Synthetic biology approaches for exploring medium chain fatty acid production in anaerobic bacterial communities

By

Madeline M Hayes

A dissertation submitted in partial fulfilment of

The requirements for the degree of

Doctor of Philosophy

(Microbiology)

At the University of Wisconsin-Madison

2025

Date of final oral examination: 01/21/2025

This dissertation is approved by the following members of the Final Oral Committee:

Ophelia Venturelli, Associate Professor, Biomedical Engineering (Duke University)

Daniel Noguera, Professor, Civil and Environmental Engineering

Jo Handelsman, Professor, Plant Pathology

Michael Thomas, Professor, Bacteriology

Sushmita Roy, Professor, Biostatistics

Abstract

Synthetic biology approaches for exploring medium chain fatty acid production in anaerobic bacterial communities

Madeline M. Hayes

Under the supervision of Dr. Ophelia Venturelli and Dr. Daniel Noguera

University of Wisconsin-Madison

Microbes have been observed in complex communities in nearly every environment on earth, and leveraging their unique chemistries for a desired functional output is an important engineering goal for increasing agricultural output, improving human health, and sustainable production of chemicals and fuels. Microbial production of carboxylic acids from lignocellulosic stillage and other waste streams aims to provide a sustainable and renewable supply of high-value chemicals. To parse individual species' functional contributions to medium chain fatty acid (MCFA) production from biofuel stillage waste, here I apply an iterative systems biology approach, using computational modeling to describe and predict community function, and high-throughput experimentation to iteratively select synthetic communities from the highly complex sample space of a 16-member anaerobic bacterial consortium. I examine two synthetic stillage media to analyze the impact of multiple carbon sources and pH on the structure and functional outputs of sub-communities. I identify *Clostridium kluyveri* as a dominant hexanoate producer within communities, and identify partner species which positively and negatively impact its growth, and support or inhibit hexanoate production. Additionally, I identify key species which digest plant sugars and reveal metabolite-mediated pH sensitivity of hexanoate production.

Acknowledgements

Thank you to Ophelia Venturelli for supporting this work and challenging me to step outside my comfort zone.

Thank you to my committee, especially Dan Noguera for stepping in as co-advisor during a hard time, and Jo Handelsman for thoughtful and kind advice.

Thank you to everyone at GLBRC – especially Mick McGee, Steve Karlan, Michael Botts and all the interns in the metabolomics lab -- for your incredible scientific support. I have always felt welcomed, valued, and respected in your company.

Thank you to my colleagues and labmates from Biochemistry, MDTP, and Plant Pathology past and present for the support, friendship, advice, feedback, and creating an environment where I could succeed. Thanks in particular to Derrick Grunwald.

Thank you to my past mentors, Torsten Eckstein and Caitilyn Allen for inspiring me to do research and supporting my goals.

Thank you to the staff and regulars of the Library Café and Bar for creating a welcoming space of respite.

Thank you to all my teammates in MUFA, especially the Ruth-Badder DiscBurns, for giving me something to look forward to every season.

Thank you to my family for supporting my work even when it confuses you and I talk too fast about it.

Thank you to my chosen family – Matt Pereyra, April MacIntyre, Alicia Truchon, CJ Jimenez, Malory Donahue, and Tom Adams-Porter for your boundless and unconditional love – you have made and continue to make me a better person.

Thank you especially to Kate Phelps, for reminding me what I'm capable of and always jumping into the deep end with me.

This material is based upon work supported by the Great Lakes Bioenergy Research Center, U.S. Department of Energy, Office of Science, Biological and Environmental Research Program under Award Number DE-SC0018409.

Table of Contents

List of Tables

Chapter 1

Chapter 2

List of Figures

Chapter 2

Chapter 3

Chapter 4

Appendix

Chapter 1: The role of microbial bioprocess engineering in sustainable chemical production

Contributions: Madeline Hayes wrote the text and created the figures and tables

The promise and challenges of microbial bioprocessing for solving anthropogenic challenges

Microbial communities are emerging as vastly untapped sources of potential biochemical labor. Especially in highly bioproductive environments, from deep sea and thermal hot springs[1] to mangrove rhizospheres[2] to the guts of unusually large rodents,[3] evidence of microbial metabolisms as mediators of diverse chemical transformations that shape global processes[4,5] is growing. This abundance of chemical capacity represents a reservoir of possible answers to humanity's grand challenges. Efforts to harness and engineer microbiomes to improve agricultural outputs,[6] promote human health,[7] and reform industrial production of chemicals[8] are facilitating a transition in global thinking surrounding how to improve the human condition in sustainable ways.

In the arena of industrially relevant chemicals, there is a particular focus on the circular bioeconomy[9] which aims to eliminate unsustainable sources of chemical products by providing sources which upgrade existing streams of waste. Many commercially manufactured materials that require high energy input to produce chemically or utilize precursors from fossil fuels are currently targets for improved sourcing and production. In pairing underutilized organic waste streams -- such as food waste,[10] water treatment waste,[11] and animal agricultural waste[12] -- with economically relevant pipelines, researchers aim to reduce the environmental impact of manufacturing while simultaneously adding value to process outputs that would otherwise be disposed of. Microbiomes are well suited to remediating the scraps and residues from everyday processes due to their wide range of digestive capabilities which naturally function to reduce complex substrates to their constituent elements in the global carbon, nitrogen, and phosphorus cycles.

Although the potential for microbiome-based solutions is obvious, harnessing these systems in precise and predictable ways has proven very challenging. Fast and inexpensive DNA sequencing has fueled a rapid rise in the characterization of various natural microbiomes, but understanding the ecological networks that drive community function often remains elusive. The crux of microbiomes as puzzles lies in their complexity: many thousands of species, often very rare, unculturable, or inactive, consuming varied and heterogenous substrates, and producing a wide array of metabolites in response to each other and their dynamic environment makes for a largely intractable system. Fortunately, approaches that aim to capture huge varieties of data (multi-omics approaches) and process many elements using computational approaches ('big data') combine to give rise to Systems Biology.[13,14] In particular, application of detailed molecular data in tightly controlled systems with synthetic microbiomes – simplified assemblages of isolated and characterized microbial species – has helped advance our understanding of microbiome function in diverse contexts.[15]

Engineering approaches for microbial community MCFA production

Medium chain fatty acids (MCFAs – organic carboxylic acids with six to ten carbons) are valuable chemicals used widely in pharmaceutical production,[16–18] as precursors for various synthetic textiles ,[19] fuel additives,[20] lubricants,[21] and biosurfactants.[22] Commercial production of MCFAs usually relies on the Ziegler-ALFOL process,[23] wherein ethylene is oligomerized using aluminum as a catalyst. Ethylene is derived from petrochemicals via steam-cracking, an energy intensive process by which hydrocarbons are liberated from naphtha (derived from crude oil) or ethane gas (derived from natural gas) as monomers at high temperatures, then rapidly cooled and purified for further use.[24] Demand for petrochemical feedstocks is projected to increase in both

developed and developing markets worldwide,[25] and thus providing renewable and sustainable sources of these oleochemicals is a critical component of reducing use of fossil fuels. Currently, a limited amount of renewable sourcing for MCFAs as well as other longer fatty acids includes soybean and palm oil,[26] neither of which are considered sustainable,[27] as they rely on clearing and replanting of land and interfere directly with human food supply. Finding a source for sustainable MCFA production is currently under intense investigation.[28,29]

As the global transportation industry grapples with the threat of supply and demand discrepancies for finite fossil fuel resources, pressure to incorporate renewable and sustainable fuel stocks has increased, with a keen focus on increasing biofuel production worldwide.[30] However, optimizing biofuel production to be not only profitable but also competitive with petrochemical equivalents has proven very challenging,[31] and biofuel production itself is not free of highly pollutant waste streams which require remediation, adding to this cost. Biofuel production in the US relies heavily on corn as plant biomass substrate and is in direct competition with human food supply and livestock feed sources. Ideally, sustainable biomass sourcing will not interfere with food supply or optimal use of agricultural land; therefore, utilization of biomass crops that can thrive on marginal land is preferred. To this end, increased use of 'bioenergy crops' to provide lignocellulosic biomass for biofuel production is on the rise.[32] The 'field to fuel' approach aims to optimize production of these crops through breeding and tuned cultivation, determine appropriate pretreatment, fermentation, and distillation methods for maximizing biofuel production, and investigate paired valorization methods for biofuel production waste streams (in this context, conversion residue or CR).[33] Transition from a fossil fuel-based economy will require integration of multiple waste recovery technologies, and pairing

microbial anaerobic digestion strategies with sustainable fuel production increases the economic viability of both.[34]

Anaerobic remediation of rich organic substrates relies on fermentative metabolic pathways, which convert carbohydrates to short chain fatty acids; these are then elongated into MCFAs via the reverse beta-oxidation (R-BOX) pathway (extensively reviewed[29,35–38]). This platform, referred to as the chain elongation (CE) platform, utilizes an electron donor such as ethanol or lactic acid to elongate acetate or butyrate two carbons at a time in a reversal of fatty acid synthesis via beta oxidation. Microbes readily engage in this process, but tuning the environmental parameters that make these reactions energetically favorable is a field that is still emerging.[39–41] Table 1 details several recent examples of wild-type and engineered strains, cocultures, and more complex communities studied for their chain elongation phenotypes. In this context, 'bioaugmented' is used to describe a naturally isolated community in concert with a well-characterized single strain. Most recent efforts have focused on *Clostridium kluyveri*, a well-established MCFA producer,[42] for its ability to utilize ethanol as an electron donor to produce hexanoate. Other species such as *Megasphaera elsdenii* utilize primarily lactate as an electron donor;[43] both these metabolites are readily produced from cellulosic plant matter, making lignocellulosic feedstocks (especially if they contain leftover ethanol, such as beer or liquor making waste streams) ideal for upgrade by these organisms. Building small cocultures and optimizing conditions for their function is a bottom-up engineering approach that has proven quite effective, with maximum observed titers for hexanoate reaching >18 g/L.[44]

Uncharacterized or environmentally isolated microbial communities, while more difficult to parse functionally, leverage pre-existing and unknown metabolisms that may be present in

open cultures. The diversity of organisms in natural inoculants could provide novel findings, as well as being more robust to contamination and perturbation. Top-down engineering approaches often utilize these complex communities to compare conditional screens and understand how communities respond to changes in conditions such as pH. While *C. kluyveri* apparently prefers neutral pH for CE, it survives and produces hexanoate under a variety of conditions,[45] and there are some species which apparently prefer acidic pH for CE.[46] The specific impacts of partner species, pH, and chain elongation favorability are being explored for small communities,[47] but not in the context of complex waste substrates that are not amended with carbon sources and electron donors designed to improve functionality. Understanding these factors in more complex systems remains unexplored.

Computational Modeling Offers Insight into complex community environments

Quantifying and predicting behavior of microbial communities remains the primary challenge in engineering microbiomes. A wide array of modeling approaches have been applied to microbial community data, both statistical and machine learning. Using computational methods to assign quantitative values to both observable properties of communities (such as abundance of species) as well as potential interactions that may not have an easily detected impact (such as growth promotion or inhibition by partner species) is the basis for extracting biologically relevant information from the dense systems.

Predicting the growth and assembly of microbial communities by inferring the species-species interactions at play is a widely used strategy. The generalized Lotka-Volterra model has been used to predict community dynamics from lower-order assemblies of species.[48–50] However, as a series of ordinary differential equations, estimating these parameters can be computationally

expensive, and lower order interactions can fail to capture community behavior that is not mediated by biomass accumulation alone.[51]

Modeling approaches that integrate community growth with functional output have been successful in identifying key parameters for rational microbiome design.[52,53] Especially in the field of machine learning, flexible model architectures that do not require prior system knowledge to be implemented can be used to discover novel community functions and predict microbiome phenotypes.[54,55] Unfortunately, these models are often 'black box' models which have no simple interpretation of the parameters or methods used to make predictions. For pure optimization problems, this may not be a hindrance, but in understanding mechanisms by which community behavior arises, interpreting model parameters remains a strong advantage. Applying modular modeling approaches, which can leverage the advantages of interpretability while still maintaining flexible architecture, is the cutting edge of microbiome modeling.[53]

While each of these approaches has its advantages and disadvantages, knowing with certainty which approach will be most successful with any given system is not currently possible. Application of diverse models to different systems provides the opportunity to gain a better understanding of modeling as a technique, but also the potential for different model structures to reveal different information. Ultimately, an understanding of which model is appropriate for any system is still a question that must be explored empirically. Building large enough datasets to train and test multiple models on real data has emerged as a leading approach,[56] and leveraging high throughput experimental pipelines as a platform for developing models will continue to support the evolution of synthetic and systems biology approaches.

Conclusion

Microbial MCFA production is a complicated, cascading process that is impacted by community composition, substrate, and conditions. Step-wise transformations of carbon sources and intermediates represent areas of particular interest, and deciphering how the flow of carbon and energy is transferred and transformed by microbial metabolism is a crucial step in engineering these systems for scale up and industrial use. Computational models show promise in deciphering these interactions and providing basis for testable hypotheses regarding molecular mechanisms.

In this thesis I will demonstrate the utility of computational models in deciphering bacterial community function and their advantages for hypothesis generation in highly complex sample spaces. By applying these modeling techniques to a synthetic MCFA-producing community, I identify subsets of communities which support the growth and function of key MCFA producers and maximize hexanoate production (Chapter 2). By following the implications of the interpretations from the modeling phase, I reveal metabolite-mediated interactions which impact the favorability of MCFA production metabolism (Chapter 3). These findings reveal novel community interactions and control knobs for engineering MCFA production, providing targets for future strain engineering and support community construction to support robust and efficient bioprocessing platforms

Table 1: Summary of recent engineering of MCFA production in single strains, co cultures, and communities.

	Substrate	Media additives	Inoculum	hexanoate g/L	total	pH	Time	Notes	Citation Year
Single strains	Roth Basal Medium	fructose and butyrate	*Caprociciproducens* sp. 7D4C2	17		5	6d	optimizing for low pH	[57] 2020
	C. kluyveri medium	ethanol and acetate	*C. kluyveri*	12.9		6.8	225h	optimizing EtOH:acetate	[58] 2024
	Modified Turbo GCM media	ethanol	*C. kluyveri*	12.65		7 (controlled)	48hr		[44] 2022
	custom media	CO_2 sparging	*C. kluyveri*	10.5		6.8	150h	investigating role of CO2	[59] 2024
	NRBC medium 213		*C. kluyveri*	3.2		7	33d +	investigating methanogenesis, EtOH:acetate	[28] 2018
	DSM 13528	CO_2 and H_2 sparging	*Clostridium ljungdahlii* (engineered)	0.1		6	240hr	engineering production of hexanol	[60] 2022
	TY	glucose and lactate	*M. elsdenii* T81	0.58		6.5	48hr	investigating glucose and lactate metabolism	[61] 2013
	LB		*Escherichia coli* (engineered)		1.8 (multiple lengths)	not adjusted	48hr	engineering MCFA production from glucose	[62] 2018
	undefined 'rich medium'		*E. coli* (engineered)		3.8 (multiple lengths)	not adjusted	48hr	optimizing R-BOX pathway	[63] 2017

Co-cultures	Modified Turbo GCM media		*C. kluyveri* + *Clostridium saccharolyticum*	18.4		7 (controlled)	60hr	partner strain produces acetate + lactate	[44] 2022
	Modified Turbo GCM media		*C. kluyveri* + *Clostiridium acetobutylicum*	11.5		6-7 (spiked)	60hr	partner strain produces EtOH	[44] 2022
	GL-45 media	glucose	*C. kluyveri* + *C. acetobutylicum*	18.5		7 (controlled)	140hr	partner strain produces EtOH	[44] 2022
	Hurst	glucose	*C. kluyveri* + *Clostridium carboxidivorans*	8.5		not adjusted	56hr	developing FISH probes for cocultures	[64] 2022
	DSM52		*C. kluyveri* + *Clostridium autoethanologen*	0.23 g/hr (continuous)		6.2	continuous	developing GEM model for combined potential	[65] 2020
Bioaugmented communities	acid whey sludge	lactate	*M. elsdenii* and *Eubacterium limosum,* + open culture acid whey from dairy processing	13.9		6.8	10days	Understanding effect of glycerol	[66] 2024
	Dried Switchgrass or alfalfa		*C. kluyveri* + ruminal contents	4.84		not controlled	72hr		[45] 2015
Natural isolates	Ginkgo leaves and grass		Anaerobic sludge	7.48		7			[67] 2022

	xylan, lactate		maize sillage	8.2gCOD/L/d		5.5	60d +	continuous	[68] 2020
	conversion residue		wastewater treatment sludge	15gCOD/L		5.5	120d	continuous	[69] 2018

Table 1: Recent publications describing MCFA production. Categories: single strains are individual species or isolates grown in monoculture. Co-cultures are groups of two species grown together. Bioaugmented communities are natural communities with addition of known, characterized strains. Natural isolates are complex communities. Media are listed, and if any carbon source or additive was added to the media. Maximum hexanoate observed is listed in g/L or gCOD/L or production rate. pH values were either not adjusted, adjusted at time 0, spiked (period adjustment), or continuously controlled. Time given in units listed. Notes describes any engineering or optimization goal, or special condition.

References

1. Mehetre, G., Shah, M., Dastager, S. G. & Dharne, M. S. Untapped bacterial diversity and metabolic potential within Unkeshwar hot springs, India. *Arch. Microbiol.* **200**, 753–770 (2018).

2. Bushra, R. *et al.* Untapped rich microbiota of mangroves of Pakistan: diversity and community compositions. *Folia Microbiol. (Praha)* **69**, 595–612 (2024).

3. Cabral, L. *et al.* Gut microbiome of the largest living rodent harbors unprecedented enzymatic systems to degrade plant polysaccharides. *Nat. Commun.* **13**, 629 (2022).

4. Bahram, M. *et al.* Structure and function of the global topsoil microbiome. *Nature* **560**, 233–237 (2018).

5. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, (2015).

6. Martinez-Feria, R. *et al.* Genetic remodeling of soil diazotrophs enables partial replacement of synthetic nitrogen fertilizer with biological nitrogen fixation in maize. *Sci. Rep.* **14**, 27754 (2024).

7. Shepherd, E. S., DeLoache, W. C., Pruss, K. M., Whitaker, W. R. & Sonnenburg, J. L. An exclusive metabolic niche enables strain engraftment in the gut microbiota. *Nature* **557**, 434–438 (2018).

8. Jiang, L.-L., Zhou, J.-J., Quan, C.-S. & Xiu, Z.-L. Advances in industrial microbiome based on microbial consortium for biorefinery. *Bioresour. Bioprocess.* **4**, 11 (2017).

9. Mesa, J. A., Sierra-Fontalvo, L., Ortegon, K. & Gonzalez-Quiroga, A. Advancing circular bioeconomy: A critical review and assessment of indicators. *Sustain. Prod. Consum.* **46**, 324–342 (2024).

10. Gazzola, G. *et al.* Biorefining food waste through the anaerobic conversion of endogenous lactate into caproate: A fragile balance between microbial substrate utilization and product inhibition. *Waste Manag.* **150**, 328–338 (2022).

11. Wu, S.-L., Wei, W., Wang, Y., Song, L. & Ni, B.-J. Transforming waste activated sludge into medium chain fatty acids in continuous two-stage anaerobic fermentation: Demonstration at different pH levels. *Chemosphere* **288**, 132474 (2022).

12. Zhang, W. *et al.* Bioconversion of swine manure into high-value products of medium chain fatty acids. *Waste Manag.* **113**, 478–487 (2020).

13. Lawson, C. E. Retooling Microbiome Engineering for a Sustainable Future. *mSystems* **6**, 10.1128/msystems.00925-21 (2021).

14. Common principles and best practices for engineering microbiomes | Nature Reviews Microbiology. https://www.nature.com/articles/s41579-019-0255-9.

15. Leggieri, P. A. *et al.* Integrating Systems and Synthetic Biology to Understand and Engineer Microbiomes. *Annu. Rev. Biomed. Eng.* **23**, 169–201 (2021).

16. Huang, C. B., Alimova, Y., Myers, T. M. & Ebersole, J. L. Short- and medium-chain fatty acids exhibit antimicrobial activity for oral microorganisms. *Arch. Oral Biol.* **56**, 650–654 (2011).

17. Desbois, A. P. & Smith, V. J. Antibacterial free fatty acids: activities, mechanisms of action and biotechnological potential. *Appl. Microbiol. Biotechnol.* **85**, 1629–1642 (2010).

18. Seo, J.-H., Lee, S.-M., Lee, J. & Park, J.-B. Adding value to plant oils and fatty acids: Biological transformation of fatty acids into ω-hydroxycarboxylic, α,ω-dicarboxylic, and ω-aminocarboxylic acids. *J. Biotechnol.* **216**, 158–166 (2015).

19. Li, G. *et al.* Advances in microbial production of medium-chain dicarboxylic acids for nylon materials. *React. Chem. Eng.* **5**, 221–238 (2020).

20. R. Beller, H., Soon Lee, T. & Katz, L. Natural products as biofuels and bio-based chemicals: fatty acids and isoprenoids. *Nat. Prod. Rep.* **32**, 1508–1526 (2015).

21. Wagner, H., Luther, R. & Mang, T. Lubricant base fluids based on renewable raw materials: Their catalytic manufacture and modification. *Appl. Catal. Gen.* **221**, 429–442 (2001).

22. Hou, C. T. *Handbook of Industrial Biocatalysis*. (CRC Press, 2005).

23. Tesser, R. *et al.* Oleochemistry Products. in *Industrial Oil Plant: Application Principles and Green Technologies* (eds. Li, C., Xiao, Z., He, L., Serio, M. D. & Xie, X.) 201–268 (Springer, Singapore, 2020). doi:10.1007/978-981-15-4920-5_8.

24. Gao, Y. *et al.* Recent Advances in Intensified Ethylene Production—A Review. *ACS Catal.* **9**, 8592–8621 (2019).

25. World Energy Outlook 2020 – Analysis. *IEA* https://www.iea.org/reports/world-energy-outlook-2020.

26. Oilseeds and oilseed products | OECD-FAO Agricultural Outlook 2021-2030 | OECD iLibrary. https://www.oecd-ilibrary.org/sites/5c08c50a-en/index.html?itemId=/content/component/5c08c50a-en.

27. Fargione, J., Hill, J., Tilman, D., Polasky, S. & Hawthorne, P. Land Clearing and the Biofuel Carbon Debt. *Science* **319**, 1235–1238 (2008).

28. Reddy, M. V., Mohan, S. V. & Chang, Y.-C. Medium-Chain Fatty Acids (MCFA) Production Through Anaerobic Fermentation Using *Clostridium kluyveri*: Effect of Ethanol and Acetate. *Appl. Biochem. Biotechnol.* **185**, 594–605 (2018).

29. Kim, H., Kang, S. & Sang, B.-I. Metabolic cascade of complex organic wastes to medium-chain carboxylic acids: A review on the state-of-the-art multi-omics analysis for anaerobic chain elongation pathways. *Bioresour. Technol.* **344**, 126211 (2022).

30. OECD-FAO Agricultural Outlook 2021-2030. https://www.oecd-ilibrary.org/agriculture-and-food/oecd-fao-agricultural-outlook-2021-2030_19428846-en.

31. Yue, D., You, F. & Snyder, S. W. Biomass-to-bioenergy and biofuel supply chain optimization: Overview, key issues and challenges. *Comput. Chem. Eng.* **66**, 36–56 (2014).

32. Yadav, P., Priyanka, P., Kumar, D., Yadav, A. & Yadav, K. Bioenergy Crops: Recent Advances and Future Outlook. in *Prospects of Renewable Bioprocessing in Future Energy Systems* (eds. Rastegari, A. A., Yadav, A. N. & Gupta, A.) 315–335 (Springer International Publishing, Cham, 2019). doi:10.1007/978-3-030-14463-0_12.

33. Slater, S., Keegstra, K. & Donohue, T. J. The US Department of Energy Great Lakes Bioenergy Research Center: Midwestern Biomass as a Resource for Renewable Fuels. *BioEnergy Res.* **3**, 3–5 (2010).

34. Scarborough, M. J. *et al.* Increasing the economic value of lignocellulosic stillage through medium-chain fatty acid production. *Biotechnol. Biofuels* **11**, 200 (2018).

35. De Groof, V., Coma, M., Arnot, T., Leak, D. J. & Lanham, A. B. Medium Chain Carboxylic Acids from Complex Organic Feedstocks by Mixed Culture Fermentation. *Molecules* **24**, 398 (2019).

36. Kallscheuer, N., Polen, T., Bott, M. & Marienhagen, J. Reversal of β-oxidative pathways for the microbial production of chemicals and polymer building blocks. *Metab. Eng.* **42**, 33–42 (2017).

37. Scarborough, M. J., Hamilton, J. J., Erb, E. A., Donohue, T. J. & Noguera, D. R. Diagnosing and Predicting Mixed-Culture Fermentations with Unicellular and Guild-Based Metabolic Models. *mSystems* **5**, e00755-20.

38. Spirito, C. M., Richter, H., Rabaey, K., Stams, A. J. & Angenent, L. T. Chain elongation in anaerobic reactor microbiomes to recover resources from waste. *Curr. Opin. Biotechnol.* **27**, 115–122 (2014).

39. San-Valero, P., Abubackar, H. N., Veiga, M. C. & Kennes, C. Effect of pH, yeast extract and inorganic carbon on chain elongation for hexanoic acid production. *Bioresour. Technol.* **300**, 122659 (2020).

40. Xie, S. *et al.* Anaerobic caproate production on carbon chain elongation: Effect of lactate/butyrate ratio, concentration and operation mode. *Bioresour. Technol.* **329**, 124893 (2021).

41. Zagrodnik, R., Duber, A., Łężyk, M. & Oleskowicz-Popiel, P. Enrichment Versus Bioaugmentation—Microbiological Production of Caproate from Mixed Carbon Sources by Mixed Bacterial Culture and *Clostridium kluyveri*. *Environ. Sci. Technol.* **54**, 5864–5873 (2020).

42. Bornstein, B. T. & Barker, H. A. The energy metabolism of *Clostridium kluyveri* and the synthesis of fatty acids. *J. Biol. Chem.* **172**, 659–669 (1948).

43. Cabral, L. da S. & Weimer, P. J. *Megasphaera elsdenii*: Its Role in Ruminant Nutrition and Its Potential Industrial Application for Organic Acid Biosynthesis. *Microorganisms* **12**, 219 (2024).

44. Otten, J. K., Zou, Y. & Papoutsakis, E. T. The potential of caproate (hexanoate) production using *Clostridium kluyveri* syntrophic cocultures with *Clostridium acetobutylicum* or *Clostridium saccharolyticum*. *Front. Bioeng. Biotechnol.* **10**, (2022).

45. Weimer, P. J., Nerdahl, M. & Brandl, D. J. Production of medium-chain volatile fatty acids by mixed ruminal microorganisms is enhanced by ethanol in co-culture with *Clostridium kluyveri*. *Bioresour. Technol.* **175**, 97–101 (2015).

46. Candry, P. *et al.* Mildly acidic pH selects for chain elongation to caproic acid over alternative pathways during lactic acid fermentation. *Water Res.* **186**, 116396 (2020).

47. Richter, H., Molitor, B., Diender, M., Sousa, D. Z. & Angenent, L. T. A Narrow pH Range Supports Butanol, Hexanol, and Octanol Production from Syngas in a Continuous Co-culture of *Clostridium ljungdahlii* and *Clostridium kluyveri* with In-Line Product Extraction. *Front. Microbiol.* **7**, (2016).

48. Connors, B. M. *et al.* Control points for design of taxonomic composition in synthetic human gut communities. *Cell Syst.* **14**, 1044-1058.e13 (2023).

49. Sulaiman, J. E. *et al.* Human gut microbiota interactions shape the long-term growth dynamics and evolutionary adaptations of *Clostridioides difficile*. *bioRxiv* 2024.07.15.603560 (2024) doi:10.1101/2024.07.15.603560.

50. Bucci, V. *et al.* MDSINE: Microbial Dynamical Systems INference Engine for microbiome time-series analyses. *Genome Biol.* **17**, 121 (2016).

51. Sanchez-Gorostiaga, A., Bajić, D., Osborne, M. L., Poyatos, J. F. & Sanchez, A. High-order interactions distort the functional landscape of microbial consortia. *PLOS Biol.* **17**, e3000550 (2019).

52. Skwara, A. *et al.* Statistically learning the functional landscape of microbial communities. *Nat. Ecol. Evol.* **7**, 1823–1833 (2023).

53. Thompson, J. C., Zavala, V. M. & Venturelli, O. S. Integrating a tailored recurrent neural network with Bayesian experimental design to optimize microbial community functions. *PLOS Comput. Biol.* **19**, e1011436 (2023).

54. Thompson, J., Johansen, R., Dunbar, J. & Munsky, B. Machine learning to predict microbial community functions: An analysis of dissolved organic carbon from litter decomposition. *PLOS ONE* **14**, e0215502 (2019).

55. Beck, D. & Foster, J. A. Machine Learning Techniques Accurately Classify Microbial Communities by Bacterial Vaginosis Characteristics. *PLOS ONE* **9**, e87830 (2014).

56. Venturelli, O. S. *et al.* Deciphering microbial interactions in synthetic human gut microbiome communities. *Mol. Syst. Biol.* **14**, (2018).

57. Esquivel-Elizondo, S. *et al.* The Isolate Caproiciproducens sp. 7D4C2 Produces n-Caproate at Mildly Acidic Conditions From Hexoses: Genome and rBOX Comparison With Related Strains and Chain-Elongating Bacteria. *Front. Microbiol.* **11**, (2021).

58. Fernández-Blanco, C., Pereira, A., Veiga, M. C., Kennes, C. & Ganigué, R. Comprehensive comparative study on *n*-caproate production by *Clostridium kluyveri*: batch vs. continuous operation modes. *Bioresour. Technol.* **408**, 131138 (2024).

59. Fernández-Blanco, C., Veiga, M. C. & Kennes, C. Carbon dioxide as key player in chain elongation and growth of *Clostridium kluyveri*: Insights from batch and bioreactor studies. *Bioresour. Technol.* **394**, 130192 (2024).

60. Lauer, I., Philipps, G. & Jennewein, S. Metabolic engineering of *Clostridium ljungdahlii* for the production of hexanol and butanol from CO_2 and H_2. *Microb. Cell Factories* **21**, 85 (2022).

61. Weimer, P. J. & Moen, G. N. Quantitative analysis of growth and volatile fatty acid production by the anaerobic ruminal bacterium *Megasphaera elsdenii* T81. *Appl. Microbiol. Biotechnol.* **97**, 4075–4081 (2013).

62. Mehrer, C. R., Incha, M. R., Politz, M. C. & Pfleger, B. F. Anaerobic production of medium-chain fatty alcohols via a β-reduction pathway. *Metab. Eng.* **48**, 63–71 (2018).

63. Wu, J., Zhang, X., Xia, X. & Dong, M. A systematic optimization of medium chain fatty acid biosynthesis via the reverse beta-oxidation cycle in *Escherichia coli*. *Metab. Eng.* **41**, 115–124 (2017).

64. Bäumler, M., Schneider, M., Ehrenreich, A., Liebl, W. & Weuster-Botz, D. Synthetic co-culture of autotrophic *Clostridium carboxidivorans* and chain elongating *Clostridium kluyveri* monitored by flow cytometry. *Microb. Biotechnol.* **15**, 1471–1485 (2022).

65. Benito-Vaquerizo, S. *et al.* Modeling a co-culture of *Clostridium autoethanogenum* and *Clostridium kluyveri* to increase syngas conversion to medium-chain fatty-acids. *Comput. Struct. Biotechnol. J.* **18**, 3255–3266 (2020).

66. Duber, A., Zagrodnik, R., Juzwa, W., Gutowska, N. & Oleskowicz-Popiel, P. Simultaneous medium chain carboxylic acids and 1,3-propanediol production in a bioaugmented lactate-based chain elongation induced with glycerol. *Bioresour. Technol.* **393**, 130123 (2024).

67. Yin, Y., Hu, Y. & Wang, J. Co-fermentation of sewage sludge and lignocellulosic biomass for production of medium-chain fatty acids. *Bioresour. Technol.* **361**, 127665 (2022).

68. 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. *Front. Microbiol.* **11**, (2020).

69. Scarborough, M. J., Lawson, C. E., Hamilton, J. J., Donohue, T. J. & Noguera, D. R. Metatranscriptomic and Thermodynamic Insights into Medium-Chain Fatty Acid Production Using an Anaerobic Microbiome. *mSystems* **3**, (2018).

Chapter 2: Computational models for predicting MCFA production and exploring complex sample spaces

A modified version of this manuscript will be submitted for publication as:

Hayes, M. M., Nightingale, N., Overmeyer, K., Thompson, J, Feng, J., Coon, J., Venturelli, O.S. Exploring the landscape of medium chain fatty acid production of a synthetic anaerobic bacterial community using a bottom-up systems biology approach. Nature Communications.

Contributions: Madeline Hayes conducted the experiments, wrote the text, and created the figures. Nicole Nightingale and Katie Overmeyer processed metabolomics samples. Jaron Thompson wrote portions of the modeling code and contributed to the text. Jun Feng determined the culturing conditions. Josh Coon oversaw the metabolomics sample processing. Ophelia Venturelli revised the text and advised on the experiments.

Abstract

Microbial communities are highly complex dynamical systems, consisting of thousands of species and metabolites that interact with each other and their environment. In order to effectively leverage the power of these systems through engineering, we must understand how individual species in the community interact with each other, and how those interactions are mediated. Microbial bioprocessing is currently being explored as a production method for industrially relevant chemicals, incorporating organic waste streams for sustainable production of valuable products. Naturally occurring microbiomes are not tractable for this purpose at scale, as they consist of many uncultured or unknown organisms that cannot be readily manipulated, and the conditions which drive community output are often unknown. While certain well-characterized chassis species have been explored for production of medium-chain fatty acids (MCFAs), understanding and harnessing communities for this purpose can improve productivity, reduce required inputs, and boost stability of the system. Here, we investigate a synthetic anaerobic bacterial community built to emulate a naturally assembled microbiome for MCFA production. We leverage computational models to quantify, describe, and predict microbial behavior, with a focus on the species-species interactions that impact community composition and functional output. In combination with high-throughput experimentation, we explore a potential sample space of over 65,000 possible species combinations, and identify communities which show reliably high production of the medium chain fatty acid hexanoate, leveraging the function of the well-known MCFA producer *Clostridium kluyveri*. Interpretation of the model parameters identified specific species interactions which positively and negatively impact community phenotype, specifically lactobacilli. We identified a diverse panel of high-producing communities for further study.

Importance

The development of fast, inexpensive, high throughput sequencing of microbial communities has allowed for an unprecedented phase of characterization of the microbial world which surrounds us. Identification of new species and unique profiles has led to the will to leverage newly discovered chemistries for engineering applications. In order to understand how to recapitulate the observed phenotypes, a clear understanding of the interaction network of individual species must be achieved. By demonstrating a generalizable framework for translating biological information in a synthetic community context into quantitative representations which can then be understood using statistical models, we provide a useful tool for microbiome researchers to apply to disparate systems. By identifying key functional interactions in a MCFA-producing community, we reveal mechanistic themes that can inform hypotheses for future testing.

Introduction

Microbes live and thrive in a wide variety of environments, carrying out diverse biochemical reactions which shape the environment around them. Growing and living together, multispecies microbial communities are dynamic systems where millions of individual cells carrying out metabolic reactions can collectively transform the landscape they inhabit by distributing metabolic labor across species.[1–4] Interactions among species define the functionality of communities, improving metabolic efficiency and resiliency to perturbations.[5,6] Harnessing and directing their biochemical potential towards anthropogenic challenges in agriculture,[7–9] healthcare,[10,11] and industry[12,13] is a promising approach for developing sustainable and effective solutions, but the manipulation of natural microbiomes presents a practical challenge due to their complexity. To overcome this obstacle, simplified synthetic microbiomes made of well-characterized species can serve as proxies for understanding species-species interactions that drive functions.[14]

For industrial applications, the generation of high-value chemicals by microbes represents a sustainable alternative to traditional production methods. For example, medium chain fatty acids (MCFAs) are valuable chemicals used widely in pharmaceutical production,[15–17] as precursors for various synthetic textiles,[18] fuel additives,[19] lubricants,[20] and biosurfactants,[21] and are currently produced using petrochemical precursors. Identifying economically viable and sustainable sources for MCFAs is a key step in transitioning away from a fossil-fuel based economy. In particular, microbial production of these compounds holds tremendous potential. Microbial production of MCFAs is a long-known and well-established phenomenon, with diverse species capable of producing high concentrations of both even and odd-chain fatty acids as a byproduct of their metabolism.[22–24] Hexanoate (for example, produced by *Clostridium kluyveri,*[25]

(Ck) *Megasphaera elsdenii,*[26] (Me) *Eubacterium limosum,*[27] (Elim), *Pseudoramibacter alactolyticus,*[28] (Pa) the species included in this study) is often the most abundant product, and therefore is often the focus for engineering[29,30] and optimizing individual species for functional applications. However, attempts to utilize microbial communities with multiple species providing functional redundancy -- which may be more efficient at utilizing heterogenous substrates, stable over time, and robust to perturbation than monocultures -- are still largely exploratory and not well-developed for scale up.[31,32]

Identifying broad functional themes in MCFA-producing microbiomes is a critical first step in logically constructing synthetic communities that recapitulate this process. By utilizing multi-omics characterizations of naturally occurring communities under defined conditions, researchers have been able to build a general framework for MCFA production from a variety of substrates, including swine manure,[33,34] beer-making wastewater[13,35] and liquor fermentation mash,[36,37] and of particular interest, biofuel stillage waste.[38] Conversion of rich organic substrates high in plant material relies on the step-wise transformation of complex carbohydrates through fermentative pathways, which convert carbohydrates to short chain fatty acids; these are then elongated into MCFAs via the reverse beta-oxidation (R-BOX) pathway (extensively reviewed[39]). This platform, referred to as the chain elongation platform, utilizes an electron donor such as ethanol or lactic acid to elongate acetate or butyrate two carbons at a time in a reversal of fatty acid catabolism via beta oxidation. Scarborough (2018)[40] provides an informative example of a naturally occurring microbiome being leveraged for the conversion of lignocellulosic biofuel waste to MCFA. The authors utilized a combination of metagenomics and metatranscriptomics to characterize dominant metagenome-assembled genomes and map the most highly expressed genes to these taxonomic units, building an understanding of active

functional species groups and the metabolic pathways most active during peak MCFA production, providing us with a data-driven basis for selecting species for a synthetic community. As in all microbiomes, the presence of individual species or metabolic pathways is not sufficient to drive functional outputs; tuning environmental parameters that make these reactions energetically favorable is a field that is still emerging.[25,41,42] With regards to CE, challenges include the variability of optimal pH[25,43–48] depending on dominant MCFA producer and substrate, and gas headspace composition and partial pressures.[49,50] Use of uncharacterized or environmentally isolated microbial communities make this metabolism difficult to reliably leverage, and an ecological understanding of how multiple species in a community divide the metabolic labor of depolymerizing plant sugars, producing short-chain intermediates, and elongating carbon chains under different environmental regimes is lacking.[51]

Although synthetic microbiomes are emerging as a reasonable proxy for studying more complex natural systems, they are not without their own challenges. Even a relatively simplified synthetic consortia of ten species can be combined to form over one thousand different communities, and the number of possible combinations scales exponentially with the number of species.[52] Including enough species to capture nature's functional redundancy – a key component of microbiome resiliency to perturbation and resistance to invasion – quickly results in an intractable number of communities which cannot all be tested empirically. Narrowing the possible experimental space requires predictive capabilities which can guide data collection. Computational models have been employed to quantify many aspects of microbial communities, from high-level community composition and response to perturbation, to cellular metabolic flux. By leveraging ecological modeling frameworks and high throughput experimentation together, testable hypotheses can emerge from these otherwise intractable experimental spaces.[53–56]

Assembling communities from individual species and assessing their function is known as bottom-up engineering, an approach recently applied by Clark et. al.[57] Using a synthetic community comprised of well-characterized species with known functions to approximate the human gut microbiome, the authors aimed to identify communities which maximize butyrate production, a key metabolite in human gut health. Using a two-stage modeling approach, the authors aimed to estimate positive and negative interactions within the community to predict butyrate production as a function of community composition. First, the generalized Lotka-Volterra model was used to estimate how species-species interactions impacted the abundance of community members and predict composition of untested communities. Then linear regression was used to predict butyrate production from community composition, allowing for the directed exploration of a large sample space. Because statistical models have the advantage of interpretable parameters, the interaction network driving community composition and butyrate production was inferred. This approach and others[58] allows for the identification of interactions stemming from exchange of metabolites between community members and identifies changes in environmental parameters that impact the strength and direction of those interactions, providing a foundation for the approach we use in this study. While successful in predicting communities which rely on a single cross-feeding step (e.g., byproducts of dietary fiber breakdown by one species fuels butyrate production by another), we aim to expand our understanding of the suitability of this approach to multi-step transformations, such as those observed in chain elongation communities, improving our understanding of the interactions that enhance energetic efficiency, resistance to invasion, and robustness to environmental uncertainties.

Here, we integrate the above industrial motivation for producing MCFAs from bioenergy waste streams with high throughput experimental methods that utilize computational models to

drive data collection. The aims of this work are twofold: (1) Establish the viability of two-stage modeling approaches to decipher community function where the desired output requires multiple chemical transformations of substrate and precursors, and identify pairwise species-species interactions within the community that drive this functionality; and (2) Identify high hexanoate producing communities to serve as a basis for understanding mechanisms and favorability of hexanoate production. We utilize two synthetic media designed to approximate biofuel stillage waste (hereafter referred to as conversion residue and the synthetic equivalents, SCR) with differing compositions and pH to identify interactions and mechanisms that are context-dependent. We find that data-driven iterative sample collection leads to successive increases in observed hexanoate production by communities and that our models can accurately predict community composition and function. Further, we demonstrate that different media environments result in differential community compositions with respect to highest observed hexanoate production, implying media composition as driving community assembly and function. This is verified by interpreting model parameters and identifying species-species interactions within larger communities that differentially impact community growth or community function in each medium. Successful application of this generalizable framework builds on a growing body of research that shows the integration of computational modeling with synthetic community experiments can reveal drivers of microbiome functionality in a wide variety of engineering contexts.

Results

Bottom-up design leverages well-defined media and synthetic communities

We aimed to explore the molecular and ecological forces which drive hexanoate production in microbial communities. A bottom-up design approach overcomes the challenges of complex substrates and uncharacterized or inactive species found in natural microbiomes while retaining key environmental components and functional features. By exerting tight control over both substrate and inoculum, we aimed to provide a tractable and reproducible system for applying computational modeling and high throughput experimental approaches to quantitatively assess community dynamics.

Since stillage residue consists of the leftover materials from fermentation and distillation of grains or other feedstocks and can be complex with many unknown components, we used two defined media to enable precise control of components for dissecting potential mechanisms shaping hexanoate production. Our two defined media mirror lignocellulosic stillage waste, containing plant-based carbon sources and amino acids. Because previous characterizations of conversion residue[59] could not determine the chain length of the identified polymeric plant sugars, only monomers of xylose and other hemicellulose components were included. Both recipes contain additional vitamins and minerals to promote growth of fastidious anaerobes (See Supplementary Spreadsheet 1: Media).

Similarly, a synthetic, defined community enables us to leverage a bottom-up approach to interrogate the contribution of individual or combinations of species to community assembly and hexanoate production. We selected well-characterized species with known functions (Figure 1A and 1B) informed by characterization of naturally occurring communities. In particular, one

naturally assembled MCFA-producing community isolated from the acid phase of a two-stage anaerobic digestion system at the wastewater treatment plant in Madison, WI was characterized in CR[40] and served as a framework for species selection, which identified key functional groups responsible for distinct steps in the transformation of stillage to MCFA. In general terms, the first function is to ferment plant sugar monomers to release lactate and acetate (lactobacilli, coded in blue in Figure1A and 1B), followed by fermentation of lactate and acetate to release butyrate (short chain fatty acid [SCFA's] producers, purple), and elongation SCFA's through chain elongation (MCFA producers, orange).We built our community from isolates which are closely related to those groups, opting to include multiple related species to provide functional redundancy in the community (four lactobacilli, six SCFA producers, four MCFA producers); multiple species capable of performing each hypothesized function offers multiple routes of functionality and provides an approximation of the diversity seen in natural communities. Additionally, we added an acetogen *Blautia hydrogenotrophica* (Bh, red), which consumes H_2 and CO_2 and releases acetate,[60,61] hypothesizing that a community member capable of modifying the gas headspace and contributing to the intermediate metabolite pool could support community function. We also included *Eggerthella lenta* (Elen, green), which transforms arginine into ammonia and buffer environmental pH.[62] In total, our 16 species (Figure 1B) can be assembled into >65,000 possible communities, providing a large sample space for exploration.

Model-guided data collection efficiently explores sample space and identifies productive communities

All possible combinations of our 16 species results in an intractably large potential sample space. In order to explore the space and identify high hexanoate-producing communities, we employed a previously developed two-stage iterative modeling and data collection approach[57]

referred to as ‘Design-Test-Learn Cycles’ (DTL) (Figure 1C). We aimed to identify species-species interactions that support or inhibit MCFA production, and therefore chose models which assume that pairwise microbe-microbe interactions impact the growth and metabolite production of the community. By incorporating information about both monospecies and species pairs, the models aim to estimate the importance of species-species interactions in community assembly, metabolite production, and ecological niche availability. In the first stage, species growth is predicted using the Generalized Lotka-Volterra (gLV) model. This model is a set of coupled ordinary differential equations to predict the temporal changes in species abundance as a function of the intrinsic growth of each species and pairwise interaction among all constituent community members. Parameters are estimated using Bayesian parameter inference (see methods). In the second stage, community MCFA production is predicted as a function of community composition (previously predicted by the gLV) using linear regression (LR) with interactions. This model is based on the phenomenological model of metabolite production used in bioprocess engineering, capturing basal metabolite production of each monospecies, as well as the impact of species interactions on metabolite production per unit biomass. We fit a linear regression model using the ElasticNetCV function from Scikit-Learn to perform hyper-parameter optimization of the L1 and L2 parameter penalty (see methods).

With these two models we aim to predict two aspects of community phenotypes: community composition after 48 hours of growth (predicted by gLV), and MCFA production at this 48-hour time point (predicted by linear regression). Data used to train the models consists of two components: sequencing data to determine the identity and proportion of each species in the community, and metabolomics data to measure the production of hexanoate (C6) and octanoate (C8). Species absolute abundances are calculated by measuring the fractional (relative)

abundance of each species by 16S rRNA gene sequencing and multiplying by the total community absorbance at 600nm (Abs600). Targeted metabolomics to measure C6 and C8 was performed by GC-MS (see methods). After an initial round of data collection to inform the initial model, the models are trained on the data and the prediction performance assessed (see methods). The trained models are used to predict the composition and MCFA production of all remaining untested communities. From these predictions, communities of interest are selected ('Design' phase) for characterization (see next section for design criteria). These communities are assembled by liquid handling robot from each species grown in monoculture, allowing for simultaneous testing of >100 communities, and sequencing and metabolomics data are collected ('Test' phase). The models are then updated; a new parameter set for the gLV and the linear regression model are determined using the full experimental dataset, and all remaining untested communities are predicted using the updated parameters ('Learn' phase). Predictions and interpretation of the parameters then inform the next Design phase. Using this systematic approach, we can efficiently explore a large combinatorial space to fill knowledge gaps in the model and identify communities with high C6 and C8 production. In sum, we used data-driven computational models to systematically explore the community design space to identify important inter-species interactions that mediate MCFA production.

Diverse community selection criteria allows for effective exploration of complex sample space

We completed four DTL cycles in each media, aiming to iteratively improve C6 and C8 production. Design criteria for selecting communities in each cycle balanced choosing communities predicted to maximize metabolite production, and providing the models with informative community examples (Figure 2A and Figure 2F). The initial dataset (cycle 1) was chosen to establish a baseline dataset containing a mix of community sizes to provide the models

with a wide variety of data for training, with 1-4 member communities (low complexity) providing examples of lower order interactions and 14-16 member communities (high complexity) providing examples of higher-order interactions to inform predictions of all other possible combinations. After model training and parameter estimation, we predicted the composition and MCFA production of all remaining untested communities, ranked them in order of C6 or C8 production, and chose communities for cycle 2 testing. Communities with the highest predicted metabolite production (top 37 communities for C6, top 40 communities for C8) were selected as Exploitation Selections (EXS), maximizing predicted metabolite production. Other Selections (OS) were made based on the inclusion of species and species pairs estimated to positively impact C6 and C8 production based on their parameter values, where larger parameter values generally indicate positive impact on C6 or C8 production (see Supplementary Table 2 and next section for parameter interpretations). Cycle 3 repeated these selection criteria. For cycle 4, only EXS communities were chosen (29 C6 predictions and 29 C8 predictions for SCRv1, and 30 C6 predictions and 29 C8 predictions for SCRv2). Additionally, in order to rigorously test whether the model was identifying improved communities and learning relevant interactions, we characterized a set of randomly selected communities to serve as an independent dataset for model validation (validation set, VS). This dataset is not subject to the biases[63–65] of our design cycle and thus provides a unique dataset for determine the prediction performance of our models.

By balancing community selections between Exploitation and informative selections, we aimed to provide the models with informative training data to encourage exploration of the sample space, identifying communities which improve metabolite production through iterartive cycles. We expected to see differences between selection groups, where if the model was

learning effectively from the training data, we would see significant improvement in metabolite production of the EXS groups over cycles. We observed a significant increase in C6 production, especially by EXS communities by cycle 4 (Figure 2B and Figure 2G). Importantly, a significant difference in mean C6 production was observed between cycle 4 EXS and VS selections. This shows that communities identified by the model to maximize C6 production are better than communities chosen at random, indicating an effective exploration of the sample space with regards to highest possible C6 production. No increase in C8 production was observed over DTL cycles (see Supplementary Figures S1 and S2). Interestingly, in SCRv1, both the C6 and C8 predictions tested in cycle 4 showed strong C6 production. In contrast, C8 predictions in cycle 4 in SCRv2 do not show the high C6 production. Given that C6 and C8 have been observed to be coproduced by natural communities in bioreactor settings,[40,66,67] this raises the possibility of conditional control knobs that could push communities towards different outputs.

In order to understand the phylogenetic distribution of observed communities within the design space, we visualized the dataset using MDS, incorporating relatedness of species in each community Figure 2C and Figure 2G). Communities can be represented as binary vectors, where each element of the vector is one or zero depending on whether the corresponding species is present or absent, respectively. However, when considering only the presence or absence of species, the full design space cannot be optimally projected into lower dimensions in a way that preserves distances between points. By augmenting each community vector to capture information about species phylogeny, different community vectors can vary in their alignment depending on phylogenetic similarity (see methods). In this way, we aimed to visualize the phylogenetic relationship between communities selected by the model (i.e. closer clustering indicates greater relatedness). We expect to see the model identifying closely related

communities as it becomes more confident in estimating which species contribute positively to function. We observed that by cycle 4, communities were beginning to show signs of clustering (Figure 2C and 2H, purple points). Interestingly, in SCRv2, cycle 4 EXS communities show strong overlap with communities tested during previous cycles, indicating that C6 production is not further. In contrast, the cycle 4 EXS communities in SCRv1 do not show strong overlap with previous cycle selections, indicating further optimization may be achieved in this media.

To further explore the properties of communities in the dataset with regards to their observed MCFA production and their relationship to other sampled communities, we modified the MDS plot to emphasize functional elements (Figure 2D, Figure 2E, Figure 2I, and Figure 2J). Using the same phylogenetic relationships, we aimed to identify whether C6 and C8 EXS communities were diverging, and how their MCFA production values varied. In SCRv1, we see a cluster of closely related communities which all have strong C6 production, some of which are C6 predictions and some of which are C8 predictions (Figure 2D and Figure 2E). In SCRv2, we see distinct clusters of C6 predictions (Figure 2I) and C8 predictions (Figure 2J). This contrast in relationships between model-identified communities between media types, taken together with the significant difference in maximum observed MCFA production, indicates that environmental factors or media composition impacts the behavior of species as learned by the models, resulting in diverging predictions for MCFA production.

Although the improvement in C6 production meets the goals of this project, we did not observe a similar optimization of C8 production. We therefore chose to continue analysis of the dataset for C6 only. Model performance assessment and interpretation of model parameters was focused on the C6 models, and all subsequent experiments focused on understanding the biology of C6 production. Understanding whether adjustments to the conditions or continued

optimization of community composition could improve C8 production is an avenue of future directions.

Modeling approach accurately learns species-species interactions from data

In order to assess whether this two-stage modeling approach is appropriate for this system, model performance was rigorously assessed during every cycle of data collection (Figure 3A). Once data was collected, both models were trained and tested using a k-folds cross validation approach (see methods). This allows for the models to be trained many times on multiple subsets of the data generating many parameter sets. The ability of these parameter sets to predict the datapoints not used for training gives insight into how accurately the model structure captures information from the data and applies it to unseen observations, known as the Out of Fold performance. Additionally, models can be trained on the whole dataset, and then tested for fit (Best Fit performance), which measures the flexibility of the model to fil all available data (see methods). For both fits, when there is a strong correlation (e.g. Spearman rho) between the observed values for the withheld data and the model's predictions, model performance is said to be good (i.e., rho >0.5). Both fits are important for building confidence in the predictive power of the modeling approach. We expect the performance (correlation between observed and predicted values) to improve over data collection as the models have more data to learn from, but critically, the difference in performance between the two fits should close if the model is not overfitting, and accurately learning generalizable features of the data; calculating both for comparison builds confidence in the interpretations extracted. For fits of both models in both media types, we observe that model performance improves over DTL cycles, and by cycle 4 is consistently high for in-cycle data (assessed by Spearman rho >0.8) (Figure 3B, Figure 3C, Figure 3F, Figure 3G, purple solid and hashed bars).

In addition to assessing performance on the data collected in-cycle, we also assessed the ability of the parameter sets calculated at each cycle to predict the independent Validation Set. The accumulation of communities which are similar in composition may bias the performance of the models even on held out data, so providing a separate dataset that does not resemble the training set is an important step in demonstrating that the model predictions are accurate across the whole sample space. For both models in both media, we observe a VS performance of Spearman rho >0.5 (Figure 3B, Figure 3C, Figure 3F, Figure 3G, grey bars). Understanding why performance on VS communities is lower than in-cycle data is an important step in building confidence in our interpretations. If the VS communities do not resemble the bulk of the dataset with regards to composition, then the parameters generated from training on the bulk dataset will not accurately predict those communities. While we see that the VS is evenly distributed among the dataset when phylogeny is accounted for (Figure 2C and 2H), the VS does tend to differ in the fraction of communities which contain individual species from the bulk dataset (Supplementary Figure S3). The lack of training examples would then help explain why performance on the VS is lower than in-cycle out of fold. For gLV, we can take this analysis one step further by examining the prediction accuracy of individual species abundances, to identify if certain species are particularly difficult to predict (Supplementary FigureS4). Some species show higher MSE on their predictions, particularly Cc, Er, Lb, Ld, and Me, which may be bringing overall performance down for this model. Accruing more examples of these species in diverse contexts may improve model performance. Regardless, the performance on the VS (Spearman rho >0.5) for all models is considered acceptable, and in addition to the high in-cycle performance metrics, this gives us confidence that interpretation of the model parameters for biological information. Additional gLV and C6 LR modeling results which confirm these results

with additional analysis (Pearson correlation, MSE) can be found in Supplementary Figures S5-S7.

Lastly, in order to maximize the amount of information available for biological interpretation, we trained and tested the models on the Whole Dataset (all in-cycle data, plus the validation set) (Figure 3B, Figure 3C, Figure 3F, Figure 3G, green solid and hashed bars). This provides us with the largest dataset for interpreting modeling parameters, improving our ability to discern generalizable features of the compositional and functional landscape. Including the validation data set did not universally improve the model performance, but both models for both media maintained performance of Spearman rho >0.75. The Whole Dataset Best Fit (Figure 3D, Figure 3E, Figure 3H, Figure 3I) was used for all parameter interpretation. The rigor of model performance assessment presented here provides us with confidence that interpretation of the model parameters will identify testable hypotheses in the functional landscape.

Interpretation of model parameters provides basis for hypothesis generation

One advantage of using statistical models to investigate the biology of this system is that the model parameters are directly interpretable (Figure 4A and 4B). The gLV parameters represent intrinsic growth rate, intra-species and inter-species interactions. These interactions are directional, and can be assigned a ‘donor’ species and a ‘receiver’ species. Examples of this kind of interaction include donor species production of a metabolite that can be consumed by the receiver (positive) or production of a toxin (negative). The LR parameters represent basal metabolite production, enhancement of metabolite production (if the species cannot produce the metabolite itself), and production inhibition. Because these are non-directional, donor and receiver relationships cannot be discerned from the parameter value alone. In both cases, strong positive parameters (growth enhancement, production, or production enhancement) are depicted

in blue, while strong negative parameters (growth inhibition or production inhibition) are depicted in red.

For interpreting parameters we focus on two functional groups: the lactobacilli and MCFA producers. For gLV (Figure 4C and Figure 4D), we see that lactobacilli donate and receive many negative growth interactions, except for Ld, which receives several positive interactions in both media, and Lv, which donates a few positive interactions in SCRv2 (Figure 4D) but not SCRv1 (Figure 4C). This indicates that not all members of this functional group behave similarly, implying non-redundancy of their functions. For MCFA producers, one species which stands out is Me. The strength of donated interactions for this species differs greatly between media, with near-zero values dominating SCRv1 (Figure 4C), but many strong values in SCRv2 (Figure 4D). This underscores the idea that behavior of individual species is very environmentally dependent.

For LR parameters (Figure 4E and Figure 4F), MCFA producers and lactobacilli again stand out. To begin, Ck is identified as the largest single contributor to C6 production in either media. This is not surprising, given the well-established ability of Ck to produce C6, but does underscore that it is likely the primary producer in both media. In SCRv1 (Figure 4E), Me and Pa also have strong positive values, but this is not true in SCRv2 (Figure 4F). The differential behavior of some, but not all, MCFA producers could be a target for understanding how environment shapes MCFA production metabolism in different species. We can more deeply probe this question by examining which partner species have positive and negative impacts on MCFA producers. In SCRv2 (Figure 4F), Ck-lactobacilli interactions are all negative, and in SCRv1, two of the four parameters are very strongly negative. This indicates that in general, lactobacilli can negatively impact Ck C6 production, regardless of their *growth* impact; in

SCRv2, Lv *donates* a strong positive growth interaction to Ck, but when these two species appear together, they negatively contribute to community C6 production. This highlights that interactions which improve growth of individual species may not be beneficial to their metabolite production, and tradeoffs between biomass and target function are context-dependent.

By comparing the SCRv1 and SCRv2 parameter values (Figure 4G and 4H), we can start to discern patterns which will inform functional hypotheses, especially with regards to the MCFA producers-lactobacilli dynamics described above. For gLV parameters (Figure 4G), we see that parameters Ck is receiving a growth interaction are usually negative in SCRv1 (left half of plot), but can be either positive or negative in SCRv2. For this reason we would expect Ck to be competing with more species in SCRv1, and individual species dynamics can be tested with leave-one-out experiments (see next Chapter). Further, we see Lb- and Ld- donating interactions mostly negative in both media types (lower left quadrant). For this reason we would expect to see the removal of these species positively impact multiple other species in the community. These patterns are also present in the LR parameters (Figure 4H), where most MCFA-lactobacilli pairs have negative parameter values in SCRv2 (bottom half), but near-zero values in SCRv1. We would expect to see any lactobacilli-mediated effect more prominently in SCRv2 in this case. All of these hypotheses are investigated in depth in the next Chapter.

Discussion

Integration of computational modeling and high-throughput experimentation has the potential to combine two advances in microbiology experimentation: greatly improved computational power capable of analysis of vast and previously intractable sample spaces, and rapid, accurate characterization of defined communities through next generation sequencing. In combination with metabolomics, we can use these tools to begin to parse the relationships between community composition and functional output. In this Chapter I aimed to demonstrate that a two-stage model and iterative data collection approach is effective for exploring the complex sample space of a 16 member synthetic bacterial community. This approach has been demonstrated in a synthetic human gut community for understanding production of butyrate, but whether it is appropriate to model a more complicated, multi-step metabolite transformation was unknown. Further, to our knowledge, the use of such a large community for understanding MCFA production is novel, and use of a synthetic community to approximate more complex, naturally occurring MCFA-producing communities has not been attempted.

We aimed to model the production of two target metabolites: hexanoate (C6) and octanoate (C8). Both metabolites are industrially valuable and co-produced by R-BOX metabolism, and we aimed to understand if there were distinct mechanisms or species-species interactions which could shift community production towards either metabolite. While the modeling and optimization of C6 production proved successful, we were unable to identify any communities capable of improving C8 production. Because so few communities produced C8, the modeling results were also unable to capture key interactions and demonstrate good predictive performance, and thus attempting to interpret the parameter values is not an appropriate approach for improving production of this metabolite. Given the opportunity to

explore a wider range of environmental parameters in a more controlled environment such as a benchtop bioreactor, it may be possible to start to understand if the synthetic community explored here is even capable of reliably producing octanoate, but whether conditions can be reliably achieved in a high enough throughput to apply a modeling approach which relies on many observations is unknown. This demonstrates a key limitation of this modeling approach.

We also measured and modeled the production of butyrate (Supplementary Figures S8-S9). The synthetic community contains many butyrate producers, and we expected to see this as a normal product of metabolism. Given that Scarborough 2018 identified SCFA's as important intermediates, we reasoned that modeling this metabolite could be helpful in understanding the flow of carbon and energy in the system. The modeling approach for butyrate showed good out of fold performance (cycle 4 and whole dataset Spearman rho > 0.75), the validation set performance was less reliable (Spearman rho <0.5 at cycle 4). Although we did observe co-production of butyrate and hexanoate by Ck monospecies specifically (Figure S10), we did not optimize for this function and obvious hypotheses to test did not emerge. Understanding butyrate as a product of R-BOX represents an avenue of future investigation.

Flow of metabolites from plant sugars to MCFA through short chain intermediates was a key element of Scarborough 2018. Species predicted to produce acetate and lactate could play a key role in supporting C6 production. Through parameter interpretation, we identified the lactobacilli as a functional group of interest, and hypothesize that their role in community composition and function could be revealed by Leave-One-Out experiments. Lactobacilli produce two key intermediates for this system: acetate and lactate. Acetate is a primary carbon source for Ck, and lactate is a primary carbon source for Me. Interestingly, the LR parameter values containing both lactobacilli and MCFA producers (specifically lacto-Ck interactions) tend

to be negative (Figure S10), refuting the hypothesis that intermediate production will support MCFA output. Understanding how these species are negatively impacting hexanoate production will require expanding our understanding of community function to encompass environmental conditions in which MCFA production is favorable in addition to the available intermediates.

Other LR parameter value interpretations indicate that some species could play a role in supporting MCFA production (Figure S10). For example, SCRv1 shows Bh-Ck as the second highest parameter value and positive contributor to C6 production, only behind Ck. Bh produces acetate from H_2 and CO_2; Ck consumes acetate to produce C6, and this could be a source of cross feeding. While communities which contain Bh do on average produce more C6 than communities without Bh (Figure S12), this specific interaction will need to be investigated more rigorously to draw conclusions.

Although interpreting parameter values provides a starting point for understanding community ecology, the values alone do not offer any true molecular mechanistic insight. The integration of prior knowledge regarding the biology of individual species and their roles in communities is required to transform these insights into actionable hypotheses. synthetic biology methods are suited to narrowing the community sample space to our highest producing communities while broadening the depth and amount of data collected per community will expand the picture of the MCFA production landscape for this community.

Methods

Bacterial Growth and Sample Collection

Strain Maintenance and Preculturing – All anaerobic culturing was performed in an anaerobic chamber (Coy) with an atmosphere of 2.5 ± 0.5% H_2, 15 ± 1% CO_2 and balance N_2. All prepared media and materials were transferred into the chamber at least overnight before use to equilibrate to anaerobic atmosphere. Supplementary Table 1 details strain source and individual culturing conditions. Permanent stocks of each species were maintained at -80C in 25% glycerol. Batches of single use glycerol stocks (SUGS) were produced for each strain by first isolating a single colony from the permanent stock on anaerobic basal broth (ABB) or MRS solid agar plate, then inoculating 5-7mL liquid media and incubated at 37°C until log phase, mixed with equal volume 50% glycerol, aliquoted (400µL) into Matrix Tubes (Thermo Fisher), and stored at -80°C until use. SUGS parent cultures were 16S Illumina sequenced to verify cµLture purity. For each experiment, a serial culturing strategy was used to ensure ample growth and similar growth phase of each species prior to experimental inoculation (see Supplementary Methods).

Monoculture dynamic culturing – Precultured cells were pelleted at 4000RPM then resuspended in experimental media (compositions in Supplement spreadsheet 1: media recipes) at normalized OD600=0.5 for inoculation. Briefly, cultures were diluted to a final OD of 0.1 and volume of 1mL and incubated at 37°C for 48 hours covered with a semipermeable membrane, and sampled (100µL) every four hours for OD600 measurement (Tecan F200). OD was measured in two ways to ensure samples were in dynamic range: a) aliquoting 100µL into 96-well microplate (96M) or aliquoting 20µL into 180µL blank media in 96M. Individual replicates were used for each time

point to avoid adverse effects from disturbing the culture during growth. Supernatant and cells were harvested at 48 hours (see Sample Collection below).

Community batch culturing - For each experiment, precultured cells were pelleted at 4000rpm then resuspended in experimental media (compositions in Supplement spreadsheet 1: media recipes) at normalized OD600=0.5 for inoculation. Community combinations were arrayed in 96DW plates by pipetting each species at equal volume into the appropriate well (Tecan Evo Liquid Handling Robot). Each community was then diluted to a final OD of 0.1 and volume of 1.0mL and incubated at 37°C for 48 hours covered with a semipermeable membrane (Breathe Easy). At 48 hours, final OD600 was measured in two ways to ensure samples were in dynamic range: a) aliquoting 100µL into 96M or aliquoting 20µL into 180µL blank media. Supernatant and cells were harvested at 48 hours (see below).

Sample collection – After each experiment, cells and supernatant were harvested and stored at -80°C. Briefly, cells were pelleted in 96 deep well plate at 4000rpm and supernatant removed without disturbing the pellet. Both cells and supernatant were stored at -80°C until processing.

Sample Processing

Genomic DNA extraction and sequencing library preparation – Collected cell pellets were lysed for extraction and purification of genomic DNA using a modified Qiagen Blood and Tissue kit (cat). Briefly, cells thawed in a room temperature water bath, then resuspended in 20 mg/mL lysozyme (from chicken egg whites Sigma) in enzymatic lysis buffer (20mM Tris-HCl [Invitrogen], 2mM sodium EDTA [Sigma], 1.2% Triton X-100 [Sigma]). Plates were covered in a foil seal and allowed to incubate for 30 min at 37°C with orbital shaking at 600rpm. Then, 25µL 20 mg/mL proteinase K solution and 200 µL Qiagen Buffer AL were added and incubated

for 30min at 56°C with orbital shaking at 600rpm. DNA was precipitated by adding 200µL 100% EtOH, then purified over Pall DNA-binding column plates. Qiagen AW1 and AW2 500µL washes were performed and the column allowed to dry for 5min at room temperature. DNA was eluted with 110µL Qiagen buffer AE preheated to 56°C. Samples were stored at -20°C until further use.

Genomic DNA concentrations were measured using a SYBR Green fluorescence assay (Invitrogen) according to the manufacturer's instructions and then normalized to a concentration of 1 ng/µL by diluting in molecular grade water using a Tecan Evo Liquid Handling Robot. Briefly, genomic DNA samples were removed from −20 °C and thawed in a room temperature water bath and combined with 95 µL of SYBR Green diluted by a factor of 100 in TE buffer (Integrated DNA Technologies) in a black 384-well microplate. A standard curve was constructed in triplicate using 5µL of standard concentrations of 0, 0.5, 1, 2, 4, and 6 ng/µL. Fluorescence with an excitation/emission of 485/535 nm was measured using a Tecan Spark plate reader. Concentrations of each sample were calculated using the standard curve and a custom Python script was used to compute the dilution factors and write a worklist for the Tecan Evo Liquid Handling Robot to normalize each sample to 1 ng µL−1 in molecular grade water. Samples with DNA concentration <1 ng µL−1 were not diluted. Normalized genomic DNA samples were stored at −20 °C until further processing.

Normalized genomic DNA was then used for 16S rRNA gene amplification by PCR using Invitrogen Phusion and custom dual-indexed primers (see Supplementary spreadsheet 2: primers). Primers were arrayed in skirted 96-well PCR plates (Thomas Scientific) using an acoustic liquid handling robot (Labcyte Echo 550) such that each well received a different combination of one forward and one reverse primer (0.1 µL of each). After liquid evaporated, dry

primers were stored at −20 °C. Primers were resuspended in 15 µL PCR master mix (0.2 µL Phusion High Fidelity DNA Polymerase (Thermo Scientific), 0.4 µL 10 mM dNTP solution (New England Biolabs), 4 µL 5× phusion HF buffer (Thermo Scientific), 4 µL 5 M Betaine (Sigma-Aldrich), 6.4 µL Water) and 5 µL of normalized genomic DNA to give a final concentration of 0.05 µM of each primer. Primer plates were sealed with Microplate B seals (Bio-Rad) and PCR was performed using a Bio-Rad C1000 Thermal Cycler with the following program: initial denaturation at 98 °C (30 s); 25 cycles of denaturation at 98 °C (10 s), annealing at 60 °C (30 s), extension at 72 °C (60 s); and final extension at 72 °C (10 min). 2 µL of PCR products from each well were pooled and purified using the DNA Clean & Concentrator (Zymo) and eluted in water. The resulting libraries were sequenced on an Illumina MiSeq using a MiSeq Reagent Kit v3 (600-cycle) according to the manufacturer's instructions to generate 2 × 300 paired-end reads.

Quantification of species abundance - Sequencing data were demultiplexed using Basespace Sequencing Hub's FastQ Generation program. Custom python scripts were used for further data processing (method adapted from Venturelli et al. Mol. Syst. Biol., 2018) Paired end reads were merged using PEAR (v0.9.10)83 after which reads without forward and reverse annealing regions were filtered out. A reference database of the V3–V5 16S rRNA gene sequences was created using consensus sequences from next-generation sequencing data or Sanger sequencing data of monospecies cultures. Sequences were mapped to the reference database using the mothur (v1.40.5)84 command classify.seqs (Wang method with a bootstrap cutoff value of 60). Relative abundance was calculated as the read count mapped to each species divided by the total number of reads for each condition. Absolute abundance of each species was calculated by multiplying the relative abundance by the OD600 measurement for each sample. Samples were excluded from further analysis if >1% of the reads were assigned to a species not expected to be

in the community or if they had >1% unclassified reads (indicating contamination). In-house validation of this method in response to previous reviewer comments has confirmed that neither number of RNA gene copies nor dead cell debris in the media does not impact abundance calculation, and this is a reasonable proxy for absolute abundance (data not shown).

Organic extraction of MCFAs – MCFAs were extracted from supernatants using a liquid-liquid method. Briefly, supernatants were thawed in a room-temperature water bath and 75µL of supernatant was combined with 150µL chilled acidified acetonitrile (Sigma) (0.002% v/v glacial hydrochloric acid [Fisher]) in 96DW (foil sealed) and incubated at 4 °C for 20min with orbital shaking at 600rpm. Each sample was then combined with 150µL ethyl acetate (Fisher) and mixed by pipetting. Samples were then spun down at 4000rpm at 4 °C for 5 minutes, separating the organic and aqueous phases. Then, 100µL of the organic (top) phase was aliquoted into HPLC vials (ResTek) and sealed with screw cap with rubber septa (ResTek) and immediately proceeded to GC-MS analysis.

Measurements of MCFAs

Standard Stock Preparation - Individual solutions of Butyric Acid, Hexanoic Acid, and Octanoic Acid were purchased from Sigma-Aldrich with purity no less than 99%. 10 µL of each organic acid was aliquoted into a 1.5 mL vial followed by 970 µL of acidified acetonitrile (0.2% HCl) solution to create a 10000 ppm stock solution. The stock solution was used to prepare standards at concentrations of 5 ppm, 12.5 ppm, 25 ppm, 50 ppm, 75 ppm, 100 ppm, 150 ppm, 250 ppm, 500 ppm, 1000 ppm, 2000 ppm, 4000 ppm and 5000 ppm in 100 µL acidified acetonitrile (0.2% HCl). The standards were further diluted by the addition of 200 µL ethyl acetate. Samples were

thoroughly mixed and centrifuged at 12,000 x g for 5 min at 4°C. 100 µL of upper layer was transferred to amber glass autosampler vial with fused glass insert for analysis.

GC-MS Analysis - Samples were analyzed using a GC-MS instrument set up to perform chemical ionization comprising a Trace 1310 GC coupled to an ISQ series mass spectrometer (Thermo Scientific) with methane as the ionization gas. Analytes were injected into a split/splitless heated injector at a temperature of 200 °C using an AI1310 autosampler. The samples were split 1:5 by the injector and then injected onto a 30 m Stabilwax DA column (Restek) using helium at a flow of 1.20 mL/min. A ramped temperature mode was employed using the following gradient:

Number	Retention time [min]	Rate [°C/min]	Target value [°C]	Hold time [min]
1	2	0	50	2
2	12	12.5	175	0
3	17	50	225	4

The mass spectrometer transfer line and ion source temperatures were set to 200 °C with a mass scan range of 43 – 250 m/z.

Data Analysis - Peak area extraction was performed through TraceFinder 4.0 using an internally created processing method. Calculations of sample concentrations were performed manually based on the peak area counts of each fatty acid standard at 0-5000 ppm concentrations. A regression model was fit to all standards consisting of at least four consecutive standard levels for lower concentration curve and at least three standard levels for higher concentration curve. If the peak area of the analyte was below the peak area of the lowest standard used in the curve, the resulting calculated concentration is stated as 0.

Computational Modeling

gLV – The generalized Lotka-Volterra (gLV) model is a set of coupled ordinary differential equations that describe the growth of interacting species over time,

$$\frac{dx_i}{dt} = x_i\left(r_i + \sum_{j=1}^{n_s} a_{ij}x_j\right)$$

where x_i is the abundance of species i and n_s is the total number of species. Model parameters include the species growth rate, denoted as r_i, and coefficients that determine how species j affects the growth of species i, denoted as a_{ij}. Measurements of each species are assumed to be subject to the addition of zero-mean Gaussian random noise with a species-specific variance parameter. A Gaussian prior over the parameter distribution is set so that growth rates have a mean of .3, self-interaction terms have a mean of -1, and inter-species interaction terms have a mean of zero. The variance of the prior and the measurement variance of each species are model hyperparameters and can be estimated using the training data using the Expectation-Maximization (EM) algorithm (see Supplementary methods: gLOVE). For gLV, a 10-fold K-folds cross validation was utilized. Out of fold prediction performance (figure 3) is the performance for each of these folds. Best fit prediction is the model trained on the entire dataset predicting the entire dataset.

Linear Regression – Linear regression was used to predict concentrations of MCFAs produced during each 48hour 1mL batch culture as previously described[57]. Briefly, we used a regression model with terms that parameterize microbial production of target metabolites, both basal release of individual species and pairwise interaction terms. Because we did not collect time-resolved data, the model assumes a linear approximation of metabolite release. Interaction terms were

added for both species OD (representing biomass) and also categorical presence to capture any metabolite production not associated linearly with biomass (e.g. production of metabolites early in culture before the metabolite producer experiences a death phase). Model fitting was used using custom Python scripts using scikitlearn. We used ElasticNet to optimize two hyperparamters, scanning values for L1:L2 regularization ratio and α. Hyperparamter values were chosen using a leave-one-out K-folds cross validation approach. Out of fold prediction (figure 3) is the prediction performance for each of these folds, which generate unique individual parameter sets. In order to calculate best fit (figure 3), each parameter set was used to predict every point in the dataset, and these predictions were averaged. Each sample in the dataset was the average of n=1-4 biological replicates of each community, with additional replicates of some communities being added at each cycle (e.g. the full 16-member community, which was used as an internal benchmark to ensure consistency in day-to-day variation). In order to simplify parameter value interpretation (Figure 4), we simplified each parameter by combining the values for all parameters that contained either monospecies or species pairs. For example, species x_i and species x_j appeared in four parameterized terms as described: x_i presence • x_j presence, x_i OD • x_j OD, x_i presence • x_j OD, and x_i OD • x_j presence, to account for all combinations of potential mechanism. In order to interpret parameter values to extract meaningful biological information, all parameter values for these terms were summed, to understand their collective contribution. This was only completed for interpretation purposes.

MDS With Phylogeny

All communities in each media type were transformed into vectors where present =1 and absent =1. Community vectors that contain a species that is a descendant of a parent node in the

phylogenetic tree will populate the element of the vector corresponding to that node. Consequently, the alignment of two vectors corresponding to two different communities depends on the phylogenetic relationships between species. To create the functional landscape of the full set of experimentally characterized communities, we performed multidimensional scaling (MDS) on the set of vectors augmented with phylogenetic information using a Euclidean distance metric to project the data onto 2-dimensions using Scikit-Learn's MDS function.

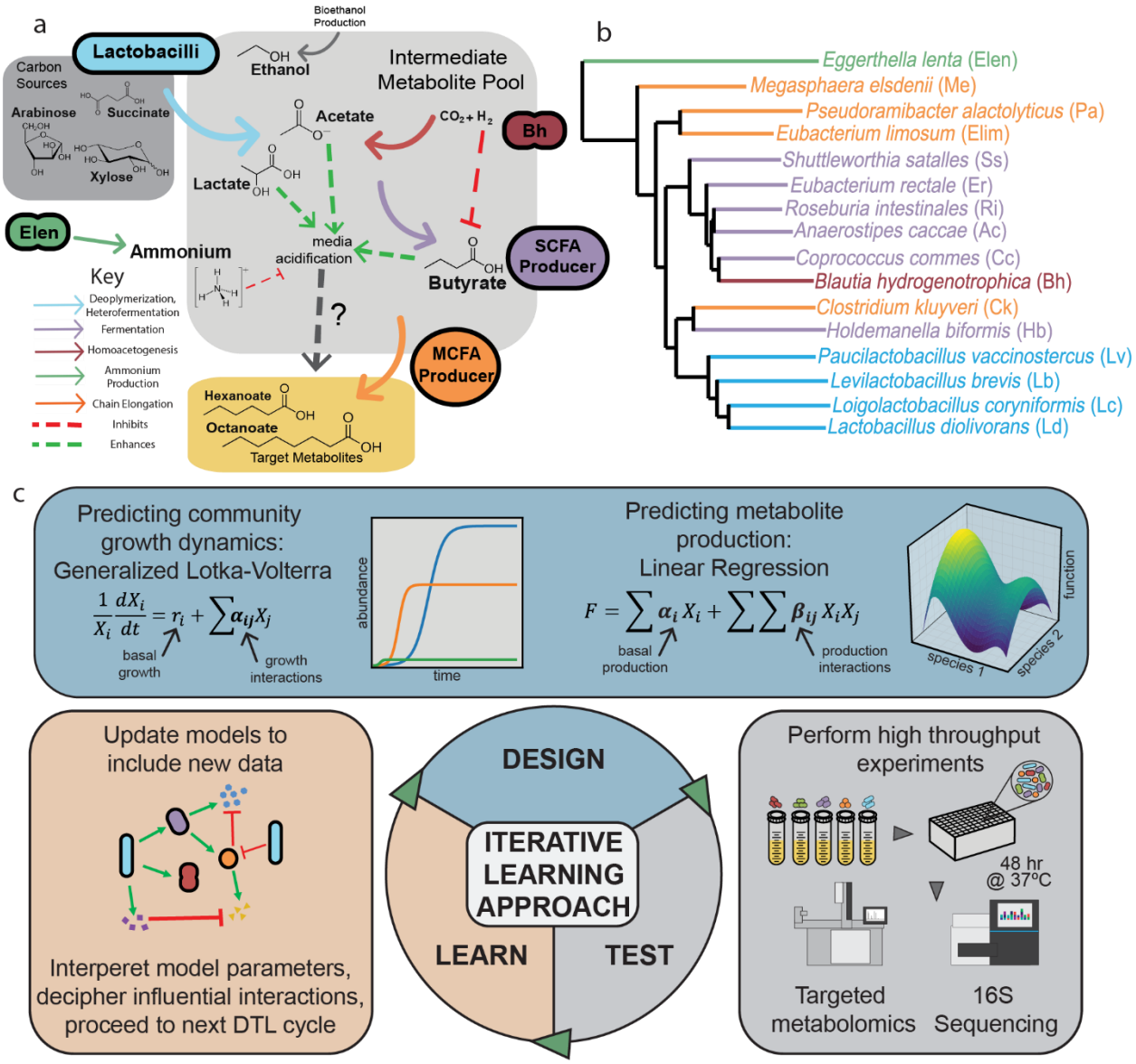

Figure 1: System summary and experimental approach.

A. Hypothesized system functionality. Stillage waste contains plant sugars, ethanol, and other carbon sources, which are transformed step-wise into MCFAs via fermentation and chain elongation by major functional groups (lactobacilli, SCFA producers, MCFA producers). Other included species modify the gas headspace (homoacetogenesis) and media buffering capacity

(ammonium production). Gas headspace and pH have context-dependent effects on organic acid production, and it is unknown how these factors will interact to impact overall MCFA production in communities. **B. Phylogenetic tree of synthetic community.** Tree constructed from 16S rRNA genes. Species are color-coded by functional group from panel A. **C. Iterative modeling approach for exploring community space.** Community behavior was predicted in two stages: (1) the Generalized Lotka-Volterra model was used to predict community composition at 48 hours and (2) Linear Regression was used to predict MCFA production (F, function) from 48-hour compositional data. Models are combined to predict growth and production of all possible combinations of organisms, and communities with highest predicted MCFA production can be targeted for testing. After predicting all possible combinations, high-throughput experimentation can empirically test target communities, with community composition and MCFA production data collected. After data collection, models are updated including the new data. Parameter values are interpreted to extract biological information and infer potential mechanisms. After models are updated, the next DTL cycle commences with predictions on all remaining untested combinations.

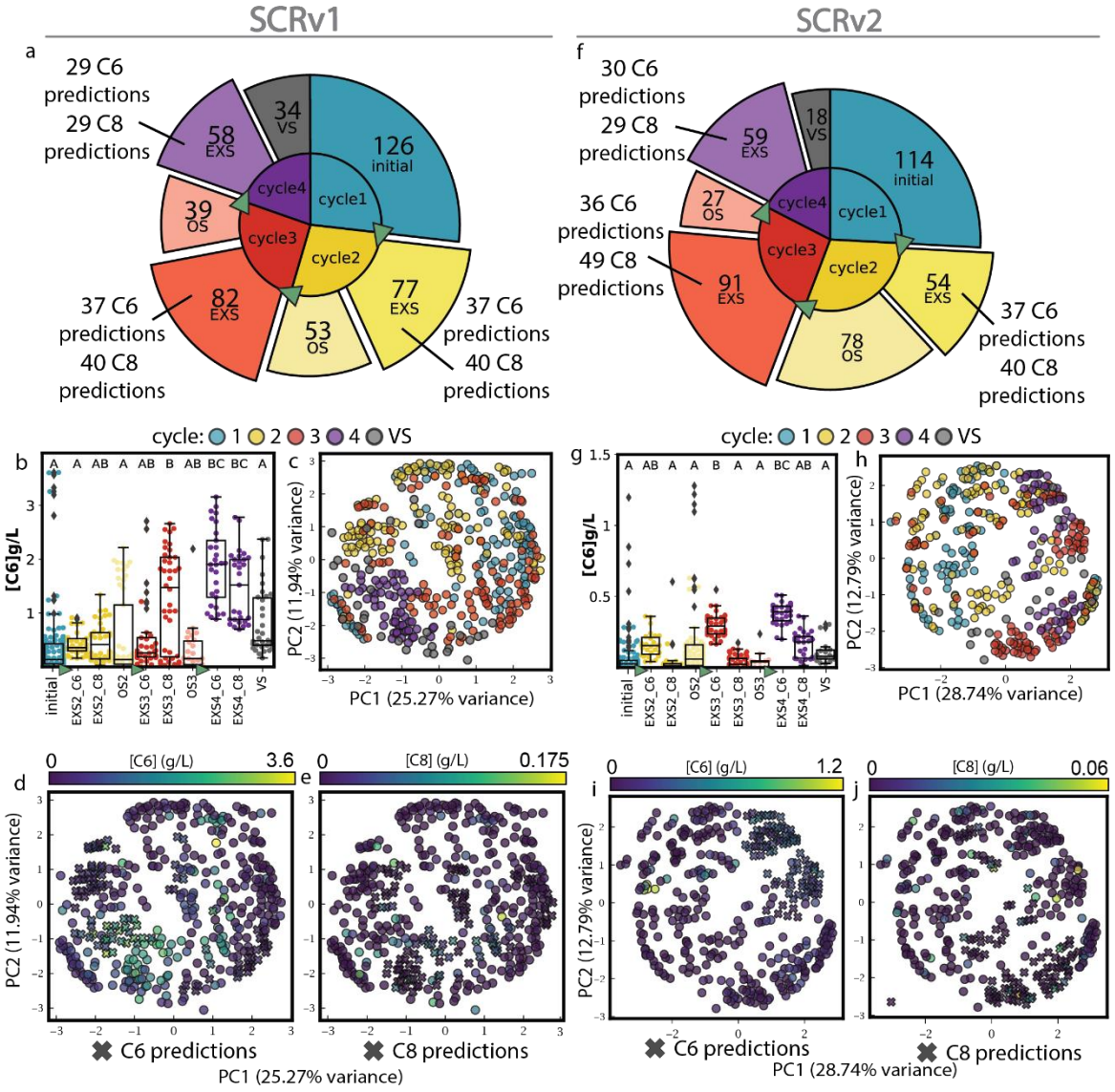

Figure 2: Community selection criteria and exploration of the sample space.

A/F: Community selection criteria. For both media, community selections for each design phase balanced Exploitation Selections (EXS - predictions for both C6 and C8 production) and Other Selections (OS – based on parameter values). A random selection of communities was also chosen as a Validation Set (VS). **B/G: Hexanoate production of communities over cycles**. X-axis shows community selections from panels A/F. Y-axis shows measured C6 (g/L). Mean C6

production was compared between groups using two-way ANOVA and Tukey's HSD yardstick ($p<0.001$) **C/H: MDS with phylogeny.** Composition and phylogeny of each community was projected onto 2D sample space using MDS. Each point represents one community, and the distance between points indicates similarity in composition including relatedness of constituent species. Colors correspond to cycle in which each community was collected. **D/E/I/J: MDS with phylogeny and function.** MDS plots from C/H were transformed to show functional output ([C6] or [C8] observed, color bars, top of panels) and whether communities were selected for testing as predictions for maximum C6 production (D/I) or C8 production (E/J), indicated by X shape.

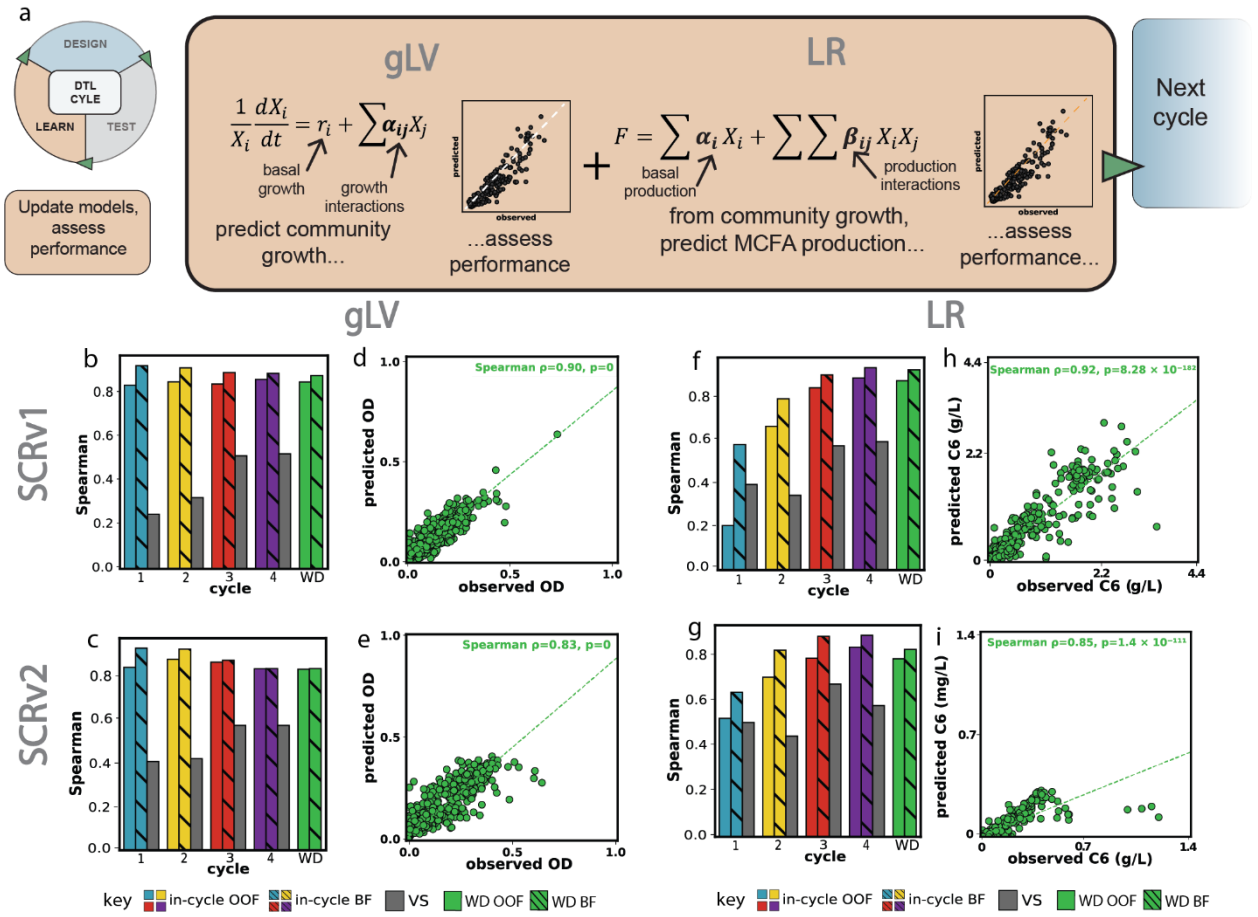

Figure 3: Assessment of model performance.

A. Two stage modeling predicts community growth then MCFA production. During Learn phases, models were trained on the collected data to estimate growth and production interactions. Both models were evaluated for performance. After model training, predictions on unseen communities were made to begin next cycle Design phase. **B/C: gLV model performance for each cycle out of fold, best fit, and validation set.** At each cycle (colors), out of fold (OOF, solid bars) and best fit (BF, hatched bars) performance was assessed. Parameter values estimated were also applied to the Validation Set (VS, grey bars) after this data was collected to assess performance. The Whole Dataset (WD) includes all in-cycle data and the VS (green bars), and was also fit with OOF and BF. **D/E: Whole dataset gLV best fit performance.** Performance can

be visualized by regression – x-axis shows the measured OD value for each species in each community, y-axis shows the model-predicted value. **F/G: LR model performance for each cycle out of fold, best fit, and validation set.** At each cycle (colors), out of fold (OOF, solid bars) and best fit (BF, hatched bars) performance was assessed. Parameter values estimated were also applied to the Validation Set (VS, grey bars) after this data was collected to assess performance. The Whole Dataset (WD) includes all in-cycle data and the VS (green bars), and was also fit with OOF and BF. **H/I: Whole dataset LR best fit performance.** Performance can be visualized by regression – x-axis shows the measured C6 value for each species in each community, y-axis shows the model-predicted value.

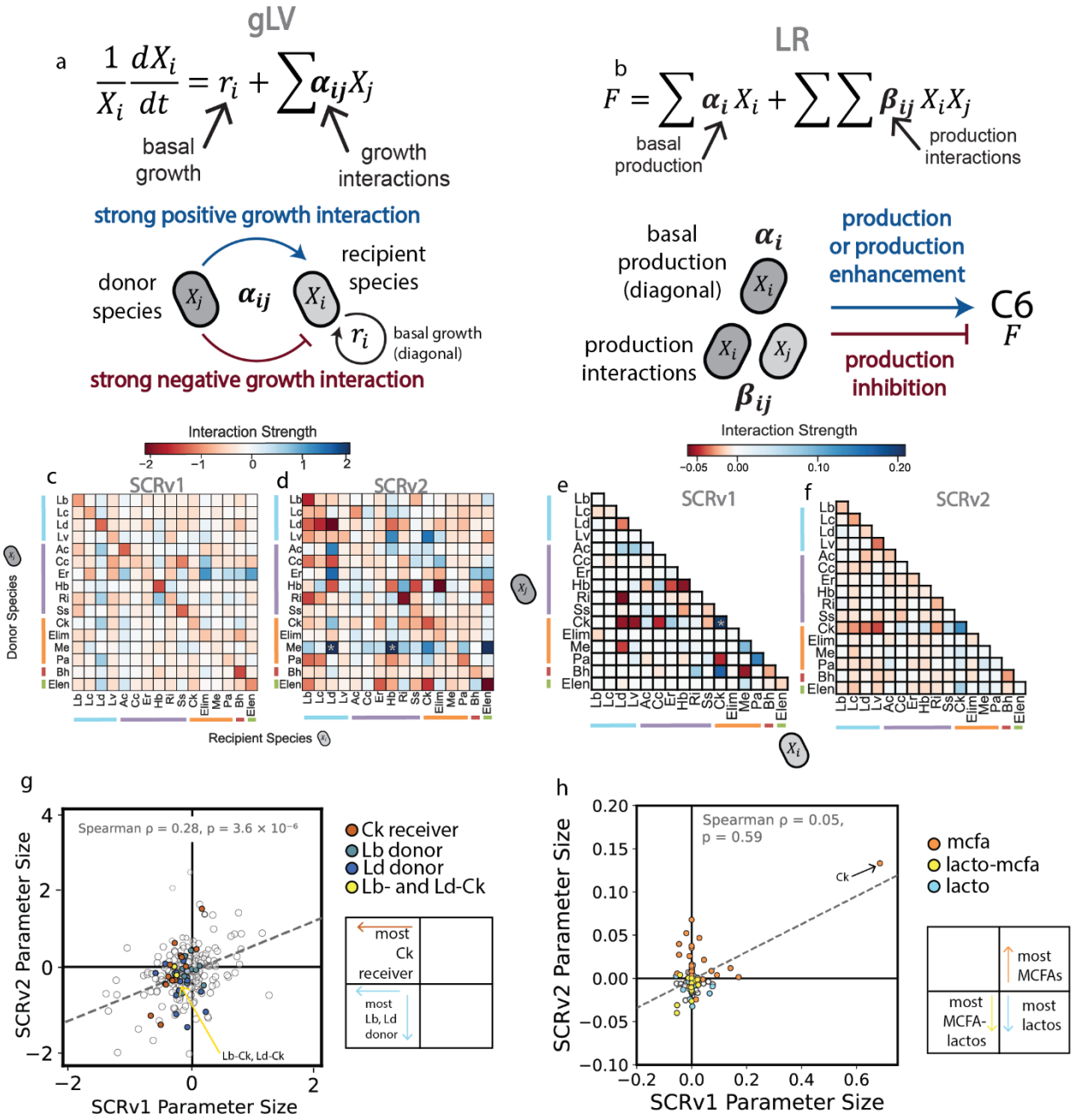

Figure 4: Interpretation and analysis of model parameters.

A. gLV parameter interpretation. gLV estimates community composition given each species (X_i) intrinsic growth rate (r_i) and growth interactions (α_{ij}). Interpretation of these parameters is directional, and influence of donor species (X_j) on X_i can be inferred. **B. LR parameter interpretation.** LR regression estimates the functional output a community given each species

(X_i) basal metabolite production (α_i) and production interactions (β_{ij}). Interpretation of these parameters is not directional, and only assess the impact of species pairs when they appear in a community together. **C/D: gLV parameter values for SCRv1 (C) and SCRv2 (D)**. Values are given by colorbar (top). Species are organized by functional groups (lactobacilli = blue, SCFA producers = purple, MCFA producers = orange, Bh = red, Elen = green). Two values are outliers (SCRv2 Me-Ld=3.6, Me-Ld=2.39, *). Values should be read as donor species (y-axis) directional impact on recipient species (x-axis). Diagonal represents r_i and is limited by the carrying capacity of the media for that species. **E/F: LR parameter values for SCRv1 (E) and SCRv2 (F).** Values are given by colorbar (top). Species are organized by functional groups (lactobacilli = blue, SCFA producers = purple, MCFA producers = orange, Bh = red, Elenta = green). One value is an outlier (SCRv1 Ck * = 0.6). Values should be read across rows and then down columns. Diagonal represents the monospecies contribution. **G. Regression of gLV parameters between SCRv1 (x-axis) and SCRv2 (y-axis).** Each point represents a directional parameter. Dotted line is linear regression, Spearman rho and p value indicated. Parameters where Ck is the receiver colored orange, parameters where Lb or Ld is the donor are colored blue, Lb-Ck and Ld-Ck colored yellow. Inset indicates pattern. **H. Regression of LR parameters between SCRv1 (x-axis) and SCRv2 (y-axis).** Each point represents a non-directional parameter. Dotted line is linear regression, Spearman rho and p value indicated. Parameters which contain at least one MCFA producer colored orange, parameters which contain at least one lactobacilli colored blue, parameters which contain one lactobacilli and one MCFA producer colored yellow. Inset indicates pattern.

Supplementary Information I: Methods and Tables

Strains, preculture media, serial preculturing scheme

General protocol – Single use glycerol stocks (SUGS) were prepared as described in materials and methods. In order to grow high enough volume for experiments and synchronize the growth phase of cultures, a serial preculture strategy was utilized. For all species but Ck and Hb, one SUGS was added to anaerobic media as described for first preculture, then 400uL of this preculture was transferred to a fresh tube after the above indicated time for the experimental culture. For Ck and Hb, two SUGS were added to anaerobic media and allowed to grow for four days. Then, 1mL of this culture was added to 5mL of fresh media and allowed to grow for 48 hours. Then, the entire 5mL volume was added to 25 or 30mL fresh media and allowed to grow for 48 hours. See Table 1 below.

Table 1: Strains and preculture conditions.

Strains were acquired from DSMZ. Recipe for DM35 given in Supplement spreadsheet 1: media recipes. ABB = HiMedia ABB cat# M1636. MRS = MRS Lactobacillus broth, Millipore cat# 69966

Species/Strain	**Media/ Reference or cat#**	**Preculture setup**	**Time from SUGS (hrs)**	**Time from Preculture (hrs)**
Anaerostipes caccae DSMZ14662	DM35 + yeast	1 preculture (5mL), 400uL to final culture	24	16

Blautia hydrogenotrophica DSMZ10507	DM35 + yeast	1 preculture (5mL) , 400uL to final culture	24	16
Coprococcus commes ATCC27758	DM35 + yeast	1 preculture (5mL) , 400uL to final culture	24	16
Clostridium kluyveri DSMZ555	DM35 + yeast	2 precultures: 2 SUGS in 5mL, then 1mL transferred to 5mL; full 5mL to 25mL final culture	4 days, then 48 hours	48
Eggerthella lenta DSMZ2243	DM35 + yeast	1 preculture (5mL), 400uL to final culture	48	48
Eubacterium limosum ATCC8486	MRS	1 preculture (5mL), 400uL to final culture	24	24
Eubacterium rectale ATCC33656	DM35 + yeast	1 preculture (5mL), 400uL to final culture	24	24
Holdemanella biformis DSMZ12042	ABB	2 precultures: 2 SUGS in 5mL, then 1mL transferred to 5mL; full	4 days, then 48 hours	48

		5mL to 25mL final culture		
Levilactobacillus brevis ATCC367	MRS	1 preculture (5mL), 400uL to final culture	24	12
Loigolactobacillus coryniforms DSMZ20001	MRS	1 preculture (5mL), 400uL to final culture	24	12
Lentilactobacillus diolvorans DSMZ14421	MRS	1 preculture (5mL), 400uL to final culture	24	12
Paucilactobacillus vaccinostercus* DSMZ20634	MRS	1 preculture (5mL), 400uL to final culture	24	12
Megasphaera elsdenii DSMZ20460	DM35 + yeast	1 preculture (5mL), 400uL to final culture	24	16
Pseudoramibacter alactolyticus DSMZ2980	DM35 + yeast	1 preculture (5mL), 400uL to final culture	48	48
Roseburia intestinalis DSMZ14610	DM35 + yeast	1 preculture (5mL), 400uL to final culture	24	16

Shuttleworthia satalles DSMZ14600	ABB	1 preculture (5mL), 400uL to final culture	48	48

* Note on Lactobacilli naming scheme: two letter code ‘Lv’ refers to former name (*Lactobacillus vaccinostercus*), prior to major reclassification of lactobacilli (Zheng et. al. 2020) to maintain consistent coding throughout experimental materials.

Community Selection Information

Definitions

Parameter of Interest: At each cycle, simplified LR parameters were calculated as described in materials and methods. These were then ranked highest to lowest and parameters which were of the highest value were prioritized when selecting communities.

Leave One Out – In the context of the full 16-member community, leave one outs are all 15-member sub communities.

QC – Quality Control for sequencing as described in materials and methods.

Internal Benchmark – Full community tested every plate, every experiment to assure consistency between plates and days.

Random Selections - In order to validate the performance of the model, we selected 30 random untested communities. We reasoned that as the dataset accumulates community selections that are specifically designed to maximize MCFA production, communities of this kind will likely dominate the training set, and this similarity between training and test sets could bias model performance. In contrast, a random selection of communities would include communities less likely to resemble the training data in composition, and would mitigate this bias. Additionally, a random control group is necessary to determine whether the model-guided community selection is more effective at identifying high producing communities than choosing communities from the sample space at random.

See Table 2 below for implementation.

Table 2: Community Selection Criteria

This table describes criteria for selecting communities at each cycle. Broadly, these categories are the Exploitation Selections (communities selected because the model predicted that they would have the highest MCFA production (either C6 or C8)) or Other Selections (all other selection criteria). Number of communities to be tested was limited by experimental throughput.

Media	Cycle	Community choice descriptions
SCRv1	1	All monospecies, leave-one-out's (15 member communities), low complexity communities (2-5 members) previously observed in pilot experiments, full community for internal benchmark
	2	1. Any communities targeted in cycle 1 that did not pass QC 2. Repeat high C6 and C8 observations (top 10 each) 3. Parameters of interest: collect pair data for species contained in high-value parameters - Ck pairs (C6), Lb pairs (C8), Ld pairs (outliers contain Ld), Pa pairs (C8) 4. Model predictions: - C6: top 30 predictions, then next 14 predictions that contain Er (parameter of interest) - C8: top 30 predictions (if not overlapping with C6), then next 15 predictions that contain Bh-Lv (parameter of interest) 5. Full community as internal benchmark
	3	1. Any communities targeted in cycle 2 that did not pass QC 2. Repeat high C6 and C8 observations (top 10 each)

		3. Parameters of interest: collect pair data for species contained in high-value parameters - Bh pairs (C8), Er pairs (C6) 4. Model predictions: - C6: top 20 predictions, next 10 that had Ck-Me (parameter of interest) - C8: top 20 predictions, next 15 that had Ck-Ri (parameter of interest) 5. Full community as internal benchmark
	4	1. Model predictions: - C6: top 30 predictions - C8: top 30 predictions 2. Random selections
SCRv2	1	Same communities as SCRv1 cycle 1
	2	1. Any communities targeted in cycle 1 that did not pass QC 2. Repeat high C6 and C8 observations (top 10 each) 3. Parameters of interest: collect pair data for species contained in high-value parameters - Me pairs (C6), Ac pairs (C8), Er pairs (C6) 4. Model predictions: - C6: top 27 predictions - C8: top 27 predictions (if not overlapping with C6) 5. Full community as internal benchmark
	3	1. Any communities targeted in cycle 2 that did not pass QC 2. Repeat high C6 and C8 observations (top 10 each) 3. Parameters of interest: collect pair data for species contained in high-value parameters – Bh pairs (C8)

		4. Model predictions: - C6: top 19 predictions, then next 20 that had Ac-Pa or Ck-Ld - C8: top 19 predictions (if not overlapping with C6), and next 49 that had Pa-Elen, Er-Bh, Ck-Elen, or Cc-Er 5. Full community as internal benchmark
	4	1. Model predictions: - C6: top 30 predictions - C8: top 30 predictions 2. Random selections

Supplementary Information II: Figures

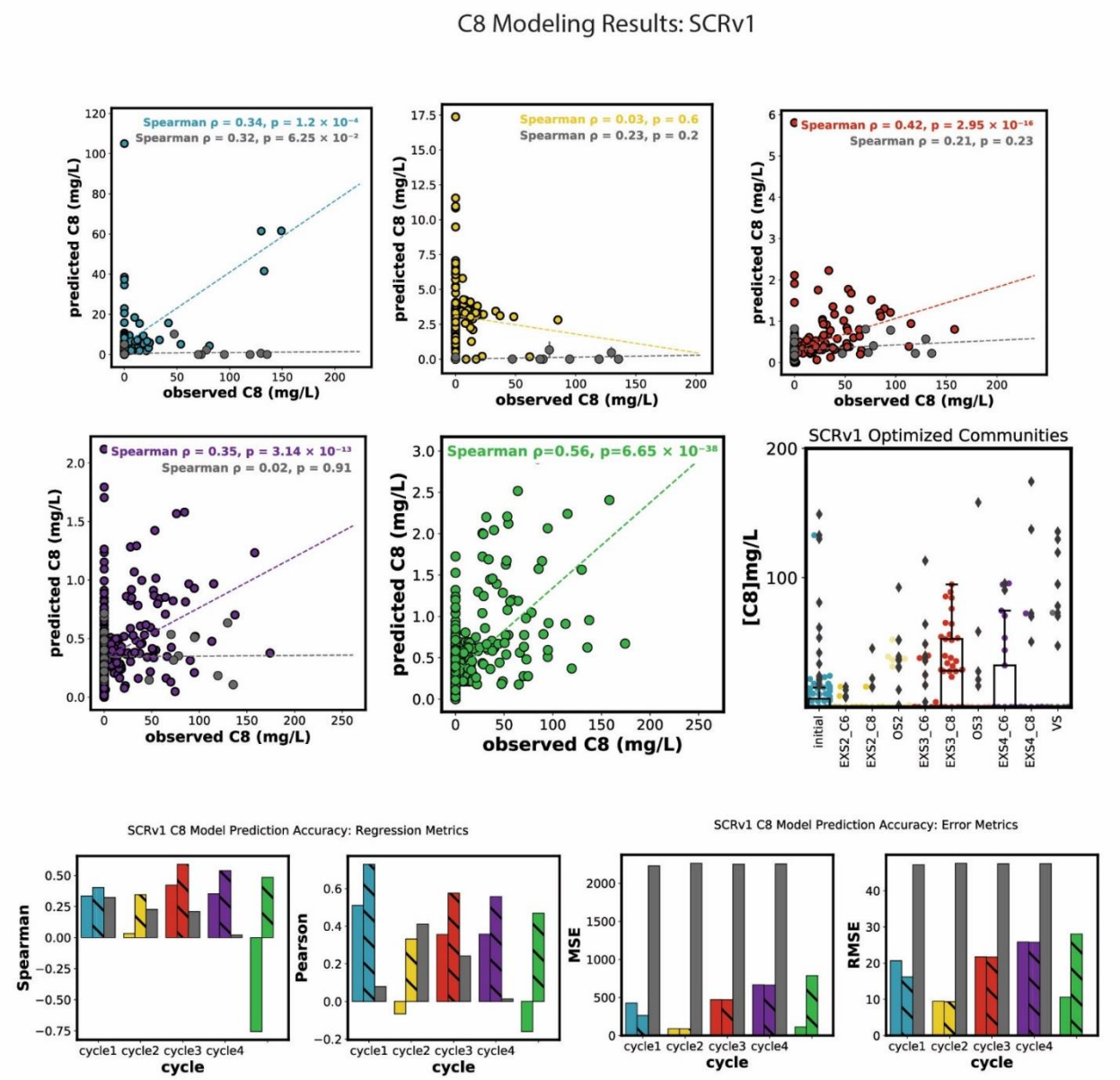

Figure S1: SCRv1 C8 modeling results.

Scatterplots show in cycle out of fold regression for observed (x-axis) and predicted (y-axis) C8 measurements. Box plot shows measured C8 (mg/L, y axis) for each community selection group.

Bar plots show performance (Spearman rho, Pearson R) and error metrics (MSE, RMSE) for each cycle and validation set.

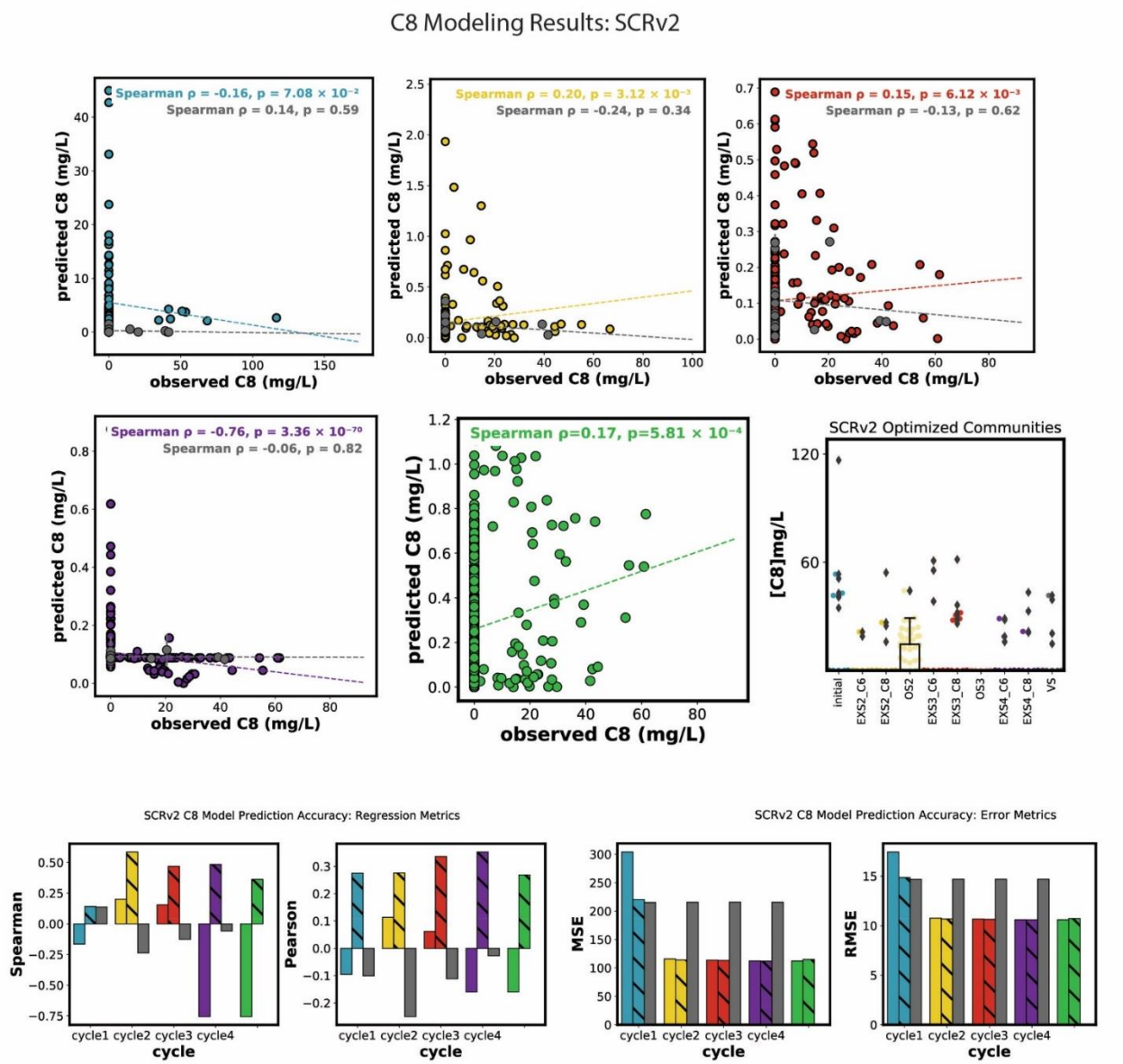

Figure S2: SCRv2 C8 modeling results.

Scatterplots show in cycle out of fold regression for observed (x-axis) and predicted (y-axis) C8 measurements. Box plot shows measured C8 (mg/L, y axis) for each community selection group. Bar plots show performance (Spearman rho, Pearson R) and error metrics (MSE, RMSE) for each cycle and validation set.

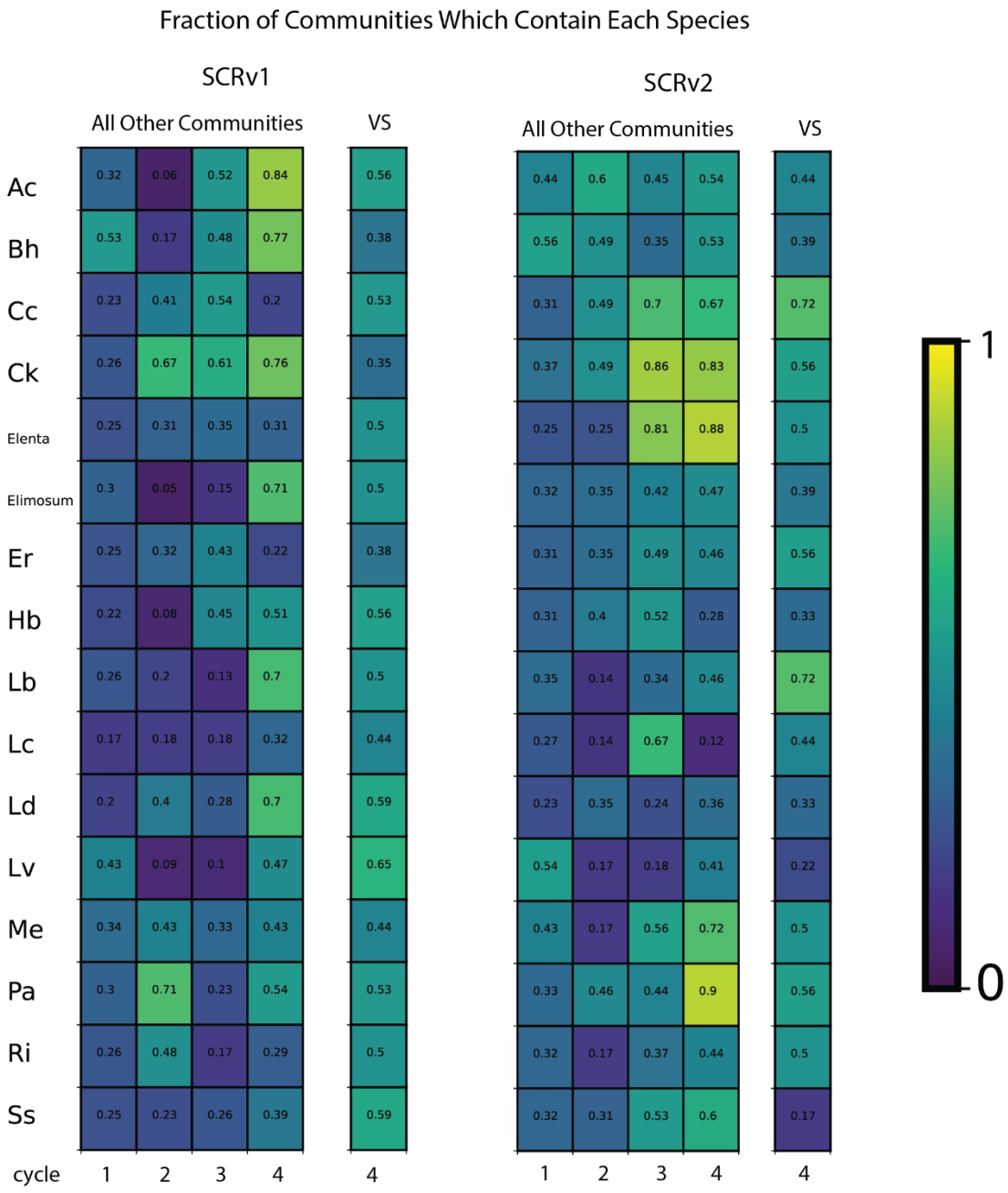

Figure S3: Fraction of communities which contain each species, by cycle.

In order to assess whether the dataset was accumulating communities of similar composition which might bias model performance, we constructed a heatmap which shows the fraction of communities (0-1.0, colorbar) which contain each species (y-axis) at each cycle (x-axis) for each media type (SCRv1, left, and SCRv2, right). The single column (right) represents the Validation

Set, a random selection of communities. All other selection types (initial dataset, parameters of interest, and exploitation selections) are grouped together (left).

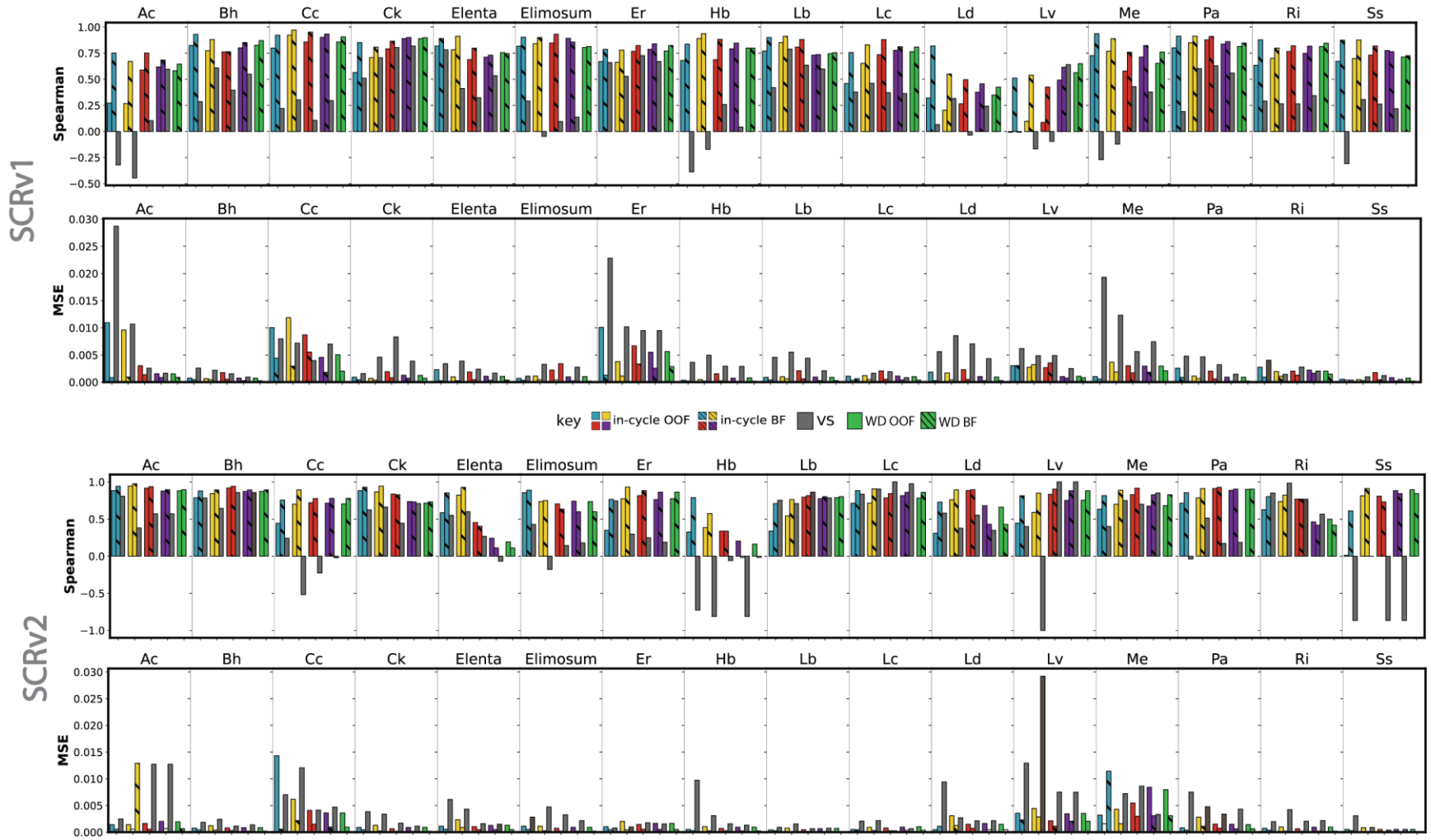

Figure S4: gLV prediction performance and MSE for individual species over cycles

Bar plots show Spearman rho and MSE (y-axis) for each species (x-axis) as predicted by gLV for SCRv1 (top) and SCRv2 (bottom). Since gLV makes a prediction for every species in each community, the accuracy of predictions for each species can be assessed. Predictions for in cycle OOF, BF, VS, and WD shown.

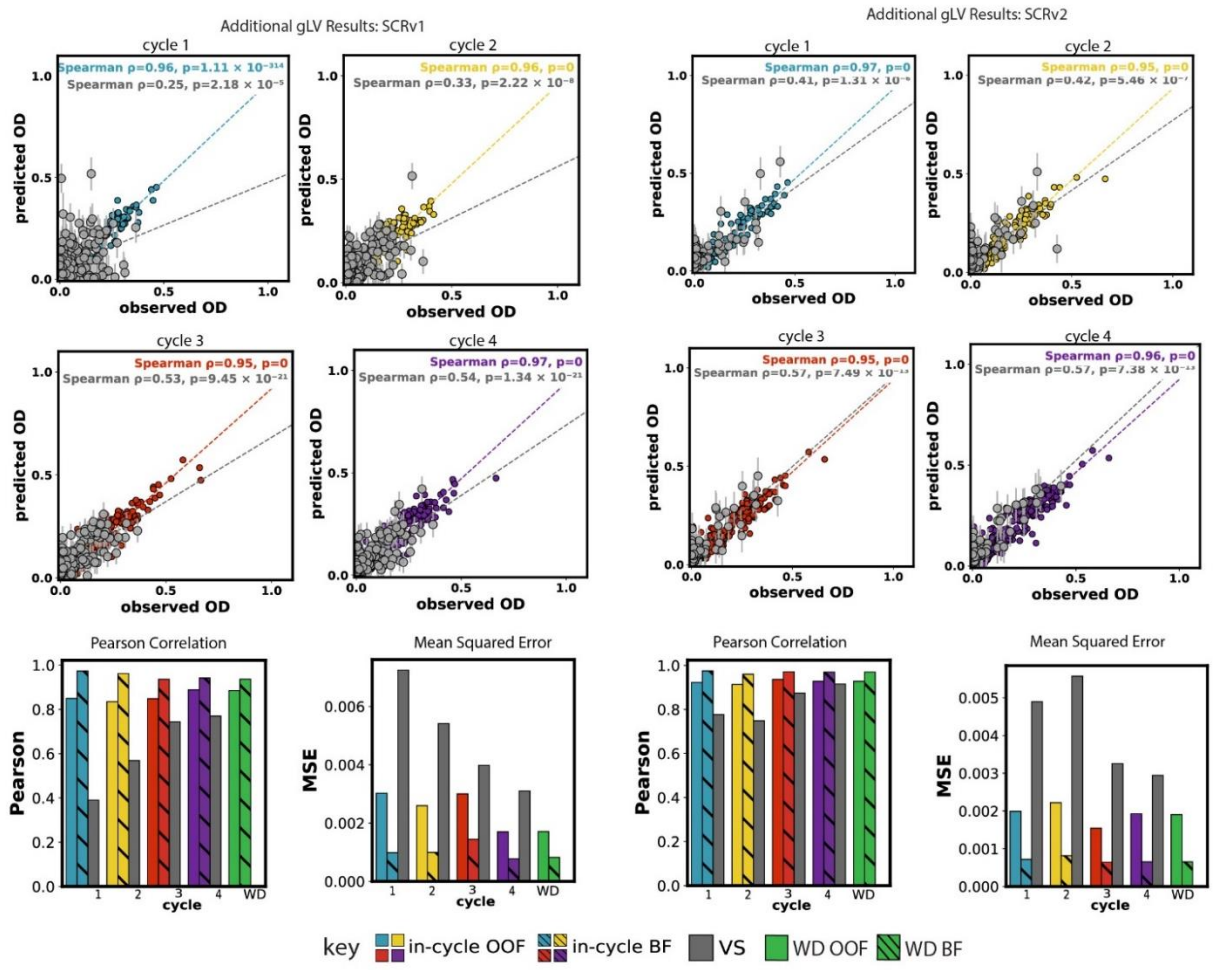

Figure S5: Additional gLV results.

Additional modeling performance results for SCRv1 (left) and SCRv2 (right). Scatter plots show correlation between observed OD for each species (x-axis) and model predicted value (y-axis) for each species for each cycle. Bar plots show Pearson R correlation for in cycle OOF, BF, VS and WD OOF and BF.

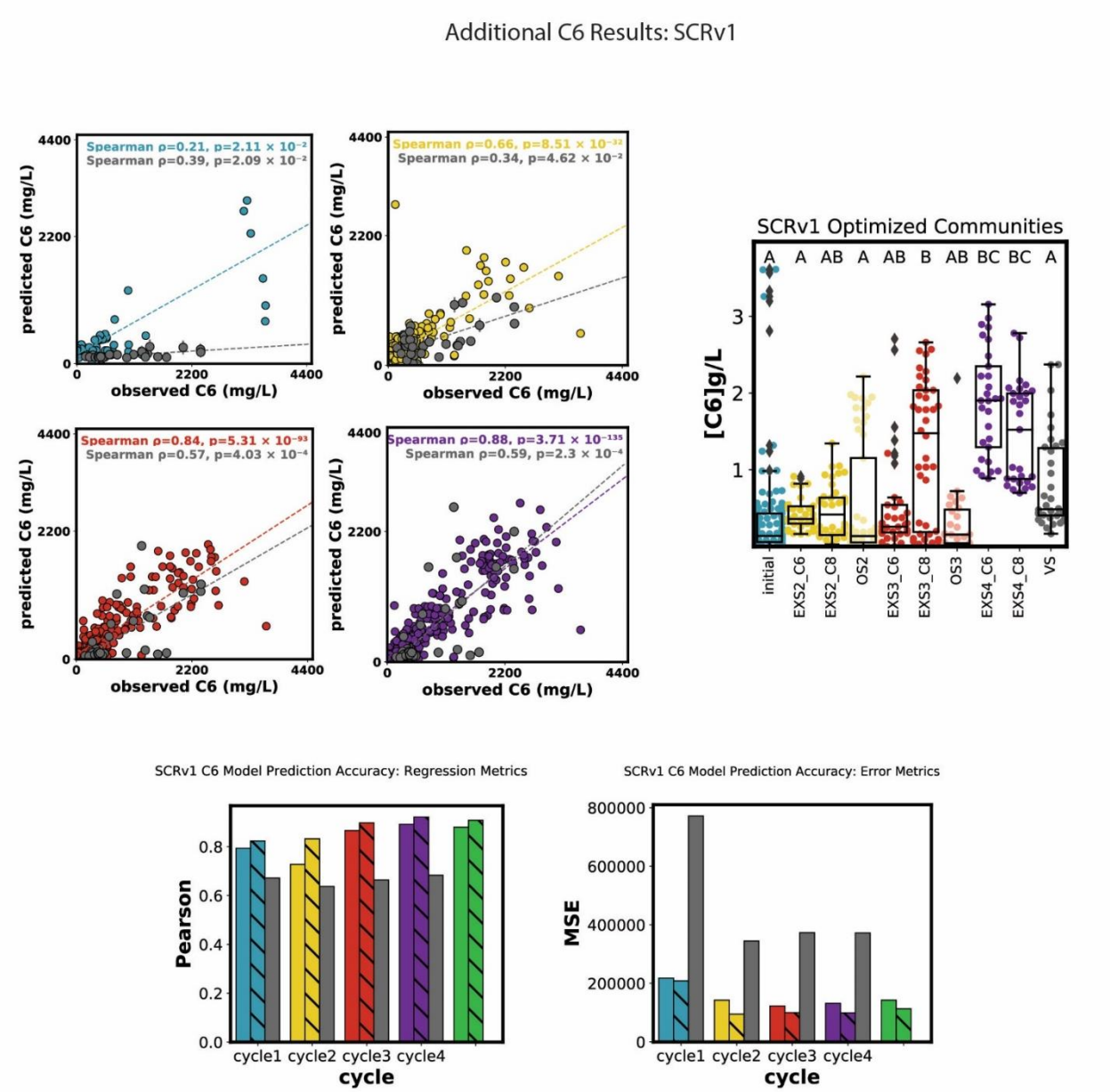

Figure S7: Additional C6 modeling results (SCRv1).

Scatter plots show correlations between observed C6 values (x-axis) and model-predicted values (y-axis) over cycles. Box plot shows C6 production values for each selection type. Bar plots show Pearson R and MSE for in-cycle OOF, BF, VS, and WD OOF and BF.

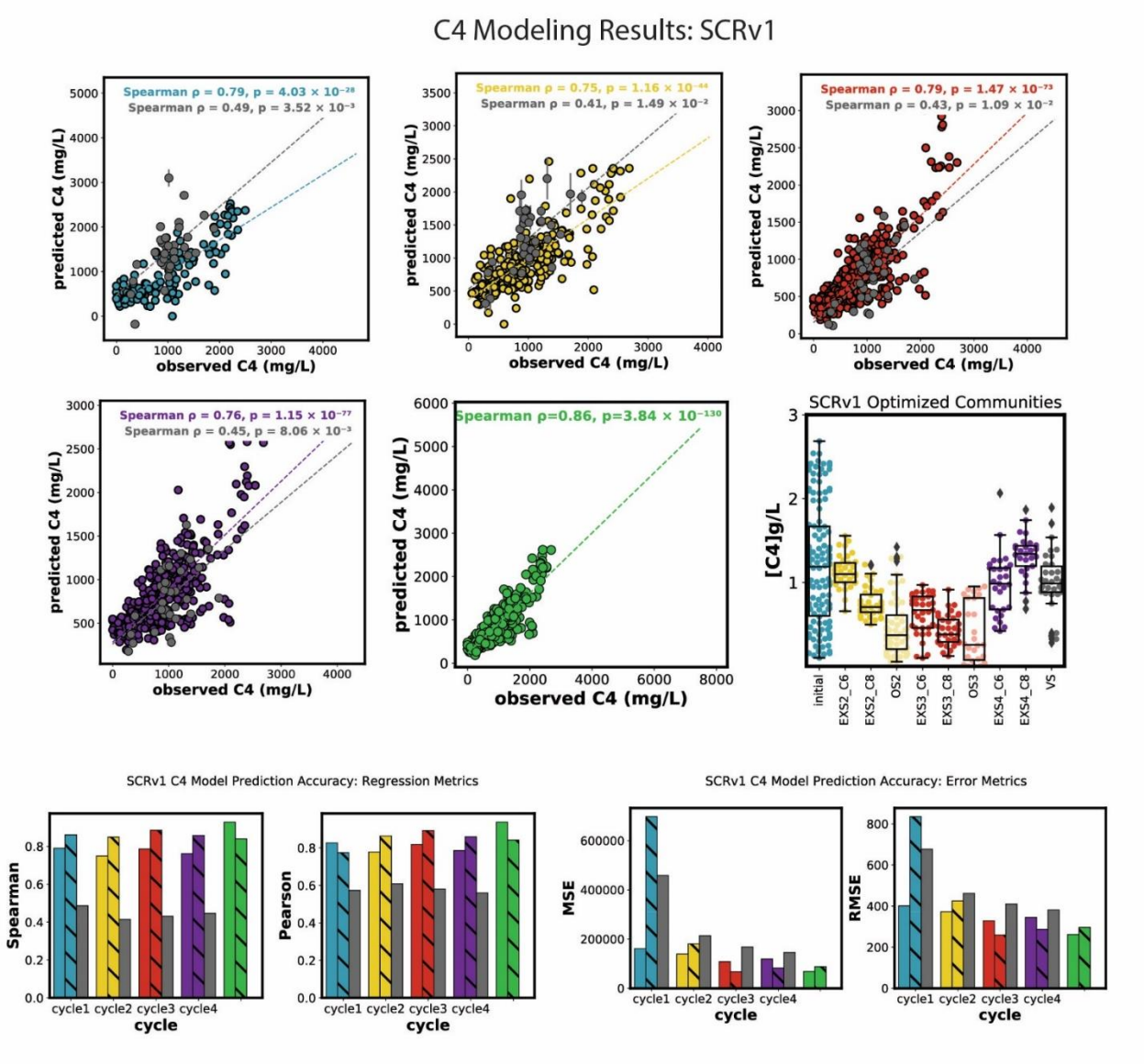

Figure S8: C4 modeling results (SCRv1).

Scatterplots show correlations between observed C4 values (x-axis) and model-predicted values (y-axis) over cycles. Box plot shows C4 production values for each selection type. Bar plots show Spearman rho, Pearson R, MSE and RMSE for in-cycle OOF, BF, VS, and WD OOF and BF.

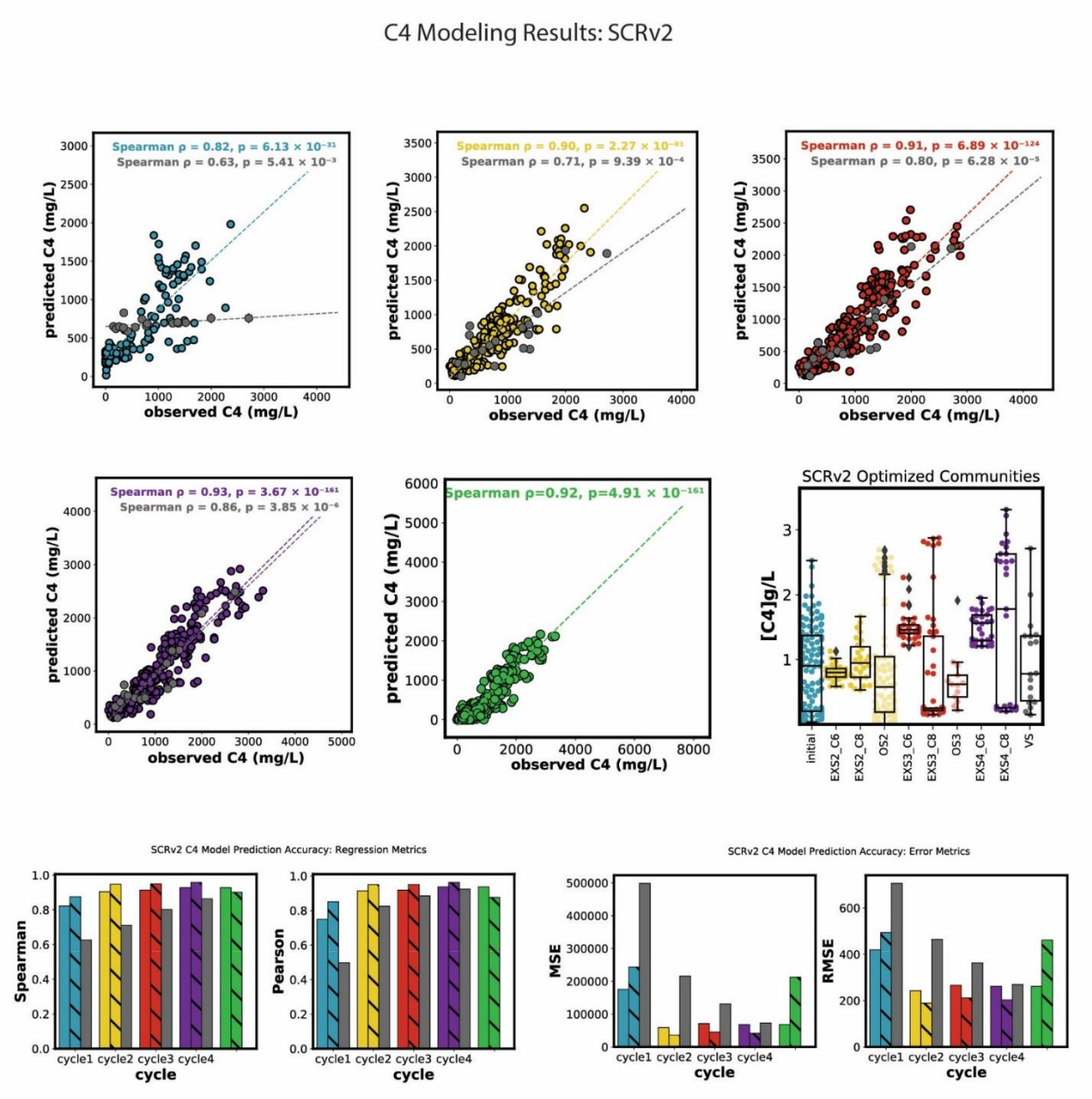

Figure S9: C4 modeling results (SCRv2).

Scatterplots show correlations between observed C4 values (x-axis) and model-predicted values (y-axis) over cycles. Box plot shows C4 production values for each selection type. Bar plots show Spearman rho, Pearson R, MSE and RMSE for in-cycle OOF, BF, VS, and WD OOF and BF.

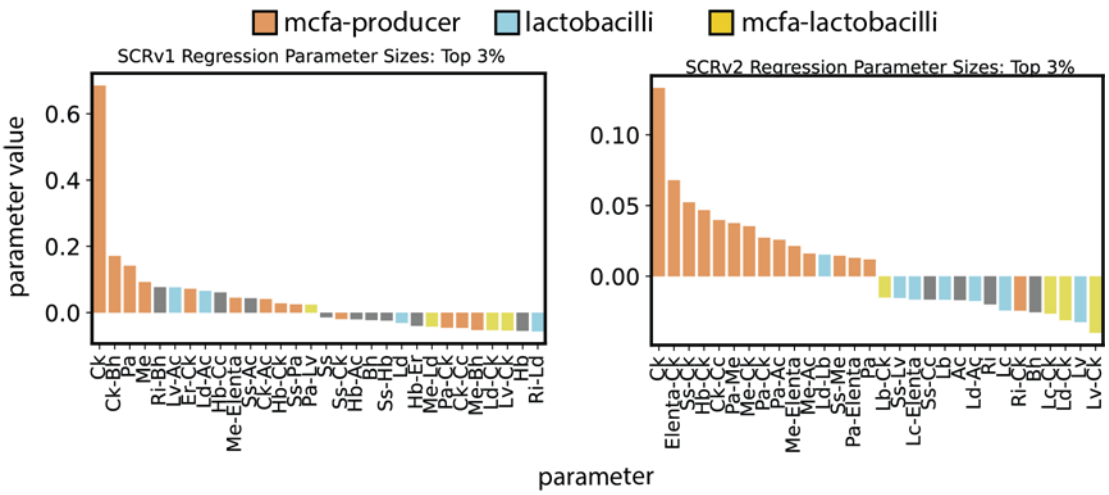

Figure S10: Top and bottom 3% of LR parameter spaces.

LR simplified parameter sizes (y-axis) for SCRv1 (left) and SCRv2 (right). Bars are colored to highlight parameters containing MCFA producers (orange), lactobacilli (blue), and both (yellow).

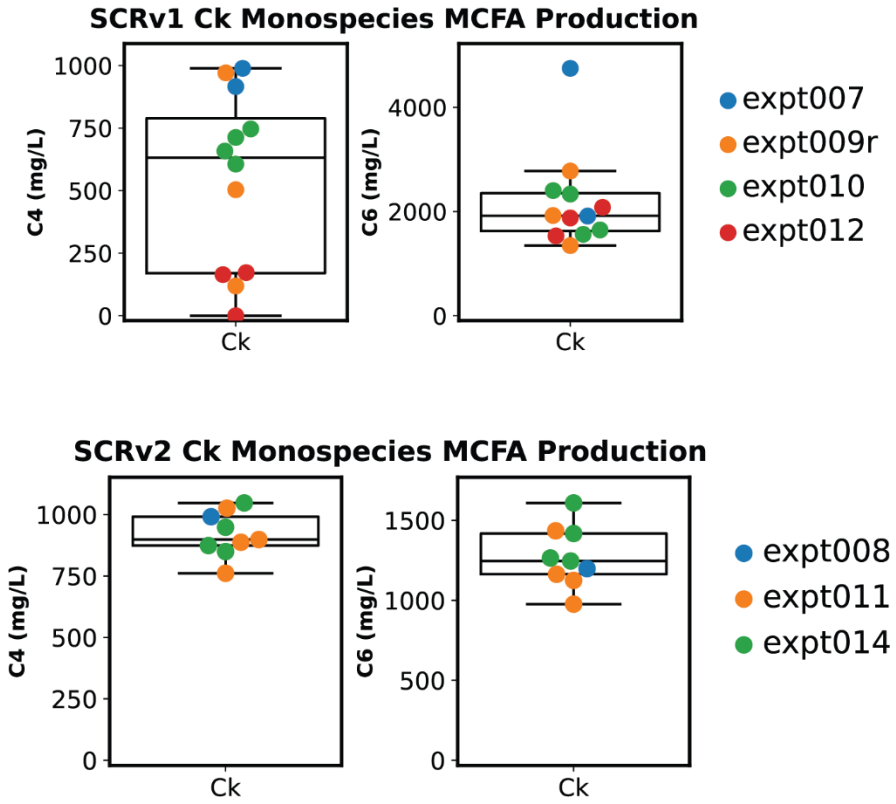

Figure S11: Ck monospecies C4 and C6 production.

Top = SCRv1, Bottom = SCRv2, C4 = left column, C6 = right column. Each point represents one biological replicate collected during one experimental day (color, legend).

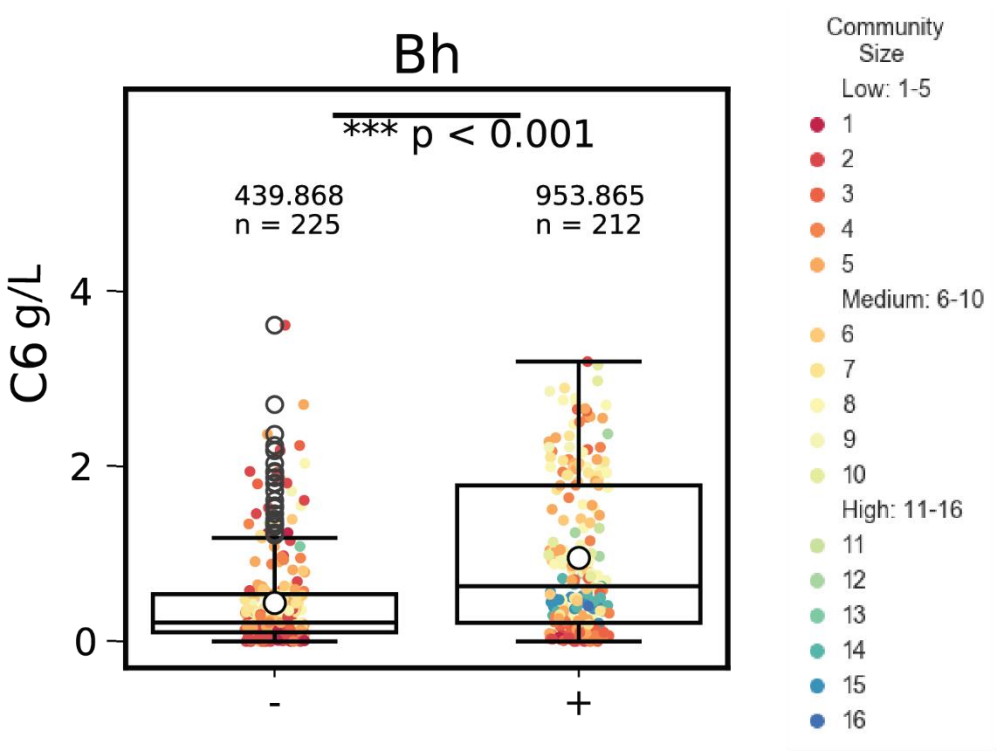

Figure S12: SCRv1 presence and absence of Bh in tested communities and C6 production.

Each point represents one community, colored by community size. Production of C6 (y-axis) of communities where Bh was present (right, +) were compared to communities where Bh was absent (left, -) by Mann-Whitney U. Mean and number of replicates indicated.

References

1. Rillig, M. C. *et al.* The role of multiple global change factors in driving soil functions and microbial biodiversity. *Science* **366**, 886–890 (2019).

2. Kuypers, M. M. M., Marchant, H. K. & Kartal, B. The microbial nitrogen-cycling network. *Nat. Rev. Microbiol.* **16**, 263–276 (2018).

3. Abatenh, E., Gizaw, B., Tsegaye, Z. & Tefera, G. Microbial Function on Climate Change - A Review. *Environ. Pollut. Clim. Change* **02**, (2018).

4. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, (2015).

5. Burmølle, M. *et al.* Enhanced Biofilm Formation and Increased Resistance to Antimicrobial Agents and Bacterial Invasion Are Caused by Synergistic Interactions in Multispecies Biofilms. *Appl. Environ. Microbiol.* **72**, 3916–3923 (2006).

6. Dejonghe, W. *et al.* Synergistic Degradation of Linuron by a Bacterial Consortium and Isolation of a Single Linuron-Degrading *Variovorax* Strain. *Appl. Environ. Microbiol.* **69**, 1532–1541 (2003).

7. Jacoby, R., Peukert, M., Succurro, A., Koprivova, A. & Kopriva, S. The Role of Soil Microorganisms in Plant Mineral Nutrition—Current Knowledge and Future Directions. *Front. Plant Sci.* **8**, (2017).

8. Compant, S., Samad, A., Faist, H. & Sessitsch, A. A review on the plant microbiome: Ecology, functions, and emerging trends in microbial application. *J. Adv. Res.* **19**, 29–37 (2019).

9. Pankievicz, V. C. S., Irving, T. B., Maia, L. G. S. & Ané, J.-M. Are we there yet? The long walk towards the development of efficient symbiotic associations between nitrogen-fixing bacteria and non-leguminous crops. *BMC Biol.* **17**, 99 (2019).

10. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).

11. Yissachar, N. *et al.* An Intestinal Organ Culture System Uncovers a Role for the Nervous System in Microbe-Immune Crosstalk. *Cell* **168**, 1135-1148.e12 (2017).

12. Briones, A. & Raskin, L. Diversity and dynamics of microbial communities in engineered environments and their implications for process stability. *Curr. Opin. Biotechnol.* **14**, 270–276 (2003).

13. Werner, J. J. *et al.* Bacterial community structures are unique and resilient in full-scale bioenergy systems. *Proc. Natl. Acad. Sci.* **108**, 4158–4163 (2011).

14. Venturelli, O. S. *et al.* Deciphering microbial interactions in synthetic human gut microbiome communities. *Mol. Syst. Biol.* **14**, (2018).

15. Huang, C. B., Alimova, Y., Myers, T. M. & Ebersole, J. L. Short- and medium-chain fatty acids exhibit antimicrobial activity for oral microorganisms. *Arch. Oral Biol.* **56**, 650–654 (2011).

16. Desbois, A. P. & Smith, V. J. Antibacterial free fatty acids: activities, mechanisms of action and biotechnological potential. *Appl. Microbiol. Biotechnol.* **85**, 1629–1642 (2010).

17. Seo, J.-H., Lee, S.-M., Lee, J. & Park, J.-B. Adding value to plant oils and fatty acids: Biological transformation of fatty acids into ω-hydroxycarboxylic, α,ω-dicarboxylic, and ω-aminocarboxylic acids. *J. Biotechnol.* **216**, 158–166 (2015).

18. Li, G. *et al.* Advances in microbial production of medium-chain dicarboxylic acids for nylon materials. *React. Chem. Eng.* **5**, 221–238 (2020).

19. R. Beller, H., Soon Lee, T. & Katz, L. Natural products as biofuels and bio-based chemicals: fatty acids and isoprenoids. *Nat. Prod. Rep.* **32**, 1508–1526 (2015).

20. Wagner, H., Luther, R. & Mang, T. Lubricant base fluids based on renewable raw materials: Their catalytic manufacture and modification. *Appl. Catal. Gen.* **221**, 429–442 (2001).

21. Hou, C. T. *Handbook of Industrial Biocatalysis*. (CRC Press, 2005).

22. De Groof, V., Coma, M., Arnot, T., Leak, D. J. & Lanham, A. B. Medium Chain Carboxylic Acids from Complex Organic Feedstocks by Mixed Culture Fermentation. *Molecules* **24**, 398 (2019).

23. Kallscheuer, N., Polen, T., Bott, M. & Marienhagen, J. Reversal of β-oxidative pathways for the microbial production of chemicals and polymer building blocks. *Metab. Eng.* **42**, 33–42 (2017).

24. Scarborough, M. J., Hamilton, J. J., Erb, E. A., Donohue, T. J. & Noguera, D. R. Diagnosing and Predicting Mixed-Culture Fermentations with Unicellular and Guild-Based Metabolic Models. *mSystems* **5**, e00755-20.

25. San-Valero, P., Abubackar, H. N., Veiga, M. C. & Kennes, C. Effect of pH, yeast extract and inorganic carbon on chain elongation for hexanoic acid production. *Bioresour. Technol.* **300**, 122659 (2020).

26. Zhang, C. *et al.* Efficient caproate production from ethanol and acetate in open culture system through reinforcement of chain elongation process. *J. Clean. Prod.* **383**, 135394 (2023).

27. Flaiz, M. *et al.* Refining and illuminating acetogenic *Eubacterium* strains for reclassification and metabolic engineering. *Microb. Cell Factories* **23**, 24 (2024).

28. Holdeman, L. V., Cato, E. P. & Moore, W. E. C. Amended description of *Ramibacterium alactolyticum* Prévot and Taffanel with proposal of a neotype strain1. *Int. J. Syst. Evol. Microbiol.* **17**, 323–341 (1967).

29. Steen, E. J. *et al.* Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* **463**, 559–562 (2010).

30. Wu, J., Zhang, X., Xia, X. & Dong, M. A systematic optimization of medium chain fatty acid biosynthesis via the reverse beta-oxidation cycle in *Escherichia coli*. *Metab. Eng.* **41**, 115–124 (2017).

31. Gazzola, G. *et al.* Biorefining food waste through the anaerobic conversion of endogenous lactate into caproate: A fragile balance between microbial substrate utilization and product inhibition. *Waste Manag.* **150**, 328–338 (2022).

32. Han, W., He, P., Shao, L. & Lü, F. Metabolic Interactions of a Chain Elongation Microbiome. *Appl. Environ. Microbiol.* **84**, e01614-18 (2018).

33. Ge, S., Usack, J. G., Spirito, C. M. & Angenent, L. T. Long-Term n-Caproic Acid Production from Yeast-Fermentation Beer in an Anaerobic Bioreactor with Continuous Product Extraction. *Environ. Sci. Technol.* **49**, 8012–8021 (2015).

34. Liang, S. & Wan, C. Carboxylic acid production from brewer's spent grain via mixed culture fermentation. *Bioresour. Technol.* **182**, 179–183 (2015).

35. Yang, P., Li, X. & Leng, L. Microbial dynamics with the introduction of brewery waste in a long-term chain elongation process for caproate production. *J. Chem. Technol. Biotechnol.* **98**, 2283–2294 (2023).

36. Zhu, X. *et al.* The synthesis of n -caproate from lactate: a new efficient process for medium-chain carboxylates production. *Sci. Rep.* **5**, 14360 (2015).

37. Wu, Q. *et al.* Upgrading liquor-making wastewater into medium chain fatty acid: Insights into co-electron donors, key microflora, and energy harvest. *Water Res.* **145**, 650–659 (2018).

38. Scarborough, M. J. *et al.* Increasing the economic value of lignocellulosic stillage through medium-chain fatty acid production. *Biotechnol. Biofuels* **11**, 200 (2018).

39. Spirito, C. M., Richter, H., Rabaey, K., Stams, A. J. & Angenent, L. T. Chain elongation in anaerobic reactor microbiomes to recover resources from waste. *Curr. Opin. Biotechnol.* **27**, 115–122 (2014).

40. Scarborough, M. J., Lawson, C. E., Hamilton, J. J., Donohue, T. J. & Noguera, D. R. Metatranscriptomic and Thermodynamic Insights into Medium-Chain Fatty Acid Production Using an Anaerobic Microbiome. *mSystems* **3**, (2018).

41. Xie, S. *et al.* Anaerobic caproate production on carbon chain elongation: Effect of lactate/butyrate ratio, concentration and operation mode. *Bioresour. Technol.* **329**, 124893 (2021).

42. Yang, P., Leng, L., Zhuang, H. & Lee, P.-H. Significant enhancement by casamino acids of caproate production via chain elongation. *Biochem. Eng. J.* **193**, 108879 (2023).

43. Liu, B. *et al.* Functional Redundancy Secures Resilience of Chain Elongation Communities upon pH Shifts in Closed Bioreactor Ecosystems. *Environ. Sci. Technol.* **57**, 18350–18361 (2023).

44. Ganigué, R., Sánchez-Paredes, P., Bañeras, L. & Colprim, J. Low Fermentation pH Is a Trigger to Alcohol Production, but a Killer to Chain Elongation. *Front. Microbiol.* **7**, (2016).

45. Candry, P. *et al.* Mildly acidic pH selects for chain elongation to caproic acid over alternative pathways during lactic acid fermentation. *Water Res.* **186**, 116396 (2020).

46. Wu, Q., Feng, X., Guo, W., Bao, X. & Ren, N. Long-term medium chain carboxylic acids production from liquor-making wastewater: Parameters optimization and toxicity mitigation. *Chem. Eng. J.* **388**, 124218 (2020).

47. Wu, L. *et al.* Medium chain fatty acids production from anaerobic fermentation of food wastes: The role of fermentation pH in metabolic pathways. *Chem. Eng. J.* **472**, 144824 (2023).

48. Wu, S.-L., Wei, W., Wang, Y., Song, L. & Ni, B.-J. Transforming waste activated sludge into medium chain fatty acids in continuous two-stage anaerobic fermentation: Demonstration at different pH levels. *Chemosphere* **288**, 132474 (2022).

49. Baleeiro, F. C. F., Raab, J., Kleinsteuber, S., Neumann, A. & Straeuber, H. Mixotrophic chain elongation with syngas and lactate as electron donors. *Microb. Biotechnol.* **16**, 322–336 (2023).

50. Fernández-Blanco, C., Veiga, M. C. & Kennes, C. Carbon dioxide as key player in chain elongation and growth of *Clostridium kluyveri*: Insights from batch and bioreactor studies. *Bioresour. Technol.* **394**, 130192 (2024).

51. Wang, J. & Yin, Y. Biological production of medium-chain carboxylates through chain elongation: An overview. *Biotechnol. Adv.* **55**, 107882 (2022).

52. Maynard, D. S., Miller, Z. R. & Allesina, S. Predicting coexistence in experimental ecological communities. *Nat. Ecol. Evol.* **4**, 91–100 (2020).

53. Widder, S. *et al.* Challenges in microbial ecology: building predictive understanding of community function and dynamics. *ISME J.* **10**, 2557–2568 (2016).

54. Connors, B. M. *et al.* Control points for design of taxonomic composition in synthetic human gut communities. *Cell Syst.* **14**, 1044-1058.e13 (2023).

55. Gowda, K., Ping, D., Mani, M. & Kuehn, S. Genomic structure predicts metabolite dynamics in microbial communities. *Cell* **185**, 530-546.e25 (2022).

56. Cheng, A. G. *et al.* Design, construction, and in vivo augmentation of a complex gut microbiome. *Cell* **185**, 3617-3636.e19 (2022).

57. Clark, R. L. *et al.* Design of synthetic human gut microbiome assembly and butyrate production. *Nat. Commun.* **12**, 3254 (2021).

58. Ostrem Loss, E., Thompson, J., Cheung, P. L. K., Qian, Y. & Venturelli, O. S. Carbohydrate complexity limits microbial growth and reduces the sensitivity of human gut communities to perturbations. *Nat. Ecol. Evol.* **7**, 127–142 (2023).

59. Wadler, C. S. *et al.* Utilization of lignocellulosic biofuel conversion residue by diverse microorganisms. *Biotechnol. Biofuels Bioprod.* **15**, 70 (2022).

60. Bernalier, A., Willems, A., Leclerc, M., Rochet, V. & Collins, M. D. *Ruminococcus hydrogenotrophicus* sp. nov., a new H_2/CO_2-utilizing acetogenic bacterium isolated from human feces. *Arch. Microbiol.* **166**, 176–183 (1996).

61. Liu, C., Finegold, S. M., Song, Y. & Lawson, P. A. Y. 2008. Reclassification of *Clostridium coccoides, Ruminococcus hansenii, Ruminococcus hydrogenotrophicus, Ruminococcus luti, Ruminococcus productus* and *Ruminococcus schinkii* as *Blautia coccoides gen. nov., comb. nov., Blautia hansenii comb. nov., Blautia hydrogenotrophica comb. nov., Blautia luti comb. nov., Blautia producta comb. nov., Blautia schinkii comb. nov.* and description of *Blautia wexlerae* sp. nov., isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* **58**, 1896–1902.

62. Liu, Y. *et al.* Shaping human gut community assembly and butyrate production by controlling the arginine dihydrolase pathway. 2023.01.10.523442 Preprint at https://doi.org/10.1101/2023.01.10.523442 (2024).

63. Esbensen, K. H. & Geladi, P. Principles of Proper Validation: use and abuse of re-sampling for validation. *J. Chemom.* **24**, 168–187 (2010).

64. Westad, F. & Marini, F. Validation of chemometric models – A tutorial. *Anal. Chim. Acta* **893**, 14–24 (2015).

65. Lopez, E., Etxebarria-Elezgarai, J., Amigo, J. M. & Seifert, A. The importance of choosing a proper validation strategy in predictive models. A tutorial with real examples. *Anal. Chim. Acta* **1275**, 341532 (2023).

66. J. Steinbusch, K. J., M. Hamelers, H. V., M. Plugge, C. & N. Buisman, C. J. Biological formation of caproate and caprylate from acetate : fuel and chemical production from low grade biomass. *Energy Environ. Sci.* **4**, 216–224 (2011).

67. Spirito, C. M., Marzilli, A. M. & Angenent, L. T. Higher Substrate Ratios of Ethanol to Acetate Steered Chain Elongation toward n-Caprylate in a Bioreactor with Product Extraction. *Environ. Sci. Technol.* **52**, 13438–13447 (2018).

Chapter 3: Community production of hexanoate is sensitive to starting pH and mediated by species-specific metabolite dynamics

This manuscript appears as part of a larger paper in combination with the previous chapter:

Hayes, M. M., Nightingale, N., Overmeyer, K., Thompson, J, Feng, J., Coon, J., Venturelli, O.S. Exploring the landscape of medium chain fatty acid production of a synthetic anaerobic bacterial community using a bottom-up systems biology approach. Nature Communications.

Contributions: Madeline Hayes conducted the experiments, wrote the text, and created the figures. Nicole Nightingale and Katie Overmeyer processed metabolomics samples. Jaron Thompson wrote portions of the modeling code and contributed to the text. Jun Feng determined the culturing conditions. Josh Coon oversaw the metabolomics sample processing. Ophelia Venturelli revised the text and advised on the experiments.

Abstract

pH is a key determinant of microbial metabolism that impacts a variety of cellular features, from the stability of the membrane, to the efficiency of metabolite import, to the function of ATP generating proteins. For microbial communities, species can modify their local pH environment and impact other community members through the production and consumption of organic acids produced during fermentation. For chain elongating communities, pH is a large determinant of the concentration and length of carboxylic acid output, but how individual species perceive and respond to changes in pH is unknown. Further, understanding how individual species impact each other through metabolite exchange mediated by pH sensitivity is a key step in understanding the larger metabolic network of microbiomes. Here we aimed to understand how changes in pH impact the medium chain fatty acid (MCFA) production of a panel of communities optimized for maximum hexanoate production in two media environments. We determined that communities show differential sensitivity to media starting pH, and that *Lactobacillus diolivorans* (Ld) is a key community member which increases the sensitivity of the community to pH changes. We examined a large metabolic panel for one promising community, revealing that Ld digests xylose, producing acetate and lactate, and at pH 5.7, abolishes the ability of key MCFA producer *Clostridium kluyveri* (Ck) to effectively produce hexanoate. We also show that two pH-mediating species, *Eggerthella lenta* (Elen) and *Anaerostipes caccae* (Ac), offers a small but detectable buffering against this pH change. Deciphering the specific interactions within this larger synthetic community is a novel look at chain elongating community function.

Importance

Microbial communities are sensitive to their environments, receiving molecular signals and responding with changes in their metabolism. pH is an environmental parameter that has fundamental impacts on the harvesting of energy from metabolites within the cell, and is a large determinant of the metabolic output of any cell. Microbial communities can influence their local pH environment by producing and consuming organic acids, a common biproduct of fermentative pathways. This dynamic flux of acids in and out of the system mediates interspecies interactions which drive community phenotype. For MCFA producing communities, pH determines whether species in the community will utilize chain elongation to survive, and leveraging the influence of pH on this useful metabolism is a chief goal of many engineering approaches. By identifying key species and metabolites that directly impact the favorability of chain elongation, here we offer a novel target for engineering production of MCFA using microbial bioprocessing.

Introduction

Microbial metabolisms are impacted by a wide range of environmental parameters, including temperature, salinity, and nutrient availability. In particular, pH has been observed to be a determining factor in community composition and function, correlating with community similarity across a diverse set of geographical locations[1] and influencing metabolism.[2] pH directly impacts the ability of bacteria to produce ATP via proton pumps,[3] and therefore plays a key role ecological niche differentiation.

Natural systems have dynamic pH ranges which shift in response to abiotic and biotic factors. In the context of fermentative communities, the production of acidic compounds acts to lower environmental pH. Carboxylates and organic acids are natural products of microbial communities in many different environments. Fermentative metabolisms can produce a wide variety of lengths and branch arrangements depending on the substrate,[4] temperature,[5] and gas headspace,[6] and this production can induce feedback loops that can give rise to oscillatory behavior.[7] Identifying deterministic factors and control points for metabolism output is an important element in understanding community functionality and predicting community behavior.

Chain elongation as a metabolism is sensitive to pH. While key MCFA producer *Clostridium kluyveri* has been shown to favor neutral pH for chain elongation,[8] there are many examples of complex communities[9–11] and strains[11] that favor a lower pH. This includes *Megasphaera elsdenii*[12] (pH6.5). However, low pH may induce increased MCFA toxicity,[13] limiting the overall productivity. The interplay between carbon source availability, extracellular pH, and metabolism is highly complex, and the impact of dynamic acid production and consumption impacts on hexanoate production is not well understood.

In bioprocessing, pH is an important parameter for controlling reactions. In practice, pH is most often used to inhibit undesirable metabolisms, such as methanogenesis,[14] which converts carbon to products in direct competition with the optimal output. In bioreactors, pH can be tightly controlled, mitigating the negative impacts of product inhibition and toxicity. However, requiring additional inputs to industrial systems necessarily decreases the profit margin and increases the carbon footprint, as those inputs must be produced, transported, and then the waste streams mitigated. Ideally, a community or organism that is robust to changes in pH would be selected for large-scale applications. Identifying communities which maximize substrate utilization and production efficiency without the need for maintaining a selective environment is ultimately a goal of bioprocess engineering.

Here we aim to investigate the impact of pH on hexanoate production of a panel of synthetic bacterial communities previously identified to produce high hexanoate. The panel consists of communities optimized for hexanoate production at two pH levels (7.0 and 5.7), and understanding how changes in pH from their optimized environment affects their performance will inform our understanding of how species-species interactions and community composition interfaces with environmental parameters. Re-examining the highly productive communities at non-optimal pH reveals that *Lactobacillus diolivorans* (Ld) plays a unique role in substrate utilization and hexanoate production. Ld's presence is associated with increased sensitivity to pH, and when removed from the community, hexanoate production at low pH is effectively rescued. A broad panel of metabolites identifies lactate as a potential mediator of this interaction. Additionally, two other species (Ac and Elenta) are observed to have small but statistically significant positive impacts on community hexanoate production at low pH, offering additional insight.

Results

Hexanoate production is sensitive to starting media pH

We previously utilized computational models to optimize community composition to maximize C6 production in two synthetic media (Chapter 2 of this work). These media differed in composition and also pH, one with a starting pH of 7.0 (SCRv1), and the other with a starting pH of 5.7 (SCRv2). The media were not adjusted to match pH because a difference in pH allowed for the possibility of optimizing differential communities that produced hexanoate at different pH levels. By providing two environments for optimization, we allowed for a wider range of environmental factors to influence the model-guided design of communities, hoping to identify environment-specific species interactions that support hexanoate production.

Model guided exploration of the sample space generated a large dataset of communities tested in each media type: 421 communities were tested in SCRv1, and 371 communities were tested in SCRv2. One key goal of the modeling was to identify high-hexanoate producing communities to explore further, providing a platform for mechanistic investigations. For this purpose, we narrowed our original datasets from their full size to the top 46 hexanoate-producing communities in each media. This number maximizes the high-throughput setup while maintaining tractability (i.e., 46 communities fills one plate with two replicates per community, with two replicates of an internal benchmark), while also providing a wide range of examples from each media type. Both sets contained communities of various complexities, ranging from monospecies to high complexity (11-15 members). By maintaining some variety in community composition but limiting this work to a few key factors, we aimed to test hypotheses generated from interpreting model parameters and answer questions about the favorability of MCFA

production and the capability of multiple MCFA producers to cumulatively contribute to functional output, and investigate the potential for functions unique to individual species or redundant among functional groups.

In order to isolate the importance of starting pH as a variable for hexanoate production in these media we retested the top-producing communities from each dataset at a different starting pH (Figure 1A). For SCRv1, the native (original) pH is 7.0, and the non-native (new) pH was 5.7. For SCRv2, the native (original) pH is 5.7, and the non-native (new) pH was 7.0. We then measured community composition and MCFA production as previously described (see Chapter 2 methods). To categorize the response profiles, here we define sensitivity as the percent change in hexanoate production from the native pH level (i.e. -100% is abolishment of production when pH is changed, +100% is doubling of production when pH is changed). For each medium, the tested communities were ranked by their sensitivity (Figure 1B and 1D, lower bar plots). For SCRv1 (Figure 1B), communities show negative sensitivity to a drop in pH: most tested communities show little to no hexanoate production when pH is dropped from 7.0 (native) to 5.7 (nonnative). This preference for neutral pH is mirrored in the SCRv2 communities (Figure 1D), and sensitivity is largely positive: SCRv2 communities show a dramatic increase in hexanoate production when pH is raised to 7.0.

Lactobacillus diolivorans *demonstrates unique effect on pH sensitivity in SCRv1*

In order to parse any patterns in community composition with regards to pH sensitivity, the species present in each community are shown as a simple present/absent mapping above the sensitivity bar plots (present = blue, absent = white). Based on our interpretation of the gLV and LR parameter values, we hypothesized that lactobacilli may be impacting MCFA production, and

when arranged by sensitivity, patterns in lactobacilli emerge (Figure 1C and Figure 1E, red bars). For SCRv1, Lb and Ld appear in roughly the same number of communities, but Ld does not appear in any of the most resilient communities. For SCRv2, lactobacilli are largely absent from *any* of the top producing communities. This reinforces the interpretations of the LR parameter values from Chapter 2 (Figure S10): lactobacilli have a strong negative effect on hexanoate production in both media, but Ld may be uniquely detrimental in SCRv1, as it appears in four of the fifteen parameters which make up the most negative parameter space.

To further understand any compositional similarities and probe the relationship of Ld and pH sensitivity in SCRv1, we visualized the observed communities using MDS, incorporating relatedness of species in each community. Communities can be represented as binary vectors, where each element of the vector is one or zero depending on whether the corresponding species is present or absent, respectively. However, when considering only the presence or absence of species, the full design space cannot be optimally projected into lower dimensions in a way that preserves distances between points. By augmenting each community vector to capture information about species phylogeny, different community vectors can vary in their alignment depending on phylogenetic similarity (as in Chapter 2 methods). In this way, we aimed to visualize the phylogenetic relationship between communities which show sensitivity or resilience to pH changes (i.e. closer clustering indicates closer relatedness). We expect to see communities with similar sensitivity profiles clustering. SCRv1 communities show organization across their pH sensitivity gradient (Figure 2A). When color coded according to the same sensitivity ranking as in Figure 1, we see that the least sensitive communities occupy a similar space. These communities also tend to lack Ld (Figure 2B); most (but not all) communities which show less sensitivity to pH changes do not contain Ld. To quantitatively evaluate the relationship between

Ld presence/absence and C6 production at different pH levels, we compared average C6 production of communities in SCRv1 which did nor did not contain Ld at each pH level (Figure 2C), and found that Ld presence did result in a statistically significant decrease in C6 production, but only at pH5.7. This comparison was also made for Lb, and the same impact was not observed (Supplementary Figure S1).

Individual species demonstrate nonredundant functions, distinct impact on measured metabolites

In order to investigate individual species contributions to hexanoate production and sensitivity to pH, we constructed communities that exclude individual species ('Leave-One-Out' experiments). We expect that if any individual species in the community is impacting the growth of other community members or community hexanoate production, a change in community composition or hexanoate production will be observed when the species is left out (Figure 3A). For this approach, we selected a 10-member community from the SCRv1 panel (white starred community in Figure 1B). This community was identified as a promising target for several reasons: (1) it contained all four MCFA producers, allowing for an assessment of the contribution of each species to overall MCFA production, (2) it contained Bh and *E. lenta*, allowing for the testing of previously generated hypotheses regarding these species' positive impacts on Ck, and (3) it showed consistent hexanoate production with low variation between replicates.

This community was tested again at both pH 7.0 and 5.7, along with each nine-member subcommunity (eleven communities total), and growth and hexanoate production was assessed. To gain a more complete picture of the metabolite space, a broader panel of metabolites were tested (Figure 3B, columns), including xylose, succinate (components in the medium), acetate (a component in the medium and also an intermediate), lactate (an intermediate), and C6 (the

target). We also measured endpoint pH. We see that for any individual community, growth is largely unchanged between pH levels (Figure 3B, left column, 'Absolute Abundance'). This indicates community composition is not sensitive to changes in starting media pH, and that differences in production or metabolite concentrations is unlikely to be directly related to biomass accumulation.

In the metabolic landscape, we observe four key metabolic shifts in response to a species being excluded. When Ld is left out (-Ld, boxed), no xylose appears to be consumed, indicating that Ld is the key xylose consumer in this community. When Ck is left out (-Ck), no succinate is consumed, which aligns with the known fact that Ck can consume succinate as a primary carbon source,[15] and hexanoate production is significantly lower, confirming Ck as the primary hexanoate producer in this community. When Me is left out (-Me), lactate accumulates in the medium, which also aligns with the well-established fact that Me consumes lactate.[12] Finally, Ld- hexanoate production shows far less sensitivity to starting pH than other communities. We also measured the endpoint pH of each community, and noted that Ld- had a significantly higher pH at 48 hours than the full 10-member community.

From these observations, we conclude that there is a complex interplay of intermediate metabolites, starting and ending pH, and hexanoate production. When xylose is not degraded (Figure 3B, xylose column), no lactate or additional acetate is added to the intermediate metabolite pool (Figure 3B, acetate column and Supplementary Figure S2), and the absence of these acids results in less media acidification and a higher endpoint pH (Figure 3B, pH column). However, Ld is not the only species responsible for media acidification, as the Ck- community also showed slight but significantly higher pH at 48 hours than the 10-member community at both starting pHs. This may be due to butyrate production by Ck (see Supplement S2). The more

robust hexanoate production of Ld- with regards to starting pH could be partially explained by media acidification, where addition of lactate to the system drops the pH to a level too low for favorable hexanoate production by Ck. Given that this effect is only seen in communities (see Figure 2 pH sensitivity ranking – Ck monospecies shows ~ -40% sensitivity as compared to -100% in this community), we conclude that this must be mediated by species-species interactions.

Discussion

Elucidating the molecular mechanisms of community function requires an understanding of the environmental factors which impact the favorability of output metabolisms. In communities which rely on acidic intermediates for functionality, pH plays an important role in how species-species interactions are mediated. Here we demonstrate that bacterial communities show varying degrees of sensitivity to pH with regards to their hexanoate production. Using communities previously identified in two synthetic media, we demonstrated that communities show phylogenetic similarity with regards to their sensitivity level. Zooming in on one 10-species community, *L. diolivorans* was identified as the key xylose degrader, and also responsible for pH sensitivity. When removed, hexanoate production was improved. This demonstrates that metabolites digested in the multi-step cascade of MCFA production from plant sugars contribute to modifying the pH environment.

While not observed directly, it seems likely that lactate is the key intermediate that mediates this interaction. Ld is a heterofermentative organism, producing both acetate and lactate from xylose.[16] When Ld is absent, Me, a known lactate consumer, has dramatically decreased growth, as does *Anaerostipes caccae* (Ac), another known lactate consumer.[17] When Me is absent, lactate accumulates to some degree in the media, indicating that it exists in transient concentrations between being produced by Ld and consumed by other community members. Lactic acid is a strong organic acid, and its production likely acidifies the media significantly. This is supported by the higher endpoint pH of the Ld- community. Intermittent observations would be required to confirm this proposed model. Because Ck in monoculture does not demonstrate the same abolishment of hexanoate production at pH 5.7 (Figure 1), we can assume that there is some lower threshold below which hexanoate production is totally unfavorable, and

for the time that the media occupies this space, Ck produces no C6. This raises the possibility of consumption of lactate by other species increases the pH eventually, returning the environment to a favorable regime. Extending the timeframe of observation would be required to probe this possibility.

Interrogating other subcommunities reveals possible positive interactions as well. When *E. lenta* is left out, hexanoate production is slightly but significantly lower than the 10-member community at pH 5.7. *E. lenta* was incorporated for its ability to modulate pH through arginine metabolism, which produces ammonia.[18] This indicates that a small amount of buffering by this species may be contributing to improving hexanoate production favorability at low pH. This is consistent with Ck-Elen parameter values from the model. This same profile is observed in the Ac- community. While the linear regression parameters did show a slight positive contribution for Ac-Ck, it seems more likely that any positive effect is mediated through lactate consumption; Ac-Ld and Ac-Lv are stronger positive parameters, indicating that consumption of lactate produced by those species may be beneficial. However, why the Me- community does not show a similar profile is unknown.

While stark differences in hexanoate production are observed between pH levels, the abundance of Ck does not appear to differ significantly. This is in line with previous observations of Ck metabolism, where hexanoate production is not linear with biomass accumulation under all conditions.[8,19,20] This raises the question of how appropriate a modelling approach that explicitly parameterizes species abundance is for understanding this process; our linear regression model incorporates both species abundance and binary presence/absence, but if emergent or higher order interactions cause these parameters to disagree in value, this may impact prediction accuracy and interpretability.

Computational methods have been developed for understanding community function as mediated by pH. Ecological frameworks incorporating pH[21] indicate that this parameter plays an important role in predicting biodiversity in soil environments. Understanding how local dynamic shaping of pH environments is critical to predicting stability of communities as well.[22] On a species level, attempts have been made to predict the preferred pH range based on genomic evidence,[23] as well as using pH as a proxy for species-species interactions without molecular knowledge of cross-feeding substrates.[24] Incorporating these methods into the study of microbial bioprocessing holds tremendous promise of further elucidating and controlling the processes useful for solving anthropogenic challenges.

Methods

Bacterial Growth and Sample Collection

Strain Maintenance and Preculturing – All anaerobic culturing was performed in an anaerobic chamber (Coy) with an atmosphere of 2.5 ± 0.5% H_2, 15 ± 1% CO_2 and balance N_2. All prepared media and materials were transferred into the chamber at least overnight before use to equilibrate to anaerobic atmosphere. Supplementary Table 1 details strain source and individual culturing conditions. Permanent stocks of each species were maintained at -80C in 25% glycerol. Batches of single use glycerol stocks (SUGS) were produced for each strain by first isolating a single colony from the permanent stock on anaerobic basal broth (ABB) or MRS solid agar plate, which were then used 5-7mL liquid media and incubated at 37°C until log phase, mixed with equal volume 50% glycerol, aliquoted (400µL) into Matrix Tubes (Thermo Fisher), and stored at -80°C until use. SUGS parent cultures were 16S rRNA gene Illumina sequenced to verify culture purity. For each experiment, a serial culturing strategy was used to ensure ample growth and similar growth phase of each species prior to experimental inoculation (see Supplementary Methods).

Monoculture dynamic culturing – Precultured cells were pelleted at 4000RPM then resuspended in experimental media (compositions in Supplement spreadsheet 1: media recipes) at normalized OD600=0.5 for inoculation. Briefly, diluted to a final OD of 0.1 and volume of 1mL and incubated at 37°C for 48 hours covered with a semipermeable membrane, and sampled (100µL) every 4 hours for OD600 measurement (Tecan F200). was measured in two ways to ensure samples were in dynamic range: a) aliquoting 100µL into 96-well microplate or aliquoting 20µL into 180µL blank media. Individual replicates were used for each time point to avoid adverse

effects from disturbing the culture during growth. Supernatant and cells were harvested at 48 hours (see Sample Collection below).

Community batch culturing - For each experiment, precultured cells were pelleted at 4000rpm then resuspended in experimental media (compositions in Supplement spreadsheet 1: media recipes) at normalized OD600=0.5 for inoculation. Community combinations were arrayed in 96DW plates by pipetting each species at equal volume into the appropriate well (Tecan Evo Liquid Handling Robot). Each community was then diluted to a final OD of 0.1 and volume of 1.0mL and incubated at 37°C for 48 hours covered with a semipermeable membrane (Breathe Easy). At 48 hours, final OD600 was measured in two ways to ensure samples were in dynamic range: (a) aliquoting 100µL into 96-well microplate or (b) aliquoting 20µL into 180µL blank media. Supernatant and cells were harvested at 48 hours (see below).

Sample collection – After each experiment, cells and supernatant were harvested and stored at -80°C. Briefly, cells were pelleted in 96 deep well plate at 4000rpm and supernatant removed without disturbing the pellet. Both cells and supernatant were stored at -80°C until processing.

Sample Processing

Genomic DNA extraction and sequencing library preparation – Collected cell pellets were lysed for extraction and purification of genomic DNA using a modified Qiagen Blood and Tissue kit. Briefly, cells thawed in a room temperature water bath, then resuspended in 20 mg/mL lysozyme (from chicken egg whites Sigma) in enzymatic lysis buffer (20mM Tris-HCl [Invitrogen], 2mM sodium EDTA [Sigma], 1.2% Triton X-100 [Sigma]). Plates were covered in a foil seal and allowed to incubate for 30 min at 37°C with orbital shaking at 600rpm. Then, 25µL 20 mg/mL proteinase K solution and 200 µL Qiagen Buffer AL were added and allowed to incubate for

30min at 56°C with orbital shaking at 600rpm. DNA was precipitated by adding 200µL 100% EtOH, then purified over Pall DNA-binding column plates. Qiagen AW1 and AW2 500µL washes were performed and the column allowed to dry for 5min at room temperature. DNA was eluted with 110µL Qiagen buffer AE preheated to 56°C. Samples were stored at -20°C until further use.

Genomic DNA concentrations were measured using a SYBR Green fluorescence assay (Invitrogen) according to the manufacturer's instructions and then normalized to a concentration of 1 ng μL^{-1} by diluting in molecular grade water using a Tecan Evo Liquid Handling Robot. Briefly, genomic DNA samples were removed from −20 °C and thawed in a room temperature water bath and combined with 95 µL of SYBR Green diluted by a factor of 100 in TE buffer (Integrated DNA Technologies) in a black 384-well microplate. A standard curve was constructed in triplicate using 5µL of standard concentrations of 0, 0.5, 1, 2, 4, and 6 ng μL^{-1}. Each sample was then measured for fluorescence with an excitation/emission of 485/535 nm using a Tecan Spark plate reader. Concentrations of each sample were calculated using the standard curve and a custom Python script was used to compute the dilution factors and write a worklist for the Tecan Evo Liquid Handling Robot to normalize each sample to 1 ng μL^{-1} in molecular grade water. Samples with DNA concentration <1 ng μL^{-1} were not diluted. Normalized genomic DNA samples were stored at −20 °C until further processing.

Normalized genomic DNA was then used for 16S PCR amplification using Invitrogen Phusion and custom dual-indexed primers (see Supplementary spreadsheet 2: primers). Primers were arrayed in skirted 96-well PCR plates (Thomas Scientific) using an acoustic liquid handling robot (Labcyte Echo 550) such that each well received a different combination of one forward and one reverse primer (0.1 µL of each). After liquid evaporated, dry primers were stored at

−20 °C. Primers were resuspended in 15 µL PCR master mix (0.2 µL Phusion High Fidelity DNA Polymerase (Thermo Scientific), 0.4 µL 10 mM dNTP solution (New England Biolabs), 4 µL 5× phusion HF buffer (Thermo Scientific), 4 µL 5 M Betaine (Sigma-Aldrich), 6.4 µL Water) and 5 µL of normalized genomic DNA to give a final concentration of 0.05 µM of each primer. Primer plates were sealed with Microplate B seals (Bio-Rad) and PCR was performed using a Bio-Rad C1000 Thermal Cycler with the following program: initial denaturation at 98 °C (30 s); 25 cycles of denaturation at 98 °C (10 s), annealing at 60 °C (30 s), extension at 72 °C (60 s); and final extension at 72 °C (10 min). 2 µL of PCR products from each well were pooled and purified using the DNA Clean & Concentrator (Zymo) and eluted in water. The resulting libraries were sequenced on an Illumina MiSeq using a MiSeq Reagent Kit v3 (600-cycle) according to the manufacturer's instructions to generate 2 × 300 paired-end reads.

Quantification of species abundance - Sequencing data were demultiplexed using Basespace Sequencing Hub's FastQ Generation program. Custom python scripts were used for further data processing (method adapted from Venturelli et al. Mol. Syst. Biol., 2018) Paired end reads were merged using PEAR (v0.9.10)83 after which reads without forward and reverse annealing regions were filtered out. A reference database of the V3–V5 16S rRNA gene sequences was created using consensus sequences from next-generation sequencing data or Sanger sequencing data of monospecies cultures. Sequences were mapped to the reference database using the mothur (v1.40.5)84 command classify.seqs (Wang method with a bootstrap cutoff value of 60). Relative abundance was calculated as the read count mapped to each species divided by the total number of reads for each condition. Absolute abundance of each species was calculated by multiplying the relative abundance by the OD600 measurement for each sample. Samples were

excluded from further analysis if >1% of the reads were assigned to a species not expected to be in the community or if they had >1% unclassified reads (indicating contamination).

Measurement of MCFA for pH swap (Figure 1):

Standard Stock Preparation - Individual solutions of Butyric Acid, Hexanoic Acid, and Octanoic Acid were purchased from Sigma-Aldrich with purity no less than 99%. 10 µL of each organic acid was aliquoted into a 1.5 mL vial followed by 970 µL of acidified acetonitrile (0.2% HCl) solution to create a 10000 ppm stock solution. The stock solution was used to prepare standards at concentrations of 5 ppm, 12.5 ppm, 25 ppm, 50 ppm, 75 ppm, 100 ppm, 150 ppm, 250 ppm, 500 ppm, 1000 ppm, 2000 ppm, 4000 ppm and 5000 ppm in 100 µL acidified acetonitrile (0.2% HCl). The standards were further diluted by the addition of 200 µL ethyl acetate. Samples were thoroughly mixed and centrifuged at 12,000 x g for 5 min at 4°C. 100 µL of upper layer was transferred to amber glass autosampler vial with fused glass insert for analysis.

GC-MS Analysis - Samples were analyzed using a GC-MS instrument set up to perform chemical ionization comprising a Trace 1310 GC coupled to an ISQ series mass spectrometer (Thermo Scientific) with methane as the ionization gas. Analytes were injected into a split/splitless heated injector at a temperature of 200 °C using an AI1310 autosampler. The samples were split 1:5 by the injector and then injected onto a 30 m Stabilwax DA column (Restek) using helium at a flow of 1.20 mL/min. A ramped temperature mode was employed using the following gradient:

Number	Retention time [min]	Rate [°C/min]	Target value [°C]	Hold time [min]
1	2	0	50	2
2	12	12.5	175	0
3	17	50	225	4

The mass spectrometer transfer line and ion source temperatures were set to 200 °C with a mass scan range of 43 – 250 m/z.

Data Analysis - Peak area extraction was performed through TraceFinder 4.0 using an internally created processing method. Calculations of sample concentrations were performed manually based on the peak area counts of each fatty acid standard at 0-5000 ppm concentrations. A regression model was fit to all standards consisting of at least four consecutive standard levels for lower concentration curve and at least three standard levels for higher concentration curve. If the peak area of the analyte was below the peak area of the lowest standard used in the curve, the resulting calculated concentration is stated as 0.

Measurement of MCFA for LOO (Figure 3):

Standard stock preparation – A mixed standard stock solution containing butanoic, pentanoic, hexanoic, and octanoic acid (0.318 mM, 0.271mM, 0.074mM, 0.117mM respectively) in their natural abundances as well as a separate solution of the same deuterated compounds (62.4mM, 53.1mM, 14.6mM, 22.8mM) in acetonitrile was prepared. The natural and deuterated standards were used to run a 10-point calibration curve tailored in dilution to the expected concentrations of the experimental samples (based on a dilution series of one of the samples). Each experimental sample was diluted into internal standard solution (1%v/v or as determined by needs, depending on the concentration of the experimental sample) to ensure accurate peak detection.

GC-MS Analysis

Samples were analyzed using an Agilent 7890 GC with LECO L-PAL 3 autosampler and LECO Pegasus BT MS. The mobile phase was helium at a constant flow. Samples were injected into a

split/splitless heated injector at temperature 225°C. The samples were split 10:1 by the injector then injected onto a 20m Stabilwax DA column using the helium flow. A ramped temperature mode was employed using a temperature gradient with 50°C hold for 2.5min, ramp 10°C/min to 250°C. The MS transfer line was set to 250°C with a mass scan range of 20-250m/z. Blank samples were injected ever ~8-10 samples to prevent column overload.

Data analysis

Peak extraction was performed using LECO software using an internally created processing method. Calculations of the sample concentrations were automatically calculated by the software. A linear standard curve was fit to the 10-point calibration curve.

Measurement of broader metabolite panel (Figure 3):

Standard stock preparation – A combined standard stock solution of all measured compounds was prepared by adding each compound at the highest expected concentration from the experimental samples (based on pilot experiments). Compounds were divided into non-volatile (xylose, succinate, lactate) and volatile (acetate) solutions. A 10-point standard curve was prepared from each of these solutions.

HPLC Analysis – Samples were analyzed using a Aginlent 1260 infinity HPLC with quartenary pump and chilled autosampler. The mobile phase was 0.02N sulfuric acid (filtered). 50uL samples were injected at 50°C onto an AMinex HPX-87H column with cation H guard column (300x7.8mm). A constant temperature of 50°C with a run time of 28min was used for detection.

Data analysis – all data analysis was completed using ChemStationV (Agilent). Concentrations were calculated from 10-point standard curve and comparison against individual QC samples.

Endpoint pH: pH endpoint was collected with a pH probe (MetlerToledo).

MDS With Phylogeny

All communities in each media type were transformed into vectors where present =1 and absent =1. Community vectors that contain a species that is a descendant of a parent node in the phylogenetic tree will populate the element of the vector corresponding to that node. Consequently, the alignment of two vectors corresponding to two different communities depends on the phylogenetic relationships between species. To create the functional landscape of the full set of experimentally characterized communities, we performed multidimensional scaling (MDS) on the set of vectors augmented with phylogenetic information using a Euclidean distance metric to project the data onto 2-dimensions using Scikit-Learn's MDS function.

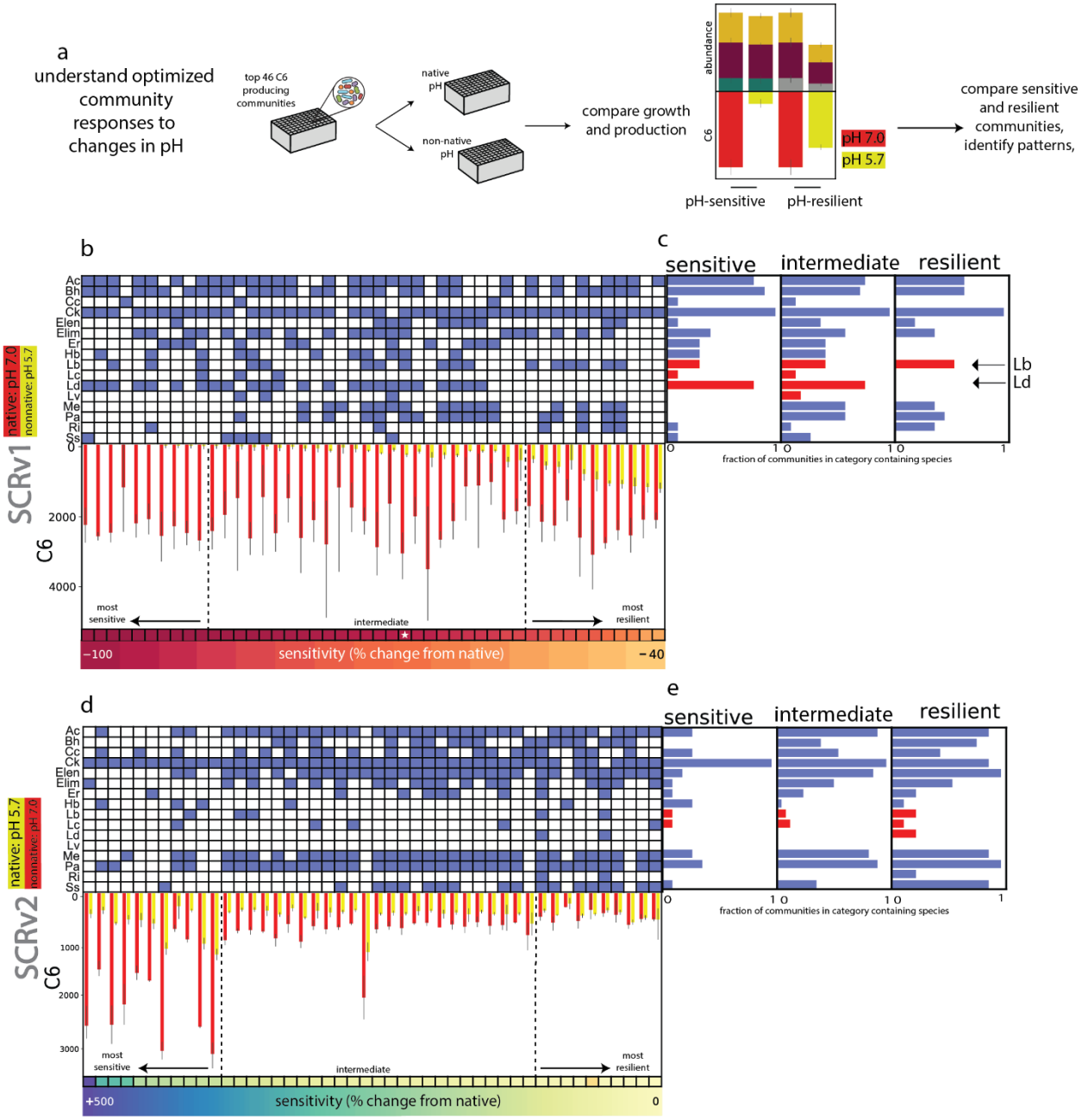

Figure 1: Highest-hexanoate producing communities show sensitivity to pH.

A. Experimental approach scheme. The top 46 hexanoate-producing communities from each medium were selected for further testing. These communities were tested again in a non-native pH (SCRv1: native = pH 7.0, nonnative = pH 5.7; SCRv2: native = pH 5.7, nonnative = pH 7.0). Community growth and production were compared to assess sensitivity (sensitive = large change

in hexanoate production, resilient = small or no change in hexanoate production). **B and D: Communities ranked by pH sensitivity. Top:** species added (blue) to each community. **Bottom**: hexanoate production of each community in pH 7.0 (red bar) and pH 5.7 (yellow bar). Communities are ranked by sensitivity (bottom color bar, % change in C6 production from native pH; SCRv1 shows generally reduction in production, SCRv2 shows generally improvement or no change in production). **C and E: Ranked communities show species-specific composition as resiliency improves.** Number of communities in each category (sensitive, neutral, resilient; from panels B and E) which contain each species. Lactobacilli are highlighted. In SCRv1 (panel C), Ld does not appear in resilient communities, but Lb does not share this pattern. In SCRv2 (panel E), lactobacilli are generally absent from highest-producing communities.

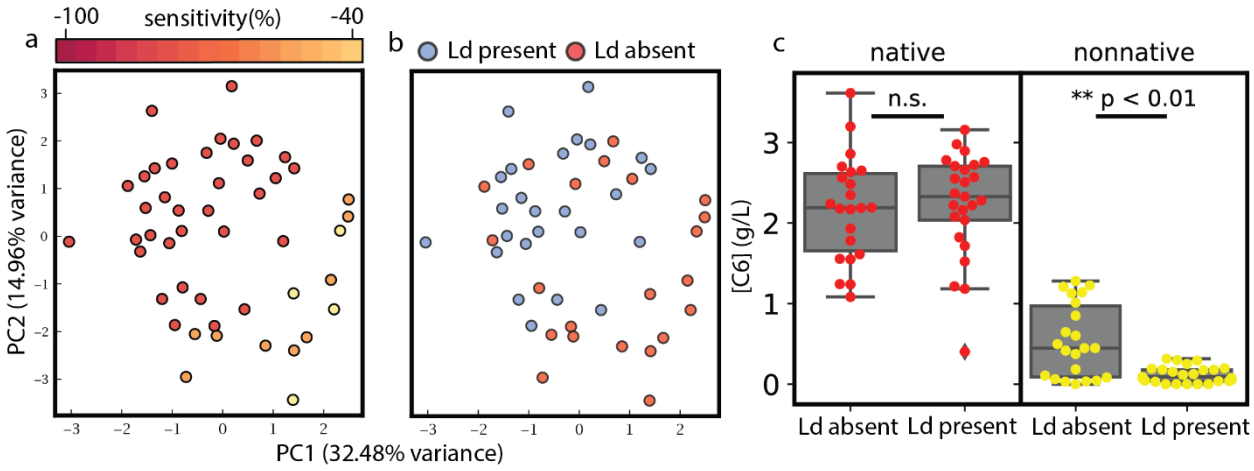

Figure 2: Communities sensitive to pH change in SCRv1 show compositional similarity with Ld as a key determinant

A. MDS with phylogeny. Composition and phylogeny of each community was projected onto 2D sample space using MDS. Each point represents one community, and the distance between points indicates similarity in composition including relatedness of constituent species. Colors correspond to pH sensitivity (color bar, top). **B. MDS with phylogeny and Ld presence/absence**. MDS from panel A was transformed to show communities which contained (blue) or lacked (red) Ld. **C. Ld is associated with increased pH sensitivity in SCRv1.** Comparison of hexanoate production (y-axis) between communities in which Ld is present or absent (x-axis) reveals no relationship at pH7.0 (red), but a significant decrease in hexanoate production at pH5.7. $p>0.001$, Mann-Whitney U test.

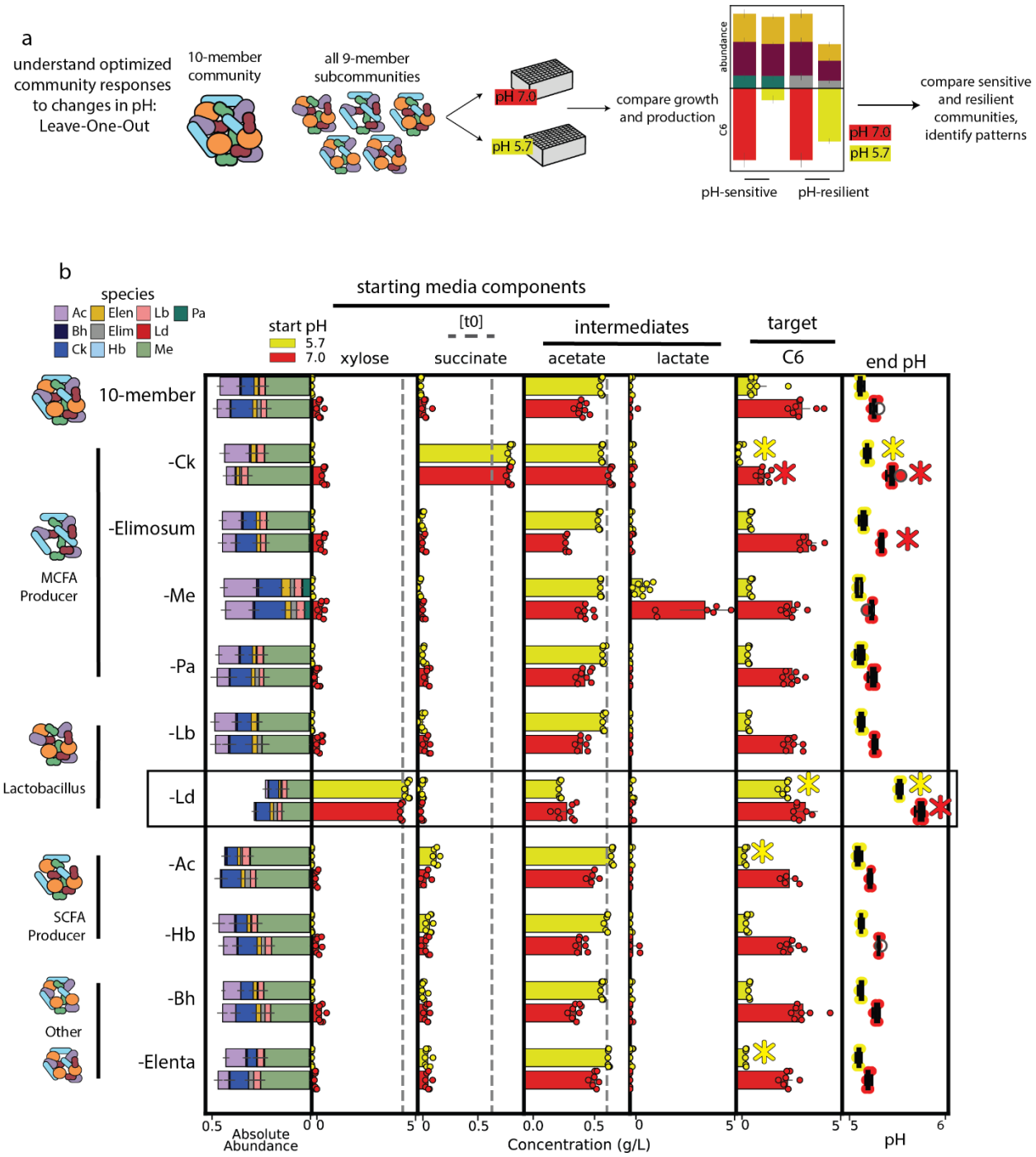

Figure 3: Leave-One-Out experiment reveals species-specific contributions to metabolite production and pH sensitivity.

A. Experimental setup. One 10-member community was chosen for further investigation (white starred community from Figure 1B). This community and each 9-member sub-community was

tested again at pH 5.7 and 7.0, and profiles for MCFA production, xylose, acetate, succinate, and lactate concentrations, and end pH at 48 hours were compared. Statistical evaluation limited to hexanoate and pH only for clarity. **B. Community growth, metabolite profiles and end pH. Left:** comparison of full 10-member community (top) and each 9-member sub-community (descending, organized by functional groups) growth at 48 hours (calculated absolute abundance, x-axis). **From left,** xylose, succinate, lactate, acetate, and hexanoate concentrations for each community at 48 hours (g/L, x-axes). Endpoint pH was also measured (f**ar right panel**). If medium contained the metabolite at t0, concentration is indicated by grey hashed line. All communities are shown as the average of 8 biological replicates, represented by circles. Statistical analysis is two-way ANOVA using Tukey's HSD $p<0.001$ where each community was compared within pH levels (i.e. only communities with the same starting pH were compared). * represents significant difference between the starred community and the 10-member community (see Supplementary Figure S2 for all comparisons).

Supplementary Information

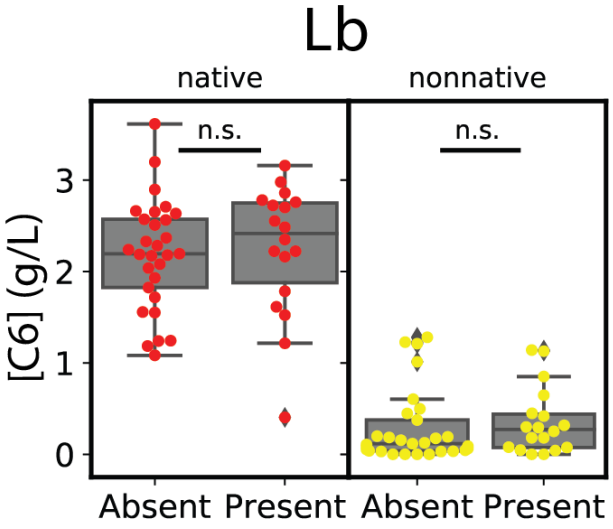

Figure S1: Lb impact on pH sensitivity.

The presence or absence of Lb in communities (each point) measured for C6 production (y-axis) in SCRv1 native pH (7.0, left, red) and SCRv1 non-native pH (5.7, right, yellow). Communities were compared by Mann-Whitney U test.

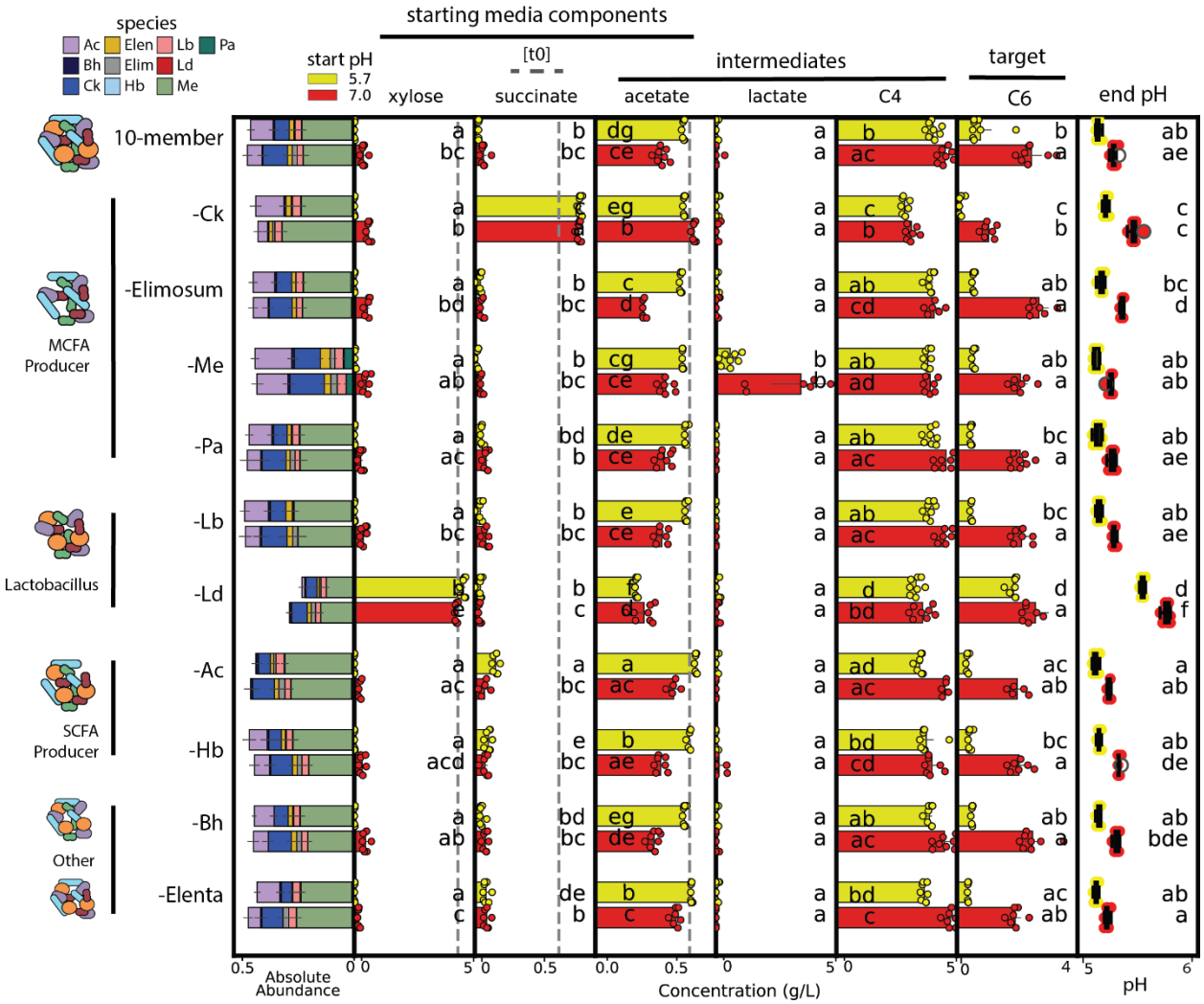

Figure S2: Additional statistics for F3, including measured butyrate.

Left: comparison of full 10-member community (top) and each 9-member sub-community (descending, organized by functional groups) growth at 48 hours (calculated absolute abundance, x-axis). **From left,** xylose, succinate, lactate, acetate, butyrate, and hexanoate concentrations for each community at 48 hours (g/L, x-axes). Endpoint pH was also measured (f**ar right panel**). If medium contained the metabolite at t0, concentration is indicated by grey hashed line. All communities are shown as the average of 8 biological replicates, represented by circles.

Statistical analysis is two-way ANOVA using Tukey's HSD $p<0.001$ where each community was compared within pH levels (i.e. only communities with the same starting pH were compared).

References

1. Lauber, C. L., Hamady, M., Knight, R. & Fierer, N. Pyrosequencing-Based Assessment of Soil pH as a Predictor of Soil Bacterial Community Structure at the Continental Scale. *Appl. Environ. Microbiol.* **75**, 5111–5120 (2009).

2. Temudo, M. F., Muyzer, G., Kleerebezem, R. & van Loosdrecht, M. C. M. Diversity of microbial communities in open mixed culture fermentations: impact of the pH and carbon source. *Appl. Microbiol. Biotechnol.* **80**, 1121–1130 (2008).

3. Albert, L. S. & Brown, D. G. Variation in bacterial ATP concentration during rapid changes in extracellular pH and implications for the activity of attached bacteria. *Colloids Surf. B Biointerfaces* **132**, 111–116 (2015).

4. Murali, N., Srinivas, K. & Ahring, B. K. Biochemical Production and Separation of Carboxylic Acids for Biorefinery Applications. *Fermentation* **3**, 22 (2017).

5. Garcia-Aguirre, J., Aymerich, E., González-Mtnez. de Goñi, J. & Esteban-Gutiérrez, M. Selective VFA production potential from organic waste streams: Assessing temperature and pH influence. *Bioresour. Technol.* **244**, 1081–1088 (2017).

6. Wang, M., Sun, X. Z., Janssen, P. H., Tang, S. X. & Tan, Z. L. Responses of methane production and fermentation pathways to the increased dissolved hydrogen concentration generated by eight substrates in *in vitro* ruminal cultures. *Anim. Feed Sci. Technol.* **194**, 1–11 (2014).

7. Ratzke, C., Denk, J. & Gore, J. Ecological suicide in microbes. *Nat. Ecol. Evol.* **2**, 867–872 (2018).

8. Wang, Y. *et al.* pH-dependent medium-chain fatty acid synthesis in waste activated sludge fermentation: Metabolic pathway regulation. *J. Environ. Manage.* **373**, 123722 (2024).

9. Wu, L. *et al.* Medium chain fatty acids production from anaerobic fermentation of food wastes: The role of fermentation pH in metabolic pathways. *Chem. Eng. J.* **472**, 144824 (2023).

10. Candry, P. *et al.* Mildly acidic pH selects for chain elongation to caproic acid over alternative pathways during lactic acid fermentation. *Water Res.* **186**, 116396 (2020).

11. Esquivel-Elizondo, S. *et al.* The Isolate *Caproiciproducens* sp. 7D4C2 Produces n-Caproate at Mildly Acidic Conditions From Hexoses: Genome and rBOX Comparison With Related Strains and Chain-Elongating Bacteria. *Front. Microbiol.* **11**, (2021).

12. Chen, L. *et al. Megasphaera elsdenii* Lactate Degradation Pattern Shifts in Rumen Acidosis Models. *Front. Microbiol.* **10**, (2019).

13. Shrestha, S., Colcord, B., Fonoll, X. & Raskin, L. Fate of influent microbial populations during medium chain carboxylic acid recovery from brewery and pre-fermented food waste streams. *Environ. Sci. Water Res. Technol.* **8**, 257–269 (2022).

14. Angenent, L. T. & Wrenn, B. A. Optimizing Mixed-Culture Bioprocessing To Convert Wastes into Bioenergy. in *Bioenergy* 179–194 (John Wiley & Sons, Ltd, 2008). doi:10.1128/9781555815547.ch15.

15. Kenealy, W. R. & Waselefsky, D. M. Studies on the substrate range of *Clostridium kluyveri*; the use of propanol and succinate. *Arch. Microbiol.* **141**, 187–194 (1985).

16. Pflügl, S., Marx, H., Mattanovich, D. & Sauer, M. Heading for an economic industrial upgrading of crude glycerol from biodiesel production to 1,3-propanediol by *Lactobacillus diolivorans*. *Bioresour. Technol.* **152**, 499–504 (2014).

17. Sato, T. *et al.* Isolation of lactate-utilizing butyrate-producing bacteria from human feces and in vivo administration of *Anaerostipes caccae* strain L2 and galacto-oligosaccharides in a rat model. *FEMS Microbiol. Ecol.* **66**, 528–536 (2008).

18. Liu, Y. *et al.* Shaping human gut community assembly and butyrate production by controlling the arginine dihydrolase pathway. 2023.01.10.523442 Preprint at https://doi.org/10.1101/2023.01.10.523442 (2024).

19. Candry, P. *et al.* A novel high-throughput method for kinetic characterisation of anaerobic bioproduction strains, applied to *Clostridium kluyveri*. *Sci. Rep.* **8**, 9724 (2018).

20. Candry, P., Huang, S., Carvajal-Arroyo, J. M., Rabaey, K. & Ganigue, R. Enrichment and characterisation of ethanol chain elongating communities from natural and engineered environments. *Sci. Rep.* **10**, 3682 (2020).

21. Luan, L. *et al.* Integrating pH into the metabolic theory of ecology to predict bacterial diversity in soil. *Proc. Natl. Acad. Sci.* **120**, e2207832120 (2023).

22. Mougi, A. pH Adaptation stabilizes bacterial communities. *Npj Biodivers.* **3**, 1–7 (2024).

23. Ramoneda, J. *et al.* Building a genome-based understanding of bacterial pH preferences. *Sci. Adv.* **9**, eadf8998 (2023).

24. Ratzke, C. & Gore, J. Modifying and reacting to the environmental pH can drive bacterial interactions. *PLOS Biol.* **16**, e2004248 (2018).

Chapter 4: Summary and Future Directions

Contributions: Madeline Hayes wrote the text and generated the figures.

Summary: On the efficacy of synthetic biology approaches for exploration of MCFA-producing communities

In this dissertation, I aimed to apply a previously developed, two-stage modeling approach[1] to a novel system with two key goals: demonstrate the utility of this approach in a system more complex than its development (and improve the rigor), and identify communities which produce abundant medium chain fatty acids (MCFAs) for further study. By combining the high-throughput experimental setup which can assemble hundreds of synthetic communities with computational models designed to survey microbial data to extract biological information, we aimed to cover a portion of the potential sample space and efficiently identify communities demonstrating interesting phenotypes. The synthetic biology approaches applied here are, to our knowledge, the first large synthetic community designed to approximate a chain elongating microbiome.[2]

In Chapter 2, I demonstrated that the gLV+LR models utilized in a design-test-learn iterative data collection approach identifies successively increasing MCFA-producing communities in two media, and that by the fourth cycle, those communities produce higher C6 than a random selection. This demonstrates effective exploration of the sample space. The modeling performance, as assessed by multiple metrics, is sufficiently high to build confidence in the interpretation of parameters for real biological information. The models identify Ck as the key MCFA producer in this system, consistent with the literature, and lactobacilli as key functional species that impact the growth of multiple other species as well as MCFA production. Parameters also hint at possible roles for Bh and *E. lenta* as positively impacting Ck. We hypothesized that the underlying biology driving the model-identified interactions would become apparent with more rigorous testing. Because pH is known to be a key factor in the favorability

of chain elongation,[3] we aimed to move forward with a subset of communities and focus on the impact of pH.

In Chapter 3, I narrowed the focus to less than 100 communities: two elite panels from each synthetic media type containing a diverse range of community sizes were examined specifically for their response to changes in pH. Both panels consistently show that lactobacilli are not associated with communities which are resilient to changes in pH, and specifically in SCRv1, Ld showed a strong association with pH sensitivity. I proceeded to narrow even further to a single SCRv1 community which contained 10 species – all four MCFA producers, two lactobacilli including Ld, Bh, and *E. lenta* -- to test previously generated hypotheses and begin to understand its pH sensitive profile. I demonstrated that Ld is responsible for pH sensitivity in this community, and identify lactate as the probable mediator of this effect. I also show that E. lenta has a small but detectable effect on pH resiliency, offering some protection against pH-induced abolishment of hexanoate production. Interestingly, Bh did not show any impact on Ck or hexanoate production in this context, and how this parameter value may manifest is still a mystery.

In this way, the main goals of this work were achieved. There is a wide variety of communities still available for investigation, including ideal candidates for broader metabolomic studies, gene expression analysis, stability analysis, and scale up.

Future Directions

During this work there were many experimental options that we did not have the resources to pursue, as we were severely limited by the throughput of our metabolomics facilities. Here I present a few options for future work.

Broaden metabolic panel for existing communities

One key goal of this body of work was to identify communities which differed in their composition and output in order to provide a basis for comparative studies. In doing so, we now have the opportunity to expand the knowledge of the existing communities. For example, the SCRv2 communities which showed resilience to changes in pH did not improve in hexanoate production when the pH was increased. Very few of these communities contained lactobacilli, and the lack of hexanoate production at either pH level could be explained by other species-species interactions. Expanding the metabolomic dataset for these communities to include the media component and intermediates panel could reveal whether inefficient substrate consumption or some other factor is prohibiting strong production by these communities.

Similarly, a broader look at the SCRv1 panel could elucidate unique mechanisms at play, as there are some sensitive communities which do not contain Ld, and some high producing communities which do not contain Ck. The fact that there are communities with such divergent compositions which still show similar profiles strongly implies more than one mechanism of MCFA production, and investigating how these communities differ from the bulk of the dataset could reveal more community interactions.

Control of the media environment for elucidating metabolite production and consumption by cross-feeding species

In this work, we were able to closely control the composition of the communities in order to map the contribution of individual species to their function. By similarly controlling the composition of the media, we could map how individual components impact species within the community, and the larger rippling effects towards function. Two obvious targets stand out: xylose and succinate. Xylose is the primary target of degradation for the lactobacilli, and its

breakdown products contribute to the intermediate metabolite pool. By removing or increasing this component, we could begin to track the flow of carbon through the system, and understand any bottlenecks preventing full substrate utilization. Succinate is a key carbon source for Ck, and understanding whether it is necessary for Ck functionality would identify a key control knob for promoting Ck in the community context.

In addition to controlling the amount of any given media component, differentiating how carbon is transformed from one component to the next is a key step in understanding where metabolisms overlap and diverge within the cell. Many of the media components have 'heavy' counterparts, and carbon incorporated into downstream metabolites can be readily differentiated by the metabolomics methods employed here. By replacing components one at a time, the proportion of carbon incorporated from different sources could be revealed, contributing to the specific stoichiometric landscape.

Lastly, we detected ethanol in the media and traced its source to fumes from decontamination of surfaces during experimental setup. While the concentrations appeared relatively consistent, and our internal benchmarks indicate consistent conditions from day to day, more tightly controlling ethanol is a key step in wrangling all the interfacing conditions in this system. Ck utilizes ethanol for fermentation,[4] and the presence of this metabolite is certainly contributing to the observed hexanoate production. Control of this substrate would likely improve the consistency of results, and allow for controlled exploration of optimizing this parameter.

Understanding finer resolution of time point data

Given that we only harvested cells and supernatant at a single time point, the first 48 hours and any following community growth and function remain unknown. It is extremely likely

that the maximum output of Ck in this community is not being captured; some sources indicate that in large volume batch culture, it could take up to 160 hours[5] for this very slow growing species to fully leave lag phase. Simply collecting the existing panel of communities at a slightly later time point could allow for a better modeling approach by capturing a more productive time point.

Additionally, the story of dynamic pH as a player in MCFA production favorability requires confirmation that pH changes predictably in response to acid addition. Understanding how the production of different acids – including butyrate – impact pH over time, Expanding the view of the system to incorporate multiple time points could help identify when production and consumption of acidic intermediates changes pH. When paired with finer resolution of MCFA data, this could lend a much more detailed mechanistic insight into when reverse beta oxidation becomes active, and the metabolite switching that might precede output.

Application of diverse models to the dataset

The gLV+LR approach was a qualified success, but the potential for even further optimization should not be ignored. In order to visualize this possibility, I used MDS with phylogeny to display the model predictions for the next 50 communities I would test (‘cycle 5’) in relation to the existing dataset (Figure 1). These were the top-ranked C6 predictions from the final round of predicting all untested communities. Figure 1 shows that in SCRv1, the next cycle moves into as yet untested sample space, and the predicted C6 values continue to improve. As noted in Chapter 2, in SCRv2, overlap between cycles 3 and 4 in the maximized predictions indicate that there may not be further optimization in this media. The cycle 5 communities continue to overlap with existing samples, confirming this observation. Testing these

communities is a much smaller experiment than any individual DTL cycle, and represents an attainable next step in experimentation.

In concert with this, the possibility of applying diverse model architectures could allow for more flexibility in the assumptions about system functionality. Currently, the model strictly assumes that OD and community output are linearly linked; sources indicate that this is not necessarily true under all conditions for Ck, and allowing for more flexibility for identifying key parameters may help to capture that. For example, neural networks[6] and random forest models[7] are both easy to implement and shown to be discerning of community functions. By comparing the predictions from these models to the gLV+LR cycle 5 predictions, we could begin to understand if other models are picking up on different interactions that drive community function. Comparatively testing these could be a straightforward computational paper.

Scale up: define and control headspace composition and pH using benchtop bioreactors

To bring all the above methods together, I would aim for a scaling up of the best communities. Apply the different models to ascertain if maximum community productivity has been reached, perform one more high-throughput cycle to compare how accurately the models predicted community function, expand the metabolomics panels, and control all media components and pH in batch and continuous runs to understand the temporal dynamics of the system. From there, comparing gene expression of key pathways under conditions where variable outputs are observed would provide an excellent mechanistic insight into how we can control MCFA production in communities. In reality, this was an original proposed aim of this project, and effectively reaching a point where the targets are within reach is an indicator that the approach we have employed is effective for exploring microbiomes.

Lastly, the opportunity to take all the data from the synthetic system and compare the performance of our defined communities to their performance in genuine conversion residue could be an excellent opportunity to start to parse the translatability of these results. Our first attempt at moving synthetic communities into conversion residue yielded some unexpected results (see Appendix I), with communities underperforming relative to our expectations. The uncharacterized components of conversion residue are a huge challenge for successfully designing solutions that are consistent and reliable, and understanding how engineered communities respond to the challenges of conversion residue would offer valuable insight into the road to implementing this as a genuinely effective solution.

Figures

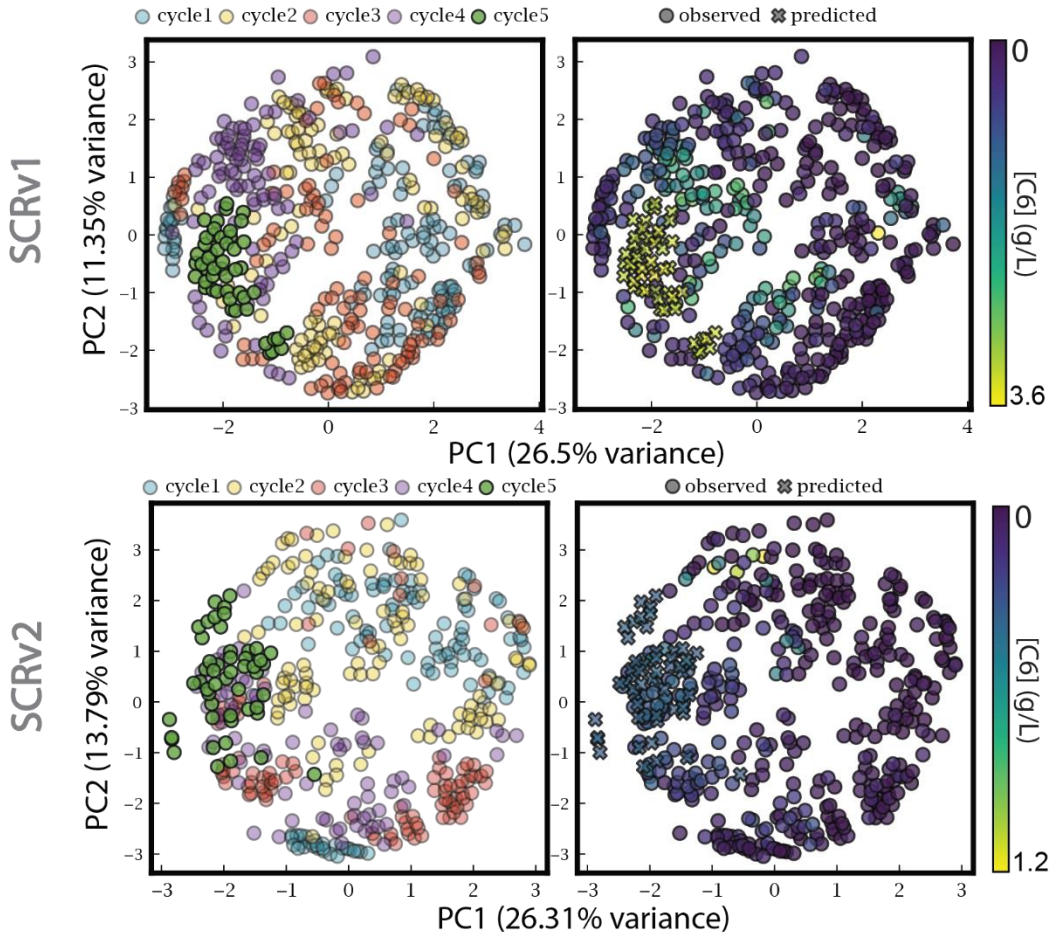

Figure 1: Hypothetical next round of data collection following gLV+LR predictions

Using the parameter sets from the last round of model training (fit on the whole dataset), a final round of predictions on the untested community space was made for both media (SCRv1, top, and SCRv2, bottom). These 'cycle 5' (green) shown in relation to the previous cycles (left column, where cycle 1 = blue, cycle 2 = yellow, cycle 3 = red, and cycle 4 = purple), as well as the rest of the dataset C6 production (right column, concentrations indicated with color bar). Previously observed points are circles, new predictions are X's.

References

1. Clark, R. L. *et al.* Design of synthetic human gut microbiome assembly and butyrate production. *Nat. Commun.* **12**, 3254 (2021).

2. Scarborough, M. J., Lawson, C. E., Hamilton, J. J., Donohue, T. J. & Noguera, D. R. Metatranscriptomic and Thermodynamic Insights into Medium-Chain Fatty Acid Production Using an Anaerobic Microbiome. *mSystems* **3**, (2018).

3. Allaart, M. T., Stouten, G. R., Sousa, D. Z. & Kleerebezem, R. Product Inhibition and pH Affect Stoichiometry and Kinetics of Chain Elongating Microbial Communities in Sequencing Batch Bioreactors. *Front. Bioeng. Biotechnol.* **0**, (2021).

4. Yin, Y., Zhang, Y., Karakashev, D. B., Wang, J. & Angelidaki, I. Biological caproate production by *Clostridium kluyveri* from ethanol and acetate as carbon sources. *Bioresour. Technol.* **241**, 638–644 (2017).

5. Fernández-Blanco, C., Veiga, M. C. & Kennes, C. Carbon dioxide as key player in chain elongation and growth of *Clostridium kluyveri*: Insights from batch and bioreactor studies. *Bioresour. Technol.* **394**, 130192 (2024).

6. Thompson, J. C., Zavala, V. M. & Venturelli, O. S. Integrating a tailored recurrent neural network with Bayesian experimental design to optimize microbial community functions. *PLOS Comput. Biol.* **19**, e1011436 (2023).

7. Hernández Medina, R. *et al.* Machine learning and deep learning applications in microbiome research. *ISME Commun.* **2**, 98 (2022).

Appendix I: Community performance in conversion residue at multiple pH levels

Contributions: Madeline Hayes designed and executed the experiments, wrote the text, generated the figures

Summary

Optimization of community performance in synthetic media equivalents was a necessary limitation for this project approach. DTL cycles must keep media environment between cycles or risk not being able to adequately quantify species-species interactions as examples are added to the dataset. Testing the performance of optimized communities in complex conversion residue (CR) to assess how generalizable the community function is in different contexts is an important test to understand whether synthetic communities will be robust to batch to batch variation that comes with substrates from large-scale processes. To this end, we tested the performance of 22 communities from each dataset in conversion residue. We reasoned that the most pH-resilient and pH-sensitive communities from each media type would provide a diverse panel of performance for testing, and provided these communities with a range of pH values (5.5, 6.0, 6.5, 7.0) for growth and MCFA production. Ideally, we would see robust production with regards to pH changes that aligned with the previous observations of community sensitivity. In completing these experiments in genuine CR, we aimed to assess the translatability of this work for application.

A comparison of the media compositions shows that the batch of CR resembles SCRv1 more closely than SCRv2 (Figure A1). Measurements of CR components were provided by Great Lakes Bioenergy Research Center as part of their standard QC pipeline for orders. For this reason we expected the SCRv1 communities to perform better than the SCRv2 communities, with the caveat that since SCRv2 communities were optimized at a lower pH, they may be superior in the lower pH range tested.

We found that across the board, low complexity communities that contained Ck performed the best, but only at pH 6.0 or above (Figure A2). We also found that the more

complex SCRv2 communities outperformed the SCRv1 communities, against our expectations, and improving on their own performance in SCR. The biggest difference in community growth was that the lactobacilli occupied a much higher fraction of the community in CR, especially at low pH. While it is not surprising that low pH supported the growth of the lactobacilli, the lower concentration of xylose in this media would not have led me to hypothesize that they would be highly successful in this environment, no matter the pH, especially when there are other species competing for the limited acetate. These surprising result underscores the variability of the optimization process.

Methods

Samples were collected and processed as in previous chapters with the following modifications:

Preparation of CR: CR is an opaque media containing abundant solid particulate matter. The particulates were allowed to settle and the media decanted into separate bottles. Each bottle was transferred to the anaerobic chamber to equilibrate then pH'd once anaerobic.

Inoculation of species: Each species was precultured as previously described, but resuspended in PBS at OD=1.0, then inoculated at 1:10 volume in the experimental plate. This was to minimize the dilution of the media.

During cell harvest: cell pellets were rinsed thoroughly by resuspending and repelleting with PBS three times to remove any residual fibers or matter likely to clog the DNA column filter plate. Rinsing pellets greatly improves the filter time and yield for gDNA extraction.

During abundance quantification: OD is not able to be read in this media, and only relative abundance of each species is reported.

Figures

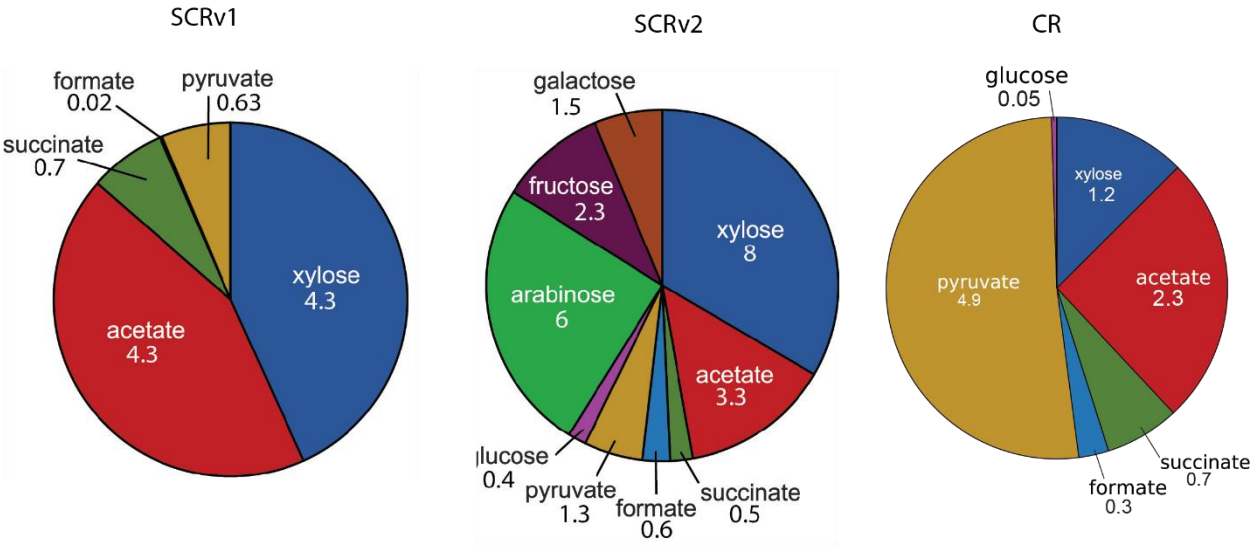

Figure 1: Media composition comparison

SCRv1 (left) and SCRv2 (center) were prepared as usual. CR (right) was ordered in three, 1-L batches to provide enough media for experiments, then combined to make one large batch. Measurements provided by GLBRC metabolomics QC for the three batches were averaged for the final concentrations.

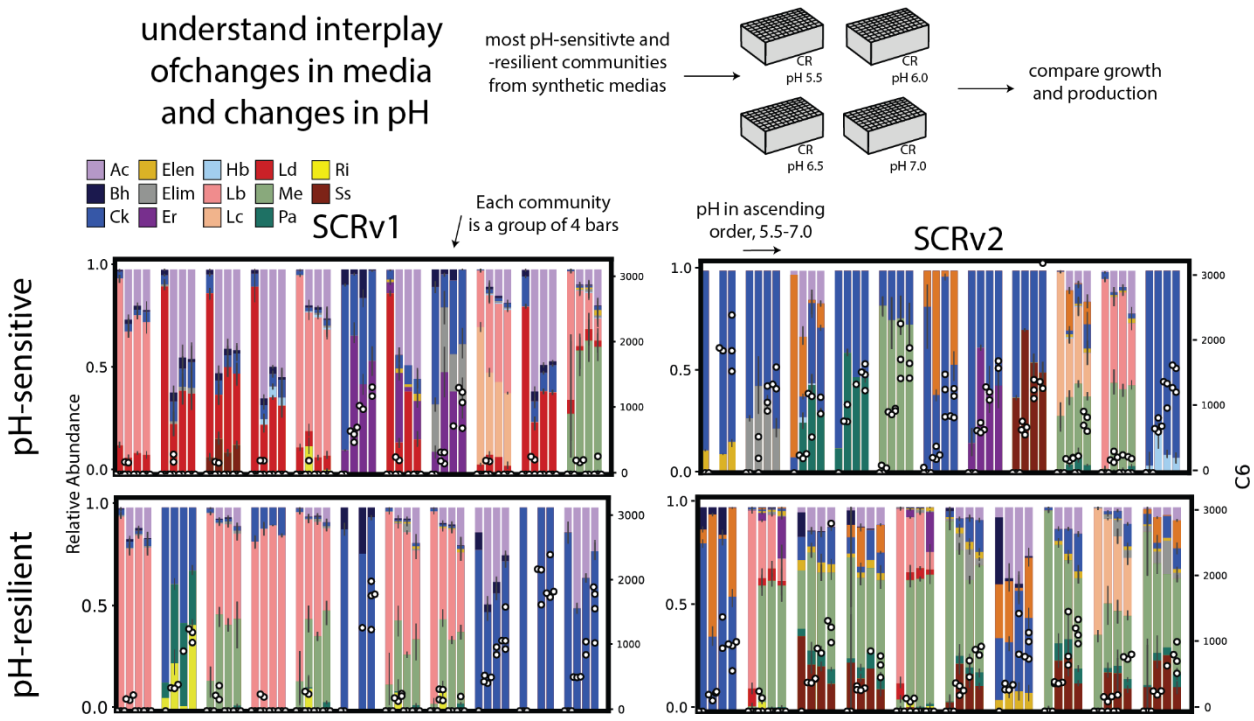

Figure 2: Community growth and MCFA production

Top: Experimental design. In order to understand the impact of media and pH changes, the 11 most and least pH sensitive communities as found in Chapter 3 were tested in CR at four different pH levels.

Bottom: Relative abundance (left y axis) and C6 production (right y axis) for pH sensitive (top row) and pH resilient (bottom row) communities optimized in SCRv1 (left column) and SCRv2 (right column). Each group of bars represents one community (n=2 biological replicates per pH level). Growth at each pH in ascending order, left to right for each group.