted in Gephi (v0.9.2) [75].

16S rRNA phylogenetic analysis. The 16S rRNA gene sequences of ASVs were aligned using the SINA alignment algorithm [76] via the SILVA web interface [77]. We additionally used SINA to search and classify the sequences with the least common ancestor method based on the SILVA taxonomy. For each query sequence, the minimum identity was set to 0.95 and the five nearest neighbours were considered. The tree was reconstructed based on the aligned sequences and their neighbours, with RAxML using the GTRCAT model of evolution. Later only ASV species of this study were kept in the generated tree for an easier viewing. The tree was visualised using iTOL [78].

Metagenomic analysis. Six samples from the enrichment period were selected for whole-genome sequencing, which was performed by StarSEQ GmbH (Mainz, Germany) using the Illumina NextSeq 500 system (NEBNext Ultra II FS DNA library prep kit; 2 × 150 bp) with at a minimum of 20 million reads per library generated. Quality check and reads trimming were

551 performed using metaWRAP (v0.7, raw read QC module) [79] and TrimGalore (v0.4.3) [80].
552 Reads of human origin were discriminated from microbial reads using BMTagger (v3.101) [81].
553 All adapters were removed and the resulting reads were assembled using metaSPAdes (v3.11.1)
554 [82]. Paired-end reads were aligned back to the assembly using BWA (v0.7.15, mem algorithm)
555 [83]. Binning of assembled contigs was performed using the metaWRAP modules metaBAT
556 (2.12.1) [84], MaxBin (2.2.4) [85] and CONCOCT (1.0.0) [86]. The metaWRAP-Bin_refinement
557 module was applied to separate the overlaps between two bins. Quality of MAGs was checked
558 using CheckM (v1.0.7) [87]. MAGs were classified in high or medium quality regarding
559 completeness, contamination, quality score (completeness - 5 × contamination) and strain
560 heterogeneity [88]. The following thresholds were used for high quality: quality score > 50,
561 completeness > 80, contamination < 5 and strain heterogeneity < 50; and for medium quality:
562 quality score > 50, completeness > 50 and contamination < 10. One bin with lower quality was
563 removed from the analysis. The taxonomy was assigned using GTDB-Tk (v0.3.2) [89]. Genome
564 metrics were calculated with the statswrapper tool in the BBTools suite [90]. A phylogenomic
565 tree based on Mash distances was generated with Mashtree (V1.1.2) [91] and visualised in iTOL
566 [78]. Miscellaneous visualisations of the dataset metrics were performed in R with the packages
567 ggplot2 (v3.3.0) and DataExplorer (v0.8.1). Species differentiation was performed using fastANI
568 [92] and aniSplitter.R (http://github.com/felipborim789/aniSplitter/). Genomes were annotated
569 with Prokka (v1.14.6) [93]. Functional annotation of genes relevant to xylan hydrolysis, xylose
570 fermentation and chain elongation was curated using Swiss-Prot, COG and GenBank [94–96].
571 Default settings were chosen for all tools unless otherwise specified.
572
573 **Determination of bioindicators of HRT changes.** To select the machine learning algorithm for

differentiating the HRT phases of 8 days and 2 days, the mAML automated machine learning pipeline [26] was used to heuristically test several different algorithms on our microbiome data. We selected the algorithm with the highest prediction accuracy that can rank feature relevance. ASV relative abundances were used as features to train and test the different classifiers included in the mAML pipeline. After the initial algorithm selection process, the random forest algorithm (randomForest R package, v4.6-14) [97] was chosen to determine the HRT bioindicators due to its high accuracy and ability to rank feature relevance. Considering how we replicated the HRT changing mode in both bioreactors (Additional file 1: Table S5), the whole operation period was divided into four sampling intervals: 0-50 days, 51-100 days, 101-140 days and 141-211 days. Based on the results of community analysis, we chose the ASV data of both bioreactors in the sampling intervals of 0-50 days and 141-211 days to determine the HRT bioindicators, and we used data of all samples in the four HRT phases as controls. To delineate the model prediction strength, we trained the classifier with ASV data of one bioreactor and tested in the other bioreactor and vice versa. For random forest classification analysis, importance of the different features (ASVs) was measured by the Gini index (mean decrease in Gini, default in randomForest R package; where larger values indicate a variable to be more important for accurate classification [98]).

The random forest classifier was trained on the training set, with 2,000 trees and 40 variables (with lowest out-of-bag estimated error rates achieved) being selected randomly for each tree. Explained variance (% Var. explained in R) was used to measure the model performance on the training set [97]. We predicted the accuracy by measuring how well the features can classify the HRT phases on the test set [98]. We first computed the feature importance of all 71 ASVs. Then in each step, the ASVs having the smallest importance were eliminated and a new forest was

built with the remaining ASVs. For both bioreactors, the features were selected when their Gini scores were higher than 1% of the sum of the Gini scores of all ASVs (Additional file 1: Table S6). Feature selection based on the random forest classifier with its associated Gini index has shown abilities to identify optimal feature subsets in high-dimensional data [99]. Finally, we selected the 15 top-ranked ASVs leading to the model of smallest error rate for classifying the HRT phases of 8 days and 2 days. In each bioreactor, the 15 ASVs that best discriminated between HRT phases were referred to as A-HRT bioindicators or B-HRT bioindicators (bioreactors A and B, respectively). ASVs common to both sets were defined as HRT bioindicators (workflow of random forest classification in Additional file 1: Figure S17).

Quantitative predictions based on HRT and non-HRT bioindicators. The data of bioreactor A and bioreactor B were used for training and testing the regression models independently. Due to the unbalanced ratio of HRT 8 days (14 samples with 26%) and HRT 2 days (40 samples with 74%), we also created models using balanced training datasets. The artificially balanced datasets were created based on the HRT class information and using the SMOTE technique implemented in the R package UBL (v0.0.6) [100]. The balanced datasets had 52 and 48 samples for HRT 2 days and 8 days, respectively. For the process parameters to be predicted, four training datasets were considered: only with samples from bioreactor A, only with samples from bioreactor B and the balanced version of these two datasets. Initially, three algorithms including linear regression, support vector machine with radial kernel and random forest for regression (implemented in R package ranger, v0.12.1) [101] were employed as a heuristic approach to evaluate their predictive performance based on the metric root mean square error. The training and benchmarking processes were performed using the R package mlr (v2.18.0) [102]. All

algorithms were validated using a 5-fold cross validation approach. We selected the algorithm presenting better overall prediction performance and trained it with another round of 5-fold cross validation. After, the random forest regression analysis was used to predict the process parameters specified as concentrations of lactate, C4, C6 and C8, and productivities as well as yields of C4, C6 and C8. Here, the relevance of the different ASVs to the prediction was determined by residual sum of squares (IncNodePurity, default in randomForest) for the regressions. Explained variance (% Var. explained in R) was used to measure the model performance on the training set [97]. We predicted the accuracy by measuring how well the features can explain the variance of these process parameters on the test set [98]. The hyperparameters of random forest trained models (e.g., number of trees) were tuned heuristically during cross validation.

We performed the quantitative prediction by applying a two-step regression analysis with 5-fold cross-validation (workflow in Additional file 1: Figure S18). First, HRT bioindicators were used to predict the data of different process parameters in the sampling intervals of 0-50 days and 141-211 days. Data of all samples in the four HRT phases were considered as controls. Relative abundance dataset of bioreactor A was used as training set and that of bioreactor B was used as test set and vice versa. Next, considering community assembly caused by time, we determined the ASVs (non-HRT bioindicators) that could predict the numeric values of each process parameter, using data of samples in the intervals of 0-50 days and 141-211 days. For each process parameter, we started with computing the feature importance of all ASVs and further selected the 15 top-rated ASVs as the bioindicators of this non-HRT parameter. Datasets of bioreactors A and B were independently used for training and testing. As controls, we used the non-HRT bioindicators of each parameter to predict the corresponding data of all samples in the

four HRT phases. The final set of ASVs presented in HRT bioindicators and not in non-HRT bioindicators were considered HRT bioindicators irrespective of time.

Evaluation of prediction accuracy. When in both training sets the HRT bioindicators and non-HRT bioindicators explained more than 80% of the variance in a process parameter, we proceeded only with those parameters. To compare the predicted and measured values for these process parameters, we considered the following performance metrics for reflecting the error of the model in predicting consecutive data: RRMSE, cutoff < 10%; R squared, slope and intercept of the least squares line of best fit. The final values of RRMSE were averaged among the 100 random forest replicates, with four ASVs for HRT bioindicators and five for non-HRT bioindicators randomly sampled at each replicate.

List of abbreviations

ASVs: Amplicon Sequence Variants, C4: *n*-butyrate, C6: *n*-caproate, C8: *n*-caprylate, CE: chain elongation, FDR: false discovery rate, GTDB: Genome Taxonomy Database, HRT: hydraulic retention time, MAGs: metagenome-assembled genomes, NMDS: nonmetric multidimensional scaling, OD: optical density, PERMANOVA: permutational multivariate analysis of variance, RRMSE: relative root mean square error, SMOTE: Synthetic Minority Oversampling Technique.

Declarations

Availability of data and materials

All data described in this study are available in the paper or in the Supplementary material. Raw

reads of amplicon sequencing data (ERR4158761 to ERR4158850) and metagenome sequencing data (ERR4183110 to ERR4183115) have been deposited in the European Nucleotide Archive (ENA) under study no. PRJEB38353. The MAGs are publicly available in ENA under the sample accession nos. ERS4594296 to ERS4594324.

Funding

The study was supported by the Initiative and Networking Fund of the Helmholtz Association. B.L. was supported by the China Scholarship Council (# 201606350010). J.S. and U.R. were financed by the Helmholtz Young Investigator grant VH-NG-1248 Micro 'Big Data'. H.S., H.H. and S.K. were financed by the BMBF – German Federal Ministry of Education and Research (# 031B0389B and # 01DQ17016) and the Helmholtz Association (Program Renewable Energies). S.G.S. was the recipient of a PhD scholarship conceded by FCT (PD/BD/143029/2018). J.C.K was supported by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) [2019/03396-9].

Authors' contributions

B.L., H.S., J.S., S.K. and U.R. designed the study and the experiments. B.L. performed the experiments and analysed the reactor data as well as sequencing data. B.L., J.S., J.C.K. and U.R. performed the machine learning analysis. B.L., H.S., J.S., S.G.S., J.C.K., S.K. and U.R. contributed to data analysis and interpretation. H.H. contributed to the discussion of the results. All authors critically contributed to the preparation of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

690 The authors thank Ute Lohse for her technical assistance in amplicon sequencing, and the
691 colleagues from DBFZ Deutsches Biomasseforschungszentrum GmbH for their technical support
692 in analyses of bioreactor process parameters. We thank Rodolfo Brizola Toscan and Felipe
693 Borim Corrêa for their help with data analysis. We also thank Masun Nabhan Homsi for valuable
694 discussions regarding our machine learning analysis.

695

696 **Ethics approval and consent to participate**

697 Not applicable.

698

699 **Consent for publication**

700 Not applicable.

701

702 **Competing interests**

703 The authors declare no competing interests.

704

705 **References**

706

707 1. Banerjee S, Schlaeppi K, van der Heijden MGA. Keystone taxa as drivers of microbiome
708 structure and functioning. Nature Reviews Microbiology. Springer US; 2018;16:567–76.

709 2. de los Reyes FL. Challenges in determining causation in structure-function studies using
710 molecular biological techniques. Water Research. Elsevier Ltd; 2010;44:4948–57.

711 3. Koch C, Müller S, Harms H, Harnisch F. Microbiomes in bioenergy production : From
712 analysis to management. Current Opinion in Biotechnology. 2014;27:65–72.

713 4. Verstraete W, Wittebolle L, Heylen K, Vanparys B, de Vos P, van de Wiele T, et al. Microbial
714 Resource Management: The road to go for environmental biotechnology. Engineering in Life
715 Sciences. 2007;2:117–26.

5. Kleerebezem R, van Loosdrecht MC. Mixed culture biotechnology for bioenergy production. Current Opinion in Biotechnology. 2007;18:207–12.

6. Lawson CE, Harcombe WR, Hatzenpichler R. Common principles and best practices for engineering microbiomes. Nature Review Microbiology. 2019;17:725–41.

7. Goldford JE, Lu N, Bajić D, Estrela S, Tikhonov M, Sanchez-Gorostiaga A, et al. Emergent simplicity in microbial community assembly. Science. 2018;361:469–74.

8. Zuñiga C, Li CT, Yu G, Al-Bassam MM, Li T, Jiang L, et al. Environmental stimuli drive a transition from cooperation to competition in synthetic phototrophic communities. Nature Microbiology. Springer US; 2019;4:2184–91.

9. Angenent LT, Richter H, Buckel W, Spirito CM, Steinbusch KJJ, Plugge CM, et al. Chain elongation with reactor microbiomes: open-culture biotechnology to produce biochemicals. Environmental Science and Technology. 2016;50:2796–810.

10. Liu B, Kleinsteuber S, Centler F, Harms H, Sträuber H. Competition between Butyrate Fermenters and Chain-elongating Bacteria Limits the Efficiency of Medium-chain Carboxylate Production. Frontiers in Microbiology. 2020;11:336.

11. Lambrecht J, Cichocki N, Schattenberg F, Kleinsteuber S, Harms H, Müller S, et al. Key sub-community dynamics of medium-chain carboxylate production. Microbial Cell Factories. BioMed Central; 2019;18:92.

12. Kucek L, Spirito CM, Angenent LT. High n-caprylate productivities and specificities from dilute ethanol and acetate: chain elongation with microbiomes to upgrade products from syngas fermentation. Energy Environ Sci. 2016;9:3482–94.

13. Kucek LA, Nguyen M, Angenent LT. Conversion of L-lactate into n-caproate by a continuously fed reactor microbiome. Water Research. Elsevier Ltd; 2016;93:163–71.

14. Duber A, Jaroszynski L, Zagrodnik R, Chwialkowska J, Juzwa W, Ciesielski S, et al. Exploiting the real wastewater potential for resource recovery – n -caproate production from acid whey. Green Chemistry. Royal Society of Chemistry; 2018;20:3790–803.

15. Grootscholten TIM, Steinbusch KJJ, Hamelers HVM, Buisman CJN. Improving medium chain fatty acid productivity using chain elongation by reducing the hydraulic retention time in an upflow anaerobic filter. Bioresource Technology. Elsevier Ltd; 2013;136:735–8.

16. Nzeteu CO, Trego AC, Abram F, O'Flaherty V. Reproducible, high-yielding, biological caproate production from food waste using a single-phase anaerobic reactor system. Biotechnology for Biofuels. BioMed Central; 2018;11:108.

17. Mansfeldt C, Achermann S, Men Y, Walser JC, Villez K, Joss A, et al. Microbial residence time is a controlling parameter of the taxonomic composition and functional profile of microbial communities. ISME Journal. Springer US; 2019;13:1589–601.

18. Bonk F, Popp D, Weinrich S, Sträuber H, Becker D, Kleinsteuber S, et al. Determination of Microbial Maintenance in Acetogenesis and Methanogenesis by Experimental and Modeling Techniques. Frontiers in Microbiology. 2019;10:166.

19. Werner JJ, Knights D, Garcia ML, Scalfone NB, Smith S, Yarasheski K, et al. Bacterial community structures are unique and resilient in full-scale bioenergy systems. Proceedings of the National Academy of Sciences of the United States of America. 2011;108:4158–63.

20. Oyetunde T, Bao FS, Chen JW, Martin HG, Tang YJ. Leveraging knowledge engineering and machine learning for microbial bio-manufacturing. Biotechnology Advances. Elsevier Inc; 2018;36:1308–15.

21. Lopatkin AJ, Collins JJ. Predictive biology: modelling, understanding and harnessing microbial complexity. Nature Reviews Microbiology. Springer US; 2020;

22. Astudillo-García C, Hermans SM, Stevenson B, Buckley HL, Lear G. Microbial assemblages and bioindicators as proxies for ecosystem health status: potential and limitations. Applied Microbiology and Biotechnology. Applied Microbiology and Biotechnology; 2019;103:6407–21.

23. Bodein A, Chapleur O, Droit A, Lê Cao KA. A Generic Multivariate Framework for the Integration of Microbiome Longitudinal Studies With Other Data Types. Frontiers in Genetics. 2019;10:963.

24. Seshan H, Goyal MK, Falk MW, Wuertz S. Support vector regression model of wastewater bioreactor performance using microbial community diversity indices: Effect of stress and bioaugmentation. Water Research. Elsevier Ltd; 2014;53:282–96.

25. Hermans SM, Buckley HL, Case BS, Curran-Cournane F, Taylor M, Lear G. Using soil bacterial communities to predict physico-chemical variables and soil quality. Microbiome. Microbiome; 2020;8:79.

26. Yang F, Zou Q. mAML: an automated machine learning pipeline with a microbiome repository for human disease classification. Database. 2020;2020:baaa050.

27. Temudo MF, Mato T, Kleerebezem R, Van Loosdrecht MCM. Xylose anaerobic conversion by open-mixed cultures. Applied Microbiology and Biotechnology. 2009;82:231–9.

28. Zhu X, Zhou Y, Wang Y, Wu T, Li X, Li D, et al. Production of high-concentration n-caproic acid from lactate through fermentation using a newly isolated Ruminococcaceae bacterium CPB6. Biotechnology for Biofuels. BioMed Central; 2017;10:102.

29. Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, et al. Introducing EzBioCloud: A taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. International Journal of Systematic and Evolutionary Microbiology. 2017;67:1613–7.

30. Sträuber H, Bühligen F, S Kleinsteuber, Dittrich-Zechendorf M. Carboxylic acid production from ensiled crops in anaerobic solid-state fermentation - trace elements as pH controlling agents support microbial chain elongation with lactic acid. Eng Life Sci. 2018;0:447–58.

31. Xu J, Hao J, Guzman JJL, Spirito CM, Harroff LA, Angenent LT. Temperature-Phased Conversion of Acid Whey Waste Into Medium-Chain Carboxylic Acids via Lactic Acid: No External e-Donor. Joule. Elsevier Inc.; 2018;2:1–16.

32. Scarborough MJ, Lynch Griffin, Dickson Mitch, McGee Mick, Donohue TJ, Noguera DR. Increasing the economic value of lignocellulosic stillage through medium-chain fatty acid production. Biotechnology for Biofuels. BioMed Central; 2018;11:200.

33. Khor WC, Andersen S, Vervaeren H, Rabaey K. Electricity-assisted production of caproic acid from grass. Biotechnology for Biofuels. BioMed Central; 2017;10:180.

34. Andersen SJ, de Groof V, Khor WC, Roume H, Props R, Coma M, et al. A Clostridium group IV species dominates and suppresses a mixed culture fermentation by tolerance to medium chain fatty acids products. Frontiers in Bioengineering and Biotechnology. 2017;5:8.

35. Contreras-Dávila CA, Carrión VJ, Vonk VR, Buisman CNJ, Strik DPBTB. Consecutive lactate formation and chain elongation to reduce exogenous chemicals input in repeated-batch food waste fermentation. Water Research. 2020;1:115215.

36. Vrancken G, Gregory AC, Huys GRB, Faust K, Raes J. Synthetic ecology of the human gut microbiota. Nature Reviews Microbiology. Springer US; 2019;17:754–63.

37. Maus I, Klocke M, Derenkó J, Stolze Y, Beckstette M, Jost C, et al. Impact of process temperature and organic loading rate on cellulolytic / hydrolytic biofilm microbiomes during biomethanation of ryegrass silage revealed by genome-centered metagenomics and metatranscriptomics. Environ Microbiome. 2020;15:7.

38. Detman A, Mielecki D, Pleśniak Ł, Bucha M, Janiga M, Matyasik I, et al. Methane-yielding microbial communities processing lactate-rich substrates: a piece of the anaerobic digestion puzzle. Biotechnol Biofuels. 2018;11:116.

39. Zhu X, Feng X, Liang C, Li J, Jia J, Feng L, et al. Microbial Ecological Mechanism for Long-Term Production of High Concentrations of n-Caproate via Lactate-Driven Chain Elongation. Appl Environ Microbiol. 2021;87.

40. Westerholm M, Müller B, Isaksson S, Schnürer A. Trace element and temperature effects on microbial communities and links to biogas digester performance at high ammonia levels. Biotechnol Biofuels. 2015;8:154.

41. Candry P, Radić L, Favere J, Carvajal-Arroyo JM, Rabaey K, Ganigué R. Mildly acidic pH selects for chain elongation to caproic acid over alternative pathways during lactic acid fermentation. Water Research. 2020;186:116396.

42. Wu L, Yang Y, Chen S, Zhao M, Zhu Z, Yang S, et al. Long-term successional dynamics of microbial association networks in anaerobic digestion processes. Water Research. Elsevier Ltd; 2016;104:1–10.

43. Topçuoğlu BD, Lesniak NA, Ruffin M, Wiens J, Schloss PD. A framework for effective application of machine learning to microbiome-based classification problems. mBio. 2020;11:e00434-20.

44. Fortino V, Wisgrill L, Werner P, Suomela S, Linder N, Jalonen E, et al. Machine-learning–driven biomarker discovery for the discrimination between allergic and irritant contact dermatitis. PNAS. National Academy of Sciences; 2020;117:33474–85.

45. Wirbel J, Zych K, Essex M, Karcher N, Kartal E, Salazar G, et al. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. Genome Biology. 2021;22:93.

46. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making. 2019;19:281.

47. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5:221–32.

48. Bokulich NA, Ziemski M, Robeson MS, Kaehler BD. Measuring the microbiome: Best practices for developing and benchmarking microbiomics methods. Computational and Structural Biotechnology Journal. 2020;18:4048–62.

49. Breiman L. Random forests. Machine Learning. 2001;45:5–32.

50. Zhou YH, Gallins P. A review and tutorial of machine learning methods for microbiome host trait prediction. Frontiers in Genetics. 2019;10:579.

51. Pasolli E, Truong DT, Malik F, Waldron L, Segata N. Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLOS Computational Biology. Public Library of Science; 2016;12:e1004977.

52. Xiao J, Chen L, Johnson S, Yu Y, Zhang X, Chen J. Predictive modeling of microbiome data using a phylogeny-regularized generalized linear mixed model. Frontiers in Microbiology. 2018;9:1391.

53. Xiao J, Chen L, Yu Y, Zhang X, Chen J. A Phylogeny-regularized sparse regression model for predictive modeling of microbial community data. Frontiers in Microbiology. 2018;9:3112.

54. Saraiva JP, Worrich A, Karakoç C, Kallies R, Chatzinotas A, Centler F, et al. Mining Synergistic Microbial Interactions: A Roadmap on How to Integrate Multi-Omics Data. Microorganisms. Multidisciplinary Digital Publishing Institute; 2021;9:840.

55. D'hoe K, Vet S, Faust K, Moens F, Falony G, Gonze D, et al. Integrated culturing, modeling and transcriptomics uncovers complex interactions and emergent behavior in a three-species synthetic gut community. eLife. 2018;7:e37090.

56. Mei R, Liu W-T. Quantifying the contribution of microbial immigration in engineered water systems. Microbiome. Microbiome; 2019;7:144.

857 57. Sträuber H, Lucas R, Kleinsteuber S. Metabolic and microbial community dynamics during
858 the anaerobic digestion of maize silage in a two-phase process. Applied Microbiology and
859 Biotechnology. Springer Berlin Heidelberg; 2016;100:479–91.

860 58. Scarborough MJ, Lawson CE, Hamilton JJ, Donohue TJ, Noguera DR. Metatranscriptomic
861 and Thermodynamic Insights into Medium-Chain Fatty Acid Production Using an Anaerobic
862 Microbiome. mSystems. 2018;3:e00221-18.

863 59. Shahab RL, Brethauer S, Davey MP, Smith AG, Vignolini S, Luterbacher JS, et al. A
864 heterogeneous microbial consortium producing short-chain fatty acids from lignocellulose.
865 Science. 2020;369:eabb1214.

866 60. Chase JM. Stochastic Community Assembly Causes Higher Biodiversity in More Productive
867 Environments. science. 2010;328:1388–91.

868 61. Ofiteru ID, Lunn M, Curtis TP, Wells GF, Criddle CS, Francis CA, et al. Combined niche
869 and neutral effects in a microbial wastewater treatment community. Proceedings of the National
870 Academy of Sciences. 2010;107:15345–50.

871 62. Urban C, Xu J, Sträuber H, dos Santos Dantas TR, Mühlenberg J, Härtig C, et al. Production
872 of drop-in fuel from biomass by combined microbial and electrochemical conversions. Energy
873 Environ Sci. 2017;10:2231–44.

874 63. Lucas R, Kuchenbuch A, Fetzer I, Harms H, Kleinsteuber S. Long-term monitoring reveals
875 stable and remarkably similar microbial communities in parallel full-scale biogas reactors
876 digesting energy crops. FEMS Microbiology Ecology. 2015;91:fiv004.

877 64. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of
878 general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-
879 based diversity studies. Nucleic Acids Research. 2013;41:e1.

880 65. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Chase J, Cope EK, et al. Reproducible,
881 interactive, scalable and extensible microbiome data science using QIIME 2. Nature
882 Biotechnology. 2019;37:852–7.

883 66. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-
884 resolution sample inference from Illumina amplicon data. Nature Methods. 2016;13:581–3.

885 67. McIlroy SJ, Kirkegaard RH, McIlroy B, Nierychlo M, Kristensen JM, Karst SM, et al.
886 MiDAS 2.0: An ecosystem-specific taxonomy and online database for the organisms of
887 wastewater treatment systems expanded for anaerobic digester groups. Database. 2017;2017:1–9.

888 68. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naıve Bayesian classifier for rapid assignment of
889 rRNA sequences. Applied and Environmental Microbiology. 2007;73:5261–7.

890 69. Shannon CE. A Mathematical Theory of Communication. Bell System Technical Journal.
891 1948;27:379–423.

70. McMurdie PJ, Holmes S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE. 2013;8:e61217.

71. Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin. Ecological Monographs. 1957;27:325–49.

72. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecology. 2001;26:32–46.

73. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple Testing. Journal of the Royal Statistical Society B (Methodological). 1995;57:289–300.

74. Ju F, Xia Y, Guo F, Wang Z, Zhang T. Taxonomic relatedness shapes bacterial assembly in activated sludge of globally distributed wastewater treatment plants. Environmental Microbiology. 2014;16:2421–32.

75. Bastian M, Heymann S, Jacomy M. Gephi: An open source software for exploring and manipulating networks. BT - International AAAI Conference on Weblogs and Social. International AAAI Conference on Weblogs and Social Media. 2009;8:361–2.

76. Pruesse E, Peplies J, Glöckner FO. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics. 2012;28:1823–9.

77. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research. 2007;35:7188–96.

78. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic acids research. 2016;44:W242-245.

79. Uritskiy G V., Diruggiero J, Taylor J. MetaWRAP - A flexible pipeline for genome-resolved metagenomic data analysis 08 Information and Computing Sciences 0803 Computer Software 08 Information and Computing Sciences 0806 Information Systems. Microbiome. Microbiome; 2018;6:158.

80. Galore K. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files [Internet]. 2015. Available from: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

81. Rotmistrovsky, K. Agarwala R. BMTagger: best match tagger for removing human reads from metagenomics datasets [Internet]. 2011. Available from: ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/

82. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. MetaSPAdes: A new versatile metagenomic assembler. Genome Research. 2017;27:824–34.

83. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

84. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;3:e1165.

85. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32:605–7.

86. Alneberg J, Bjarnason BS, De Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. Nature Methods. 2014;11:1144–6.

87. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research. 2015;25:1043–55.

88. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology. Springer US; 2017;2:1533–42.

89. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2019;36:1925–7.

90. Bushnell B. BBMap short read aligner, and other bioinformatic tools [Internet]. Available from: http://sourceforge.net/projects/bbmap

91. Katz L, Griswold T, Morrison S, Caravas J, Zhang S, Bakker H, et al. Mashtree: a rapid comparison of whole genome sequence files. Journal of Open Source Software. 2019;4:1762.

92. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications. 2018;9:5114.

93. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.

94. Bateman A. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Research. Oxford University Press; 2019;47:D506–15.

95. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin E V., et al. The COG database: An updated vesion includes eukaryotes. BMC Bioinformatics. 2003;4:41.

96. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Research. 2016;44:D67–72.

97. Liaw A, Wiener M. Classification and Regression with Random Forest. R News. 2002;2:18–22.

98. Huang BFF, Boutros PC. The parameter sensitivity of random forests. BMC Bioinformatics. BMC Bioinformatics; 2016;17:331.

99. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics. 2009;10:213.

100. Branco P, Ribeiro RP, Torgo L. UBL: an R package for Utility-based Learning. arXiv preprint. 2016;arXiv:1604.08079.

101. Wright MN, Ziegler A. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software. 2017;77:i01.

102. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, et al. Mlr: Machine Learning in R. Journal of Machine Learning Research. 2016;17:1–5.

Additional files

Additional file 1: Figure S1. Gas production of bioreactors. **Figure S2.** Biomass production of bioreactors. **Figure S3.** Microbial community composition profiles of bioreactors. **Figure S4.** Alpha diversity metrics of bioreactor communities. **Figure S5.** Predictive performance of three machine learning algorithms using HRT bioindicators. **Figure S6.** Predictive performance of three machine learning algorithms using non-HRT bioindicators for considering community assembly caused by time. **Figure S7.** Prediction results of C6 and C8 productivities using non-HRT bioindicators for considering community assembly caused by time. **Figure S8.** Prediction results of C6 and C8 productivities for all samples in the four HRT phases using HRT bioindicators. **Figure S9.** Prediction results of C6 and C8 productivities for all samples in the four HRT phases using non-HRT bioindicators for considering community assembly caused by time. **Figure S10.** Random forest feature importance of A-HRT bioindicators and B-HRT bioindicators used to predict C6 and C8 productivities. **Figure S11.** Random forest feature importance of the non-HRT bioindicators used to predict C6 and C8 productivities. **Figure S12.** Metabolic pathways involved in converting lactate and xylan to *n*-caproate and *n*-caprylate. **Figure S13.** Correlation network of environmental factors, process performance and microbial community. **Figure S14.** Prediction results of C6 and C8 productivities for all samples in the

four HRT phases using the four ASVs of HRT bioindicators irrespective of time. **Figure S15.** Reducing HRT increases abundances of HRT bioindicators driving the catabolism of xylan and lactate to *n*-caproate and *n*-caprylate. **Figure S16.** Alpha rarefaction curves. **Figure S17.** Workflow of the random forest classification analysis. **Figure S18.** Workflow of a two-step random forest regression analysis. **Table S1.** Mean carboxylate yields (i.e., C mole product to substrate ratios) at HRTs of 8 days and 2 days (stable production period). **Table S2.** Explained variances of the training set in the regression-based prediction of process parameters using A-HRT bioindicators and B-HRT bioindicators. **Table S3.** Explained variances of the training set in the regression-based prediction of process parameters using non-HRT bioindicators for considering community assembly caused by time. **Table S4.** Growth medium used for the reactor operation. **Table S5.** Daily feeding of bioreactors A and B during the four HRT phases. **Table S6.** Gini scores of all ASVs in the classification-based prediction of HRT phases.

Additional file 2: Comparison of prediction accuracy of different algorithms using the mAML machine learning pipeline for classification.

Additional file 3: Dataset S1. MAGs taxonomy and genome metrics.

Additional file 4: Dataset S2. Functional annotations of xylose fermentation for MAGs with the same taxonomy as HRT bioindicators.

Additional file 5: Dataset S3. Functional annotations of chain elongation for MAGs with the same taxonomy as HRT bioindicators.

Additional file 6: Dataset S4. Functional annotations of xylose fermentation for all MAGs.

Additional file 7: Dataset S5. Functional annotations of chain elongation for all MAGs.

Figure legends

Figure 1. Performance of bioreactors. Concentrations of chain elongation products and lactate, as well as productivities and yields of chain elongation products in bioreactors A (**a**) and B (**b**) during the four HRT phases. Chain elongation products: C4, *n*-butyrate; C6, *n*-caproate; C8, *n*-caprylate.

Figure 2. Dissimilarities in bacterial community composition (beta-diversity). Non-metric multidimensional scaling (NMDS) based on Bray-Curtis dissimilarities of microbial community composition in bioreactors. **a,** All samples in the four HRT phases were considered for dissimilarity calculation. **b,** Samples in the 8-day HRT phase classified to the sampling interval 0-50 days and in the 2-day HRT phase classified to the interval 141-211 days were included.

Figure 3. Random forest feature importance of ASVs used to classify the HRT phases (A-HRT bioindicators and B-HRT bioindicators). The top-ranked 15 ASVs reducing the uncertainty in the prediction of HRT phases (HRT of 8 days and 2 days). The order of features (from top to bottom) was based on their mean decrease in Gini scores, according to their ASV abundances distribution, with HRT as the response variable. **a,** Feature importance of A-HRT

bioindicators. The ASV importance was calculated using the relative abundance data of bioreactor A as a training set and data of bioreactor B as a test set. **b,** Feature importance of B-HRT bioindicators. Similar to A-HRT bioindicators, ASV importance of B-HRT was calculated using the relative abundance data of bioreactor B as a training set and data of bioreactor A as a test set. The taxonomic classification of ASVs assigned at the genus level is provided in parentheses.

Figure 4. Prediction results of C6 and C8 productivities using HRT bioindicators. a,b, Prediction performance of C6 productivity. **c,d,** Prediction performance of C8 productivity. Results in **a** and **c** were obtained by using relative abundance data of bioreactor A for training the model and data of bioreactor B for testing. Results using the data of bioreactor B for training and bioreactor A for testing are shown in **b** and **d**. The red lines and grey shaded areas depict the best-fit trendline and the 95% confidence interval of the least-squares regression, respectively. C6, *n*-caproate; C8, *n*-caprylate; %Var., explains the variance (%) in C6/C8 productivity of the training set; RRMSE, relative root mean square error.

Figure 5. Phylogeny of HRT bioindicators and non-HRT bioindicators for considering community assembly caused by time. a,b, A maximum likelihood 16S rRNA gene tree showing the ASV species based on the rarefied sequencing data. ASVs are coloured according to the class (**a**, first inner ring) and family (**b**, second inner ring). **c**, The third inner ring shows the 11 HRT bioindicators identified in both reactors for the prediction of HRT phases of 8 days and 2 days. The ASVs identified as HRT bioindicators are shown in bold. Their taxonomic assignments at the genus level are provided in the legend. **d**, The four ASVs of HRT

bioindicators irrespective of time are shown in red in the outer ring. The ASVs only present in non-HRT bioindicators of C6/C8 productivity are shown in pink in the outer ring. **e**, Relative abundance dynamics of HRT bioindicators during the whole reactor operation period. In the legend, A and B stand for bioreactors A and B, respectively. The four ASVs (in bold) of HRT bioindicators, irrespective of time, assigned at the genus level are indicated in parentheses. C6, *n*-caproate; C8, *n*-caprylate.

Figure 6. Genetic potential of metagenome-assembled genomes (MAGs) with the same taxonomy as HRT bioindicators driving the catabolism of xylan and lactate to *n*-caproate and *n*-caprylate. These catabolic steps were categorised into four main functions of the anaerobic mixed culture fermentation. **a,** Hydrolysis of xylan. **b,** Xylose fermentation producing acetate and lactate. **c,** Butyrate formation from lactate and acetate. **d,** Chain elongation with lactate as electron donor producing *n*-butyrate, *n*-caproate and *n*-caprylate. Numbers represent the 18 different MAGs with similar phylogenies as the HRT bioindicators at the genus level (details in Table 1). The enzyme abbreviations are provided in red letters next to the pathways (solid lines). Dashed lines represent multi-enzyme reactions between the two indicated molecules. In **(d)**, "cycle" refers to the reverse β-oxidation cycle. The complete metabolic pathways are depicted in Additional file 1: Figure S12. un., unclassified; XL, xylanase (EC 3.2.1.8); XylT, xylose transporter (EC 7.5.2.10, EC 7.5.2.13); LacP, lactate permease (TC 2.A.14); CoAT, butyryl-CoA:acetate CoA-transferase (EC 2.8.3.-); PTB, phosphate butyryltransferase (EC 2.3.1.19); BUK, butyrate kinase (EC 2.7.2.7); ACT, acyl-CoA thioesterase (EC 3.1.2.20).

Figure 7. Phylogenetic tree of the recovered metagenome-assembled genomes (MAGs). a,b,

A phylogenomic tree based on mash distances showing the MAGs taxonomy determined by GTDB-Tk at phylum (**a**) and family (**b**) levels. A total of 108 MAGs were recovered and differentiated into 29 species based on the ANI values. We defined the representative MAG for each species as that showing high quality. Only the representative MAG for each species is depicted in the tree. The ENA accession numbers of the representative MAGs are shown in parentheses. MAGs with similar phylogenies as HRT bioindicators are indicated by a star.

Figure 8. Overview on the quantitative prediction of process performance in the anaerobic bioreactor system. a, Anaerobic mixed culture fermentation of lactate and xylan for the production of *n*-caproate (C6) and *n*-caprylate (C8) by lactate-based chain elongation. Based on the recovery of metagenome-assembled genomes, the left panel shows the bioindicators capable of performing key steps of the fermentation. **b,** Reducing the hydraulic retention time (HRT) as an operation-based strategy to optimise the process performance and to manage the reactor microbiota towards desired functions. Shortening the HRT from 8 days to 2 days enhanced productivities of C4, C6 and C8. The enriched reactor microbiota comprised functional groups involved in xylan hydrolysis, xylose fermentation and chain elongation with lactate, presented by a co-occurrence network of environmental factors (controlled conditions with only reducing the HRT), ecosystem functioning (process performance) and microbial community. The full network is shown in Additional file: Figure S13. **c**, Predicting performance of ecosystem processes with random forest analysis. We developed a random forest two-step workflow to qualitatively predict the HRT phases and to quantitatively predict carboxylate production by using relative abundance data of the 16S rRNA-derived species (ASVs, Amplicon Sequence Variants).

1103 **Table 1. Summary of metagenome-assembled genomes (MAGs) with the same taxonomy as HRT bioindicators.**

HRT bioindicators	Number of MAGs		Taxonomic classification						Representative MAG
	High quality	Medium quality	Phylum	Class	Order	Family	Genus	Species	
Olsenella sp. ASV034	2	3	*Actinobacteriota*	*Coriobacteriia*	*Coriobacteriales*	*Atopobiaceae*	*Olsenella*_B	*Olsenella*_B sp000752675	UMB00010
Olsenella sp. ASV057	4	2	*Actinobacteriota*	*Coriobacteriia*	*Coriobacteriales*	*Atopobiaceae*	*Olsenella*_C	unclassified	UMB00003
***Olsenella* sp. ASV058**	1	0	*Actinobacteriota*	*Coriobacteriia*	*Coriobacteriales*	*Atopobiaceae*	*Olsenella*	unclassified	UMB00074
unclassified *Erysipelotrichaceae* sp. ASV002	4	1	*Firmicutes*	*Bacilli*	*Erysipelotrichales*	*Erysipelotrichaceae*	*unclassified*	unclassified	UMB00059
Bulleidia sp. ASV010	0	1	*Firmicutes*	*Bacilli*	*Erysipelotrichales*	*Erysipelotrichaceae*	*Solobacterium*	unclassified	UMB00050
	6	0	*Firmicutes*	*Bacilli*	*Erysipelotrichales*	*Erysipelotrichaceae*	*Solobacterium*	*Solobacterium* sp900343155	UMB00007
	5	1	*Firmicutes*	*Bacilli*	*Erysipelotrichales*	*Erysipelotrichaceae*	*Solobacterium*	*Solobacterium* sp900290205	UMB00011
Lachnospiracea incertae sedis ASV053	3	0	*Firmicutes*_A	*Clostridia*	*Lachnospirales*	*Lachnospiraceae*	UBA4285	unclassified	UMB00063
***Syntrophococcus* sp. ASV060**	*Eubacterium cellulosolvens* 6		*Firmicutes*_A	*Clostridia*	*Lachnospirales*	*Lachnospiraceae*	*Eubacterium*_H	*Eubacterium*_H *cellulosolvens*	
	Eubacterium cellulosolvens LD2006		*Firmicutes*_A	*Clostridia*	*Lachnospirales*	*Lachnospiraceae*	*Eubacterium*_H	*Eubacterium*_H *cellulosolvens*_A	
	5	0	*Firmicutes*_A	*Clostridia*	*Lachnospirales*	*Lachnospiraceae*	*Eubacterium*_H	unclassified	UMB00012
	6	0	*Firmicutes*_A	*Clostridia*	*Lachnospirales*	*Lachnospiraceae*	*Eubacterium*_H	unclassified	UMB00020
***Clostridium* IV sp. ASV073**	*Caproiciproducens galactitolivorans* BS-1		*Firmicutes*_A	*Clostridia*	*Oscillospirales*	*Acutalibacteraceae*	MS4	unclassified	
	5	0	*Firmicutes*_A	*Clostridia*	*Oscillospirales*	*Acutalibacteraceae*	UBA1033	UBA1033 sp002399935	UMB00014
	1	0	*Firmicutes*_A	*Clostridia*	*Oscillospirales*	*Acutalibacteraceae*	UBA1033	UBA1033 sp002407675	UMB00060
	3	0	*Firmicutes*_A	*Clostridia*	*Oscillospirales*	*Acutalibacteraceae*	UBA1033	UBA1033 sp002409675	UMB00097
	Ruminococcaceae bacterium CPB6		*Firmicutes*_A	*Clostridia*	*Oscillospirales*	*Acutalibacteraceae*	UBA4871	UBA4871 sp002119605	

	6	0	*Firmicutes_A*	*Clostridia*	*Oscillospirales*	*Acutalibacteraceae*	UBA4871	UBA4871 sp002399445	UMB00016
Clostridium sensu stricto sp. ASV008	*Clostridium luticellarii* DSM29923		*Firmicutes_A*	*Clostridia*	*Clostridiales*	*Clostridiaceae*	*Clostridium_B*	*Clostridium_B luticellarii*	
	3	0	*Firmicutes_A*	*Clostridia*	*Clostridiales*	*Clostridiaceae*	*Clostridium_B*	*Clostridium_B* sp003497125	UMB00080
***Lactobacillus* sp. ASV074**	6	0	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus_H*	*Lactobacillus_H mucosae*	UMB00017
	0	1	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	unclassified	UMB00041
	2	2	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	*Lactobacillus amylovorus*	UMB00015
unclassified *Coriobacteriaceae* sp. ASV082	0	0							

Taxonomy refers to the GTDB (Genome Taxonomy Database) phylogenomic classification. ASVs in bold represent the four HRT bioindicators irrespective of time. Sequence datasets of genomes in red letters were taken from the databases of NCBI and EzBioCloud. These genomes (in red) were used to affiliate the MAGs of *Syntrophococcus*, *Clostridium* IV and *Clostridium sensu stricto*, since their genomes are not available in GTDB. See details of MAGs in Additional file 3: Dataset S1. ASV: amplicon sequencing variant.

Figures

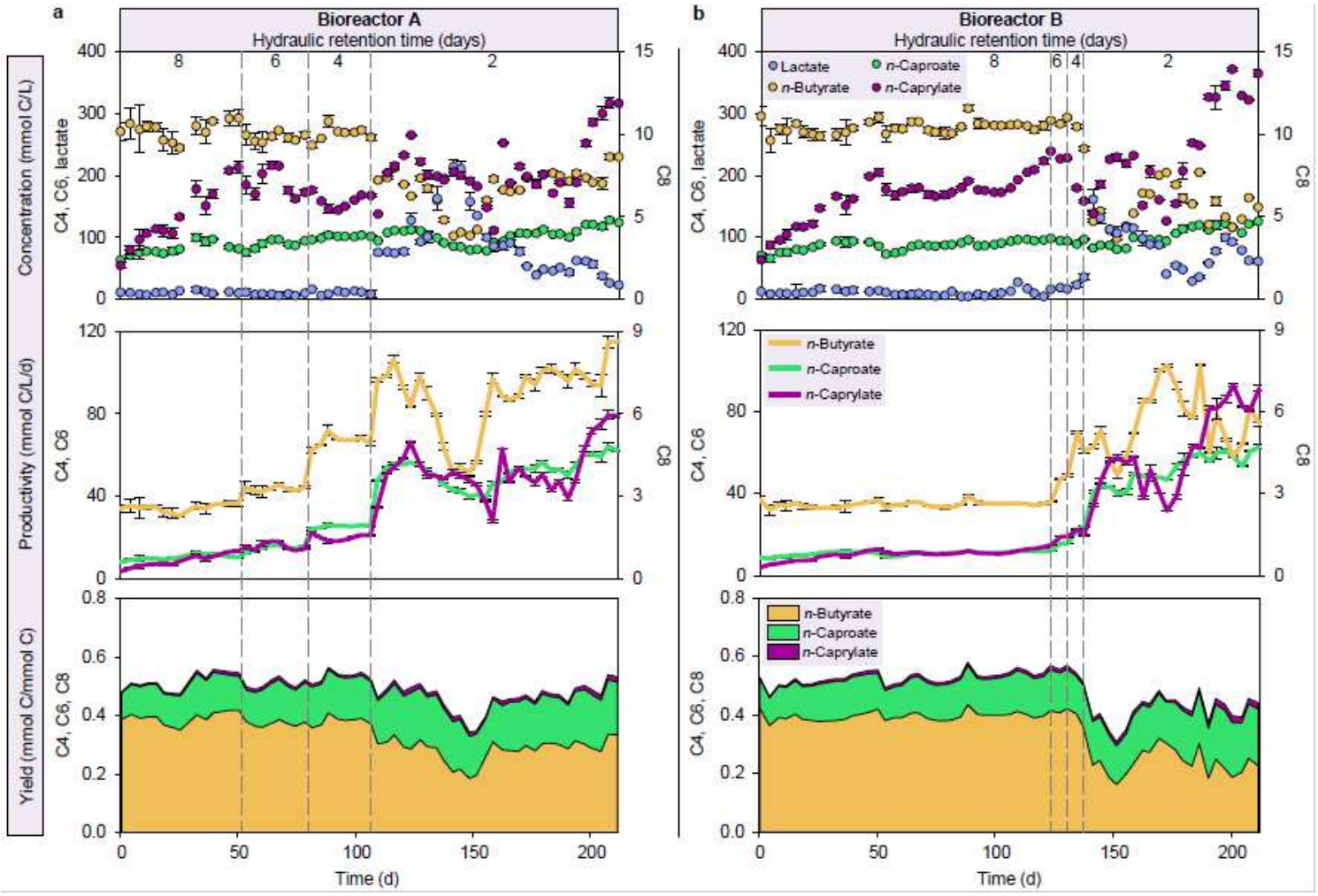

Figure 1

Performance of bioreactors. Concentrations of chain elongation products and lactate, as well as productivities and yields of chain elongation products in bioreactors A (a) and B (b) during the four HRT phases. Chain elongation products: C4, n-butyrate; C6, n-caproate; C8, n-caprylate.

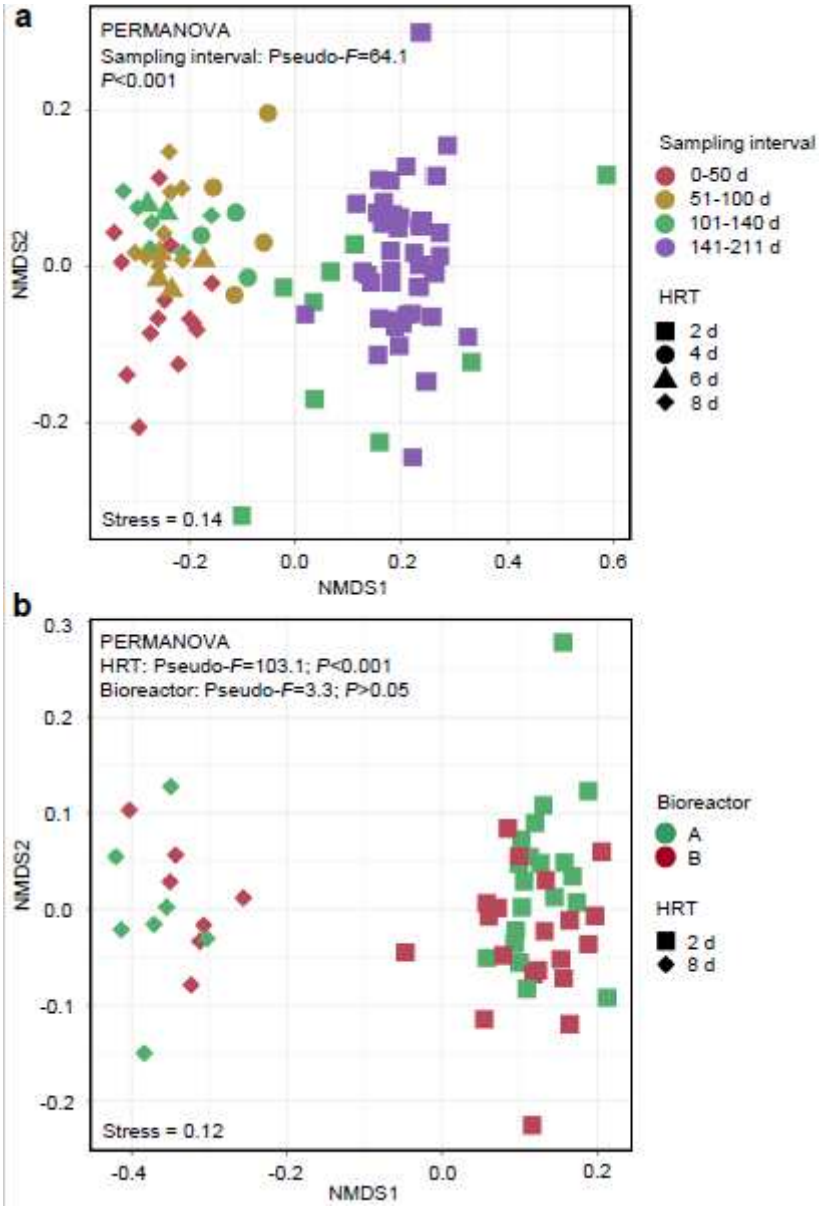

Figure 2

Dissimilarities in bacterial community composition (beta-diversity). Non-metric multidimensional scaling (NMDS) based on Bray-Curtis dissimilarities of microbial community composition in bioreactors. a, All samples in the four HRT phases were considered for dissimilarity calculation. b, Samples in the 8-day HRT phase classified to the sampling interval 0-50 days and in the 2-day HRT phase classified to the interval 141-211 days were included.

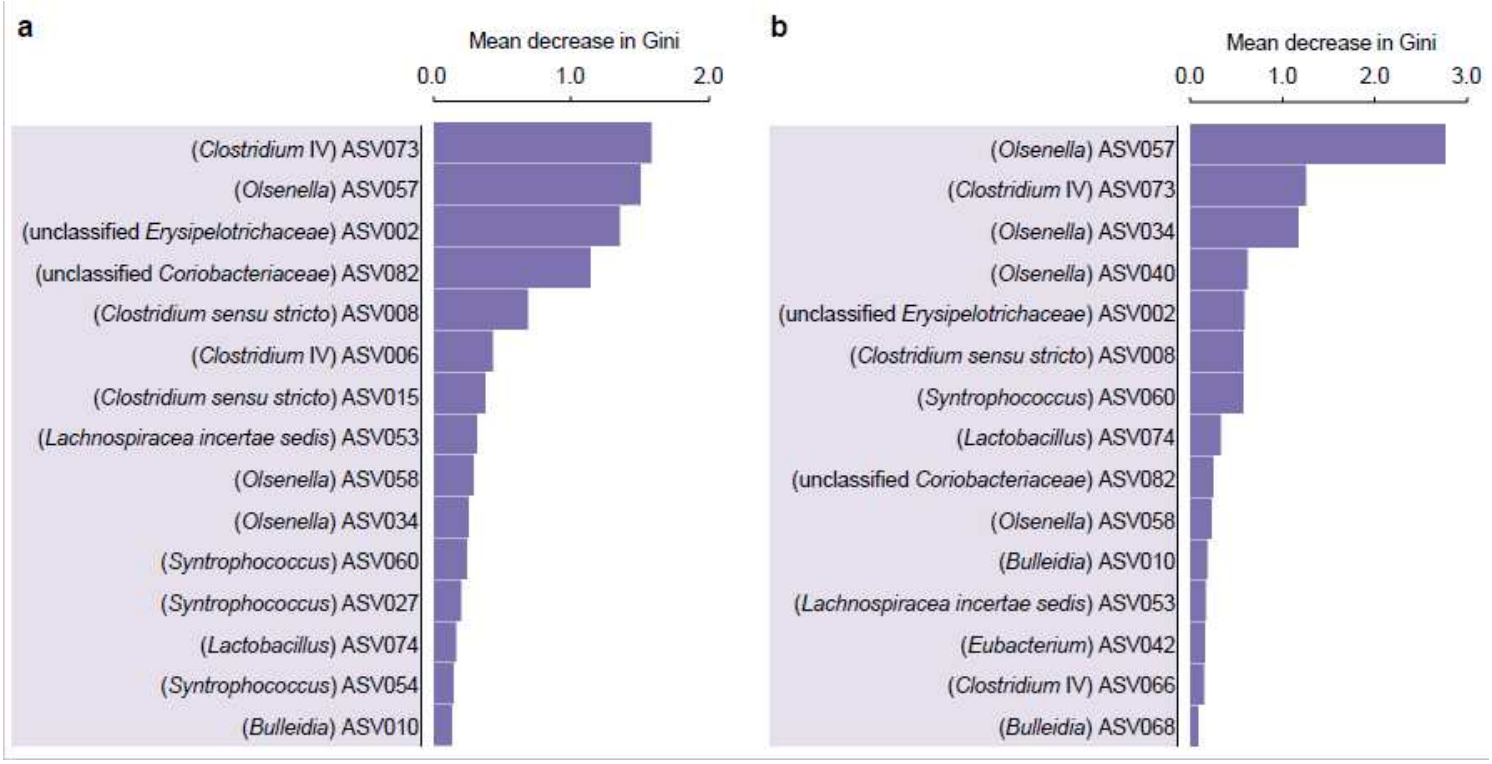

Figure 3

Random forest feature importance of ASVs used to classify the HRT phases (A-HRT bioindicators and B-HRT bioindicators). The top-ranked 15 ASVs reducing the uncertainty in the prediction of HRT phases (HRT of 8 days and 2 days). The order of features (from top to bottom) was based on their mean decrease in Gini scores, according to their ASV abundances distribution, with HRT as the response variable. a, Feature importance of A-HRT bioindicators. The ASV importance was calculated using the relative abundance data of bioreactor A as a training set and data of bioreactor B as a test set. b, Feature importance of B-HRT bioindicators. Similar to A-HRT bioindicators, ASV importance of B-HRT was calculated using the relative abundance data of bioreactor B as a training set and data of bioreactor A as a test set. The taxonomic classification of ASVs assigned at the genus level is provided in parentheses.

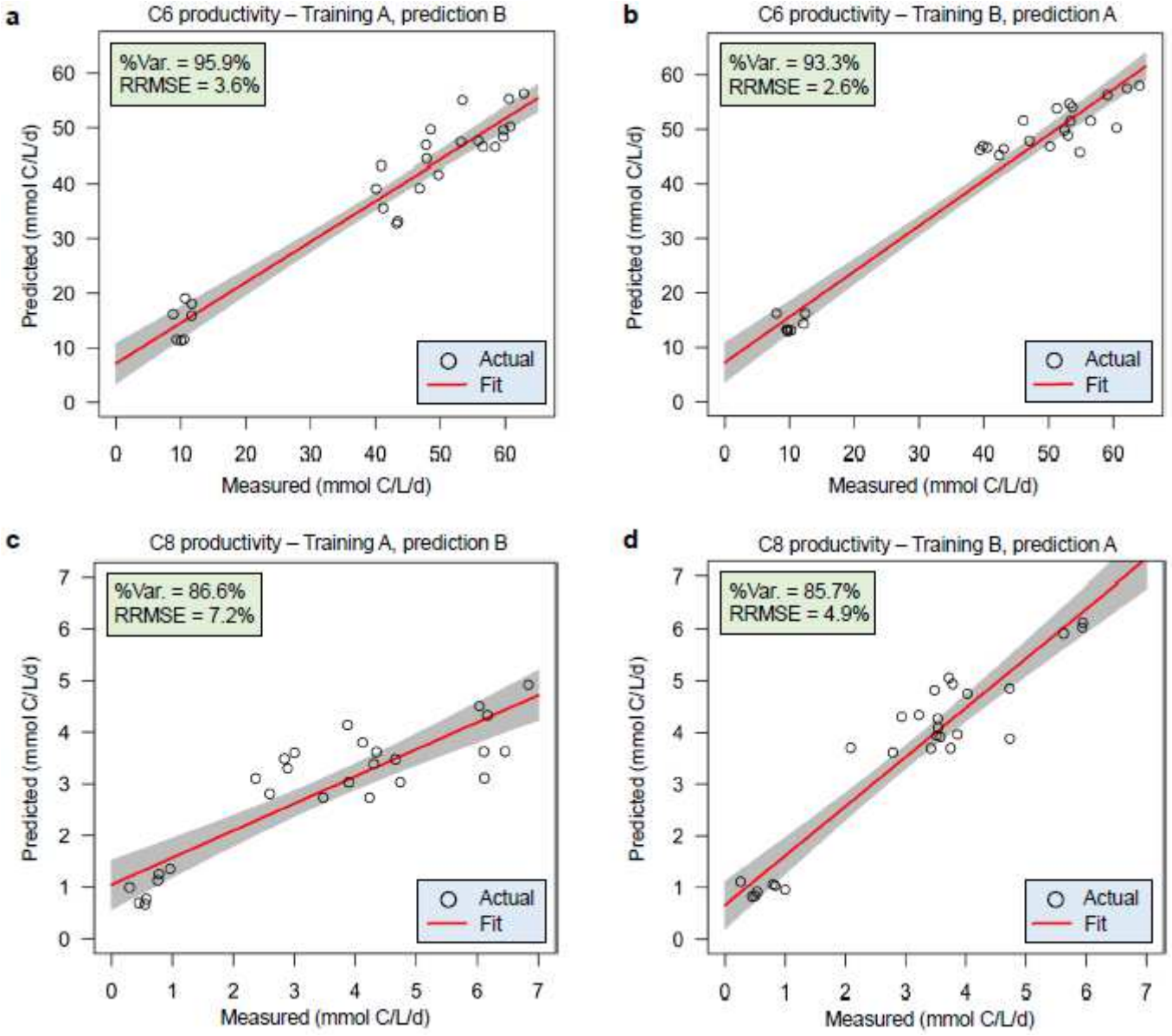

Figure 4

Prediction results of C6 and C8 productivities using HRT bioindicators. a,b, Prediction performance of C6 productivity. c,d, Prediction performance of C8 productivity. Results in a and c were obtained by using relative abundance data of bioreactor A for training the model and data of bioreactor B for testing. Results using the data of bioreactor B for training and bioreactor A for testing are shown in b and d. The red lines and grey shaded areas depict the best-fit trendline and the 95% confidence interval of the least-squares regression, respectively. C6, n-caproate; C8, n-caprylate; %Var., explains the variance (%) in C6/C8 productivity of the training set; RRMSE, relative root mean square error.

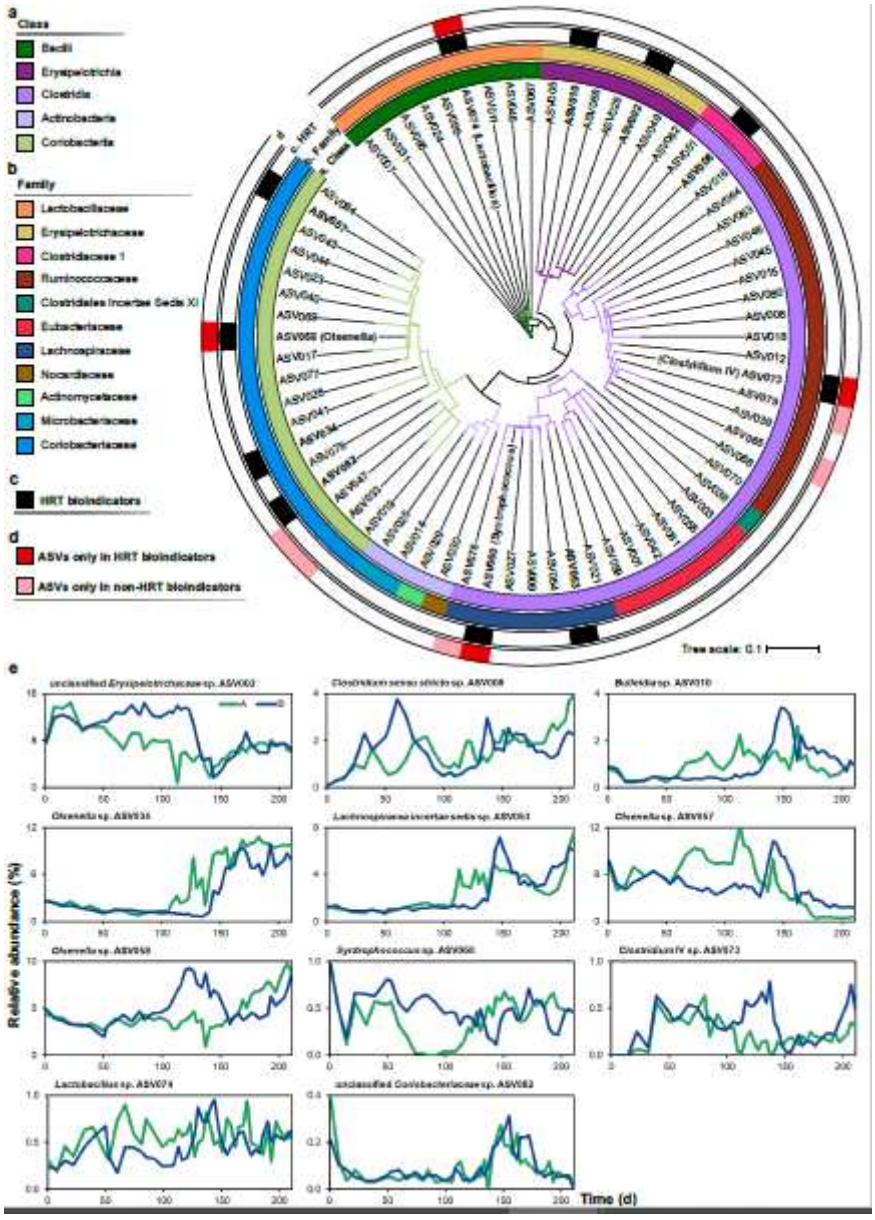

Figure 5

Phylogeny of HRT bioindicators and non-HRT bioindicators for considering community assembly caused by time. a,b, A maximum likelihood 16S rRNA gene tree showing the ASV species based on the rarefied sequencing data. ASVs are coloured according to the class (a, first inner ring) and family (b, second inner ring). c, The third inner ring shows the 11 HRT bioindicators identified in both reactors for the prediction of HRT phases of 8 days and 2 days. The ASVs identified as HRT bioindicators are shown in bold. Their taxonomic assignments at the genus level are provided in the legend. d, The four ASVs of HRT bioindicators irrespective of time are shown in red in the outer ring. The ASVs only present in non-HRT bioindicators of C6/C8 productivity are shown in pink in the outer ring. e, Relative abundance dynamics of HRT bioindicators during the whole reactor operation period. In the legend, A and B stand for bioreactors A and B, respectively. The four ASVs (in bold) of HRT bioindicators, irrespective of time, assigned at the genus level are indicated in parentheses. C6, n-caproate; C8, n-caprylate.

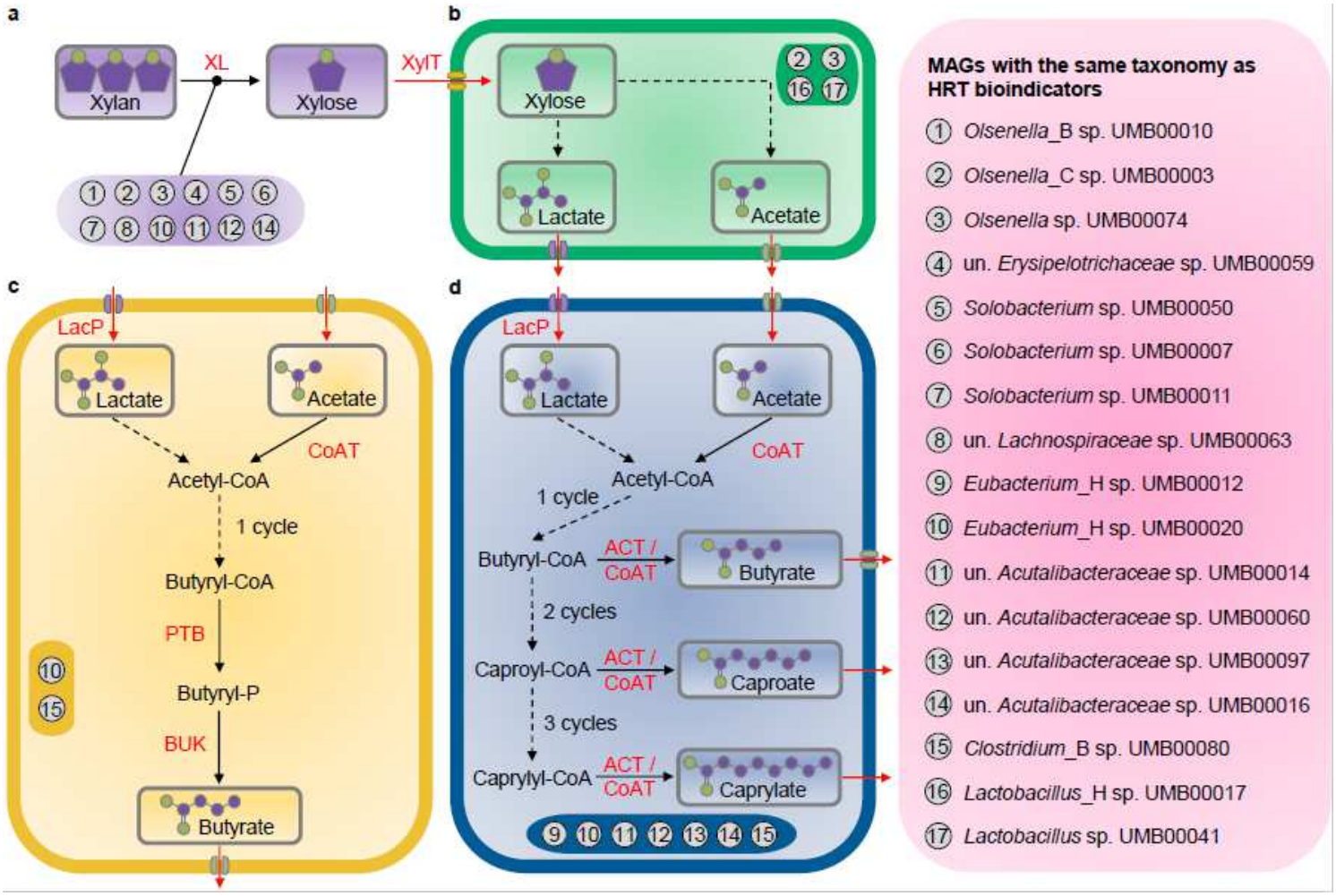

Figure 6

Genetic potential of metagenome-assembled genomes (MAGs) with the same taxonomy as HRT bioindicators driving the catabolism of xylan and lactate to n-caproate and n-caprylate. These catabolic steps were categorised into four main functions of the anaerobic mixed culture fermentation. a, Hydrolysis of xylan. b, Xylose fermentation producing acetate and lactate. c, Butyrate formation from lactate and acetate. d, Chain elongation with lactate as electron donor producing n-butyrate, n-caproate and n-caprylate. Numbers represent the 18 different MAGs with similar phylogenies as the HRT bioindicators at the genus level (details in Table 1). The enzyme abbreviations are provided in red letters next to the pathways (solid lines). Dashed lines represent multi-enzyme reactions between the two indicated molecules. In (d), “cycle” refers to the reverse β-oxidation cycle. The complete metabolic pathways are depicted in Additional file 1: Figure S12. un., unclassified; XL, xylanase (EC 3.2.1.8); XylT, xylose transporter (EC 7.5.2.10, EC 7.5.2.13); LacP, lactate permease (TC 2.A.14); CoAT, butyryl-CoA:acetate CoA-transferase (EC 2.8.3.-); PTB, phosphate butyryltransferase (EC 2.3.1.19); BUK, butyrate kinase (EC 2.7.2.7); ACT, acyl-CoA thioesterase (EC 3.1.2.20).

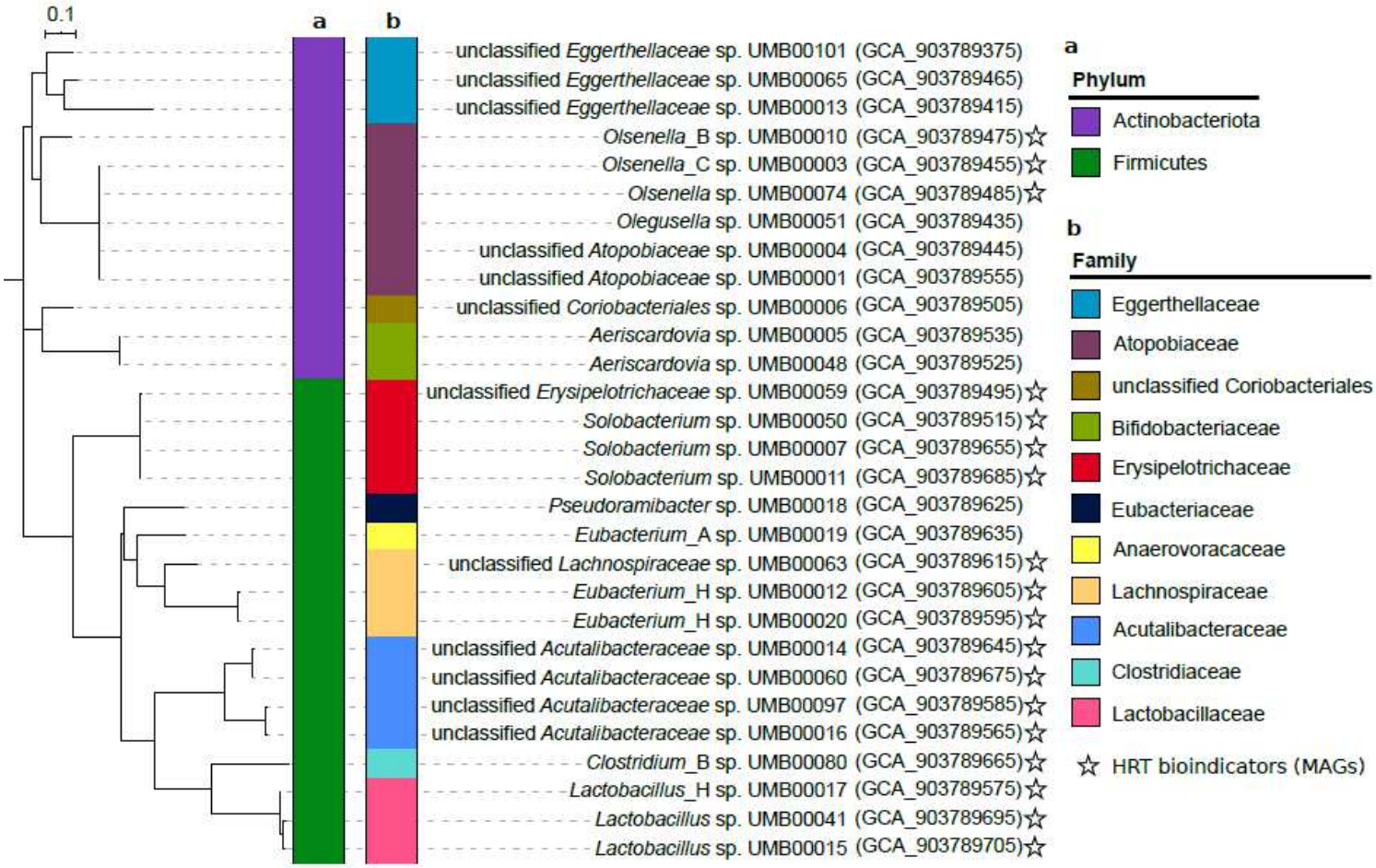

Figure 7

Phylogenetic tree of the recovered metagenome-assembled genomes (MAGs). a,b, A phylogenomic tree based on mash distances showing the MAGs taxonomy determined by GTDB-Tk at phylum (a) and family (b) levels. A total of 108 MAGs were recovered and differentiated into 29 species based on the ANI values. We defined the representative MAG for each species as that showing high quality. Only the representative MAG for each species is depicted in the tree. The ENA accession numbers of the representative MAGs are shown in parentheses. MAGs with similar phylogenies as HRT bioindicators are indicated by a star.

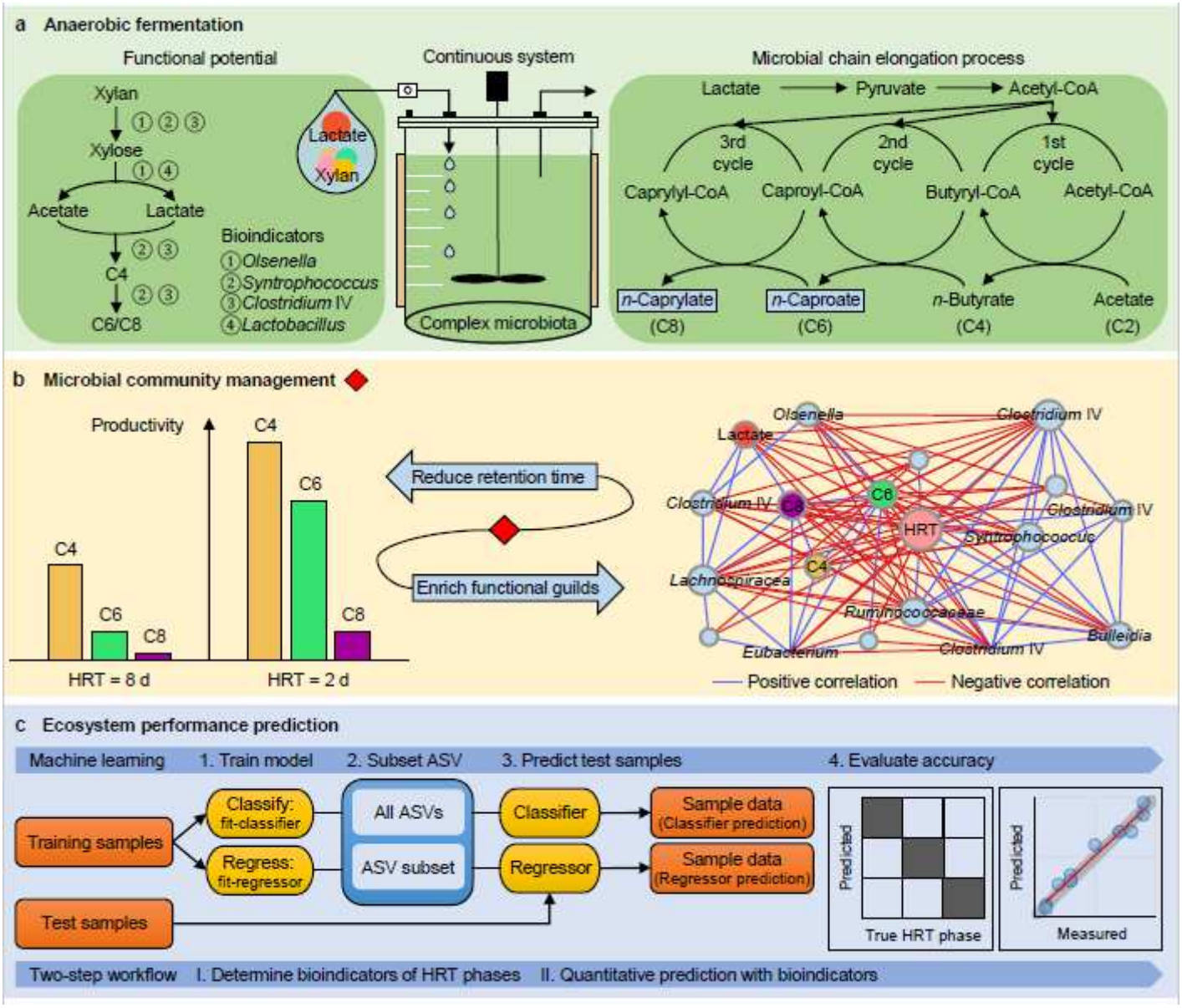

Figure 8

Overview on the quantitative prediction of process performance in the anaerobic bioreactor system. a, Anaerobic mixed culture fermentation of lactate and xylan for the production of n-caproate (C6) and n-caprylate (C8) by lactate-based chain elongation. Based on the recovery of metagenome-assembled genomes, the left panel shows the bioindicators capable of performing key steps of the fermentation. b, Reducing the hydraulic retention time (HRT) as an operation-based strategy to optimise the process performance and to manage the reactor microbiota towards desired functions. Shortening the HRT from 8 days to 2 days enhanced productivities of C4, C6 and C8. The enriched reactor microbiota comprised functional groups involved in xylan hydrolysis, xylose fermentation and chain elongation with lactate, presented by a co-occurrence network of environmental factors (controlled conditions with only reducing the HRT), ecosystem functioning (process performance) and microbial community. The full network is shown in Additional file: Figure S13. c, Predicting performance of ecosystem processes with random forest analysis. We developed a random forest two-step workflow to qualitatively predict the HRT phases

and to quantitatively predict carboxylate production by using relative abundance data of the 16S rRNA-derived species (ASVs, Amplicon Sequence Variants).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 12Additionalfile1.FigureS1S18andTableS1S6.pdf
- 13Additionalfile2.PredictionaccuracymAML.xlsx
- 14Additionalfile3.MAGstaxonomyandgenomemetrics.xlsx
- 15Additionalfile4.xlsx
- 16Additionalfile5.xlsx
- 17Additionalfile6.xlsx
- 18Additionalfile7.xlsx