THE ECOLOGY OF MICROBIAL METABOLIC PATHWAYS

by

Stilianos Louca

B.Sc., Mathematics, Friedrich Schiller University of Jena, Germany, 2010
Diplom, Physics, Friedrich Schiller University of Jena, Germany, 2012

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Applied Mathematics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

September 2016

© Stilianos Louca, 2016
Abstract

Microbial metabolic activity drives biogeochemical cycling in virtually every ecosystem. Yet, microbial ecology and its role in ecosystem biochemistry remain poorly understood, partly because the enormous diversity found in microbial communities hinders their modeling. Despite this diversity, the bulk of global biogeochemical fluxes is driven by a few metabolic pathways encoded by a small set of genes, which through time have spread across microbial clades that can replace each other within metabolic niches. Hence, the question arises whether the dynamics of these pathways can be modeled regardless of the hosting organisms, for example based on environmental conditions. Such a pathway-centric paradigm would greatly simplify the modeling of microbial processes at ecosystem scales.

Here I investigate the applicability of a pathway-centric paradigm for microbial ecology. By examining microbial communities in replicate “miniature” aquatic environments, I show that similar ecosystems can exhibit similar metabolic functional community structure, despite highly variable taxonomic composition within individual functional groups. Further, using data from a recent ocean survey I show that environmental conditions strongly explain the distribution of microbial metabolic functional groups across the world’s oceans, but only poorly explain the taxonomic composition within individual functional groups. Using statistical tools and mathematical models I conclude that biotic interactions, such as competition and predation, likely underlie much of the taxonomic variation within functional groups observed in the aforementioned studies. The above findings strongly support a pathway-centric paradigm, in which the distribution and activity of microbial metabolic pathways is strongly determined by energetic and stoichiometric constraints, whereas additional mechanisms shape the taxonomic composition within metabolic guilds.

These findings motivated me to explore concrete pathway-centric mathematical models for specific ecosystems. Notably, I constructed a biogeochemical model for Saanich Inlet, a seasonally anoxic fjord with biogeochemistry analogous to oxygen minimum zones. The model describes the dynamics of individual microbial metabolic pathways involved in carbon, nitrogen and sulfur cycling, and largely explains geochemical depth profiles as well as DNA, mRNA and protein sequence data. This work yields insight into ocean biogeochemistry and demonstrates the potential of pathway-centric models for microbial ecology.
Preface

A number of sections in this work are partly or wholly published, in press or in review. Copyright licenses to all works were obtained and are listed where appropriate.

• Chapter 2: Stilianos Louca conceived the project. Stilianos Louca performed the sequence analysis and statistical analysis with input from Laura W. Parfrey and Michael Doebeli. Stilianos Louca wrote a first draft of the manuscript. All coauthors contributed to the final preparation of the manuscript. Michael Doebeli and Laura Parfrey supervised the project.

A version of this chapter is published in Science as a copyrighted article. Copyright (2016), American Association for the Advancement of Science. Reprinted with permission from AAAS, from:


• Chapter 3: Stilianos Louca, Vinicius F. Farjalla, Saulo M. S. Jacques, Aliny P. F. Pires, Juliana S. Leal performed the field work. Vinicius F. Farjalla and Saulo M. S. Jacques performed the chemical measurements in the laboratory. Stilianos Louca performed the molecular work in the laboratory, the DNA sequence analysis and the statistical analyses. Stilianos Louca, Michael Doebeli, Vinicius F. Farjalla, Diane S. Srivastava, and Laura W. Parfrey interpreted the statistical findings. Stilianos Louca wrote a first draft of the manuscript, and all authors contributed to the final preparation of the manuscript. Michael Doebeli and Vinicius F. Farjalla supervised the project.

A version of this chapter is under peer review for publication:


• Chapter 4: Stilianos Louca conceived and developed MCM. Stilianos Louca designed the E. coli example and performed the simulations. S.D. and Michael Doebeli analyzed the simulation results. Stilianos Louca wrote the manuscript, with editorial support from Michael Doebeli. Michael Doebeli supervised the project.

Software developed in this chapter is available under the GNU Lesser General Public License (http://www.gnu.org/licenses/lgpl.html) at: http://www.zoology.ubc.ca/MCM

A version of this chapter is published in eLife under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0):

• Chapter 5: Stilianos Louca conceived the project, ran the simulations and performed the statistical analysis. Stilianos Louca wrote the manuscript, with editorial support from Michael Doebeli. Michael Doebeli supervised the project.

A version of this chapter is published in Environmental Microbiology as a copyrighted article. Copyright (2015) Society for Applied Microbiology and John Wiley & Sons Ltd. Reprinted, with permission, from:


• Chapter 6: Stilianos Louca conceived the project, ran the simulations and performed the statistical analysis. Stilianos Louca wrote the manuscript, with editorial support from Michael Doebeli. Michael Doebeli supervised the project.

A version of this chapter is under peer review for publication:


• Chapter 7: Sean A. Crowe and Steven J. Hallam had the idea for the research. Stilianos Louca constructed the mathematical models with input from Sergei Katsev, and Stilianos Louca performed the simulations. Stilianos Louca, Sergei Katsev, Alyse Hawley, Sean A. Crowe and Steven J. Hallam analyzed the model predictions. Alyse Hawley, Maya P. Bhatia, Monica Torres-Beltran and Steven J. Hallam performed the sequencing. Alyse Hawley, Maya P. Bhatia, Stilianos Louca and Steven J. Hallam analyzed the sequence data. Monica Torres-Beltran, Alyse Hawley, Celine Michiels, Dave Capelle, Sergei Katsev, Gaute Lavik, Sean A. Crowe and Steven J. Hallam collected the chemical and physical data. Stilianos Louca wrote the manuscript with Sean A. Crowe and Steven J. Hallam. All authors contributed to the final preparation of the manuscript. Sergei Katsev, Sean A. Crowe, Michael Doebeli and Steven J. Hallam supervised the project.

A version of this chapter is published in PNAS as a copyrighted article. Copyright (2016) National Academy of Sciences. Reprinted, with permission, from:


• Chapter 8: Stilianos Louca conceived the project, ran the simulations and performed the statistical analysis. Stilianos Louca wrote the manuscript, with editorial support from Michael Doebeli. Michael Doebeli supervised the project.

A version of this chapter is published in Ecological Modelling as a copyrighted article. Copyright (2016) Elsevier. Reprinted, with permission, from:

Throughout this dissertation the word “we” refers to Stilianos Louca unless otherwise stated. None of the work encompassing this dissertation required consultation with the UBC Research Ethics Board.
# Table of Contents

Abstract ................................................................. ii
Preface ................................................................ iii
Table of Contents ........................................................ vi
List of Tables ............................................................... xi
List of Figures .............................................................. xii
List of Abbreviations ...................................................... xv
Acknowledgements ......................................................... xvi

1 Opening chapter ........................................................... 1
   1.1 Introduction ......................................................... 1
   1.2 Problem statement .................................................. 3
   1.3 Objectives and overview of this dissertation ...................... 6

2 The decoupling of function and taxonomy in the global ocean microbiome 8
   2.1 Synopsis ............................................................... 8
   2.2 Introduction .......................................................... 8
   2.3 Environmental conditions mainly affect microbial function ...... 10
   2.4 Causes of variation within functional groups .................... 11
   2.5 Beyond taxonomic profiling ....................................... 13
   2.6 Conclusions .......................................................... 15

3 Functional stability despite high taxonomic variability across microbial communities in bromeliad tanks 21
   3.1 Synopsis ............................................................... 21
   3.2 Introduction .......................................................... 22
   3.3 Results and discussion .............................................. 23
      3.3.1 Functional stability contrasts with taxonomic variability .... 23
      3.3.2 Causes of variation within functional groups ................ 24
   3.4 Conclusions .......................................................... 27

4 Calibration and analysis of cell-metabolic models for microbial ecology 32
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Synopsis</td>
<td>32</td>
</tr>
<tr>
<td>4.2 Introduction</td>
<td>32</td>
</tr>
<tr>
<td>4.3 Model overview</td>
<td>34</td>
</tr>
<tr>
<td>4.4 Results and discussion</td>
<td>36</td>
</tr>
<tr>
<td>4.4.1 Successional dynamics of a microbial community</td>
<td>36</td>
</tr>
<tr>
<td>4.4.2 Experimental calibration</td>
<td>37</td>
</tr>
<tr>
<td>4.4.3 Predicting microbial community dynamics</td>
<td>38</td>
</tr>
<tr>
<td>4.5 Conclusions</td>
<td>40</td>
</tr>
<tr>
<td>5 Transient dynamics of competitive exclusion in microbial communities</td>
<td>48</td>
</tr>
<tr>
<td>5.1 Synopsis</td>
<td>48</td>
</tr>
<tr>
<td>5.2 Introduction</td>
<td>48</td>
</tr>
<tr>
<td>5.3 Results and discussion</td>
<td>49</td>
</tr>
<tr>
<td>5.3.1 Competition for a common resource</td>
<td>49</td>
</tr>
<tr>
<td>5.3.2 Bioreactors as model systems</td>
<td>52</td>
</tr>
<tr>
<td>5.3.3 Bioreactor community dynamics</td>
<td>53</td>
</tr>
<tr>
<td>5.3.4 Variable does not mean unstable</td>
<td>55</td>
</tr>
<tr>
<td>5.3.5 Model limitations</td>
<td>56</td>
</tr>
<tr>
<td>5.3.6 Alternative destabilizing factors</td>
<td>56</td>
</tr>
<tr>
<td>5.3.7 Towards a pathway-centric microbial ecology</td>
<td>57</td>
</tr>
<tr>
<td>5.4 Conclusions</td>
<td>58</td>
</tr>
<tr>
<td>6 Taxonomic variability and functional stability in microbial communities infected by phages</td>
<td>63</td>
</tr>
<tr>
<td>6.1 Synopsis</td>
<td>63</td>
</tr>
<tr>
<td>6.2 Introduction</td>
<td>64</td>
</tr>
<tr>
<td>6.3 Results and discussion</td>
<td>66</td>
</tr>
<tr>
<td>6.3.1 Bioreactor dynamics in the absence of phages</td>
<td>66</td>
</tr>
<tr>
<td>6.3.2 Effects of phages on community dynamics</td>
<td>67</td>
</tr>
<tr>
<td>6.3.3 Functional redundancy promotes functional stability</td>
<td>68</td>
</tr>
<tr>
<td>6.3.4 Statistical averaging or dynamic stabilization?</td>
<td>71</td>
</tr>
<tr>
<td>6.3.5 Phages promote prokaryotic diversity</td>
<td>72</td>
</tr>
<tr>
<td>6.4 Conclusions</td>
<td>73</td>
</tr>
<tr>
<td>7 Gene-centric modeling of the Saanich Inlet oxygen minimum zone</td>
<td>78</td>
</tr>
<tr>
<td>7.1 Synopsis</td>
<td>78</td>
</tr>
<tr>
<td>7.2 Introduction</td>
<td>79</td>
</tr>
<tr>
<td>7.3 Construction and calibration of a gene-centric model</td>
<td>80</td>
</tr>
<tr>
<td>7.4 Results and Discussion</td>
<td>81</td>
</tr>
<tr>
<td>7.4.1 DNA profiles and process rates</td>
<td>81</td>
</tr>
<tr>
<td>7.4.2 mRNA and protein profiles</td>
<td>84</td>
</tr>
<tr>
<td>7.4.3 Consequences for geobiology</td>
<td>86</td>
</tr>
<tr>
<td>7.5 Conclusions</td>
<td>88</td>
</tr>
<tr>
<td>8 Reaction-centric modeling of microbial community metabolism</td>
<td>92</td>
</tr>
</tbody>
</table>
8.1 Synopsis .................................................. 92
8.2 Introduction ............................................. 93
8.3 Derivation of the reaction-centric framework ............................................ 95
  8.3.1 One reaction per cell ........................................ 95
  8.3.2 Multiple reactions per cell .................................... 97
8.4 Demonstration of the reaction-centric framework .......................................... 99
  8.4.1 Overview .................................................. 99
  8.4.2 Example 1: Urea hydrolysis and nitrification in a batch-fed incubator .......... 99
  8.4.3 Example 2: Nitrification in a flow-through bioreactor .......................... 102
  8.4.4 Estimating concentrations of other organic compounds ...................... 106
  8.4.5 Limitations and extensions of reaction-centric models ...................... 107
8.5 Conclusions .............................................. 108

9 Closing chapter ............................................... 116
  9.1 Conclusions ............................................... 116
  9.2 So, what is life? ........................................... 118

Summary of software contributions ....................................................... 121

Bibliography .................................................... 122

Appendices

A Chapter 2: Supplemental material ...................................................... 173
  A.1 Methods ................................................... 173
    A.1.1 Sequencing data ........................................... 173
    A.1.2 Functional annotation of prokaryotic taxa .......................... 173
    A.1.3 Statistical Analysis ....................................... 174
  A.2 Resolving ambiguities in gene-centric metagenomics ............................ 178
  A.3 Comparison with Sunagawa et al. (2015) ....................................... 179

B Chapter 3: Supplemental material ...................................................... 202
  B.1 Methods ................................................... 202
    B.1.1 Biological sample collection .................................. 202
    B.1.2 Chemical analysis of tank water .................................. 202
    B.1.3 Measurement of other physicochemical variables ...................... 204
    B.1.4 16S sequencing .......................................... 205
    B.1.5 Functional annotation of prokaryotic taxa (FAPROTAX) ............... 206
    B.1.6 Metagenomic sequencing .................................... 207
    B.1.7 Comparing functional and taxonomic variability ...................... 208
    B.1.8 Metric multidimensional scaling and coloring ....................... 209
    B.1.9 Phylogenetic community structure ................................ 210
    B.1.10 Comparing OTU proportions to environmental variables ............ 211
    B.1.11 Comparing community dissimilarities to geographical distances ...... 212
    B.1.12 Comparing OTU co-occurrences to a null model .................... 212
    B.1.13 Sequence data availability .................................. 213
### Table of Contents

**C Chapter 4: Supplemental material** .......................................................... 235
C.1 Methods ........................................................................................................ 235
C.1.1 MCM overview .......................................................................................... 235
C.1.2 Calibration of *E. coli* cell models ............................................................ 239
C.1.3 Simulation of the microbial community model .......................................... 240
C.1.4 Robustness of the SS-FS coexistence ....................................................... 240
C.1.5 Seasonal restriction of the SS-FS co-cultures ............................................. 240

**D Chapter 5: Supplemental material** ............................................................... 244
D.1 Methods ........................................................................................................ 244
D.1.1 Computational framework ......................................................................... 244
D.1.2 Construction of the cell models .................................................................. 245
D.1.3 Calibration of the template cell models ..................................................... 245
D.1.4 Nitrifying membrane bioreactor model ..................................................... 246
D.1.5 Statistics of community convergence ......................................................... 247
D.2 Elaboration on the competition model .......................................................... 247
D.3 Details on the bioreactor model ...................................................................... 248
D.3.1 Construction of cell models ....................................................................... 248
D.3.2 Metabolites ................................................................................................. 249
D.3.3 Reaction network for AOB ......................................................................... 251
D.3.4 Reaction network for NOB ......................................................................... 253
D.3.5 Uptake kinetics ........................................................................................... 254
D.3.6 Community-scale dynamics ....................................................................... 255

**E Chapter 6: Supplemental material** ............................................................... 257
E.1 Methods ........................................................................................................ 257
E.1.1 Model overview ......................................................................................... 257
E.1.2 Reaction rates and metabolite dynamics .................................................... 257
E.1.3 Gibbs free energy and cell production ....................................................... 258
E.1.4 Cell and phage population dynamics ......................................................... 259
E.1.5 Parameterization and simulations .............................................................. 260
E.1.6 Statistical analysis ...................................................................................... 261
E.1.7 Deterministic vs stochastic competitive exclusion ..................................... 262

**F Chapter 7: Supplemental material** ............................................................... 269
F.1 Methods ........................................................................................................ 269
F.1.1 Model overview ......................................................................................... 269
F.1.2 Mathematical model structure .................................................................... 269
F.1.3 Considered pathways .................................................................................. 270
F.1.4 Model calibration and data ......................................................................... 272
F.1.5 mRNA and protein models ........................................................................ 272
F.1.6 Inverse linear transport modeling .............................................................. 273
F.2 Data acquisition .............................................................................................. 274
F.2.1 Sampling site and time ............................................................................... 274
F.2.2 Chemical and physical depth profiles ......................................................... 275
Table of Contents

F.2.3 Metagenomics, metatranscriptomics and metaproteomics ........................................ 276
F.2.4 Quantifying metagenomic and metatranscriptomic data using RPKM .......................... 277
F.2.5 Process rate measurements .................................................................................. 278
F.2.6 qPCR quantification of SUP05 cell counts ......................................................... 279
F.3 Mathematical model ................................................................................................. 280
  F.3.1 Overview ......................................................................................................... 280
  F.3.2 Considered pathways ....................................................................................... 282
  F.3.3 Pathway stoichiometry ..................................................................................... 283
  F.3.4 Reaction kinetics .............................................................................................. 284
  F.3.5 Gibbs free energy and gene growth ................................................................... 285
  F.3.6 Boundary conditions ....................................................................................... 286
  F.3.7 Model parameterization .................................................................................... 286
  F.3.8 Calibrating reaction-kinetic parameters to data .............................................. 290
  F.3.9 Calibrating multi-omic data units .................................................................... 291
  F.3.10 Predicting metatranscriptomic and metaproteomic profiles ............................. 292
  F.3.11 Calculating metabolic fluxes between pathways ............................................ 297
  F.3.12 Local sensitivity analysis ............................................................................... 298
F.4 Caveats and special notes ......................................................................................... 300
  F.4.1 The role of sulfate reduction .......................................................................... 300
  F.4.2 The role of DNRA .......................................................................................... 302
  F.4.3 The role of aerobic sulfide oxidation .............................................................. 302
  F.4.4 Planctomycetes and nxr .................................................................................. 303
F.5 Inverse linear transport modeling (ILTM) ............................................................... 303

G Chapter 8: Supplemental material ........................................................................... 307
  G.1 Methods ............................................................................................................. 307
    G.1.1 Details on example 1 (batch-fed incubator) ................................................... 307
    G.1.2 Details on example 2 (flow-through bioreactor) .......................................... 310
    G.1.3 Computational methods: Using MCM for reaction-centric models ............. 312
  G.2 Mathematical proofs ............................................................................................. 316
    G.2.1 On specific maintenance rates ..................................................................... 316
    G.2.2 On the concentration of organic components ............................................... 317
    G.2.3 Properties of the amplification matrix .......................................................... 319
List of Tables

8.1 Overview of symbols and units for cell-centric and reaction-centric models . 111
A.1 Overview of oceanographic variables . . . . . . . . . . . . . . . . . . . . . . 181
A.2 Overview of considered Tara oceans samples . . . . . . . . . . . . . . . . . . 196
A.3 KOG-function associations for the ocean microbiome . . . . . . . . . . . . 200
A.4 OTUs per functional group in the global ocean microbiome . . . . . . . . . 201
B.1 OTU overlap between bromeliad microbiomes . . . . . . . . . . . . . . . . . . 227
B.2 Coefficients of variation for gene abundances across bromeliad microbiomes 228
B.3 OTU co-occurrence patterns across bromeliad microbiomes. . . . . . . . . 229
B.4 Geographical distances vs taxonomic dissimilarities across bromeliad micro-
biomes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 229
B.5 Physicochemical environmental variables in bromeliads . . . . . . . . . . . . 230
B.6 OTUs per functional group in bromeliad microbiomes . . . . . . . . . . . . 231
B.7 KOG-function associations for bromeliad microbiomes . . . . . . . . . . . . 232
C.1 Fitted parameters for the individual E. coli cell models . . . . . . . . . . . . . 241
E.1 Pathway stoichiometry in a methanogenic bioreactor . . . . . . . . . . . . . 267
E.2 Model parameters for a methanogenic bioreactor . . . . . . . . . . . . . . . 268
F.1 KEGG-function associations for the global ocean microbiome . . . . . . . . 277
F.1 Boundary conditions for the gene-centric model . . . . . . . . . . . . . . . . . 287
F.2 Reaction-kinetic parameters used in the gene-centric model . . . . . . . . . 289
F.3 Rescaling factors for metagenomic profiles . . . . . . . . . . . . . . . . . . . 293
F.4 Proportionality factors for mRNA and protein models . . . . . . . . . . . . 295
G.1 Model parameters for a ureolytic bioreactor . . . . . . . . . . . . . . . . . . . 309
G.2 Model parameters for a nitrifying bioreactor . . . . . . . . . . . . . . . . . . . 311
List of Figures

2.1 Linking functional and taxonomic composition to environmental conditions. 16
2.2 Environmental filtering of microbial communities in the global ocean. 17
2.3 Functional redundancy in the global ocean microbiome. 18
2.4 Microbial community differences vs geographical distance. 19
2.5 Regression of the taxonomic composition of the global ocean microbiome. 20

3.1 Bromeliad species used. 28
3.2 Functional stability vs taxonomic variability of bromeliad microbiomes. 29
3.3 OTU-function associations. 30
3.4 Phylogenetic dispersion in bromeliad microbiomes. 30
3.5 Relating OTU proportions to environmental variables across bromeliads. 31

4.1 Conceptual framework used by MCM. 42
4.2 Overview of MCM functionalities. 43
4.3 Relative cell densities during evolution experiments with E. coli. 44
4.4 Calibration of E. coli cell models using monoculture experiments. 44
4.5 Dynamics of the E. coli microbial community model. 45
4.6 Metabolic differentiation of E. coli types. 46
4.7 Predicted relative cell densities in co-culture when restricted to only one season. 47

5.1 Calibration and validation of AOB and NOB models. 60
5.2 Simulations of the nitrifying bioreactor under constant conditions. 61
5.3 Simulations of the perturbed nitrifying bioreactor. 62

6.1 Modeling methanogenic communities. 75
6.2 Phage predation drives OTU turnover. 76
6.3 Effects of functional redundancy on methane production. 76
6.4 Effects of functional redundancy on functional community composition. 77

7.1 Metabolic network and selected chemical time series in Saanich Inlet. 89
7.2 Measured vs predicted chemical depth profiles in Saanich Inlet. 90
7.3 Molecular and reaction rate profiles in Saanich Inlet, compared to model predictions. 91

8.1 Modeling urea hydrolysis and nitrification. 110
8.2 Model predictions and data for a ureolytic bioreactor. 112
8.3 Reconstructing a bioreactor’s state using chemical time series. 113
8.4 Information needed to estimate the state of a reaction-centric model. 114
8.5 Model predictions and data for a nitrifying bioreactor. 115

9.1 Steady state in the Saanich Inlet OMZ. 118
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.2</td>
<td>Community dynamics in a methanogenic bioreactor, compared to turbulent thermal convection</td>
<td>119</td>
</tr>
<tr>
<td>A.1</td>
<td>Correlation analysis at various taxonomic levels</td>
<td>182</td>
</tr>
<tr>
<td>A.2</td>
<td>Environmental filtering at higher taxonomic levels</td>
<td>183</td>
</tr>
<tr>
<td>A.3</td>
<td>Functional vs taxonomic community profiles of the ocean microbiome</td>
<td>184</td>
</tr>
<tr>
<td>A.4</td>
<td>Functional dissimilarities vs geographical distances in the ocean microbiome</td>
<td>185</td>
</tr>
<tr>
<td>A.5</td>
<td>Phenotype-based vs metagenomic functional profiles</td>
<td>186</td>
</tr>
<tr>
<td>A.6</td>
<td>Correlations between functional groups in the global ocean microbiome</td>
<td>187</td>
</tr>
<tr>
<td>A.7</td>
<td>Taxonomic compositions within functional groups in the ocean microbiome</td>
<td>188</td>
</tr>
<tr>
<td>A.8</td>
<td>Community dissimilarities vs geographical distances in the surface and DCM</td>
<td>189</td>
</tr>
<tr>
<td>A.9</td>
<td>Sampling locations of the Tara oceans survey</td>
<td>190</td>
</tr>
<tr>
<td>A.10</td>
<td>Correlations between environmental variables across the global ocean</td>
<td>191</td>
</tr>
<tr>
<td>A.11</td>
<td>Functional group overlaps in the global ocean microbiome</td>
<td>192</td>
</tr>
<tr>
<td>A.12</td>
<td>Correlations between functional groups in the global ocean microbiome</td>
<td>193</td>
</tr>
<tr>
<td>A.13</td>
<td>Taxonomic dissimilarities vs geographical distances in the global ocean microbiome</td>
<td>194</td>
</tr>
<tr>
<td>A.14</td>
<td>Functional vs taxonomic determinism in the ocean surface microbiome</td>
<td>195</td>
</tr>
<tr>
<td>B.1</td>
<td>Genus-level composition within functional groups across bromeliad microbiomes</td>
<td>214</td>
</tr>
<tr>
<td>B.2</td>
<td>Family-level composition within functional groups across bromeliad microbiomes</td>
<td>215</td>
</tr>
<tr>
<td>B.3</td>
<td>Order-level composition within functional groups across bromeliad microbiomes</td>
<td>216</td>
</tr>
<tr>
<td>B.4</td>
<td>Class-level composition within functional groups across bromeliad microbiomes</td>
<td>217</td>
</tr>
<tr>
<td>B.5</td>
<td>Geographic distances vs dissimilarities within functional groups across bromeliad microbiomes</td>
<td>218</td>
</tr>
<tr>
<td>B.6</td>
<td>Geographic location vs composition within functional groups in bromeliad microbiomes</td>
<td>219</td>
</tr>
<tr>
<td>B.7</td>
<td>PARAFAC model components of bromeliad DOC</td>
<td>220</td>
</tr>
<tr>
<td>B.8</td>
<td>Modeling EEMs of bromeliad DOC with PARAFAC</td>
<td>221</td>
</tr>
<tr>
<td>B.9</td>
<td>16S rDNA rarefaction curves for bromeliad microbiomes</td>
<td>222</td>
</tr>
<tr>
<td>B.10</td>
<td>Detailed functional profiles of bromeliad microbiomes</td>
<td>223</td>
</tr>
<tr>
<td>B.11</td>
<td>Functional redundancy in bromeliad microbiomes at the genus level</td>
<td>224</td>
</tr>
<tr>
<td>B.12</td>
<td>Functional redundancy in bromeliad microbiomes at the family level</td>
<td>225</td>
</tr>
<tr>
<td>B.13</td>
<td>Functional redundancy in bromeliad microbiomes at the class level</td>
<td>226</td>
</tr>
<tr>
<td>C.1</td>
<td>Predicted relative cell densities of the A and FS E. coli types in co-culture</td>
<td>242</td>
</tr>
<tr>
<td>C.2</td>
<td>Robustness of the predicted stable coexistence of the SS and FS E. coli types in co-culture</td>
<td>242</td>
</tr>
<tr>
<td>C.3</td>
<td>Measured relative cell densities of the SS and FS types in batch co-culture restricted to only one “season”</td>
<td>243</td>
</tr>
<tr>
<td>E.1</td>
<td>Predicted metabolite uptake rates in a methanogenic bioreactor</td>
<td>264</td>
</tr>
<tr>
<td>E.2</td>
<td>Predicted metabolite export rates in a methanogenic bioreactor</td>
<td>265</td>
</tr>
<tr>
<td>E.3</td>
<td>Predicted phage-host trajectories in a methanogenic bioreactor</td>
<td>266</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>F.1</td>
<td>Temperature, salinity and POM profiles in Saanich Inlet</td>
<td></td>
</tr>
<tr>
<td>F.2</td>
<td>Overview of the gene-centric modeling approach for the Saanich Inlet OMZ</td>
<td></td>
</tr>
<tr>
<td>F.3</td>
<td>Local sensitivity analysis of the gene-centric model</td>
<td></td>
</tr>
<tr>
<td>F.1</td>
<td>Comparison of multi-omic depth profiles for various sulfur and nitrogen cycling genes in the Saanich Inlet OMZ</td>
<td></td>
</tr>
<tr>
<td>G.1</td>
<td>Comparison of models for ammonium production in a ureolytic bioreactor</td>
<td></td>
</tr>
<tr>
<td>G.2</td>
<td>Comparison of models with and without ure-amo cross-amplification in a ureolytic bioreactor</td>
<td></td>
</tr>
</tbody>
</table>
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>amo</td>
<td>Aerobic ammonium oxidation</td>
</tr>
<tr>
<td>AOB</td>
<td>ammonia oxidizing bacteria</td>
</tr>
<tr>
<td>CH&lt;sub&gt;4&lt;/sub&gt;</td>
<td>methane</td>
</tr>
<tr>
<td>CO&lt;sub&gt;2&lt;/sub&gt;</td>
<td>carbon dioxide</td>
</tr>
<tr>
<td>COG</td>
<td>clusters of orthologous genes</td>
</tr>
<tr>
<td>DNRA</td>
<td>dissimilatory nitrate reduction to ammonium</td>
</tr>
<tr>
<td>FAPROTAX</td>
<td>functional annotation of prokaryotic taxa</td>
</tr>
<tr>
<td>FBA</td>
<td>flux balance analysis</td>
</tr>
<tr>
<td>H&lt;sub&gt;2&lt;/sub&gt;</td>
<td>hydrogen</td>
</tr>
<tr>
<td>H&lt;sub&gt;2&lt;/sub&gt;S</td>
<td>hydrogen sulfide</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td>KOG</td>
<td>KEGG orthologous groups</td>
</tr>
<tr>
<td>KTW</td>
<td>killing the winner</td>
</tr>
<tr>
<td>MCM</td>
<td>Microbial Community Modeler</td>
</tr>
<tr>
<td>MDS</td>
<td>multidimensional scaling</td>
</tr>
<tr>
<td>N&lt;sub&gt;2&lt;/sub&gt;O</td>
<td>nitrous oxide</td>
</tr>
<tr>
<td>NH&lt;sub&gt;3&lt;/sub&gt;</td>
<td>ammonia</td>
</tr>
<tr>
<td>NH&lt;sub&gt;4&lt;/sub&gt;&lt;sup&gt;+&lt;/sup&gt;</td>
<td>ammonium</td>
</tr>
<tr>
<td>NMDS</td>
<td>non-metric multidimensional scaling</td>
</tr>
<tr>
<td>NO&lt;sub&gt;2&lt;/sub&gt;</td>
<td>nitrite</td>
</tr>
<tr>
<td>NO&lt;sub&gt;3&lt;/sub&gt;&lt;sup&gt;-&lt;/sup&gt;</td>
<td>nitrate</td>
</tr>
<tr>
<td>NOB</td>
<td>nitrite oxidizing bacteria</td>
</tr>
<tr>
<td>nxr</td>
<td>Aerobic nitrite oxidation</td>
</tr>
<tr>
<td>O&lt;sub&gt;2&lt;/sub&gt;</td>
<td>oxygen</td>
</tr>
<tr>
<td>OMZ</td>
<td>oxygen minimum zone</td>
</tr>
<tr>
<td>PDNO</td>
<td>partial denitrification to nitrous oxide</td>
</tr>
<tr>
<td>PO&lt;sub&gt;4&lt;/sub&gt;&lt;sup&gt;-&lt;/sup&gt;</td>
<td>phosphate</td>
</tr>
<tr>
<td>ROM</td>
<td>remineralization of organic matter</td>
</tr>
<tr>
<td>OTU</td>
<td>operational taxonomic unit</td>
</tr>
<tr>
<td>SNTZ</td>
<td>sulfate-nitrate transition zone</td>
</tr>
<tr>
<td>SO&lt;sub&gt;4&lt;/sub&gt;&lt;sup&gt;-2&lt;/sup&gt;</td>
<td>sulfate</td>
</tr>
<tr>
<td>SSU rRNA</td>
<td>small subunit of the 16S rRNA gene</td>
</tr>
<tr>
<td>TEA</td>
<td>terminal electron acceptor</td>
</tr>
<tr>
<td>ure</td>
<td>Urea hydrolysis</td>
</tr>
</tbody>
</table>
Acknowledgements

I would like to express my deep appreciation to my supervisor, Prof. Michael Doebeli, who gave me the freedom and tremendous support to chase pretty much any white rabbit that crossed my way. I would like to thank all my collaborators, from whom I have learned a lot and with whom I was able to accomplish much of the work presented here: Alyse Hawley, Prof. Sergei Katsev, Captain Monica Torres-Beltran, Maya Bhatia, Celine Michiels, David Capelle, Gaute Lavik, Prof. Sean Crowe, Aria Hahn, Prof. Laura Parfrey, Sam Kheirandish, Saulo Jacques, Aliny Pires, Juliana Leal and Prof. Diane Srivastava. Special thanks to Prof. Steven Hallam for giving me an entrance to the world of microbial ecology and for plenty of stimulating discussions. Special thanks to Prof. Vinicius Farjalla for tremendous support on my field work and for being an amazing companion during my stay in Brazil. My gratitude to Mellisa Chen for teaching me a multitude of laboratory techniques. Thanks to Andreas Mueller for teaching me DNA extraction. Thanks to the Parfrey lab for hosting and supporting me along the way. Thanks to the Hallam lab for being such great collaborators. Thanks to Prof. Rebecca Tyson, Prof. Mary O’Connor and Prof. Daniel Coombs for their numerous advises and pleasant discussions. Thanks to Jan Finke, Andy Loudon, Prof. Angélica González, Sarah Perez and Matthew Osmond for numerous scientific discussions and for providing feedback on my manuscripts. Thanks to Prof. Sally Otto for being such a thought-provoking committee member and for providing feedback on my manuscripts. I am grateful to the Department of Mathematics (UBC), to the Pacific Institute for the Mathematical Sciences (PIMS) and to NSERC for funding. Thanks to my parents, brother and sister for supporting me at all times.
Chapter 1

Openning chapter

For such a model there is no need to ask the question “Is the model true?”. If “truth” is to be the “whole truth” the answer must be “No”. The only question of interest is “Is the model illuminating and useful?”

George Box, 1979

1.1 Introduction

Microorganisms are the most ancient, the most diverse and the most abundant form of life on Earth (518). The biomass of prokaryotes alone is comparable to that of all plants combined (518), and their distribution extends far beyond that of multicellular organisms (111, 404). The metabolic activity of microorganisms drives the bulk of biogeochemical fluxes in virtually every natural ecosystem (116), including marine sediments (388), soil (231, 494), the open ocean (79, 455) and freshwater lakes (126). Cyanobacteria, for example, perform up to 35% of global photosynthesis (356), turning solar energy into a redox disequilibrium between carbon and oxygen that powers much of current heterotrophic life. Microorganisms strongly shape the marine nitrogen and sulfur cycles, thereby modulating global ocean productivity and climate (136). For example, denitrification and anaerobic ammonia oxidation (anammox), two microbial pathways that utilize nitrogen compounds as alternative terminal electron acceptors for respiration, can lead to a significant net loss of bioavailable nitrogen to N$_2$ (501). On the other hand, nitrification, an exclusively prokaryotic process by which reduced nitrogen compounds are aerobically oxidized to nitrate for energy, plays a central role in soil productivity (375) and industrial processes such as wastewater treatment (520). Understanding the spatiotemporal dynamics of microbial metabolic processes is therefore central to understanding overall ecosystem biochemistry and towards optimizing the performance of microbially driven industrial processes (316).

Until recently, difficulties in culturing and therefore characterizing the majority of microor-
ganisms has been a bottleneck to microbial ecology (21, 203, 346). With the advent of high-throughput molecular techniques, notably marker-gene sequencing (353, 529), shotgun DNA sequencing (metagenomics; 534), shotgun RNA sequencing (metatranscriptomics; 157, 511) and mass-spectrometry based protein sequencing (metaproteomics; 298, 359), we are now entering a new era of biological inference (169). These culture-independent techniques generate massive amounts of data and provide unprecedented insight into the composition, metabolic potential and activity of microbial communities (98). However, the generation of these data is still rarely theory-driven and in most cases their mechanistic interpretation remains elusive (209, 241, 380, 401, 459). For example, while taxonomic community profiling can reveal intriguing variation, for example along the ocean water column (460) or across seasons (540), the reasons for the observed taxonomic variation remain largely unknown because the ecological role of most taxa is unknown and can vary strongly even between closely related clades (3, 308).

The mechanistic modeling of microbial communities is further complicated by the sheer diversity of microorganisms in any particular environment (377, 401). For example, a single gram of soil can harbor several thousands of different — and potentially interacting — microbial species (399). Conventional reductionist approaches would require a careful physiological characterization of each member species. However, even the simplest life forms cannot be studied in isolation because the ability to catalyze different steps of metabolic pathways or synthesize required biomolecules is partitioned across different organisms, and hence each species inevitably only constitutes a small link in the overall reaction network sustaining life in an environment (183, 200, 238, 311). An incorporation of each species into a comprehensive mathematical model similar to some mechanistic “macro-ecological” models (230, 412) is thus impractical for most natural microbial communities, although such approaches have been suggested in the past (258). In most cases, a different approach is needed for describing microbial metabolic processes at ecosystem scales.

Despite an enormous microbial diversity, the bulk of global biogeochemical fluxes is driven by a core set of metabolic pathways, which have evolved and proliferated in response to various redox disequilibria available for sustaining life (116). Through time and notably via horizontal gene transfer, these pathways have spread across microbial clades that can co-occupy or replace each other within metabolic niches (116, 270, 483). The growth of each clade is inevitably coupled to the activity of its energy-yielding pathways, and this activity is subject to environmental energetic and stoichiometric constraints such as the availability of specific electron donors and acceptors (56). For example, the abundance and expression of genes linked to nitrogen and sulfur metabolism in marine oxygen minimum
zones typically reflect the varying redox conditions along the water column (181, 448). On the other hand, microbial — notably prokaryotic — genomes rapidly lose pathways that are not required in their natural environment, presumably due to strong selective pressure for genome streamlining (41, 149, 250, 334). It is therefore tempting to theorize that these pathways — or more precisely, the genes and operons encoding them — constitute largely independent units of replication and selection (41, 93), and that environmental physicochemical conditions prescribe the structure of the community metabolic network regardless of which organisms happen to occupy each metabolic niche (126). The taxonomic composition within each niche may of course be subject to additional selection processes, such as tolerance to particular environmental stressors or susceptibility to specialist phage populations (397, 423), however the resulting variation in taxonomic composition may have little effect on overall metabolic functioning. Such a “pathway-centric” paradigm, if applicable, would greatly simplify the modeling of microbially mediated processes at ecosystem scales and would provide holistic insight into global biogeochemistry. Further, modeling microbial communities at the level of pathways or genes would enable a direct integration of metagenomic, metatranscriptomic and metaproteomic sequence data, which to date remain largely unutilized in quantitative ecosystem models (258).

1.2 Problem statement

The pathway-centric paradigm assumes that the metabolic function of a community somehow becomes decoupled from its specific taxonomic makeup, so that overall community metabolism is strongly controlled by physicochemical environmental factors alone, while additional processes shaping taxonomic composition have little effect on metabolic function. It is a priori unclear what conditions would promote such a decoupling and how appropriate a pathway-centric paradigm is for natural microbial communities. For communities with low taxonomic richness, or in which certain pathways are only performed by a small set of organisms, metabolic activity may depend strongly on the particular genomes present. In particular, multiple pathways co-occurring in the same genomes will not behave as independent replicating units, and this will likely lead to deviations from the pathway-centric paradigm. On the other hand, at the community level the potential presence of the same pathways in alternative configurations could enable pathway independence. Hence, functional redundancy, i.e., the presence of multiple alternative clades capable of filling a specific metabolic niche, may promote the decoupling between the taxonomic makeup of a community and its metabolic function, however this remains to be tested.
A key prediction of the pathway-centric paradigm is that physicochemically similar environments would promote similar metabolic functional community structure, while allowing for strong taxonomic variation within individual functional groups. This prediction is indeed supported by observations in engineered ecosystems, such as bioreactors exhibiting strong taxonomic fluctuations while maintaining constant biochemical performance (122, 495), although it remains largely untested for natural microbial communities (25). More generally, one would expect that variable environmental conditions correlate more strongly with community function than with taxonomic composition, especially taxonomic composition within functional groups. An analysis of metagenomes from the Global Ocean Survey (406) indeed revealed strong correlations between pairwise environmental differences on the one hand, and pairwise metagenomic (but not taxonomic) community differences on the other hand (381). Similarly, bacterial community composition on the macroalgae *Ulva australis* was best explained in terms of metagenomic content rather than species content (52). These findings are consistent with a pathway-centric paradigm, however they do not explicitly address the taxonomic composition within individual functional groups, perhaps because assigning metagenomic sequences to specific taxa is a notoriously hard problem (471; which, as we show in Chapter 2, can be circumvented).

Even if not absolutely accurate, a pathway-centric paradigm would constitute an elegant and potentially insightful null model for microbial ecology because, as discussed above, it makes concrete predictions about microbial community metabolism and its interaction with the environment. Hence, the question is not whether a pathway-centric paradigm would be true; indeed, the short answer is “No”. Rather, how much of the “truth” would we really be missing in a pathway-centric paradigm, in which pathways are no longer associated with a specific host but constitute self-serving replicators that directly interact with their environment? The real question of interest is, would such a paradigm be “illuminating and useful?” (43) and if so, under which conditions? For example, it may seem intuitive that a high functional redundancy would promote the decoupling between environmentally-driven pathway dynamics on the one hand, and the particular taxonomic composition of a community on the other hand, but this remains to be rigorously examined.

Further, any pathway-centric description of microbial communities would only capture part of the full story because it would make no statement about the assembly within individual metabolic guilds, which may be driven by a multitude of additional mechanisms. Such mechanisms may include biotic interactions such as competition or predation (258, 276, 455), random population drift (349), random colonization order (“lottery effects” 52), spatially limited random dispersal (306) and microbial chemical warfare (237). For example, adapta-
tion of bacteriophages to specific hosts can strongly influence bacterial species composition (423) and promote variation of microbial communities through so-called “killing the winner dynamics” (397, 462). Further, trade-offs between environmental stress tolerance and competition may lead to additional environmental filtering within functional groups (159, 362). The question then arises, at what point do these additional processes significantly influence community metabolism?

Apart from questions regarding the appropriateness and limitations of a pathway-centric paradigm, questions also emerge regarding the precise mathematical formulation of an ecological theory for microbial metabolic pathways. Specifically, how exactly would an environment “determine” the distribution and activity of particular pathways, and how should the current activity of pathways feed back to the “population dynamics” of these pathways? Further, how could multi-omic (metagenomic, metatranscriptomic and metaproteomic) sequence data be quantitatively incorporated into such models? Inspiration may be taken from existing ecosystem models in which broad microbial processes are represented by homogenous functional groups, such as photoautotrophs or detritivores. In these models, the activity of functional groups is determined by resource-dependent responses, such as Michaelis-Menten kinetics (211), and their growth is determined by simple biomass-per-substrate yield factors (193, 360). These “functional group models” are much coarser than potential descriptions of metabolic networks at the pathway level, and this coarseness may explain why multi-omic data are yet to be incorporated in such models. Further, such models rarely account for energetic constraints of microbial metabolism because yield factors are considered to be fixed model parameters. In reality, however, the energy that can be gained from a metabolic reaction depends on the specific physicochemical state of the local environment, and thus local reaction energetics strongly shape the structure of the microbial metabolic network (34, 56, 256). A recent biogeochemical model by Reed et al. (386), which predicts gene growth rates based on the energy yield from their associated metabolic pathways, constitutes a first attempt to construct a thermodynamics-based pathway-centric model. While compelling, the model by Reed et al. (386) stops short of a quantitative integration between geochemical and multi-omic sequence data. For example, while the model allowed for a qualitative comparison between modeled gene production rates and selected transcript profiles, it does not provide any explicit mechanistic links (386). In fact, the lack of a quantitative validation against process rate measurements or other proxies for activity (e.g., proteins) begs the question whether such pathway-centric models are adequate descriptions for microbial community metabolism.
1.3 Objectives and overview of this dissertation

In this dissertation, I examine the appropriateness of a pathway-centric paradigm for microbial ecology using microbial community profiling and mathematical modeling. Further, I examine concrete biogeochemical pathway-centric models for specific ecosystems, which I evaluate using chemical and molecular sequence data.

Specifically, in Chapter 2, I use metagenomic data from the Tara Oceans survey (460) to investigate the functional structure of bacterial and archaeal communities across the global ocean, as well as the taxonomic composition within each of 28 metabolically defined functional groups. For that purpose, I constructed a custom database for functional annotations of prokaryotic taxa (“FAPROTAX”) based on extensive literature search. Using this database, I find that environmental physicochemical conditions strongly predict the distribution of microbial metabolic functional groups, but only poorly predict the taxonomic composition within individual functional groups, in line with a pathway-centric paradigm.

In Chapter 3, I use metagenomic sequencing and 16S rRNA marker gene sequencing of bacterial and archaeal communities within the foliage of bromeliad plants, a model system for community ecology (117, 438), to show that similar aquatic environments can indeed sustain similar microbial metabolic networks, despite a highly variable taxonomic composition within individual metabolic functional groups. I then use statistical tools from community ecology to elucidate the potential mechanisms shaping the taxonomic composition within functional groups. I find that deterministic biotic interactions and additional environmental filtering, but not random drift or dispersal limitation, likely underlie the observed taxonomic variation between bromeliad microbiomes.

In Chapter 4, I present a computational framework (“Microbial Community Modeler”, short MCM) that I developed for modeling microbial communities. This framework enables the construction of both pathway-centric models as well as cell-centric models (using genome-based metabolic cell models). I validate this framework by modeling previous laboratory evolution experiments with Escherichia coli (188, 262), during which an ancestral strain diversified into two coexisting ecotypes. The models are able to reproduce the successional dynamics in the evolution experiments, and yield detailed insights into the metabolic processes that drove bacterial diversification. Thus, apart from demonstrating the potential of MCM, this work provides a unifying quantitative perspective on a multitude of co-culturing experiments performed in our lab over the last 10 years.
In Chapters 5 and 6 I use MCM to examine specific mechanisms by which biotic interactions and functional redundancy could potentially affect community composition and metabolism, based on models for nitrifying as well as methanogenic bioreactors. I find support for the interpretation that biotic interactions, such as competition (Chapter 5) and predation by phages (Chapter 6), may underlie a large part of the taxonomic variation within functional groups, reported in Chapters 2 and 3. Further, these models explicitly demonstrate how a higher functional redundancy can lead to a decoupling between functional stability and taxonomic variability under constant environmental conditions.

Taken together, the work presented in Chapters 2, 3, 5 and 6 provides strong support for a pathway-centric paradigm, according to which the distribution and activity of metabolic pathways is strongly determined by physicochemical conditions, whereas additional mechanisms shape the taxonomic composition within metabolic guilds. These results motivated me to develop and test specific pathway-centric biogeochemical models for a series of natural as well as engineered ecosystems.

Specifically, in Chapter 7 I describe a biogeochemical model that I constructed for the oxygen-depleted water column in Saanich Inlet (540), a seasonally anoxic fjord with biogeochemistry analogous to oxygen minimum zones (OMZs). The model describes the activity and distribution of individual microbial metabolic pathways involved in carbon, nitrogen and sulfur cycling in Saanich Inlet and integrates geochemical depth profiles, process rate measurements as well as DNA, mRNA and protein sequence data.

In Chapter 8, I present an alternative pathway-centric modeling framework, in which the metabolic activity of a microbial community is described purely in terms of reaction rates and the “capacity” to perform particular reactions. The benefits of such a “reaction-centric” approach, when compared to models explicitly keeping track of cell (or gene) densities, is the reduced number of physiological parameters required for constructing a model and of biotic measurements required to estimate the current state of a system. I validate this approach using data from previous bioreactor experiments (94, 109). Further, I examine how the co-occurrence of multiple pathways in the same organisms, rather than in separate organisms, can affect overall community metabolism and thus the accuracy of the pathway-centric paradigm.

Methodological details for each chapter are provided as supplemental material in Appendices A–G.
Chapter 2

The decoupling of function and taxonomy in the global ocean microbiome

2.1 Synopsis

Here we use statistical analyses of taxonomic and functional community profiles to determine the factors that shape marine bacterial and archaeal communities across the global ocean. Through an extensive literature search we classified >30,000 microbial organisms into metabolic functional groups, which allowed us to disentangle functional from taxonomic community variation. We find that environmental conditions strongly influence the distribution of functional groups by shaping metabolic niches, but barely influence the taxonomic composition within individual functional groups. Hence, the bulk of environmentally driven variation in community composition is attributable to functional properties, while the remaining variation is enabled by a high global functional redundancy across taxa.

2.2 Introduction

Microbial communities power global biogeochemical cycling and form the most important interface between abiotic and biotic processes on Earth (116). Bacteria and Archaea, in particular, drive marine nitrogen and sulfur cycling, thereby modulating global ocean productivity and climate (136). Elucidating the processes shaping microbial communities over space and time presents a missing link towards understanding the integrated biotic-abiotic system we call our planet, and is key for predicting how biogeochemical cycles will change with changing environmental conditions.

The majority of microbial biogeographical studies focus on taxonomic community compo-

1A version of this chapter has been published in Science (see the Preface for author contributions): Louca, S., Parfrey, L.W., Doebeli, M. (in press). Decoupling function and taxonomy in the global ocean microbiome. Science. 353:1272–1277. DOI:10.1126/science.aaf4507.
sition (306). Taxonomic community profiling can reveal intriguing variation between environments, and functional differences between organisms are generally thought to cause these patterns. Distantly related microbes, however, can often perform similar metabolic functions and, reciprocally, closely related taxa may occupy separate metabolic niches (308). This leads to a disconnect between taxonomic community structure and function (122, 466). As a consequence, the ecological reasons for the observed taxonomic variation usually remain unknown. Other studies directly estimate functional potential based on community gene content using environmental shotgun sequencing—or metagenomics—and have indeed revealed strong correlations between the distribution of particular metabolic pathways and environmental conditions (99, 181, 381). This suggests that environmental conditions shape the functional potential of microbial communities in terms of community gene content by constraining metabolic niches. However, it is yet unknown how this relates to community assembly rules, and which aspect of the variation in taxonomic composition is relevant to ecosystem functions. In addition to these niche effects (also known as environmental filtering), microbial populations are subject to complex community-level processes such as predation or mutualistic interactions (258, 455), as well as to potential limits to their dispersal across spatial scales (306). Given this complexity, it is important to establish basic principles determining microbial community composition.

Here we present an analysis of over 100 bacterial and archaeal communities across the global ocean, combining taxonomic and functional community profiling to elucidate the role of environmental filtering, global functional redundancy and dispersal limitation in shaping natural microbial communities. Taxonomic profiles were generated based on shotgun DNA sequences of the 16S ribosomal gene, a standard marker gene in microbial ecology. Functional profiles were generated by associating individual organisms with metabolic functions of particular ecological relevance, such as photoautotrophy and sulfate respiration, using an annotation database that we created through extensive literature search. This information, which we validate by comparing it to metagenomic gene profiles, allowed us to explore multiple facets of microbial community structure — taxonomic composition, metabolic functional potential and taxonomic composition within individual functional groups — in relation to environmental conditions and geographical location.
2.3 Environmental conditions mainly affect microbial function

To assess the effects of environmental conditions on various aspects of community structure, we performed regression and correlation analysis of the relative abundances of metabolic functional groups, as well as the proportions of various operational taxonomic units (OTU) within each functional group, against 13 key abiotic oceanographic variables that included dissolved oxygen, salinity, temperature and depth (Table A.1). Both regression and correlation analyses generally revealed that environmental conditions had very strong effects on the functional profiles of microbial communities, but only minor effects on the taxonomic composition within each functional group. In particular, the cross-validated coefficient of determination ($R^2_{cv}$, a measure of the predictive power of a model) for the relative abundances of most functional groups greatly exceeds the average $R^2_{cv}$ achieved for the OTU proportions within the same groups (Fig. 2.1A). Similarly, correlations between relative functional group abundances and environmental variables are generally greater in magnitude, compared to the correlations between OTU proportions within each group and environmental variables (Fig. 2.1B). These differences persist even when OTUs are combined at higher taxonomic levels (e.g., genus, family or order; Figs. A.1 and A.2). Hence, the poor correlation between the taxonomic composition within functional groups and environmental conditions is not due to a sub-optimal choice of taxonomic resolution, but rather reflects a lack of environmental effects on the non-functional variation in community composition.

Regression modeling of taxonomic profiles of entire communities (at any taxonomic resolution) against environmental variables achieved an average $R^2_{cv}$ (Fig. 2.5) that is lower than the $R^2_{cv}$ for the relative abundances of most functional groups, but higher than the mean $R^2_{cv}$ achieved for the taxonomic compositions within functional groups. This further supports the interpretation that deterministic environmental effects on function only partly shape overall taxonomic community structure, due to taxonomic variation within functional groups that is much less affected by environmental conditions. Accordingly, clustering samples by taxonomic as well as functional community composition (Bray-Curtis dissimilarity metric, Figs. 2.2BC and A.3) shows that a distinction between water column zones based on function is comparable in strength to a distinction based on taxonomy. In fact, the fraction of functional groups exhibiting statistically significant segregation between water column zones (e.g., mesopelagic vs surface) is usually higher than the fraction of significantly segregated OTUs or higher taxa (Figs. 2.2E–G). Hence, the bulk of deterministic variation in community composition across different zones is well captured by the variation of purely functional
Decoupling between function and taxonomy in the ocean microbiome

properties.

These results strongly suggest that environmental conditions influence microbial community structure in the global ocean primarily by shaping metabolic niches. In fact, we find that the productivity of a particular metabolic niche – represented here by the relative abundance of a functional group – generally only weakly influences the taxonomic composition within that niche (Fig. 2.3B). The importance of niche effects in structuring marine microbial communities is further underlined by our finding that OTUs sharing a higher number of functions tend to co-occur more frequently (Fig. 2.2H). In contrast, if competition and species assortment were dominant forces we would expect lower co-occurrences among OTUs with more shared functions. In a similar way metabolic niche effects were shown to dominate human gut microbiome assembly (271), suggesting that this may be a general pattern in natural microbial communities.

The decoupling between environmental conditions and niche productivity on the one hand, and the taxonomic composition within individual niches on the other hand, is consistent with previous smaller-scale observations. For example, in a wastewater treatment plant the ratio of aerobic ammonia oxidizing bacteria and heterotrophic bacteria remained constant over time, while the taxonomic composition within each functional group varied markedly and appeared to be only weakly explained by environmental factors (349). Moreover, metatranscriptomics in two distinct ocean regions revealed strongly conserved diurnal cyclic succession patterns of community gene expression, despite largely non-overlapping taxonomic affiliations of transcripts between the two regions (25). Our work suggests that these previous observations are in fact just the tip of an iceberg. Energetic and stoichiometric constraints generally drive ocean microbial metabolic activity, but not the identity of the microbes involved in that activity.

2.4 Causes of variation within functional groups

If environmental conditions mainly interact with functional community structure, the question arises as to what drives the taxonomic variation within individual functional groups. High functional redundancy on a global ocean scale (Fig. 2.3) presumably enables the high diversity within functional groups and a decoupling between taxonomic composition and constraints on function. It is possible that the “unexplained” taxonomic variation within functional groups may be partly due to unconsidered physicochemical variables combined with unconsidered phenotypic differences, driving location-dependent growth differences be-
tween competing species. However, it is unlikely that latent environmental variables alone could explain the widespread apparent randomness seen within such a large number of functional groups (Fig. A.7). Alternatively, spatially limited dispersal is often suggested as a cause of random distribution patterns not attributable to environmental conditions. Dispersal limitation has been shown to be important for larger organisms such as plants (487), but its importance for microorganism is generally thought to be low compared to environmental filtering, and to depend strongly on the environments considered (e.g., lakes vs open ocean; (29, 306, 466)). To test whether the variation in taxonomic composition in the ocean communities considered here can be explained by dispersal limitation, we compared geographical sample distances to dissimilarities in functional and taxonomic community composition, as well as to differences in composition within individual functional groups. Mantel correlation tests between geographical distances and various dissimilarity metrics revealed no significant positive correlations, neither at the level of the whole community (Figs. 2.4AB) nor within functional groups (Figs. A.4). The apparent absence of a distance-decay of similarity is also reflected in ordination plots (Figs. 2.4CD), in which sample clustering appears to be independent of ocean region, with the notable exception that polar samples are clearly distinct. These observations suggest that range-limited random dispersal does not play a significant role in shaping taxonomic community differences across the oceans, even when restricted to within metabolic niches. In principle, dispersal limitation may cause a distance-decay of similarity at much smaller geographical scales than the ones considered here, and thus remain undetected by our analysis (307). This scenario, however, is unlikely for marine microorganisms that can be dispersed at global scales by large ocean currents (125), and that indeed appear to be recruited from a global marine microbial seed bank (147).

Instead, the unexplained taxonomic variation may be driven predominantly by community-level processes such as metabolic interdependencies, chemical warfare, or predation by viruses and eukaryotes. Previous work has highlighted the central role of such processes in microbial community assembly (258, 276, 455). For example, adaptation of bacteriophages to specific hosts can cause continuous replacement of competing species (423) and promote spatial structuring of microbial communities (462). Hence, functional community structure and taxonomic composition within functional groups appear to constitute roughly complementary axes of variation, with the former being affected more strongly by environmental conditions and the latter shaped by community-level processes. In this perspective, the observed overall microbial community structure would result from the superposition of both environmental constraints and community-level processes (276, 455).
2.5 Beyond taxonomic profiling

Taxonomic microbial community profiles can reveal intriguing differences between environments, for example across mammalian guts (335) or along the ocean water column (460). However, microbial taxonomy and metabolic potential can exhibit significant inconsistencies (3); the extent of these inconsistencies is strongly trait-dependent and often considerable (308). Inconsistencies between taxonomy and metabolic potential are driven by diverse evolutionary processes including adaptive loss of function (165, 334) and metabolic convergence of distinct clades accelerated by frequent horizontal gene transfer (116). As a result, many metabolic phenotypes can be shared by distant microbial clades (Figs. 2.3AC) and, reciprocally, members of the same clade can fill separate metabolic niches (165). This misalignment between taxonomy and metabolic potential complicates the mechanistic interpretation of taxonomic biogeographical patterns and the development of predictive ecological theories (381). Thus, while modern marker-gene sequencing techniques can yield detailed taxonomic community profiles, they suffer from the crucial limitation that these high-dimensional profiles are obtained along axes of sub-optimal ecological relevance. Sophisticated statistical techniques such as principal coordinate analysis or multidimensional scaling are often used to reduce the dimensionality of these data, in the hope of detecting axes of variation bearing an ecological meaning (382). Alternatively, detected species may be combined at higher taxonomic levels that more closely resemble the depth at which traits vary across lineages. However, the optimal taxonomic level is highly trait-dependent (308).

Here we have shown that a straightforward but thorough binning of organisms into functional groups can reveal strong patterns relevant to biogeochemistry and ecosystem function. In particular, the comparison of functional profiles to environmental variables (Fig. 2.1) yields insight into the processes driving variation in community composition along geochemical gradients and, reciprocally, gives information about the effects of that variation on ecosystem processes. For example, our correlation analysis revealed a particularly strong influence of water depth on function (Fig. 2.1B), which is also reflected in the clear separation of the mesopelagic zone from the upper sunlit zones (surface and deep chlorophyll maximum; Figs. 2.2AB). This functional zonation with depth is consistent with metagenomic profiles of the same samples (Fig. A.5). The partitioning of function along the water column is presumably driven by depth-dependent factors known to influence microbial life history strategies and productivity, such as light intensity and temperature (236). Furthermore, in deeper zones oxygen becomes a limiting resource for respiration that can be partly replaced by alternative electron acceptors such as nitrate and sulfate (531). Such redox gradients underlie the
spatial zonation of metabolic pathways frequently observed in gene-centric metagenomic, metatranscriptomic and metaproteomic profiles across ecosystems (99, 181). Accordingly, our functional profiles reveal that deeper samples exhibit an over-representation of several groups capable of fermentation as well as nitrate and sulfate respiration (Fig. 2.2A). These alternative metabolic modes lead to an increased community richness (in terms of detected functional groups and OTUs), especially in the mesopelagic zone (Fig. 2.2D, after rarefying at equal sequence count). While an increase of taxonomic richness with depth has been noted previously (368, 460), our analysis now provides evidence that this richness gradient is strongly related to the number of available metabolic niches. Hence, the systematic integration of taxonomic and functional information demonstrated here can help answer long standing questions regarding the relation between microbial taxonomic and functional diversity and variability in a given environment (283). Similarly, functional group co-occurrence patterns (e.g., Fig. A.6) may trigger novel hypotheses on the interaction of metabolic pathways at ecosystem scales.

The translation of taxonomic information into functional profiles based on phenotypic characterizations, as demonstrated here, provides a powerful complement to gene-centric metagenomics, the current de facto standard for functional community profiling (486). Metagenomic profiles suffer from the conceptual limitation that community gene content generally does not directly and unambiguously translate to functional potential (376). Phenotype-based profiles on the other hand have the potential to resolve ambiguities inherent to metagenomics, because experimental evidence is used to identify the actual metabolic capabilities of organisms (see supplementary text for examples). We emphasize that the full potential of phenotype-based profiling remains underutilized due to our current inability to associate several detected taxa with any functional group. For example, a large fraction of the ubiquitous but poorly studied Thaumarchaeota phylum potentially involved in ammonia oxidation (441) was excluded from our functional annotations. Similarly, microeukaryotes such as fungi likely contribute to some of the considered metabolic functions, including cellulose and chitin degradation (91). The restriction to Bacteria and Archaea may thus explain our inability to relate the distribution of these functional groups to environmental conditions (Fig. 2.1). Future functional profiling will greatly benefit from an inclusion of eukaryotic microorganisms and from an effort to phenotypically characterize underexplored (e.g., hard-to-culture) clades of potentially high relevance to ecosystem functioning.
2.6 Conclusions

Despite an enormous microbial diversity, the bulk of global biogeochemical fluxes is driven by a core set of metabolic pathways that evolved in response to past geochemical conditions (116). Through time, these pathways have spread across microbial clades that can co-occupy and compete within metabolic niches. The decoupling of taxonomic composition within metabolic niches from environmental conditions, as demonstrated here, suggests that Baas Becking’s famous hypothesis “everything is everywhere and the environment selects” (28) should be refined towards the more conservative formulation that “every function is everywhere, and the environment selects” (381). This realization has implications for the interpretation of differences in community structure across environments or across time, because differences in taxonomic composition that do not affect functional composition have little relevance to ecosystem processes (122). Functional descriptions of microbial communities should therefore constitute the baseline for interpreting biogeographical patterns, particularly across transects where geochemical gradients shape microbial niche distribution (162, 386). The remaining variation within functional groups can then be analyzed separately to elucidate additional community assembly processes. As the number of phenotypically characterized organisms increases, the potential of functional annotations of microbial taxa can only further improve. An incorporation of global microbial functional profiles — and their response to potentially changing environmental conditions — into future biogeochemical models will greatly benefit reconstructive and predictive modeling of Earth’s elemental cycles.
Figure 2.1: Linking functional and taxonomic composition to environmental conditions. (A) Cross-validated coefficients of determination ($R^2_{cv}$) for relative functional group abundances (left bars), as well as for OTU proportions within each functional group (right bars), achieved by regression models with environmental predictor variables. (B) Spearman rank correlations between environmental variables and relative functional group abundances (left box) or OTU proportions within each group (right box). Circle surface area and color saturation are proportional to the absolute correlation.
Decoupling between function and taxonomy in the ocean microbiome

Figure 2.2: **Environmental filtering of microbial communities in the global ocean.** (A) Functional community profiles, with samples ordered according to water column zone (SRF: surface water layer; DCM: deep chlorophyll maximum; MIX: subsurface epipelagic mixed layer; MES: mesopelagic zone). A darker color corresponds to a higher relative abundance of a functional group. (B,C) Metric multidimensional scaling of microbial communities (one point per sample), based on Bray-Curtis dissimilarities in terms of (B) functional groups and (C) OTUs. Points in greater proximity correspond to more similar communities. (D) Community richness in terms of functional groups and OTUs (one point per sample), after rarefaction. (E–G): Fraction of OTUs, taxa or functional groups significantly segregated between water column zones (E: DCM vs MES, F: DCM vs SRF, G: MES vs SRF). (H) Box-plots of pairwise Spearman rank correlations between relative OTU abundances, depending on the number of shared functions. Vertical bars show 66% percentiles.
Figure 2.3: **Functional redundancy in the global ocean microbiome.** (A) Number of bacterial and archaeal taxa represented within each functional group (one point per group), at various taxonomic levels. At the species and genus level, aerobic chemoheterotrophs present by far the richest group. (B): OTU proportions within the group of aerobic ammonia oxidizers (one color per OTU). Samples are sorted according to the relative abundance of the entire functional group. For OTU proportions within other functional groups, see Fig. A.7. (C) Association of functional groups (columns) with members of microbial classes (rows). A darker color corresponds to a higher relative contribution of a class (in terms of the number of OTUs) to a functional group. Rows and columns are sorted accorded to the number of non-zero entries within them.
Figure 2.4: **Community differences vs geographical distance.** (A,B): Bray-Curtis dissimilarities between microbial communities, compared with geographical distances (one point per sample pair). Community dissimilarities are calculated in terms of (A) relative functional group abundances and (B) relative OTU abundances. Plot imprints indicate Spearman rank correlations and their statistical significance. Samples are restricted to the mesopelagic zone; for other water column zones see Fig. A.8. For other taxonomic resolutions (e.g., at genus or family level) see Fig. A.13. (C,D): Metric multidimensional scaling of microbial communities (one point per sample), based on Bray-Curtis dissimilarities in terms of (C) functional groups (C) OTUs. Points in greater proximity correspond to more similar communities. Points are shaped and colored by ocean region.
Figure 2.5: **Regression of taxonomic community composition.** Mean cross-validated coefficients of determination ($R^2_{cv}$) for relative taxon abundances at the community level (for various taxonomic resolutions), achieved by regression models with environmental predictor variables.
Chapter 3

Functional stability despite high taxonomic variability across microbial communities in bromeliad tanks

3.1 Synopsis

According to the pathway-centric paradigm suggested in the previous chapter, the metabolic functional structure of microbial communities is strongly shaped by environmental conditions constraining the metabolic pathways that sustain growth. Apart from metabolic niche effects, however, several additional processes such as biotic interactions or dispersal limitation can influence overall community composition. Thus, similar habitats could exhibit very different microbial communities despite a similar functional structure. To test this prediction, we determined the bacterial and archaeal community composition in 22 replicate “miniature” aquatic ecosystems, contained within the foliage of wild bromeliads. We used 16S rDNA marker gene sequencing for inferring the taxonomic composition within 9 metabolically defined functional groups, as well as shotgun environmental DNA sequencing for estimating the overall abundances of these groups. We find that all bromeliads exhibit remarkably similar functional community structure, but a highly variable taxonomic composition within individual functional groups. Using a variety of statistical analyses from community ecology, we find evidence that the taxonomic turnover within functional groups is driven by a combination of environmental filtering and biotic interactions. We find no effect of dispersal limitation or random population drift on community composition, and conclude that complex deterministic processes — rather than neutral assembly — shape community variation within functional groups.

1A version of this chapter is currently under review for publication (see the Preface for author contributions): Louca, S., Jacques, S.M.S., Pires, A.P.F., Leal, J.S., Srivastava, D.S., Parfrey, L.W., Farjalla, V.F., Doebeli, M. (in review). Functional stability despite high taxonomic variability across microbial communities.
3.2 Introduction

Microbial metabolism drives the bulk of biogeochemical fluxes in virtually every natural ecosystem and has shaped Earth’s surface chemistry through geological time (116). Natural microbial communities can display complex variation in composition across space or time, such as through the ocean water column (460) or across seasons (540), and this variation can have profound effects on ecosystem functions (500, 540). The mechanisms driving this variation remain poorly understood, because the entanglement of multiple mechanisms severely complicates the identification of direct causal relationships (258). Potential mechanisms of microbial community assembly suggested previously include adaptation to local environmental conditions (“environmental filtering” 371), biotic interactions such as predation or syntrophy (258, 276, 455), random population drift (349), random colonization order (“lottery effects” 52) and spatially limited random dispersal (306). Recent work suggests that the bulk of environmentally driven variation in the global ocean microbiome is closely related to its metabolic function, while the taxonomic variation within individual functional groups is only poorly explained by environmental conditions (289, 381). This points towards a promising and elegant paradigm for microbial ecology, in which community function is strongly shaped by energetic and stoichiometric constraints such as the availability of light or electron acceptors for respiration (25, 381), while the composition within functional groups is modulated by additional deterministic or stochastic mechanisms. According to this paradigm, one would predict that physicochemically similar environments will promote similar functional community structure, while allowing for strong taxonomic variation within individual functional groups. This prediction is supported by observations in engineered ecosystems such as bioreactors exhibiting strong taxonomic fluctuations while maintaining constant biochemical performance (122, 495), but it remains largely untested for natural microbial communities.

Here we test this prediction in natural prokaryote (i.e., bacterial and archaeal) communities across 22 replicate natural aquatic environments, harbored within the foliage (“tanks”) of bromeliads in the Jurubatiba coastal sand dune National Park, Brazil (Fig. 3.1). Bromeliad tanks accumulate rain water and organic material (such as dead leaves) from their surrounding environment, and intense decomposition of this material sustains a high richness of microorganisms and macroinvertebrates (152, 395). Apart from constituting regional biodiversity hotspots, bromeliads are often used as “miniature” model systems for microbial and invertebrate ecology (117, 438). Microbial communities in bromeliads tend to be highly distinct from the surrounding environments (e.g., soil), exhibiting a strong shift towards
fermenting and methanogenic organisms (151, 152, 304).

To ensure a high similarity between systems, we only surveyed mature plants of a single bromeliad species (*Aechmea nudicaulis*) from the same region (296). We used amplicon DNA sequencing of the 16S ribosomal gene, a standard marker gene in microbial ecology (536), to estimate the taxonomic richness and variability within 9 metabolically defined functional groups of potential ecological importance in bromeliads, such as fermentation, dissimilatory reduction of nitrogen compounds (“nitrogen respiration”) and methanogenesis (151, 152, 345). Detected taxa were assigned to these metabolic functional groups, whenever possible, based on available literature. In parallel, we used environmental shotgun DNA sequencing (metagenomics) to estimate the overall abundance of each functional group in terms of one or multiple proxy genes. We find that all communities exhibited a remarkably similar functional composition, which contrasts with a highly variable taxonomic composition within individual functional groups. Further, we examined phylogenetic community structure and species co-occurrence patterns, and compared community composition to abiotic environmental conditions and geographical location, to elucidate potential mechanisms driving this variation within functional groups.

### 3.3 Results and discussion

#### 3.3.1 Functional stability contrasts with taxonomic variability

We found that the metabolic functional potential of tank prokaryotic communities, as measured by gene abundance profiles, is consistent across bromeliads (Fig. 3.2A,B). This consistency in metabolic functional potential is presumably promoted by strong stoichiometric balancing between coupled metabolic pathways, the majority of which serve to break down large organic compounds to simpler organic molecules and gradually move electrons from reduced organic carbon to terminal electron acceptors such as protons (H⁺), carbon dioxide (CO₂), sulfate (SO₄²⁻), nitrate (NO₃⁻) and oxygen (O₂) (56). These metabolic pathways are distributed across multiple organisms and link the breakdown of dead organic matter captured in the bromeliads to the eventual release of CO₂ (22), methane (CH₄)(304) and presumably molecular nitrogen (N₂). Each step along these pathways thus appears to sustain highly constrained microbial productivities, resulting in specific proportions of metabolic functional groups that are conserved across bromeliads.

On the other hand, we find that the taxonomic composition within individual functional
groups is highly variable across bromeliads, both in terms of the occurrence of operational
taxonomic units (OTU, at 99% 16S rDNA similarity, see the Methods for justification) as
well as their relative abundances (Figs. 3.2C–K). For example, within any given functional
group, any two bromeliads share only ∼20–60% of their OTUs (Table B.1), and this overlap
is significantly lower than would be expected solely due to insufficient sampling effort (
P < 0.001). In fact, within any given functional group, OTUs detected in all of the samples
(“core microbiome”) only make up ∼0–1% of total OTUs across all samples (“regional pool”).
Further, coefficients of variation for OTU proportions within functional groups are typically
an order of magnitude higher than coefficients of variation of relative gene abundances (∼2–3
vs ∼0.2–0.6, respectively; Table B.2). This taxonomic variability within functional groups
persists to a considerable extent even when OTUs are combined at higher taxonomic levels
e.g., genus, family, order or class level; Figs. B.1, B.2, B.3 and B.4) and is in strong
contrast to the much more stable relative gene abundances. Hence, in each bromeliad the
same metabolic niches appear to be occupied by very different species assemblages, even if the
occupancy of each niche — in terms of its relative abundance — remains almost unchanged.
This variability within metabolic niches explains the previously observed strong variation
in overall microbial community composition across bromeliads (117) and underlines the fact
that high taxonomic variability between replicate ecosystems need not imply differences in
metabolic function.

3.3.2 Causes of variation within functional groups

The strong taxonomic variability within functional groups is presumably enabled by a high
functional redundancy in the regional microbial species pool (Fig. 3.3), allowing for potential
colonization of each bromeliad by multiple metabolically similar OTUs. Although we do not
yet know the precise mechanisms determining the subset of OTUs that eventually establish
in each bromeliad and within each metabolic niche, we can discount certain explanations.
For example, random population drift combined with random dispersal within the sampled
area would result in negligible associations between OTUs and be perceived as a random
subsampling of the regional OTU pool. To test this scenario for each functional group, we
compared OTU co-occurrences, as defined by their “C-score” (a measure for mutual OTU
segregation, averaged over all OTU pairs; 159), to a null model corresponding to random
OTU sampling from the functional group’s regional pool. The null model preserved the
total number of OTUs per sample and per functional group as well as the total number of
samples containing each OTU, in order to avoid spurious co-occurrence patterns caused by
differences in OTU richness or OTU frequency. Within 6 out of 9 functional groups (aero-
bic chemoheterotrophs, cellulose degraders, fermenters, nitrogen respirers, photoautotrophs and sulfate respirers), we find that OTUs are significantly co-segregated with respect to each other, that is, C-scores are higher than expected by chance ($P < 0.05$, Table B.3). The remaining functional groups also display OTU segregation, although differences from the null model are not statistically significant. This general segregation of OTUs beyond the null expectation rules out random subsampling and drift as important causes of OTU turnover within functional groups. We note that, when combined with spatially limited dispersal, neutral population drift could in principle produce non-random OTU co-occurrence patterns (490), because bromeliads in greater proximity would tend to exhibit more similar (i.e., correlated) community composition. However, spatially limited dispersal is likely not important at this scale: We did not find any significant correlations between geographical distance and community dissimilarity for any of the functional groups and for any of the considered dissimilarity metrics (Mantel tests with Spearman rank correlations; Fig. B.5; detailed results in Table B.4). In fact, bromeliads at opposite ends of our study site often contained more similar communities than immediately adjacent bromeliads (Fig. B.6). These results are consistent with previous work that found negligible effects of spatial distance on bacterial, zooplankton and macroinvertebrate communities in bromeliads at similar spatial scales (117). Hence, the OTU co-occurrence patterns observed here likely reflect a deterministic mutual exclusion between OTUs that is potentially caused by environmental filtering or biotic interactions (159), rather than spatially correlated or uncorrelated random assembly.

To further verify the importance of deterministic assembly processes, we examined the phylogenetic structure within functional groups in each sample. Specifically, within any given functional group, we assessed whether OTUs found in the same sample tend to be phylogenetically underdispersed (“clustered”) or overdispersed in terms of their mean phylogenetic distance, when compared to the expectation based on random OTU sampling from the regional pool of that functional group. Conventionally, underdispersion is interpreted as a sign of environmental filtering acting similarly on closely related clades (196), while overdispersion is interpreted as a sign of increased competition between close relatives, although alternative mechanisms, such as specialist predation by phages, may also create non-random patterns (362). Of the 9 functional groups, we find that 4 show a significant tendency towards underdispersion and 2 functional groups demonstrate overdispersion ($P < 0.05$, Fig. 3.4). The detection of a statistically significant phylogenetic structure in 6 out of 9 functional groups is unlikely the result of a false positive detection rate ($P < 0.000001$). This supports our previous interpretation that community assembly is generally not random within func-
Functional stability and taxonomic variability in bromeliad microbiomes

tional groups, but is subject to deterministic processes that are sensitive to phylogenetic relationships. Note that an absence of phylogenetic structuring, on the other hand, does not rule out deterministic processes (362). Moreover, the fact that some groups exhibit phylogenetic underdispersion, while others exhibit phylogenetic overdispersion, suggests that different ecological processes influence phylogenetic structure in different functional groups. In particular, the strong overdispersion of methanogens as well as methylotrophs suggests that competition may correlate strongly with relatedness in these groups and that other factors, such as environmental filtering, may be unimportant or only weakly correlate with phylogeny. On the other hand, frequent horizontal transfer of genes for the degradation of particular organic compounds reduces the correlation between phylogenetic relatedness and metabolic similarity in aerobic chemoheterotrophs and fermenters (308). This may explain why in these two functional groups mechanism causing underdispersion — rather than overdispersion — seem to dominate.

The above findings suggest that, although taxonomic composition within functional groups is highly variable, it is not random in terms of OTU co-occurrences or phylogenetic relationships. This determinism might be caused by environmental filtering, by biotic interactions, or by a combination of these, such as trade-offs between environmental stress tolerance and competition (159, 362). For example, recent work demonstrates that microbial communities can exhibit complex but deterministic responses to extremely weak environmental fluctuations (130), and that environmental turnover — when carefully characterized — can explain community turnover (371, 450). To determine whether environmental filtering partly drives the variation in OTU composition within functional groups, we examined the predictive ability of an extensive set of physicochemical variables (overview in Table B.5). We considered standard limnological variables such as pH, salinity and multivariate characterization of dissolved organic carbon, as well as other potentially important variables such as detrital volume and vegetative cover (“shading”). Many of these variables are known to influence macroinvertebrate communities in bromeliads (102, 279, 296). We find that a subset of environmental variables — including pH, salinity, detrital volume and shading — exhibit high and statistically significant correlations to relative OTU abundances within several functional groups, suggesting that these variables may be particularly influential (Fig. 3.5A). Regression models generally exhibit low to moderate predictive power when compared against novel data, as indicated by cross-validated coefficients of determination ($R^2_{cv}$), although we note that predictive power varies greatly among different OTUs (Fig. 3.5B). We therefore conclude that the environmental variables considered here can explain some of the variation within functional groups, but that additional factors are also important. In fact, the predictive
power of environmental variables is similarly low within functional groups showing either phylogenetic underdispersion or overdispersion (Fig. 3.4), indicating that the non-random phylogenetic community structure may be shaped mostly by biotic interactions rather than environmental filtering.

Taken together, our results suggest that, in addition to environmental filtering, biotic interactions also play a significant role in shaping these communities while maintaining functional similarity across bromeliads. The potential importance of biotic interactions, such as competitive exclusion, predation by phages or protists as well as metabolic interdependencies, in shaping microbial communities has been pointed out previously (258, 276, 455). For example, adaptation of bacteriophages to specific hosts can strongly influence bacterial species composition (423) and promote spatial as well as temporal variation of microbial communities (397, 462). Consequently, seemingly random taxonomic variation across locations may result from biotic interactions driving complex but deterministic population dynamics or, alternatively, from biotic interactions generating complex community responses to subtle environmental variation (130, 450, 451). This conclusion is consistent with previous findings that the distribution of cyanobacterial taxa across coexisting bromeliads was driven both by physicochemical factors as well as protozoans and invertebrates (63). Here we have not considered possible effects of invertebrates, although we note that we detected almost no insects in the sampled bromeliads. Further, given the available data it is impossible to determine whether microbial communities within single bromeliads exhibited high temporal variability, or if communities were near steady state. Previous work shows that even strongly controlled engineered ecosystems can exhibit high temporal fluctuations in microbial taxonomic composition (122, 349) and that these fluctuations can be deterministic (130, 495). Hence, the highly variable community profiles observed here could be mere “snapshots” along similar successional trajectories far from steady state (285).

3.4 Conclusions

Here we have shown that replicate natural ecosystems exhibit highly variable taxonomic composition of bacterial and archaeal communities, despite very similar metabolic functional structure. This points to a fundamental and important difference between functional and taxonomic community structure, which arises because mechanisms leading to a convergence of functional structure (e.g., nutrient limitation, stoichiometric balancing between coupled metabolic pathways) do not necessarily lead to a convergence of taxonomic composition. Reciprocally, strong taxonomic turnover may only weakly affect ecosystem
functioning (254, 517; but see 454). We suggest that functional community profiles, either based on gene-centric metagenomics (255, 381) or on a functional classification of detected taxa (289), should be the baseline of microbial biogeographical studies, particularly in cases where geochemical gradients shape microbial niche distribution (381, 387). The residual variation within individual functional groups can then be extracted and analyzed separately, as demonstrated here, in order to elucidate additional community assembly processes that act in superposition to metabolic niche effects. Our analysis suggests that in bromeliad tank ecosystems the variation within individual functional groups is the result of multiple deterministic processes, including environmental filtering and biotic interactions, while random processes such as dispersal limitation or neutral population drift (427) appear to be less relevant. This is in line with recent work on the global ocean microbiome (289) and suggests a general paradigm for microbial ecology. A careful distinction between functional composition and taxonomic composition within functional groups thus enables deeper insight into microbial community assembly processes and will be an important step towards a truly mechanistic microbial ecology.

Figure 3.1: Bromeliad species used in this study Large picture: Aechmea nudicaulis, the bromeliad species considered in this study. The foliage forms a deep central cavity (“tank”, small picture) that accumulates rainwater and dead organic material, such as leaves from nearby trees. The decomposition of this material sustains a highly productive and diverse food web inside the tank.
Figure 3.2: **Functional stability vs taxonomic variability.** (A) Relative abundances of proxy genes in prokaryotic metagenomic sequences (genes grouped by function, one color per gene group, one column per sample). For details on associating genes with functions see the Methods. (B) Subplot of (A) focusing on the rarer genes for better illustration. (C–K) Prokaryotic OTU proportions within individual functional groups (one color per OTU, one column per sample, one plot per functional group), as determined from 16S rDNA sequences. Due to ambiguities in gene function, for some functional groups (D, H) we considered multiple proxy genes. For each functional group, proxy genes are indicated via color codes (corresponding to colors in A and B) next to the functional group’s name. For more detailed metagenomic profiles see Supplemental Fig. B.10. For the taxonomic composition within functional groups at higher taxonomic levels (genus, family or order) see Figs. B.1, B.2 and B.3.
Figure 3.3: **Functional redundancy in the regional OTU pool.** Associations of functional groups (rows) with OTUs (columns), indicated by blue table cells. Functional groups are sorted according to their number of OTUs (indicated in brackets). Some OTUs were associated with more than one functional group. For analogous plots at the genus, family and class level, see Figs. B.11, B.12 and B.13, respectively.

Figure 3.4: **Phylogenetic dispersion.** Standardized effect sizes (SES) of mean phylogenetic distances between OTUs within individual functional groups (one point per sample, one row per functional group), as an indicator of phylogenetic overdispersion (SES > 0) or underdispersion (SES < 0). The vertical line at zero corresponds to the expectation under the null model, and is shown for reference. Functional groups displaying statistically significant overdispersion or underdispersion (i.e., a strong tendency towards positive or negative SES across samples, respectively) are highlighted in bold (P-values are given in brackets).
Figure 3.5: Relating OTU proportions to environmental variables. (A) Average magnitude (i.e., average absolute value) of correlations between OTU proportions within functional groups and measured environmental variables (one column per environmental variable, one row per functional group). A larger and darker circle corresponds to a larger average absolute correlation, and indicates a stronger relation between an environmental variable and a functional group’s taxonomic composition. Statistically significant correlations ($P < 0.05$) are written inside the circles. (B) Distribution of cross-validated coefficients of determination ($R_{cv}^2$, a measure for a model’s predictive power) for regression models of OTU proportions within each functional group using environmental variables as predictors (one box per functional group). Horizontal bars comprise 95% of the $R_{cv}^2$ across OTUs. The vertical grey line at zero is shown for reference.
Chapter 4

Calibration and analysis of cell-metabolic models for microbial ecology

4.1 Synopsis

Microbial ecosystem modeling is complicated by the large number of unknown parameters and the lack of appropriate calibration tools. Here we present a novel computational framework for modeling microbial ecosystems, which combines genome-level cell models into a microbial community in which the metabolic activity of each cell can affect the shared metabolite pool and thus potentially the metabolism of other cells. The framework, which we called MCM (“Microbial Community Modeler”), automates statistical analysis and model calibration to experimental data. To demonstrate the potential of MCM, we examined the dynamics of a community of *Escherichia coli* strains that emerged in laboratory evolution experiments, during which an ancestral strain diversified into two coexisting ecotypes. We constructed a microbial community model comprising the ancestral and the evolved strains, which we calibrated using separate monoculture experiments. Simulations reproduced the successional dynamics in the evolution experiments, and pathway activation patterns observed in microarray transcript profiles. Our approach yielded detailed insights into the metabolic processes that drove bacterial diversification, involving acetate cross-feeding and competition for organic carbon and oxygen.

4.2 Introduction

Metabolic interactions are an emergent property of microbial communities (70, 333). Even the simplest life forms can only be understood in terms of biological consortia characterized by shared metabolic pathways and distributed biosynthetic capacities (200, 238, 311).

---

1A version of this chapter has been published (see the Preface for author contributions): Louca, S., Doebeli, M. 2015. *Calibration and analysis of genome-based models for microbial ecology*. eLife. 4:e08208. DOI:10.7554/eLife.08208
For example, glucose catabolism to carbon dioxide or methane is a multi-step process often involving several organisms that indirectly exchange intermediate products through their environment (442). Microbial communities are thus complex systems comprising several interacting components that cannot be fully understood in isolation. In fact, metabolic interdependencies between organisms are at least partially responsible for our current inability to culture the great majority of prokaryotes (410). Understanding the emergent dynamics of microbial communities is crucial to harnessing these multicomponent assemblages and using synthetic ecology for medical, environmental and industrial purposes (44).

Genome sequencing has enabled the reconstruction of full-scale cell-metabolic networks (184), which have provided a firm basis for understanding individual cell metabolism (107, 238, 496). Recent work indicates that multiple cell models can be combined to understand microbial community metabolism and population dynamics (70, 177, 238, 453, 543). These approaches assume knowledge of all model parameters such as stoichiometric coefficients, maintenance energy requirements or extracellular transport kinetics, a requirement that is rarely met in practice (118, 177). Experiments and monitoring of environmental samples could provide valuable data to calibrate microbial community models, e.g., via statistical parameter estimation, but appropriate tools are lacking. So far, the standard approach has been to obtain each parameter through laborious specific measurements or from the available literature, or to manually adjust parameters to match observations (70, 177, 292). Furthermore, statistical model evaluation and sensitivity analysis is typically performed using ad-hoc code, thus increasing the effort required for the construction of any new model. Consequently, the experimental validation of genome-based microbial community models and their application to biological questions are rare (177, 318).

We have developed MCM (Microbial Community Modeler), a mathematical framework and computational tool that unifies model construction with statistical evaluation, sensitivity analysis and parameter calibration. MCM is designed for modeling multi-species microbial communities, in which the metabolism and growth of individual cell species is predicted using genome-based metabolic models. Cells in the community interact in a dynamical environment in which metabolite concentrations and other environmental variables influence, and are influenced by, microbial metabolism. Unknown model parameters can be automatically calibrated (fitted) using experimental data such as cell densities, nutrient concentrations or rate measurements. To demonstrate the potential of MCM, we modeled a bacterial community that has emerged from in-vitro evolution experiments, during which an ancestral strain repeatedly diversified into two distinct ecotypes. Experiments with microbes have an established tradition as model systems for understanding ecological and evolutionary processes.
We show that the predictions derived from MCM are in very good agreement with the outcomes of several monoculture and co-culture experiments. While the experimental results described below have been found over the course of several years (132, 188, 262, 436), it is only now that a mechanistic model has managed to unify them in a clear, unambiguous and synergistic manner. The analysis presented here thus provides a unifying quantitative interpretation for a large body of experimental work performed in our lab over the course of roughly a decade.

4.3 Model overview

In MCM, a microbial community model is a set of differential equations for the population densities of the cell species comprising the community and of the ambient concentrations of utilized nutrients (metabolites), coupled to optimization problems for the cell-specific rates of reactions involving these metabolites. Each cell is characterized by its metabolic potential, that is, the genetically determined subset of reactions it can catalyze, as well as any available metabolite transport mechanisms. The reaction rates and metabolite exchange rates (i.e., the metabolism) of each cell are assumed to depend on its metabolic potential as well as on the current environmental conditions, such as metabolite concentrations. Through their metabolism, in turn, cells act as sinks and sources of metabolites in the environment. Additional metabolite fluxes, such as oxygen diffusion from the atmosphere into the growth medium of a modeled bacterial culture, can be included in the model.

At any point in time, individual cell metabolism is determined using flux balance analysis (FBA) (354), a widely used framework in cell-metabolic modeling (70, 107, 129, 238, 496). In FBA, cell metabolism is assumed to be regulated in such a way that the rate of biosynthesis is maximized (119, 496). The chemical state of cells is assumed to be steady, leading to stoichiometric constraints that need to be satisfied for any particular combination of intracellular reaction rates. Reaction rates, on the other hand, are limited due to finite enzyme capacities. Metabolite uptake/export rates can also be limited due to finite diffusion rates or limited transmembrane transporter efficiency. For example, uptake rates can be Monod-like functions of substrate concentrations (177, 292). Taken together, cell-metabolic potential, stoichiometric consistency, reaction rate limits and transport rate limits define the constraints of a linear optimization problem for each cell species at each point in time. The optimized biosynthesis rate is translated into a cell production rate by dividing by the cell’s mass, thus defining the species’ population growth (Fig 4.1).
The central assumption of individual cells maximizing biosynthesis, subject to environmental and physiological constraints, is rooted in the idea that evolution has shaped regulatory mechanisms of unicellular organisms in such a way that they strive for maximum growth whenever possible. Biosynthesis has been experimentally verified as an objective for *Saccharomyces cerevisiae* and *Escherichia coli* (51, 146, 176). The assumption of maximized biosynthesis is less valid for genetically engineered organisms or those exposed to environments that are radically different from the environments that shaped their evolution (417). Despite its limitations, FBA has greatly contributed to the understanding of several genome-scale metabolic networks and metabolic interactions between cells (70, 129, 177, 238, 354, 453).

One advantage of FBA models over full biochemical cell models is their independence of intracellular kinetics and gene regulation, which limits the number of required parameters to stoichiometric coefficients and uptake kinetics.

The combination of FBA with a varying environmental metabolite pool, as implemented by MCM, is known as dynamic flux balance analysis (DFBA) (70, 177, 292). In contrast to conventional FBA, DFBA models are dynamical because cell densities and environmental metabolite concentrations both change with time, and the rate of change of each cell density and metabolite concentration depends on the current cell densities and metabolite concentrations (177, 292). Because metabolites can be depleted or produced by several cell species, the environmental metabolite pool mediates the metabolic interactions between cells (410). For example, oxygen uptake rates might depend on environmental oxygen concentrations, which in turn are reduced by cellular respiration. Similarly, cells might excrete acetate as a byproduct of glucose catabolism, which then becomes available to other cells. The metabolic optimization of individual cells striving for maximal growth, while modifying their environment, leads to non-trivial community dynamics that can include competition, cooperation and exploitation. The cell-centric nature of DFBA differs fundamentally from other flux balance analyses of microbial communities that assume an optimization of a community-wide objective such as total biomass synthesis (239, 453, 545). Such an assumption is, however, questionable from an evolutionary perspective and likely not appropriate for natural communities comprising several species. Community-level optimality will typically conflict with optimality for individual competing lineages, and configurations that optimize overall biosynthesis at the expense of individual “cooperators” would be vulnerable to exploitation (327).

Recent work suggests that DFBA is a promising approach to microbial ecological modeling (70, 177, 318). For example, Harcombe et al. (177) designed a computational tool (COMETS) based on DFBA, which was able to accurately predict equilibrium compositions.
of mixed bacterial cultures grown on petri dishes. However, COMETS offers limited model versatility in terms of uptake and reaction kinetics and only has few environmental feedback mechanisms (namely, varying extracellular metabolite concentrations). Furthermore, it assumes complete knowledge of all required model parameters and provides no generic statistical model analysis. Hence, while COMETS sets an important precedent, considerable work is still needed to make DFBA a practical approach in microbial ecosystem modeling. MCM extends Harcombe et al.’s framework to more versatile microbial ecological models that include arbitrary reaction kinetics (e.g., subject to product-inhibition) as well as dynamical environmental variables (e.g., pH) that influence, and are influenced by, microbial metabolism. In addition, MCM supports cell models in which internal molecules act as dynamical constraints that further restrict the FBA solution space, for example to account for post-transcriptional regulation or delays in enzyme synthesis (38). These so called regulatory FBA models have been shown to improve the fidelity of conventional FBA models for *E. coli* and *S. cerevisiae* (81–83, 187), however their application to microbial communities remains untested. MCM can statistically evaluate models against data, analyze their sensitivity to varying parameters (61), and estimate the uncertainty of model predictions in the face of stochasticity (171). Perhaps most importantly, MCM can automatically calibrate unknown model parameters to data, for example obtained from monoculture experiments (as demonstrated below), from bioreactor experiments involving multiple species (285) or from environmental samples of unculturable communities (Fig 4.2; see section C.1 for details). MCM can thus be used to understand the dynamics of realistic microbial ecosystems, ranging from the soil or groundwater to mixed laboratory cultures and bioreactors.

### 4.4 Results and discussion

#### 4.4.1 Successional dynamics of a microbial community

In a series of laboratory evolution experiments with *E. coli* (strain B REL606; 538) in glucose-acetate supplemented medium, two metabolically distinct strains consistently evolved from the ancestral (A) strain (188, 262, 437). When grown in monoculture with the same medium composition, all three strains exhibit diauxic growth curves with a fast glucose-driven growth phase followed by slower growth on acetate. However, the three strains differ in their efficiencies to catabolize glucose and acetate: Strain SS (slow switcher) is a better glucose utilization when compared to strain A, and the depletion of glucose only leads to a slow switch to acetate consumption. On the other hand, the FS (fast switcher) strain has evolved to be a better acetate utilization, initiating acetate consumption at higher remnant glucose concen-
tations than strains A and SS. This acetate specialization is based on a tradeoff in the citric acid cycle and comes at the cost of being a less competitive glucose consumer.

Replicated serial dilution experiments starting with strain A monocultures have shown a consistent phenotypic diversification, involving an initial invasion of the SS phenotype and a subsequent invasion of the FS phenotype, leading to the eventual extinction or near-extinction of the ancestor and the stable coexistence of the SS and FS phenotypes (Fig 4.3) (188, 262, 437, 488). Genome sequencing revealed that this metabolic diversification can be attributed to point-mutations in genes linked to glucose and acetate uptake kinetics and metabolism (188). The successional dynamics of the three phenotypes are thus likely driven by adaptations to a changing metabolic niche space, defined by fluctuating glucose, acetate and, potentially, oxygen availabilities (188, 262, 488). An understanding of the underlying ecological processes would shed light on the ecology and evolution of natural microbial communities with shared catabolic pathways.

To mechanistically explain the observed community dynamics, we used MCM to construct a model comprising the ancestral and the two evolved *E. coli* types. By keeping track of pathway activation, cell densities, metabolic fluxes and nutrient concentrations, we gained detailed insight into the processes driving the successional dynamics of metabolic diversification.

### 4.4.2 Experimental calibration

Based on a published cell-metabolic template for the ancestral *E. coli* strain comprising over 2000 reactions (538), we first constructed three separate cell models for the phenotypes A, SS and FS, respectively. In these preliminary models, cells grew on a substrate pool that resembled previous batch-fed monoculture experiments with glucose-acetate supplemented minimal medium (262). Cell-specific oxygen, acetate and glucose uptake rate limits were Monod-like functions of substrate concentrations (114, 325). We calibrated several physiological parameters for each cell type to measured chemical concentration and cell density profiles, using least squares fitting (Fig 4.4). MCM automatically calibrates free parameters to data through an optimization algorithm that involves step-wise exploration of parameter space and repeated simulations (see Appendix C.1.2).

We then constructed the microbial community (MC) model by combining the three calibrated cell models into a community growing in a common substrate pool. The environmental context resembles Herron and Doebeli’s evolution experiments (188). In particular, the model
includes realistic oxygen depletion-repletion dynamics (167), glucose and acetate depletion by microbial consumption, as well as daily dilutions into fresh glucose-acetate supplemented medium at a factor 1:100. The microbial community initially consists mostly of type A ($10^{10}$ cells/L), while both SS as well as FS cells are assumed to be rare (1 cell/L). Because the model is deterministic, the invasion or extinction of each type only depends on its growth rate in a possibly changing environment, but not on random mutation events, nor on demographic stochastic fluctuations.

### 4.4.3 Predicting microbial community dynamics

Simulations of the MC model reproduced the successional dynamics observed in Herron and Doebeli’s experiments: An initial replacement of the ancestor by the SS type is followed by an invasion of the FS type, leading to the eventual coexistence of the SS and FS types and extinction of the ancestral strain (Fig 4.5A). Interestingly, FS can also invade in the absence of SS, however invasion occurs much slower and FS reaches lower densities than in the presence of SS (Supplemental Fig. C.1). This is consistent with an early presence of the FS lineage at low densities in the evolution experiments (Fig 4.3), indicating that some of the first FS mutations already confer a slight advantage over the ancestor when FS is rare (188).

Time series of acetate concentrations (Fig 4.5B) link the observed successional dynamics of the three types to a gradually changing metabolic niche space: The replacement of type A by the more efficient glucose specialist SS leads to an accumulation of acetate and facilitates the invasion of the FS type. The specialization of the SS and FS types on glucose and acetate, respectively (Fig 4.6A), enables their long-term coexistence on glucose-acetate enriched medium through frequency dependent competition (132, 188, 262). In fact, cell-specific acetate exchange rates reveal that the SS type temporarily excretes acetate during short intervals, which is concurrently and subsequently consumed by the FS type (Fig 4.5G). This periodic acetate cross-feeding is an evolutionarily emergent property of the microbial community (484). The temporary production of acetate by the SS type is consistent with previous SS-FS co-culture experiments, which revealed slightly increased acetate concentrations towards the end of the SS exponential growth phase (436). An evolved increase of acetate excretion by E. coli in glucose minimal medium has also been reported by Harcombe et al. (176).

It should be noted that cell metabolism depends on substrate concentrations and is subject to strong temporal variation. In particular, acetate excretion by SS cells correlates strongly
with oxygen limitation (Figs. 4.5G,K). The excretion of acetate by *E. coli* as a byproduct of oxygen-limited glucose catabolism has been observed experimentally and explained using flux balance analysis (292). In the absence of oxygen limitation, complete aerobic glucose catabolism to carbon dioxide is preferred over incomplete glucose catabolism with acetate excretion. On the other hand, oxygen limitation leads to an energetic tradeoff between complete glucose catabolism and efficient oxygen utilization, resulting in the excretion of acetate.

Furthermore, the depletion of oxygen during cell growth makes oxygen a temporary limiting resource for all cells (Fig 4.5K). Shortly after dilution into fresh medium, the exponential growth of the SS type on glucose leads to a rapid drop of oxygen to nanomolar concentrations. Despite oxygen diffusion into the medium, oxygen remains at sub-saturation levels for several more hours because the slow-growing acetate-consuming FS cells still consume oxygen after the growth of SS cells has halted. Differences in SS and FS growth rates (Figs. 4.5C,E) thus mitigate competition for oxygen through temporal niche separation. Hence, oxygen likely plays an important role in the metabolic diversification, as previously hypothesized by Le Gac et al. (262). This shows that the splitting of metabolic pathways across specialists can be caused by the composite effects of competition for electron donors and electron acceptors.

Consistent with differential substrate usage, average cell-specific reaction rates (Fig 4.6B) reveal differences in pathway activation: The transformation of acetate into acetyl-CoA by acetyl-CoA synthetase (*acs*) is predicted to be decreased in type SS and increased in type FS, when compared to the ancestral type. Furthermore, the conversion of phosphoenolpyruvate to oxaloacetate (*ppc*), the conversion of phosphoenolpyruvate to pyruvate (*pyk*) and the decarboxylation of pyruvate to acetyl-CoA (*pdh*), linking the glycolysis pathway to the citric acid cycle, are all predicted to be upregulated in the SS type when compared to the FS type. Similar differences in pathway activation are also predicted during early exponential growth in monoculture (Fig 4.6C,D), because FS grows partly on acetate and SS excretes acetate (Fig 4.4F,J). Previous microarray profiles of mRNA concentrations during exponential growth in monocultures (262) found an upregulation of acetate consumption genes in FS and acetate excretion genes in SS compared to A, qualitatively confirming our predictions (Fig 4.6C,D). Interestingly, our simulations suggest a significant downregulation of glucose catabolism (*pyk*, *pdh* and *ppc*) in FS compared to A, which contradicts the transcript profiles (Fig 4.6D). This discrepancy may be explained by the fact that mRNA was harvested from well-aerated flasks, while the monoculture experiments (Fig 4.4) and evolution experiments (Fig 4.3) were performed in test tubes where oxygen can become limiting (10). Oxygen becomes particularly scarce in the FS tubes (Fig 4.4K) and temporarily limits glucose catabolism,
which would explain the strong downregulation not reflected in the transcript profiles (262). Furthermore, while broad pathway activation patterns could be qualitatively reproduced in our system, this might be harder in other cases due to post-transcriptional regulation or post-translational modifications (38).

The periodic (seasonal) changes in glucose and acetate concentrations in batch culture have previously been shown to promote coexistence of the SS and FS types, in analogy to the maintenance of phytoplankton diversity via fluctuations of resource availability (432, 436). Experiments with SS-FS batch co-cultures revealed that the SS type quickly dominates over the FS type, when restricted to the first glucose-rich season through frequent dilution into fresh growth medium. Reciprocally, when SS and FS are grown in solution resembling the second glucose-depleted acetate-rich season, the FS type quickly dominates over the SS type (436). Accordingly, in a full batch cycle the relative SS cell density has been shown to culminate within 4-8 hours and to gradually decrease afterwards (132, Fig 6B), consistent with our simulations (Fig 4.5D). Simulations of the SS and FS batch co-culture restricted to the first or second season, analogous to Spencer et al.’s experiments, reproduce these observations and verify the role of periodic variation of glucose and acetate concentrations in maintaining the coexistence of both types (Fig 4.7, see Appendix C.1 for details).

4.5 Conclusions

The models presented here make detailed predictions about the microbial dynamics in the considered experiments. First, after calibration the cell models largely explain the data from the monoculture experiments (Fig 4.4). Second, the predictions for pathway activation in the three strains (Fig 4.6) are qualitatively consistent with most transcription profiles. Third, simulations of the microbial community consisting of all three strains (Fig 4.5) reproduce the successional dynamics of diversification observed in the evolution experiments (Fig 4.3). Fourth, simulations of the SS-FS co-cultures restricted to either the glucose-rich or glucose-depleted season reproduce the dominance of the SS or FS type (Fig 4.7), respectively, in consistence with previous co-culture experiments. It is important to note that only data from monoculture experiments were used to calibrate the cell models for the three strains (A, SS and FS). In particular, no information from co-culture experiments was used in the setup of the microbial community model, and thus there was no a priori knowledge about what the emergent community dynamics would be. Hence, our work conceptually produced non-trivial predictions that could be compared to experimental observations, although all experiments had already been performed.
Our work sheds light on the fundamental problem of metabolic diversification and the emergence of shared catabolic pathways. In particular, our microbial community model allowed quantitative predictions for the metabolic fluxes for each strain in co-culture, revealing temporary cross-feeding as an emergent property of the evolved community (484). Cross-feeding, conventionally seen as a beneficial interaction (333), thus emerged as a form of niche segregation driven by competition for organic carbon and oxygen. Because both evolved types prefer glucose whenever available at high concentrations, but exchange acetate under oxygen limitation, the community constantly switches between competitive and beneficial interactions. Natural microbial populations might thus also oscillate between negative and positive interactions, for example depending on oxygen levels.

The models considered here were completely deterministic, in the sense that the growth and metabolic activity of each strain were completely determined by the conditions in the test tubes. In particular, both evolved strains were included in the simulations right from the start at low densities, while the invasion or extinction of individual strains was contingent upon their growth rates in an environment that changed in response to the activity of each strain. Our findings thus support previous suggestions that microbial evolution can be driven by deterministic ecological processes (188, 358, 530). In this case, the observed diversification is due to competition for limiting resources whose use is constrained by basic metabolic tradeoffs. Other instances of ecological diversification in microbial evolution experiments, e.g., as reported by Plucain et al. (367), might be explained using a similar approach. We emphasize that at longer time scales and in more diverse communities evolution may not be as predictable as here, because horizontal gene transfer and rare but complex mutations could introduce substantial stochasticity (39, 348). The transitions in community structure and activity following such rare events may still be understood in a deterministic framework such as ours.

We have demonstrated how MCM can be used to experimentally calibrate and combine genome-based cell models to predict the emergent dynamics of microbial communities. Our framework thus provides a starting point for designing microbial communities with particular metabolic properties, such as optimized catabolic performance. While MCM is designed for genome-based metabolic models, it can also accommodate conventional functional group models. In these models, different ecological functions such as photosynthesis, heterotrophy or nitrification are performed by distinct populations whose metabolic activity is determined, for example, by Michaelis-Menten kinetics and whose growth is described by simple substrate-biomass yield factors (Chapter 6; 193, 386). Hence, natural microbial communities could be modeled even if annotated genomes are not available for each member species.
While functional group models generally require fewer parameters, their calibration remains a challenge (360). In MCM, model calibration becomes analogous to coefficient estimation in conventional multivariate regression and can be used to estimate poorly known parameters such as stoichiometric coefficients, growth kinetics or extracellular transport coefficients. To our knowledge, no existing comparable framework offers the flexibility combined with the statistical functionality of MCM. In view of the increasing availability of genome-scale metabolic models (118), our work provides a missing link to a predictive and synthetic microbial ecology.

Figure 4.1: (A) Conceptual framework used by MCM. Cells (colored shapes) optimize their metabolism for maximal growth and influence their environment via metabolite exchange (small colored arrows). Additional external fluxes can also affect the environment (large grey arrows). The environment, in turn, influences each cell’s metabolism. (B) Computational framework used by MCM. Each iteration consists of four steps: flux balance analysis (FBA) is used to translate cell-metabolic potentials and environmental conditions (1) into a linear optimization problem for the growth rate of each cell species (2). The set of possible reaction rates corresponds to a polytope in high-dimensional space. Solving the optimization problems (3) yields predictions on microbial metabolite exchange rates (4). Metabolic fluxes and cell growth rates are used to predict metabolite and cell concentrations in the next iteration (1).
Figure 4.2: Overview of MCM’s working principle and functionalities: A microbial community model is specified using human-readable configuration files in terms of metabolites, reactions, the metabolic potential of cell species and any additional environmental variables. Models with multiple ecosystem compartments are also possible. A script with MCM commands controls the analysis of the model and, if needed, its calibration using experimental data. The calibrated model can also be used to create new, more complex models (as exemplified in this chapter).
Figure 4.3: Relative cell densities of the A, SS and FS types during three replicated evolution experiments by Herron and Doebeli (188, Fig. 2), starting with the same ancestral *E. coli* strain. Within each experiment, the illustrated SS or FS lineage comprises several strains with varyingly pronounced SS or FS phenotypes, respectively. Cell generations were translated to days by assuming an average of 6.7 generations per day (188).

Figure 4.4: Calibration of *E. coli* cell models using monoculture experiments. Continuous curves: Time course of cell densities, glucose concentration, acetate concentration and oxygen concentration (columns 1–4, respectively) predicted by MCM for monocultures of strain A, SS and FS (rows 1–3, respectively) grown on glucose-acetate medium. Points are data used for model calibration and were obtained from analogous monoculture growth experiments (262). Oxygen data were not available for strain A.
Figure 4.5: **Dynamics of the *E. coli* microbial community model.** (A) Relative cell densities of the A, SS and FS types over time. (B) Acetate concentration over time. (C), (D) and (E): SS and FS cell densities, relative cell densities and growth rates over time, respectively, during stable coexistence. (F), (G) and (H): Cell-specific glucose, acetate and oxygen uptake rates over time, respectively. Negative values correspond to export. (I), (J) and (K): Glucose, acetate and oxygen concentrations over time, respectively. Diurnal fluctuations in all figures are due to daily dilutions into fresh medium. Tics on the time axes in (c–k) mark points of dilution.
Figure 4.6: Metabolic differentiation of the A, SS and FS types. (A) Predicted cell-specific net metabolite uptake rates in co-culture. (B) Predicted cell-specific reaction rates in co-culture, for *acs* (acetyl-CoA synthesis), *ack* (acetate synthesis), *pta* (acetyl phosphate synthesis), *ppc* (oxaloacetate synthesis from phosphoenolpyruvate), *pdh* (decarboxylation of pyruvate to acetyl-CoA) and *pyk* (pyruvate synthesis from phosphoenolpyruvate). Rates in (A) and (B) are averaged over all time points within the first 100 days of evolution, with reversed reactions or net metabolite export represented by negative rates. (C) and (D): Simplified model subset of *E. coli* acetate and glucose metabolism, showing pathway activations in type SS (C) and FS (D) relative to type A during exponential growth in monoculture. Non-bracketed numeric values are ratios of predicted fluxes in the evolved types over fluxes in type A. Bracketed values are ratios of mRNA harvested from monoculture experiments by Le Gac et al. (262), for comparison. A ratio of 0/0 indicates zero flux in both the evolved and ancestral type, a ratio of 1 corresponds to an unchanged flux or mRNA, a ratio of 0 corresponds to complete deactivation in the evolved type. Darker arrows indicate increased predicted fluxes in the evolved type. Flux predictions correspond to the time points of mRNA measurements, i.e., 3.5 hours after dilution for SS and 4 hours after dilution for A and SS (262).
Figure 4.7: Predicted relative cell densities of the SS and FS types in batch co-culture when restricted to either the first glucose-rich (A) or second glucose-depleted (B) season. In (A), restriction to the first season was achieved by shorter dilution periods which prevented the complete depletion of glucose. In (B), restriction to the second season was achieved by using the glucose-depleted acetate-rich solution, produced by the full-batch co-culture, as growth medium (see the Methods for details).
Chapter 5

Transient dynamics of competitive exclusion in microbial communities

5.1 Synopsis

Molecular profiling in bioreactors has revealed that microbial community composition can be highly variable while maintaining constant functional performance, in accordance with a pathway-centric paradigm for microbial ecology. Similarly, following perturbation bioreactor performance typically recovers rapidly while community composition only slowly returns to its original state. At this point we still lack a detailed understanding of the actual mechanisms causing the discrepancy between functional and compositional stability of microbial communities. Using a mathematical model for microbial competition, as well as simulations of a model for a nitrifying bioreactor, we explain these observations on grounds of slow nonequilibrium dynamics eventually leading to competitive exclusion. In the presence of several competing strains, metabolic niches are rapidly occupied by opportunistic populations, while subsequent species turnover and the eventual dominance of top competitors proceeds at a much slower rate. Hence, functional redundancy causes a separation of the time scales characterizing the functional and compositional stabilization of microbial communities. This effect becomes stronger with increasing richness, because greater similarities between top competitors lead to longer transient population dynamics.

5.2 Introduction

Microbial metabolism drives the biochemistry of virtually all ecosystems and plays a central role in industrial processes such as biofuel production and wastewater treatment (16, 116). Thus, understanding the mechanisms that shape the dynamics and metabolic performance

---

1A version of this chapter has been published (see the Preface for author contributions): Louca, S., Doebeli, M. 2015. Transient dynamics of competitive exclusion in microbial communities. Environmental Microbiology. 18:1863–1874. DOI:10.1111/1462-2920.13058
of microbial communities is of great practical importance. Experiments with bioreactors have shown that bioreactor performance can be constant despite highly variable microbial communities (506). For example, following functional stabilization, methanogenic or nitri-fying bioreactors can exhibit species turnover for several more years (122, 524). In some cases non-equilibrium community trajectories have been reproduced across replicated experiments, suggesting that the underlying processes are deterministic (24, 495). Even when communities converge to a steady composition, recovery of community composition following perturbation can take several months. This is in contrast to metabolic throughput, which recovers within a few days (144, 461). Fluctuations and the rate of stabilization of microbial communities are thus multifaceted properties that depend greatly on whether the focus is on metabolic function or taxonomic composition. An improved understanding of these properties in microbial communities is crucial for optimizing microbially driven industrial processes and interpreting the response of ecosystems to anthropogenic perturbations.

It has been hypothesized that functional redundancy and non-equilibrium population dynamics within each metabolic compartment could promote fast stabilization of performance with slow convergence of community composition (46). Here we show that temporary population dynamics leading to an eventual steady community composition via competitive exclusion can indeed last much longer than the time required for the stabilization of overall metabolic performance. We first formalize our reasoning using a microbial community model, in which multiple strains compete for a common resource. The model illustrates in a simple way how taxonomic community composition can vary almost independently of the community’s metabolic performance. We then construct a more realistic model of a nitrifying bioreactor and use simulations to demonstrate the validity of our arguments and their consistency with previous experimental observations.

5.3 Results and discussion

5.3.1 Competition for a common resource

Microbial community richness can be disproportionally high compared to the metabolic complexity of bioreactors (123, 523). In fact, the coexistence of organisms with similar metabolic function in these systems has been contrasted to the “competitive exclusion principle” (122, 178). In the case of a single limiting resource, Tilman’s competition theory predicts that at equilibrium the only persisting competitor will be the one that can survive at the lowest resource concentration (479). However ecosystems can be subject to long tran-
sient dynamics, i.e., temporary population dynamics far from equilibrium, and convergence to equilibrium might occur at much longer time scales than assumed (180). For example, slow species turnover has been suggested to be responsible for the perplexingly high diversity seen in many microbial systems (67) and, as we show here, can explain the discrepancy between functional and taxonomic stability in bioreactors.

To formalize our argument, we use a model for multiple populations competing for the same limiting resource. We focus on the transient dynamics eventually leading to steady state, where resource input is balanced by microbial consumption. The cell density of strain \( i \), denoted \( N_i \), as well as the resource concentration, denoted \( R \), are described by the following differential equations:

\[
\frac{dN_i}{dt} = N_i \left[ \beta_i \Phi_i(R) - \lambda_i \right], \quad (5.3.1)
\]

\[
\frac{dR}{dt} = f_o - \sum_i N_i \Phi_i(R). \quad (5.3.2)
\]

Here, \( f_o \) is the constant resource supply rate, \( \lambda_i \) is the decay rate of strain \( i \) in the absence of growth, \( \Phi_i(R) \) is the rate at which cells take up the resource as a function of \( R \), and \( \beta_i \) is a biomass yield factor. We assume that \( \Phi_i(R) \) increases with \( R \). For example, \( \Phi_i(R) \) could be a Monod function that increases linearly with \( R \) at low concentrations but saturates at high concentrations, an assumption often made in geobiological and bioengineering models (211). The last term in Eq. (5.3.2) is a sum over all strains accounting for resource depletion by microbial metabolism.

The growth rate of strain \( i \) is positive whenever \( R \) is greater than the threshold concentration, \( R^o_i \), defined by \( \Phi_i(R^o_i) = \lambda_i/\beta_i \). In general, the equilibrium of Equations (5.3.1) and (5.3.2) is characterized by the extinction of all but one strain, namely the strain with the lowest survival threshold \( R^o_i \). To elucidate the transient dynamics preceding this competitive exclusion, we consider the total cell density \( N = \sum_i N_i \) and the relative cell densities \( \eta_i = N_i/N \). Using the community-average growth kinetics (denoted \( \bar{\Phi}, \bar{\beta} \Phi \) and \( \bar{\lambda} \)), one can derive the dynamics

\[
\frac{d\eta_i}{dt} = \varepsilon_i \cdot \eta_i(\bar{\beta} \Phi - \bar{\lambda}) \quad (5.3.3)
\]

for the relative cell densities,

\[
\frac{dN}{dt} = N \left( \bar{\beta} \Phi - \bar{\lambda} \right) \quad (5.3.4)
\]
for the total cell density $N$, and

$$\frac{dR}{dt} = f_0 - N\Phi$$  \hspace{1cm} (5.3.5)$$

for the resource concentration $R$ (details in Appendix D). Here, the $\varepsilon_i$ account for deviations of strain $i$ growth kinetics from the community average. For example, if $N$ is growing and $\varepsilon_i$ is positive, then strain $i$ grows faster than average and thus increases in relative abundance.

As the resource is depleted, weaker competitors decay and the average growth kinetics are determined by a few remaining competitors of similar efficiency, for which the deviations $\varepsilon_i$ from the average become very small ($\varepsilon_i \ll 1$). Hence, while the dynamics of $N$ and $R$ are determined by the community-average growth kinetics (Eqs. (5.3.4) and (5.3.5)), the relative cell densities are slowed down by the factors $\varepsilon_i$. This means that while metabolic niches are quickly filled, establishing a high rate of resource uptake, some of the competing populations can coexist during prolonged transition periods until eventual competitive exclusion.

In agreement with these predictions, Gentile et al. (144) reports a quick functional stabilization and long transient periods in community composition following mechanical shock, and Vanwonterghem et al. (495) reports a gradually decreasing richness in anaerobic digesters over the course of several months following inoculation.

The probability of similar strains being present in a random inoculum, or a microbial community in general, increases with the number of strains. In particular, the expected dissimilarity between top-competitors decreases with increasing community richness. The underlying assumption is that growth kinetic parameters are bound within some natural finite range. Hence, one should expect longer transient dynamics of competitive exclusion and slower convergence to a steady community composition at higher inoculum richness.

It has been previously hypothesized that as richness increases, the variability of ecosystem functions decreases whereas the variability of individual populations increases (267, 283, 480). The proposed mechanisms typically involve stochastic fluctuations of independent populations, so that the total community biomass and functional performance become more stable when more populations contribute to them. This statistical inevitability (105), which has been criticized on grounds of interspecific interactions (481), differs fundamentally from the deterministic mechanisms explored here. Namely, competition between strains leads to a slow decay of weaker competitors, which is compensated by the growth of other populations that stabilize overall functional performance.
5.3.2 Bioreactors as model systems

The above competition model explains how populations occupying a common metabolic niche can, in principle, undergo long transient periods of coexistence. The actual duration and nature of these transients depend on the similarity between competing strains, as well as their typical intrinsic growth kinetics. To test the relevance of our predictions to realistic microbial communities, we examined a separate numerical model for a nitrifying bioreactor (524). Apart from their practical relevance to industrial processes such as sewage treatment and biofuel production (16), bioreactors are also ideal model systems for understanding microbial ecology and processes shaping microbial community structure (123, 160, 495). The bioreactor considered here is a flow-through chemostat continuously fed with ammonium (\( \text{NH}_4^+ \)), which is aerobically oxidized to nitrate (\( \text{NO}_3^- \)) in a two-step process. Oxidation occurs in a microbial community that consists of chemoautotrophic ammonium-oxidizing bacteria (AOB), which transform ammonium to nitrite (\( \text{NO}_2^- \)), and chemoautotrophic nitrite-oxidizing bacteria (NOB), which transform nitrite to nitrate. Nitrate is exported from the bioreactor as part of a continuous outflow through a filter membrane that retains cells within the bioreactor. The substrate feed rate and the hydraulic dilution rate are kept constant and in line with previous bioreactor experiments (524), allowing the establishment of a steady metabolic throughput following an initial startup period.

The bioreactor’s microbial community is modeled using differential equations for the cell population densities and the ambient ammonium, nitrite and nitrate concentrations. These metabolites are subject to microbial production and depletion, as well as physical addition and removal from the bioreactor. The metabolic activity of individual cells is determined using flux balance analysis (FBA), a widely used framework in cell-metabolic modeling (354). In FBA, the chemical state of cells is assumed to be steady, leading to stoichiometric constraints that need to be satisfied for any particular combination of intracellular reaction rates. These rates are assumed to be regulated by the cell in such a way that some objective function, commonly associated with biomass production, is maximized subject to additional constraints on substrate uptake rates (119). In our case, the optimized biosynthesis rate is translated to a growth rate by dividing by the cell mass. Ammonium and nitrite uptake rates are limited by substrate concentrations in a Monod-like fashion, thus constraining the achievable growth rates depending on the bioreactor’s chemical state (177, 292).

The assumption of cells maximizing biosynthesis, subject to environmental and physiological constraints, is rooted in the idea that evolution has shaped regulatory mechanisms to induce maximum growth whenever possible (51, 176). This assumption is less valid for genetically
engineered organisms or those exposed to environments that are radically different from the environments that shaped their evolution, and other objectives such as ATP production or metabolic efficiency have been proposed (146, 417). Biosynthesis has been experimentally verified as an objective for, among others, *Saccharomyces cerevisiae*, *Escherichia coli* and *Nitrosomonas europaea* (107, 119, 364). Despite its limitations, FBA with maximization of growth has greatly contributed to the understanding of several single-cell metabolic networks as well as metabolic interactions between cells (70, 129, 238, 354). One advantage of FBA models over full biochemical cell models is their independence of intracellular kinetics and gene regulation, which limits the number of required parameters to stoichiometric coefficients and uptake kinetics. Recent work has shown that FBA-based models with maximization of growth can accurately predict microbial community dynamics (70, 177, 284, 318).

Our bioreactor model comprises multiple AOB and NOB strains, which are constructed by randomly choosing several cell parameters around those of two template AOB and NOB models. The AOB and NOB templates were calibrated and validated beforehand using data from previous bioreactor experiments (Fig. 5.1; see Section D.1 for details). Because metabolites can be depleted or produced by several cells, the environmental metabolite pool mediates the metabolic interactions between cells (410). For example AOB deplete ammonium from their environment, rendering it a limiting resource that mediates competition between AOB strains. The excretion of nitrite as a by-product, in turn, enables the growth of nitrite-limited NOB. The metabolic optimization of individual cells striving for maximal growth, while modifying their environment, thus leads to non-trivial community dynamics that can include cooperation, competition and extinction.

### 5.3.3 Bioreactor community dynamics

Following inoculation of the bioreactor, two phases can generally be distinguished (Fig. 5.2). Initially, the concentration of inflowing ammonium increases until AOB populations have grown to sufficient densities to balance ammonium supply by ammonium consumption. The accumulation of nitrite as an AOB waste product, in turn, triggers the growth of NOB populations until nitrite production is eventually balanced by nitrite consumption. This initial startup phase is dominated by fast-growing opportunists that benefit from an excess of substrates and little competition. The duration of this phase is mainly determined by the hydraulic renewal rate, ammonium supply rate and bacterial growth rates, and the duration predicted by our model (roughly 3 weeks) is in line with typical nitrifying bioreactor experiments (109, 302).
As ammonium and nitrite consumption increase, their concentrations decrease to near or below the survival thresholds for an increasing number of strains (Fig. 5.2C). This second saturation phase is characterized by low and relatively stable substrate concentrations, stagnation of growth, a gradual extinction of less competitive strains and a long coexistence of similar top competitors (Fig. 5.2A). The microbial community slowly converges to a stable composition of decreased diversity in which each metabolic niche is occupied by a single strain, with transient periods occasionally lasting up to several thousands of days. A gradual decrease in diversity is expected under the competitive exclusion principle of equilibrium ecology (178) and is consistent with similar observations in previous bioreactor experiments (495). On the other hand, the total cell densities of metabolically similar strains (e.g., AOB) stabilize much faster and only vary weakly during the saturation phase (Fig. 5.2B). Hence, each of the two available metabolic niches is rapidly filled by several competing and temporarily coexisting strains, which are only slowly replaced by the top competitor.

Our results show how transient dynamics of competitive exclusion can lead to a separation of time scales characterizing functional and compositional stabilization of communities. This separation of time scales is also expected to be reflected in the community’s response to perturbations. Perturbations such as mechanical biofilm removal (144) or nutrient shocks (461) can alter the relative abundances of individual clades or lead to a temporary collapse of the community. Such a collapse would initiate a race for the (re-)occupation of metabolic niches and a subsequent saturation phase, analogous to the dynamics following inoculation. For example, Gentile et al. (144) observed that, after the shearing of biofilms inside a fluidized bed reactor, community composition recovered much slower, i.e., it had lower resilience (419), than the bioreactor’s performance.

Simulations of the bioreactor model including a strong pulse perturbation, applied simultaneously to the entire community, reproduced these observations (Fig. 5.3). The modeled perturbation corresponds to an increased mortality for one day, with a strength chosen randomly for each strain and resulting in a temporary collapse of the community by several orders of magnitude. Consistent with experimental observations, the bioreactor’s performance quickly recovers within a few days to weeks (Figs. 5.3C,D), while the community’s recovery to its original composition typically takes several months to years (Figs. 5.3A,E,F). Metabolic niches are reoccupied rapidly and concurrently with the bioreactor’s functional stabilization (Fig. 5.3B), however metabolic niches can be temporarily shared by several coexisting strains. Non-equilibrium processes, particularly following perturbation, are frequently thought to maintain high diversity, for example in rain forests (75) or phytoplankton (432). Furthermore, a meta-analysis by Shade et al. (419) found more studies reporting re-
covery of microbial community function than composition, following pulse perturbation.

As predicted above by our competition model, the discrepancy between functional and taxonomic stability should be stronger for communities with high richness because the likelihood of two similar top competitors increases, thus delaying competitive exclusion. Simulations of bioreactors inoculated with different numbers of random strains verify this prediction. For example, the time until compositional convergence following inoculation, i.e., reaching a 90% Bray-Curtis similarity to the steady state (266), ranges from roughly 600 days for 20 strains to 1300 days for 100 strains (median values, Figs. 5.2E,F). Moreover, richer communities are expected to be more prone to temporary changes in composition during perturbation because of a greater reservoir of opportunistic strains that could temporarily invade (524). This is reflected in our simulations, where a greater number of strains correlates with a stronger change in community composition following the pulse perturbation (Figs. 5.3E,F).

The insensitivity to disturbance is known in ecology as resistance and is, together with resilience, a common measure of community stability (419). Our work suggests that microbial communities with higher functional redundancy have lower resilience and lower resistance to pulse perturbation in terms of taxonomic composition.

5.3.4 Variable does not mean unstable

Previous bioreactor experiments have revealed variable community composition despite stable bioreactor performance over hundreds of days following inoculation (122, 495, 524, 548), while others have reported convergence to steady compositions within a few months (144, 314, 533). Fluctuating community compositions are often interpreted as unstable, non-convergent or even chaotic. However, the observed dynamics may be mere transients of slowly converging communities. Typical richness in bioreactors can range from hundreds to thousands of operational taxonomic units (OTUs, a species analog based on rDNA similarity) (234, 440, 495). As shown here, at these richness scales transient dynamics of competitive exclusion can last several years. Much longer operation times might thus be needed to actually observe an eventual community convergence in typical bioreactors. However, at these time scales other destabilizing processes, such as the invasion of new strains introduced by contamination, could prevent community convergence.
5.3.5 Model limitations

The simple models considered in this chapter focus on generic ingredients of microbial ecosystems, namely substrate-limited growth and competition, stoichiometric constraints on coexisting pathways, as well as physical substrate repletion and waste removal (e.g., in continuous-flow bioreactors). In particular, we have assumed that microbial growth increases with increasing substrate concentrations, thus ignoring the possibility of substrate inhibition. For example, substrate inhibition can occur during nitrification by excess ammonia and nitrous acid (15), resulting in reduced bioreactor performance (425). Similarly, growth may also be subject to product inhibition, e.g., when the partial pressure of accumulating waste products renders a pathway unfavorable (225). Accurately modeling specific industrial setups or natural systems may thus require a consideration of more complicated kinetics, e.g., including substrate and product inhibition. Further, metabolic niche structure in natural systems may be more complex than considered here since functional groups may partly or completely overlap, for example if a single organism performs both nitrification steps (89).

Our main point is that long transient dynamics can emerge even in the simple cases considered here, acknowledging that more complex communities are likely subject to further destabilizing mechanisms (see below).

5.3.6 Alternative destabilizing factors

Transient dynamics of competitive exclusion provide a simple explanation for the discrepancy between functional and taxonomic stability of microbial communities, and our simulations underline the relevance of these processes at least to typical bioreactor setups. However, other mechanisms likely contribute to a long-term variability of community composition. For example, time lags associated with the degradation of organic matter, such as cellulose hydrolysis in anaerobic digesters (495), can result in slow changes of the metabolic landscape and optimal electron flow, in turn driving adaptive changes in community composition (122). More complicated non-sequential pathways, ubiquitous in organic carbon catabolism, could also lead to positive feedback loops that further destabilize community dynamics. Furthermore, in contrast to well-controlled bioreactors, many natural ecosystems are subject to intense environmental variation that can drive adaptation and succession in microbial communities. For example, annual deep-water renewal in a seasonally anoxic fjord has been shown to cause significant changes in microbial community structure (540).

We emphasize that mechanisms that destabilize community composition need not necessarily destabilize community function. For example, in open systems such as wastewater treatment
Transients of competitive exclusion

Plants (506) occasional invasion by novel competitors could drive species turnover without significantly affecting ecosystem functioning, however this scenario is unlikely in bioreactors with a sterile feed (495). Similarly, repeated adaptation of bacteriophages to dominant hosts (“killing the winner” dynamics) has been shown to sustain bacterial diversity and drive continuous species turnover (423, 475). Collapsing populations could be replaced by less susceptible but functionally similar populations that ensure the overall stability of biochemical fluxes.

Reciprocally, negative feedback mechanisms stabilizing biochemical fluxes may only weakly affect community composition. For example, substrate build-up can promote the growth of functional groups benefiting from the underutilized resource, in turn counteracting the processes causing substrate build-up. This stabilizing mechanism, perhaps comparable to La Châtelier’s principle of an “opposing force” (411), a priori acts on functional groups rather than taxonomy.

5.3.7 Towards a pathway-centric microbial ecology

Marker gene-based taxonomic community profiling has become a standard approach in microbial ecology (148, 524). However, metabolic functions may be performed by several competing clades and, conversely, members of the same clade can fill separate metabolic niches (3, 308). Such irregular metabolic trait distributions across clades are caused by diverse evolutionary processes, including adaptive loss of function and metabolic convergence accelerated by frequent horizontal gene transfer (116). As a result, taxonomic profiles, at any taxonomic level, often obscure the relationship between community structure and function. For example, environmental conditions and ecological function can show a stronger correlation to particular metabolic pathways or even individual genes, than to the distribution of particular taxa (Chapter 2; 52, 349, 381). In fact, even physicochemically similar systems with similar functional community structure can exhibit markedly different taxonomic composition (Chapter 3; 122). Consistent with this, as we have shown here, compositional stability can be independent from functional stability while specific functional groups (for example AOB) remain synchronized with the community’s metabolic activity. Hence, prokaryotic taxa or OTUs should be questioned as ecologically meaningful units for describing community structure, at least when the focus is on ecosystem functioning (106, 145, 510). Microbial ecology and biogeography might be best understood using pathway-centric theories in which individual genes, operons or pathways are considered as basic reproductive and functional units, particularly under conditions where metabolic function defines the microbial niche.
space (109, 386). Accordingly, metagenomic, metatranscriptomic and metaproteomic profiling would be more suitable than taxonomic profiling for monitoring or predicting fluctuations in ecosystem functioning (106, 181, 510). Alternatively, functional community structure may be estimated from marker gene sequences by binning known OTUs into functional groups, as demonstrated in Chapters 2 and 3.

5.4 Conclusions

Convergence of microbial community composition is a gradual process that can last much longer than typical bioreactor experiments and environmental surveys. Transient dynamics of competitive exclusion explain why microbial communities can remain variable long after inoculation or perturbation, while exhibiting high functional stability. The correct interpretation of observed community dynamics in bioreactors and natural ecosystems thus requires a proper consideration of the involved time scales. Previous work has highlighted the general mismatch between the duration of typical experiments and the time scales assumed by conventional steady-state ecological theories (180), and our work demonstrates some of the implications of this mismatch. Fluctuations in natural and less controlled microbial communities likely result from several destabilizing processes, however the effects of these processes could be augmented by transient dynamics of competitive exclusion.

Furthermore, less resilient and more flexible communities need not imply a compromised functional stability, and previous experiments have indeed indicated a positive correlation between flexible community structure and stable performance (123). Several competing strains can rapidly and concurrently fill a metabolic niche when opportunities arise, while slowly replacing each other and maintaining constant performance during saturation. The time required for convergence or recovery of community composition correlates positively with functional redundancy, because more competitors are likely to have similar efficiencies under substrate limitation.

The extreme case in which each functional group consists of equal competitors is comparable to the so called emergent group theory in ecology, according to which assembly within each group is subject to neutral dynamics (185, 197). In that limit transient periods of competitive exclusion can be extremely long, while community composition appears dissociated from environmental conditions and driven by purely stochastic factors. While exact neutrality is an extreme idealization, some natural communities may indeed include functional groups consisting of almost-equal competitors. For example, previous work on a wastewater
treatment plant found that fluctuations within the group of ammonia oxidizing bacteria, as well as within the heterotrophic community, were predominantly explained by neutral processes rather than environmental factors (349). Similarly, a global study of desert microbial communities by Caruso et al. (64) found that climatic effects were detectable at the whole community level, but became undetectable when restricted to variations within the photosynthetic or heterotrophic groups. Frossard et al. (134) found that spatiotemporal variations of microbial community structure in stream catchments were best described by a neutral assembly model, whereas potential activities of several carbon-, nitrogen- and phosphorus-acquiring enzymes showed clear seasonal patterns. This disconnect between potential enzyme activities and the composition of their host communities indicates high functional redundancy and a decoupling between community function and taxonomy. Similarly, Yin et al. (537) showed significant functional redundancy in soil microbial communities by measuring population responses to enrichment with individual carbon sources. However, it is unclear whether all detected OTUs were active prior to the enrichment, and at any point in time a significant fraction of functionally similar clades may have only been present at low densities (418). Furthermore, subtle partitioning along additional non-functional axes such as moisture or pH may create micro-niches that enable the long-term coexistence of functionally similar populations, particularly in spatially or temporally heterogenous environments. For example, the coexistence of hundreds of subpopulations of the marine cyanobacterium *Prochlorococcus* is likely enabled by subtle niche differentiation such as adaptations to different nutrient availabilities (224). The role of neutrality in natural microbial communities and its proper reconciliation with niche theory remains controversial, and patterns that appear to result from neutral drift may in fact have underlying deterministic causes (see Chapter 6). Nevertheless, our work shows that approximate neutrality within ecological niches can explain several patterns of microbial community assembly in engineered environments and should also be considered when interpreting the dynamics of natural microbial communities.
Figure 5.1: Calibration and validation of the template AOB and NOB cell models to data from an experiment with a nitrifying batch-fed bioreactor (A) (94). Ammonium (NH$_4^+$) was added at the beginning of the experiment, and was sequentially oxidized to nitrite (NO$_2^-$) and nitrate (NO$_3^-$) by a growing nitrifier community. (B) and (C): Ammonium (B) and nitrate (C) concentration time series data (dots), compared to the calibrated model (continuous lines). (D): AOB and NOB cell densities over time, as predicted by the model. (E): Nitrite concentration over time, as predicted by the model. Note that while the template cell models were calibrated using batch-bioreactor experiments, for our subsequent analysis we consider continuously-fed flow-through bioreactors because these can support metabolically active microbial communities at steady state.
Figure 5.2: Simulations of the nitrifying bioreactor under constant conditions. (A): Simulated cell densities over time in the ammonium-fed nitrifying membrane bioreactor, inoculated with 100 random strains (AOB in variations of red, NOB in variations of blue). (B) Corresponding total cell densities per functional group (AOB or NOB). (C) Corresponding ammonium ($\text{NH}_4^+$), nitrite ($\text{NO}_2^-$) and nitrate ($\text{NO}_3^-$) concentrations. (D) Corresponding community-wide ammonium and nitrite uptake rates. (E,F): Distance of the community from the long-term steady state (in terms of the Bray-Curtis dissimilarity index; 266), following inoculation with 20 (E) or 100 (F) random strains. Shown as a probability distribution over 100 random simulations (colors correspond to centiles). Notice the faster rate of convergence to steady state (i.e., resilience) in (E) compared to (F). The two intervals on the top of figures (A,C,E) indicate rough startup and saturation phases, respectively.
Figure 5.3: **Simulations of the perturbed nitrifying bioreactor.** (A): Simulated cell densities over time in the ammonium-fed nitrifying membrane bioreactor, inoculated with 100 random strains (AOB in variations of red, NOB in variations of blue). A strong perturbation during day 5000 (grey arrow) causes a temporary collapse of the microbial community. (B) Corresponding total cell densities per functional group (AOB or NOB). (C) Corresponding ammonium (NH$_4^+$), nitrite (NO$_2^-$) and nitrate (NO$_3^-$) concentrations, at times near the perturbation. (D) Corresponding community-wide ammonium and nitrite uptake rates, at times near the perturbation. (E,F): Bray-Curtis dissimilarity of the community to the state shortly prior to the perturbation, in bioreactors inoculated with 20 (E) or 100 (F) random strains. Shown as a probability distribution over 100 random simulations (colors correspond to percentiles). Notice the greater resistance to perturbation and greater resilience in (E), compared to (F).
Chapter 6

Taxonomic variability and functional stability in microbial communities infected by phages

6.1 Synopsis

As shown in Chapters 2 and 3, microbial communities can display intense variation in taxonomic composition across space or time, and yet this taxonomic variation can coincide with stable metabolic functional community structure and constant biochemical performance. This decoupling between taxonomic and functional community structure is presumably enabled by a high functional redundancy in the global microbiome, however the mechanisms driving the sustained taxonomic variation within functional groups remain largely unknown. Predation by specialist lytic phages leading to “killing the winner” dynamics has been suggested as a potential cause of host turnover, however the plausibility and required conditions for this scenario have not been rigorously examined. Further, it is unknown how predation by phages affects community metabolic processes and whether these effects are actually mitigated by functional redundancy in the host populations. Here we address these issues using a model for a methanogenic microbial community that includes several interacting metabolic functional groups. Each functional group comprises multiple competing strains with distinct physiological parameters, and each strain is subject to predation by a specialist lytic phage. We find that phages induce intense taxonomic turnover within each functional group, resembling the variability observed in past experiments. The functional composition and metabolic throughput of the community are also disturbed by phage predation, but they become more stable as the functional redundancy in the community increases. Our work reveals explicit mechanisms by which functional redundancy stabilizes ecosystem performance.

1A version of this chapter is under peer review for publication (see the Preface for author contributions): Louca, S., Doebeli, M. (in review). Taxonomic variability and functional stability in microbial communities infected by phages.
and supports the interpretation that biotic interactions — rather than random population drift, as often suggested — drive taxonomic turnover within functional guilds.

6.2 Introduction

Microbial metabolism powers global biogeochemical fluxes (116) and is a key component in many engineered ecosystems, such as biofuel production units or wastewater treatment plants (16). Understanding the mechanisms shaping microbial community composition and function is thus paramount towards predicting ecosystem responses to anthropogenic change and towards optimizing the performance of microbially driven industrial processes (316). Microbial communities can exhibit strong variation in taxonomic composition, both across time and space, and yet this taxonomic variation can coincide with remarkably stable functional community structure (109, 288, 349, 397). For example, the proportions of several metabolic functional groups, such as nitrifiers, photoautotrophs and sulfate reducers, were found to be very similar across replicate natural aquatic ecosystems despite strong taxonomic turnover within individual functional groups (288). Similarly, methanogenic and nitrifying bioreactors operated under constant conditions were found to exhibit intense turnover of bacterial operational taxonomic units (OTUs) despite constant overall biochemical performance (122, 123, 506, 524, 548). These observations point towards an elegant paradigm in microbial ecology, in which energetic and stoichiometric constraints determine functional community structure and performance, but a high degree of functional redundancy in the global microbiome enables taxonomic variability within individual functional groups. The precise mechanisms causing this taxonomic variability remain largely unknown. Neutral population drift between equivalent competitors is sometimes suggested as a possible cause (349, 427), however non-random phylogenetic structure and co-occurrence patterns within functional groups point towards deterministic community assembly mechanisms, notably biotic interactions (288). In some cases complex non-equilibrium community trajectories have been reproduced across replicate isolated systems under constant conditions, further suggesting that the taxonomic variation within functional groups is not random (24, 495).

Host-specific predation by lytic phages has been suggested as a potential mechanism promoting host succession through “killing the winner” (KTW) dynamics, in which abundant host populations eventually collapse due to increased specialist predation, giving way to opportunistic competitors (322, 397, 398, 423, 499). For example, previous experiments revealed strong OTU turnover in a bioreactor, in which the temporary emergence of specific OTUs was followed by the temporary increase in their specific phages, consistent with KTW dynamics.
Microbial communities infected by phages (424). Predation by lytic phages is increasingly recognized as an important contributor to microbial mortality in natural and engineered ecosystems (135, 322, 476), and viral lysis has been shown to significantly reduce the flux of microbially assimilated organic carbon to higher trophic levels (e.g., to protist grazers; 135, 324). On the other hand, the effects of phages on dissimilatory carbon transformations (e.g., via respiration and fermentation) are much less understood (423), despite the fact that in many (e.g., methanogenic) environments most organic carbon is metabolized to byproducts for energy gain rather than assimilated into cell mass (312). Further, it is unclear whether — and under what conditions — phage-driven KTW dynamics can actually explain the extreme discrepancy between taxonomic variability and functional stability observed in microbial communities (109, 288, 349, 397). While this role of KTW dynamics is often hypothesized, in practice time lags involved in the recovery or replacement of collapsed populations, as well as potentially destabilizing interdependencies between metabolic pathways, could prevent the functional stability of communities.

To elucidate the effects of phage predation and microbial functional redundancy on community structure and metabolic functioning, we constructed a mechanistic model for a methanogenic microbial community subject to predation by lytic phages, hosted within an anaerobic flow-through bioreactor. Bioreactors constitute powerful model ecosystems for microbial ecology, because physicochemical conditions can be closely monitored and controlled, enabling replicate time series experiments. It is thus not surprising that bioreactor experiments have greatly contributed to our mechanistic understanding of microbial community assembly and of phage-host dynamics in particular (122, 312, 349, 423, 491). We chose methanogenic bioreactors as a template for our model because methanogenic metabolic networks are well understood and of great industrial relevance (77, 179) and because this allows for comparisons with previous experiments (122, 123).

Our model considers the population dynamics of multiple microbial functional groups involved in the anaerobic catabolism of glucose (the input substrate) to methane (CH4), as well as the concentrations of any intermediate metabolites (77; overview in Fig. 6.1). Specifically, in the first step input glucose is fermented to short-chain fatty acids, lactate and alcohols by several bacterial functional groups. These fermentation products are then further catabolized to hydrogen (H2) and acetate by “syntrophs”, i.e., bacteria that rely on the rapid consumption of H2 and acetate by hydrogenotrophic (“H2/CO2”) and acetoclastic methanogenic archaea. Each functional group initially comprises one or more distinct cell lineages — henceforth referred to as operational taxonomic units (OTU), which catalyze the same reaction but differ in several of their physiological parameters, such as their substrate half-saturation constants. We use “OTU” as an abstraction representing a taxonomic group
Microbial communities infected by phages (such as a strain or species) that is sufficiently narrow so that reaction kinetics are similar across members, and sufficiently broad so that different OTUs are infected by different specialist phages (69, 174, 521). The number of OTUs initially present in each functional group (termed “functional redundancy”) is a key parameter in our analysis and accounts for the presence of multiple functionally similar lineages in many bioreactors and natural environments (128, 224, 288, 289, 349, 526, 537). Note that here functional redundancy only refers to the number of redundant OTUs at the beginning of our simulations, while we make no assumptions on the long-term persistence of OTUs. Each OTU is associated with a distinct phage population that infects cells and causes increased mortality through cell lysis. Cell infection rates are proportional to phage concentrations and phage concentrations are, in turn, driven by cell lysis rates. Physiological parameters were chosen randomly for each OTU and each phage within realistic ranges (Table E.2), to account for the variation typically seen between strains or species (207, 328, 352). As we describe below, our model successfully reproduces previous experimental observations and yields novel insight into the effects of phage predation and functional redundancy on microbial community composition and function.

6.3 Results and discussion

6.3.1 Bioreactor dynamics in the absence of phages

Following “startup” of the bioreactor, and in the absence of phages, the successive growth of fermenters, syntrophs and methanogens quickly leads to the stabilization of metabolic activity within a few weeks (Fig. 6.2A). At this stage, microbial metabolism balances glucose supply and residual substrate loss from the bioreactor, although the exact steady state metabolite concentrations and community composition depend on the random parameters chosen. Competitive exclusion between reactions that are limited by the same substrates eventually leads to the persistence of only a subset of possible pathways driving glucose catabolism to CH₄ and CO₂. Each reaction is eventually performed by at most one remaining OTU, characterized by its ability to persist at the lowest substrate concentration (Figs. 6.2B,C). The bulk of biomass is attributable to fermenters and, to a lesser extent, syntrophs. Methanogens only account for a small fraction (< 1%) of the community because most of the energy available from glucose catabolism is extracted in the preceding steps (77).
6.3.2 Effects of phages on community dynamics

When phages are included in the model, specialist predation by phages leads to intense and irregular fluctuations of individual host populations in accordance to KTW dynamics (Fig. 6.2D). The duration of each infection cycle (i.e., the period from the initial detection to the eventual collapse of a phage population) varies greatly between phage-host pairs and between simulations, but is typically on the order of 20–150 days, consistent with similar durations observed in activated sludge (423). For many phage-host pairs these fluctuations closely resemble classical predator-prey cycles, although most cycles are irregular in their phase and amplitude (Fig. E.3), owing to their strong indirect interactions. Such complex — often chaotic — dynamics are common in systems composed of interacting oscillating components with distinct random frequencies (161). When averaged over time, predation by phages has detrimental effects on individual cell populations as well as on overall reaction rates. In particular, in the absence of any functional redundancy, average methane production drops down to less than 1% of the production that would typically be achieved in the absence of phages. Our model suggests that this reduction in performance can occur in at least two ways: First, increased cell mortality through cell lysis results in fewer cells that could consume a particular substrate before it is lost from the bioreactor. Second, because phage predation is biased towards dominant OTUs, it skews selection towards potentially less competitive (in terms of metabolic efficiency) OTUs within each functional group, leading to residual substrate concentrations that are higher than the equilibrium substrate concentrations of the top competitor (479). Due to the temporal delays involved in the recovery of populations or the opportunistic invasion by competitors, phages not only reduce the average metabolic throughput, but also induce fluctuations around that average (Fig. 6.3C). This may explain previously observed fluctuations of bioreactor performance that could not be fully explained using purely energetic and reaction-kinetic models (109, 286, 429).

Neutral demographic drift between equivalent competitors has previously been suggested as an explanation for seemingly random OTU turnover within functional groups in a wastewater treatment plant (349). In such systems, however, cell densities can be extremely high (e.g., $\sim 10^9$ cells $\cdot$ L$^{-1}$ in lakes and up to $10^{13}$ cells $\cdot$ L$^{-1}$ in bioreactors; 249, 518) and hence deterministic dynamics such as competitive exclusion between OTUs with even slight physiological differences are expected to dominate over pure demographic drift. For example, even at a population size of only $10^5$ cells and a constant difference in growth rates of only 1% between two competing OTUs, stochastic trajectories accounting for demographic drift would closely resemble the deterministic trajectory of exponential decline of the weaker competitor (e.g., $R^2 \sim 0.98$ for a birth-death model with constant combined population size, see
Methods for details). In contrast, our model shows that simple deterministic mechanisms can easily explain the sustained turnover between non-equivalent competitors observed in realistic settings, without the need for unrealistic neutral models.

6.3.3 Functional redundancy promotes functional stability

When considering multiple degrees of functional redundancy, we find a clear trend towards higher as well as more stable overall methane production rates at elevated functional redundancies (Figs. 6.3A,B). This suggests that the opportunistic growth of functionally similar OTUs may mitigate the detrimental effects of phage predation on community function by filling underutilized metabolic niches, thereby increasing and stabilizing overall community function. The probability that an appropriate alternative OTU is able to quickly replace a collapsing competitor increases with the functional redundancy in the initial inoculum (i.e., the available “seed bank”). Functional niche complementation is generally thought to promote a positive correlation between community richness and functional stability against external environmental perturbations (46, 365, 480, 522). Our work suggests that functional complementation in microbial communities also mitigates the detrimental effects that intrinsically emerging (rather than externally driven) fluctuations can have on ecosystem functioning. This may explain the occasionally observed positive relationship between microbial species diversity and biochemical performance, particularly at low diversities, even in the absence of external perturbations (163, 516, 526). The mechanism proposed here is fundamentally different from known mechanisms leading to a positive diversity-stability relation in classical food webs, as these mechanisms either involve a differential response of competitors to environmental perturbations (282) or are based on a skewed distribution towards weaker consumer-resource interactions (310). We mention that phage-driven KTW dynamics have been previously hypothesized to stabilize bioreactor performance by preventing competitive exclusion and hence maintaining a high functional redundancy that would, in turn, increase resilience to perturbations (423). Our analysis suggests that this interpretation may be slightly misleading, because phage predation itself can severely reduce and destabilize community function, while functional redundancy (e.g., ensured by a rich inoculum or exposure to a large pool of potential colonizers) acts to mitigate these destabilizing effects.

At increased functional redundancy, our model predicts that the proportions of various functional groups (in terms of relative cell abundances) become more stable, although the extent of this stabilization depends on the functional groups considered (Fig. 6.4I–L). The sta-
Microbial communities infected by phages

The stabilization of functional community structure is especially pronounced at a “coarse” level, i.e., when considering the proportions between fermenters, syntrophs and methanogens (Fig. 6.4I). Specifically, the coefficient of variation of these coarse groups drops from ∼0.9 (median value) in the absence of functional redundancy down to ∼0.2 at 100-fold functional redundancy. Since these groups represent indispensable and stoichiometrically coupled catabolic steps in the bioreactor, it is not surprising that their relative productivities are subject to strong stabilizing forces. Our model thus provides an explicit explanation for the discrepancy between stable functional community profiles and the highly variable taxonomic profiles often observed under constant environmental conditions (109, 288, 349, 397), and highlights the central role of functional redundancy coupled with biotic interactions in promoting this discrepancy. In particular, our results support the interpretation that the strong OTU turnover observed within functional groups in previous studies and in Chapter 3 was at least partly caused by biotic interactions including (but not necessarily limited to) predation by phages, rather than by neutral drift. Prolonged transients of competitive exclusion may also lead to slow OTU turnover within functional groups, especially following environmental perturbation (Chapter 5), although the detection of correlated OTU and phage succession in bioreactors provides additional evidence for KTW dynamics in these systems (397, 424).

On a finer resolution, i.e., when considering the proportions of single-reaction functional groups (e.g., acetoclastic vs H₂/CO₂ methanogens), the extent of stabilization depends on the specific set of functions considered (Figs. 6.4J–L). For example, individual fermenting functional groups (A–F in Fig. 6.1) appear interchangeable even at high functional redundancies (Fig. 6.4J), presumably because these groups represent strongly overlapping metabolic niches (they all ferment glucose). Fluctuations between these reactions, in turn, are predicted to drive comparably strong fluctuations in the proportions of syntrophic functional groups (G–J in Fig. 6.1) specializing on different fermentation products (Fig. 6.4K). Irregular transitions between alternative (“parallel”) catabolic electron flows — congruently with stable overall catabolic performance — have indeed been observed in previous experiments (122). In contrast, the proportions of the two methanogenic groups (K,L in Fig. 6.1) quickly stabilize at increasing functional redundancy. These findings reveal that the stability of functional community structure depends on the precise definition of functional groups, because non-identical functional groups may be interchangeable in case of overlap (365). A distinction between parallel and sequential functional groups is particularly crucial for “branched” metabolic networks such as organic carbon catabolism but may be less relevant for more sequential functions such as nitrification (oxidizing ammonium to nitrite and then nitrate) or denitrification (reducing nitrate to nitrite, nitric oxide, nitrous oxide and...
Microbial communities infected by phages

eventually nitrogen gas; 144).

Although not modeled, phage-host co-evolution could further destabilize population dynamics within functional groups, for example due to the repeated emergence of new resistant strains. In addition, rapid evolution of host resistance, for example via clustered regularly interspaced short palindromic repeats (CRISPRs; 13), could buffer the impacts of viruses on overall ecosystem functioning (269). While evolution is likely an important component of phage-host dynamics in natural systems (156, 322, 424), our work suggests that deterministic ecological dynamics are sufficient to explain the succession of OTUs often observed during constant community function. In fact, Shapiro et al. (424) observed rapid succession between distantly related OTUs and their associated phages in a bioreactor, suggesting that replacement by non-related competitors — rather than adaptive evolution of resistance — was indeed the main mode of host succession.

We note that the destabilizing role of phage predation and the stabilizing role of functional redundancy, as predicted by our model, rely on the assumption that lytic phages are specialized on single host populations (a prerequisite for KTW dynamics). High host specificities of phage infectivity (e.g., at the strain or species level) are generally considered to be the rule (69, 174, 521), however it is becoming apparent that host specificity may be more variable than previously assumed (191, 423). In reality multiple microbial clades may be susceptible to the same phages and hence the effective functional redundancy in the system may be much lower than the actual number of bacterial and archaeal strains. In cases where host specificity is the exception rather than the rule, the dynamics are expected to be markedly different than predicted here. For example, experimental and theoretical work shows that generalist predation by protist grazers can severely and permanently reduce community function even at high prey diversity (213, 366, 474, 476). Further, temperate phage strategies, not considered here, likely have less severe effects on host populations and metabolic throughput than predicted by our model. For example, recent findings suggest that phages may be switching from lytic to lysogenic during high host abundances, a behavior termed “piggy-backing the winner” (243), and this switching may dampen oscillations in host abundances. Hence, in environments characterized by temperate — rather than lytic — phage-host interactions (as may be the case for the human gut, 391), KTW dynamics will be less pronounced and hence microbial communities may be more stable in terms of their taxonomic composition. Further, recent work suggests that metabolic host reprogramming by viruses can direct energy and nutrients towards viral replication, potentially altering biogeochemical cycling (199, 405). It is likely, however, that the time scales of such host-phage co-evolutionary dynamics are much longer than — and thus of limited relevance to — the ecological successional dynamics.
considered here.

6.3.4 Statistical averaging or dynamic stabilization?

As discussed above, a high functional redundancy in the microbial community can reverse the destabilizing effects that KTW dynamics have on net metabolic activity and on the abundances of functional groups (Figs. 6.4A–L). Such stabilizing effects at increased species richness have been hypothesized in the past, based on the expectation that the effects of multiple fluctuating populations on broad community properties (such as total biomass or overall catabolic activity) ought to “average” each other out (267, 283, 480). According to this interpretation, which assumes that populations fluctuate independently, the stabilization of broad community properties at elevated functional redundancy is a simple statistical necessity (46, 105). In reality, however, populations are inevitably coupled due to competition for resources and metabolic interdependencies, and it is a priori unclear whether these interactions lead to stronger or weaker averaging effects (481). For example, fluctuations in metabolite concentrations caused by changes in specific microbial populations will affect the growth of all functional groups consuming or producing the particular metabolites, and these effects will typically act in a similar direction on all members within a functional group. Such positive correlations in the response of competing populations to chemical perturbations would act against stabilization by averaging. On the other hand, the collapse of a particular population due to cell lysis eventually frees a metabolic niche that can be occupied by competing populations, and such a “dynamic stabilization” would likely lead to a higher functional stability when compared to mere statistical averaging. A distinction between the two stabilizing mechanisms — statistical averaging vs dynamic stabilization — is key to understanding the effects of (metabolic) community interactions on overall functional stability.

To assess whether the functional stabilization observed in our simulations at high functional redundancy is a mere averaging effect or dynamic, we compared the coefficients of variation of functional group abundances to a null model in which the time course of each cell population was randomly shifted in time. This null model resembles the hypothetical scenario in which populations fluctuate independently of one another, while preserving the broad features of these fluctuations. We defined the “degree of dynamic stabilization” (DDS) in a particular simulation as the probability that the null model would lead to a higher coefficient of variation (averaged across functional groups) than observed, estimated through repeated random trials. We find that coarse functional group proportions exhibit a high DDS that
approaches 1.0 at high degrees of functional redundancy (Fig. 6.4M), consistent with the interpretation that opportunistic competitors quickly replace collapsing populations. On the other hand, we find contrasting effects for the proportions of single-reaction functional groups. Specifically, the proportions of various fermenting groups exhibit the lowest DDS (~0.1 – 0.6, Fig. 6.4N), while the proportions of the two methanogenic groups exhibit the highest DDS (up to ~0.9 – 1.0, Fig. 6.4P). This is consistent with the interpretation that groups consuming similar substrates (e.g., glucose in the case of fermenters) are highly interchangeable and hence their proportions are only weakly stabilized. On the other hand, the rapid stabilization of H₂ and acetate concentrations at high functional redundancies appears to promote a stable ratio between H₂/CO₂ and acetoclastic methanogen populations.

We find that the total cell concentration is only occasionally dynamically stabilized and is often less stable than predicted purely based on statistical averaging (DDS~0.1–0.7 at 100-fold functional redundancy). This means that fluctuations of single populations can lead to synchronized and similarly signed responses across multiple functional groups (e.g., a collapse of the dominant glucose fermenters also induces a temporary collapse of methanogens). In consequence, the absolute abundances of individual functional groups are less stable than their relative proportions, even at elevated functional redundancies. This means that care must be taken when assessing the stability of functional community profiles, for example based on metagenomic sequences (97), because such profiles generally only reflect the relative but not the absolute abundances of functional groups.

6.3.5 Phages promote prokaryotic diversity

Classical competition theory predicts that at steady state only a single competitor can persist within a metabolic functional group that is limited by a single substrate, and this competitor is determined by its ability to survive at the lowest possible substrate concentration (178, 479). Consistent with this competitive exclusion principle, when we excluded phages from our model each functional group was eventually occupied by at most one OTU (e.g., Fig. 6.2C). In reality, however, microbial richness can be extremely high even in simple engineered ecosystems, such as nitrifying bioreactors (109), methanogenic digesters (123, 150) or activated sludge (234), where it can range from hundreds to thousands of OTUs. Such high richness is in apparent contradiction to the competitive exclusion principle. Mechanisms proposed in the past to explain this discrepancy include slow transient dynamics of competitive exclusion far from steady state (285), negative frequency dependence through non-linear substrate dependencies (281, 449), externally driven fluctuations
Microbial communities infected by phages of resource availability or physical conditions (432, 449), as well as phage-driven KTW dynamics (48, 320, 424, 474, 475, 512). Our work provides support for KTW dynamics as a likely explanation for sustained high OTU richness within metabolic functional groups, especially when high richness is observed in the absence of obvious environmental fluctuations (122, 349, 424). For example, in our simulations with 20-fold functional redundancy, in the long term each functional group is typically occupied by 5–10 OTUs, each of which occasionally and temporarily increases in relative abundance (e.g., Figs. 6.2E,F). Similarly, as mentioned earlier, the repeated disruption of reactions prevents the long-term mutual exclusion between alternative reactions limited by the same substrates, resulting in a “parallelization” of catabolic fluxes. Notably, during most simulations multiple fermenting groups would contribute (occasionally or at all times) to the catabolism of glucose. Phage-induced catabolic parallelization can thus maintain functional diversity in addition to diversity within functional groups and may explain the cooccurrence of alternative intermediate products (e.g., multiple fatty acids) sometimes observed in methanogenic digesters (179, 533). We note that in certain cases (e.g., at high functional redundancy) slow transients of competitive exclusion can also lead to prolonged coexistence of OTUs with similar metabolic efficiencies (e.g., Fig. 6.2A,B and Chapter 5), and in reality multiple mechanisms likely contribute to the maintenance of richness within a community.

6.4 Conclusions

Microbial communities can display strong taxonomic variation across space (288, 307) or time (109, 349, 397) even under similar environmental conditions, while exhibiting relatively constant functional community structure. In fact, recent work showed that physical and chemical conditions strongly predict the distribution of metabolic functional groups across the world’s ocean, but only poorly explain the taxonomic composition within individual functional groups (289). Both Louca et al. (288) and Louca et al. (289) found no effect of spatial dispersal limitation on community differences, and Louca et al. (288) showed that OTUs within individual functional groups displayed non-random phylogenetic relationships and non-random co-occurrence patterns that were at least partly caused by biotic interactions. These findings suggest an elegant paradigm for microbial ecology, in which the functional structure and the taxonomic composition within functional groups constitute two separate facets of community composition, with the former being driven by stoichiometric and energetic environmental constraints, while the latter is heavily shaped by complex biotic interactions. Our mechanistic model provides further support for such an interpretation by explicitly demonstrating how — and under which conditions — predation by specialist
phages can drive OTU turnover while maintaining constant functional community structure and metabolic performance. Experimental evidence for phage-driven KTW dynamics remains limited (322, 357, 397, 398, 424, 499), mostly due to the technical difficulties involved in virome profiling and in linking particular phages to their specific hosts in natural ecosystems. Nevertheless, many natural environments, including the open ocean (462), soil (18) and the human gut (391), exhibit high phage densities and it is likely that phages also contribute to the decoupling between taxonomic and functional community structure in these environments (134, 288, 289, 307, 379, 381).

The decoupling between functional and taxonomic community structure, especially at high functional redundancy, has important implications for the interpretation of microbial biogeographical patterns, because variation in taxonomy need not imply differences in community function. Reciprocally, environmental constraints determining community function may only poorly explain the distribution of individual taxa. If phage-induced KTW dynamics indeed strongly shape the taxonomic variation within functional groups, as suggested by this and previous work, then the aspiration of accurate microbial species distribution models (258, 455) may turn out to be a Sisyphean struggle, especially when important functional traits are non-monophyletic (308). Disentangling the functional and taxonomic variation in microbial communities is thus an important prerequisite for a truly predictive microbial ecology.
Figure 6.1: **Modeling methanogenic communities.** Simplified metabolic network of anaerobic methanogenic communities (77), as modeled in this study. Each circle represents one of 12 functional groups specialized on a particular metabolic reaction and each reaction can be performed by multiple competing OTUs (“functional redundancy”). The network is roughly structured into 3 sequential “coarse” functional groups based on the type of substrate used (fermentation, syntrophy and methanogenesis). Detailed reaction stoichiometry in reference to the inscribed letters is provided in Table E.1.
Figure 6.2: **Phage predation drives OTU turnover.** (A) OTU abundances (one color per OTU) during a single simulation without phage predation, at 20-fold functional redundancy. Competitive exclusion leads to the extinction of all but a single fermenter and all but two very similar syntrophs (although eventually only one syntroph remains, not shown). Methanogens exhibit much lower abundances than fermenter and syntrophs and are thus not visible. (B,C): Proportions of acetoclastic methanogens (B) and proportions of $\text{H}_2/\text{CO}_2$ methanogens (C) (one color per OTU) during the same simulation as (A). (D–F): Analogous to (A–C), but for a simulation including phage predation. Phage-host interactions drive variation in overall cell abundances (D) as well as in the OTU composition within functional groups (E,F).

Figure 6.3: **Functional redundancy increases and stabilizes methane production in the presence of phages.** (A,B): Temporal (A) averages and (B) coefficients of variation of effluent methane concentration, for various degrees of functional redundancy. Box plots represent the distribution across 100 random simulations. Vertical bars comprise 95% percentiles. (C,D): Methane concentrations during a single simulation at (C) 20-fold and (D) 100-fold functional redundancy.
Figure 6.4: Effects of functional redundancy on functional community composition in the presence of phages. Row 1: Proportions of (A) coarse functional groups (total methanogens vs total fermenters vs total syntrophs), (B) fermenting groups, (C) syntrophic groups and (D) methanogenic groups over time (one color per functional group), during a random simulation at 20-fold redundancy. Note that each functional group consists of multiple OTU populations (individual OTU concentrations not shown). Row 2: Similar to row 1, but for a simulation at 100-fold functional redundancy. Row 3: Coefficients of variation (CV) for the proportions of (I) coarse functional groups, (J) fermenting groups, (K) syntrophic groups and (L) methanogenic groups, at various levels of functional redundancy. Row 4: Degree of dynamic stabilization (DDS) of functional group proportions (corresponding to I–L), at various levels of functional redundancy. Box plots (I–P) represent the distribution of CVs (I–L) and DDSs (M–P) across 100 random simulations; vertical bars indicate 95% percentiles.
Chapter 7

Gene-centric modeling of the Saanich Inlet oxygen minimum zone

7.1 Synopsis

The work presented above provides strong support for a pathway-centric paradigm of microbial ecology, according to which the distribution and activity of microbial metabolic pathways is strongly shaped by environmental physicochemical conditions, regardless of the detailed community assembly mechanisms. The question then arises, what would be an appropriate mathematical model to describe the interaction between metabolic pathways and their environment, and how could such a model incorporate modern molecular sequence data? To address these questions, here we construct a quantitative biogeochemical model that describes metabolic coupling along the redox gradient in Saanich Inlet—a seasonally anoxic fjord with biogeochemistry analogous to oxygen minimum zones (OMZ). The model reproduces measured biogeochemical process rates as well as DNA, mRNA, and protein concentration profiles across the redox cline. Simulations make rigorous but unexpected predictions about the role of ubiquitous OMZ microorganisms in mediating carbon, nitrogen, and sulfur cycling. For example, nitrite “leakage” during incomplete sulfide-driven denitrification by SUP05 Gammaproteobacteria is predicted to support inorganic carbon fixation and intense nitrogen loss via anaerobic ammonia oxidation (anammox). Moreover, this coupling creates a metabolic niche for nitrous oxide reduction that completes denitrification by currently unidentified community members. These results quantitatively improve previous conceptual models describing microbial metabolic networks in OMZs. Beyond OMZ-specific predictions, model results indicate that geochemical fluxes are robust indicators of microbial community structure and, reciprocally, that gene abundances and geochemical conditions

largely determine gene expression patterns. The integration of geochemical profiles, process rate measurements, as well as metagenomic, metatranscriptomic and metaproteomic sequence data into a biogeochemical model, as demonstrated here, enables holistic understanding of the distributed microbial metabolic network driving nutrient and energy flow at ecosystem scales.

### 7.2 Introduction

Microbial communities catalyze Earth’s biogeochemical cycles through metabolic pathways that couple fluxes of matter and energy to biological growth (116). These pathways are encoded in evolving genes that over time spread across microbial lineages and today help shape the conditions for life on Earth. High-throughput sequencing and mass spectrometry platforms are generating multi-omic sequence information (DNA, mRNA and protein) that is transforming our perception of this microcosmos, yet the vast majority of environmental sequencing studies lack a mechanistic link to geochemical processes. At the same time, mathematical models are increasingly used to describe local and global scale biogeochemical processes or to predict future changes in global element cycling and climate (138, 392). While these models typically incorporate the catalytic properties of cells, they fail to integrate the information flow between DNA, mRNA, proteins and process rates as described by the central dogma of molecular biology (85). Hence, a mechanistic framework integrating multi-omic data with geochemical information has remained elusive.

Recent work based on metagenomics and quantitative PCR suggests that biogeochemical processes may be described by models focusing on the population dynamics of individual genes (386, 387). In such gene-centric models, genes are used as proxies for particular metabolic pathways, with gene production rates being determined solely by the Gibbs free energy released by the catalyzed reactions. While compelling, these models stop short of incorporating the entire central dogma of molecular biology, and therefore do not achieve a truly quantitative integration of multi-molecular information and geochemical processes. For example, while such models allowed for a qualitative comparison between modeled gene production rates and selected transcript profiles, they do not provide any explicit mechanistic links (386). In fact, the lack of a quantitative validation against process rate measurements or other proxies for activity (e.g., proteins) begs the question whether gene-centric models are adequate descriptions for microbial activity.

Here we construct a gene-centric model for Saanich Inlet, a seasonally anoxic fjord on the
Gene-centric modeling of the Saanich Inlet OMZ coast of Vancouver Island, Canada (12), and a tractable analogue for the biogeochemistry of oxygen minimum zones (OMZs). Oxygen minimum zones are widespread and expanding regions of the ocean in which microbial community metabolism drives coupled carbon, nitrogen and sulfur cycling (489, 531). These processes exert a disproportionate influence on global N budgets with resulting feedbacks on marine primary productivity and climate (252, 509). Extensive time series monitoring in Saanich Inlet provides an opportunity to interrogate biogeochemical processes along defined redox gradients extensible to coastal and open ocean OMZs (503, 540). Our model explicitly describes DNA, mRNA, and protein dynamics as well as biogeochemical reaction rates at ecosystem scales. We use geochemical depth profiles, rate measurements for N cycling processes, metagenomic, metatranscriptomic and metaproteomic sequence data, as well as qPCR-based cell count estimates for SUP05 Gammaproteobacteria — the dominant denitrifiers in Saanich Inlet (181), to calibrate and validate our model. All data were obtained from a single location in Saanich Inlet across several depths in early or mid-2010.

7.3 Construction and calibration of a gene-centric model

A recurring cycle of deep water renewal and stratification in Saanich Inlet results in an annual formation of an oxygen-depleted ($O_2$) zone in deep basin waters (Fig. 7.1B). As this oxygen minimum zone slowly expands upwards during spring, anaerobic carbon remineralization in the underlying sediments leads to an accumulation of ammonium ($NH_4^+$) and hydrogen sulfide ($H_2S$) at depth (340). The oxidation of sulfide using nitrate produced through nitrification in the upper water column fuels chemoautotrophic activity and promotes the formation of a narrow sulfide-nitrate transition zone (SNTZ) at intermediate depths (218, 504, 540).

Our model describes the dynamics of several genes involved in carbon, nitrogen and sulfur cycling across the Saanich Inlet redox cline. Each gene is a proxy for a particular redox pathway that couples the oxidation of an external electron donor to the reduction of an external electron acceptor. The model builds on a large reservoir of previous work that provides conceptual insight into the microbial metabolic network in Saanich Inlet (11, 181, 218, 503, 504, 540) and includes aerobic remineralization of organic matter (ROM), nitrification, anaerobic ammonium oxidation (anammox) and denitrification coupled to sulfide oxidation (Fig. 7.1A). Certain pathways found in other OMZs, such as aerobic sulfide oxidation (415), dissimilatory nitrate reduction to ammonium (DNRA; 253) and sulfate reduction (386), were excluded from our model based on information from previous studies (11, 181, 218, 503, 504, 540) as well as preliminary tests with model variants (as explained below). Reaction rates (per
gene) depend on the concentrations of all used metabolites according to 1st order or 2nd order (Michaelis-Menten) kinetics (211). In turn, the production or depletion of metabolites at any depth is determined by the reaction rates at that depth. The production of genes is driven by the release of energy from their catalyzed reactions, and is proportional to the Gibbs free energy multiplied by the reaction rate (396).

The model was evaluated at steady state between 100 and 200 m depth. Accordingly, free parameters were calibrated using geochemical profiles obtained in early 2010 during a period of intense water column stratification, which resulted in an extensive anoxic zone that approached a steady state (Fig. 7.2; Appendix F.3.7). After calibration, most residuals to the data are associated with an upward offset of the predicted SNTZ (Figs. 7.2B,C,F) and an underestimation of nitrous oxide (N₂O) concentrations in deep basin waters (Fig. 7.2E). These residuals can be explained by subtle deviations from a steady state. Such deviations were revealed in subsequent time series measurements during which the SNTZ continued to migrate upwards in the water column (Fig. 7.2).

7.4 Results and Discussion

7.4.1 DNA profiles and process rates

The calibrated model makes predictions about gene abundance and process rates, which can be validated using metagenomic sequence data and N process rate measurements from the same location and period as the geochemical calibration data (Fig. 7.3). Consistent with metagenomic depth profiles, the model predicts a redox-driven partitioning of pathways across the water column. Genes associated with ROM (ABC transporters mapped to dominant aerobic heterotrophs), aerobic ammonium oxidation to nitrite (ammonia monooxygenases, amo) and aerobic nitrite oxidation to nitrate (nitrite oxidoreductases, nxr) decline precipitously in deep basin waters, whereas genes associated with partial denitrification to nitrous oxide (PDNO, represented by nitric oxide reductases, norBC), nitrous oxide reduction (nitrous oxide reductases, nosZ) and anammox (hydrazine oxidoreductases, hzo) are most abundant in the SNTZ (Fig. 7.1A).

The similarity of the PDNO, nosZ and hzo gene profiles is indicative of their strong interaction. In particular, the co-occurring peaks of PDNO and nosZ abundances in the absence of N₂O accumulation (Fig. 7.2E) reflect a quantitative coupling between the two denitrification steps and imply that both steps support extensive productivity at the SNTZ. This
is intriguing because genomic reconstruction from both uncultivated and cultivated SUP05, the dominant denitrifier in Saanich Inlet, has not identified the *nosZ* gene (181, 420, 504). This suggests that incomplete nitrate reduction by SUP05 and reduction of nitrous oxide by unidentified community members constitute separate and complementary metabolic niches in Saanich Inlet under suboxic and anoxic conditions.

The superposition of electron donor-acceptor pairs in redox transition zones supports chemical energy transfer in stratified water columns (45, 547), and previous studies revealed relatively high cell abundances and chemosynthetic activity in such zones (164, 515, 540). At the SNTZ in Saanich Inlet, the simultaneous availability of NO$_3^-$ and H$_2$S fuels chemosynthetic nitrate respiration coupled to sulfide oxidation, in turn supplying anammox with NO$_2^-$ via “leaky” denitrification (up to 88% of NO$_2^-$ supplied by PDNO; Appendix F.3.11). Most of the NH$_4^+$ utilized by anammox (0.3 mmol·m$^{-2}·$d$^{-1}$), on the other hand, is predicted to originate from the underlying sediments and reach the SNTZ via diffusion. Accordingly, both anammox and denitrification rates are predicted to peak around the SNTZ and lead to a significant release of N$_2$. This prediction is consistent with process rate measurements from discrete depth intervals during subsequent cruises in 2010 (Fig. 7.3B), as well as elevated SUP05 cell counts at the SNTZ (Fig. 7.3A, estimated via qPCR). In fact, the good agreement between predicted PDNO gene counts and observed SUP05 cell counts suggests energy fluxes associated with denitrification can be accurately translated to denitrifier biosynthesis rates. Predicted peak sulfide-driven denitrification rates are somewhat higher than peak anammox rates, although depth-integrated nitrogen loss rates are comparable for both pathways ($\sim$ 0.3 mmol-N$_2$·m$^{-2}·$d$^{-1}$). These predictions are partly consistent with rate estimates derived directly from the geochemical profiles using inverse linear transport modeling (ILTM, Fig. 7.3B; see Methods for details). Hence, near steady state conditions, coupled sulfide-driven denitrification and anammox can concurrently drive significant nitrogen loss in the water column.

The fraction of NO$_2^-$ leaked during denitrification, compared to the total NO$_3^-$ consumed ($L_{PDNO} = 0.352$; Appendix F.3.3), could be calibrated as a free model parameter based on the observed geochemical profiles. Such a high NO$_2^-$ leakage may result from an optimization of energy yield under electron donor limitation. Further experimental work is needed to understand the mechanisms controlling this leakage by SUP05. Heterotrophic denitrification and nitrification are conventionally thought of as the primary sources of both nitrite and ammonium for anammox in OMZs (253, 501), and so far evidence for a direct coupling between sulfide-driven denitrification and anammox has been scarce (515). Our results indicate that sulfide-driven denitrification is an important precursor for anammox, particularly
Gene-centric modeling of the Saanich Inlet OMZ under conditions of organic carbon limitation (172).

Steady state gene production rates for chemoautotrophic pathways are predicted to peak around the SNTZ, reaching $\sim 3.2 \times 10^6$ genes $\cdot$ L$^{-1} \cdot$ d$^{-1}$. This gene production rate corresponds to a dark carbon assimilation (DCA) rate of $\sim 60$ mmol-C $\cdot$ L$^{-1} \cdot$ d$^{-1}$, assuming a carbon:dry weight ratio of 0.45 (115) and a dry cell mass of $m = 5 \times 10^{-13}$ g (396). Previously measured peak DCA rates in the Saanich Inlet OMZ reached $2 \mu$mol-C $\cdot$ L$^{-1} \cdot$ d$^{-1}$ (218), which is significantly higher than the values predicted here. The potential activity of autotrophic pathways not considered here, such as methane oxidation (531), may explain some of these differences. Moreover, the model assumes steady state conditions, while redox conditions were far from steady state during previous DCA measurements (218). Transient dynamics in Saanich Inlet can exhibit significantly higher nitrogen fluxes and chemoautotrophic activity (295), which might further explain discrepancies between the model and measured DCA rates. Accounting for chemoautotrophic productivities based on oxidant and reductant supply in redox transition zones is generally difficult due to limited knowledge on active pathways, the possibility of cryptic nutrient cycling and potential lateral substrate intrusions, and discrepancies similar to our study are frequently reported for other OMZs (217, 274, 341, 546).

Previously detected amino acid motifs similar to those found in proteins catalyzing DNRA, suggested that SUP05 may also be providing $\text{NH}_4^+$ to anammox through DNRA (181). DNRA, not included in the model, is known to fuel anammox in anoxic sediments and water columns (253, 374). So far incubation experiments have not revealed any DNRA activity in Saanich Inlet, and measured ammonium profiles do not indicate any significant ammonium source at or below the SNTZ (Fig. 7.2B). Nevertheless, DNRA could be active in Saanich Inlet and remain undetected due to rapid ammonium consumption by anammox (374). An extension of the model that included DNRA as an additional pathway, and which we calibrated to the same geochemical data (January–March 2010), predicted negligible DNRA rates compared to denitrification and anammox and consistently converged to the same predictions as the simpler model. This suggests that DNRA may be absent from the Saanich Inlet water column — at least near steady state conditions in late spring — and that the hydroxylamine-oxidoreductase homolog encoded by SUP05 plays an alternative role in energy metabolism (504).

DNA concentration profiles of anammox and denitrification genes appear more diffuse and are skewed towards deep basin waters, compared to their corresponding rate profiles and the SNTZ (Fig. 7.3). The model explains this apparent discrepancy based on turbulent diffusion.
and sinking, which both transport genes away from their replication origin. Hence, community composition at any depth is the combined result of local as well as non-local population dynamics. Metabolic flexibility encoded in the genomes of microorganisms mediating these processes may also contribute to broader distributions of individual genes than their predicted activity range (172). This disconnect between local metabolic potential and activity needs to be considered when interpreting metagenomic profiles in a functional context, especially in environments with strong redox gradients such as OMZs (531) or hydrothermal vents (387).

The concentration maxima of anammox and denitrification genes are predicted at shallower depths than measured (Fig. 7.3A). This observation is consistent with the upward offset of the predicted SNTZ and highlights an important limitation of steady state models applied to dynamic ecosystems. Indeed, process rate maxima predicted via ILTM at multiple time points continue to move upwards beyond the time interval used for model calibration (Fig. 7.3B). In reality, an electron donor/electron-acceptor interface as narrow as predicted by the model would only develop after sufficient time for transport processes and microbial activity to reach a true steady state. Such narrow interfaces do appear in permanently stratified meromictic lakes (435) or the Baltic Sea, where stagnation periods can persist for many years (172).

### 7.4.2 mRNA and protein profiles

Metagenomics can yield insight into the metabolic potential of microbial communities, however it is ill suited as a means to infer actual pathway activity because the latter depends, among others, on environmental conditions. Metatranscriptomics and metaproteomics present powerful means to assess community metabolic activity, and each method comes with its own set of advantages (181, 330). For example, while transcripts present immediate proxies for gene up-regulation (e.g., in response to changing redox conditions), proteins represent the immediate catalytic potential of a community, and in vitro characterization of enzyme kinetics can facilitate the projection of protein abundances to in situ process kinetics (502). Transcript abundances need not always correlate strongly with protein abundances, for example in cases of transcriptional control or protein instability (265, 502), and hence metatranscriptomics and metaproteomics can in principle provide different perspectives on community activity. This warrants a systematic evaluation of the consistency between these alternative layers of information in real ecosystems. In fact, a unifying mechanistic framework describing the processes that control environmental mRNA and protein
distributions is crucial for the correct interpretation of multi-omics data (330).

While DNA replication and process rates are predicted by our gene-centric model merely based on environmental redox conditions, it is uncertain to what extent intermediate stages of gene expression (transcription and translation) can be explained based on such a paradigm. For example, environmental mRNA concentrations measured via qPCR have previously been directly compared to predicted reaction rates (386), but such a heuristic comparison ignores other mechanisms controlling environmental biomolecule distributions, such as physical transport processes. Here, in an attempt to mechanistically describe mRNA and protein dynamics at ecosystem scales, we hypothesized that both mRNA and protein production rates at a particular depth are proportional to the total reaction rate at that depth (calculated using the calibrated model). This premise is motivated by observations of elevated transcription and translation rates during high metabolic activity or growth (9, 154, 229). Furthermore, we assumed that mRNA and protein molecules are subject to the same hydrodynamic dispersal processes as DNA, while decaying exponentially with time, post synthesis. The decay time of each molecule, as well as the proportionality factor between the reaction rate and synthesis, were statistically estimated using metatranscriptomic and metaproteomic data (Appendix F.3.10).

The general agreement between this model and the molecular data (Fig. 7.3A) suggests that the production-degradation dynamics of several of these molecules are, at the ecosystem level, dominated by the mechanisms described above. The best fit (in terms of the coefficient of determination, Table F.4) is achieved for nosZ and nxr mRNA, as well as amo, norBC, nxr and ROM proteins. The greater number of protein over mRNA profiles that can be explained by the model suggests that the proteins considered here are indeed simply produced on demand and slowly degrade over time, while mRNA dynamics are subject to more complex regulatory mechanisms (153, 330). In particular, the decay times of some transcripts and proteins were estimated to be as high as several weeks (Table F.4). For proteins these estimates fall within known ranges (153), however for transcripts these estimates are much higher than decay times determined experimentally in cells (465). One reason for this discrepancy appears to be the underestimation of the SNTZ depth range by the model, which in turn leads to longer estimates for mRNA decay times needed to explain the detection of these molecules outside of the SNTZ. Alternatively, transcripts and proteins might persist in the cells in inactive states for a significant period of time even after dispersal into areas with low substrate concentrations. For example, stable but silent transcripts have been found in bacteria following several days of starvation (245, 299). Further, gene expression may not change immediately upon a change of external stimulus (181). For example, for
some prokaryotic transcription cascades the basic time unit may be the cell doubling time (which can reach several weeks in anoxic environments; 518), due to regulation by long-lived transcription factors (402). Hence, the decay times estimated here may reflect a hysteresis in gene down-regulation following nutrient depletion, perhaps in anticipation of potential future opportunities for growth (326). Overall, these observations suggest that future models for environmental mRNA dynamics and future metatranscriptomics would benefit from a consideration of additional mRNA control mechanisms, for example derived from cell-centric transcription models (403, 542).

7.4.3 Consequences for geobiology

Gene-centric models have the potential to integrate biogeochemical processes with microbial population dynamics (386, 387). According to the central dogma of molecular biology (85), gene transcription and translation are intermediate steps that regulate metabolic processes in individual cells, but the appropriate projection of the central dogma to ecosystem scales remained unresolved because transcription and translation were not explicitly considered in previous models (386, 387). We have developed a biogeochemical model that explicitly incorporates multi-omic sequence information and predicts pathway expression and growth in relation to geochemical conditions. In particular, when mRNA and protein dynamics are omitted, the gene-centric model only includes 4 calibrated parameters and yet is able to largely reproduce geochemical profiles (Fig. 7.2), relative metagenomic profiles (Fig. 7.3A) and SUP05 cell abundances, which means that the good fit is unlikely the mere result of overfitting. In fact, as we refined and calibrated our model to the geochemical profiles, we observed that the metagenomic profiles were well reproduced as soon as the model’s geochemical predictions roughly aligned with the data, even if the calibrated parameters were still being adjusted. This further strengthens the interpretation that fluxes of matter and energy are robust predictors of microbial productivity and functional community structure.

Our model successfully explains a large fraction of environmental mRNA and protein distributions based on DNA concentration profiles and biogeochemical reaction rates in the OMZ. These results are consistent with the idea that DNA is a robust descriptor of an ecosystem’s biological component (104, 201) which, in conjunction with the geochemical background, determines pathway expression and process rates (257). This implies that the co-occurrence of a metabolic niche with cells able to exploit it is sufficient to predict microbial activity. Under this paradigm, DNA may be regarded as directly coupled to the extracellular geochemical environment, while the production of mRNA and proteins is an inevitable consequence of
the biologically catalyzed flow of matter and energy. This interpretation is supported by the overall consistency between the metatranscriptomic and metaproteomic profiles for N and S cycling genes (Fig. 7.3A) and suggests that mRNA and proteins may each be adequate proxies for pathway activity in Saanich Inlet. Further work is needed to test this paradigm in other ecosystems, especially under non-steady state conditions. Nevertheless, many marine and freshwater water columns are permanently or almost permanently anoxic (416, 435, 489) and hence, our approach and conclusions are expected to be particularly applicable to these systems.

In addition to providing a systematic calibration and validation of the model we identified processes that need to be considered when interpreting multi-omic data. Conventional proxies for activity, such as mRNA and proteins, are themselves subject to complex population dynamics that include production, active or passive degradation as well as physical transport processes. Consequently, the close association between process rates and biomolecule production suggested above does not imply that biomolecule distributions per se are equivalent to local microbial activity rates. In Saanich Inlet, for example, the wide distribution of DNA, mRNA and proteins across the OMZ, in contrast to a relatively narrow metabolic “hot zone” at the SNTZ, is predicated on a balance between spatially confined production and dispersal across the water column. This so-called mass effect (268) indicates that geochemical or biochemical information is needed to assign actual activity to genes or pathways identified in multi-omic data, especially for components mediating energy metabolism. This conclusion is generalizable and should be applied to other ecosystems exhibiting dispersal across strong environmental gradients, such as freshwater-marine transitions (87) or hydrothermal vents (387). Moreover, in dynamic ecosystems with changing geochemical conditions past population growth rates can influence future community structure and biomolecular patterns and hence, cross-sectional community profiles may not reflect current dynamics (430). In such systems, an incorporation of multiple layers of geochemical and biological information into a mechanistic framework — as demonstrated here — will be crucial for disentangling the multitude of processes underlying experimental observations.

The gene-centric model constructed here, although evaluated at steady state, is in fact a spatiotemporal model that could, in principle, predict gene population dynamics and process rates over time. A spatiotemporal analysis of the model would require multi-omic time series coverage and knowledge of non-stationary physical processes, such as seasonal patterns in surface productivity and hydrodynamics during deep water renewal events, which were unavailable at the time of this study. Multi-omic time series would be especially useful for improving the mRNA and protein models introduced here, due to the high number of cur-
rently unknown parameters. For example, integrating metatranscriptomic, metaproteomic and geochemical time series during rapid environmental changes into our model would allow for a more direct determination of in situ transcriptional and translational responses and biomolecule decay times. Advances in multi-omic sequencing will undoubtedly lead to increased spatiotemporal coverage in the future (241), thus allowing for the much needed integration of such data into biogeochemical models.

The multi-omic profiles that we used to validate our model are only given in terms of relative — rather than absolute — biomolecule concentrations. Hence, the observed abundance of each biomolecule may be influenced by the abundances of other biomolecules, which could explain some of the discrepancies between the model and the multi-omic data. Unfortunately, this limitation is currently pervasive across environmental shotgun sequencing studies, largely due to technical challenges in estimating in situ DNA, mRNA and protein concentrations (376). These challenges will likely be overcome in the future (428), and this will undoubtedly improve model testing and refinement. Given this current caveat, the general agreement of the model with the shape of the multi-omic profiles (Fig. 7.3A) is all too remarkable, and suggests that the spatial structuring of the metabolic network is well captured by the model. In fact, our qPCR-based estimates for absolute SUP05 cell concentrations are consistent with absolute PDNO gene concentrations predicted by the model, as well as with the shape of the PDNO abundance profiles in the metagenomes (Fig. 7.3A). This double agreement suggests that — at least for PDNO — both our model as well as our metagenomic data sets reflect the actual gene distributions.

7.5 Conclusions

Most major metabolic pathways driving global biogeochemical cycles are encoded by a core set of genes, many of which are distributed across distant microbial clades (116). These genes are expressed and proliferate in response to specific redox conditions and, in turn, shape Earth’s surface chemistry. Here we have shown that the population dynamics of genes representative of specific metabolic pathways, their expression and their catalytic activity at ecosystem scales can all be integrated into a mechanistic framework for understanding coupled carbon, nitrogen and sulfur cycling in OMZs. This framework largely explained DNA, mRNA and protein concentration profiles and resolved several previous uncertainties in metabolic network structure in Saanich Inlet, including a direct coupling of sulfide-driven denitrification and anammox through “leaky” nitrite production by SUP05, as well as the presence of a metabolic niche for nitrous oxide reduction contributing to nitrogen loss. Be-
yond OMZ-specific predictions, model results indicate that geochemical fluxes are robust indicators of microbial community structure and, reciprocally, that gene abundances and geochemical conditions largely determine gene expression patterns. Such integrated modeling approaches offer the potential to understand microbial community metabolic networks and to predict the responses of elemental cycles in a changing world.

Figure 7.1: **Metabolic network and selected chemical time series.** (A) Coupled carbon, nitrogen and sulfur redox pathways considered in the model: Remineralization of organic matter (ROM), aerobic ammonia oxidation (amo), aerobic nitrite oxidation (nxr), anaerobic ammonia oxidation (hzo), as well as partial denitrification to nitrous oxide (PDNO) and reduction of nitrous oxide (nosZ) coupled to hydrogen sulfide oxidation (see Methods for details). Major taxonomic groups encoding specific pathways are indicated. (B) Water column oxygen (O$_2$), nitrate (NO$_3^-$) and hydrogen sulfide (H$_2$S) concentrations measured at Saanich Inlet station SI03 from January 2008 to December 2011. The shaded interval and the dates at the bottom indicate the chemical measurements that were used for model calibration. The vertical white line marks the time of molecular sampling. The model considers depths between 100 m and 200 m.
Figure 7.2: Measured and predicted geochemical profiles. (A) oxygen, (B) ammonium, (C) nitrate, (D) nitrite, (E) nitrous oxide and (F) hydrogen sulfide concentrations as predicted by the calibrated model at steady state (thick blue curves). Dots: Data used for the calibration, measured during cruise 41 on January 13, 2010 (SI041_01/13/10, rectangles), cruise 42 (SI042_02/10/10, rhomboids) and cruise 43 (SI043_03/10/10, triangles). Oxygen profiles were not available for cruises 41 and 43, hence data from cruise 44 (SI044_04/07/10, stars) were used instead. Thin black curves: Data measured during cruise 47 (SI047_07/07/10), shortly before deep water renewal. Details on data acquisition in Appendix F.2.
Figure 7.3: **Molecular and rate profiles.** (A) Predicted DNA, mRNA and protein concentrations (rows 1–3) for ROM, *amo, nxr, norBC, hzo* and *nosZ* genes (thick curves), compared to corresponding metagenomic, metatranscriptomic and metaproteomic data (circles, February 10, 2010). The dashed curve under PDNO genes (row 1, column 4) shows concurrent qPCR-based cell count estimates for SUP05, the dominant denitrifier in Saanich Inlet. (B) Denitrification and anammox rates predicted by the model (thick blue curves), compared to rate measurements (circles) during cruises 47 (SI047 07/07/10) and 48 (SI048 08/11/10), as well as rates estimated from geochemical concentration profiles using inverse linear transport modeling (ILTM; Appendix F.5). The ILTM estimates “calibr.” in the 3rd and 6th plot are based on the same geochemical data as used for model calibration (Fig. 7.2).
Chapter 8

Reaction-centric modeling of microbial community metabolism

8.1 Synopsis

The growth of microbial populations catalyzing biochemical reactions leads to positive feedback loops and self-amplifying process dynamics at ecosystem scales. Hence, the state of a biocatalyzed process is not completely determined by its physicochemical state but also depends on current cell or enzyme concentrations that are often unknown. Here we propose an approach to modeling reaction networks of natural and engineered microbial ecosystems that is able to capture the self-amplifying nature of biochemical reactions without explicit reference to the underlying microbial populations. This is achieved by an appropriate combination of parameters and variables, that allows keeping track of a system’s “capacity” to perform particular reactions, rather than the cell populations actually catalyzing them. Our reaction-centric approach minimizes the need for cell-physiological parameters such as yield factors and provides a suitable framework for describing a system’s dynamics purely in terms of chemical concentrations and fluxes. We demonstrate our approach using data from an incubation experiment involving urea hydrolysis and nitrification, as well as time series from a long-term nitrifying bioreactor experiment. We show that reaction-centric models can capture the dynamical character of microbially catalyzed reaction kinetics and enable the reconstruction of bioprocess states using solely chemical data, hence reducing the need for laborious biotic measurements in environmental and industrial process monitoring.

\footnote{A version of this chapter has been published (see the Preface for author contributions): Louca, S., Doebeli, M. 2016. \textit{Reaction-centric modeling of microbial ecosystems}. Ecological Modelling. 335:74–86. DOI:10.1016/j.ecolmodel.2016.05.011}
8.2 Introduction

Microbial metabolism powers biochemical fluxes in natural and engineered ecosystems (116, 312). Reciprocally, biochemical fluxes sustain biosynthesis and thus drive microbial population dynamics (210). Changes in the microbial populations, in turn, influence the reaction kinetics at ecosystem scales because system-wide reaction rates depend not only on substrate concentrations but also on the density of catalyzing cells or of extracellular enzymes (426). Thus, the dynamics of microbial communities emerge from the continuous interplay between metabolic activity, changes in the extracellular metabolite pool and microbial population growth (433). In particular, and in contrast to purely abiotic chemical processes (297), the state and future trajectory of a biocatalyzed process cannot be determined solely based on the system’s chemical state (210, 426). For example, empirical mineralization curves that describe the degradation rate of organic matter as a function of substrate density can vary strongly in shape, and this variation historically resulted partly from the interaction of substrate concentrations and cell population densities in experiments (426).

In deterministic or stochastic differential equation models (233, 390, 433), the dynamical character of microbially catalyzed reaction kinetics is typically incorporated by including additional variables representing cell densities, whose growth is proportional to the rates of the processes that they catalyze and determined by cell-per-substrate (or sometimes biomass-per-substrate) yield factors (210). In turn, system-wide reaction kinetics are modulated by current cell densities and extracellular metabolite concentrations. Such cell-centric models are widely used and can capture the typical self-amplifying character of biocatalyzed processes (68). Likewise, deterministic as well as stochastic individual-based models, which keep track of multiple individual organisms and their metabolic activity, can also capture the feedback loops within microbial metabolic networks because the metabolic or trophic activity of each organism eventually leads to the production of new copies of that organism (124, 258). All of these cell-centric models, however, depend on physiological parameters such as yield factors, cell masses or maximum cell-specific reaction rates, and require knowledge of cell or enzyme concentrations (in addition to physicochemical variables) for describing a system’s current state. As we explain below, some of these parameters also introduce redundancies from a reaction kinetic point of view that can lead to strong uncertainties in parameter estimation (242, 426).

Flux-balance models, a popular alternative to dynamical models (354), reduce the number of required parameters by ignoring cell population dynamics and by assuming that metabolite concentrations are constant through time (i.e., metabolite fluxes are “balanced”). In these
models, reaction rates (and sometimes metabolite turnover rates; 72) are the only independent variables, and their values are calculated by optimizing some objective function (e.g., ATP production) in the presence of constraints (e.g., on maximum reaction rates). Flux balance models have been very successful in elucidating metabolic network properties such as the feasibility of certain reactions or the prediction of metabolic interactions between species (238, 453, 545) but — being steady-state models — they fail to capture the dynamical nature of microbial communities. Hence, current model frameworks either ignore the temporal and self-amplifying character of biocatalyzed processes or require an extensive set of — often poorly estimated — physiological parameters.

To address the above limitations, here we develop a new framework for dynamical bioprocess modeling with a focus on system-wide reaction kinetics. Our objective was to reduce the reliance on physiological parameters and to reduce the need for biotic measurements for state reconstruction and model calibration, while still accounting for the self-amplifying character of metabolic reactions at the ecosystem level. Such a “reaction-centric” model would ideally make predictions purely in terms of metabolite concentrations and reaction rates at the ecosystem level, without the need to consider the underlying cell populations. As we show below, this can be achieved by keeping track of a system’s “capacity” to perform particular reactions (or pathways), rather than the cell populations actually catalyzing them. Microbial ecosystem metabolism can then be described similarly to abiotic reaction networks, with the addition of so-called self- and cross-amplification factors between reactions. These amplification factors are specific to a particular microbial community and translate the system’s metabolic fluxes into changes of the system’s reaction capacities. Hence, a system’s state and dynamics can be inferred using solely physicochemical measurements, bypassing laborious biotic measurements for example in environmental and industrial process monitoring. Furthermore, reaction-centric models minimize the reliance on cell-physiological parameters, allowing for model calibration even when biotic data are scarce. Reaction-centric models thus provide an elegant alternative to many conventional cell-centric models, particularly when the ultimate focus is on a system’s reaction kinetics.

We begin with a derivation of the reaction-centric framework and show how it relates to conventional, cell-centric models. We focus on differential equation models, however we note that our reasoning can also be applied to other cell-centric frameworks. We demonstrate the potential of reaction-centric models using data from a previous short-term incubation experiment with a ureolytic and nitrifying microbial community (94), as well as long-term time series from a flow-through nitrifying bioreactor (109). Bioreactors provide ideal model ecosystems for testing new theories for microbial ecology, due to their higher controllability.
and measurability when compared to natural ecosystems. Ureolysis and nitrification were chosen as examples because of their conceptual simplicity as well as their great relevance to ecosystem productivity, industry and agriculture (375, 520). Our entire analysis was performed with MCM (Chapter 4; 284), which we extended to accommodate reaction-centric models.

8.3 Derivation of the reaction-centric framework

8.3.1 One reaction per cell

Conventional cell-centric microbial ecosystem models consider the extracellular concentrations of metabolites as well as the cell densities of microbial populations catalyzing various reactions. In the simplest and most common case each reaction is catalyzed by a distinct microbial population, the growth of which is proportional to the rate of the reaction (210, 258, 426). More precisely, the population density of cells catalyzing reaction \( r \) (\( N_r \), cells per volume) and the concentration (\( C_m \)) of each metabolite \( m \) are described by differential equations similar to the following:

\[
\frac{dN_r}{dt} = N_r Y_r V_r h_r(C) - \lambda_r N_r, \quad (8.3.1)
\]

\[
\frac{dC_m}{dt} = F_m(t, C) + \sum_r S_{mr} N_r V_r h_r(C). \quad (8.3.2)
\]

In Eq. (8.3.1), \( Y_r \) is a cell yield factor (cells produced per substrate used), \( V_r \) is the maximum cell-specific reaction rate (flux per cell per time) and \( C \) is the vector representing all metabolite concentrations (overview of symbols in Table 8.1). We note that in models where \( N_r \) is alternatively measured in biomass (rather than cells) per volume, \( Y_r \) is typically a biomass yield factor and \( V_r \) is a maximum biomass-specific reaction rate. The dependence of cell-specific reaction kinetics on \( C \) is encoded by the unitless function \( h_r(C) \), which is normalized to unity at those \( C \) that maximize the cell-specific reaction rate. The last term in Eq. (8.3.1) corresponds to the decay of biomass at an exponential rate \( \lambda_r \) (with units time\(^{-1}\)), for example due to cell death. Alternatively, \( \lambda_r \) can account for reduced biosynthesis due to maintenance energy requirements, in which case it is sometimes called the “specific maintenance rate” (210). In Eq. (8.3.2), \( F_m \) accounts for abiotic metabolite fluxes, such as substrate supply in a bioreactor, and \( S_{mr} \) is the stoichiometric coefficient of metabolite \( m \) in reaction \( r \). The sum in Eq. (8.3.2) iterates through all reactions and accounts for microbial metabolic fluxes.
In the above cell-centric model the system’s state depends on the current metabolite concentrations ($C_m$) as well as the current cell densities ($N_r$), the prediction of which, in turn, requires knowledge of physiological parameters such as $Y_r$ and $V_r$. As we show below, this focus on cell populations can be avoided if one is solely interested in the system’s reaction kinetics. Observe that the product $M_r = N_r V_r$, henceforth referred to as the system’s current “reaction capacity”, is the maximum system-wide rate of reaction $r$ (flux per volume per time) that could possibly be attained at favorable metabolite concentrations (i.e., when $h_r(C) = 1$). On the other hand, the product $H_r = N_r V_r h_r = M_r h_r$ is the actual system-wide rate of reaction $r$. Note that $H_r$ depends both on the reaction capacity $M_r$ as well as the normalized kinetics $h_r(C)$, which encodes the dependence of the reaction rate on the system’s chemical state. Rewriting Eqs. (8.3.1) and (8.3.2) in terms of the reaction capacities $M_r$ yields the reaction-centric model

$$\frac{dM_r}{dt} = A_r M_r h_r(C) - M_r \lambda_r,$$  \hspace{1cm} (8.3.3)

$$\frac{dC_m}{dt} = F_m(t, C) + \sum_r S_{mr} H_r(C),$$  \hspace{1cm} (8.3.4)

$$H_r = M_r h_r,$$  \hspace{1cm} (8.3.5)

where we introduced the so called self-amplification factor $A_r = V_r Y_r$ in Eq. (8.3.3). This model describes biochemical reactions at the scale of the ecosystem, without explicit reference to biotic quantities such as cell densities or physiological parameters such as yield factors.

The structure of Eq. (8.3.3) emphasizes the self-amplifying nature of biochemical reactions at the ecosystem level, with the self-amplification factors $A_r$ mediating the conversion of reaction rates to a growth of reaction capacities. In the context of cell-centric models, $A_r$ is the maximum specific growth rate of cells performing reaction $r$ (in units time$^{-1}$). In the reaction-centric model, $A_r$ becomes the maximum exponential growth rate of the reaction capacity $M_r$. Note that $A_r$ only depends on the product $V_r Y_r$, but not on the individual $V_r$ or $Y_r$. Hence, the system’s biochemical dynamics can be modeled without knowledge of the $V_r$ and $Y_r$ because the system’s trajectory is completely determined by the self-amplification factors and the reaction capacities at some point in time. This collapse of unknown parameters into fewer ones, without loosing any predictive power, means that fewer parameters are needed for practical purposes than often assumed. In fact, the redundancy inherent to the simultaneous inclusion of $V_r$ and $Y_r$ in conventional models was previously pointed out by Simkins and Alexander (426). This redundancy can lead to strong negative correlations between estimated $Y_r$ and $V_r$, particularly when parameter estimation is based solely on non-
biotic chemical time series, because such time series cannot differentiate between alternative combinations of $V_r$ and $Y_r$ yielding the same product $V_rY_r$ (242).

### 8.3.2 Multiple reactions per cell

So far we assumed that each cell performs exactly one reaction, which means that each modeled reaction only induces the growth of its own capacity. While this assumption is widespread in ecosystem modeling (258, 386), in reality several alternative pathways may be performed by the same cells. For example, members of the ammonium oxidizing *Nitrosospira* genus are also able to hydrolyze urea (300), and urea hydrolysis in incubation experiments with *Nitrosospira* was shown to promote ammonium oxidation by the same population (94). In the simplest case, the combined effects of several metabolic pathways on cell population growth can be assumed to be additive, so that each reaction $r$ has a contribution $Y_rH_r$ to the total growth of the cell population:

$$\frac{dN_r}{dt} = Y_rH_r + \sum_{q \neq r} Y_qH_q - N_r\lambda_r.$$  \hfill (8.3.6)

Here, the sum in Eq. (8.3.6) iterates over all additional reactions attributable to cells performing reaction $r$. If two reactions $r$ and $q$ are performed by the same population then $N_r = N_q$ and, reciprocally, if two populations share a common reaction, that reaction will need to be represented twice using two separate indices $r$. The assumption of additive effects on growth is common in conventional microbial population models. For example, Courtin and Spoelstra (80) model a population of acetic acid bacteria utilizing multiple organic substrates by assuming that each pathway has an additive effect on the total population growth. More sophisticated models of microbial metabolism based on flux balance analysis and optimization of a linear utility function also assume additive effects of various metabolic fluxes, although the functions $h_r(C)$ may not be explicit in $C$, but instead specified in terms of an optimization algorithm (354).

The cell-centric model in Eq. (8.3.6) corresponds to a reaction-centric model in which multiple reactions amplify each other’s capacities whenever they are performed by the same cells:

$$\frac{dM_r}{dt} = A_rH_r + \sum_{q \neq r} A_{rq}H_q - \lambda_rM_r.$$  \hfill (8.3.7)

Here, the so-called “cross-amplification” factors $A_{rq} = V_rY_q$ correspond to the positive effects...
of the flux through some reaction \( q \) on the capacity of some other reaction \( r \) and hence, the sum in Eq. (8.3.7) iterates through all additional reactions \( q \) performed by the same cell population as reaction \( r \). The amplification matrix, whose diagonal entries are the self-amplification factors \( A_r \) and whose off-diagonal entries are the cross-amplification factors \( A_{rq} \), defines a linear transformation of the vector containing all reaction rates to a vector containing changes in reaction capacities. Note that regardless of any amplifications of the reaction capacities, actual rates may still be limited by low substrate concentrations or the presence of inhibitors, as determined by the normalized kinetics \( h_r(C) \). Also note that since \( A_{rq} = V_r Y_q \) and \( N_r = N_q \) for any two reactions \( q \) and \( r \) performed by the same cells, the following consistency conditions apply:

\[
A_{qr} = \frac{A_r A_q}{A_{rq}}, \quad M_r = M_q \frac{A_{rq}}{A_q}. \tag{8.3.8}
\]

Regardless of any cell-centric interpretation, the system’s reaction dynamics only depend on the amplification factors \( A_{rq} \), but not on any \( Y_q \) or \( V_r \).

The above discussion illustrates how conventional cell-centric models can be used to derive reaction-centric models and foster confidence in their realism. For example, amplification factors can be seen as a combination of — and a replacement for — cell-centric parameters. However, as we demonstrate below, in practice a reaction-centric model can be taken as an alternative self-contained description of a system’s reaction kinetics. Under such a paradigm, the amplification matrix becomes a set of standalone system-specific parameters and the reaction capacities become independent state variables whose dynamics are shaped by the amplification matrix. We note that while here we focus on linear growth dynamics, nonlinear generalizations are also possible with additional amplification coefficients mediating the higher order effects of biochemical fluxes on reaction capacities.

Apart from the elegance of a reaction-centric description, an added benefit is that all parameters and state variables can be inferred from purely physicochemical time series. For example, at high substrate concentrations and in the absence of inhibitors, reaction capacities \( (M_r) \) are approximately equal to actual reaction rates \( (H_r) \) and can thus be estimated directly from the derivative (slope) of chemical concentration time series. Similarly, if the normalized reaction kinetics \( h_r \) (or equivalently, the half-saturation constants in case of Monod kinetics) are known, then reaction rates estimated from concentration time series can be divided by \( h_r \) to yield the reaction capacities. In general, however, reaction capacities may constitute unknown system variables which must be estimated indirectly, for example by repeated observation of metabolite concentrations (as demonstrated below).
8.4 Demonstration of the reaction-centric framework

8.4.1 Overview

To exemplify our approach, we constructed reaction-centric models for two separate engineered microbial ecosystems used in previously published experiments. Specifically, in the first example we consider urea hydrolysis and nitrification in a batch-fed incubation experiment previously described by de Boer and Laanbroek (94). The structure of our model, described in detail below, was chosen to closely resemble the physicochemical conditions in the experiment as well as the metabolic network involved in dissimilatory nitrogen transformations — as inferred from the experiment. We test the adequacy of our reaction-centric model by assessing its “goodness of fit” after calibrating unknown parameters to the experimental data. Further, we demonstrate the importance of cross-amplification factors for accounting for pathway co-occurrences in cells by comparing two variants of the model, namely one variant with and one variant without the cross-amplification factors.

In the second example we consider a reaction-centric model for a flow-through ammonium-fed nitrifying bioreactor, operated under varying conditions over the course of several months (109). Similarly to the first example, our model is constructed to closely resemble the physicochemical conditions of the bioreactor. In this example, we demonstrate how purely chemical time series can be used to calibrate a reaction-centric model and to infer the full biochemical state of the bioreactor (i.e., $M_r$ and $C_m$) in “real-time”. In addition, to further assess the fidelity of the model, we use independent biomass concentration measurements from the original experiment, which we compare to the hypothetical biomass concentrations that would correspond to the reaction capacities in the reaction-centric model.

All time series analysis, simulations and parameter calibrations in this study were performed using MCM (Chapter 4; 284). The construction and analysis of the models in MCM is explained in Appendix G.

8.4.2 Example 1: Urea hydrolysis and nitrification in a batch-fed incubator

Overview of experimental results

The microbial community in the incubator was dominated by *Nitrosospira* sp., which are ammonium oxidizing bacteria (AOB) also capable of hydrolyzing urea to ammonium, and
Nitrobacter sp., which are nitrite oxidizing bacteria (NOB; Fig. 8.1A). The incubator was batch-fed with urea, the complete hydrolysis of which by the AOB led to a temporary accumulation of ammonium ($\text{NH}_4^+$) within roughly one week. Concurrently to its production, $\text{NH}_4^+$ was also oxidized by the AOB into nitrite ($\text{NO}_2^-$), which was in turn oxidized by the NOB into nitrate ($\text{NO}_3^-$). Nitrification continued after complete urea hydrolysis until $\text{NH}_4^+$ concentration dropped to about 0.5 mM. The high energy requirements for maintaining a more neutral internal pH than the external environment (pH = 5) could presumably not be met at lower $\text{NH}_4^+$ concentrations, eventually leading to a halt of nitrification (94).

**Inferred model structure**

The model focuses on dissimilatory nitrogen fluxes encompassing urea hydrolysis ($\text{ure}$), ammonium oxidation ($\text{amo}$) and nitrite oxidation ($\text{nxr}$). All nitrogen metabolism is assumed to be entirely dissimilatory. Specifically, we assume that each mol urea is converted by $\text{ure}$ to 2 mol $\text{NH}_4^+$, of which a small fraction $\rho_{\text{ure}}$ is immediately oxidized ("recycled") to $\text{NO}_2^-$ within the same cell, while the remaining $\text{NH}_4^+$ leaks to the extracellular medium:

$$
\text{ure} : \text{urea} \rightarrow 2(1 - \rho_{\text{ure}}) \times \text{NH}_4^+ + 2\rho_{\text{ure}} \times \text{NO}_2^-.
$$

We assume that extracellular $\text{NH}_4^+$ taken up by AOB is completely oxidized to $\text{NO}_2^-$, and that all $\text{NO}_2^-$ taken up by NOB is completely oxidized to $\text{NO}_3^-$:

$$
\text{amo} : \text{NH}_4^+ \rightarrow \text{NO}_2^-, \quad \text{nxr} : \text{NO}_2^- \rightarrow \text{NO}_3^-.
$$

The recycling term $\rho_{\text{ure}}$ was included in order to explain the early appearance of $\text{NO}_3^-$ in the incubator (Fig. 8.2C). Despite the increased model complexity (one additional free parameter), preliminary statistical model selection tests (based on AIC and BIC; 246) showed a clear preference for the inclusion of $\rho_{\text{ure}}$ (Supplemental Fig. G.1). $\text{amo}$ rates were assumed to be limited by ammonia ($\text{NH}_3$) concentrations, rather than $\text{NH}_4^+$ concentrations, in accordance to findings by Suzuki et al. (463). Due to a lack of further information, potential oxygen limitation in the incubator was ignored.

The co-occurrence of $\text{ure}$ and $\text{amo}$ genes in the same AOB cells leads to a direct coupling of the population dynamics of these genes and enzymes, and therefore the incubator’s $\text{amo}$ and $\text{ure}$ reaction capacities (Fig. 8.1B). In the model, this coupling corresponds to positive cross-amplification factors that measure the mutual effects of $\text{ure}$ flux on $\text{amo}$ capacity and vice versa. Hence, based on the model structure introduced in Section 8.3.2, the differential
Reaction-centric modeling of community metabolism

equations for the reaction capacities $M_{ure}$, $M_{amo}$ and $M_{nxr}$ take the form

$$\frac{dM_{ure}}{dt} = A_{ure} \cdot H_{ure} + A_{ure,amo} \cdot H_{amo} - M_{ure} \cdot \lambda_{AOB}, \quad (8.4.3)$$

$$\frac{dM_{amo}}{dt} = A_{amo} \cdot H_{amo} + A_{amo,ure} \cdot H_{ure} - M_{amo} \cdot \lambda_{AOB}, \quad (8.4.4)$$

$$\frac{dM_{nxr}}{dt} = A_{nxr} \cdot H_{nxr} - M_{nxr} \cdot \lambda_{NOB}. \quad (8.4.5)$$

Preliminary tests indicated that $nxr$ decay could be omitted from the model because within the time span of the experiment NOB cell densities were mostly limited by NO$_2^-$ supply, hence on grounds of parsimony we set $\lambda_{NOB} = 0$. On the other hand, our tests indicated that the term $\lambda_{AOB}$ was mostly attributable to AOB maintenance rates (210) that caused a reduced growth of $ure$ and $amo$ compared to a simple proportionality with respect to $ure$ and $amo$ rates. These maintenance requirements result in a substrate threshold below which dissimilatory metabolism can no longer sustain growth. That threshold is reached when

$$A_{amo}h_{amo} + A_{ure}h_{ure} \leq \lambda_{AOB}, \quad (8.4.6)$$

at which point we assumed a complete halt of $ure$ and $amo$ activity (Appendix G.2.1). Note that care needs to be taken to ensure consistency between the cross-amplification terms $A_{amo,ure}$ and $A_{ure,amo}$, as well as between the initial reaction capacities $M^o_{ure}$ and $M^o_{amo}$. As explained previously, we need to have

$$A_{ure,amo} = \frac{A_{amo}A_{ure}}{A_{amo,ure}}, \quad M^o_{amo} = M^o_{ure} \frac{A_{amo,ure}}{A_{ure}}. \quad (8.4.7)$$

The normalized reaction kinetics $h_{ure}$, $h_{amo}$ and $h_{nxr}$ are Monod-functions of substrate concentrations (211), i.e., linear at lower and saturating at higher concentrations:

$$h_{ure} = \frac{C_{urea}}{K_{ure} + C_{urea}}, \quad h_{amo} = \frac{C_{NH_3}}{K_{amo} + C_{NH_3}}, \quad h_{nxr} = \frac{C_{NO_2^-}}{K_{nxr} + C_{NO_2^-}}. \quad (8.4.8)$$

Here, $K_{ure}$, $K_{amo}$ and $K_{nxr}$ are half-saturation constants. Note that no cell yield factors or cell-specific rates appear in the model; instead, growth dynamics are completely captured by the cross-amplification factors $A_{ure}$, $A_{amo}$, $A_{nxr}$ and $A_{amo,ure}$. 

101
Model goodness of fit

We fixed 4 out of 11 model parameters to values from the literature. For example, the self-amplification factors $A_{amo}$ and $A_{nxr}$ were set to 1.2 d$^{-1}$ and 1.03 d$^{-1}$, according to typical maximum growth rates of *Nitrosospira* (30) and *Nitrobacter* (227), respectively. The initial *ure* rate, $M_{ure}^0$, was determined directly from the slope of the urea time series at time $t = 0$, eliminating the need for initial cell counts as in conventional cell-centric models. The remaining free parameters were calibrated to the experimental data using a maximum-likelihood approach (see Methods for details and Table G.1 for parameter values).

Upon calibration, the model largely explains the experimental data and is able to capture the self- and cross-amplifying character of the incubator’s dynamics (Fig. 8.2). In particular, the *ure-amo* cross-amplification causes an increase of the system’s *amo* capacity during urea hydrolysis, even when *amo* rates are still slow. This results in a fundamentally different behavior of the system than could have been explained by a cell- or reaction-centric model not accounting for the co-occurrence of *ure* and *amo* in the same cells. To verify this interpretation, we tested a variation of the model in which the cross-amplification factors $A_{ure,amo}$ and $A_{amo,ure}$ were set to zero. In this model variant, the initial capacities $M_{ure}^0$ and $M_{amo}^0$ became independent parameters. Similarly, $\lambda_{AOB}$ was split into two independent maintenance rates, $\lambda_{amo}$ and $\lambda_{nxr}$. The resulting larger set of free parameters was fitted to the same data as above. This model variant was unable to explain the $\text{NH}_4^+$ and $\text{NO}_3^-$ time series, despite the higher number of calibrated parameters (Supplemental Fig. G.2).

We concluded that the early increase of *amo* reaction capacity cannot be explained solely on grounds of *amo* self-amplification but was indeed partly fueled by *ure* activity. This highlights the importance of taking into account pathway co-occurrences and interactions in cells and suggests that cross-amplification factors in reaction-centric models may be an adequate means to that end.

### 8.4.3 Example 2: Nitrification in a flow-through bioreactor

#### The problem of state reconstruction

In principle, reaction-centric models predict future system trajectories ($M(t)$, $C(t)$) given initial conditions ($M(0)$, $C(0)$). In practice, uncertainty in initial conditions or model parameters, as well as neglected secondary processes, lead to uncertainties in the predicted system state that can increase with time. Selected measurements can provide crucial information to ensure model proximity to reality, however typically only a subset of state variables may be
measurable. Inferring a system’s full state from a smaller set of observations is a common problem, for example in oceanography or engineering, and generally multiple sequential measurements are used to gradually improve state reconstruction and model predictions (32, 53). In this example we demonstrate how long-term, purely abiotic chemical time series can be combined with a reaction-centric model in order to infer the full state of a bioreactor in real time.

**Model structure**

The model describes a flow-through ammonium-fed nitrifying bioreactor, resembling the experimental setup by Dumont et al. (109, Bioreactor B). In our model we assume that each mol \( \text{NH}_4^+ \) is oxidized to one mol \( \text{NO}_2^- \) by \( \text{amo} \) and subsequently to one mol \( \text{NO}_3^- \) by \( \text{nxr} \) (520). The model thus keeps track of the bioreactor’s \( \text{amo} \) and \( \text{nxr} \) reaction capacities as well as extracellular \( \text{NH}_4^+ \), \( \text{NO}_2^- \) and \( \text{NO}_3^- \) concentrations over time:

\[
\frac{dM_{\text{amo}}}{dt} = A_{\text{amo}} H_{\text{amo}} - M_{\text{amo}} \cdot \mu, \tag{8.4.9}
\]

\[
\frac{dM_{\text{nxr}}}{dt} = A_{\text{nxr}} H_{\text{nxr}} - M_{\text{nxr}} \cdot \mu, \tag{8.4.10}
\]

\[
H_{\text{amo}} = M_{\text{amo}} h_{\text{amo}}, \quad H_{\text{nxr}} = M_{\text{nxr}} h_{\text{nxr}}, \tag{8.4.11}
\]

\[
\frac{dC_m}{dt} = S_{m,\text{amo}} H_{\text{amo}} + S_{m,\text{nxr}} H_{\text{nxr}} + (C_m^{\text{in}} - C_m) \cdot \mu. \tag{8.4.12}
\]

Here, \( C_m \) is the concentration of the \( m \)-th metabolite (\( \text{NH}_4^+ \), \( \text{NO}_2^- \) or \( \text{NO}_3^- \)), \( S_{m,\text{amo}} \) and \( S_{m,\text{nxr}} \) are the known (520) stoichiometric coefficients for metabolite \( m \) in \( \text{amo} \) and \( \text{nxr} \), respectively, \( \mu \) is the bioreactor’s controlled dilution rate (causing the bulk of biomass decay in the bioreactor), and \( C_m^{\text{in}} \) is the metabolite concentration in the input medium (zero for all metabolites except \( \text{NH}_4^+ \)). During the original experiment, the dilution rate as well as the input \( \text{NH}_4^+ \) concentration were varied on several occasions (Figs. 8.3C,D), resulting in non-equilibrium bioreactor dynamics. Hence, in our model both \( \mu \) and \( C_m^{\text{in}} \) depend explicitly on time in the same way as in the original experiment (Figs. 8.3C,D). The normalized reaction kinetics, \( h_{\text{amo}} \) and \( h_{\text{nxr}} \), are Monod-functions of \( \text{NH}_3 \) and \( \text{NO}_2^- \) concentrations, respectively, as in the previous example.

**Model calibration and “real-time” state reconstruction**

The concentrations of \( \text{NH}_4^+ \), \( \text{NO}_2^- \) and \( \text{NO}_3^- \) in the bioreactor were monitored throughout, providing a subset of the bioreactor’s state variables. The remaining state variables (i.e., the
reaction capacities) were inferred through gradual assimilation of these time series into the model, as follows. At each point in time the rates of change of NH$_4^+$ and NO$_3^-$ concentrations (inferred from the NH$_4^+$ and NO$_3^-$ time series) were used to infer the reaction rates ($H_{amo}$ and $H_{nxr}$), while accounting for the known reaction stoichiometries and the part explained by the known dilution and substrate supply rates (Figs. 8.3C,D). Next, we inserted the inferred reaction rates into Eqs. (8.4.9) and (8.4.10) to predict the growth rates of $amo$ and $nxr$ capacities that would correspond to these reaction rates:

$$
\frac{d\hat{M}_r}{dt} = A_r H_r - \hat{M}_r \cdot \mu.
$$

Integrating Eq. (8.4.13) over time yields estimates, $\hat{M}_{amo}(t)$ and $\hat{M}_{nxr}(t)$, for the reaction capacities (Figs. 8.3H,I). Due to the decay rate $\mu$, any initial discrepancies between the estimated and true capacities quickly decay exponentially regardless of initial conditions, provided that model parameters are correctly chosen (see below):

$$
\frac{d}{dt}(\hat{M}_r - M_r) = -\mu \cdot (\hat{M}_r - M_r).
$$

This method of gradual state reconstruction (Fig. 8.3B) is analogous to the use of so-called “observers” in control theory, which gradually approach the system’s unknown state with time by combining sequential observations with concurrent model predictions (434). In general, finding appropriate observers for the available data and ensuring their convergence can be challenging, and our example shows that the special structure of reaction-centric models mitigates this problem. Note that the temporal resolution of the chemical data, as opposed to single snapshots, is key to estimating the reaction rates needed for full-dimensional state reconstruction (Fig. 8.4). We note that our reaction-centric approach presents an alternative to the approach taken in the original experiment, where biomolecular time series data are assimilated into a cell-centric model (108, 109).

To validate the estimated bioreactor state, we used Eqs. (8.4.11,8.4.12) to predict the time courses of the metabolite concentrations corresponding to the estimated $\hat{M}_r$, and these predictions were then compared to the measured NH$_4^+$, NO$_2^-$ and NO$_3^-$ concentrations. The $amo$ and $nxr$ half-saturation constants ($K_{amo}$ and $K_{nxr}$), as well as the self-amplification factors ($A_{amo}$ and $A_{nxr}$), were a priori unknown and were calibrated via least-squares fitting of the predicted metabolite concentrations to the data (see Appendix G.1.2 for details and Table G.2 for fitted values). Hence, the chemical time series were used both for model calibration as well as state reconstruction. Only data from days 1–250 were used for the calibration; the
remaining data (days 250–525) were used to assess the adequacy of the model for explaining the experimental observations.

Within the calibration period the model is able to reproduce most major patterns of NO\textsubscript{2}, NO\textsubscript{3} and, to a lesser extend, NH\textsuperscript{4} concentrations (Fig. 8.5A–C). This indicates that the bioreactor’s state is well estimated by the model during that time. The agreement between the model and the NH\textsuperscript{4} and NO\textsubscript{2} data decreases outside of the calibration period, although NO\textsubscript{3} predictions remain accurate. In particular, the model overestimates the temporary accumulation of NH\textsuperscript{4} on days 337–380, during which a higher dilution rate was applied to the bioreactor (Fig. 8.3D). An increase of residual substrate concentration at higher dilution rates, as predicted by our model, is consistent with standard bioreactor theory (312). An explanation for the absence of NH\textsuperscript{4} accumulation in the data could be the potential appearance of an alternative opportunistic ammonium oxidizer that achieves faster growth rates at high substrate concentrations, thus maintaining the residual NH\textsuperscript{4} below the model’s predictions. Indeed, this scenario is supported by molecular analyses in the original experiment, which showed that a previously rare phylotype had emerged temporarily during that period (109).

**Comparison with biomass concentration profiles**

The reaction-centric model in the above example does not, a priori, require or predict biomass concentrations or cell densities. After all, its purpose is to shift the focus towards system-wide reaction kinetics, and away from the microbial populations that catalyzed them. Nevertheless, biotic data (if available) can be used as an additional means to test the accuracy of a reaction-centric model. In the following we shall compare our model’s predicted reaction capacities (which are proportional to biomass concentrations) to independent dry biomass concentrations measured during the original experiment (109).

We assumed that the bulk of biomass can be attributed to ammonium oxidizers, an assumption typically met in practice (109, 520). It then follows that \( Y_{amo} \times \frac{M_{amo}}{A_{amo}} \) should be comparable to the biomass concentration, with \( Y_{amo} \) being an unknown biomass yield factor. Note that \( Y_{amo} \) simply rescales the predicted time profile of \( \frac{M_{amo}}{A_{amo}} \). Hence, \( Y_{amo} \) can be estimated in retrospect by choosing \( Y_{amo} \) such that \( Y_{amo} \times \frac{M_{amo}}{A_{amo}} \) best resembles the measured biomass profile. Ordinary linear least-squares fitting yields an estimate of \( Y_{amo} \approx 3.2 \text{ g dW/mol N} \) (Fig. 8.5D). This estimate is greater than typical yield factors for AOB (e.g., 2.1 g dW/mol N for *Nitrosomonas europaea*; 520), although higher yield factors have also been reported (321). Other microbial groups such as NOB or non-nitrifiers likely also contribute to total biomass, resulting in an overestimate of \( Y_{amo} \). For example,
heterotrophic bacteria were detected in the original experiment using molecular methods (109).

While the model is consistent with chemical measurements during most of the experiment as discussed previously (Fig. 8.5), it clearly overestimates biomass concentrations during days 380–420 (Fig. 8.5D). At that time, the input substrate concentration was high and the dilution rate was low (Fig. 8.3C,D), in principle allowing for high equilibrium cell densities. Previous models for this system based on molecular data show a similar discrepancy (108). Both Dumont’s and our model assume a constant yield factor, ignoring the fact that the microbial community is subject to continuous taxonomic turnover (109). Previous bioreactor experiments have repeatedly revealed rapid taxonomic turnover and fluctuations in biomass densities, despite stable metabolic performance (123, 524). This discrepancy between reaction rates and community composition is often attributed to functional redundancy within microbial communities (46; 285), and highlights an important limitation of reaction-centric models: Namely, reaction-centric models may explain ecosystem reaction rates, but they can fail to detect microbial community changes when functional performance remains stable. Multiple reaction capacities representing equivalent reactions may be included in a model to account for functional redundancy, however this will typically compromise parameter identifiability.

8.4.4 Estimating concentrations of other organic compounds

In the last example above we assessed the adequacy of our reaction-centric model using independent biomass concentration measurements by introducing an additional biomass yield factor, which related dissimilatory nitrogen fluxes to biosynthesis rates. Similarly, reaction-centric models may also predict the concentration of other organic compounds or elements, either for model validation using additional data or for addressing particular ecological questions. For example, organic nitrogen or carbon concentration profiles can yield insight into nitrogen fixation rates and productivity at ecosystem scales (62, 447). The concentrations of various compounds in living cells (e.g., organic N) can be derived from the reaction capacities using so called assimilation factors, which represent the amount of compound assimilated or synthesized per reaction flux (520). More precisely, the concentration of a particular organic compound is given by the matrix product

$$ T^T A^{-1} M, \quad (8.4.15) $$
where \( \mathbf{M} \) is the column vector containing all reaction capacities, \( \mathbf{T} \) is the column-vector containing the assimilation factors for the compound for the various reactions, \( \mathbf{T}^T \) is the transpose of \( \mathbf{T} \) and \( \mathbf{A}^{-1} \) is the inverse of the amplification matrix (see Appendix G.2.2 for a derivation). For example, the stoichiometries of N-metabolism and anabolism in the ammonium oxidizer \( \textit{N. europaea} \) and nitrite oxidizer \( \textit{Nitrobacter winogradskyi} \) are usually summarized by

\[
55\text{NH}_4^+ + 76\text{O}_2 + 109\text{HCO}_3^- \rightarrow C_5\text{H}_7\text{NO}_2 + 54\text{NO}_2^- + 57\text{H}_2\text{O} + 104\text{H}_2\text{CO}_3 \quad (8.4.16)
\]

and

\[
400\text{NO}_2^- + \text{NH}_4^+ + 4\text{H}_2\text{CO}_3 + \text{HCO}_3^- + 195\text{O}_2 \rightarrow C_5\text{H}_7\text{NO}_2 + 3\text{H}_2\text{O} + 400\text{NO}_3^- \quad (8.4.17)
\]

respectively (520). Here, \( C_5\text{H}_7\text{NO}_2 \) represents biomass. Hence, for organic N the assimilation factors are \( T_{amo} = 1 : 55 \approx 0.018 \) (1 mol N assimilated per \( \text{NH}_4^+ \) consumed) for dissimilatory ammonium oxidation

\[
amo: \text{NH}_4^+ + \frac{3}{2}\text{O}_2 \rightarrow \text{NO}_2^- + \text{H}_2\text{O} + 2\text{H}^+ , \quad (8.4.18)
\]

and \( T_{nxr} = 1 : 400 \approx 0.0025 \) for dissimilatory nitrite oxidation

\[
xxr: \text{NO}_2^- + \frac{1}{2}\text{O}_2 \rightarrow \text{NO}_3^- . \quad (8.4.19)
\]

In other cases (e.g., when stoichiometries are unknown) assimilation factors may be estimated through linear least-squares fitting, as demonstrated above for total biomass.

### 8.4.5 Limitations and extensions of reaction-centric models

The reaction-centric models presented in this study were formulated in terms of ordinary differential equations that describe the temporal evolution of the chemical and reaction-kinetic state of a well-mixed (i.e., spatially homogenous) system. Spatial extensions, for example comprising multiple interacting compartments or formulated as partial (i.e., spatiotemporal) differential equations, are equally possible. Such extensions may be used to describe the biogeochemistry in the ocean water column (386) or in multi-stage industrial processes (373).

For simplicity, we only considered Monod-type reaction kinetics, which capture the non-linear
and saturating dependence of reaction rates on single substrate concentrations, but which ignore potential substrate inhibition effects or multi-substrate dependencies. For example, excess ammonia and nitrous acid concentrations in nitrifying bioreactors can cause inhibition of the very pathways that consume them (15), and this substrate inhibition can result in reduced bioreactor performance (425). Similarly, the accumulation of metabolic products can inhibit pathway activity, e.g., by rendering pathways energetically unfavorable (256), thereby slowing down reaction rates or even causing a decline of reaction capacities due to cell death (77, 225). In reaction-centric models, substrate or product inhibition as well as multi-substrate dependencies can be incorporated through appropriate normalized reaction kinetics, \( h_r(C) \), for example in the form of multi-substrate Michaelis-Menten functions with inhibition terms (e.g. 478).

We note that reaction-centric models are not appropriate for capturing complex heterogeneities in the physiology or metabolic activity within populations that may be caused, for example, by stochastic regulatory switching (2). Simple heterogeneities, e.g., involving a small set of alternative metabolic phenotypes, may be accounted for by including multiple reactions whose capacities are coupled through cross-amplification factors. However, when variation between individual cells involves multiple traits or spans a continuum of values, individual-based models (124, 258) may be more appropriate for incorporating that variation. Moreover, reaction-centric descriptions eliminate cell-centric information (e.g., cell densities of particular species or strains) that is potentially needed to model additional community-level processes such as predation by protists (166) or bacteriophages (423). For example, bacteriophages adapted to specific bacterial taxa can exert strong control on their host populations and can drive rapid turnover of competing bacterial taxa through “killing the winner” dynamics (423, 462). Such taxonomic turnover within microbial “metabolic guilt” cannot be captured by reaction-centric models, although in several previous bioreactor experiments the overall biochemical throughput remained constant despite rapid taxonomic turnover (122, 179, 506, 524) and hence, reaction-centric models may be adequate for such systems. Other biotic interactions, such as chemical warfare (393) or quorum sensing (137) may also necessitate the use of cell-centric (e.g., individual-based) models.

### 8.5 Conclusions

Marker gene-based profiling of taxonomic community composition has become a standard tool in microbial ecology and bioengineering (109, 524). However, taxonomic profiles can lead to ambiguous conclusions about metabolic processes due to functional redundancy across mi-
microbial clades, fine-scale ecological differentiations and poor functional characterization of species (109, 224). In fact, microbial communities can have highly variable taxonomic composition while maintaining stable overall reaction rates, as has been repeatedly demonstrated in bioreactors (123, 506, 524). Furthermore, the measurement of biotic variables such as enzyme concentrations and taxonomic profiles often presents practical challenges (497). These observations motivate the pursuit for reaction-centric descriptions of microbial ecosystems that can fully utilize abiotic physicochemical data and minimize the need for laborious biotic measurements. This is particularly important in bioprocess and environmental engineering, where the need for real-time and unambiguous state reconstruction imposes strong requirements on the data (261).

Here we have shown how a reaction-centric model enables the inference of a bioreactor’s state, from a reaction kinetic point of view, based solely on chemical data. Reaction-centric models can capture the self- and cross-amplifying nature of biocatalyzed processes that so strongly sets them apart from most non-living systems. This is achieved through an amplification matrix that translates system-wide reaction rates to changes in system-wide reaction capacities. Because the amplification matrix can contain off-diagonal entries it can account for pathway co-occurrences in cells, as we have demonstrated for the case of urea hydrolysis and ammonium oxidation. Reaction-centric models share a conceptual similarity to gene-centric models (Chapter 7), in that both only consider the dynamics of pathways with disregard to the various organisms hosting a particular pathway. We note that here reaction-centric models were derived from cell-centric models, and hence their realism and accuracy are at most as good as those of cell-centric models. In principle, however, reaction-centric models may be taken as a starting point for further generalizations, e.g., using non-linear self- or cross-amplification terms; this remains an avenue for future research. The elegance of reaction-centric models makes them a potentially powerful alternative to cell-centric (80, 433) or gene-centric models for describing microbial metabolic fluxes at ecosystem scales, especially in the absence of molecular data or when the focus is entirely on reaction kinetics.
Figure 8.1: **Modeling urea hydrolysis and nitrification.** (A) Microbial ecosystem model for urea hydrolysis and subsequent nitrification by ammonium ($\text{NH}_4^+$) oxidizing bacteria (AOB) and nitrite ($\text{NO}_2^-$) oxidizing bacteria (NOB), in a batch-fed incubator. (B) Corresponding reaction-centric model comprising urea hydrolysis (ure), ammonium oxidation (amo) and nitrite oxidation (nxr) with explicit self- and cross-amplifications (continuous arrows): The flux through each reaction (dashed arrows) powers biosynthesis by the cells performing the reaction, leading to the growth of the rate capacity of that reaction and of other reactions catalyzed by the same cells.
Table 8.1: **Overview of symbols and units.** The indices $r$ and $q$ enumerate reactions or cell species, while $m$ enumerates metabolites. Parameters or variables specific to cell-centric models are indicated by “†”, those specific to reaction-centric models are indicated by “⋆”. Parameter values used in the examples are given in Tables G.1 and G.2.

<table>
<thead>
<tr>
<th>symbol and description</th>
<th>units</th>
<th>used as</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$ time</td>
<td>days (d)</td>
<td>-</td>
</tr>
<tr>
<td>$N_r$ cell density</td>
<td>† cells/L</td>
<td>independent variable</td>
</tr>
<tr>
<td>$N$ all cell densities (vector)</td>
<td>† cells/L</td>
<td>independent variable</td>
</tr>
<tr>
<td>$C_m$ metabolite concentration</td>
<td>mol/L</td>
<td>independent variable</td>
</tr>
<tr>
<td>$C$ all metabolite concentrations (vector)</td>
<td>mol/L</td>
<td>independent variable</td>
</tr>
<tr>
<td>$C_0_m$ initial metabolite concentration</td>
<td>mol/L</td>
<td>parameter</td>
</tr>
<tr>
<td>$Y_r$ cell yield factor</td>
<td>† cells/mol</td>
<td>parameter</td>
</tr>
<tr>
<td>$V_r$ maximum cell-specific reaction rate</td>
<td>† mol/(cell · d)</td>
<td>parameter</td>
</tr>
<tr>
<td>$h_r$ normalized cell-specific reaction rate</td>
<td>-</td>
<td>function of $C$</td>
</tr>
<tr>
<td>$\lambda_r$ exponential biomass decay rate</td>
<td>1/d</td>
<td>parameter</td>
</tr>
<tr>
<td>$K_r$ substrate half-saturation constant</td>
<td>mol/L</td>
<td>parameter</td>
</tr>
<tr>
<td>$F_m$ abiotic net metabolite influx</td>
<td>mol/(L · d)</td>
<td>function of $t$ and $C$</td>
</tr>
<tr>
<td>$S_{mr}$ stoichiometric coefficient</td>
<td>-</td>
<td>parameter</td>
</tr>
<tr>
<td>$H_r$ reaction rate</td>
<td>mol/(L · d)</td>
<td>dependent variable</td>
</tr>
<tr>
<td>$A_r$ self-amplification factor</td>
<td>* 1/d</td>
<td>parameter</td>
</tr>
<tr>
<td>$A_{rq}$ cross-amplification factor</td>
<td>* 1/d</td>
<td>parameter</td>
</tr>
<tr>
<td>$M_r$ reaction capacity</td>
<td>* mol/(L · d)</td>
<td>independent variable</td>
</tr>
<tr>
<td>$M$ all reaction capacities (vector)</td>
<td>* mol/(L · d)</td>
<td>independent variable</td>
</tr>
<tr>
<td>$M_{r0}$ initial reaction capacity</td>
<td>* mol/(L · d)</td>
<td>parameter</td>
</tr>
<tr>
<td>$\rho_{ure}$ ammonia recycling fraction</td>
<td>-</td>
<td>parameter</td>
</tr>
<tr>
<td>$T_r$ substrate assimilation factor</td>
<td>-</td>
<td>parameter</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>symbols specific to example 2 (flow-through bioreactor)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_{in}^m$ metabolite concentration in inflow</td>
</tr>
<tr>
<td>$\hat{M}_r$ reconstructed reaction capacity</td>
</tr>
<tr>
<td>$\hat{C}_m$ reconstructed metabolite concentration</td>
</tr>
<tr>
<td>$\mu$ hydraulic dilution rate</td>
</tr>
</tbody>
</table>
Figure 8.2: Model predictions and data for Example 1. Model predictions and data from a batch-fed incubation experiment involving urea hydrolysis and nitrification: (A) Urea, (B) ammonium and (C) nitrate concentrations over time, following incubation of a mixed *Nitrosospira* AHB1 and *Nitrobacter* NHB1 community in a urea-enriched medium. Second row: (D) urea hydrolysis (\textit{ure}) (E) ammonium oxidation (\textit{amo}) and (F) nitrite oxidation (\textit{nxr}) rates over time. The rapid halt of \textit{amo} (and subsequently \textit{nxr}) around day 20 occurs when ammonia concentration falls below the threshold imposed by the maintenance energy requirements of the cells (Eq. 8.4.6). See Methods for details. Data from de Boer and Laanbroek (94).
Figure 8.3: **Reconstructing a bioreactor’s state using chemical time series.** (A) Reaction-centric illustration of a flow-through nitrifying bioreactor, corresponding to experiments by Dumont et al. (109). Continuous loop-arrows represent self-amplifications of ammonium oxidation (amo) and nitrite oxidation (nxr). (B) Methodological overview for model-based inference of the bioreactor’s state using chemical time series, as performed in this chapter. Reaction rates are inferred from the derivative of metabolite concentration time series. These reaction rates, in turn, are used in the reaction-centric model to predict the growth of the corresponding reaction capacities and hence the trajectory of the system’s state. A comparison of predictions with the original chemical concentration data can be used to calibrate and validate the model. Right panel: (C) Input NH$_4^+$ concentration, (D) dilution rate, (E) inferred amo rate, (F) inferred nxr rate, (G) estimated amo capacity and (H) estimated nxr capacity over time.
Figure 8.4: **Information needed to estimate the state of a reaction-centric model.** (A) Illustration of a nitrifying microbial community: AOB oxidize ammonium (NH$_4^+$) to nitrite (NO$_2^-$), which is subsequently oxidized by NOB to nitrate (NO$_3^-$). (B) In a cell-centric framework, both abiotic (e.g., physicochemical) and biotic (e.g., cell density) measurements are required for a complete description of the system’s state at any particular moment in time (“snapshot”). (C) In a reaction-centric framework, the system’s state can be reproduced based on purely abiotic measurements, however measurements across multiple time points are needed (“time profile”).
Figure 8.5: **Model predictions and data for Example 2.** (A) Ammonium, (B) nitrite, (C) nitrate and (D) dry biomass concentration in the flow-through nitrifying bioreactor, as predicted by the data-driven model (thick blue curve) and compared to experimental data (dots). The thin grey curves show smoothened, i.e., noise-reduced, approximations of the data (see Methods for details). The shaded regions in (a–c) mark the data that were used for model calibration. Data in the white region were ignored during calibration and serve as an independent validation of the model. The arrow in (C) indicates the delayed onset of $n_{rx}$ after the temperature of the bioreactor was reduced from $30^\circ C$ to $25^\circ C$ on day 181. The unknown biomass yield factor, required for comparing the reaction-centric model to biomass measurements in (D), was calibrated using least-squares fitting (see the main text for details). Data by Dumont et al. (109, Bioreactor B).
Chapter 9

Closing chapter

9.1 Conclusions

The work presented here provides strong support for a pathway-centric paradigm of microbial ecology, in which stoichiometric and energetic constraints (“boundary conditions”), such as the availability of light and the presence of certain electron acceptors for respiration, dictate the structure of the metabolic network in a community, while exerting much less control on the taxonomic composition within individual metabolic functional groups (Chapters 2 and 3). According to this paradigm, the metabolic functional structure of a microbial community and the taxonomic composition within individual functional groups constitute roughly independent and complementary facets of community structure, which are shaped by distinct assembly mechanisms. Although undoubtedly an idealization, this paradigm provides an elegant interpretation for several patterns of community assembly examined here. For example, the decoupling between taxonomy and function, manifested as a strong taxonomic turnover within functional groups, explains the previously reported apparent randomness in overall microbial community composition across bromeliads (Chapter 3; 117) or in bioreactors during stable performance (Chapters 5 and 6; 122).

A pathway-centric paradigm greatly simplifies conceptual and mathematical modeling of ecosystem biochemistry, because taxonomic richness and turnover become system properties that have little relevance to metabolic network structure. Constructing an appropriate model for ecosystem biochemistry remains itself a hard problem, however the two approaches demonstrated in Chapters 7 and 8 yield encouraging results. Notably, the pathway-centric biogeochemical model presented in Chapter 7 was able to largely explain the geochemical, metagenomic, metatranscriptomic and metaproteomic depth profiles in the Saanich Inlet OMZ. This supports the idea that genes provide a robust description of an ecosystem’s biological component (104, 201), which, in conjunction with the geochemical background, determines pathway activity (257). That is not to say that metabolic pathways simply respond to an externally imposed and independent geochemical background, since microbial
metabolism itself can strongly affect environmental conditions and hence metabolic niche space. For example, most major transitions in Earth’s surface redox state were the combined result of coupled geological and microbial metabolic processes (54, 116). In the Saanich Inlet OMZ, the formation of a narrow sulfide-nitrate transition zone at intermediate depths during periods of stratification results from the gradual balancing between oxidant supply from the upper layers, reductant supply from the sediments, physical transport processes and microbially catalyzed electron flow (Fig. 9.1).

Taken together, the work presented in this dissertation strongly suggests that functional profiling of microbial communities — either via multi-omics or via a functional annotation of detected taxa (Chapter 2) — should constitute the baseline for microbial biogeography, especially across transects where geochemical gradients shape microbial niche space. The remaining taxonomic variation within functional groups can then be analyzed separately (e.g., Chapter 3) to elucidate additional community assembly processes. For example, the statistical analyses and mathematical models presented in Chapters 3, 5 and 6 suggest that biotic interactions, such as competition and predation by phages, play a major role in driving the taxonomic variation within functional groups observed even under constant environmental conditions. In contrast, neither across the global ocean (Chapter 2) nor between replicate bromeliad tanks (Chapter 3) did spatial distances have any significant effect on community differences. This suggests that in these cases dispersal limitation may be negligible compared to other microbial community assembly mechanisms. Similarly, non-random phylogenetic associations and OTU co-occurrence patterns across bromeliad samples (Chapter 3) suggest that random demographic drift and random colonization events are unlikely the cause of the strong taxonomic variation observed within functional groups.

The complete decoupling between community metabolism and community assembly processes other than metabolic niche effects, as implied by a pathway-centric paradigm, is of course only an idealization, whose accuracy will depend on the particular system at hand. As shown in Chapter 6, biotic interactions such as predation by lytic phages can significantly disturb the metabolic throughput of a bioreactor even under constant operating conditions, because fluctuations in individual populations (Fig. 9.1A) inevitably lead to fluctuations in community function. In such cases, a high functional redundancy can stabilize overall community function over time despite strong taxonomic turnover within individual functional groups. This situation is analogous to convective heat transport through a fluid trapped between a hot and a cold plate (aka. Rayleigh-Bénard convection; 532), where spontaneous plumes and vortices in the fluid can disturb overall heat transport rates; an increased plate surface area stabilizes overall heat transport despite small-scale turbulence (Fig. 9.2).
9.2 So, what is life?

This work provides strong support for the idea that the activity and distribution of prokaryotic metabolic pathways is dictated by environmental physicochemical conditions in a way that can become decoupled from the particular combination of organisms in a community. Given their frequent horizontal transfer across prokaryotic clades (116, 133, 270, 385, 483), these pathways — or more precisely, the genes encoding them — resemble self-serving living entities that proliferate when environmental conditions are favourable (93). Strong mutual interdependencies, for example between rRNA genes and genes involved in metabolism, in turn promote the association of genes into genomes encapsulated inside a single container (the prokaryotic cell). The conceptual similarity of this interpretation to obligate symbiotic “organelle” assemblages, structurally stabilized in the form of eukaryotic cells (195, 280), is...
Figure 9.2: **Complex community dynamics in a methanogenic bioreactor, compared to turbulent thermal convection.** (A) Simulated cell concentrations for 240 different OTUs (one color per OTU) over time, in a methanogenic glucose-fed bioreactor operated under constant conditions (Chapter 6). Phage-driven “killing-the-winner” dynamics lead to repeated complex fluctuations of cell populations and community metabolic performance. Increasing the functional redundancy stabilizes the overall metabolic throughput of the bioreactor (see Fig. 6.3 in Chapter 6), although fluctuations in individual cell populations remain at the same amplitudes. (B) Turbulent convective heat transport across a thermal gradient, through a fluid trapped between a hot and a cold plate and subject to gravity, also known as Rayleigh-Bénard convection (532). The flux of energy through this system involves the spontaneous and repeated formation of complex vortices, analogous to the complex community dynamics sustained by the supply of glucose in the bioreactor. Increasing the plate surface area stabilizes the overall heat transport between the plates, although plumes and vortices remain at the same spatiotemporal scales. Simulation performed using software by The Concord Consortium (473).

striking. In fact, primordial cells were likely modular assemblages of loosely interdependent components subject to frequent horizontal exchange (525).

A precise definition of life has long been an elusive intellectual endeavor (413, 482), although the ability to “self-reproduce” and the ability to evolve are arguably the two most commonly stated prerequisites (485). The truth of the matter is that the majority of organisms can only reproduce as part of a broader biological consortium that is characterized by distributed biosynthetic capacities and a metabolic network that is partitioned across multiple organisms (169, 200, 238, 311). Chemoheterotrophs, for example, by definition depend on other organisms for reduced carbon supply. Auxotrophy, i.e., the inability to produce specific required biomolecules such as cofactors, is widespread even amongst autotrophs (183, 408, 409). Syntrophy (333), in which the metabolic activities of two organisms are mutually dependent, is yet another manifestation of the fact that “life needs life”. For example,
hydrogen-consuming methanogens enable bacterial hydrogenogenic fermentation by lowering partial hydrogen pressure to levels at which fermentation becomes exergonic (27, 315, 453). In the Saanich Inlet water column, SUP05 chemolithoautotrophic Gammaproteobacteria oxidize hydrogen sulfide produced by sulfate reducers and reduce nitrate produced by nitrifiers, in turn supplying nitrous oxide to nitrous oxide reducers and nitrite to nitrite-respiring anammox bacteria (Fig. 7.1, Chapter 7). Viruses, the most abundant “lifeform” on Earth (462), rely on the translational machinery of their hosts for replication. In fact, bacteriophages occasionally reprogram the metabolic machinery of their hosts to increase viral genome replication (199). Prions, i.e., proteinaceous infectious particles that propagate by transmitting their misfolded state to other proteins, can even undergo Darwinian evolution (272). Plants rely on the photosynthetic capacity of their cyanobacterial endosymbionts and, reciprocally, these endosymbionts benefit from a steady source of nutrients and protection (313). The fluent conceptual transition between simple evolving molecules (proteins, genes) that depend on each other for replication, up to obligately symbiotic cell assemblages (170), underlies much of the difficulty of agreeing on a single definition for “life” (485). In fact, a precise separation of the living from the abiotic based on a threshold of “self-sufficiency” for replication appears as meaningful as the definition of a bacterial “species” based on a threshold of genomic similarity (422, 482). In any case, it seems that a coerced definition of “life” may well be irrelevant to an understanding of the integrated biotic-abiotic Earth System.
A large part of the work presented here involved the development of novel computational tools. Notably, examining the taxonomic composition within individual metabolic functional groups in Chapters 2 and 3 required the manual construction of a database for the functional annotation of prokaryotic taxa (“FAPROTAX”), based on extensive literature search. Currently, FAPROTAX includes over 7400 annotations (e.g., at the species or genus level) covering more than 80 functional groups (e.g., anoxygenic photosynthesis, anammox or sulfate respiration). FAPROTAX thus constitutes a cheap alternative to metagenomics, the current de facto standard for functional community profiling (97). In fact, because FAPROTAX is based on experimental evidence for metabolic phenotypes, it can resolve ambiguities in the interpretation of community gene content inherent to metagenomics (376). Further, since FAPROTAX provides information on taxon-function associations within a community, it enables the separate analysis of community assembly processes within individual functional groups (Chapter 3). I thus anticipate that FAPROTAX (and its potential successors) will push microbial ecology into exciting new directions. FAPROTAX will be freely available upon publication of Chapter 2 at: http://www.zoology.ubc.ca/louca/FAPROTAX

The work presented in Chapters 5, 6 and 8 required the construction of high-dimensional dynamical models for microbial communities and their calibration to experimental data. A framework tailored to this task was unavailable when work on this dissertation began and hence MCM, a software for modeling microbial communities, was developed (Chapter 4). Compared to recently published comparable tools (177), MCM remains unique in its ability to automatically calibrate arbitrary model parameters using a plethora of experimental data, such as cell counts and chemical time series. As demonstrated in Chapter 4 using data from previous E. coli evolution experiments, MCM can yield insight into the metabolic activity of individual populations in a community context and provide mechanistic explanations for previously observed eco-evolutionary dynamics. MCM, including a thorough user manual and multiple examples, is freely available at: http://www.zoology.ubc.ca/MCM
Bibliography


phenetic trees based on metabolic capabilites across the three domains of life. Journal
of Molecular Biology 340:491–512.

anoxic sediments of Saanich Inlet, British Columbia, Canada: a case for highly evolved

metabolism in a thermophilic, anaerobic, butyrate-degrading triculture. Applied and
Environmental Microbiology 53:434–439.

uptake. Marine Ecology Progress Series 440:41–51.

Pages 1–13 in Proceedings of the second annual conference on commercial aquaculture,
CES, volume 240.

nitrate and nitrite reduction in denitrification by Pseudomonas fluorescens. Biotech-
nology and Bioengineering 46:476–484.


in batch cultures limited by respiration? Journal of Bacteriology 144:114–123.

niper and R. Brinkhurst, editors. Proceedings of a multidisciplinary symposium on
Saanich Inlet, 2nd February, 1983. Number 83 in Canadian technical report of hydrography and ocean sciences, Department of Fisheries and Oceans Canada, Sidney, BC.


Bibliography


[192] Holtappels, M., G. Lavik, M. M. Jensen, and M. M. Kuypers, 2011. 15N-labeling experiments to dissect the contributions of heterotrophic denitrification and anammox to nitrogen removal in the OMZ waters of the ocean. Chapter 10, pages 223–251 in


Bibliography


Bibliography


Appendix A

Chapter 2: Supplemental material

A.1 Methods

A.1.1 Sequencing data

Bacterial and archaeal taxa were identified based on 16S ribosomal DNA sequences, extracted from a previously published unassembled metagenomic data set that comprises 139 marine samples from 68 locations (460; see Table A.2 for an overview and Fig. A.9 for sampling locations). Quality-filtered 16S rDNA sequences, extracted from the metagenomes by the original authors (460), were clustered with uclust (110) by closed-reference mapping to the SILVA 119 reference database (378) at 99% similarity, yielding 49685 clusters (so-called operational taxonomic units, or OTUs) representing 10976383 sequences. OTU abundances were converted to relative proportions in each sample by total sum scaling. Metagenomic KEGG orthologous group (221) profiles for the same samples were taken directly from the original publication (460) and were normalized using total sum scaling. Whenever possible, KEGG orthologs associated with similar enzymatic functions (e.g., dissimilatory nitrite reduction to ammonium, nirBD and nrfAH) were combined into functional groups comparable to the ones considered in this chapter (Fig. A.5). A list of KEGG orthologs associated with each function is provided in Table F.1.

A.1.2 Functional annotation of prokaryotic taxa

To estimate the functional potential of the bacterial and archaeal communities, we classified each OTU into one or more functional groups based on current literature, such as announcements of cultured representatives or manuals of systematic microbiology, whenever possible. More precisely, a taxon (e.g., species or genus) was associated with a particular function (e.g., denitrification) if all cultured species within the taxon have been shown to exhibit that function. While the risk of false generalizations was minimized via extensive manual investigation of available literature, we point out that as more organisms are cul-
tured in the future some of these generalizations may turn out to be false. Furthermore, strain-specific variations within species (514) were ignored in favor of type strains, which may have led to further inaccuracies in our functional annotations. Nevertheless, a comparison of the so-obtained functional community profiles with metagenomic profiles shows that this approach is able to reveal most major differences in functional potential between samples (Fig. A.5). A total of 30633 OTUs (66.1 %) were assigned to at least one functional group, yielding a total of 46313 functional annotations (see Table A.4 for an overview). OTUs without any functional annotation were excluded from the analysis. Our complete database for the functional annotation of prokaryotic taxa (FAPROTAX) is available online at www.zoology.ubc.ca/louca/FAPROTAX.

Some taxa were associated with multiple functions: For example *Sulfurospirillum arsenophilum* was associated with, among others, nitrate respiration and fermentation. Hence, the detection of *S. arsenophilum* was interpreted as the detection of a putative nitrate respirer and a putative fermenter. Furthermore, functional groups could be nested: For example, all nitrate denitrifiers were also associated with nitrate respiration as well as nitrate reduction. Relative functional group abundances in each sample were calculated as the cumulative abundances of OTUs assigned to each functional group, after normalizing by the cumulative abundances of OTUs associated with at least one function (i.e., total-sum-scaling restricted to functionally annotated OTUs). For example, if in a particular sample all functionally annotated OTUs together accounted for \(10^4\) reads, and a single functionally annotated OTU accounted for 10 sequences, then that OTU contributed a value of \(10/10^4 = 0.001\) to each functional group it was associated with.

### A.1.3 Statistical Analysis

**Relating community composition to environmental variables**

To assess the overall effect of environmental conditions on community structure, we performed multivariate regression of relative functional group abundances, as well as the proportions of various taxa within each functional group, against 13 key abiotic oceanographic variables that included dissolved oxygen, salinity, temperature and depth (see Table A.1 for an overview; see below for details). Regression was done using non-linear kernel ridge regression (KRR) with Gaussian radial basis function kernels, as implemented by the scikit-learn machine learning package (363). In short, KRR models combine non-linear regression with regularization of coefficients. Regularization limits the complexity of a model and reduces the risk of overfitting by penalizing excessive coefficients, and generally presents a more robust
alternative to step-wise model selection methods (88). The final complexity of a fitted model depends on a penalization parameter that mediates the trade-off between model simplicity and a better fit to the data. Non-linearities, on the other hand, are accounted for using the “kernel-trick”, which embeds predictor variables into higher dimensions using Gauss-shaped functions prior to linear regression (363). The predictive power of each model (i.e., for each functional group or taxon) was measured using 10-fold Monte-Carlo cross-validation with 500 random iterations, which estimates the achievable coefficient of determination ($R^2_{CV}$) when only a random subset (90%) of the samples are used for fitting and the remaining samples (10%) are used for independent testing (421). The $R^2_{CV}$ is typically used to assess the risk of data overfitting and inaccurate extrapolation, and provides a more conservative estimate of a model’s predictive power than the classical coefficient of determination ($R^2$). The degree of penalization as well as the Gaussian kernel bandwidth were optimized separately for each model using grid-search by maximizing the achievable $R^2_{CV}$. Hence, the final models minimize complexity and the risk of overfitting while optimizing expected predictive power.

We mention that we had initially considered multivariate linear regression instead of KRR models. However, not surprisingly, the linear models resulted in very unreliable extrapolations and revealed strong non-linearities in the data not accounted for by the linear model structure. Nevertheless, the overall results and conclusions reported in this chapter remained unchanged.

To assess the influence of individual environmental variables on community structure, we performed Pearson and Spearman rank correlation analysis between the relative abundances of functional groups as well as the taxon proportions within individual functional groups on the one hand, and environmental variables on the other hand. Both Pearson and Spearman rank correlations yielded similar conclusions, hence we focus our reports on the latter.

To verify the robustness of the regression and correlation analyses, we considered the taxonomic composition within individual functional groups at various taxonomic resolutions (species, genus, family, order and class level); all taxonomic resolutions yielded similar results (e.g., Figs. A.1 and A.2). Functional groups or taxa represented in less than 10 samples were excluded from regression and correlation analysis. Because each functional group could contain hundreds or thousands of different taxa, the achieved $R^2_{cv}$ and absolute correlations for the taxon proportions in the functional group were averaged across all taxa.
Environmental variables

Environmental variables considered for regression and correlation analysis were standard abiotic oceanographic variables that are generally known to influence marine microbial distributions (235). These variables included in-situ temperature, salinity, dissolved oxygen concentration, apparent oxygen utilization, nitrate ($\text{NO}_3^-$), phosphate ($\text{PO}_4^{3-}$), dissolved silicate, sample depth, distance to the thermocline, surface total $\text{CO}_2$, surface pH, daily insolation and duration of day. 15 out of 139 samples, for which some of these metadata were unavailable (e.g., due to limited depth coverage), were excluded from the regression and correlation analysis but still included in all other investigations. Any environmental metadata provided as part of the original data set (460) were used without modification; missing data were obtained from public global gridded databases (140, 141, 278, 342, 464, 549) and linearly interpolated between grid points if needed. No extrapolation outside of the available grids was done. An overview of environmental variables is given in Table A.1. Spearman rank correlations between environmental variables are shown in Fig. A.10.

Correlations between functional groups

To detect putative interactions between different functional groups, we calculated Spearman rank correlations between the relative abundances of functional groups across all 139 samples (Fig. A.6). To ensure that any positive correlations between two given functional groups were not merely due to their overlap (in terms of shared OTUs; Fig. A.11), we only considered the fraction of each functional group due to OTUs not shared with the other functional group. For example, when comparing the abundances of cellulolytic and xylanolytic cells, we only in fact compared the abundances of non-xylanolytic cellulolytic cells and non-cellulolytic xylanolytic cells. For correlations between the entire functional groups, i.e., not accounting for overlaps, see Fig. A.12.

Functional and taxonomic richness

To compare the taxonomic and functional richness of the microbial communities we first rarified all samples at equal sequencing depth, at the maximum depth possible (24644 sequences, picked without replacement), to eliminate richness differences purely based on varying sampling effort. Upon rarefaction, the total number of detected OTUs or functional groups was taken as a measure of richness. Rarefaction was repeated 1000 times, and the OTU richness as well as functional richness was averaged over all rarefactions.
Assessing segregation between water column zones

To assess the extent to which OTUs, higher taxa or functional groups exhibit significantly different abundances between environments, we compared their mean abundances between water column zones. Specifically, for each OTU (or taxon or functional group) and for any two water column zones, we used the Welch test statistic (513) to compare the mean relative abundances within each zone to a null model corresponding to random sample permutations. The size of each sample group (i.e., corresponding to either zone) is maintained by the permutations. In contrast to other similar tests (e.g., Student’s t-test), this null model does not assume any particular probability distribution of the data. The statistical significance of a difference in mean abundances was defined as the probability that the Welch statistic would be more extreme (in either direction) than observed, and was estimated using 1000 random permutations. Here we report the fraction of OTUs (or taxa or functional groups) that were identified to exhibit significantly different abundances between zones (Figs. 2.2E–G). Note that under the null model (i.e., no segregation) the false detection rate would be 5%. Hence, the high fraction of functional groups identified as significantly segregated between zones (e.g., up to 80% of functional groups when comparing the mesopelagic and surface zones, Fig. 2.2G) is indicative of strong differences in metabolic niche structure between zones.

Geographical variation in community structure

To assess the extent to which dispersal limitation may promote differences in community composition we performed Pearson and Spearman rank Mantel correlation tests between the geographical distances of sample pairs and their dissimilarities in terms of taxonomic as well as functional community composition. Geographical distances were geodesic distances calculated based on sample latitude and longitude. Considered dissimilarity metrics between community profiles (taxonomic or functional) were Bray-Curtis dissimilarity, Canberra distance and Hellinger distance. These dissimilarity metrics are widely used in biogeographical surveys (266), and we considered all three of them to verify the robustness of our results. When calculating dissimilarities in taxonomic composition, multiple taxonomic levels were considered (species, genus, family, order, class); all yielded similar results. To control for strong variations across depth, presumably driven by environmental conditions rather than dispersal limitation (see discussion in the main text), we restricted our analysis to samples within individual zones (i.e., mesopelagic, surface layer, mixed layer and deep chlorophyll maximum).

The statistical significance of correlations between geographical distance and community dis-
similarity was estimated using repeated random permutations of the rows and columns in the distance matrices (1000 trials, rows and columns were permuted similarly). All dissimilarity metrics, and both Pearson and Spearman rank correlations, yielded no statistically significant correlations between geographical distance and community dissimilarity. The comparisons between geographical distance and functional as well as taxonomic (OTU) dissimilarities are shown in Figs. 2.4AB for the mesopelagic zone, and Fig. A.8 for the surface layer and deep chlorophyll maximum. For other taxonomic resolutions (e.g., at genus and family level) see Fig. A.13. For comparisons between geographical distances and taxonomic differences within individual functional groups, see Fig. A.4.

**OTU co-correlations within functional groups**

To test whether OTUs with a higher number of shared functions correlated more strongly, we calculated the mean Spearman rank correlation between OTUs with 0 shared functions (90587017 OTU pairs), as well as between OTUs with exactly 1 shared function (34415335 OTU pairs), between OTUs with exactly 2 shared functions (410333 OTU pairs) and so on (Fig. 2.2H). Only OTUs occurring in at least 10 samples were considered (15840 OTUs in total). Mean correlations obtained over less than 10 OTU pairs were omitted.

**Metric multidimensional scaling**

Pairwise dissimilarities between functional or taxonomic community profiles were calculated using three metrics — Bray-Curtis, Canberra and Hellinger (266). Each resulting dissimilarity matrix was used to visualize sample differences via metric multidimensional scaling with stress majorization (42). In this procedure, sample points are embedded into two-dimensions such that the pairwise Euclidean sample distances in the embedding “best match” the original sample dissimilarities. Hence, points appearing closer to each other in the embedding correspond to samples with more similar microbial communities. Fitting was done by minimizing the Kruskal stress function using the Scikit-lean package (363). All dissimilarity metrics yielded similar results.

**A.2 Resolving ambiguities in gene-centric metagenomics**

In this work we have shown that a classification of community members into functional groups can reveal ecologically meaningful differences between microbial communities across environments. The translation of taxonomic information into community functional potential thus
Chapter 2: Supplemental material provides a powerful alternative to environmental shotgun sequencing (486). Gene-centric metagenomic profiles, in particular, suffer from the conceptual limitation that community gene content generally doesn’t directly translate to community functional potential. This is because the same or highly similar genes can be involved in several pathways and because the functionality of individual genes typically depends on their genomic context—which remains unknown in shotgun metagenomics (376). These limitations are inherent to metagenomic profiling, and apply equally to recent algorithms that estimate community gene content by projecting detected marker genes to closely related sequenced genomes (255).

In contrast, functional profiles based on experimental phenotypic characterizations, as constructed here, can resolve ambiguities in the interpretation of community gene-content (376). For example, variants of the dissimilatory sulfite reductase (dsrAB) genes can be involved in respiratory sulfur reduction, energy-yielding chemolithotrophic sulfur oxidation or electron-yielding sulfur oxidation for anoxygenic photosynthesis, depending on the host microorganism (336). In fact, metagenomic dsrAB abundances in our samples do not significantly differ between water column zones. On the other hand, the abundance of identified sulfide-oxidizing organisms is greater than the abundance of sulfate respirers and peaks in the mesopelagic (Fig. A.5), where sulfide may be used to support chemolithoautotrophic growth, notably in oxygen minimum zones (415, 489).

Similarly, gene sequences coding for ammonia monooxygenase (amo) and the homolog particulate methane monooxygenase (pmo) are generally indistinguishable by many current gene annotation databases (e.g., KEGG orthologs K10944, K10945, K10946 (221)). In fact, both genes code for enzymes that can oxidize ammonia and methane, even though ammonia oxidation (by nitrifiers) and methane oxidation (by methanotrophs) constitute trophic strategies for separate microbial groups (173, 190). Consequently, the peak of amo/pmo-related metagenomic sequences observed in the mesopelagic zone (Fig. A.5B) cannot a priori be unambiguously attributed to nitrifiers or methanotrophs. In contrast, our phenotype-based functional profiles strongly suggest that the over-representation of these genes in the mesopelagic is due to methanotrophs (Fig. A.5A).

A.3 Comparison with Sunagawa et al. (2015)

Our analysis revealed a clear positive correlation of community richness with depth, both in terms of detected functional groups as well as taxa, with the mesopelagic zone exhibiting particularly high richness (Figs. 2.2D). These patterns are consistent with an increased
genetic as well as taxonomic richness at depth, as reported by Sunagawa et al. (460). Furthermore, our analysis revealed strong correlations of functional community profiles to depth and depth-correlated environmental variables such as temperature, nitrate and phosphate, but much weaker correlations to salinity (Fig. 2.1B). Previous metagenomic analysis of a subset of the same samples by Sunagawa et al. (460) also showed a strong correlation of depth and temperature, and an insignificant correlation of salinity, with functional profiles. In apparent contrast to our results, that study found only weak effects of nitrate and phosphate on metagenomic composition. However, Sunagawa et al. only considered surface layer samples for calculating correlations with nutrient concentrations, whereas our analysis also includes samples from the mesopelagic zone where light becomes less relevant and nitrate becomes an important terminal electron acceptor for anaerobic growth. A restriction of our correlation analysis to surface samples revealed much weaker effects of nitrate and phosphate on functional profiles (Fig. A.14), consistent with Sunagawa et al.’s metagenomic analysis.
Table A.1: **Oceanographic variables.** Overview of oceanographic variables and sources.

<table>
<thead>
<tr>
<th>variable</th>
<th>depth-specific</th>
<th>units</th>
<th>source</th>
</tr>
</thead>
<tbody>
<tr>
<td>oxygen (dissolved)</td>
<td>yes</td>
<td>mL/L</td>
<td>a Sunagawa et al. (460) if available, otherwise monthly average by WOA 2013 V2 (140).</td>
</tr>
<tr>
<td>apparent oxygen utilization</td>
<td>yes</td>
<td>mL/L</td>
<td>b monthly average by WOA 2013 V2 (140).</td>
</tr>
<tr>
<td>salinity</td>
<td>yes</td>
<td>PSU</td>
<td>c Sunagawa et al. (460) if available, otherwise monthly average by WOA 2013 V2 (549).</td>
</tr>
<tr>
<td>temperature (in-situ)</td>
<td>yes</td>
<td>°C</td>
<td>d Sunagawa et al. (460) if available, otherwise monthly average by WOA 2013 V2 (278).</td>
</tr>
<tr>
<td>nitrate ($\text{NO}_3^-$)</td>
<td>yes</td>
<td>$\mu$M</td>
<td>e Sunagawa et al. (460) if available, otherwise monthly average by WOA 2013 V2 (141).</td>
</tr>
<tr>
<td>phosphate ($\text{PO}_4^{3-}$)</td>
<td>yes</td>
<td>$\mu$M</td>
<td>f Sunagawa et al. (460) if available, otherwise monthly average by WOA 2013 V2 (141).</td>
</tr>
<tr>
<td>silicate (dissolved)</td>
<td>yes</td>
<td>$\mu$M</td>
<td>g Sunagawa et al. (460) if available, otherwise monthly average by WOA 2013 V2 (141).</td>
</tr>
<tr>
<td>depth</td>
<td>yes</td>
<td>m</td>
<td>h monthly average by Sunagawa et al. (460).</td>
</tr>
<tr>
<td>total inorganic carbon (total CO$_2$)</td>
<td>no (surface)</td>
<td>$\mu$mol/kg</td>
<td>i monthly average by Takahashi et al. (464).</td>
</tr>
<tr>
<td>pH</td>
<td>no (surface)</td>
<td>-</td>
<td>j monthly average by Takahashi et al. (464).</td>
</tr>
<tr>
<td>insolation</td>
<td>no (surface)</td>
<td>kWh $\cdot$ m$^{-2} \cdot$ d$^{-1}$</td>
<td>k monthly average by NASA Earth Observations (342).</td>
</tr>
<tr>
<td>duration of day</td>
<td>no</td>
<td>hours</td>
<td>-</td>
</tr>
<tr>
<td>distance to thermocline</td>
<td>yes</td>
<td>m</td>
<td>l thermocline defined as depth where temperature drops 0.5°C below reference depth 10 m (95); using monthly average temperature profile from WOA 2013 V2 (278).</td>
</tr>
</tbody>
</table>
Figure A.1: **Correlation analysis at various taxonomic levels.** Mean absolute Spearman rank correlations between taxon proportions within each functional group and environmental variables. Taxa are collapsed at the (A) genus or (B) family level. Circle surface area and color darkness are proportional to the absolute correlation, averaged across all taxa within a functional group.
Figure A.2: **Environmental filtering at higher taxonomic levels.** Cross-validated coefficients of determination ($R^2_{cv}$) for taxon proportions within each functional group, achieved by regression models with environmental predictor variables. Taxa are collapsed at various levels (genus, family, order and class).
Figure A.3: **Functional vs taxonomic profiles.** (A) Functional and (B) taxonomic (class-level) community profiles, both based on 16S rDNA sequences. Samples are ordered according to water column zone (SRF: surface water layer; DCM: deep chlorophyll maximum; MIX: subsurface epipelagic mixed layer; MES: mesopelagic zone). A darker color corresponds to a greater relative abundance.
Figure A.4: **Dissimilarities in functional group composition vs geographical distances.** Bray-Curtis dissimilarities between microbial communities in the mesopelagic zone, compared to geographical distances (one point per sample pair). Dissimilarities are calculated in terms of OTU proportions within various functional groups (one plot per functional group).
Figure A.5: **Phenotype-based vs metagenomic functional profiles.** Functional community profiles calculated based on (A) functional annotations of bacterial and archaeal taxa and (B) metagenomic KEGG orthologous groups (221). A darker color corresponds to a higher relative abundance. Samples (columns) are clustered by ocean layer (SRF: surface water layer; DCM: deep chlorophyll maximum; MIX: subsurface epipelagic mixed layer; MES: mesopelagic zone), and functional groups (rows) are hierarchically clustered using UPGMA. Several, but not all, functional groups are comparable between (A) and (B).
Figure A.6: **Correlations between functional groups (corrected).** Spearman rank correlations between relative functional group abundances, after correcting for group overlaps (in terms of shared OTUs). Blue and red colors correspond to positive and negative correlations, respectively. White corresponds to zero or statistically non-significant correlations. Rows and columns are hierarchically clustered (UPGMA). For correlations not correcting for overlaps see Fig. A.12.
Figure A.7: **Taxonomic compositions within functional groups.** OTU proportions within various functional groups (one plot per functional group; one color per OTU within each plot). For each functional group, samples are sorted according to the relative abundance of the entire functional group, as indicated by the horizontal scale.
Figure A.8: **Community dissimilarities vs geographical distances.** Bray-Curtis dissimilarities between microbial communities compared to geographical distances (one point per sample pair). Samples are restricted to the surface layer (top row) and the deep chlorophyll maximum (bottom row). Community dissimilarities are calculated in terms of relative functional group abundances (left column) and relative OTU abundances (right column).
Figure A.9: **Sampling locations.** Most locations include samples at multiple depths. Data from Sunagawa et al. (460).
Figure A.10: **Correlations between environmental variables.** Blue and red colors correspond to positive and negative correlations, respectively. White corresponds to zero or statistically non-significant correlations. Rows and columns are hierarchically clustered (UPGMA).
Figure A.11: **Functional group overlaps.** Overlaps between functional groups in terms of shared OTUs (Jaccard similarity index). A darker color corresponds to a greater overlap. An overlap of 1.0 corresponds to identical groups.
Figure A.12: **Function correlations (uncorrected)**. Spearman rank correlations between relative functional group abundances, not accounting for functional group overlaps. Blue and red colors correspond to positive and negative correlations, respectively. White corresponds to zero or statistically non-significant correlations. Rows and columns are hierarchically clustered (UPGMA). For correlations accounting for functional group overlaps see Fig. A.6.
Figure A.13: **Community dissimilarities vs geographical distances at higher taxonomic levels.** Bray-Curtis dissimilarities between microbial communities compared to geographical distances (one point per sample pair), for samples in the mesopelagic zone. Community dissimilarities are calculated in terms of relative (A) genus, (B) family, (C) order and (D) class abundances.
Figure A.14: **Functional vs taxonomic in the surface layer.** Spearman rank correlations between environmental variables and (A) relative functional group abundances or (B) OTU proportions within individual functional groups, restricted to 63 surface layer samples. Circle surface area and color darkness are proportional to absolute correlations.
Table A.2: **Overview of considered Tara oceans samples.** Samples used in this study, obtained from Sunagawa et al. (460). The last column indicates whether a sample was used in the regression and correlation analyses, depending on the availability of metadata.

<table>
<thead>
<tr>
<th>#</th>
<th>PANGAEA accession</th>
<th>INSDC accessions</th>
<th>regr. &amp; corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TARA_A100000164</td>
<td>ERS488330</td>
<td>yes</td>
</tr>
<tr>
<td>2</td>
<td>TARA_A100001011</td>
<td>ERS478040</td>
<td>yes</td>
</tr>
<tr>
<td>3</td>
<td>TARA_A100001015</td>
<td>ERS478017</td>
<td>yes</td>
</tr>
<tr>
<td>4</td>
<td>TARA_A100001035</td>
<td>ERS488569</td>
<td>yes</td>
</tr>
<tr>
<td>5</td>
<td>TARA_A100001037</td>
<td>ERS488599</td>
<td>yes</td>
</tr>
<tr>
<td>6</td>
<td>TARA_A100001234</td>
<td>ERS488621</td>
<td>yes</td>
</tr>
<tr>
<td>7</td>
<td>TARA_A100001388</td>
<td>ERS488545</td>
<td>yes</td>
</tr>
<tr>
<td>8</td>
<td>TARA_A200000113</td>
<td>ERS477931</td>
<td>yes</td>
</tr>
<tr>
<td>9</td>
<td>TARA_A200000159</td>
<td>ERS477953</td>
<td>yes</td>
</tr>
<tr>
<td>10</td>
<td>TARA_B000000437</td>
<td>ERS490029</td>
<td>yes</td>
</tr>
<tr>
<td>11</td>
<td>TARA_B000000441</td>
<td>ERS490085</td>
<td>yes</td>
</tr>
<tr>
<td>12</td>
<td>TARA_B000000460</td>
<td>ERS490065</td>
<td>no</td>
</tr>
<tr>
<td>13</td>
<td>TARA_B000000475</td>
<td>ERS490124</td>
<td>yes</td>
</tr>
<tr>
<td>14</td>
<td>TARA_B000000477</td>
<td>ERS490163</td>
<td>yes</td>
</tr>
<tr>
<td>15</td>
<td>TARA_B000000532</td>
<td>ERS489877</td>
<td>yes</td>
</tr>
<tr>
<td>16</td>
<td>TARA_B000000557</td>
<td>ERS489846</td>
<td>yes</td>
</tr>
<tr>
<td>17</td>
<td>TARA_B000000565</td>
<td>ERS489733</td>
<td>yes</td>
</tr>
<tr>
<td>18</td>
<td>TARA_B000000609</td>
<td>ERS489712</td>
<td>yes</td>
</tr>
<tr>
<td>19</td>
<td>TARA_B100000003</td>
<td>ERS488649</td>
<td>yes</td>
</tr>
<tr>
<td>20</td>
<td>TARA_B100000029</td>
<td>ERS488685</td>
<td>yes</td>
</tr>
<tr>
<td>21</td>
<td>TARA_B100000035</td>
<td>ERS488747</td>
<td>yes</td>
</tr>
<tr>
<td>22</td>
<td>TARA_B100000073</td>
<td>ERS488830</td>
<td>yes</td>
</tr>
<tr>
<td>23</td>
<td>TARA_B100000085</td>
<td>ERS488916</td>
<td>yes</td>
</tr>
<tr>
<td>24</td>
<td>TARA_B100000123</td>
<td>ERS489087</td>
<td>yes</td>
</tr>
<tr>
<td>25</td>
<td>TARA_B100000131</td>
<td>ERS489134</td>
<td>yes</td>
</tr>
<tr>
<td>26</td>
<td>TARA_B100000161</td>
<td>ERS489236</td>
<td>yes</td>
</tr>
<tr>
<td>27</td>
<td>TARA_B100000212</td>
<td>ERS489529</td>
<td>yes</td>
</tr>
<tr>
<td>28</td>
<td>TARA_B100000214</td>
<td>ERS489585</td>
<td>yes</td>
</tr>
<tr>
<td>29</td>
<td>TARA_B100000242</td>
<td>ERS489315</td>
<td>yes</td>
</tr>
<tr>
<td>30</td>
<td>TARA_B100000282</td>
<td>ERS489043</td>
<td>yes</td>
</tr>
<tr>
<td>31</td>
<td>TARA_B100000287</td>
<td>ERS489074</td>
<td>yes</td>
</tr>
</tbody>
</table>
Table A.2: Continued from previous page.

<table>
<thead>
<tr>
<th>#</th>
<th>PANGAEA accession</th>
<th>INSDC accession(s)</th>
<th>regr. &amp; corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>TARA_B100000315</td>
<td>ERS488769</td>
<td>yes</td>
</tr>
<tr>
<td>33</td>
<td>TARA_B100000378</td>
<td>ERS489727</td>
<td>no</td>
</tr>
<tr>
<td>34</td>
<td>TARA_B100000401</td>
<td>ERS489917</td>
<td>yes</td>
</tr>
<tr>
<td>35</td>
<td>TARA_B100000405</td>
<td>ERS490002</td>
<td>yes</td>
</tr>
<tr>
<td>36</td>
<td>TARA_B100000408</td>
<td>ERS489987</td>
<td>no</td>
</tr>
<tr>
<td>37</td>
<td>TARA_B100000424</td>
<td>ERS490433</td>
<td>yes</td>
</tr>
<tr>
<td>38</td>
<td>TARA_B100000427</td>
<td>ERS490476</td>
<td>yes</td>
</tr>
<tr>
<td>39</td>
<td>TARA_B100000446</td>
<td>ERS490373</td>
<td>no</td>
</tr>
<tr>
<td>40</td>
<td>TARA_B100000459</td>
<td>ERS490327</td>
<td>yes</td>
</tr>
<tr>
<td>41</td>
<td>TARA_B100000470</td>
<td>ERS490230</td>
<td>no</td>
</tr>
<tr>
<td>42</td>
<td>TARA_B100000475</td>
<td>ERS490265</td>
<td>yes</td>
</tr>
<tr>
<td>43</td>
<td>TARA_B100000482</td>
<td>ERS490296</td>
<td>yes</td>
</tr>
<tr>
<td>44</td>
<td>TARA_B100000497</td>
<td>ERS490183</td>
<td>yes</td>
</tr>
<tr>
<td>45</td>
<td>TARA_B100000508</td>
<td>ERS490507</td>
<td>no</td>
</tr>
<tr>
<td>46</td>
<td>TARA_B100000513</td>
<td>ERS490542</td>
<td>yes</td>
</tr>
<tr>
<td>47</td>
<td>TARA_B100000519</td>
<td>ERS490597</td>
<td>yes</td>
</tr>
<tr>
<td>48</td>
<td>TARA_B100000524</td>
<td>ERS490659</td>
<td>yes</td>
</tr>
<tr>
<td>49</td>
<td>TARA_B100000530</td>
<td>ERS490691</td>
<td>yes</td>
</tr>
<tr>
<td>50</td>
<td>TARA_B100000575</td>
<td>ERS492321</td>
<td>yes</td>
</tr>
<tr>
<td>51</td>
<td>TARA_B100000579</td>
<td>ERS492357</td>
<td>yes</td>
</tr>
<tr>
<td>52</td>
<td>TARA_B100000586</td>
<td>ERS492381</td>
<td>yes</td>
</tr>
<tr>
<td>53</td>
<td>TARA_B100000609</td>
<td>ERS493044</td>
<td>yes</td>
</tr>
<tr>
<td>54</td>
<td>TARA_B100000614</td>
<td>ERS493098</td>
<td>yes</td>
</tr>
<tr>
<td>55</td>
<td>TARA_B100000674</td>
<td>ERS492821, ERS492814</td>
<td>yes</td>
</tr>
<tr>
<td>56</td>
<td>TARA_B100000676</td>
<td>ERS492863</td>
<td>yes</td>
</tr>
<tr>
<td>57</td>
<td>TARA_B100000678</td>
<td>ERS492680</td>
<td>yes</td>
</tr>
<tr>
<td>58</td>
<td>TARA_B100000683</td>
<td>ERS492733</td>
<td>yes</td>
</tr>
<tr>
<td>59</td>
<td>TARA_B100000686</td>
<td>ERS492778</td>
<td>yes</td>
</tr>
<tr>
<td>60</td>
<td>TARA_B100000700</td>
<td>ERS492699</td>
<td>yes</td>
</tr>
<tr>
<td>61</td>
<td>TARA_B100000745</td>
<td>ERS490714</td>
<td>no</td>
</tr>
<tr>
<td>62</td>
<td>TARA_B100000749</td>
<td>ERS490633</td>
<td>no</td>
</tr>
<tr>
<td>63</td>
<td>TARA_B100000767</td>
<td>ERS490928</td>
<td>yes</td>
</tr>
<tr>
<td>64</td>
<td>TARA_B100000768</td>
<td>ERS490885</td>
<td>yes</td>
</tr>
<tr>
<td>65</td>
<td>TARA_B100000780</td>
<td>ERS491001</td>
<td>yes</td>
</tr>
</tbody>
</table>
Table A.2: Continued from previous page.

<table>
<thead>
<tr>
<th>#</th>
<th>PANGAEA accession</th>
<th>INSDC accession(s)</th>
<th>regr. &amp; corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>66</td>
<td>TARA_B100000787</td>
<td>ERS491044</td>
<td>yes</td>
</tr>
<tr>
<td>67</td>
<td>TARA_B100000795</td>
<td>ERS491095</td>
<td>yes</td>
</tr>
<tr>
<td>68</td>
<td>TARA_B100000809</td>
<td>ERS491110</td>
<td>no</td>
</tr>
<tr>
<td>69</td>
<td>TARA_B100000886</td>
<td>ERS491804</td>
<td>yes</td>
</tr>
<tr>
<td>70</td>
<td>TARA_B100000900</td>
<td>ERS491938</td>
<td>yes</td>
</tr>
<tr>
<td>71</td>
<td>TARA_B100000902</td>
<td>ERS492012</td>
<td>yes</td>
</tr>
<tr>
<td>72</td>
<td>TARA_B100000925</td>
<td>ERS492145</td>
<td>yes</td>
</tr>
<tr>
<td>73</td>
<td>TARA_B100000927</td>
<td>ERS492177</td>
<td>yes</td>
</tr>
<tr>
<td>74</td>
<td>TARA_B100000929</td>
<td>ERS492205</td>
<td>yes</td>
</tr>
<tr>
<td>75</td>
<td>TARA_B100000941</td>
<td>ERS492408</td>
<td>yes</td>
</tr>
<tr>
<td>76</td>
<td>TARA_B100000945</td>
<td>ERS492445</td>
<td>yes</td>
</tr>
<tr>
<td>77</td>
<td>TARA_B100000949</td>
<td>ERS492471</td>
<td>no</td>
</tr>
<tr>
<td>78</td>
<td>TARA_B100000953</td>
<td>ERS491980</td>
<td>yes</td>
</tr>
<tr>
<td>79</td>
<td>TARA_B100000959</td>
<td>ERS491913</td>
<td>yes</td>
</tr>
<tr>
<td>80</td>
<td>TARA_B100000963</td>
<td>ERS491836</td>
<td>yes</td>
</tr>
<tr>
<td>81</td>
<td>TARA_B100000965</td>
<td>ERS491874</td>
<td>yes</td>
</tr>
<tr>
<td>82</td>
<td>TARA_B100000989</td>
<td>ERS491525</td>
<td>yes</td>
</tr>
<tr>
<td>83</td>
<td>TARA_B100001013</td>
<td>ERS491767</td>
<td>yes</td>
</tr>
<tr>
<td>84</td>
<td>TARA_B100001027</td>
<td>ERS491699</td>
<td>yes</td>
</tr>
<tr>
<td>85</td>
<td>TARA_B100001029</td>
<td>ERS491740</td>
<td>yes</td>
</tr>
<tr>
<td>86</td>
<td>TARA_B100001057</td>
<td>ERS491492</td>
<td>yes</td>
</tr>
<tr>
<td>87</td>
<td>TARA_B100001059</td>
<td>ERS491463</td>
<td>yes</td>
</tr>
<tr>
<td>88</td>
<td>TARA_B100001063</td>
<td>ERS491421</td>
<td>yes</td>
</tr>
<tr>
<td>89</td>
<td>TARA_B100001079</td>
<td>ERS492294</td>
<td>yes</td>
</tr>
<tr>
<td>90</td>
<td>TARA_B100001093</td>
<td>ERS493390</td>
<td>yes</td>
</tr>
<tr>
<td>91</td>
<td>TARA_B100001094</td>
<td>ERS493431</td>
<td>yes</td>
</tr>
<tr>
<td>92</td>
<td>TARA_B100001105</td>
<td>ERS493460</td>
<td>no</td>
</tr>
<tr>
<td>93</td>
<td>TARA_B100001109</td>
<td>ERS492228</td>
<td>yes</td>
</tr>
<tr>
<td>94</td>
<td>TARA_B100001113</td>
<td>ERS492264</td>
<td>yes</td>
</tr>
<tr>
<td>95</td>
<td>TARA_B100001115</td>
<td>ERS492642</td>
<td>yes</td>
</tr>
<tr>
<td>96</td>
<td>TARA_B100001121</td>
<td>ERS492888</td>
<td>yes</td>
</tr>
<tr>
<td>97</td>
<td>TARA_B100001123</td>
<td>ERS492926</td>
<td>yes</td>
</tr>
<tr>
<td>98</td>
<td>TARA_B100001142</td>
<td>ERS494170</td>
<td>yes</td>
</tr>
<tr>
<td>99</td>
<td>TARA_B100001146</td>
<td>ERS494208</td>
<td>yes</td>
</tr>
<tr>
<td>#</td>
<td>PANGAEA accession</td>
<td>INSDC accession(s)</td>
<td>regr. &amp; corr.</td>
</tr>
<tr>
<td>----</td>
<td>-------------------</td>
<td>--------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>100</td>
<td>TARA_B100001167</td>
<td>ERS494274</td>
<td>no</td>
</tr>
<tr>
<td>101</td>
<td>TARA_B100001173</td>
<td>ERS494579</td>
<td>yes</td>
</tr>
<tr>
<td>102</td>
<td>TARA_B100001175</td>
<td>ERS494628</td>
<td>yes</td>
</tr>
<tr>
<td>103</td>
<td>TARA_B100001179</td>
<td>ERS494616</td>
<td>no</td>
</tr>
<tr>
<td>104</td>
<td>TARA_B100001245</td>
<td>ERS493372</td>
<td>yes</td>
</tr>
<tr>
<td>105</td>
<td>TARA_B100001248</td>
<td>ERS493300</td>
<td>yes</td>
</tr>
<tr>
<td>106</td>
<td>TARA_B100001250</td>
<td>ERS493340</td>
<td>yes</td>
</tr>
<tr>
<td>107</td>
<td>TARA_B100001287</td>
<td>ERS493636</td>
<td>yes</td>
</tr>
<tr>
<td>108</td>
<td>TARA_B100001540</td>
<td>ERS494236</td>
<td>yes</td>
</tr>
<tr>
<td>109</td>
<td>TARA_B100001559</td>
<td>ERS494559</td>
<td>yes</td>
</tr>
<tr>
<td>110</td>
<td>TARA_B100001564</td>
<td>ERS494518</td>
<td>yes</td>
</tr>
<tr>
<td>111</td>
<td>TARA_B100001741</td>
<td>ERS494332</td>
<td>yes</td>
</tr>
<tr>
<td>112</td>
<td>TARA_B100001750</td>
<td>ERS494374</td>
<td>yes</td>
</tr>
<tr>
<td>113</td>
<td>TARA_B100001758</td>
<td>ERS494394</td>
<td>yes</td>
</tr>
<tr>
<td>114</td>
<td>TARA_B100001765</td>
<td>ERS494431</td>
<td>no</td>
</tr>
<tr>
<td>115</td>
<td>TARA_B100001769</td>
<td>ERS494445</td>
<td>yes</td>
</tr>
<tr>
<td>116</td>
<td>TARA_B100001778</td>
<td>ERS494488</td>
<td>yes</td>
</tr>
<tr>
<td>117</td>
<td>TARA_B100001939</td>
<td>ERS493914</td>
<td>yes</td>
</tr>
<tr>
<td>118</td>
<td>TARA_B100001964</td>
<td>ERS493670</td>
<td>yes</td>
</tr>
<tr>
<td>119</td>
<td>TARA_B100001971</td>
<td>ERS493705</td>
<td>yes</td>
</tr>
<tr>
<td>120</td>
<td>TARA_B100001989</td>
<td>ERS493752</td>
<td>yes</td>
</tr>
<tr>
<td>121</td>
<td>TARA_B100001996</td>
<td>ERS493788</td>
<td>yes</td>
</tr>
<tr>
<td>122</td>
<td>TARA_B100002003</td>
<td>ERS493822</td>
<td>yes</td>
</tr>
<tr>
<td>123</td>
<td>TARA_B100002019</td>
<td>ERS493877</td>
<td>yes</td>
</tr>
<tr>
<td>124</td>
<td>TARA_B100002049</td>
<td>ERS494006</td>
<td>no</td>
</tr>
<tr>
<td>125</td>
<td>TARA_B100002051</td>
<td>ERS493938</td>
<td>yes</td>
</tr>
<tr>
<td>126</td>
<td>TARA_B100002052</td>
<td>ERS493981</td>
<td>yes</td>
</tr>
<tr>
<td>127</td>
<td>TARA_E500000075</td>
<td>ERS477979</td>
<td>yes</td>
</tr>
<tr>
<td>128</td>
<td>TARA_E500000081</td>
<td>ERS477998</td>
<td>yes</td>
</tr>
<tr>
<td>129</td>
<td>TARA_E500000178</td>
<td>ERS488486</td>
<td>yes</td>
</tr>
<tr>
<td>130</td>
<td>TARA_E500000331</td>
<td>ERS488509</td>
<td>yes</td>
</tr>
<tr>
<td>131</td>
<td>TARA_S2000000501</td>
<td>ERS488346</td>
<td>yes</td>
</tr>
<tr>
<td>132</td>
<td>TARA_X000000368</td>
<td>ERS487936</td>
<td>yes</td>
</tr>
<tr>
<td>133</td>
<td>TARA_X000000950</td>
<td>ERS488119</td>
<td>yes</td>
</tr>
</tbody>
</table>
Table A.2: Continued from previous page.

<table>
<thead>
<tr>
<th>#</th>
<th>PANGAEA accession</th>
<th>INSDC accession(s)</th>
<th>regr. &amp; corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>134</td>
<td>TARA_X000001036</td>
<td>ERS488147</td>
<td>yes</td>
</tr>
<tr>
<td>135</td>
<td>TARA_Y100000022</td>
<td>ERS488714</td>
<td>yes</td>
</tr>
<tr>
<td>136</td>
<td>TARA_Y100000031</td>
<td>ERS488936</td>
<td>yes</td>
</tr>
<tr>
<td>137</td>
<td>TARA_Y100000287</td>
<td>ERS488799</td>
<td>yes</td>
</tr>
<tr>
<td>138</td>
<td>TARA_Y100000294</td>
<td>ERS488849</td>
<td>yes</td>
</tr>
<tr>
<td>139</td>
<td>TARA_Y200000002</td>
<td>ERS487899</td>
<td>yes</td>
</tr>
</tbody>
</table>

Table A.3: **KOG-function associations.** KEGG orthologous groups (KOG) associated with various functions in the metagenomic sequences, for comparison with our phenotype-based functional profiles. Whenever KOGs are associated with a single pathway step, the corresponding genes are indicated in brackets.

<table>
<thead>
<tr>
<th>function</th>
<th>KOGs</th>
</tr>
</thead>
<tbody>
<tr>
<td>oxygenic photosynthesis</td>
<td>K02703–K02714, K02716–K02720</td>
</tr>
<tr>
<td></td>
<td>K08901–K08904, K03541, K03542</td>
</tr>
<tr>
<td>nitrate respiration</td>
<td>K02567, K02568, K00370–K00373</td>
</tr>
<tr>
<td>nitrate reduction</td>
<td>K02567, K02568, K00370–K00374</td>
</tr>
<tr>
<td></td>
<td>K00367, K00360, K10534</td>
</tr>
<tr>
<td>hydrogen oxidoreduction</td>
<td>K00532–K00534, K06441, K18016, K18017, K18023</td>
</tr>
<tr>
<td></td>
<td>K00436 K05586–K05588, K18005–K18007</td>
</tr>
<tr>
<td>xylanolytic</td>
<td>K15924, K13465, K01198, K15920, K01181</td>
</tr>
<tr>
<td>anoxygenic photosynthesis</td>
<td>K08926–K08930, K08939–K08954</td>
</tr>
<tr>
<td></td>
<td>K13991, K13992, K13994</td>
</tr>
<tr>
<td>nitrite reduction to ammonium</td>
<td>K00362, K00363, K03385, K15876</td>
</tr>
<tr>
<td>sulfate-sulfide oxidoreduction (dsrAB)</td>
<td>K00394, K00395</td>
</tr>
<tr>
<td>denitrification (nosZ)</td>
<td>K00376</td>
</tr>
<tr>
<td>methanol oxidation</td>
<td>K00093, K14028, K14029, K16254–K16260, K17066</td>
</tr>
<tr>
<td>denitrification (norBC)</td>
<td>K04561, K02305</td>
</tr>
<tr>
<td>sulfate transport</td>
<td>K02045–K02048</td>
</tr>
<tr>
<td>nitrite respiration (nirKS)</td>
<td>K00368, K15864</td>
</tr>
<tr>
<td>methane or ammonia oxidation</td>
<td>K10944–K10946</td>
</tr>
<tr>
<td>cellulolytic</td>
<td>K19356, K19357, K01179, K01225, K19668</td>
</tr>
<tr>
<td>chitinolytic</td>
<td>K01183, K13381, K01452, K17523</td>
</tr>
</tbody>
</table>
Table A.4: OTUs per functional group. Number of OTUs assigned to each functional group, compared to the total number of taxonomically annotated OTUs. Some OTUs were assigned to multiple functional groups (see Fig. A.11 for functional group overlaps).

<table>
<thead>
<tr>
<th>functional group</th>
<th>OTUs</th>
<th>fraction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>aerobic ammonia oxidation</td>
<td>193</td>
<td>0.39</td>
</tr>
<tr>
<td>aerobic chemoheterotrophy</td>
<td>23565</td>
<td>47.4</td>
</tr>
<tr>
<td>aerobic nitrite oxidation</td>
<td>226</td>
<td>0.45</td>
</tr>
<tr>
<td>animal parasites or symbionts</td>
<td>3037</td>
<td>6.11</td>
</tr>
<tr>
<td>anoxygenic photoautotrophy</td>
<td>700</td>
<td>1.40</td>
</tr>
<tr>
<td>cellulolytic</td>
<td>175</td>
<td>0.35</td>
</tr>
<tr>
<td>chitinolytic</td>
<td>134</td>
<td>0.27</td>
</tr>
<tr>
<td>dark hydrogen oxidation</td>
<td>355</td>
<td>0.71</td>
</tr>
<tr>
<td>dark oxidation of reduced sulfur compounds</td>
<td>954</td>
<td>1.92</td>
</tr>
<tr>
<td>dark sulfide oxidation</td>
<td>327</td>
<td>0.66</td>
</tr>
<tr>
<td>dark sulfite oxidation</td>
<td>235</td>
<td>0.47</td>
</tr>
<tr>
<td>dark sulfur oxidation</td>
<td>282</td>
<td>0.57</td>
</tr>
<tr>
<td>fermentation</td>
<td>3146</td>
<td>6.33</td>
</tr>
<tr>
<td>hydrocarbon degradation</td>
<td>1776</td>
<td>3.57</td>
</tr>
<tr>
<td>ligninolytic</td>
<td>236</td>
<td>0.47</td>
</tr>
<tr>
<td>manganese oxidation</td>
<td>638</td>
<td>1.28</td>
</tr>
<tr>
<td>methanol oxidation</td>
<td>897</td>
<td>1.81</td>
</tr>
<tr>
<td>methanotrophy</td>
<td>104</td>
<td>0.21</td>
</tr>
<tr>
<td>methylotrophy</td>
<td>1023</td>
<td>2.06</td>
</tr>
<tr>
<td>nitrate denitrification</td>
<td>535</td>
<td>1.08</td>
</tr>
<tr>
<td>nitrate reduction</td>
<td>3399</td>
<td>6.84</td>
</tr>
<tr>
<td>nitrate respiration</td>
<td>1229</td>
<td>2.47</td>
</tr>
<tr>
<td>nitrite respiration</td>
<td>640</td>
<td>1.29</td>
</tr>
<tr>
<td>photoautotrophy</td>
<td>992</td>
<td>1.99</td>
</tr>
<tr>
<td>plant pathogen</td>
<td>511</td>
<td>1.03</td>
</tr>
<tr>
<td>sulfate respiration</td>
<td>300</td>
<td>0.61</td>
</tr>
<tr>
<td>ureolytic</td>
<td>1054</td>
<td>2.12</td>
</tr>
<tr>
<td>xylanolytic</td>
<td>58</td>
<td>0.12</td>
</tr>
</tbody>
</table>
Appendix B

Chapter 3: Supplemental material

B.1 Methods

B.1.1 Biological sample collection

Detritus from the bottom of bromeliad tanks was collected and physicochemical measurements were taken from all bromeliads in the period of January 8–10, 2015, within an area spanning roughly 0.2 km² in the Parque Nacional da Restinga de Jurubatiba, East coast Brazil (see Fig. B.6 for coordinates). At that time, weather conditions were sunny, dry and hot, and were preceded by several weeks of extreme drought (343). This drought may explain why we detected almost no insects in the bromeliad tanks that we sampled. Supernatant liquid was removed from the bromeliad’s central tank using a sterile serological pipette. The detritus at the bottom was then retrieved using a sterile syringe and a metal spatula, after cutting the bromeliad open for easier access. The entire retrieved detrital content was mixed before sampling. Samples were flash-frozen in liquid nitrogen within 10 minutes of collection and then frozen in the laboratory at −80°C until further processing. For shipment, samples were concentrated via centrifugation (40 000 g for 15 min, balanced using miliQ filtered water) and removal of the supernatant fluid, and then freeze-dried for 24 hours. The dried samples were shipped to our lab at the University of British Columbia, Canada, for further processing.

B.1.2 Chemical analysis of tank water

The water above the benthic detritus was collected using a serological pipette, stored in 25 mL centrifuge tubes on regular ice in the field and at −4°C in the lab until further analysis (within two days). Water samples for CH₄ measurement were taken separately (1.5 mL per measurement) and directly from the bromeliad, fixed using formalin (4 %) in 3 mL glass vials, stored on regular ice in the field and at −4°C in the lab until further analysis. Total dissolved phosphorus concentrations were determined as the inorganic phosphorus obtained
after a procedure of acid-digestion and autoclaving of the water samples and the ascorbic acid-molybdate reaction (155). Total dissolved nitrogen concentrations were determined as the concentration of nitrate obtained after an acid digestion procedure and autoclavage, in a Flow Injection Analysis System (FIA-Asia Ismatec™) (539). Methane concentrations were determine using a Shimadzu™ GC-2010AF chromatograph equipped with a Rt-QPLOT column (3 m × 0.32 m) and a flame ionization detector (FID-2010). Temperatures of the injection, column and detection were 120°C, 85°C and 220°C, respectively. Nitrogen (N₂) was used as the carrier gas.

Conductivity, pH, temperature and Total Suspended Solids (TSS) were measured in the field using an ExStik II EC500™ (ExTech Instruments). Salinity was calculated from conductivity and temperature using the empirical formula reported by Fofonoff and Millard-Junior (127). Water turbidity was measured in the field using a Hanna Turbidimeter HI98703. Absorption spectra were measured in the lab using a Varian 50 Bio UV-Visible Spectrophotometer™, following the manufacturer’s procedures. Dissolved organic carbon (DOC) concentrations were determined using by Pt-catalyzed high-temperature combustion with a Shimadzu TOC-VCPN Total Carbon Analyzer™, after filtering through 0.7 µm Whatman™ GF/F glass fiber filters.

For one bromeliad the retrieved supernatant water was insufficient for performing all of the chemical assays in the field. That water sample was thus diluted at a ratio 1:5 using deionized water prior to measuring conductivity, pH, TSS and turbidity. The resulting conductivity, salinity, TSS and turbidity were then corrected using the dilution factor. The pH was corrected using a standard curve constructed by serial dilution of water from another bromeliad. For several bromeliads the retrieved supernatant water was insufficient for measuring absorption spectra and DOC concentrations, as well as for excitation-emission spectrophotometry (EES; explained below). These water samples were thus diluted in the lab using deionized water as needed. All measurements were subsequently corrected for the effects of dilution.

EES of the water samples was performed using a Varian Cary Eclipse™ fluorescence spectrophotometer. In EES, each sample is exposed to light of several wavelengths while simultaneously measuring the resulting fluorescence spectrum (14). The obtained “excitation-emission matrices” (EEM) were analyzed for organic carbon profiles using parallel factor analysis (PARAFAC) with the MATLAB® package drEEM (338). EEMs were pre-processed as follows: The EEM of pure milli-Q™ water was subtracted from the sample EEMs. Rayleigh (elastic) and Raman (inelastic) scatter signals were removed by replacing them with NaN. EEM entries for emission wavelengths lower than the excitation wavelengths were set to
zero. EEM entries at the excitation wavelengths 320 nm and 365 nm were ignored because of abnormal intensity troughs at all emission wavelengths, likely resulting from imperfections of the fluorometer lamp. EEMs were corrected for inner filter effects using the sample absorption spectra and the drEEM function \texttt{fdomcorrect} as described by Murphy et al. (338).

PARAFAC model fitting was attempted for various model sizes (3–9). To avoid local PARAFAC minima, fitting for each model size was repeated 50 times with random initialization using the drEEM function \texttt{randinitanal}. Model residuals were inspected as described by Stedmon et al. (445) and Stedmon and Bro (444) to ensure that the model size was sufficient. Split-half validation (‘$S_4C_4T_2$’; 338) failed for all considered model sizes, but was ignored because of low sample size when compared to the high richness of observed EEM profiles. Instead, to constrain the model’s size and avoid overfitting, model components were inspected for physical plausibility as described by Murphy et al. (338, e.g., Fig. 7) and subsequently compared to published entries in the OpenFluor fluorophore database based on Tucker’s congruence coefficient (339). We kept the model (size 4; Fig. B.7) with the highest number of plausible components represented in OpenFluor at a congruence of at least 0.98. The best matches in the OpenFluor database were “CS-Galathea, C1” for component 1 (216), “Recycle_WRAMS, C5” for component 2 (337), “PrairieLakes, C2” for component 3 (355) and “FloridaKeys, C3” for component 4 (535). The model explained 98.2% of the variance, at a core consistency of 82.9% (Fig. B.8). For each sample and for each individual PARAFAC component we determined the maximum fluorescence intensity in the component’s EEM, and multiplied it by the component’s score in the particular sample. This yielded 4 PARAFAC component intensities per sample, each in arbitrary units that are comparable across samples but not across PARAFAC components. These component intensities were subsequently used in our analysis as 4 additional environmental variables (“PARAFAC 1–4”).

**B.1.3 Measurement of other physicochemical variables**

Light intensity (photosynthetic photon flux density) on bromeliads was measured using an LI-250A Lightmeter\textsuperscript{TM} (LI-COR Biosciences), placed on the ground next to the bromeliad at noon of a sunny day (January 10, 2015), after trimming the bromeliad’s foliage to avoid shading of the device by the bromeliad itself. The detrital volume was measured using the centrifuge tube scale after allowing for precipitation for 5 minutes, performing the read at the interface between the precipitated detritus and the supernatant transparent fluid. The
total tank volume was set to the total volume of all retrieved material (detritus and water). The total tank depth was either measured using a metal wire with engraved cm-scale, or using the serological pipette’s volume scale upon calibration. Tree cover (“shading”) above bromeliads was measured by taking a photo from the top of a bromeliad “face-up” on a sunny day, and processing the photo using ImageJ™ for contrasting objects against a blue sky background. An overview of all physicochemical environmental variables is provided in Table B.5.

### B.1.4 16S sequencing

DNA was extracted from the re-hydrated samples using the MoBio PowerSoil® DNA extraction kit, by applying the manufacturer’s suggested protocol. Amplification of the 16S rRNA gene was done using barcoded primers covering the V4 region (E. coli 515F and 806R) that included Illumina adapters, and using the Earth Microbiome Project 16S amplification protocol version 4.13 (60). Amplicon DNA from all samples was pooled into a single library, at such proportions that each sample contributed a similar amount of DNA. Primer dimers and remaining PCR enzymes were removed from the amplicon library using the MoBio UltraClean PCR Clean-Up Kit. Library quantitation was performed by Genoseq Core (University of California, Los Angeles) using a high-sensitivity Agilent Bioanalyzer™ and Kappa Biosystems’ Illumina Genome Analyzer™ (KAPA SYBR FAST Roche LightCycler 480) kit, followed by qPCR. Sequencing was performed by Genoseq Core using an Illumina MiSeq™ next generation sequencer, following the manufacturer’s standard protocol.

Sequencing yielded 15,090,774 paired-end sequences (2 × 300 base pairs each). Sequence analysis was performed using the QIIME toolbox (version 1.9.1, 59). Paired-end reads were merged after trimming forward reads at length 240 and reverse reads at length 160. Merged sequences were quality filtered using QIIME’s default settings, yielding 9481,315 sequences of median length 253. Remaining sequences were de-noised and clustered de-novo using cd-hit-otu (273) at a 99% 16S rDNA similarity threshold, generating 2884 operational taxonomic units (OTUs) representing 1,908,183 sequences across all samples. Sample B17 yielded by far the fewest sequences (5811 sequences corresponding to 677 OTUs).

We note that historically a less stringent threshold of 97% 16S rDNA similarity was recommended for delineating prokaryotic OTUs in biogeographical studies (145). However, recent work shows that greater taxonomic resolution is needed to detect signals of endemism (e.g., up to 99.5% for the cyanobacterium Prochlorococcus; 305) and signals of competitive exclusion (99–100%; 244), and that taxa defined on the basis of 97% similarity may be
underspeciated (232, 439).

We did not rarefy the OTU table so as to obtain as much of an accurate estimation of OTU proportions as possible for analyses based on quantitative abundances (317). Moreover, rarefaction prior to our presence-absence-based analyses (see details below) would have led to higher estimates of OTU turnover between samples (Table B.1) as well as higher checkerboard C-scores between OTUs (Table B.3) because OTUs would be “competing” for sequences. This would further strengthen the patterns upon which our conclusions are based. Diagnostic OTU rarefaction curves are shown in Fig. B.9.

Taxonomic assignment of representative sequences was done using uclust (110) and the SILVA reference database (release 119, 378), using the first 50 hits at a similarity threshold of at least 90% as follows: For any queried sequence, if at least one hit had a similarity $s \geq 99\%$, then all hits with similarity $s$ were used to form a consensus taxonomy. Otherwise, if at least one hit had a similarity $s \geq 90\%$, then all hits with similarity at least $(s - 1\%)$ were used to form a consensus taxonomy. If a query did not match any reference sequence at or above 90% similarity, it was considered unassigned. A total of 1965 OTUs (representing 1,874,361 sequences across all samples) were taxonomically annotated.

Representative sequences were aligned against the SILVA database using PyNAST (58, 378), and phylogenetic relationships were calculated using the FastTree algorithm (372), at standard QIIME settings. Phylogenetic distances are in nucleotide substitutions per site.

### B.1.5 Functional annotation of prokaryotic taxa (FAPROTAX)

To determine the taxonomic composition within each of the 9 considered functional groups (aerobic chemoheterotrophy, cellulolysis, fermentation, methanogenesis, methylotrophy, nitrogen respiration, sulfate respiration, photoautotrophy, ureolysis), we associated each taxonomically annotated OTU with one or more functions based on extensive literature search, whenever possible. Details of this approach, which we outline here, are provided by Louca et al. (289). In short, a taxon (e.g., strain, species or genus) and all OTUs within that taxon were associated with a particular metabolic function if all cultured representatives within the taxon are known to exhibit that function. We note that as the number of cultured strains continues to increase, some of these generalizations may turn out to be false. Furthermore, a substantial fraction of OTUs could not be assigned to any function, thus OTU proportions inside a functional group only apply to the subset of functionally characterized OTUs (although this limitation does not affect the conclusions of this study). In total 465 out of
1965 OTUs were assigned to at least one functional group, yielding in total 518 functional annotations (see Table B.6 for an overview). OTUs without any functional annotation were omitted from the analysis.

We note that FAPROTAX functional groups are not completely one-to-one comparable with metagenomic gene groups, due to ambiguities in the functions potentially performed by some genes (376). To strengthen our confidence in the stability of the 9 considered functional groups, we provide detailed gene-centric functional profiles for multiple related functions (Figs. 3.2A,B and B.10).

### B.1.6 Metagenomic sequencing

To assess the functional stability of microbial communities across samples, we performed shotgun environmental DNA sequencing (metagenomics), which allows the detection of known genes in an environment regardless of their host organisms. Extracted DNA was sequenced in 100-bp paired-end fragments on an Illumina HiSeq 2000™. Library preparation and sequencing was done by the Biodiversity Research Centre NextGen Sequencing Facility and followed standard Illumina protocols. All samples were uniquely barcoded and run together on a single lane. The resulting sequence data were preprocessed using Illumina’s CASAVA-1.8.2. Specifically, output files were converted to fastq format, and sequences were separated by barcode (allowing one mismatched base pair), using the `configureBclToFastq.pl` script. This yielded a total of 151,308,568 quality-filtered paired-end reads. Reads were trimmed at the beginning and end to increase average read quality, yielding an average forward and reverse read length of 97 and 98 bp, respectively. Sufficiently overlapping paired-end reads were merged using PEAR 0.9.8 with default options (544), yielding 17,007,327 merged reads. Non-merged read pairs were deduplicated using the SOFA pipeline (168, version 1.2) and the KEGG protein reference database (221, release 2011.06.18), in order to reduce potential double-counts during subsequent gene annotation. MetaPathways 2.5 (247) was used for ORF prediction in all merged and non-merged reads (min peptide length 30, algorithm prodigal), yielding 215,140,278 ORFs. Predicted ORFs were taxonomically annotated in MetaPathways using LAST and the NCBI RefSeq protein database (470, release 2015.12.12), and multiple taxonomic annotations were consolidated using a lowest common ancestor algorithm (247). Non-prokaryotic ORFs were excluded from subsequent analysis. LAST annotation of prokaryotic ORFs against the KEGG protein reference database was performed using MetaPathways (KEGG release 2011.06.18, min BSR 0.4, max E-value $10^{-6}$, min score 20, min peptide length 30, top hit), yielding 55,058,696
Chapter 3: Supplemental material

annotations. Metagenomic KEGG orthologous group (KOG) counts (221) were normalized using the total number of KEGG-annotated sequences per sample (total sum scaling).

To estimate the variability of the 9 functional groups considered in this study, we examined the abundances of selected proxy genes that roughly corresponded to one or more functional groups. These genes were chosen based on the KEGG reference pathway database (221) and were identified using the KOG annotations of prokaryotic ORFs. Whenever applicable, multiple KOGs associated with similar metabolic functions (e.g., dissimilatory nitrite reduction to ammonium, \textit{nirBD} and \textit{nrfAH}) were combined into a single gene group. An overview of KOGs associated with each function is provided in Table B.7. The resulting metagenomic profiles are given in Figs. 3.2A,B and B.10.

B.1.7 Comparing functional and taxonomic variability

To robustly compare the degree of functional variability and taxonomic variability within functional groups we used multiple statistical measures that are either entirely based on presences and absences (“binary”) or that take into account relative abundances. Specifically, for every functional group, we measured the binary OTU “overlap” between any two samples in terms of the Jaccard overlap index, defined as the number of OTUs detected in both samples, divided by the number of OTUs detected in any of the two samples (527). Hence, a Jaccard overlap of 1 corresponds to complete overlap (regardless of OTU proportions), while a Jaccard overlap of 0 corresponds to no overlap at all. Mean Jaccard overlaps (MJO), i.e., Jaccard overlaps averaged over all bromeliad pairs, were within the range \( \sim 0.2–0.6 \) for all functional groups (Table B.1). These low MJOs indicate substantial differences in community structure across bromeliads, however in principle they could result purely from detection stochasticity, especially of rare OTUs (i.e., due to insufficient sequencing depth)\(^189\). To determine the statistical significance (“P-value”) of these low MJOs, we compared them to random MJOs generated under a null model of random sampling from the regional OTU pool. Specifically, for any given functional group, OTUs were randomly drawn from a multinomial distribution corresponding to the OTU proportions in the regional OTU pool, while the number of draws per bromeliad was equal to the number of sequences assigned to the functional group in that bromeliad. The P-value of an observed MJO was defined as the probability that a random MJO would be lower than the observed MJO, and was estimated based on 1000 iterations. All functional groups had a significantly low MJO (\( P < 0.001 \)), showing that low overlaps are not just the result of detection stochasticity. We note that the Jaccard overlap of gene groups was 1 for all sample pairs, since all considered gene groups
were detected in all samples.

To account for OTU or gene group abundances, we also considered the Morisita overlap index (332). This index is particularly suited for cases where abundance estimates are obtained at varying sequencing depths (527), and its interpretation is analogous to the Jaccard index (results in Table B.1). We mention that the Morisita overlap of gene groups (0.98 when averaged across sample pairs) is slightly below 1, because gene group abundances do vary between samples.

To verify the robustness of our conclusions based on overlap indices, we also examined the coefficients of variation (CV, i.e., the standard deviation divided by the mean) of relative gene group abundances on the one hand, and the CVs of OTU proportions within individual functional groups, on the other hand. Because each particular functional group contained multiple OTUs, we averaged the CV over all OTUs within the functional group. We note that the considered gene groups (Fig. 3.2A) only cover a small fraction of the total detected gene pool (∼5% of annotated metagenomic sequences). Hence, to minimize the dependence of the CV of any particular gene group on the choice and coverage of other gene groups, we considered gene group abundances relative to the total number of annotated metagenomic sequences in each sample. An overview of CVs is provided in Table B.2. Observe that OTU CVs are generally an order of magnitude higher than gene group CVs, consistent with our conclusions based on overlap indices.

B.1.8 Metric multidimensional scaling and coloring

Pairwise dissimilarities between taxonomic community profiles reported here were calculated using the Bray-Curtis metric, which is widespread in biogeographical studies (266). Other dissimilarity metrics (Canberra and Hellinger) generally yielded similar conclusions, so they are not further discussed here. Each dissimilarity matrix was used to visualize differences in community composition via metric multidimensional scaling (MDS; 42). In MDS, sample points are embedded into a reduced number of dimensions (e.g., 3) such that the pairwise Euclidean sample distances in the embedding “best match” the original dissimilarities. Hence, points that are closer to each other in the embedding correspond to samples with more similar microbial communities. The embedding was performed by minimizing the Kruskal stress, using the Scikit-lean package (363). 3-dimensional MDS coordinates were mapped to color space by associating each coordinate with one color channel (red, green and blue; Fig. B.6), hence sample pairs with similar microbial communities are colored similarly, allowing easier identification. In all reported cases, the Kruskal stress of the MDS embeddings was below
B.1.9 Phylogenetic community structure

To assess whether community assembly within individual functional groups was driven by purely stochastic processes (such as lottery effects; 52) or was subject to deterministic selection, we examined the phylogenetic distances between functionally similar OTUs co-occurring in each sample. The phylogenetic distance (PD) between two OTUs was calculated as the sum of branch lengths needed to traverse the phylogenetic tree from one OTU to the other. The mean phylogenetic distance (MPD) in a community of $M$ OTUs was calculated as

$$\text{MPD} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{i-1} d_{ij} N_i N_j}{\sum_{i=1}^{M} \sum_{j=1}^{i-1} N_i N_j},$$

(B.1.1)

where $N_i$ is the abundance of OTU $i$ in the community and $d_{ij}$ is the phylogenetic distance between OTUs $i$ and $j$. Note that this definition of MPD is almost equivalent to the “phylogenetic diversity” introduced by Chave et al. (66), with the difference that Chave et al. defined $d_{ij}$ as the divergence time between two OTUs (which is half of their phylogenetic distance in most cases).

For each sample, the MPD was compared to the expected MPD ($\overline{\text{MPD}}$) under the null hypothesis of random phylogenetic relationships between OTUs. Specifically, the distribution of MPDs under the null hypothesis was estimated by randomly and repeatedly permuting OTUs in the phylogenetic tree 1000 times, while keeping their proportions in each sample fixed. OTUs were permuted independently for each sample, and permutations were restricted to OTUs within the same functional group. The standardized effect size (SES) of a sample — which quantifies the deviation of the observed MPD from the expectation of the null hypothesis, was calculated as

$$\text{SES} = \frac{\text{MPD} - \overline{\text{MPD}}}{\sigma_{\text{MPD}}},$$

(B.1.2)

where $\sigma_{\text{MPD}}$ is the standard deviation of random MPDs generated by the null hypothesis. Hence, a strongly positive or strongly negative SES corresponds to strong phylogenetic overdispersion or underdispersion, respectively. The SESs for all samples and within each functional group are shown in Fig. 3.4.
For several functional groups, SESs are either predominantly positive or predominantly negative across samples, indicating that the MPDs within these groups may not be random (i.e., are inconsistent with the null hypothesis). The statistical significance (P-value) of these imbalances (i.e., the difference between the number of positive and negative SES across samples) was defined as the probability that random SESs generated by the null hypothesis would display comparable or stronger imbalances in magnitude, and was estimated using 1000 random iterations. Hence, a low P-value means that the observed imbalance of SESs within a functional group is unlikely to have occurred by chance.

B.1.10 Comparing OTU proportions to environmental variables

To test whether the variation of taxonomic composition within functional groups can be attributed to variation in environmental conditions, we constructed multivariate non-linear regression models for each OTU in each functional group, using 21 environmental variables as predictors (overview of environmental variables in Table B.5). Specifically, we used non-linear kernel ridge regression (KRR) with Gaussian radial kernels, implemented by the scikit-learn software (363). An important advantage of KRR models over conventional (e.g., multivariate linear) regression models is the use of regularization, which reduces the risk of overfitting by penalizing excessive model coefficients, thus avoiding excessive model complexity. The final model complexity depends on a parameter that influences the extent to which coefficients are penalized. KRR models generally present a more robust alternative to step-wise model selection methods (88). Non-linearities in the data are addressed using the “kernel-trick”, which replaces predictors with higher-dimensional variables using Gauss-shaped functions prior to linear regression (363). The predictive power of each KRR model was measured in terms of the 10-fold cross-validated coefficient of determination ($R^2_{cv}$), which represents the achievable coefficient of determination when only a random subset (90%) of the samples are used for fitting and the remaining samples (10%) are used for independent testing (421). Hence, $R^2_{cv}$ provides a more conservative estimate of the predictive power than the classical coefficient of determination ($R^2$). The $R^2_{cv}$ was determined via 10-fold Monte-Carlo cross-validation with 500 random iterations. The penalization parameter as well as the Gaussian kernel radius were optimized for each KRR model using grid search and maximization of the achievable $R^2_{cv}$.

To further assess the potential influence of individual environmental variables on taxonomic community composition, we calculated Spearman rank correlations between OTU proportions within functional groups on the one hand, and environmental variables on the other.
hand. For each functional group and each environmental variable, we calculated the average magnitude of correlations across all OTUs (“average absolute correlation”; Fig. 3.5). Hence, a large average absolute correlation means that OTUs in a particular functional group tend to be strongly (negatively or positively) correlated with a particular environmental variable. The statistical significance of large average absolute correlations was estimated using 1000 random permutations of columns (i.e., samples) in the OTU table.

B.1.11 Comparing community dissimilarities to geographical distances

To examine whether dispersal limitation across bromeliads had an effect on the composition within individual functional groups, we calculated Spearman rank Mantel correlations between pairwise geographical distances and pairwise dissimilarities (in terms of OTU proportions within functional groups). To ensure the robustness of our conclusions, we considered three different dissimilarity metrics — Bray-Curtis, Canberra and Hellinger, all of which are widely used in ecology (266). We estimated the statistical significance of correlations using 1000 random permutations of the rows and columns in the geographical distance matrix (rows and columns permuted similarly). None of the considered dissimilarity metrics yielded any statistically significant correlations to geographical distance. An overview of results for the Bray-Curtis dissimilarity metric is given for illustration in Table B.4. A visual comparison of geographical distances and Bray-Curtis dissimilarities is shown in Fig. B.5.

B.1.12 Comparing OTU co-occurrences to a null model

To examine whether OTU co-occurrences across samples follow non-random patterns (e.g., resulting from competitive exclusion), we considered a statistical quantity known as the checkerboard score (“C-score”) of the OTU presence-absence matrix (159). The C-score is defined as

\[ C = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=1}^{i-1} (N_i - N_{ij})(N_j - N_{ji}), \]

where \( M \) is the total number of considered OTUs, \( N_i \) is the number of samples containing OTU \( i \) and \( N_{ij} \) is the number of samples containing both OTUs \( i \) and \( j \). Hence, for fixed \( N_i \), the C-score becomes larger if species co-occur less frequently (i.e., \( N_{ij} \) are smaller). To assess whether an observed C-score was likely due to chance (i.e., if OTUs occur independently of
each other), we compared it to the C-score distribution of several random presence-absence matrixes generated under the “fixed-fixed” null model (159). This null model preserves the number of samples containing each OTU as well as the number of OTUs present in each sample and in each functional group, and is thus suitable for detecting non-random co-occurrence patterns across samples that may differ in terms of OTU richness, while maintaining a low false positive error rate (76). Specifically, if C-scores generated by the null hypothesis are typically lower than the observed C-score, this would mean that OTUs tend to exclude each other more often than expected by chance (i.e., are segregated). We calculated the C-score and its deviation from the null model separately for each functional group. Randomized presence-absence matrixes corresponding to the null model were generated using the “curve-ball” algorithm (456). We used 1000 random matrixes to assess the statistical significance of C-scores. An overview of results is given in Table B.3.

### B.1.13 Sequence data availability

Molecular sequence data reported in this chapter have been deposited in the NCBI Bio-Project database (http://www.ncbi.nlm.nih.gov/bioproject) and will be made public upon publication of this work (BioProject no. PRJNA321235; SRA accession nos. SRP074855 and SRP074855).
Figure B.1: **Taxonomic composition within functional groups (genus level).** Proportions of prokaryote genera within individual functional groups (one color per genus, one bar stack per sample, one plot per functional group). Samples are sorted alphabetically as in Fig. 3.2.
Figure B.2: Taxonomic composition within functional groups (family level). Proportions of prokaryote families within individual functional groups (one color per family, one bar stack per sample, one plot per functional group). Samples are sorted alphabetically as in Fig. 3.2.
Figure B.3: **Taxonomic composition within functional groups (order level)**. Proportions of prokaryote orders within individual functional groups (one color per order, one bar stack per sample, one plot per functional group). Samples are sorted alphabetically as in Fig. 3.2.
Figure B.4: **Taxonomic composition within functional groups (class level).** Proportions of prokaryote classes within individual functional groups (one color per class, one bar stack per sample, one plot per functional group). Samples are sorted alphabetically as in Fig. 3.2.
Figure B.5: Geographical distances vs dissimilarities within functional groups. Bray-Curtis dissimilarities between samples (in terms of OTU proportions within individual functional groups), compared to geographical sample distances (one point per sample pair, one plot per functional group).
Figure B.6: **Geographic location vs composition within functional groups.** Each plot: Geographical sample locations in terms of longitude and latitude (one point per sample). Points are colored according to Bray-Curtis dissimilarities between samples, in terms of OTU proportions within individual functional groups (one plot per functional group). Similar colors correspond to similar compositions within functional groups (see Methods for details).
Figure B.7: **PARAFAC model components.** Left column: Excitation-emission matrixes of the 4 PARAFAC model components, estimated for the excitation-emission spectra of the bromeliad detrital samples. Right column: Excitation and emission spectra corresponding to the PARAFAC components.
Figure B.8: **Modeling EEMs of bromeliad DOC with PARAFAC.** Left column: Measured excitation-emission matrixes (EEM) for a subset of bromeliad samples that illustrates the detected fluorophore diversity (B1, B4, B12 and B31). Middle column: Corresponding EEMs modeled by the 4-component PARAFAC model. Right column: Corresponding residual EEMs. White horizontal bands cover EEM pixels that we omitted from the analysis due to spurious excitation troughs erroneously detected by the fluorometer. Diagonal white bands cover Rayleigh and Raman scatters, which we also omitted from the analysis.
Figure B.9: **16S rDNA rarefaction curves (OTU richness)**. Each plot: Expected number of observed distinct OTUs at various sequencing depths for a particular sample, determined through repeated random rarefactions.
Figure B.10: **Detailed functional community profiles.** (A) Detailed gene-centric prokaryotic functional profiles, in terms of functional group proportions inferred from metagenomic sequences (one color per function, one column per sample). (B) Same as (A), but focusing on the 10 least abundant functional groups.
Figure B.11: **Functional redundancy at the genus level.** Association of functional groups (columns) with members of various prokaryote genera (rows). A darker color corresponds to a higher relative contribution of a genus (in terms of the number of associated OTUs) to a functional group. Rows and columns are sorted according to the number of non-zero entries within them. For analogous plots at the OTU, family and class level, see Figs. 3.3, B.12 and B.13, respectively.
Figure B.12: **Functional redundancy at the family level.** Association of functional groups (columns) with members of various prokaryote families (rows) across all samples. A darker color corresponds to a higher relative contribution of a family (in terms of the number of associated OTUs) to a functional group. Rows and columns are sorted accorded to the number of non-zero entries within them. For analogous plots at the OTU, genus or class level, see Figs. 3.3, B.11 and B.13, respectively.
Figure B.13: **Functional redundancy at the class level.** Association of functional groups (columns) with members of various microbial classes (rows). A darker color corresponds to a higher relative contribution of a class (in terms of the number of associated OTUs) to a functional group. Rows and columns are sorted accorded to the number of non-zero entries within them. For analogous plots at the OTU, genus and family level, see Figs. 3.3, B.11 and B.12, respectively.
Table B.1: **OTU overlap between samples.** Overview of pairwise OTU overlaps between any two samples in terms of the Jaccard index (number of OTUs shared by both samples divided by the number of OTUs present in any of the two samples), averaged over all sample pairs, and overall overlap across all samples (number of OTUs present in all samples divided by the number of OTUs present in at least one sample), within individual functional groups. The last column lists the mean pairwise Morisita overlap index, which takes into account OTU proportions.

<table>
<thead>
<tr>
<th>group</th>
<th>Jaccard overlap</th>
<th>overall overlap</th>
<th>Morisita overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>aerobic chemoheterotrophs</td>
<td>0.42</td>
<td>0.018</td>
<td>0.20</td>
</tr>
<tr>
<td>cellulolytic</td>
<td>0.23</td>
<td>0.0</td>
<td>0.28</td>
</tr>
<tr>
<td>fermenters</td>
<td>0.39</td>
<td>0.0</td>
<td>0.28</td>
</tr>
<tr>
<td>methanogens</td>
<td>0.62</td>
<td>0.077</td>
<td>0.31</td>
</tr>
<tr>
<td>methylotrophs</td>
<td>0.57</td>
<td>0.042</td>
<td>0.37</td>
</tr>
<tr>
<td>nitrogen respirers</td>
<td>0.48</td>
<td>0.0</td>
<td>0.35</td>
</tr>
<tr>
<td>photoautotrophs</td>
<td>0.26</td>
<td>0.0</td>
<td>0.24</td>
</tr>
<tr>
<td>sulfate respirers</td>
<td>0.32</td>
<td>0.0</td>
<td>0.27</td>
</tr>
<tr>
<td>ureolytic</td>
<td>0.36</td>
<td>0.0</td>
<td>0.38</td>
</tr>
</tbody>
</table>
Table B.2: **Coefficients of variation.** Coefficient of variation (CV = standard deviation divided by mean) for the relative abundance of each gene group, and average coefficients of variation of OTU proportions within each functional group (averaged over all OTUs in a functional group).

<table>
<thead>
<tr>
<th>gene group</th>
<th>CV of relative gene abundances</th>
</tr>
</thead>
<tbody>
<tr>
<td>heterotrophy (PTS)</td>
<td>0.20</td>
</tr>
<tr>
<td>oxygen respiration (cox)</td>
<td>0.11</td>
</tr>
<tr>
<td>carbon fixation</td>
<td>0.098</td>
</tr>
<tr>
<td>monosaccharide ABC transporters</td>
<td>0.23</td>
</tr>
<tr>
<td>cellulolysis</td>
<td>0.31</td>
</tr>
<tr>
<td>fermentation</td>
<td>0.071</td>
</tr>
<tr>
<td>methanogenesis</td>
<td>0.56</td>
</tr>
<tr>
<td>methylotrophy</td>
<td>0.65</td>
</tr>
<tr>
<td>nitrogen respiration</td>
<td>0.15</td>
</tr>
<tr>
<td>ureolysis</td>
<td>0.29</td>
</tr>
<tr>
<td>photoautotrophy</td>
<td>0.54</td>
</tr>
<tr>
<td>dissimilatory sulfur metabolism</td>
<td>0.59</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>functional group</th>
<th>average CV of OTU proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td>aerobic chemoheterotrophs</td>
<td>2.9</td>
</tr>
<tr>
<td>cellulolytic</td>
<td>2.7</td>
</tr>
<tr>
<td>fermenters</td>
<td>3.1</td>
</tr>
<tr>
<td>methanogens</td>
<td>2.3</td>
</tr>
<tr>
<td>methylotrophs</td>
<td>2.5</td>
</tr>
<tr>
<td>nitrogen respirers</td>
<td>2.4</td>
</tr>
<tr>
<td>photoautotrophs</td>
<td>3.3</td>
</tr>
<tr>
<td>sulfate respirers</td>
<td>3.1</td>
</tr>
<tr>
<td>ureolytic</td>
<td>2.9</td>
</tr>
</tbody>
</table>
Table B.3: **OTU co-occurrence patterns.** Overview of checkerboard analysis of OTU co-occurrences within each functional group, including standardized effect sizes (SES) of the C-scores, statistical significances (P) and interpretation of co-occurrence patterns.

<table>
<thead>
<tr>
<th>group</th>
<th>SES</th>
<th>P</th>
<th>interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>aerobic chemoheterotrophs</td>
<td>8.07</td>
<td>&lt;0.001</td>
<td>segregated</td>
</tr>
<tr>
<td>cellulolytic</td>
<td>2.6</td>
<td>0.008</td>
<td>segregated</td>
</tr>
<tr>
<td>fermenters</td>
<td>7.4</td>
<td>&lt;0.001</td>
<td>segregated</td>
</tr>
<tr>
<td>methanogens</td>
<td>1.5</td>
<td>0.07</td>
<td>random, slightly segregated</td>
</tr>
<tr>
<td>methylotrophs</td>
<td>0.23</td>
<td>0.35</td>
<td>random</td>
</tr>
<tr>
<td>nitrogen respirers</td>
<td>1.7</td>
<td>0.05</td>
<td>segregated</td>
</tr>
<tr>
<td>photoautotrophs</td>
<td>3.6</td>
<td>0.002</td>
<td>segregated</td>
</tr>
<tr>
<td>sulfate respirers</td>
<td>6.6</td>
<td>&lt;0.001</td>
<td>segregated</td>
</tr>
<tr>
<td>ureolytic</td>
<td>0.39</td>
<td>0.31</td>
<td>random</td>
</tr>
</tbody>
</table>

Table B.4: **Geographical distances vs taxonomic dissimilarities.** Overview of Mantel Spearman rank correlation tests between pairwise geographical distances and Bray-Curtis dissimilarities (in terms of OTU proportions within functional groups).

<table>
<thead>
<tr>
<th>group</th>
<th>correlation</th>
<th>statistical significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>aerobic chemoheterotrophs</td>
<td>-0.043</td>
<td>0.30</td>
</tr>
<tr>
<td>cellulolytic</td>
<td>-0.033</td>
<td>0.36</td>
</tr>
<tr>
<td>fermenters</td>
<td>0.017</td>
<td>0.39</td>
</tr>
<tr>
<td>methanogens</td>
<td>-0.016</td>
<td>0.46</td>
</tr>
<tr>
<td>methylotrophs</td>
<td>0.067</td>
<td>0.18</td>
</tr>
<tr>
<td>nitrogen respirers</td>
<td>0.12</td>
<td>0.08</td>
</tr>
<tr>
<td>photoautotrophs</td>
<td>-0.016</td>
<td>0.43</td>
</tr>
<tr>
<td>sulfate respirers</td>
<td>0.027</td>
<td>0.31</td>
</tr>
<tr>
<td>ureolytic</td>
<td>-0.025</td>
<td>0.39</td>
</tr>
</tbody>
</table>
Table B.5: Environmental variables. Overview of physicochemical environmental variables, including mean, standard deviation and measurement units.

<table>
<thead>
<tr>
<th>variable</th>
<th>mean</th>
<th>std.</th>
<th>units</th>
</tr>
</thead>
<tbody>
<tr>
<td>absorption at 240 nm</td>
<td>2.12</td>
<td>1.39</td>
<td>-</td>
</tr>
<tr>
<td>detrital volume</td>
<td>5.34</td>
<td>4.42</td>
<td>mL</td>
</tr>
<tr>
<td>CH$_4$ concentration</td>
<td>5.34</td>
<td>4.42</td>
<td>$\mu$M</td>
</tr>
<tr>
<td>shading</td>
<td>42.3</td>
<td>16.8</td>
<td>% cover of face-up view</td>
</tr>
<tr>
<td>tank depth</td>
<td>10.6</td>
<td>2.31</td>
<td>cm</td>
</tr>
<tr>
<td>DOC</td>
<td>80.9</td>
<td>49.8</td>
<td>mg · L$^{-1}$</td>
</tr>
<tr>
<td>plant height</td>
<td>56.0</td>
<td>8.17</td>
<td>cm</td>
</tr>
<tr>
<td>light intensity</td>
<td>1064</td>
<td>822</td>
<td>$\mu$mol · m$^{-2}$ · s$^{-1}$ (photosynthetic photon flux density)</td>
</tr>
<tr>
<td>number of leaves</td>
<td>8.31</td>
<td>2.22</td>
<td>-</td>
</tr>
<tr>
<td>total nitrogen</td>
<td>83.4</td>
<td>55.9</td>
<td>$\mu$M</td>
</tr>
<tr>
<td>total phosphorous</td>
<td>3.15</td>
<td>2.76</td>
<td>$\mu$M</td>
</tr>
<tr>
<td>molar N:P ratio</td>
<td>29.6</td>
<td>6.8</td>
<td>-</td>
</tr>
<tr>
<td>PARAFAC 1</td>
<td>212</td>
<td>147</td>
<td>-</td>
</tr>
<tr>
<td>PARAFAC 2</td>
<td>64.9</td>
<td>63.9</td>
<td>-</td>
</tr>
<tr>
<td>PARAFAC 3</td>
<td>76.9</td>
<td>218</td>
<td>-</td>
</tr>
<tr>
<td>PARAFAC 4</td>
<td>31.3</td>
<td>32.4</td>
<td>-</td>
</tr>
<tr>
<td>pH</td>
<td>5.21</td>
<td>0.71</td>
<td>-</td>
</tr>
<tr>
<td>salinity</td>
<td>0.047</td>
<td>0.031</td>
<td>PSU</td>
</tr>
<tr>
<td>total suspended solids</td>
<td>73.0</td>
<td>55.1</td>
<td>mg · L$^{-1}$</td>
</tr>
<tr>
<td>total volume</td>
<td>63.6</td>
<td>24.4</td>
<td>mL</td>
</tr>
<tr>
<td>turbidity</td>
<td>145</td>
<td>508</td>
<td>NTU</td>
</tr>
</tbody>
</table>
Table B.6: **OTUs per functional group.** Number of OTUs assigned to each functional group, compared to the total number of taxonomically annotated OTUs. Some OTUs were assigned to multiple functional groups.

<table>
<thead>
<tr>
<th>functional group</th>
<th>OTUs</th>
<th>fraction (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>aerobic chemoheterotrophs</td>
<td>225</td>
<td>11.4</td>
</tr>
<tr>
<td>cellulolytic</td>
<td>14</td>
<td>0.7</td>
</tr>
<tr>
<td>fermenters</td>
<td>103</td>
<td>5.2</td>
</tr>
<tr>
<td>methanogens</td>
<td>13</td>
<td>0.7</td>
</tr>
<tr>
<td>methylotrophs</td>
<td>24</td>
<td>1.2</td>
</tr>
<tr>
<td>nitrogen respirers</td>
<td>17</td>
<td>0.9</td>
</tr>
<tr>
<td>photoautotrophs</td>
<td>61</td>
<td>3.1</td>
</tr>
<tr>
<td>sulfate respirers</td>
<td>40</td>
<td>2.0</td>
</tr>
<tr>
<td>ureolytic</td>
<td>21</td>
<td>1.1</td>
</tr>
</tbody>
</table>
Table B.7: **KOG-function associations.** KEGG orthologous groups (KOG) associated with various functions in the metagenomic sequences.

<table>
<thead>
<tr>
<th>function</th>
<th>KOG</th>
</tr>
</thead>
<tbody>
<tr>
<td>fermentation</td>
<td>K01568, K13951, K00114, K00002, K04022, K00128, K00129, K00016, K00102, K00656, K00825, K00004, K00929, K00248, K00239, K00240, K00241, K00242</td>
</tr>
<tr>
<td>carbon fixation</td>
<td>K01595, K01601, K01602, K03737</td>
</tr>
<tr>
<td>cellulolysis</td>
<td>K19356, K19357, K01179, K01225, K19668</td>
</tr>
<tr>
<td>denitrification (norBC)</td>
<td>K04561, K02305</td>
</tr>
<tr>
<td>denitrification (nosZ)</td>
<td>K00376</td>
</tr>
<tr>
<td>dissimilatory sulfur metabolism</td>
<td>K11180, K11181, K00394, K00395, K17219, K17220, K17221, K16952, K17222, K17223, K17224, K17225, K17226, K17227</td>
</tr>
<tr>
<td>hydrogen oxidoreduction</td>
<td>K00532, K00533, K00534, K06441, K18016, K18017, K18023, K00436, K05586, K05587, K05588, K18005, K18006, K18007</td>
</tr>
</tbody>
</table>
Table B.7: Continued from previous page.

<table>
<thead>
<tr>
<th>function</th>
<th>KOG</th>
</tr>
</thead>
<tbody>
<tr>
<td>methanogenesis</td>
<td>K00399, K00401, K00402, K03421, K03422, K03388, K03389, K03390</td>
</tr>
<tr>
<td>methanol oxidation</td>
<td>K14028, K14029, K17066, K14028, K16254, K16255, K14029, K16256,</td>
</tr>
<tr>
<td></td>
<td>K16257, K16258, K16259, K16260, K00093, K17066</td>
</tr>
<tr>
<td>methylotrophy</td>
<td>K14028, K14029, K17066, K14028, K16254, K16255, K14029, K16256,</td>
</tr>
<tr>
<td></td>
<td>K16257, K16258, K16259, K16260, K00093, K17066, K10713, K14028, K14029, K17066, K16157, K16158, K16159, K16160, K16161, K16162</td>
</tr>
<tr>
<td>nitrate respiration</td>
<td>K02567, K02568, K00370, K00371, K00374, K00373</td>
</tr>
<tr>
<td>nitrite reduction to ammonium</td>
<td>K00362, K00363, K03385, K15876</td>
</tr>
<tr>
<td>nitrite respiration</td>
<td>K00368, K15864, K00362, K00363, K03385, K15876</td>
</tr>
<tr>
<td>nitrogen respiration</td>
<td>K04561, K02305, K00376, K00368, K15864, K02567, K02568, K00370, K00371, K00374, K00373, K00362, K00363, K03385, K15876</td>
</tr>
</tbody>
</table>
Table B.7: Continued from previous page.

<table>
<thead>
<tr>
<th>function</th>
<th>KOG</th>
</tr>
</thead>
<tbody>
<tr>
<td>oxygen respiration (cox)</td>
<td>K02277, K02276, K02274, K15408, K02275, K02262, K02256, K02261, K02263, K02264, K02265, K02266,</td>
</tr>
</tbody>
</table>
Appendix C

Chapter 4: Supplemental material

C.1 Methods

C.1.1 MCM overview

MCM is a mathematical and computational framework for the construction, simulation, statistical analysis and calibration of microbial community models (Fig 4.2). Below we give a brief overview of MCM’s mathematical structure and functionality. A thorough user manual, including mathematical details and several step-by-step examples, was published by Louca and Doebeli (284) and is also available online at: http://www.zoology.ubc.ca/MCM.

Mathematically, microbial community (MC) models in MCM correspond to a combination of differential equations and optimization problems. In the simplest case, a model considers the concentrations of \( S \) unicellular species, the concentrations of \( M \) chemical substances (metabolites) in a single extracellular metabolite pool, and the cell-specific rates of \( R \) biologically catalyzed reactions. The environment (i.e., the medium containing the cells and extracellular metabolites) is generally assumed to be well mixed, although compartmentalized ecosystem models are also possible (see the user manual).

Each species is characterized by its metabolic potential, that is, the subset of reactions that it can catalyze as well as any metabolites that it is able to uptake or export. The rate at which each cell performs a specific reaction in a specific moment depends on the species as well as on the current extracellular metabolite concentrations. At all times, intracellular metabolite fluxes are assumed to be balanced, so that the intracellular reaction rates completely determine the rates at which metabolites are exported or taken up by cells. For each particular reaction performed by a particular species, a model may specify limits regarding the forward and/or backward reaction rate. Similarly, for each metabolite utilized or produced by a particular species, a model may specify limits regarding its uptake and/or export rate. This constraint-based metabolic modeling is also known as Flux Balance Analysis (FBA; 354).
Each reaction can contribute to a cell’s biosynthesis rate (e.g., if the reaction produces a specific amino acid), and cells are assumed to optimize their biosynthesis rate by appropriate choice of their reaction rates, within the constraints imposed by FBA. Mathematically, if $F_s$ is the vector listing the species’ cell-specific net metabolite uptake rates, $N_s$ is the species’ cell concentration, $C$ is the vector listing the extracellular concentrations of all metabolites and $f$ is the vector listing any additional net metabolite fluxes into the environment (e.g., external nutrient supply), then $C$ changes according to the differential equation

$$
\frac{dC}{dt} = - \sum_{s=1}^{S} N_s F_s + f. \quad (C.1.1)
$$

The sum in Eq. (C.1.1) iterates over all species and represents the net metabolite uptake by the entire microbial community. Since intracellular metabolite fluxes are balanced, $F_s$ is given by

$$
F_s = -S \cdot H_s, \quad (C.1.2)
$$

where $H_s$ is the vector listing the cell-specific rates of all reactions for species $s$ and $S$ is the stoichiometric matrix of all reactions. Specifically, each entry $S_{mr}$ is the stoichiometric coefficient for metabolite $m$ in reaction $r$. For example, the reaction

$$
C_6H_{12}O_6 + 6O_2 + 30ADP \rightarrow 6CO_2 + 6H_2O + 30ATP \quad (C.1.3)
$$

has stoichiometric coefficients $-1$, $-6$, $-30$, $+6$, $+6$ and $+30$ for the compounds $C_6H_{12}O_6$, $O_2$, $ADP$, $CO_2$, $H_2O$ and $ATP$, respectively. Note that some entries in $H_s$ may be constant zero, for example if the species lacks the capacity to perform certain reactions.

At any moment, $H_s$ is the solution to a linear optimization problem that maximizes the biosynthesis rate (which is a linear function of $H_s$), given the constraints imposed on $H_s$ as well as $F_s$. These constraints can depend on the metabolite concentrations $C$, and hence $H_s$ and $F_s$ depend on $C$. For example, the maximum possible $O_2$ uptake rate may be a Michaelis-Menten-type function of environmental $O_2$ concentrations (211), and hence the optimal rates of aerobic pathways may depend on environmental $O_2$ concentrations. Cell-specific biosynthesis rates are translated to per-capita cell birth rates by dividing by the cell mass, $\mu_s$. On the other hand, cell loss is described by an exponential decay rate that accounts, e.g., for cell death or dilution. Hence, the cell concentration $N_s$ changes according
to the differential equation

$$\frac{dN_s}{dt} = -\frac{N_s}{\tau_s} + \frac{N_s}{\mu_s} \cdot B(H_s(C)),$$

where $\tau_s$ is the expected life time and $B_s$ is the cell-specific biosynthesis rate as a function of $H_s$. We mention that in more general models, $H_s$, $f$ and $\tau_s$ may also depend on arbitrary environmental variables, $E$, which themselves can be additional dynamical model variables or be explicitly specified as part of the model (details in the user manual). For example, reaction rates may be limited by temperature (49) and temperature may be an explicitly controlled environmental variable.

In MCM, models are specified in plain-text configuration files that define all metabolites, reactions, cell species as well as environmental variables. MCM translates these models into differential equations and linear optimization problems and solves them numerically. MCM is controlled through special scripts, which may contain commands for running simulations, fitting parameters or simply modifying technical parameters. MCM includes tools for the conversion of conventional genome-scale FBA models, such as generated by the Model SEED pipeline (184) based on sequenced genomes, into a draft MC model.

MCM can accommodate microbial communities comprising genome-based cell models with arbitrary environmental variables, metabolite exchange kinetics and regulatory mechanisms. For example, environmental variables may be stochastic processes (e.g., representing climate), or specified using measured data (e.g., redox potential in bioreactor experiments), or depend on metabolite concentrations (e.g., pH determined by acetate concentration) or even be dynamical (e.g., temperature increasing at a rate proportional to biomass production rates). This versatility allows for the incorporation of complex environmental feedbacks, such as host immune responses in gut microbiota (223). Metabolite uptake and export rate limits can be arbitrary functions of metabolite concentrations or environmental variables. Similar interdependencies are possible for reaction rate limits, thus allowing the inclusion of inhibitory or regulatory mechanisms (84). Metabolite concentrations can be explicitly specified, e.g., using measured time series, or depend dynamically on microbial export and other external fluxes. Effects of phage predation (204), reaction energetics (386) or stochastic environments can also be incorporated.

MCM keeps track of a multitude of output variables such as cell densities, reaction rates, metabolite concentrations and metabolite exchange rates. Because each reaction can be formally associated with a particular enzyme, in turn encoded by a particular gene, MCM also
makes predictions about gene densities as a product of cell densities and gene copy numbers per cell. Metabolic activity statistics (e.g., Fig 4.6A,B) facilitate the identification of metabolic interactions such as cross-feeding (333). The predicted time courses of output variables can be statistically evaluated against time series ranging from chemical concentrations, rate measurements to cell densities and metagenomics.

MC models can include arbitrary abstract (symbolic) numeric parameters with a predefined value range or probability distribution. Symbolic parameters can represent, for example, stoichiometric coefficients, gene copy numbers, cell life expectancies, half-saturation constants or environmental variables. The inclusion of symbolic parameters enables a high-level analysis of microbial communities: For example, MCM can automatically calibrate (fit) unknown symbolic parameters to time series using maximum-likelihood parameter estimation (113). The likelihood of the data, given a particular parameter choice, is calculated by assuming a mixed deterministic-stochastic model in which the deterministic part is given by the model predictions, and the stochastic part is given by normally distributed errors. The likelihood is minimized using an iterative optimization algorithm involving step-wise parameter adjustments and repeated simulations. Other fitting algorithms are also available, such as maximization of the average coefficient of determination ($R^2$), which is equivalent to weighted least-squares fitting. Because MCM can calibrate unknown measurement units, raw uncalibrated data (e.g., optical cell densities with no calibration to colony forming units, Fig 4.4A) can also be used.

In this chapter single-cell models were calibrated to monoculture experiments, however models can also be calibrated using data from experimental or natural communities that include uncultured species. In general, fitted parameters need not be directly connected to the data used for calibration, as long as a change in the parameters influences the predictions that are being compared to the data. While this is a general principle of parameter estimation (468), in practice the uncertainty of calibrated parameters (e.g., in terms of confidence intervals) increases when their influence on the “goodness of fit” is weaker. Moreover, alternative parameter combinations can sometimes yield a comparable match to the data, especially if multiple parameters influence the same variables (inverse problem degeneracy). Local fitting optima can be detected through repeated randomly seeded calibrations (see next section), and overfitting can be partially avoided by keeping the number of free parameters at a bare minimum. Nevertheless, in certain cases good knowledge of the system or previous literature may be required to identify the most plausible calibrations. Finally, we emphasize that MCM is, after all, merely a framework enabling the construction, calibration and analysis of microbial community models. MCM models are thus limited by the same caveats and
assumptions as other constraint-based metabolic models (17, 38) and any predictions made by MCM should be subject to similar scrutiny.

C.1.2 Calibration of *E. coli* cell models

*E. coli* strains were obtained from an evolution experiment performed in a batch culture environment with daily dilutions into glucose-acetate supplemented Davis minimal medium (437, 488). For each phenotype, three clones were isolated from population 20 after 150 days and used for three independent monoculture growth experiments. Optical densities, as well as glucose, acetate and oxygen concentration data from these experiments were used to calibrate the individual cell-metabolic models for the A, SS and FS phenotypes. Oxygen measurements were not available for type A. Experimental details and results are described by Le Gac et al. (262).

In the models, the limiting nutrients are assumed to be oxygen, glucose and acetate; all other nutrients can be taken up at an arbitrary rate. Oxygen, glucose and acetate uptake rate limits were described by Monod-like kinetics. The maximum cell-specific oxygen uptake rate was set to $1.008 \times 10^{-13} \text{ mol/(d \cdot cell)}$, according to Varma and Palsson (496). The oxygen half-saturation constant was set to $1.21 \times 10^{-7} \text{ M}$ according to Stolper et al. (452). Oxygen was assumed to be initially at atmospheric saturation levels ($0.217 \text{ mM}$ at $37^\circ \text{ C}$) and repleted at a rate proportional to its deviation from saturation (167).

The fitted parameters for each cell type were the maximum cell-specific uptake rates and half-saturation constants for glucose and acetate, as well as initial cell densities and non-growth associated ATP maintenance energy requirements. The initial glucose and acetate concentrations were set to the average value measured at the earliest sampling point (1 hr after incubation) for each type. The oxygen mass transfer coefficient (M/day per M deviation) was initially fitted individually for each type together with all other parameters, and then fixed to the average of all three initial fits. All other parameters were then again fitted individually for each type. Parameter fitting was done by maximizing the average coefficient of determination ($R^2$) using the MCM command `fitMCM`. A total of 237 data points were used to fit 19 parameters (Supplemental Table C.1). To reduce the possibility of only reaching a local maximum, fitting was repeated 100 times for each strain starting at random initial parameter values and the best fit among all 100 runs was used. While some fitting runs reached alternative local maxima, the best overall fit was reached in most cases.

Cell densities were directly compared to optical density (OD) measurements. The appropri-
ate calibrations were estimated by MCM and ranged within \(8.2 \times 10^{11} - 1.3 \times 10^{12} \text{ cells/(L \cdot OD)}\). These estimates are consistent with previous experimental calibrations (260) that yielded 0.26 g dW/(L \cdot OD), i.e., \(1.4 \times 10^{12} \text{ cells/(L \cdot OD)}\) (assuming a cell dry weight of \(1.8 \times 10^{-13} \text{ g}\) in the stationary phase; 115).

### C.1.3 Simulation of the microbial community model

The microbial community model was simulated using the MCM command `runMCM`. Initial glucose and acetate concentrations were set to the average of all values measured at the earliest sampling point of the monoculture incubations. Cell death was not explicitly included, because of lack of appropriate data for calibration and because daily dilutions by far exceeded cell death as a factor of cell population reduction. The MCM files required to run this model were published by Louca and Doebeli (284) and are also available online at: [http://www.zoology.ubc.ca/MCM](http://www.zoology.ubc.ca/MCM).

### C.1.4 Robustness of the SS-FS coexistence

To verify the robustness of the stable SS-FS coexistence in co-culture, we randomly varied each fitted model parameter uniformly within an interval spanning 10% above and 10% below its calibrated value. Both types coexisted in 50 out of 50 random simulations (Supplemental Fig. C.2).

### C.1.5 Seasonal restriction of the SS-FS co-cultures

Simulations of the SS-FS co-cultures restricted to the first glucose-rich or second glucose-depleted season, as opposed to the full batch cycle, were performed in analogy to the experiments by Spencer et al. (436). More precisely, to model the first season experiment we changed the dilution rate to \(1/32\) every 5 hours, so that at the end of each batch cycle glucose was not yet completely depleted. Similarly, for the second season experiment we changed the dilution rate to \(1/32\) every 19 hours, and adjusted the growth medium to resemble the glucose-depleted acetate-rich solution reported by Spencer et al. (no glucose, 3.59 mM acetate). Initial cell densities were set to \(1 \times 10^{10} \text{ cells/L}\) for both types. All other model parameters were kept unchanged. The original experiments by Spencer et al. (436) were performed at higher dilution rates (4 and 15 hours for the first and second season experiment, respectively), however in our simulations neither the FS nor SS type could persist at these high dilution rates. We note that the strains used in our work (262) had evolved
in separate evolution experiments using a different growth medium than those by Spencer et al. (436).

Table C.1: Fitted parameters for the *E. coli* models described in the main text, together with reference values from the literature for comparison. Maximum cell-specific uptake rates ($V_{\text{max}}$) are in fmol/cell/d. Half-saturation constants for acetate ($K_{\text{half,acetate}}$) are in mM, half-saturation constants for glucose ($K_{\text{half,glucose}}$) are in $\mu$M. Initial cell densities are in $10^9$ cells/L. Non-growth associated maintenance requirements are given in fmol ATP/cell/d. The $O_2$ mass transfer coefficient is in 1/d (reference value only roughly comparable, as the transfer coefficient depends strongly on shaking frequency and flask volume (293)). Dry-weight-specific values from the literature were converted to cell-specific values by assuming a dry weight of 180 fg/cell (115). All reference values were measured for strains other than B REL606.

<table>
<thead>
<tr>
<th>parameter</th>
<th>values</th>
<th>comparison</th>
<th>reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{\text{max,acetate}}$</td>
<td>67.8 (A), 16.5 (SS), 220 (FS)</td>
<td>8.6</td>
<td>(318)</td>
</tr>
<tr>
<td>$K_{\text{half,acetate}}$</td>
<td>10.6 (A), 5.55 (SS), 12.9 (FS)</td>
<td>6.0</td>
<td>(318)</td>
</tr>
<tr>
<td>$V_{\text{max,glucose}}$</td>
<td>43.2 (A), 56.9 (SS), 29.0 (FS)</td>
<td>45</td>
<td>(496)</td>
</tr>
<tr>
<td>$K_{\text{half,glucose}}$</td>
<td>21.3 (A), 11.4 (SS), 44.6 (FS)</td>
<td>3–15</td>
<td>(158)</td>
</tr>
<tr>
<td>maintenance req.</td>
<td>18.6 (A), 11.0 (SS), 15.0 (FS)</td>
<td>32</td>
<td>(496)</td>
</tr>
<tr>
<td>$O_2$ mass transfer</td>
<td>60.9</td>
<td>180</td>
<td>(292)</td>
</tr>
<tr>
<td>init. cell density</td>
<td>8.48 (A), 11.3 (SS), 7.57 (FS)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure C.1: Predicted relative cell densities of the A and FS types in co-culture, in the absence of SS. Initial cell densities were $10^{10}$ cells/L for type A and 1 cell/L for type FS. All other model parameters are identical to the microbial community model (comprising the A, SS and FS types) described in the main text.

Figure C.2: Robustness of the predicted stable coexistence of the SS and FS types in co-culture. Shown are the probability distributions of the relative SS (A) and FS (B) cell densities over time, when calibrated model parameters are randomly chosen within an interval spanning 10% above and 10% below their fitted values. Initial cell densities were $10^{10}$ cells/L for both types, all other parameters were as described in the main text. Probability distributions were estimated using 50 Monte Carlo simulations. In all cases both the SS and FS type persisted. The analysis was performed using the MCM command UAMCM.
Figure C.3: Measured relative cell densities of the SS and FS types in batch co-culture, when restricted to either the first glucose-rich (left column) or second glucose-depleted (right column) season for three independently evolved communities (rows 1–3), as reported by Spencer et al. (436, Figs. 2A,B therein). Restriction to the first season was achieved by shorter dilution periods which prevented the complete depletion of glucose. In (B), restriction to the second season was achieved by using the glucose-depleted acetate-rich solution, produced by the full-batch co-culture, as growth medium. Initial population sizes different between experiments. Strains used by Spencer et al. (436) evolved in slightly different growth medium than in this chapter. Cell generations were translated to days by assuming an average of 6.7 generations per day (188).
Appendix D

Chapter 5: Supplemental material

D.1 Methods

D.1.1 Computational framework

Model calibration, simulations and statistical analysis were performed using MCM (Chapter 4; 284). MCM combines FBA-based cell models with a dynamical environment that influences, and is influenced by, microbial metabolism. The combination of FBA with a varying environmental metabolite pool is known as dynamic flux balance analysis (DFBA) (70, 177, 292), and has been shown to be a promising approach to microbial ecological modeling (70, 177, 284, 318). MCM can accommodate microbial community models with arbitrary environmental variables and metabolite exchange kinetics. For example, environmental variables may be stochastic processes (e.g., representing climate fluctuations) or specified using measured data (e.g., pH in bioreactor experiments). Metabolite uptake and export rate limits can be arbitrary functions of metabolite concentrations or environmental variables. Similar interdependencies are possible for reaction rate limits, thus allowing the inclusion of inhibitory or regulatory mechanisms (84). Metabolite concentrations can be explicitly specified (e.g., using measured time series) or depend dynamically on microbial export and other external fluxes.

MCM keeps track of a multitude of output variables such as cell concentrations, reaction rates, metabolite concentrations and metabolite exchange rates. Model predictions can then be compared to time series from experiments or environmental surveys, such as rate measurements, chemical profiles or optical cell densities. Reciprocally, time series data can be used to automatically calibrate unknown model parameters, e.g., using least squares fitting or maximum-likelihood estimation (e.g., Fig 5.1, see details below). Because MCM can calibrate unknown measurement units, raw uncalibrated data (e.g., optical cell densities with no calibration to colony forming units) can also be used. MCM was recently validated using laboratory experiments with bacterial communities (284). MCM is Open Source and
D.1.2 Construction of the cell models

The metabolism of each cell was modeled using flux balance analysis with optimization of biomass synthesis (354). The cell-internal reaction networks of the ammonium oxidizing bacteria (AOB) and nitrite oxidizing bacteria (NOB) are based on the core metabolic models published by Poughon et al. (370) and Perez-Garcia et al. (364). More precisely, the biomass synthesis functions of both cell types are taken from Perez-Garcia et al. (364), the energy metabolism of AOB is adopted from Perez-Garcia et al. (364) and the energy metabolism of NOB is adopted from (370). Assimilatory nitrite reduction to ammonium, required for biomass synthesis, was added to NOB (443). The constructed AOB and NOB models are comprised of 16 and 11 reactions, respectively (see Appendix D.3 for details). The nitrogen substrate half-saturation constants were set to 26 $\mu$M NH$_3$ for the AOB template according to Suzuki et al. (463), and to 229 $\mu$M NO$_2^-$ for the NOB template according to Remacle and De Leval (389). Cell masses were set to $3 \times 10^{-13}$ g dW/cell for the AOB and $4 \times 10^{-13}$ g dW/cell for the NOB, according to Keen and Prosser (227).

D.1.3 Calibration of the template cell models

The maximum cell-specific substrate uptake rates for the AOB and NOB templates ($V_{\text{max,NH}_4^+}$ and $V_{\text{max,NO}_2^-}$, respectively) were calibrated using time series from a previous experiment with an ammonium-batch-fed nitrifying bioreactor by de Boer and Laanbroek (94), inoculated with strains of the AOB *Nitrosospira* and NOB *Nitrobacter* genera. For the calibration, our bioreactor model was adjusted to de Boer and Laanbroek’s experiment (Fig. 5.1A): The initial ammonium concentration was set to 0.916 mM, nitrite and nitrate were initially absent, the pH was set to the reported profile and oxygen was assumed to be non-limiting.

We used the reported concentration profiles of the gradually depleted ammonium (Fig. 5.1B) and produced nitrate (Fig. 5.1C) to calibrate $V_{\text{max,NH}_4^+}$ and $V_{\text{max,NO}_2^-}$ via maximum-likelihood estimation (113). This approach estimates unknown parameters by maximizing the likelihood of observing the available data, given a particular candidate choice of parameter values. Maximum likelihood estimation is widely used in statistical inference such as multilinear regression, computational phylogenetics and modeling in physics (291). In our case, the likelihood of the data was calculated on the basis of a mixed deterministic-stochastic structure, in which the deterministic part is given by the microbial community model and errors available at http://www.zoology.ubc.ca/MCM.
are assumed to be normally distributed. The likelihood was maximized using the SBPLX optimization algorithm (214), which uses repeated simulations and gradual exploration of parameter space and is integrated into MCM (284). To reduce the possibility of only reaching a local maximum, fitting was repeated 100 times using random initial parameter values and the best fit among all 100 runs was used. While some fitting runs reached alternative local maxima, the best overall fit was reached in most cases. This procedure yielded the fitted values $V_{\text{max}, \text{NH}_3} = 6.48 \times 10^{-13} \text{ mol NH}_3/\text{cell/d}$ and $V_{\text{max}, \text{NO}_2^-} = 7.31 \times 10^{-13} \text{ mol NO}_2^-/\text{cell/d}$, which are consistent with the literature (375).

### D.1.4 Nitrifying membrane bioreactor model

Using the calibrated AOB and NOB template cell models, we constructed a model of an ammonium-fed nitrifying membrane bioreactor similar to the one described by Wittebolle et al. (524). The hydraulic turnover rate for all metabolites was 0.672 d$^{-1}$, ammonium input was 7.14 mM·d$^{-1}$ and pH was fixed to 7.4. The input medium was assumed to be sterile and to contain micronutrients in sufficient amounts for autotrophic growth via nitrification (109, 524). The bioreactor medium was assumed to be well mixed. Microbial communities started with an equal number of AOB and NOB strains, each with an initial density of $10^7$ cells/L. Cell death was modeled as exponential decay. Further details are provided in Appendix D.3.

Each strain was a random variation of the calibrated template cell models, with physiological parameters chosen as follows: Substrate uptake kinetic parameters, i.e., the maximum cell-specific nitrogen substrate uptake rates ($V_{\text{max}}$) and substrate affinities ($\alpha$; 6), were randomly and uniformly chosen within an interval ranging an order of magnitude above and an order of magnitude below the template values. To account for the typically assumed tradeoff between $V_{\text{max}}$ and $\alpha$, these parameters were multiplied by a factor $\kappa$ and $(1 - \kappa)$, respectively, where $\kappa$ was chosen randomly within [0, 1] for each strain (431). Cell life times were randomly chosen within $50 - 100$ d for each strain according to typical nitrifier decay rates (7).

Perturbations were modeled as a temporary increase in mortality rates, such that after one day each cell population declines by some random factor, chosen log-uniformly and independently for each strain within the interval [1, $10^{12}$].
D.1.5 Statistics of community convergence

The distance between two community compositions was expressed using the Bray-Curtis dissimilarity, which is well established in the ecological literature (266). The maximum dissimilarity between any two communities is 100\%, while identical communities have a dissimilarity of 0\%. The convergence of the bioreactor community was examined by calculating its dissimilarity to the steady composition established after a long time. This dissimilarity curve is typically decreasing in time because communities eventually converge to a steady composition in which each metabolic niche is occupied by a single strain. A steeper curve implies a faster convergence. Following inoculation, the dissimilarity curve depends on the particular strains present in the community, which are chosen randomly for each simulation. The resulting probability distribution of the dissimilarity curves (Figs. 5.2E,F, and 5.3E,F) was estimated using 100 repeated random simulations of the model.

D.2 Elaboration on the competition model

Below we elaborate on the competition model in the main text. We consider the total cell density \( N = \sum_i N_i \) and the relative cell densities \( \eta_i = N_i / N \). Using the community-average quantities

\[
\bar{\beta}\Phi = \sum_i \eta_i \beta_i \Phi_i, \quad \bar{\lambda} = \sum_i \eta_i \lambda_i, \quad \bar{\Phi} = \sum_i \eta_i \Phi_i
\]  

(D.2.1)

it is straightforward to derive the dynamics

\[
\frac{dR}{dt} = f_o - N\Phi(R)
\]  

(D.2.2)

for the resource concentration. Similarly, starting with

\[
\frac{dN}{dt} = \sum_i \frac{dN_i}{dt}
\]  

(D.2.3)

and using the original model equations for the \( N_i \) one quickly arrives at

\[
\frac{dN}{dt} = N \left( \bar{\beta}\Phi(R) - \bar{\lambda} \right).
\]  

(D.2.4)
Furthermore, by using the product rule
\[ \frac{d\eta_i}{dt} = \frac{1}{N} \frac{dN_i}{dt} - \frac{N_i}{N^2} \frac{dN}{dt} \]  
(D.2.5)
and inserting Eq. (D.2.4) one finds that
\[ \frac{d\eta_i}{dt} = \eta_i \left[ (\beta_i \Phi_i - \lambda_i) - (\overline{\beta \Phi} - \overline{\lambda}) \right]. \]  
(D.2.6)

Note that the dynamics of \( N \) and \( R \) are determined by the community-average growth kinetics. In contrast, relative cell densities change at rates that depend on the deviation of individual growth kinetics from the community average. Furthermore, one can always write
\[ \beta_i \Phi_i - \lambda_i = (\overline{\beta \Phi} - \overline{\lambda}) (1 + \varepsilon_i), \]  
(D.2.7)
where \( \varepsilon_i \) accounts for the relative deviation of individual growth kinetics from the community average. Then Eq. (D.2.6) becomes
\[ \frac{d\eta_i}{dt} = \varepsilon_i \cdot \eta_i (\overline{\beta \Phi} - \overline{\lambda}), \]  
(D.2.8)
as given in the main text.

D.3 Details on the bioreactor model

D.3.1 Construction of cell models

The reaction networks for the two cell types, AOB and NOB, only include core energy metabolism and biomass synthesis (production of ATP and NADH via nitrification and biosynthesis via consumption of ATP and NADH). The biomass synthesis functions of both cell types are taken from (364), assuming that biomass stoichiometry is similar for *Nitrosomonas* (AOB) and *Nitrobacter* (NOB) cells. In particular, energy-to-biomass conversion coefficients (ATP & NADH to g dry weight) were experimentally calibrated by (364). The energy metabolism of AOB is taken from (364). Note that the original published AOB model had 17 reactions related to energy metabolism (Table S1 in (364)). One of these reactions, “NO1”, was merely a trivial conversion reaction and was thus merged with the rest. The remaining reactions in the AOB model by (364) were pure transport reactions and are implicitly included in the MCM model (but not referred to as “reactions”). The energy
metabolism of NOB is taken from (370). Assimilatory nitrite reduction to ammonium, required for biomass synthesis but not included in the original reaction network by (370), was added to NOB according to (443). Uptake kinetic parameters for NH₃,NH₄ and HNO₂,NO₂, as well as cell life times and cell masses were chosen as described in the main text. Below we provide an overview of all considered metabolites and reactions. Please consult the references by the reactions for details. The complete model and simulation script are available at: http://www.zoology.ubc.ca/MCM

D.3.2 Metabolites

- ADP (adenosine diphosphate)
- ATP (adenosine triphosphate)
- CO₂ (carbon dioxide)
- Cytc₅₅₀ox (class I cytochrome c₅₅₀, oxidized)
- Cytc₅₅₀red (class I cytochrome c₅₅₀, reduced)
- Cytc₅₅₂ox (class I cytochrome c₅₅₂, oxidized)
- Cytc₅₅₂red (class I cytochrome c₅₅₂, reduced)
- Cytc₅₅₄ox (class I cytochrome c₅₅₄, oxidized)
- Cytc₅₅₄red (class I cytochrome c₅₅₄, reduced)
- H_c (hydrogen, cytosol)
- H_p (hydrogen, periplasm)
- H₂O (water)
- NH₃,NH₄ (ammonia + ammonium)
- NH₃ (purely extracellular, ammonia, depending on NH₃,NH₄ and pH, see notes below)
- NH₄ (purely extracellular, ammonium, depending on NH₃,NH₄ and pH, see notes below)
- HNO₂,NO₂ (nitrite + nitrous acid)
- NO₂ (purely extracellular, nitrite, depending on HNO₂,NO₂ and pH)
- $\text{HNO}_3\text{NO}_3$ (nitrate + nitric acid)
- $\text{NO}_3^-$ (purely extracellular, nitrate, depending on $\text{HNO}_3\text{NO}_3$ and pH)
- $\text{N}_2\text{O}$ (nitrous oxide)
- $\text{NAD}$ (oxidized nicotinamide adenine dinucleotide)
- $\text{NADH}$ (Reduced nicotinamide adenine dinucleotide)
- $\text{NH}_2\text{OH}$ (hydroxylamine)
- $\text{NO}$ (nitric oxide)
- $\text{NOH}$ (nitroxyle)
- $\text{O}_2$ (oxygen)
- $\text{Pi}$ (inorganic phosphate)
- protein
- $\text{UQ}$ (ubiquinone)
- $\text{UQH}_2$ (ubiquinol)
- maint (formal maintenance requirements, in ATP-equivalents)
- $\text{Q}_8\text{H}_2$ (ubiquinol-8)
- $\text{Q}_8$ (ubiquinone-8)
- $\text{Cyt}_{554}\text{e}$ (ferrocytochrome c554)
- $\text{Cyt}_{554}$ (ferricytochrome c554)
- $\text{Cyt}_{552}\text{m}\text{e}$ (membrane ferrocytochrome)
- $\text{Cyt}_{552}\text{m}$ (ferricytochrome c552)
- $\text{Cyt}_{552}\text{e}$ (periplasmic ferrocytochrome c552)
- $\text{Cyt}_{552}$ (ferricytochrome c552)
Note that in the model ammonium (NH$_4^+$) is assumed to be at dissociation equilibrium with ammonia (NH$_3$), determined by the pH and the acid dissociation constant $5.69 \times 10^{-10}$ M at standard temperature (73). pH was either adjusted to measurements for the batch reactor (94), or to constant 7.4 for the membrane continuous-flow reactor (524). The total concentration of ammonium and ammonia is represented in the model by NH$_3$.NH$_4^+$ and depends on the rate of ammonium input, the hydraulic dilution rate and microbial consumption. A similar formalism was applied to nitrite and nitrate.

### D.3.3 Reaction network for AOB

R$_\text{biomass}$ (364) (biomass coefficient 113 g dW/mol, modified to include CO$_2$ consumption):

$$
15 \text{ ATP} + 12 \text{ NADH} + 0.25 \text{ protein} + 32 \text{ maint} + 5 \text{ CO}_2 + 7 \text{ H}_c \\
\rightarrow 15 \text{ ADP} + 10 \text{ NAD} + 15 \text{ Pi} + 4 \text{ O}_2 \tag{D.3.1}
$$

R$_\text{maint}$ (364) (maintenance ATP consumption):

$$
\text{ATP} + \text{H}_2\text{O} \rightarrow \text{ADP} + \text{Pi} + \text{H}_c + \text{maint} \tag{D.3.2}
$$

R$_\text{protein}$ (364) (protein synthesis via ATP consumption):

$$
8.9 \text{ ATP} + 4 \text{ NH}_3$.NH$_4$ $\rightarrow$ 8.9 ADP + 8.9 Pi + 8.9 H$_c$ + protein \tag{D.3.3}
$$

R$_\text{amo}$ (364):

$$
\text{NH}_3$.NH$_4$ + O$_2$ + $Q_8$H$_2$ $\rightarrow$ NH$_2$OH + H$_2$O + $Q_8$ \tag{D.3.4}
$$

R$_\text{HAO\_NOH}$ (364):

$$
\text{NH}_2\text{OH} + \text{Cyt}_{554} \rightarrow \text{NOH} + \text{Cyt}_{554e} + 2 \text{ H}_p \tag{D.3.5}
$$

R$_\text{HAO\_NO}$ (364):

$$
\text{NOH} + 0.5 \text{ Cyt}_{554} \rightarrow \text{NO} + 0.5 \text{ Cyt}_{554e} + \text{H}_p \tag{D.3.6}
$$
R:\_HAO\_hno2 (364):

\[
\text{NO} + 0.5 \text{Cyt}_{554} + \text{H}_2\text{O} \rightarrow \text{HNO}_2\cdot\text{NO}_2 + 0.5 \text{Cyt}_{554e} + \text{H}_p \quad \text{(D.3.7)}
\]

R:\_Cyt\_554 (364):

\[
\text{Cyt}_{554e} + \text{Cyt}_{552m} \rightarrow \text{Cyt}_{552me} + \text{Cyt}_{554} \quad \text{(D.3.8)}
\]

R:\_Q\_8H\_2\_synt (364):

\[
\text{Q}_8 + \text{Cyt}_{552me} + 2 \text{H}_p \rightarrow \text{Q}_8\text{H}_2 + \text{Cyt}_{552m} \quad \text{(D.3.9)}
\]

R:\_NADH\_synt (364):

\[
\text{NAD} + \text{Q}_8\text{H}_2 + 2 \text{H}_p \rightarrow \text{NADH} + \text{Q}_8 + 3 \text{H}_c \quad \text{(D.3.10)}
\]

R:\_Cytbc1 (364):

\[
\text{Q}_8\text{H}_2 + 2 \text{Cyt}_{552} \rightarrow 2 \text{H}_p + \text{Q}_8 + 2 \text{Cyt}_{552e} \quad \text{(D.3.11)}
\]

R:\_Cytaa3 (364):

\[
0.5 \text{O}_2 + 4 \text{H}_c + 2 \text{Cyt}_{552e} \rightarrow \text{H}_2\text{O} + 2 \text{H}_p + 2 \text{Cyt}_{552} \quad \text{(D.3.12)}
\]

R:\_CytP460 (364):

\[
0.5 \text{NH}_2\text{OH} + 0.5 \text{NO} + 2.5 \text{Cyt}_{552} + \text{H}_2\text{O}
\rightarrow \text{HNO}_2\cdot\text{NO}_2 + 2.5 \text{Cyt}_{552e} + 2.5 \text{H}_p \quad \text{(D.3.13)}
\]

R:\_nir (364):

\[
\text{HNO}_2\cdot\text{NO}_2 + \text{Cyt}_{552e} + \text{H}_p \rightarrow \text{NO} + \text{Cyt}_{552} + \text{H}_2\text{O} \quad \text{(D.3.14)}
\]

R:\_nor (364):

\[
\text{NO} + \text{Cyt}_{552e} + \text{H}_p \rightarrow 0.5 \text{N}_2\text{O} + \text{Cyt}_{552} + 0.5 \text{H}_2\text{O} \quad \text{(D.3.15)}
\]
R_ATP_synt (364):

\[ ADP + Pi + 3 H_p \rightarrow ATP + H_2O + 3 H_c \]  \hspace{1cm} (D.3.16)

### D.3.4 Reaction network for NOB

**R_biomass (364) (biomass coefficient 113 g dW/mol):**

\[ 15 \text{ATP} + 12 \text{NADH} + 0.25 \text{protein} + 32 \text{maint} + 5 \text{CO}_2 + 7 \text{H}_c \]
\[ \rightarrow 15 \text{ADP} + 10 \text{NAD} + 15 \text{Pi} + 4 \text{O}_2 \]  \hspace{1cm} (D.3.17)

**R_maint (364):**

\[ \text{ATP} + \text{H}_2\text{O} \rightarrow \text{ADP} + \text{Pi} + \text{H}_c + \text{maint} \]  \hspace{1cm} (D.3.18)

**R_protein (364):**

\[ 8.9 \text{ATP} + 4 \text{NH}_3\text{.NH}_4 \rightarrow 8.9 \text{ADP} + 8.9 \text{Pi} + 8.9 \text{H}_c + \text{protein} \]  \hspace{1cm} (D.3.19)

**Jnrj8 (370):**

\[ 2 \text{Cytc}_{550red} + \text{HNO}_2\text{.NO}_2 + \text{H}_2\text{O} \rightarrow 2 \text{Cytc}_{550ox} + 2 \text{H}_c + \text{HNO}_3\text{.NO}_3 \]  \hspace{1cm} (D.3.20)

**Jnrj9 (370):**

\[ \text{H}_p + \text{Cytc}_{550ox} + \text{HNO}_2\text{.NO}_2 \rightarrow \text{Cytc}_{550red} + \text{H}_2\text{O} + \text{NO} \]  \hspace{1cm} (D.3.21)

**Jnrj10 (370):**

\[ 4 \text{H}_p + \text{UQ} + 2 \text{H}_2\text{O} + 2 \text{NO} \rightarrow \text{UQH}_2 + 4 \text{H}_p + 2 \text{HNO}_2\text{.NO}_2 \]  \hspace{1cm} (D.3.22)

**Jnrj11 (370):**

\[ \text{Cytc}_{550red} + \text{H}_2\text{O} + \text{NO} \rightarrow \text{Cytc}_{550ox} + \text{H}_c + \text{HNO}_2\text{.NO}_2 \]  \hspace{1cm} (D.3.23)
Jtermox_dissip (370) (dissipation (loss) of proton motive force (PMF) by proton diffusion):

\[
2 \text{Cytc}_{550\text{ox}} + 4 \text{H}_c + 0.5 \text{O}_2 + 1.12 \text{H}_p \rightarrow 2 \text{H}_p + 2\text{Cytc}_{550\text{red}} + \text{H}_2\text{O} + 1.12 \text{H}_c \quad (D.3.24)
\]

JNAD (370):

\[
4 \text{H}_p + \text{NAD} + \text{UQH}_2 \rightarrow \text{UQ} + \text{NADH} + 5 \text{H}_c \quad (D.3.25)
\]

JATP (370):

\[
3 \text{H}_p + \text{ADP} + \text{Pi} \rightarrow \text{ATP} + 3 \text{H}_c + \text{H}_2\text{O} \quad (D.3.26)
\]

R_nirBD (443) (nitrite reduction to ammonium, accounting for nitrogen assimilation, nitrite detoxification, and NAD regeneration):

\[
\text{HNO}_2 . \text{NO}_2 + 3 \text{NADH} + 4 \text{H}_c \rightarrow \text{NH}_3 . \text{NH}_4 + 3 \text{NAD} + 2 \text{H}_2\text{O} \quad (D.3.27)
\]

### D.3.5 Uptake kinetics

- Ammonium/ammonia (NH$_3$.NH$_4$) uptake rate limits are specified as a Monod function of NH$_3$:

\[
\frac{[\text{NH}_3]}{\frac{1}{\alpha} + \frac{[\text{NH}_3]}{V}}, \quad (D.3.28)
\]

where \(V = 6.48 \times 10^{-13} \text{ mol/cell/day} \) (maximum cell-specific uptake rate) and \(\alpha = 2.49 \times 10^{-8} \text{ L/cell/day} \) (affinity (6)) for the calibrated AOB model. Note that the model does not differentiate between ammonium and ammonia uptake, since in the bioreactor the two compounds are at dissociation equilibrium, determined by pH and the sum of ammonium and ammonia concentrations. Nevertheless, uptake kinetics were specified in terms of ammonia concentrations according to findings by Suzuki et al. (463) that suggest that ammonia is likely the limiting substrate.

Random AOB strains had modified kinetics:

\[
\frac{[\text{NH}_3]}{\frac{1}{\alpha_r(1-\kappa_r)} + \frac{[\text{NH}_3]}{\kappa_r V_r}}, \quad (D.3.29)
\]

where \(\kappa_r\) is chosen randomly from 0 to 1, \(V_r\) is chosen randomly from \(V/10\) to \(10V\), and
α_r is chosen randomly from α/10 and 10α. We randomly varied V and the affinity α (as opposed to the half-saturation concentration) according to Aksnes and Cao (6), who showed that affinity, rather than half-saturation concentration, is an inherent biological trait. The random parameter κ_r accounts for trait-offs between affinities and maximum cell-specific uptake rates, as suggested by (431).

- Nitrite/nitrous acid (HNO_2.NO_2) uptake rate limits are specified as a Monod function of HNO_2.NO_2:

\[
\frac{[\text{HNO}_2\text{.NO}_2]}{\frac{1}{\alpha} + \frac{[\text{HNO}_2\text{.NO}_2]}{V}},
\]

where \( V = 7.31 \times 10^{-13} \text{ mol/cell/day} \) and \( \alpha = 3.19 \times 10^{-9} \text{ L/cell/day} \) for the calibrated AOB model. Note that the model does not differentiate between nitrite and nitrous acid uptake, since in the bioreactor the two compounds are at dissociation equilibrium determined by pH.

Random NOB strains had modified kinetics:

\[
\frac{[\text{HNO}_2\text{.NO}_2]}{\frac{1}{\alpha_r(1-\kappa_r)} + \frac{[\text{HNO}_2\text{.NO}_2]}{\kappa_r V_r}},
\]

where \( \kappa_r \) is chosen randomly from 0 to 1, \( V_r \) is chosen randomly from \( V/10 \) to \( 10V \), and \( \alpha_r \) is chosen randomly from \( \alpha/10 \) and \( 10\alpha \).

D.3.6 Community-scale dynamics

The membrane bioreactor model keeps track of the cell concentrations of each AOB and NOB strain, as well as extracellular ammonia/ammonium (NH_3.NH_4), nitrite/nitrous acid (HNO_2.NO_2) and nitrate/nitric acid (HNO_3.NO_3) concentrations over time in terms of differential equations. The bioreactor is assumed to be well mixed and well oxygenated. Let \( N_i(t) \) be the cell concentration for strain \( i \) (AOB or NOB), \( C_j(t) \) the concentration of metabolite \( j \) and \( E_{ij}(t) \) the net export rate of metabolite \( j \) by each cell of strain \( i \), at time \( t \). Note that \( E_{ij}(t) \) is calculated by solving a linear FBA problem for each strain and at each time step. The constraints of the FBA problems, and thus their solutions, depend on the maximum substrate uptake rates, which are given by Monod kinetics as described above. All population sizes \( N_i \) are described by ordinary differential equations (ODEs) of the following
form:

\[
\frac{dN_i(t)}{dt} = \frac{1}{m_i} B_i(t) N_i(t) - \lambda_i N_i(t). 
\]

(D.3.32)

Here, \( \lambda_i \) is the exponential death rate of strain \( i \), i.e., the inverse of its expected life time, \( B_i(t) \) is the per-cell biosynthesis rate determined by FBA and \( m_i \) is the cell mass. Similarly, all metabolite concentrations \( C_i \) are described by ODEs of the following form:

\[
\frac{dC_j}{dt} = r \cdot [\tilde{C}_j - C_j(t)] + \sum_i E_{ij}(t) N_i(t). 
\]

(D.3.33)

Here, \( r \) is the hydraulic turnover rate (feed rate over bioreactor volume) and \( \tilde{C}_j \) is the concentration of the metabolite in the feed. The last term is a sum over all cell populations, accounting for microbial production or consumption of the metabolite.
Appendix E

Chapter 6: Supplemental material

E.1 Methods

E.1.1 Model overview

Our model describes the population dynamics of multiple bacterial and archaeal operational taxonomic units (OTUs), their reaction kinetics, the population dynamics of multiple phage populations, as well as extracellular metabolite concentrations in a flow-through bioreactor. Here, an “OTU” represents a cell lineage that is specialized on a specific metabolic function (e.g., acetoclastic methanogenesis) and predated by its own specialist phage population. Hence, an OTU represents a taxonomic group that is sufficiently narrow so that reaction kinetics are similar across members, and sufficiently broad so that different OTUs have different specialist phages. Hence, an OTU in our model is roughly analogous to a single prokaryotic species or strain (69, 174, 521).

The bioreactor model largely resembles the setup used in previous experiments (122, 533). Glucose is supplied continuously to the bioreactor as part of a sterile inflow, which is balanced by an equivalent outflow that removes residual substrates, metabolic by-products as well as cells and free phage particles at a constant hydraulic renewal rate. The bioreactor’s interior is assumed to be well mixed and anaerobic. pH and temperature are held constant.

E.1.2 Reaction rates and metabolite dynamics

The model considers a total of 12 reactions, driving the stepwise catabolism of glucose all the way to the eventual production of methane (see Table E.1 for a list of reactions and Fig. 6.1 for a schematic overview). Each OTU is associated with a single metabolic reaction, such as fermentation of glucose to ethanol or acetoclastic methanogenesis, but each reaction may be performed by multiple competing OTUs. The cell-specific rate of a reaction performed by some OTU $s$, $H_s$, is assumed to be limited by a single limiting substrate (such as glucose...
in the case of fermenters) according to classical Monod-kinetics (211):

\[ H_s = \frac{V_s C}{C + K_s}. \]  
(E.1.1)

Here, \( K_s \) is a half-saturation concentration of the limiting substrate (specific to OTU \( s \)), \( C \) is the substrate concentration and \( V_s \) is the maximum cell-specific reaction rate (specific to OTU \( s \)). Each reaction couples the uptake of a number of substrates to the export of a number of products into the extracellular medium, thereby affecting metabolite concentrations in the bioreactor. Specifically, the concentration of the \( m \)-th metabolite in the bioreactor, \( C_m \), changes according to the differential equation

\[ \frac{dC_m}{dt} = \lambda (C_m^o - C_m) + \sum_r S_{mr} \sum_{s \in J_r} N_s H_s, \]  
(E.1.2)

where \( J_r \) is the set of OTUs performing reaction \( r \), \( N_s \) is the cell concentration of OTU \( s \), \( \lambda \) is the hydraulic renewal rate, \( C_m^o \) is the metabolite’s concentration in the inflow (zero for all metabolites except glucose) and \( S_{mr} \) is the stoichiometric coefficient of metabolite \( m \) in reaction \( r \). For example, for glucose fermentation to ethanol,

\[ C_6H_{12}O_6 \rightarrow 2 \text{CH}_3\text{CH}_2\text{OH} + 2 \text{CO}_2, \]  
(E.1.3)

the stoichiometric coefficients of glucose, ethanol and \( \text{CO}_2 \) are \(-1\), \(+2\) and \(+2\), respectively. For each OTU, the total cell production rate is assumed to be proportional to the total rate of its catalyzed reaction, multiplied by the reaction’s Gibbs free energy (as described below).

**E.1.3 Gibbs free energy and cell production**

The Gibbs free energy of a reaction (\( \Delta G \)) is conventionally interpreted as the amount of energy that can be readily transformed to “work” (e.g., for ATP production) by a microbial population (492). For example, the zonation of various microbial groups along redox transition zones in sediments is strongly determined by the Gibbs free energy of their metabolic pathways (56). In anaerobic methanogenic digesters, the growth of so-called syntrophic bacteria is (in addition to substrate availability) highly dependent upon the removal of downstream products (e.g., acetate and \( \text{H}_2 \)) by methanogens, because a low partial pressure of products ensures that the Gibbs free energy of reactions is sufficient to sustain growth (77). Hence, fluctuations of methanogen populations can affect upstream metabolic activity and population dynamics by modulating the energetic yield of reactions.
A meta-analysis of several anaerobic metabolic pathways by Roden and Jin (396) showed that microbial biosynthesis rates are approximately proportional to the Gibbs free energy released by the utilized pathway. The regression formula provided by Roden and Jin (396) has been included in subsequent microbial ecological models for predicting cell growth rates based on their metabolic activity (287, 386, 387). Following previous work, we thus assume that the amount of biosynthesis supported by reaction $r$ is given by

$$Z_r = 2.08 - \frac{1}{\gamma_r^e} 0.0211 \times \Delta G_r \text{ (g dry biomass per mole e-donor consumed).} \quad (E.1.4)$$

Here, $\gamma_r^e$ is the absolute stoichiometric coefficient of the electron donor in the reaction,

$$\Delta G_r = \Delta G_o^r + R_g T \ln Q_r \quad (E.1.5)$$

is the Gibbs free energy of the reaction (in kJ per mol), $\Delta G_o^r$ is the standard Gibbs free energy of the reaction (see Table E.1), $T$ is the temperature in Kelvin, $R_g = 8.314 \text{ J} \cdot \text{K} \cdot \text{mol}^{-1}$ is the molar gas constant and

$$Q_r = \prod_m C_m^{S_{mr}} \quad (E.1.6)$$

is the so-called reaction quotient (96). The per-capita biosynthesis rate, $Z_r h_r$, is translated to a per-capita cell production rate by dividing by the dry cell mass.

We note that determining microbial growth rates based on Gibbs free energy fluxes is a non-trivial task even for engineered systems (312), and other factors — such as cell-specific maintenance requirements — can lead to deviations from regression models. Nevertheless, our model successfully captures key aspects of microbial metabolic networks, namely stoichiometric balancing between metabolic pathways and the dependence of biosynthesis on energetic yield (396).

### E.1.4 Cell and phage population dynamics

In the model, each OTU $s$ is associated with a single specialist lytic phage population, which comprises free phage particles as well as phages that have infected a host cell. The model keeps track of the infected portion, $N_i^s$, and the uninfected portion, $N_u^s$, of each cell population. The rate at which healthy cells become infected is proportional to the total number of free phage particles ($P_s^i$), multiplied by some proportionality constant ($\beta_s$) that accounts for the rate at which phages “scan” the medium via passive diffusion (volume clearance rate).
as well as the probability that an encounter with a cell would lead to infection (322). All infected cells are assumed to eventually undergo lysis, releasing new phage particles into the bioreactor (423). Lysogeny is not considered in the model. The loss of uninfected cells is assumed to be driven by hydraulic dilution, and is hence modeled as an exponential decay rate $\lambda$. In addition to hydraulic dilution, infected cells suffer from an elevated mortality rate, $\mu_s$, which is equivalent to the inverse of the time lag between infection and cell lysis. Hence, uninfected and infected cell concentrations change according to the differential equations

$$\frac{dN_s^u}{dt} = \frac{Z_s}{m_s} N_s^u H_s - \lambda N_s^u - \beta_s P_s N_s^u,$$

(E.1.7)

$$\frac{dN_s^i}{dt} = \beta_s P_s N_s^u - \lambda N_s^i - \mu_s N_s^i.$$

(E.1.8)

Here, $Z_s$ is the biomass yield of the reaction catalyzed by species $s$ (introduced above) and $m_s$ is the dry cell mass. Hence, the first term on the right hand side of Eq. (E.1.7) accounts for a variable cell growth rate, depending on the current metabolic rate and the energy currently available from a reaction. Observe that the term $\beta_s P_s N_s$ (rate of new infections of OTU $s$) increases as the concentration of cells increases, which is a key prerequisite for KTW dynamics. Phage-induced cell lysis leads to the release of new free phage particles. Phage particles that fail to infect any cells are assumed to eventually get flushed out of the bioreactor. Hence, the concentration of free phage particles associated with species $s$ satisfies the differential equation

$$\frac{dP_s}{dt} = \nu_s \mu_s N_s^i - \beta_s P_s N_s^u - \lambda P_s,$$

(E.1.9)

where $\nu_s$ is the average number of phage particles released per lysed cell.

### E.1.5 Parameterization and simulations

Model parameters were either fixed at values obtained from the literature, or chosen randomly and uniformly within an interval around values obtained from the literature (overview in Table E.2). In particular, for each OTU the reaction-kinetic parameters ($V$ and $K$; Eq. E.1.1) as well as parameters describing phage-host interactions ($\beta_s$, $\nu_s$ and $\mu_s$; Eqs. E.1.7–E.1.9) were chosen randomly and independently of other OTUs. The glucose concentration in the inflow is set to 8 mg · L$^{-1}$. This glucose concentration is comparable to dissolved organic carbon concentrations in natural methanogenic environments (215, 505), although it is lower than in typical bioreactor feeds (122, 533). Apart from influencing the overall extent of fluctuations in the bioreactor, the choice of glucose input (within ranges spanning
natural and engineered systems) did not influence our overall conclusions.

The differential equations (E.1.2), (E.1.7), (E.1.8) and (E.1.9) describe a high-dimensional deterministic dynamical system of $3S + M$ time-dependent variables, where $S$ is the number of OTUs and $M$ is the number of considered metabolites. Numerical simulations of this system were performed using MCM (Chapter 4; 284).

### E.1.6 Statistical analysis

To quantify the metabolic performance of the community (in terms of methane production), for each simulation we calculated the average effluent methane concentration over time. To quantify the variation in metabolic performance we calculated the coefficient of variation (CV, i.e., the standard deviation divided by the average) of effluent methane concentration over time. Note that for each simulation the average methane concentration and its CV were different and random because several model parameters were chosen randomly. For each degree of functional redundancy, we used 50 random simulations to estimate the distribution of average methane concentrations and their CVs (box-plots in Figs. 6.3A,B).

To quantify the variation in functional community structure across time during any particular simulation, we calculated the CVs of functional group proportions and averaged these over all considered functional groups. For example, to quantify the variation of coarse functional groups (fermenters, syntrophs, methanogens) we considered the average of (1) the CV of the fraction of fermenters, (2) the CV of the fraction of syntrophs and (3) the CV of the fraction of methanogens in the community. Similarly, to quantify the variation of methanogenic groups we considered the average of (1) the CV of the fraction of acetoclastic methanogens and (2) the CV of the fraction of $\text{H}_2/\text{CO}_2$ methanogens. Note that CVs were different and random for each simulation, because several model parameters were chosen randomly. For each degree of functional redundancy, we used 50 random simulations to estimate the distribution of CVs (box-plots in Figs. 6.4I–L).

As mentioned in the main text, to distinguish between statistical averaging and dynamic stabilization we compared the CVs of functional group proportions to a null model in which OTU populations fluctuated independently. Specifically, for each simulation the null model cyclically shifted the time series of each OTU by a random time step, resulting in hypothetical community trajectories in which each OTU population fluctuates at a random phase lag when compared to other OTUs. The shifted time series of all OTUs within each functional group were then summed to calculate the hypothetical corresponding abundance of the functional
group. For each simulation this was done 1000 times, yielding a distribution of random CVs generated by the null model. The “degree of dynamic stabilization” (DDS) of a single simulation was then defined as the fraction of random CVs that were above the actual CV of the simulation. For each degree of functional redundancy, we used 50 random simulations to estimate the distribution of DDSs (box-plots in Figs. 6.4M–P).

All statistics were calculated using the time series spanning days 500–1000, in order to avoid any transients right after inoculation.

E.1.7 Deterministic vs stochastic competitive exclusion

Demographic drift between similar competitors has been suggested previously (349) as a cause of seemingly random and sustained OTU succession observed within functional groups under constant environmental conditions. At high cell densities, however, deterministic dynamics such as competitive exclusion or predator-prey cycles are expected to dominate over stochastic demographic drift, although the importance of drift depends on the extent of competitive differences and the strength of biotic interactions, as well as the spatial scale at which populations are mixed (e.g., the entire bioreactor in case of vigorous mixing). For example, significant differences in growth rates, enzyme efficiencies, maintenance rates and stress responses are common between prokaryotic sister species or even between strains of the same species (207, 328, 352). To assess the plausibility of random demographic drift as a possible explanation for OTU succession in the face of competition, we examined a simple stochastic birth-death model (described below) for the population sizes of two competing OTUs with slightly different competitive abilities. We performed multiple simulations of the model and compared the resulting stochastic trajectories to the corresponding deterministic trajectory leading to competitive exclusion. We measured the deviation of each stochastic trajectory from the corresponding deterministic trajectory in terms of the coefficient of determination ($R^2$). As will become clear below, being conservative in terms of the assumed competitive differences between OTUs strengthens the confidence in our results.

The model considers the population sizes ($N_1$ and $N_2$) of two competing OTUs with equal death rates but slightly different birth rates, while assuming a constant combined population size ($N = N_1 + N_2$). At each time step, a random cell is lost (“death”) from one of the two populations, while another random cell is added (“birth”) to one of the two populations. The probability that the lost cell belongs to population 1 is thus $N_1/N$. The probability that the added cell belongs to population 1 is $N_1/(N_1 + (1 + s)N_2)$, where we assumed that the ratio of per-capita birth rates (OTU 1 : OTU 2) is $1 : (1 + s)$ and $s$ is the relative difference between
the two per-capita birth rates ("relative advantage"). Hence, at each time step \( N_1 \) can either decrease by 1, increase by 1 or remain unchanged, while the transition probabilities depend on the current \( N_1 \). On the other hand, the deterministic trajectory of \( N_1 \) corresponding to the above birth-death process is given by the difference equation

\[
N_1(t + 1) = N_1(t) + \frac{N_1}{N_1 + (1 + s)N_2} - \frac{N_1}{N},
\]  

(E.1.10)

where \( t \) is the number of elapsed time steps. We note that previously published neutral models for microbial communities also include random immigration of cells from a "regional pool" as an additional source of fluctuations in the considered community (349, 427). Immigration is omitted from our model, because in the systems in which OTU turnover within functional groups has been observed over time (109, 349, 397) the bulk of living cells was likely produced within the system at hand, rather than added via immigration.

Each simulation of the above stochastic birth-death model was initiated at equal population sizes \( (N_1 = N_2 = N/2) \) and was ran until population 1 dropped below a given "extinction threshold" \( (E) \), at which point competitive exclusion was considered to be complete. The coefficient of determination was then calculated as

\[
R^2 = 1 - \frac{\sum_{t=1}^{T} [\bar{N}_1(t) - N_1(t)]^2}{\sum_{t=1}^{T} [N_1(t) - \bar{N}_1]^2},
\]  

(E.1.11)

where \( T \) is the number of time steps until competitive exclusion, and \( \bar{N}_1 \) is the average of the stochastic trajectory. For the example cited in the main text, we used a combined population size of \( N = 10^5 \), a relative advantage of \( s = 1\% \) and a threshold of \( E = 0.01 \times N \). The \( R^2 \) reported in the text was averaged over 1000 random simulations. The fact that even at such a low population size and such a weak competitive advantage the stochastic trajectory closely resembles the deterministic trajectory of competitive exclusion, strengthens our argument that pure demographic drift is an unlikely explanation for the OTU turnover observed in previous experiments (109, 349, 397).
Figure E.1: **Predicted metabolite uptake rates.** Community-wide metabolite uptake rates over time during a simulation at 50-fold functional redundancy.
Figure E.2: **Predicted metabolite export rates.** Community-wide metabolite export rates over time during a simulation at 50-fold functional redundancy.
Figure E.3: **Predicted phage-host trajectories.** Cell concentrations (horizontal axes) and associated phage concentrations (vertical axes) across time (one plot per phage-host pair), during a simulation at 5-fold functional redundancy (i.e., comprising 60 cell populations). A brighter point on a trajectory indicates an earlier time in the simulation. Trajectories that appear completely bright correspond to hosts that went extinct early in the simulation.
Table E.1: **Pathway stoichiometry.** Reaction stoichiometry and standard Gibbs free energies ($\Delta G_o$, kJ · mol$^{-1}$ substrate) used in the model, taken from Conrad (77). Reaction IDs are as in Fig. 6.1.

<table>
<thead>
<tr>
<th>ID</th>
<th>Reaction</th>
<th>$\Delta G_o$</th>
</tr>
</thead>
<tbody>
<tr>
<td>glucose fermentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>$C_6H_{12}O_6 \rightarrow 2 CH_3CH_2OH + 2 CO_2$</td>
<td>-235.0</td>
</tr>
<tr>
<td>B</td>
<td>$C_6H_{12}O_6 \rightarrow 2 CH_3CHOHCOOH$</td>
<td>-198.1</td>
</tr>
<tr>
<td>C</td>
<td>$C_6H_{12}O_6 \rightarrow 3 CH_3COOH$</td>
<td>-311.2</td>
</tr>
<tr>
<td>D</td>
<td>$C_6H_{12}O_6 + 2 H_2O \rightarrow 2 CH_3COOH + 2 CO_2 + 4 H_2$</td>
<td>-216.1</td>
</tr>
<tr>
<td>E</td>
<td>$C_6H_{12}O_6 \rightarrow 4/3 CH_3CH_2COOH + 2/3 CH_3COOH + 2/3 CO_2 + 2/3 H_2O$</td>
<td>-311.4</td>
</tr>
<tr>
<td>F</td>
<td>$C_6H_{12}O_6 \rightarrow 2/3 CH_3CH_2CH_2COOH + 2/3 CH_3COOH + 2 CO_2 + 8/3 H_2$</td>
<td>-248.0</td>
</tr>
<tr>
<td>syntrophy (catabolism of short-chain fatty acids, lactate and alcohols)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>$CH_3CH_2OH \rightarrow CH_3COOH + 2 H_2$</td>
<td>+9.6</td>
</tr>
<tr>
<td>H</td>
<td>$CH_3CHOHCOOH + H_2O \rightarrow CH_3COOH + CO_2 + 2 H_2$</td>
<td>-48.7</td>
</tr>
<tr>
<td>I</td>
<td>$CH_3CH_2COOH + 2 H_2O \rightarrow CH_3COOH + CO_2 + 3 H_2$</td>
<td>+31.8</td>
</tr>
<tr>
<td>J</td>
<td>$CH_3CH_2CH_2COOH + 2 H_2O \rightarrow 2 CH_3COOH + 2 H_2$</td>
<td>+48.3</td>
</tr>
<tr>
<td>methanogenesis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>$CH_3COOH \rightarrow CO_2 + CH_4$</td>
<td>-35.6</td>
</tr>
<tr>
<td>L</td>
<td>$4H_2 + CO_2 \rightarrow 2 H_2O + CH_4$</td>
<td>-32.7</td>
</tr>
</tbody>
</table>
Table E.2: **Model parameters.** Parameter types used in the model, including substrate half-saturation concentrations and maximum cell-specific substrate uptake rates. Parameters marked with an asterisk (*) are randomly and uniformly chosen within an interval spanning $10 - 1000$ % of their default value, independently for each OTU and for each simulation. References indicated by “†”: Mass-specific rates converted to cell-specific rates based on a dry cell mass of $2.8 \times 10^{-13} \text{ g}$ (344).

<table>
<thead>
<tr>
<th>symbol and description</th>
<th>scope</th>
<th>default value</th>
<th>reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m$ dry cell mass</td>
<td>fermenters</td>
<td>280 fg</td>
<td>(344)</td>
</tr>
<tr>
<td>$m$ dry cell mass</td>
<td>syntrophs</td>
<td>280 fg</td>
<td>(344)</td>
</tr>
<tr>
<td>$m$ dry cell mass</td>
<td>H$_2$/CO$_2$ methanogens</td>
<td>440 fg</td>
<td>(5)</td>
</tr>
<tr>
<td>$m$ dry cell mass</td>
<td>acetoclastic methanogens</td>
<td>2.5 pg</td>
<td>(5)</td>
</tr>
<tr>
<td>$V$ max. cell-sp. glucose upt. rate</td>
<td>all glucose fermenters</td>
<td>* 67.2 fmol · cell$^{-1}$ · d$^{-1}$</td>
<td>(318)$^\dagger$</td>
</tr>
<tr>
<td>$K$ glucose half-saturation conc.</td>
<td>— ” —</td>
<td>* 0.53 mM</td>
<td>(318)$^\dagger$</td>
</tr>
<tr>
<td>$V$ max. cell-sp. H$_2$ upt. rate</td>
<td>H$_2$/CO$_2$ methanogens</td>
<td>* 1.43 pmol · cell$^{-1}$ · d$^{-1}$</td>
<td>(5)</td>
</tr>
<tr>
<td>$K$ H$_2$ half-saturation conc.</td>
<td>— ” —</td>
<td>* 7.65 µM</td>
<td>(5)</td>
</tr>
<tr>
<td>$V$ max. cell-sp. acetate upt. rate</td>
<td>acetoclastic methanogens</td>
<td>* 0.55 pmol · cell$^{-1}$ · d$^{-1}$</td>
<td>(5)</td>
</tr>
<tr>
<td>$K$ acetate half-saturation conc.</td>
<td>— ” —</td>
<td>* 442 µM</td>
<td>(5)</td>
</tr>
<tr>
<td>$V$ max. cell-sp. lactate upt. rate</td>
<td>lactate syntrophs</td>
<td>* 143 fmol · cell$^{-1}$ · d$^{-1}$</td>
<td>(78)$^\dagger$</td>
</tr>
<tr>
<td>$K$ lactate half-saturation conc.</td>
<td>— ” —</td>
<td>* 380 µM</td>
<td>(78)</td>
</tr>
<tr>
<td>$V$ max. cell-sp. ethanol upt. rate</td>
<td>ethanol syntrophs</td>
<td>* 536 fmol · cell$^{-1}$ · d$^{-1}$</td>
<td>(80)$^\dagger$</td>
</tr>
<tr>
<td>$K$ ethanol half-saturation conc.</td>
<td>— ” —</td>
<td>* 0.3 µM</td>
<td>(220)</td>
</tr>
<tr>
<td>$V$ max. cell-sp. butyrate upt. rate</td>
<td>butyrate syntrophs</td>
<td>* 72.8 fmol · cell$^{-1}$ · d$^{-1}$</td>
<td>(78)$^\dagger$</td>
</tr>
<tr>
<td>$K$ butyrate half-saturation conc.</td>
<td>— ” —</td>
<td>* 76 µM</td>
<td>(5)</td>
</tr>
<tr>
<td>$V$ max. cell-sp. propionate upt. rate</td>
<td>propionate syntrophs</td>
<td>* 44.8 fmol · cell$^{-1}$ · d$^{-1}$</td>
<td>(78)$^\dagger$</td>
</tr>
<tr>
<td>$K$ butyrate half-saturation conc.</td>
<td>— ” —</td>
<td>* 432 µM</td>
<td>(259)</td>
</tr>
<tr>
<td>$\beta$ phage infectivity</td>
<td>free phage particles</td>
<td>* $3 \times 10^{-10}$ L · d$^{-1}$</td>
<td>(1, 322)</td>
</tr>
<tr>
<td>$\mu$ lysis rate</td>
<td>all infected cells</td>
<td>* 2 d$^{-1}$</td>
<td>(319, 322)</td>
</tr>
<tr>
<td>$\nu$ phage particles released per lysis</td>
<td>— ” —</td>
<td>* 10</td>
<td>(319, 322)</td>
</tr>
<tr>
<td>$\lambda$ hydraulic turnover rate</td>
<td>bioreactor</td>
<td>0.1 d$^{-1}$</td>
<td>(122, 533)</td>
</tr>
<tr>
<td>$C_{g0}^{\text{bi}}$ glucose input concentration</td>
<td>bioreactor</td>
<td>16 mg · L$^{-1}$</td>
<td>(215, 505)</td>
</tr>
<tr>
<td>$T$ temperature</td>
<td>bioreactor</td>
<td>35°C</td>
<td>(122, 533)</td>
</tr>
<tr>
<td>pH</td>
<td>bioreactor</td>
<td>7.0</td>
<td>(122, 533)</td>
</tr>
</tbody>
</table>
Appendix F

Chapter 7: Supplemental material

F.1 Methods

F.1.1 Model overview

The core model is a set of differential equations for the concentrations of 8 metabolites and 6 proxy genes (DNA) across depth (100–200 m) and time. Each gene is a proxy for a particular energy-yielding pathway, which couples the oxidation of an external electron donor to the reduction of an external electron acceptor. Each gene is considered as a replicating unit that is independent of other genes. This corresponds to the simplifying assumption that each cell occupies a single metabolic niche associated with one of the modeled pathways (181, 383). Gene-specific reaction rates depend on the concentrations of all involved metabolites according to 1st or 2nd order (Michaelis-Menten) kinetics (211, 386) (Appendix F.3.4). In turn, the production or depletion of metabolites at any depth is determined by the reaction rates at that depth, taking into account reaction stoichiometry (Appendix F.3.3) and diffusive transport across the water column. The production of genes at any depth is driven by the release of energy from their catalyzed reactions, and is proportional to the Gibbs free energy multiplied by the reaction rate (396) (Appendix F.3.5). In addition, gene populations are subject to exponential decay rates, diffusive transport and sinking.

F.1.2 Mathematical model structure

The DNA concentration for gene \( r \) (\( \Gamma_r \), copies per volume) exhibits the dynamics

\[
\frac{\partial \Gamma_r}{\partial t} = -q_r \Gamma_r + \frac{1}{c} Z_r H_r \Gamma_r - v \frac{\partial \Gamma_r}{\partial z} + \frac{\partial}{\partial z} \left( K_r \frac{\partial \Gamma_r}{\partial z} \right),
\]  

(F.1.1)
while the concentration of the $m$-th metabolite ($C_m$, mole per volume) follows

$$
\frac{\partial C_m}{\partial t} = \sum_r S_{mr} H_r \Gamma_r + \frac{\partial}{\partial z} \left( K \frac{\partial C_m}{\partial z} \right).
$$

(F.1.2)

Both the gene concentrations $\Gamma_r$ and metabolite concentrations $C_m$ depend on time $t$ and depth $z$. The first term on the right-hand-side of Equation (F.1.1) corresponds to cell death, with $q_r$ being the exponential death rate in the absence of any metabolic activity for pathway $r$. The 2nd term corresponds to gene production, with $H_r$ being the per-gene reaction rate as a function of metabolite concentrations (Appendix F.3.4). The biomass production coefficient $Z_r$ is a linear function of the Gibbs free energy of reaction $r$ (Appendix F.3.5). $c$ is the average dry cell mass, which is used to convert biomass production into cell production. The 3rd term corresponds to cell sinking at speed $v$. The last term in Equations (F.1.1) and (F.1.2) corresponds to diffusive transport, with $K$ being the vertical eddy-diffusion coefficient. In Equation (F.1.2), $S_{mr}$ is the stoichiometric coefficient of metabolite $m$ in reaction $r$ (Appendix F.3.3). The sum on the right hand side of Equation (F.1.2) iterates through all reactions and thus accounts for microbial metabolic fluxes. Equations (F.1.1) and (F.1.2) specify the rates of change for the DNA and chemical concentration profiles. Steady state profiles were obtained after long simulations when all profiles had eventually stabilized.

F.1.3 Considered pathways

Redox pathways occurring in a single cell require at least two enzymes, one involved in the oxidation of the initial electron donor and one involved in the reduction of the final electron acceptor. In the model such pathways are represented by single proxy genes, chosen such that ambiguities in their functional role are minimized. For example, nitrous oxide reduction using nitrous oxide reductase ($nosZ$) coupled to sulfide oxidation is identified with $nosZ$, because many sulfur oxidizing enzymes are reversible. Other pathways considered in the model are partial denitrification of nitrate to nitrous oxide coupled to sulfide oxidation (PDNO), aerobic ammonium oxidation using ammonia monooxygenase ($amo$), aerobic nitrite oxidation to nitrate using nitrite oxidoreductases ($nxr$), anammox, i.e., the anaerobic ammonium oxidation involving hydrazine oxidoreductase ($hzo$), as well as aerobic remineralization of (dissolved) organic matter (ROM). PDNO comprises 3 denitrification steps which are thought to be predominantly performed by the same microorganisms in the SUP05 clade (181, 504): nitrate reduction to nitrite involving dissimilatory nitrate reductases ($narGHIJ$ or $napAB$), nitrite reduction to nitric oxide using nitrite reductases ($nirKS$) and nitric oxide
reduction to nitrous oxide using nitric oxide reductases \((\text{norBC})\). The first denitrification step was assumed to be leaky, so that a small fraction of nitrite is released into the extracellular environment \((253)\). We used \text{norBC} as a proxy for PDNO when interpreting molecular data (but see Figs. F.1 d,e,f in the Appendix for \text{narGHIJ}, \text{napAB} and \text{nirKS} multimolecular data, and Fig. 7.3A for coverage of the dissimilatory sulfide oxidation pathway). ROM is associated with the release of ammonium and sulfate \((\text{SO}_4^{2-})\) at ratios corresponding to marine bacterial biomass stoichiometry \((115)\). The choice of redox pathways in the model follows the hypotheses put forward by Hawley et al. \((181)\) based on molecular depth profiles, as well as reports of nitrous oxide reduction coupled to hydrogen sulfide oxidation in Saanich Inlet \((74)\).

Hydrogen sulfide is assumed to originate via diffusion from the sediments, where intense sulfate reduction occurs \((340)\) (Appendix F.4.1). Sulfate reduction was omitted from our model because both our molecular as well as chemical data suggest that sulfate reduction in the water column is negligible compared to the oxidation of sulfur compounds (see Appendix F.4.1 for a detailed discussion). In fact, when we included sulfate reduction in preliminary tests of our model the agreement between the model and the \(\text{H}_2\text{S}\) profiles decreased dramatically, providing further evidence that \(\text{H}_2\text{S}\) is largely supplied from the bottom, rather than produced in the water column.

Aerobic \(\text{H}_2\text{S}\) oxidation was omitted from the model based on extensive previous work that points towards \(\text{NO}_3^-\) and other nitrogen compounds as dominant electron acceptors for \(\text{H}_2\text{S}\) oxidation during periods of strong stratification \((11, 181, 218, 503, 540)\). For example, as shown in Fig. 7.2B, the upper boundary of \(\text{H}_2\text{S}\) concentrations closely follows the lower boundary of \(\text{NO}_3^-\) — rather than \(\text{O}_2\) — over time, especially during the period considered here (early 2010). We mention that during renewal events in Fall \(\text{O}_2\) can become an important electron acceptor for \(\text{H}_2\text{S}\) oxidation \((540)\), however this does not affect this study, which focuses on a period of intense stratification near steady state conditions. A more detailed discussion on the role of aerobic sulfide oxidation is provided in Appendix F.4.3.

Pathways for hydrogen \((\text{H}_2)\) and methane \((\text{CH}_4)\) metabolism are not included on grounds of parsimony, because these are not directly linked to the other considered pathways \((540)\) and because low hydrogen and methane fluxes into the OMZ suggest that hydrogen and methane pathways are of secondary importance \((275, 540)\).
F.1.4 Model calibration and data

Unknown parameters of the basic gene-centric model (Eq. (F.1.1) and (F.1.2), ignoring mRNA and protein dynamics) were calibrated by comparing steady state predictions to measured depth profiles of oxygen, ammonium, nitrate, nitrite, hydrogen sulfide and nitrous oxide. Chemical calibration data were acquired on January 13, February 10 and March 10, 2010 (or February 10 and April 7 for oxygen) from a single location in Saanich Inlet (123° 30.30′ W, 48° 35.50′ N; Appendix F.2.2). The calibrated parameters were the maximum cell-specific reaction rate $V_{PDNO}$, the 1st order rate constants $A_{ROM}$ and $A_{nosZ}$, as well as the PDNO leakage fraction $L_{PDNO}$ (Supplemental Table F.2). Calibration was performed by maximizing the likelihood of a statistical model, in which the deterministic part (i.e., expectation) is given by the predictions of the gene-centric model and the stochastic part (i.e., error) is normally distributed (Appendix F.3.8). This calibration method is known as maximum-likelihood estimation and is widespread in statistical regression and physics (113). Maximization of the likelihood was performed using the MATLAB® function `fmincon`, which uses repeated simulations and gradual exploration of parameter space (309). The sensitivity of the model to parameter variation was assessed via local sensitivity analysis (71), as described in Appendix F.3.12. An overview of our workflow is shown in Supplemental Fig. F.2.

Samples for molecular sequencing were collected on February 10, 2010 from the same location as the geochemical data (Appendixes F.2.3 and F.2.4). Metagenomic profiles (a priori in relative units) were rescaled to match the model scales using maximum-likelihood estimated factors (Appendix F.3.9). SUP05 cell counts for February 10, 2010 were estimated via quantitative PCR (qPCR) using SUP05-specific primers targeting the 519–1048 region of the SUP05 16S rRNA gene, following the protocol by Hawley et al. (182). 16S gene counts were corrected for the number of 16S rRNA gene copies per cell, estimated using the Tax4Fun pipeline (19) (Appendix F.2.6). Denitrification and anammox rates were measured on cruises 47 (SI047_07/07/10) and 48 (SI048_08/11/10) via ex situ incubation experiments, and were subsequently corrected for differences between in-situ and incubated substrate concentrations (Appendix F.2.5).

F.1.5 mRNA and protein models

As mentioned previously, upon calibration of the gene-centric model to the geochemical profiles, we extended the model to describe mRNA (and similarly, protein) dynamics in the water column. Specifically, the production rate of an mRNA (transcripts produced per
time and per volume of seawater) at a particular depth was assumed to be proportional to the total reaction rate (mol per time and per volume of seawater) at that depth. A linear relation, while only an approximation, can be justified by the fact that increased enzyme dilution rates at elevated cell division rates must be balanced (at the population level) by correspondingly increased translation — and hence transcription — rates (403). We also assumed that mRNA molecules disperse via diffusion and sinking similarly to genes (as they are hosted by the same cells) and decay exponentially with time. Thus, environmental mRNA concentrations satisfy the partial differential equation

\[
\frac{\partial T_r}{\partial t} = -\frac{T_r}{\tau_r} + \frac{R_r}{\alpha_r} - v \frac{\partial T_r}{\partial z} + \frac{\partial}{\partial z} \left( K \frac{\partial T_r}{\partial z} \right),
\]

where \( T_r \) is the mRNA concentration corresponding to the \( r \)-th reaction, \( \tau_r \) is the decay time of the mRNA molecule, \( R_r = H_r \Gamma_r \) is the total reaction rate and \( \alpha_r \) is an unknown proportionality constant. We considered \( T_r \) in the same units as the multi-omic data (i.e., RPKM for metatranscriptomes and NSAF for metaproteomes). Consequently, \( \alpha_r \) is the ratio between the \( r \)-th reaction rate and the corresponding RPKM (or NSAF) “production rate” (mol per time per volume of seawater per RPKM), and thus not only depends on the particular reaction, but also on our sampling protocol and sequencing pipeline. The above model was evaluated at steady state, when mRNA production, dispersal and decay are balanced at each depth (\( \partial T_r / \partial t = 0 \)). The parameters of the mRNA and protein models (proportionality factors and decay times) were calibrated by fitting to the metatranscriptomic and metaproteomic data, respectively (Appendix F.3.10). Calibration to metatranscriptomic data failed for \( amo \) mRNA. Metagenomic and metaproteomic data were not available for \( nxr \) and \( nosZ \), respectively (Appendix F.2.3). For all other mRNAs and proteins, the iterative calibration converged rapidly to an optimum and this optimum was robust against various starting values for the parameters.

**F.1.6 Inverse linear transport modeling**

In addition to the model predictions and rate measurements, denitrification and anammox rates were also estimated directly from chemical concentration profiles via inverse linear transport modeling (ILTM, Appendix F.5). ILTM provides an estimate for the metabolic fluxes in the OMZ based on the observed chemical concentration profiles. The exact shape of estimated rate profiles depends sensitively on measurement errors and the noise-reduction method applied to the concentration profiles. Hence, ILTM only serves as a rough verification of the order of magnitude of rates predicted by the model or measured experimentally. ILTM
fitting was applied separately to concentration profiles from cruises 47 and 48, as well as to the chemical profiles used for model calibration (cruises 41–44, Fig. 7.2) after averaging across replicates at each depth.

**F.2 Data acquisition**

**F.2.1 Sampling site and time**

Saanich Inlet (SI) a seasonally anoxic fjord on the coast of Vancouver Island, British Columbia Canada has been the site of intensive study for many decades (86, 186). The presence of a shallow glacial entrance sill at 75 m depth limits mixing and ventilation of basin waters below approximately 100 m, resulting in stratification and oxygen depletion during spring and summer (Fig. 7.1A in the main text). Shifts in coastal currents in late summer and fall lead to an influx of denser, oxygenated and nutrient-rich water into the Inlet shoaling anoxic basin waters upward in a process known as deep water renewal (86, 540). Consistent partitioning of the microbial community along the redox cline and similarity to other OMZ microbial communities make Saanich Inlet a model ecosystem for studying the intersection between environmental sequence information and biogeochemical activity along defined redox gradients (503, 531, 540).

The fjord has a maximal depth of 232 m at the sampling site SI03 (123° 30.300' W, 48° 35.500' N).

Sampling is conducted monthly during daylight hours using a combination of 5 and 8 L Niskin bottles and 12 L Go-Flo bottles attached to a nonconducting wire. A Sea-Bird CTD (conductivity, temperature and depth) sensor attached to the bottom of the wire provides depth profiles for temperature, salinity, PAR/Irradiance, conductivity, density, and dissolved oxygen (Sea-Bird Electronics™). Water sampling for multiple chemical and microbial parameters proceeds directly from the bottles in the following order: First, samples are taken for dissolved $O_2$ measurements via Winkler titration, followed by sampling of dissolved gases. Next, samples are taken for RNA, then protein followed by ammonium, hydrogen sulfide and nitrite. Finally, salinity is measured for a subset of depths for CTD calibration, and samples are taken for DNA.

All molecular sequencing was performed using samples collected on February 10, 2010 (cruise SI040_02/08/10). Chemical data were acquired in the same year on January 13 (SI041_01/13/10), February 10 (SI042_02/10/10), March 10 (SI043_03/10/10), April 7
F.2.2 Chemical and physical depth profiles

Temperature, salinity and depth were measured using the CTD sensor described above. The Winkler titration method was used to measure dissolved oxygen (O$_2$) concentrations and calibrate CTD measurements. Samples were collected into Winkler glass Erlenmeyer flasks using latex tubing, overflowing three times to ensure no air contamination, manganese (III) sulphate and potassium iodide were added in succession, inverted to mix and stored at room temperature. Samples were titrated using an automatic titrator. CTD data were processed and manually curated using the Sea-Bird Seasoftware.

Samples for dissolved nutrient (nitrate, nitrite, sulphate and silicate) analyses were collected into 60 mL syringes and filtered through a 0.22 µm Millipore Acrodisc™ into 15 mL falcon tubes. Prior to analysis all samples were stored on ice. Nitrate (NO$_3^-$) samples were stored at $-20^\circ$C in the laboratory, and later analyses carried out using a Bran Luebbe autoanalyser using standard colorimetric methods. For nitrite (NO$_2^-$) analysis, 2 mL of sample water were supplemented with 100 µL sulfanilamide and 100 µL nicotinamide adenine dinucleotide in 4 mL plastic cuvettes. Prepared standards were supplemented with reagents at the same time. Cuvettes were inverted for mixing, and temporarily stored on ice for not more than 4 hrs. Concentration was measured using a Cary60® spectrometer, based on absorbance at 452 nm.

Samples for ammonium (NH$_4^+$) and hydrogen sulphide (H$_2$S) were collected directly from Niskin and GoFlo bottles into 15 mL amber scintillation vials and 15 mL falcon tubes aliquoted with 200 µL 20% zinc acetate respectively. Samples were stored on ice prior to analysis. For NH$_4^+$ analysis, amber vials for standard curve and samples were pre-aliquoted with 7.5 mL O-phthaldialdehyde (OPA) reagent respectively. 5 mL of sample water in triplicate and standard solutions were transferred into OPA pre-aliquoted amber vials. Vials were inverted and stored up to 4 hours. From each standard solution and sample water vial, 300 µL were transferred into a 96 well round bottom plate. Fluorescence at 380$_{ex}$/420$_{em}$ was read using a Varioskan plate reader. For H$_2$S analysis, 300 µL samples were transferred in triplicate to a 96 well plate, and finally Hach Reagent 1 and 2 (6 µL per well) were added. Absorbance at 670 nm was read after 5 min incubation using a Varioskan™ plate reader.

Water for dissolved nitrous oxide (N$_2$O) analysis was collected using Go-flo or Niskin bottles, and was transferred via a Teflon tube into 30 mL or 60 mL borosilicate glass serum vials.
Vials were overflown three times their volume in order to remove any bubbles from the vial or tubing. Vials were subsequently spiked with 50 µL saturated mercuric chloride using a pipette. Vials were then crimp-sealed with a butyl-rubber stopper and aluminum cap, and stored in the dark at 4°C until they were analyzed. Dissolved nitrous-oxide concentrations were measured using a purge-and-trap auto-sampler coupled with a gas-chromatography mass-spectrometer (57).

F.2.3 Metagenomics, metatranscriptomics and metaproteomics

Metagenome and metaproteome datasets were generated using the same methods as described in Hawley et al. (182). Metaproteome sequence coverage was quantified using normalized spectral abundance factors (NSAF) (361). Metatranscriptome samples were filtered in the field onto 0.2 µm sterivex filter with inline pre-filter of 2.7 µm pre-filter, adding 1.8 mL of RNAlater® (Qiagen) and freezing on dry ice before transferring to −80°C. RNA later was removed by washing Sterivex filter with Ringer’s solution before proceeding with cell lysis in the filter cartridge. Total RNA was extracted using the mirVana™ miRNA extraction kit (Ambion), DNA was removed using the TURBO DNA-free™ kit (Ambion) and total RNA was purified using RNeasy™ MiniElute Cleanup Kit (Qiagen). RNA concentration and quality was determined using a Bioanalyzer. Production of cDNA libraries and sequencing was carried out at the Joint Genome Institute using the TruSeq® Stranded Total RNA Sample preparation Guide, including depletion of ribosomal RNA using Ribo-Zero. Assembled metagenomic and metatranscriptomic sequences (contigs) were run through Metapathways (248) for annotation using a combination of RefSeq (470), KEGG (221), COG (469) and MetaCyc (65) databases. Contig coverage was quantified using RPKM values (Appendix F.2.4). KEGG-annotated contigs were assigned to the selected process proxy genes of the model (Supplemental Table F.1); gene coverage at each depth was then quantified by summing all assigned contig RPKM values.

Nitrate reductase (narGHIJ) assigned to planctomycetes showed a decline with depth, suggesting that it may be acting in reverse as a nitrite oxidase (458). In fact, narGHIJ counts affiliated with planctomycetes (narGHIJ-P) dominated all other nxr-associated counts in the metagenomes, metatranscriptomes and metaproteomes. We thus associated nxr with narGHIJ-P. However, because planctomycetes perform anammox in deeper depths (181), we observed a secondary peak in the narGHIJ-P DNA closer to the SNTZ that did not dissipate completely in bottom waters. Given this ambiguity in the interpretation of detected narGHIJ-P genes, we omitted the narGHIJ-P metagenomes and only used the narGHIJ-P
metatranscriptomes and metaproteomes. For more details see Appendix F.4.4.

All nosZ-related protein sequences mapped to a nosZ homolog found in the strictly aerobic *Roseobacter Maritimibacter alkaliphilus* HTCC2654 (264, 477) and showed strong inconsistencies with nosZ metagenomic and metatranscriptomic profiles. nosZ genes have been found to be enriched on particles, likely because they constitute a more anaerobic niche (139). Our metaproteomes were pre-filtered to remove eukaryotes and particles and are expected to be impoverished in nosZ proteins, facilitating a potential masking by related but functionally different proteins. We thus omitted the nosZ metaproteomic data from our analysis.

Table F.1: KEGG orthologous groups (KOG) identified with each gene. The abundance of each gene was the sum of RPKM values (Appendix F.2.4) assigned to all included KOGs.

<table>
<thead>
<tr>
<th>gene or pathway</th>
<th>KOGs</th>
<th>restrictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROM</td>
<td>K12536, K05648</td>
<td>ABC transporters in <em>Pelagibacter</em> and <em>Roseobacter</em></td>
</tr>
<tr>
<td>amo</td>
<td>K10945, K10946</td>
<td></td>
</tr>
<tr>
<td>nxr</td>
<td>K00370, K00371</td>
<td></td>
</tr>
<tr>
<td></td>
<td>K00374, K00373</td>
<td></td>
</tr>
<tr>
<td>hzo</td>
<td>K10535</td>
<td>hao in <em>Planctomycetacea</em></td>
</tr>
<tr>
<td>PDNO(norBC)</td>
<td>K04561, K02305</td>
<td></td>
</tr>
<tr>
<td>nosZ</td>
<td>K00376</td>
<td></td>
</tr>
<tr>
<td>sat</td>
<td>K00958</td>
<td></td>
</tr>
<tr>
<td>aprAB</td>
<td>K00394, K00395</td>
<td></td>
</tr>
<tr>
<td>dsrAB</td>
<td>K11180, K11181</td>
<td></td>
</tr>
<tr>
<td>nirKS</td>
<td>K00368, K15864</td>
<td></td>
</tr>
<tr>
<td>napAB</td>
<td>K02567, K02568</td>
<td></td>
</tr>
<tr>
<td>narGHIJ</td>
<td>K00370, K00371</td>
<td></td>
</tr>
<tr>
<td></td>
<td>K00374, K00373</td>
<td></td>
</tr>
</tbody>
</table>

F.2.4 Quantifying metagenomic and metatranscriptomic data using RPKM

Relative open reading frame (ORF) abundance in the metagenomic and metatranscriptomic datasets was determined for quantitative assessment of pathway coverage. This was achieved by adapting the reads per kilobase per million mapped (RPKM) coverage measure as described by Konwar et al. (247). Briefly, unassembled Illumina reads were mapped to assembled contigs using the short-read aligner BWA-MEM. The resulting SAM file is then inputed
into the MetaPathways v2.5 software (247), which generates an RPKM value per ORF that is extended to an RPKM per pathway via summation. For the case of determining the abundance of pathways expressed in the metatranscriptome relative to those present in the metagenome, the unassembled metatranscriptome reads were mapped back to the assembled metagenome contigs. The RPKM calculation is a simple proportion of the number of reads mapped to a particular section of sequence normalized for ORF length and sequencing depth.

\section*{F.2.5 Process rate measurements}

Rate measurements for anammox and denitrification were carried out as follows: Sample water from each depth was collected anaerobically with sterile nitrile tubing directly into 200 mL glass serum bottles, six per depth, and capped with butyl-rubber stopper and aluminum cap and stored at $10^\circ C$ for approximately 1 hr while collection was completed. The protocol described by Holtappels et al. (192) and briefly outlined here, was then followed. One sample from each depth was bubbled with He for 30 min to decrease concentration of $N_2$. The following substrates were then added: $^{15}NH_4^+$ alone, $^{15}NH_4^+$ and $^{14}NO_2^-$ combined, $^{15}NO_2^-$ alone, $^{15}NO_2^-$ and $^{14}NH_4^+$ combined or $^{15}NO_3^-$ alone. A blank for each depth was also bubbled with He. Sample water was then transferred from the serum bottle into a 12 mL exetainer, capped and stored upside down. Samples in exetainers were then killed with 50 µL saturated HgCl at time intervals of 0 min, 6 hr, 12 hr, 24 hr, 48 hr and 72 hrs. Partial pressures of $^{29}N_2$ and $^{30}N_2$ evolved during the incubations were measured by gas chromatography coupled to isotope ratio mass spectroscopy. Rates of anammox and denitrification were calculated as described by Holtappels et al. (192).

Rate measurements using N isotope methods require a compromise between ensuring detection of labeled tracer elements and avoiding excessive perturbation of ambient substrate concentrations (240, §2.1). Due to the extremely low in-situ substrate levels in some of our samples (Fig. 7.2 in the main text), tracer substrate concentrations in the ex-situ incubator ($25 \mu M NH_4^+$, $2 \mu M NO_2^-$ and $5 \mu M NO_3^-$) significantly exceeded in-situ concentrations. On the other hand, denitrification and anammox-related genes were found throughout the OMZ water column (Fig. 7.3A in the main text). Hence, rates measured in the incubator are only potential rates that likely overestimate actual in-situ rates, especially in substrate-depleted regions far from the SNTZ. For example, Dalsgaard et al. (90) reports a 2–4 fold increase of anammox rates following the addition of $10 \mu M NH_4^+$ in anoxic water column experiments. Similarly, Wenk et al. (515) found high potential denitrification rates in nitrate-depleted regions of a meromictic lake. We thus corrected our rate measurements for differences between
in-situ and incubator substrate concentrations, as described below.

The simplest approach would be to multiply measured rates with the ratios of in-situ over ex-situ substrate concentrations, as has been done in previous ex-situ incubation experiments (508). However, such a linear rescaling implicitly assumes that substrate half-saturation constants are much higher than both the in-situ as well as ex-situ concentrations, an assumption that may not be justifiable in regularly substrate-depleted natural environments. For example, members of the Scalindua candidate clade, which is well represented in Saanich Inlet (181), exhibit nitrite half-saturation constants as low as \(0.45 \mu M\) (23). To avoid an implicit assumption of 1st order kinetics, and for consistency with the assumptions of our model, we corrected our rates using Michaelis-Menten kinetic curves (Appendix F.3.4) with the same half-saturation constants as used in our model (Appendix F.3.7). Specifically, if \(R^*_{\text{hzo}}(z)\) is the measured ex-situ (i.e., potential) anammox rate at some particular depth, then the corrected in-situ rate was assumed to be

\[
R_{\text{hzo}} = R^*_{\text{hzo}}(z) \cdot \frac{[\text{NH}_4^+]}{K_{\text{NH}_4^+}} \cdot \frac{[\text{NO}_2^-]}{K_{\text{NO}_2^-}} \cdot \frac{[\text{NH}_4^+]^*}{[\text{NO}_2^-]^*} 
\]

(F.2.1)

Here, \(K_{\text{NH}_4^+}\) and \(K_{\text{NO}_2^-}\) are anammox half-saturation constants for \(\text{NH}_4^+\) and \(\text{NO}_2^-\), respectively (Appendix F.3.7), \([\text{NH}_4^+]\) and \([\text{NO}_2^-]\) are the corresponding measured in-situ concentrations and \([\text{NH}_4^+]^*\) and \([\text{NO}_2^-]^*\) are the concentrations in the incubator at the beginning of the experiment, i.e., \([\text{NH}_4^+]^* = [\text{NH}_4^+] + 25 \mu M\) and \([\text{NO}_2^-]^* = [\text{NO}_2^-] + 2 \mu M\). Measured denitrification rates were corrected in a similar way to account for differences in \(\text{NO}_3^-\) concentrations.

**F.2.6 qPCR quantification of SUP05 cell counts**

All metagenomic, metatranscriptomic and metaproteomic profiles presented here only provide relative — rather than absolute — biomolecule abundances. This remains the de facto standard for multi-omic data sets, owing largely to methodological challenges involved in absolute DNA, mRNA and protein quantification (but see Smets et al. (428) for recent advancements). As we explain below (section F.3.9), multi-omic depth profiles were linearly rescaled to facilitate comparison with our model predictions — expressed in absolute gene counts, however this comes at the cost of additional rescaling parameters.

In order to perform an independent validation of modeled gene concentrations, we compared
the predicted PDNO gene concentrations to independent cell-count estimates for SUP05 (the dominant nitrate reducer in Saanich Inlet; 181), obtained through quantitative polymerase chain reaction (qPCR). qPCR quantification of SUP05 cell counts was performed for water samples collected at 8 distinct depths from the same location and time as for multi-omic sequencing (Fig. 7.3A in the main text). Water samples (volume ~ 1L) were filtered in the field onto 0.2 µm sterivex filters. Samples were not pre-filtered in order to obtain an accurate estimate of total in-situ SUP05 cell counts. We used a custom SUP05-specific primer set (Ba519F–1048R) to amplify the 519–1048 region of the SUP05 16S rRNA gene, and followed the protocol described by Hawley et al. (182) to estimate the starting template concentration. qPCR was performed in triplicate for each sample. We multiplied the average template concentration for each sample by the volume of extracted fluid (∼ 200 – 400 µL), divided by the volume of filtered seawater, to obtain an estimate for the concentration of SUP05 16S gene copies in seawater. To correct for multiple 16S gene copies in single cells, we divided this concentration by the 16S gene copy number (3.767), estimated for members of the SUP05 clade based on closely related fully sequenced reference genomes. Specifically, we used the 16S gene copy number assigned by the Tax4Fun pipeline (19) to the clade “Oceanospirillales;SUP05 cluster;uncultured gamma proteobacterium” in the SILVA 123 database (378). Note that Tax4Fun (19) uses a probabilistic model to assign multiple reference genomes with varying weights to each clade in the SILVA database. Hence, the effective 16S gene copy number assigned by Tax4Fun to each clade is the weighted harmonic mean of the 16S gene copy numbers in each reference genome assigned to that clade.

F.3 Mathematical model

F.3.1 Overview

The gene-centric model describes the spatiotemporal dynamics of 8 metabolite concentrations and 6 gene (DNA) concentrations along the Saanich Inlet water column between depths 100–200 m. Each gene is a proxy for a particular redox pathway that couples the oxidation of an external electron donor to the reduction of an external electron acceptor (Appendix F.3.2). The model assumes that each cell occupies a single metabolic niche, associated with one of the modeled pathways and thus one of the considered proxy genes. Reaction rates (per gene) depend on the concentrations of all used metabolites according to 1st order or 2nd order (Michaelis-Menten) kinetics (211, 386) (Appendix F.3.4). In turn, the production or depletion of metabolites at any depth is determined by the reaction rates at that depth, taking into account reaction stoichiometry (Appendix F.3.3). The production of genes (or
more precisely, their host cells) at any depth is driven by the release of energy from their catalyzed reactions, and is proportional to the Gibbs free energy multiplied by the reaction rate (Appendix F.3.5) (396). In addition, genes are subject to exponential decay as well as eddy-diffusion and sinking. Metabolites are also subject to eddy-diffusion.

Mathematically, the model is defined as a set of partial differential equations (PDE) for the gene and metabolite concentrations across time and depth. More precisely, the DNA concentration of the $r$-th gene ($\Gamma_r$, copies per volume) at any a given depth $z$ changes according to

$$\frac{\partial \Gamma_r}{\partial t} = -q_r \Gamma_r + \frac{1}{c} Z_r H_r \Gamma_r - v \frac{\partial \Gamma_r}{\partial z} + \frac{\partial}{\partial z} \left( K(z) \frac{\partial}{\partial z} \Gamma_r \right), \quad (F.3.1)$$

and the concentration of the $m$-th metabolite ($C_m$, mole per volume) changes according to

$$\frac{\partial C_m}{\partial t} = \sum_r S_{mr} H_r \Gamma_r + \frac{\partial}{\partial z} \left( K(z) \frac{\partial C_m}{\partial z} \right). \quad (F.3.2)$$

Both the DNA concentrations $\Gamma_r$ and metabolite concentrations $C_m$ depend on time $t$ and depth $z$. The first term in equation (F.3.1) corresponds to cell death, with $q_r$ being the exponential death rate for cells hosting gene $r$ in the absence of any metabolites. The 2nd term in (F.3.1) corresponds to gene production proportional to the per-gene reaction rate $H_r$ (which in turn depends on metabolite concentrations, see Appendix F.3.4). The biomass production coefficient $Z_r$ is a linear function of the Gibbs free energy of the reaction catalyzed by gene $r$ and depends on the reaction quotient (Appendix F.3.5). In particular, $Z_r$ increases when product concentrations are low and decreases when substrate concentrations are low. $c$ is the average dry cell mass, which is used to convert biomass production into cell production. The 3rd term in equation (F.3.1) corresponds to cell sinking at a constant speed $v$. The last term in equation (F.3.1) and equation (F.3.2) corresponds to diffusive transport (263), with $K$ being the vertical eddy-diffusion coefficient. The 1st term in equation (F.3.2) corresponds to production or depletion of metabolites due to microbial metabolism. Reaction rates are transformed into metabolite fluxes via the stoichiometric matrix $S$, with entry $S_{mr}$ corresponding to the stoichiometric coefficient of metabolite $m$ in reaction $r$ (Appendix F.3.3).

The differential equations (F.3.1) and (F.3.2) give the rate of change of each metabolite and gene profile, if the profiles are known at a given moment in time. Once all boundary conditions (Appendix F.3.6), model parameters (Appendix F.3.7) and initial profiles are specified, the model predicts the profiles at any future time point. Steady state profiles were
obtained by running simulations of the model until convergence to equilibrium. Because the predicted profiles depend on model parameters, parameters can be calibrated such that the predicted steady state profiles best reproduce the measured data. We used chemical depth profiles to fit poorly known model parameters, thus obtaining a model calibrated to Saanich Inlet’s OMZ (Appendix F.3.8). This calibrated model was then used to make predictions about steady state DNA profiles, which were compared to measured metagenomic profiles (sections F.2.3 and F.3.9). This comparison, described in the main text, serves as a test of the model’s ability to explain metagenomic profiles in Saanich Inlet’s OMZ. Reaction rates at each depth are automatically calculated using the kinetics described in Appendix F.3.4.

**F.3.2 Considered pathways**

The model considers key dissimilatory redox pathways involved in nitrogen and sulfur cycling. When comparing model predictions to molecular data, each pathway was represented by a single gene. For example, nitrous oxide reduction (nosZ gene) coupled to hydrogen sulfide oxidation (dsr, apr and sat genes) is formally represented by nosZ. Other pathways considered by the model were aerobic ammonium oxidation (amo), aerobic nitrite oxidation (nxr), partial denitrification of nitrate to nitrous oxide (PDNO) coupled to sulfide oxidation, anammox (hzo) and remineralization of organic matter via aerobic respiration (ROM). PDNO comprises 3 denitrification steps: Reduction of nitrate to nitrite (narGHIJ or napAB genes), reduction of nitrite to nitric oxide (nirKS genes) and reduction of nitric oxide to nitrous oxide (norBC genes), all three of which are suspected to be predominantly performed by SUP05 γ-proteobacteria (181, 504). The first denitrification step was assumed to be leaky, so that a small fraction of nitrite was released into the extracellular environment (253). PDNO was represented by norBC genes when comparing the model to molecular data (Fig. 7.3A in the main text, but see Figs. F.1D,E,F for narGHIJ, napAB and nirKS multimolecular data). Aerobic ammonium oxidation included a weak production of nitrous oxide (nitrifier denitrification (407)), although the inclusion of this process did not noticeably affect model predictions because most of the nitrous oxide was produced by PDNO. Aerobic respiration of organic matter included the release of ammonium and sulfate at ratios adjusted to measured C:N:S ratios for marine bacterial biomass (115). The choice of pathways follows the hypotheses made by Hawley et al. (181) based on metagenomic and metaproteomic depth profiles, as well as reports of nitrous oxide reduction in Saanich Inlet’s OMZ (74). Hydrogen sulfide is assumed to originate from the sediments via diffusion, where high rates of sulfate reduction have been observed (4, 101) (Appendix F.4.1 for a discussion of this assumption). Figure 7.1A in the main text gives an overview of the described reaction
network. The detailed reaction stoichiometry is given in section F.3.3.

**F.3.3 Pathway stoichiometry**

We list the stoichiometry of the dissimilatory redox pathways considered by the model:

- Remineralization of organic matter through aerobic respiration:

\[
\frac{1}{6}\text{POM} + \text{O}_2 \xrightarrow{\text{ROM}} \text{CO}_2 + \text{H}_2\text{O} + \nu\text{NH}_4^+ + \sigma\text{SO}_4^{2-} \tag{F.3.3}
\]

where POM corresponds to

\[
(C_6H_{12}O_6)(\text{NH}_4^+)_{6\nu}(\text{SO}_4^{2-})_{6\sigma} \tag{F.3.4}
\]

and

\[
1 : \nu : \sigma = 1 : 0.184 : 0.0113 \tag{F.3.5}
\]

correspond to typical molar C : N : S ratios in marine bacterial biomass (115).

- Aerobic ammonium oxidation:

\[
\text{NH}_4^+ + \frac{1}{2}(3 - \text{amo}) \times \text{O}_2 \xrightarrow{\text{amo}} (1 + \text{amo}/2) \times \text{H}_2\text{O} + (1 - \text{amo}) \times \text{NO}_2^- + (\text{amo}/2) \times \text{N}_2\text{O} + (2 - \text{amo}) \times \text{H}^+. \tag{F.3.6}
\]

Here \(\text{amo}\) is a parameter representing the fraction of N released as \(\text{N}_2\text{O}\) via nitrifier denitrification, compared to the total \(\text{NH}_4^+\) consumed (407). For example, if \(\text{amo} = 0\), then ammonium is completely oxidized and released as nitrite.

- Aerobic nitrite oxidation:

\[
2\text{NO}_2^- + \text{O}_2 \xrightarrow{\text{nxr}} 2\text{NO}_3^- . \tag{F.3.7}
\]

- Anaerobic ammonium oxidation (anammox):

\[
\text{NH}_4^+ + \text{NO}_2^- \xrightarrow{\text{hzo}} \text{N}_2 + 2\text{H}_2\text{O}. \tag{F.3.8}
\]
• Partial denitrification to nitrous oxide (PDNO) coupled to hydrogen sulfide oxidation:

\[
(1 - L_{PDNO}/2) \times H_2S + 2NO_3^{-}_{PDNO} \rightarrow (1 - L_{PDNO}) \times N_2O + 2L_{PDNO} \times NO_2^- \\
+ (1 - L_{PDNO}/2) \times SO_4^{2-} + (1 - L_{PDNO}) \times H_2O + L_{PDNO} \times H^+.
\]  

(F.3.9)

Here, \( L_{PDNO} \) is a parameter representing the fraction of \( NO_2^- \) leaked to the extracellular medium during PDNO, compared to the total \( NO_3^- \) consumed (253).

• Nitrous oxide reduction coupled to hydrogen sulfide oxidation:

\[
H_2S + 4N_2O \xrightarrow{\text{nosZ}} 4N_2 + SO_4^{2-} + 2H^+.
\]  

(F.3.10)

• Nitrate reduction to ammonium (DNRA, identified with the \textit{nirBD} gene):

\[
H_2S + NO_3^- + H_2O \xrightarrow{\text{DNRA}} NH_4^+ + SO_4^{2-}.
\]  

(F.3.11)

DNRA was eventually omitted from the model for reasons described in Appendix F.4.2.

F.3.4 Reaction kinetics

Respiration of organic matter involves the hydrolysis of particulate organic matter (POM) to dissolved organic matter (DOM), which is subsequently broken down to simpler organic molecules by fermenters that provide non-fermenting organotrophs with a reactive DOM pool. However, reactive DOM rarely accumulates and most of the DOM pool is expected to be refractory (350). Furthermore, POM degradation has been found to be strongly correlated to bacterial growth in subeutrophic zones, likely due to limiting POM hydrolysis rates (194). We thus modeled organic matter respiration rates as a first-order function of particulate organic carbon (POC) concentrations (323). More precisely, the gene-specific ROM reaction rate, \( H_{ROM} \), is a function of metabolite concentrations \( C \) given by

\[
H_{ROM}(C) = A_{ROM}F_T \times \frac{C_{POM}C_{O_2}}{C_{O_2} + K_{ROM,O_2}},
\]  

(F.3.12)

where \( K_{ROM,O_2} \) is the oxygen half-saturation constant, \( A_{ROM} \) is a first-order rate constant (“affinity”) and \( F_T \) is the unitless thermodynamic potential factor given by Reed et al. (386) (equation S1)
Half-saturation constants reported for nitrous oxide oxidation are typically on the order of 0.37 – 2.5 µM N₂O (414, 519) and 40 µM H₂S (212), which are well above the typical N₂O and H₂S concentrations in the Saanich Inlet OMZ (Fig. 7.2 in the main text). Sulfide-driven nitrous oxide reduction in Saanich Inlet is therefore likely limited both by electron donor as well as electron acceptor availability. We thus modeled nitrous oxide reduction using first order substrate kinetics with oxygen inhibition:

\[ H_{\text{nosZ}}(C) = A_{\text{nosZ}}F_T \times \frac{C_{\text{N}_2\text{O}}C_{\text{H}_2\text{S}}K_{\text{nosZ,O}_2}}{C_{\text{O}_2} + K_{\text{nosZ,O}_2}}, \]  

where \( K_{\text{nosZ,O}_2} \) is the oxygen half-inhibition constant and \( A_{\text{nosZ}} \) is a first-order rate constant.

All other gene-specific reaction rates (\( H_r \)) are modeled using Michaelis-Menten kinetics with possible inhibition (211, 386):

\[ H_r(C) = V_rF_T \times \prod_{\text{m reactant of reaction } r} \frac{C_m}{K_{rm} + C_m} \times \prod_{\text{n inhibitor of reaction } r} \frac{K_{rn}^*}{K_{rn}^* + C_n}. \]  

(F.3.14)

Here, \( V_r \) is the maximum gene-specific reaction rate and \( K_{rm} \) and \( K_{rn}^* \) are half-saturation and half-inhibition constants, respectively, given in Appendix F.3.7. The only explicitly modeled inhibition was oxygen inhibition for anammox (\( hzo \)), PDNO and nitrous oxide reduction (\( nosZ \)).

**F.3.5 Gibbs free energy and gene growth**

Following Roden and Jin (396) and Reed et al. (386), we set

\[ Z_r = 2.08\gamma_r^e - 0.0211\Delta G_r, \]  

(F.3.15)

(in g biomass per mole reaction flux) where \( \gamma_r^e \) is the negative stoichiometric coefficient of the electron donor in the reaction,

\[ \Delta G_r = \Delta G_r^o + R_gT \ln Q_r \]  

(F.3.16)
is the Gibbs free energy of the reaction (in kJ per mol), \( \Delta G^o_r \) is the standard Gibbs free energy of the reaction and

\[
Q_r = \prod_m C_m^{S_mr}
\]  \hspace{1cm} (F.3.17)

is the reaction quotient (96). Each \( \Delta G^o_r \) depends on the local temperature and pressure and was calculated using the CHNOSZ R package (103).

### F.3.6 Boundary conditions

Uniquely solving the partial differential equations (F.3.1) and (F.3.2) requires appropriate boundary conditions (BC) for all genes and metabolites at the top and bottom boundaries (100 m and 200 m, respectively). For all metabolites except N\(_2\), N\(_2\)O, SO\(_4^2-\) and O\(_2\), BCs were fixed values set to the average measurements from cruises 41 (SI041_01/13/10), 42 (SI042_02/10/10) and 43 (SI043_03/10/10). For N\(_2\) and N\(_2\)O, lower BCs were set to Neumann (zero flux). For O\(_2\), we used Dirichlet BCs (fixed value) with values equal to the average measurements from cruise 42 and 44 (SI044_04/07/10), because O\(_2\) data were unavailable for cruises 41 and 43. For SO\(_4^2-\) we used Dirichlet BCs set to 28 mM on both sides (323). Metabolite boundary conditions are summarized in Table F.1. These boundary conditions result in a net oxygen and nitrate influx from the top as well as an ammonium and sulfide influx from the sediments (4, 100, 340).

Gene boundary conditions were either set to fixed zero (hzo and norBC top BCs, ROM, amo and nxr bottom BCs) or to fixed relative gradients (ROM, amo, nxr, nirBD and nosZ top BCs, hzo, nirBD, norBC and nosZ bottom BCs), with the relative gradient inferred from the metagenomic profiles.

### F.3.7 Model parameterization

Half-saturation and half-inhibition constants for all involved pathways are listed in Table F.2. Maximum cell-specific reaction rates were set to \( V_{amo} = 1.23 \times 10^{-13} \) mol/(cell \cdot d) (389), \( V_{nxr} = 3.26 \times 10^{-13} \) mol/(cell \cdot d) (389) and \( V_{hzo} = 2 \times 10^{-14} \) mol/(cell \cdot d) (251, 457).

The nitrifier denitrification fraction \( L_{amo} \) was set to \( 10^{-4} \), according to nitrifier denitrification fractions of marine ammonium oxidizing archaea measured by Santoro et al. (407, Fig 2) over varying NO\(_{3}^-\) concentrations, and the fact that in Saanich Inlet NO\(_{3}^-\) concentrations are typically below 2 \( \mu \)M (Fig. 7.2 in the main text). Because of a lack of reliable information,
the rate constants $A_{\text{ROM}}$, $V_{\text{PDNO}}$ and $A_{\text{nosZ}}$, as well as the PDNO leakage fraction $L_{\text{PDNO}}$, were calibrated to chemical profiles as described in Appendix F.3.8 and in the main text. Calibration yielded $A_{\text{ROM}} = 5.11 \times 10^{-9}$ L/(cell·d), $V_{\text{PDNO}} = 2.18 \times 10^{-14}$ mol/(cell·d), $A_{\text{nosZ}} = 0.098$ L/(cell·d) and $L_{\text{PDNO}} = 0.352$. An overview of fixed and calibrated reaction-kinetic parameters is provided in Table F.2. The sensitivity of the model to parameter variation is illustrated in Appendix F.3.12.

The dry cell mass was assumed to be $c = 5 \times 10^{-13}$ g, for consistency with the mass used by Roden and Jin (396) to obtain the regression formula (F.3.15). Cell death rates were set to $q_{\text{ROM}} = 0.063$ d$^{-1}$ in accordance with turnover times estimated by Whitman et al. (518) for marine prokaryotic heterotrophs above 200 m; to $q_{\text{amo}} = 0.024$ d$^{-1}$ in accordance with average values reported for ammonium oxidizing bacteria (143); to $q_{\text{nxr}} = 0.054$ d$^{-1}$ corresponding to values estimated for nitrite oxidizers (175) and to 0.0033 d$^{-1}$ for all other genes, in accordance with turnover times estimated by Whitman et al. (518) for marine prokaryotes below 200 m.

The concentration of H$^+$ was fixed to 8.5 nM, corresponding to pH$= 8.07$ (384). The total dissolved inorganic carbon (DIC) was fixed to 2141 μM, corresponding to a surface DIC of 2180 μmol/kg (528) and a surface water density of 1018 kg/m$^3$. Accordingly, the dissolved CO$_2$ concentration was fixed at 28 μM according to aquatic carbonate equilibrium at the given pH and DIC (541). The POC profile was calculated from data reported for February 2011 by Luo et al. (290) and POM was set to $(1/6) \times \text{POC}$ (Fig. F.1C in the Appendix). Fixing the POM profile circumvents poorly understood physical processes contributing to
organic matter fluxes in Saanich Inlet (290). CO₂, H⁺ and POM concentrations, while fixed, were still included in the reaction quotients (Appendix F.3.5) as well as the reaction-kinetics (Appendix F.3.4).

The diapycnal eddy diffusion coefficient \( K \) was set to \( N^{-2} \cdot 3.7 \times 10^{-10} \text{ W} \cdot \text{kg}^{-1} \), where \( N \) is the buoyancy frequency (120, 142). The latter was calculated using temperature and salinity profiles from January 13, 2010, using the \texttt{oce} R package (228) (Fig. F.1 in the supplement) after loess-smoothing temperature at degree 2 and salinity at degree 1, with a span of 75%. We chose this time point because the two subsequent temperature and salinity measurements (February 10th and March 10th) were unreliable due to technical problems with our CTD. The cell sinking speed \( v \) was set to 0.1 m/day, in accordance with previous marine microbial ecological models (26, 121).
Table F.2: Reaction-kinetic parameters used in the gene-centric model, either calibrated or taken from the literature: Half-saturation substrate concentrations ($K$), half-inhibition concentrations ($K^\star$), cell-specific maximum rates for 2nd order kinetics ($V$), 1st order kinetic constants ($A$, “affinities”), nitrifier denitrification fraction ($L_{amo}$) and PDNO leakage fraction ($L_{PDNO}$). The exact role of each parameter is explained in Appendix F.3.4. Additional (non-kinetic) fixed model parameters are provided in Appendix F.3.7. Clades with members that have been found active in the Saanich Inlet OMZ (181) are marked with a “†”.

<table>
<thead>
<tr>
<th>reaction</th>
<th>parameter</th>
<th>value</th>
<th>units</th>
<th>organism/region</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROM</td>
<td>$K_O_2$</td>
<td>0.121</td>
<td>$\mu$M</td>
<td><em>Escherichia coli</em></td>
<td>(452)</td>
</tr>
<tr>
<td></td>
<td>$A$</td>
<td>5.11</td>
<td>nL/(cell · d)</td>
<td>calibr.</td>
<td></td>
</tr>
<tr>
<td>amo</td>
<td>$K_{NH_4^\star}$</td>
<td>0.133</td>
<td>$\mu$M</td>
<td><em>Ca. Nitrosopumilus maritimus</em></td>
<td>(302)</td>
</tr>
<tr>
<td></td>
<td>$K_O_2$</td>
<td>3.91</td>
<td>$\mu$M</td>
<td><em>Ca. Nitrosopumilus maritimus</em></td>
<td>(277)</td>
</tr>
<tr>
<td></td>
<td>$V$</td>
<td>123</td>
<td>fmol/(cell · d)</td>
<td><em>Nitrosomonas spp.</em></td>
<td>(389)</td>
</tr>
<tr>
<td></td>
<td>$L_{amo}$</td>
<td>$10^{-4}$</td>
<td>–</td>
<td>marine ammonia oxidizing archaea</td>
<td>(407)</td>
</tr>
<tr>
<td>nxr</td>
<td>$K_{NO_2^-}$</td>
<td>11.7</td>
<td>$\mu$M</td>
<td><em>Nitrospira spp.</em></td>
<td>(37)</td>
</tr>
<tr>
<td></td>
<td>$K_O_2$</td>
<td>0.78</td>
<td>$\mu$M</td>
<td>Chilean OMZ</td>
<td>(47)</td>
</tr>
<tr>
<td></td>
<td>$V$</td>
<td>326</td>
<td>fmol/(cell · d)</td>
<td><em>Nitrobacter</em> sp.</td>
<td>(389)</td>
</tr>
<tr>
<td>hzo</td>
<td>$K_{NH_4^\star}$</td>
<td>3</td>
<td>$\mu$M</td>
<td><em>Ca. Scalindua sp.</em></td>
<td>(23)</td>
</tr>
<tr>
<td></td>
<td>$K_{NO_2^-}$</td>
<td>0.45</td>
<td>$\mu$M</td>
<td><em>Ca. Scalindua sp.</em></td>
<td>(23)</td>
</tr>
<tr>
<td></td>
<td>$K^\star_O_2$</td>
<td>0.2</td>
<td>$\mu$M</td>
<td>Peruvian OMZ</td>
<td>(219, 386)</td>
</tr>
<tr>
<td></td>
<td>$V$</td>
<td>20</td>
<td>fmol/(cell · d)</td>
<td><em>Planctomycetales</em></td>
<td>(251, 457)</td>
</tr>
<tr>
<td>PDNO</td>
<td>$K_{NO_3^-}$</td>
<td>2.9</td>
<td>$\mu$M</td>
<td>marine anoxic basin</td>
<td>(206)</td>
</tr>
<tr>
<td></td>
<td>$K_{H_2S}$</td>
<td>2</td>
<td>$\mu$M</td>
<td>Saanich Inlet OMZ</td>
<td>(40)</td>
</tr>
<tr>
<td></td>
<td>$K^\star_O_2$</td>
<td>0.1</td>
<td>$\mu$M</td>
<td>Eastern South Pacific OMZ</td>
<td>(26, 472)</td>
</tr>
<tr>
<td></td>
<td>$V$</td>
<td>21.8</td>
<td>fmol/(cell · d)</td>
<td>calibr$^1$.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$L_{PDNO}$</td>
<td>35.2</td>
<td>%</td>
<td>calibr$^2$.</td>
<td></td>
</tr>
<tr>
<td>nosZ</td>
<td>$K^\star_O_2$</td>
<td>0.971</td>
<td>$\mu$M</td>
<td>low-oxygen activated sludge</td>
<td>(498)</td>
</tr>
<tr>
<td></td>
<td>$A$</td>
<td>0.098</td>
<td>L/(cell · d)</td>
<td>calibr.</td>
<td></td>
</tr>
</tbody>
</table>

$^1$ Frey et al. (131) reports cell-specific thiosulphate-driven denitrification rates for *Sulfurimonas gotlandica* in the range $24.2 - 74.3$ fmol/(cell · d).

$^2$ Reported fractions of nitrite leakage during incomplete denitrification ($L_{PDNO}$) range from 0% to 87% (8, 35, 351).
F.3.8 Calibrating reaction-kinetic parameters to data

As described in the previous section, most model parameters were obtained from the literature, however a subset of reaction-kinetic parameters ($A_{\text{ROM}}, V_{\text{PDNO}}, L_{\text{PDNO}}$ and $A_{\text{nosZ}}$; overview in Table F.2) had to be calibrated due to the lack of available information. Here we describe the statistical methods used to calibrate unknown reaction-kinetic model parameters to available chemical depth profile data. The steady state solution of the model defines a mapping from a given choice of parameter values (collectively written as a vector $p$) to predicted depth profiles for metabolite concentrations, $C_1, C_2, \ldots$. We assumed that measured concentrations ($\tilde{C}_1, \tilde{C}_2, \ldots$) are normally distributed:

$$\tilde{C}_i = C_i + \sigma_i \cdot \varepsilon_i.$$

(F.3.18)

Here, $\varepsilon_i$ is a standard-normally distributed error and $\sigma_i$ is the (unknown) standard deviation of measurement errors (henceforth referred to as error scale). We allowed for a different $\sigma_i$ for each metabolite to account for variations in the magnitude of measurement errors.

In the context of our spatial model, the concentrations $C_i$ are predicted as functions of depth, $z$, i.e., $C_i = C_i(z; p)$. Calibration data is given as tuples $(z_{ij}, \tilde{C}_{ij})$, where each $\tilde{C}_{ij}$ is a measurement of the $i$-th concentration at some depth $z_{ij}$ and $j$ enumerates all measurements.
Chapter 7: Supplemental material

of the \(i\)-th concentration. The overall log-likelihood function for such a data set is given by

\[
l(\sigma, p) = - \sum_{i,j} \ln \left( \sigma_i \sqrt{2\pi} \right)
- \sum_{i,j} \frac{1}{2\sigma_i^2} \left[ \tilde{C}_{ij} - C_i(z_{ij}; p) \right]^2.
\]  

(F.3.19)

The model was calibrated by maximizing the log-likelihood \(l(\sigma, p)\) by choice of the error scales \(\sigma_i\) and the parameter values \(p\). Maximization of the log-likelihood was performed using the MATLAB\textsuperscript{®} function \texttt{fmincon}, which uses repeated simulations and gradual exploration of parameter space (309). The following chemical concentration data were used for calibration: \(\text{NH}_4^+, \text{NO}_3^-, \text{NO}_2^-, \text{N}_2\text{O}\) and \(\text{H}_2\text{S}\) from cruises 41–43, and \(\text{O}_2\) from cruises 42 and 44.

F.3.9 Calibrating multi-omic data units

Metagenomic, metatranscriptomic and metaproteomic data are given only in relative units. For example, the correspondence between metagenomic RPKM values and actual DNA concentrations in the water column is, a priori, unknown. In fact, RPKM values for different genes may correspond to different gene concentrations due to detection biases (33, 331, 369). Furthermore, model predictions regarding RNA and protein abundances are in arbitrary units because the transcriptional, translational and enzymatic efficiency of proteins is unknown and differs between proteins.

In order to compare model predictions to multi-omic sequence data, we assumed that each measured DNA, mRNA and protein abundance profile is related to the corresponding model prediction by a constant linear conversion factor. Conversion factors were estimated via maximum-likelihood estimation, separately for each molecule to account for detection biases. More precisely, for each data set we assumed a normal error distribution as already described in Appendix F.3.8. Hence, measured environmental biomolecule concentrations, for example \(\text{amo}\) DNA concentrations, are distributed as

\[
\tilde{\Gamma}_i = \Gamma_i / \beta_i + \sigma_i \cdot \varepsilon_i,
\]  

(F.3.20)

where \(\varepsilon_i\) are uncorrelated standard-normally distributed errors, scaled by an unknown factor \(\sigma_i\), and \(\beta_i\) is the unknown proportionality factor between \(\text{amo}\) metagenomic RPKM values \(\tilde{\Gamma}_i\) and actual DNA concentrations. The log-likelihood of a measured depth profile comprising
$N_i$ data points, $(z_{i1}, \Gamma_{i1}), \ldots, (z_{iN_i}, \Gamma_{iN_i})$, is thus

$$l_i(\sigma_i, p) = - \sum_{j=1}^{N_i} \ln \left( \sigma_i \sqrt{2\pi} \right) - \sum_{j=1}^{N_i} \frac{1}{2\sigma_i^2} \left[ \Gamma_{ij} - \Gamma_i(z_{ij}; p)/\beta_i \right]^2. \quad (F.3.21)$$

For any fixed model parameter choice $p$ (and therefore fixed predictions $\Gamma_i$), the log-likelihood $l_i(\sigma_i; p)$ is maximized by choosing

$$\beta_i = \sqrt[N_i]{\frac{N_i}{\prod_{j=1}^{N_i} \frac{\Gamma_i(z_{ij}; p)}{\Gamma_{ij}}}}, \quad (F.3.22)$$

(i.e., the geometric mean of model predictions over measurements) and

$$\sigma_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} \left[ \Gamma_{ij} - \Gamma_j(z_{ij}; p)/\beta_i \right]^2. \quad (F.3.23)$$

Choosing $\beta_i$ as in equation (F.3.22) yields maximum-likelihood estimates for the appropriate conversion factors between metagenomic units (RPKM) and actual DNA concentrations (genes/L) (see table F.3 in the supplement). Inserting the estimated $\beta_i$ and $\sigma_i$ back into equation (F.3.21) yields the log-likelihood of the particular metagenomics data set and for a particular choice of model parameters $p$. A similar approach was used to compare metatranscriptomic and metaproteomic data sets to model predictions (Appendix F.3.10).

The estimated proportionality factors $\beta_i$ are listed in table F.3 of the supplement, and range from $3.9 \times 10^4$ genes · L$^{-1}$ · RPKM$^{-1}$ for norBC up to $3.3 \times 10^7$ genes · L$^{-1}$ · RPKM$^{-1}$ for ROM. These differences may be due to variable DNA extraction efficiencies across cells, uneven community sampling due to filter-size partitioning (139) or differences in gene copy numbers per cell. Additionally, the assumption of a common cell mass for all modeled genes may have resulted in an inaccurate conversion of predicted biomass production to gene production. However, the good overall agreement between predicted functional gene concentrations and flow cytometry cell counts (Fig. 7.3 in the main text) suggests that this may only be a minor problem.

### F.3.10 Predicting metatranscriptomic and metaproteomic profiles

A priori, the gene-centric model makes no predictions regarding mRNA or protein dynamics; in fact transcription and translation are circumvented by assuming that the release of energy manifests directly as DNA replication. To explore the possibility of explaining mRNA and
Chapter 7: Supplemental material

Table F.3: Proportionality factors ($\beta$) between environmental gene abundances and metagenomic RPKM values (in genes·L$^{-1}$·RPKM$^{-1}$), as defined in Appendix F.3.9. Estimated by comparing the predictions of the calibrated model with metagenomic data from February 10, 2010. Unambiguous metagenomic data was not available for nxr (see Appendix F.4.4).

<table>
<thead>
<tr>
<th>gene</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROM</td>
<td>$4.1 \times 10^7$</td>
</tr>
<tr>
<td>amo</td>
<td>$1.0 \times 10^6$</td>
</tr>
<tr>
<td>nxr</td>
<td>NA</td>
</tr>
<tr>
<td>hzo</td>
<td>$3.2 \times 10^5$</td>
</tr>
<tr>
<td>norBC</td>
<td>$3.4 \times 10^4$</td>
</tr>
<tr>
<td>nosZ</td>
<td>$4.5 \times 10^4$</td>
</tr>
</tbody>
</table>

Protein distributions in Saanich Inlet’s OMZ, we extended the model to a set of hypothetical mechanisms driving the production, decay and dispersal of these molecules. More precisely, we assumed that mRNA and protein production rate at a particular depth is proportional to the total reaction rate at that depth ($H_r\Gamma_r$), and that mRNA and proteins disperse similarly to genes (Appendix F.3.10). The assumption that mRNA and protein production rates are proportional to reaction rates is motivated by observations of a positive relation between transcription and translation rates and metabolic activity or growth (9, 154, 229). A linear relation, in particular, may be justified by the fact that increased enzyme dilution rates at elevated cell growth must be balanced (at the population level) by correspondingly increased translation (and hence transcription) rates (403).

This simple description introduces two unknown parameters per mRNA or protein: The proportionality factor that converts reaction rates to molecule production rates, and the decay time of molecules following production. We calibrated both parameters using metatranscriptomic and metaproteomic depth profiles and then checked how well the latter could be reproduced. Our methodology is described for mRNA in detail below. Protein dynamics were modeled and compared to metaproteomic data in a similar way.

As mentioned, our first assumption was that the mRNA production rate (transcripts produced per time and per volume of seawater) at a particular depth is proportional to the total reaction rate (mol per time and per volume of seawater) at that depth. We also assumed that mRNA molecules disperse via diffusion and sinking similarly to genes, as they are hosted by the same cells. Thus, environmental mRNA concentrations satisfy the partial differential
equation

\[ \partial_t T_r = -\frac{T_r}{\tau_r} + \frac{R_r}{\alpha_r} - v \partial_z T_r + \partial_z K \partial_z T_r, \]  

(F.3.24)

where \( T_r(t, z) \) is the mRNA concentration corresponding to the \( r \)-th reaction, \( \tau_r \) is the decay time of the mRNA molecule, \( R_r(t, z) = H_r(t, z) \Gamma_r(t, z) \) is the total reaction rate at depth \( z \) and \( \alpha_r \) is an unknown proportionality constant. We considered \( T_r \) in the same units as the multi-omic data (i.e., RPKM for metatranscriptomes and NSAF for metaproteomes). Consequently, \( \alpha_r \) is the ratio between the \( r \)-th reaction rate and the corresponding RPKM (or NSAF) “production rate” (mol per time per volume of seawater per RPKM), and thus not only depends on the particular reaction, but also on our sampling protocol and sequencing pipeline.

For each gene \( r \), the transcript profile \( T_r \) will satisfy the same boundary conditions as the DNA profile \( \Gamma_r \), provided that the latter are either zero value (Dirichlet), zero flux (Neumann) or fixed relative gradient boundary conditions (Appendix F.3.6). We calculated the steady state solution of equation (F.3.24), \( T_r^\star \), by solving the time-invariant equation

\[ 0 = -\frac{T_r^\star}{\tau_r} + \frac{R_r}{\alpha_r} - v \partial_z T_r^\star + \partial_z K \partial_z T_r^\star \]  

(F.3.25)

using the MATLAB function \texttt{bvp4c} (309). This was done after the gene-centric model had already reached steady state, at which point the reaction rates \( R_r \) are time-independent functions of depth.

Note that the steady state profile \( T_r^\star(z) \) is proportional to \( 1/\alpha_r \), all else being equal. Hence, by comparing \( T_r^\star \) to metatranscriptomic data (for some given \( \tau_r \)), the constant \( \alpha_r \) can be calibrated via maximum-likelihood estimation as described in Appendix F.3.9. On the other hand, maximizing the log-likelihood in equation (F.3.21) (separately for each gene) by choice of \( \alpha_r, \tau_r \) and the corresponding error scale, yields an estimate of the decay time \( \tau_r \). This was done through repeated solutions of equation (F.3.25) with varying \( \tau_r \) and using the interior-point optimization algorithm implemented by the MATLAB function \texttt{fmincon} (309). We confined the fitted \( \tau_r \) to between \( 10^{-4} \) and \( 10^5 \) days.

After calibration of the decay time \( \tau_r \) and proportionality factor \( \alpha_r \), we calculated the coef-
ficients of determination,

\[ R^2_r = 1 - \frac{\sum_j \left[ \bar{T}_{rj} - T_r(z_{rj}) \right]^2}{\sum_j \left[ \bar{T}_{rj} - \bar{T}_r \right]^2} \]  

(F.3.26)

to evaluate how well the mRNA model explained the metatranscriptomic data. Here, \( \bar{T}_{r1}, \bar{T}_{r2}, \ldots \) are measured mRNA abundances at depths \( z_{r1}, z_{r2}, \ldots \) and \( \bar{T}_r \) is their average. For any given gene \( r \), \( R^2_r \) is a measure for the goodness of fit of the above model to the multi-omic data. Table F.4 in the supplement lists the results for all genes for which \( R^2_r \geq 0.5 \).

The statistical significance (\( P\)-value) of the obtained \( R^2 \) was defined as the probability of obtaining the same or greater \( R^2 \) by applying the same procedure to a random data set, with independent normally distributed values with mean and standard deviation set to the original sample mean and standard deviation. We estimated the \( P\)-values for cases where \( R^2_r \geq 0.9 \) using Monte Carlo simulations of 1000 random data sets: all of them were estimated below 0.005.

Table F.4: Proportionality factors (\( \alpha \)) between mRNA or protein abundances and reaction rates (in mol/(L \( \cdot \) d \( \cdot \) RPKM) or mol/(L \( \cdot \) d \( \cdot \) NSAF), respectively), exponential mRNA or protein decay times (\( \tau \)) and coefficients of determination (\( R^2 \)), estimated as described in Appendix F.3.10. Only cases with \( R^2 \geq 0.5 \) are shown.

<table>
<thead>
<tr>
<th>molecule</th>
<th>type</th>
<th>( \alpha )</th>
<th>( \tau ) (days)</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( nxx )</td>
<td>mRNA</td>
<td>( 9.7 \times 10^{-10} )</td>
<td>52</td>
<td>0.93</td>
</tr>
<tr>
<td>( nosZ )</td>
<td>mRNA</td>
<td>( 2.5 \times 10^{-8} )</td>
<td>222</td>
<td>0.95</td>
</tr>
<tr>
<td>ROM</td>
<td>protein</td>
<td>( 2.7 \times 10^{-3} )</td>
<td>67</td>
<td>0.59</td>
</tr>
<tr>
<td>amo</td>
<td>protein</td>
<td>( 3.0 \times 10^{-5} )</td>
<td>42</td>
<td>0.92</td>
</tr>
<tr>
<td>( nxx )</td>
<td>protein</td>
<td>( 2.1 \times 10^{-4} )</td>
<td>510</td>
<td>0.65</td>
</tr>
<tr>
<td>( norBC )</td>
<td>protein</td>
<td>( 4.6 \times 10^{-4} )</td>
<td>145</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Calibrated gene-centric model (predicts reaction rates and DNA distributions)

Metatranscriptomic and metaproteomic data

Postulated mRNA and protein dynamics

Estimation of unknown mRNA and protein parameters

Evaluation of postulated mRNA and protein dynamics

Figure F.2: **Overview of the modeling approach for the Saanich Inlet OMZ.** Previous geochemical and multi’omic investigations provide conceptual information on the metabolic network in the Saanich Inlet OMZ (11, 181, 218, 503, 504, 540). This information was used to construct a gene-centric biogeochemical mathematical model, which describes the population dynamics of individual genes and metabolic process rates. Unknown reaction-kinetic parameters of the model were calibrated using geochemical depth profiles. The predictions of the calibrated gene-centric model were then validated using independent metagenomic sequence data, qPCR-based cell count estimates for SUP05 as well as process rate measurements. A subsequent extension of the model describes the production, dispersal and decay of mRNA and protein molecules based on the reaction rates predicted by the calibrated gene-centric model. Unknown parameters for the mRNA and protein dynamics are estimated using metatranscriptomic and metaproteomic data. The “goodness of fit” to these multi-omic data is used to further evaluate the gene-centric model, to assess the adequacy of the postulated mRNA and protein dynamics and to gain insight into potentially important but omitted mechanisms of mRNA and protein regulation at ecosystem scales.
F.3.11 Calculating metabolic fluxes between pathways

Dissimilatory metabolic reactions can be interpreted as sources and sinks of metabolites distributed along the water column, producing and consuming metabolites at rates given by the first term in equation (F.3.2). Due to diffusive transport (2nd term in equation (F.3.2)), metabolite fluxes from sources to sinks need not be localized and can span across different depths. Furthermore, some metabolites are partly transported across the OMZ boundaries, towards or from the top layers or the sediments. In the following we describe our approach for calculating steady-state metabolite fluxes across individual reactions.

Let us focus on a particular metabolite and consider a single hypothetical particle created at time 0 at depth $x$. Let $G(t, x, y)$ be the Green’s function of the dispersal-destruction model, so that $G(t, x, \cdot)$ is the distribution density of a particle (created at depth $x$) at depth $y$ and after time $t$. Note that $G(t, x, \cdot)$ may integrate to less than unity if the particle has a positive probability of being consumed anywhere in the water column. The probability rate at which that particle is consumed by any sink $j$ at time $t$ is then

$$\int dy \ G(t, x, y) \frac{\lambda_j(y)}{C(y)},$$

where $\lambda_j(y)$ gives the rate at which sink $j$ consumes particles at depth $y$ and $C(y)$ is the steady state metabolite concentration at that depth. Since each sink corresponds to a pathway consuming the metabolite, $\lambda_j(y)$ is given by the community-wide reaction rate at $y$ multiplied by the appropriate stoichiometric coefficient. The probability that the particle will eventually be destroyed by sink $j$ is given by

$$\int_0^\infty dt \int dy \ G(t, x, y) \frac{\lambda_j(y)}{C(y)}.$$  

The total rate at which particles created by source $i$ are destroyed by sink $j$ across the entire OMZ, denoted $F_{ij}$, is

$$F_{ij} = \int dx \ b_i(x) \int_0^\infty dt \int dy \ G(t, x, y) \frac{\lambda_j(y)}{C(y)},$$

where $b_i(x)$ is the rate at which the metabolite is produced by source $i$ at depth $x$. Switching integrals in (F.3.29) gives

$$F_{ij} = \int dy \ \frac{\lambda_j(y)}{C(y)} \int_0^\infty dt \ \vartheta_i(t, y),$$
where

\[ \vartheta_i(t, y) = \int dx \, b_i(x) G(t, x, y) \]  \hspace{1cm} (F.3.31)

is the solution to the dispersal-destruction model with initial distribution \(b_i(x)\):

\[ \partial_t \vartheta_i(t, y) = -\vartheta_i(t, y) \frac{\vartheta_i(t, y)}{C(y)} \sum_j \lambda_j(y) + \partial_y [K(y) \partial_y \vartheta_i(t, y)], \]  \hspace{1cm} (F.3.32)

\[ \vartheta_i(0, y) = b_i(y). \]

Particles crossing the domain boundary are considered to be lost. Hence, Dirichlet (Neumann) boundary conditions in the original model correspond to zero-value (zero-flux) boundary conditions for \(\vartheta_i\). The total boundary loss rate of particles created by source \(i\) is the remainder

\[ F_{i,o} = \int dx \, b_i(x) - \sum_j F_{ij}. \]  \hspace{1cm} (F.3.33)

Similarly, the rate at which particles flow in at the boundary and are destroyed by sink \(j\) is given by

\[ F_{o,j} = \int dx \, \lambda_j(x) - \sum_i F_{ij}. \]  \hspace{1cm} (F.3.34)

We solved equation (F.3.32) using the MATLAB® function `pdepe` and evaluated all integrals in equation (F.3.30), (F.3.33) and (F.3.34) using the trapezoid integration scheme (309).

**F.3.12 Local sensitivity analysis**

We evaluated the sensitivity of the model predictions to small changes in model parameters using normalized local sensitivity coefficients (NLSC) (71). NLSCs compare the relative changes in model output variables (\(V_j\), integrated over all depths) to the relative changes of model parameters (\(p_i\)) by means of partial derivatives, evaluated at the default (e.g., fitted) parameter values:

\[ \text{NLSC}_{ij} = \left| \frac{p_i \frac{\partial V_j}{\partial p_i}}{V_j \frac{\partial V_j}{\partial p_i}} \right|. \]  \hspace{1cm} (F.3.35)

Hence, \(\text{NLSC}_{ij}\) is a measure for the relative effects that parameter \(i\) has on the output variable \(j\). The partial derivative in equation (F.3.35) was approximated numerically by
changing \( p_i \) by 1% from its default value. The results are summarized in figure F.3 in the supplement.

The sensitivity of the model varied strongly among parameters. For example, the kinetic constants for ROM (aerobic remineralization of organic matter) had a relatively strong effect on chemical as well as gene concentration profiles by modulating the availability of oxygen and ammonium near and above the SNTZ. On the other hand, the kinetic constants for PDNO and \( nosZ \) (which constitute the denitrification pathway) had relatively little effects on the predicted chemical profiles, as long as both were increased or decreased in unison. Similar observations were made for \( amo \) and \( nxr \), which constitute the nitrification pathway. Moreover, the total predicted gene concentrations (Fig. 7.3 in the main text) were robust against parameter changes and only varied within an order of magnitude as long as the calibrated geochemical profiles matched the data moderately well. This suggests that geochemical fluxes are good predictors for microbial growth, but less suited for estimating reaction-kinetic parameters, especially when these are correlated (242).

Figure F.3: Local sensitivity heatmap of the calibrated model by means of normalized local sensitivity coefficients. A brighter color corresponds to a higher sensitivity. “Khalf” stands for half-saturation constants, “Kinh” for half-inhibition constants, “V” for maximum cell-specific reaction rates and “q” for cell death rates. The heatmap is hierarchically clustered using UPGMA linkage and Euclidean metric. Methodological details are given in Appendix F.3.12.
F.4 Caveats and special notes

F.4.1 The role of sulfate reduction

The choice of pathways included in the model was based on metaproteomics data by Hawley et al. (181). None of the proteins associated with sulfur-metabolism were mapped to known sulfate reducers, suggesting that these proteins may act in sulfur oxidation and that sulfate reduction only played a minor role in Saanich Inlet’s OMZ at the time of sampling. In particular, an NCBI BLASTP search mapped all detected \textit{dsrA} and \textit{aprAB} proteins to \textit{SUP05} (301). All other taxonomically resolved sulfite reductase proteins were mapped to \textit{Candidatus} Ruthia magnifica, a sulfur-oxidizing endosymbiont (400). The mRNA depth profiles of \textit{sat}, \textit{aprAB} and \textit{dsrAB} (Supplemental Figs. F.1A,B,C), which comprise the dissimilatory sulfide oxidation pathway (or sulfate reduction pathway when reversed), show a clear peak at the SNTZ, consistent with the metatranscriptomic profiles of the denitrification genes \textit{norBC} and \textit{nosZ} (Fig. 7.3 in the main text). These multimolecular data suggest that the \textit{sat}, \textit{aprAB} and \textit{dsrAB} enzymes act predominantly in sulfur oxidation. The high \textit{sat}, \textit{aprAB} and \textit{dsrAB} DNA concentrations at the bottom might be due to sediment resuspension, cell sinking from the more productive SNTZ or cell diffusion from the sulfate reducing sediments (290, 347).

Due to the much higher organic matter concentrations in the sediments, heterotrophic sulfate reduction and anaerobic remineralization is correspondingly higher in the sediments than in the water column (4, 101). Hence, most of the \( \text{H}_2\text{S} \) and \( \text{NH}_4^+ \) in the sulfidic part of the OMZ is expected to originate from the adjacent sediments via diffusion. An influx of \( \text{H}_2\text{S} \) and \( \text{NH}_4^+ \) predominantly from the sediments is compatible with the measured steep \( \text{H}_2\text{S} \) and \( \text{NH}_4^+ \) gradients (Figs. 7.2B,F in the main text), as well as the gradual upward progression of the \( \text{H}_2\text{S} \) and \( \text{NH}_4^+ \) fronts following annual renewal (Figs. 7.1B and 7.2B,F in the main text). Sediments have previously been indicated as the main sulfide sources in other OMZs, such as the the Eastern Boundary upwelling system (415) or the central Namibian coastal upwelling zone (50).

Due to the lack of rate measurements heterotrophic sulfate reduction and cryptic sulfur cycling cannot be completely ruled out. However, calibrating the above model to the chemical data (Fig. 7.2 in the main text), while including sulfate reduction as an additional pathway, dramatically decreases the \textit{goodness of fit}. This is because an additional \( \text{H}_2\text{S} \) source in the OMZ shifts the SNTZ further up, thereby increasing the main discrepancy between the model and the data. Hence, on grounds of parsimony, we eventually omitted sulfate reduction from
Figure F.1: Metagenomic, metatranscriptomic and metaproteomic depth profiles of (A) sulfate adenylyltransferase (sat), (B) adenylylsulfate reductase (aprAB) and (C) sulfite reductase (dsrAB) genes, which together comprise the sulfide oxidation pathway (or sulfate reduction pathway, if reversed), as well as (D) periplasmic nitrate reductase napAB, (E) nitrate reductase narGHIJ and (F) NO-forming nitrite reductase nirKS. Data taken on February 10, 2014. All of the dsrAB, aprAB and most of the napAB protein sequences were mapped to the γ-proteobacterial SUP05 clade (504). All detected narGHIJ protein sequences were either mapped to SUP05 or to the anammox planctomycete bacteria Candidatus Scalindua profunda and KSU-1 (493) (only SUP05 proteins are shown). Similarly, only non-planctomycete-annotated narGHIJ and nirKS DNA abundances are shown.

the model and assumed that H$_2$S originates from the sediments via diffusion.

We note that similar theoretical work by Reed et al. (386) did suggest the existence of a cryptic sulfur cycle in the Arabian Sea OMZ. However, the latter is located more than 1 km above the sediments and hydrogen sulfide influx from the sediments into the OMZ is not possible due to elevated oxygen levels below the OMZ (489).
F.4.2 The role of DNRA

It has been previously hypothesized that dissimilatory nitrate reduction to ammonium (DNRA) might be active in Saanich Inlet’s OMZ, possibly providing ammonium to anammox bacteria (181, 253, 374). So far DNRA was not detected in any of our incubation experiments, although we cannot rule out cryptic DNRA due to rapid ammonium consumption by anammox (374). Measured ammonium profiles in Spring 2010 did not indicate a significant ammonium source at or below the SNTZ (Fig 7.2 b in the main text). Similarly, Schunck et al. (415) reports negligible DNRA for a sulfidic OMZ off the coast of Peru.

Nevertheless, we tested an extension of our model with DNRA as an additional pathway. Calibrating the model to the same data (January–March 2010) consistently predicted negligible DNRA rates, and the goodness of fit (in terms of the log-likelihood) did not significantly improve with the inclusion of DNRA. On grounds of parsimony we thus eventually omitted DNRA from the model. We mention that calibrating the model to chemical data from September 2009 (181) indicated significant DNRA as well as anammox rates (both in the order of 1 mmol N/(m²⋅d)), suggesting that DNRA-fed anammox activity fluctuates strongly throughout the year. High spatiotemporal variability of N-loss activities are known for other OMZs and may be associated with fluctuations in surface primary production, as well as fluctuations in electron acceptor availability driven by annual deep water renewal (12, 205, 252).

F.4.3 The role of aerobic sulfide oxidation

Extensive previous work points towards NO$_3^-$ and other nitrogen compounds as dominant electron acceptors for H$_2$S oxidation in Saanich Inlet during periods of strong stratification (11, 181, 218, 503, 540). For example, as shown in Fig. 7.1B in the main text, the upper boundary of H$_2$S concentrations closely follows the lower boundary of NO$_3^-$ — rather than O$_2$ — over time, especially during the period considered in this study (early 2010). The strong similarity between sulfur cycling gene profiles and denitrification gene profiles (February 10, 2010; Fig. F.1) provides further evidence for the tight coupling between denitrification and sulfide oxidation at that time. Similarly, nitrogen compounds have been shown to be the dominant electron acceptors for sulfide oxidation in the Peruvian OMZ (415), and Canfield et al. (55) established a strong link between sulfide oxidation and nitrate reduction in the Chilean OMZ. Note that during renewal events in Fall, O$_2$ can indeed become an important electron acceptor for H$_2$S oxidation in Saanich Inlet (540). This does not, however, affect this study, which focuses on periods of intense stratification near steady state conditions.
We note that we had initially considered aerobic sulfide oxidation as an additional reaction in our model. Preliminary calibrations to geochemical data showed that the model’s explanatory power was significantly compromised by this reaction, because diffusive $O_2$ fluxes into the sulfidic zone could not account for the $O_2$ needed for sulfide oxidation (in addition to $O_2$ needed for nitrification). In fact, in our simulations ammonium ended up competing with $H_2S$ for $O_2$, which in turn negatively affected the accuracy of the predicted $NO_3^-$ profile. While lateral intrusions of oxygenated water could in principle account for the additional $O_2$ needed for sulfide oxidation, spatiotemporal $O_2$ profiles do not provide any indication of such intrusions during this period of intense stagnation (Fig. 7.1B in the main text). We thus omitted aerobic sulfide oxidation from our final model.

**F.4.4 Planctomycetes and nxr**

Our molecular data suggest that the anammox bacteria planctomycetes (90) are also aerobically oxidizing nitrite to nitrate in the oxycline (181) using the nitrate oxidoreductase narGHIJ (458). Metatranscriptomic and metaproteomic profiles of planctomycete-associated narGHIJ sequences peak at about 120 m depth and decrease rapidly below that (Fig. 7.3 in the main text), while planctomycete-associated HAO (anammox-associated hydroxylamine-oxidoreductase (458)) sequences are most abundant at 150 m depth and at appreciable levels all the way down to 200 m. As a consequence, narGHIJ is expected to also proliferate in regions where it is not actually being transcribed. Indeed, metagenomic data show a bimodal profile of Planctomycete-associated narGHIJ sequences, with local maxima at 120 m and 150 m depths, corresponding to the putative maxima of nitrite oxidation and anammox activity. Due to this bimodality we did not include narGHIJ nor nxr metagenomic profiles in our analysis.

**F.5 Inverse linear transport modeling (ILTM)**

Chemical concentration profiles were used to estimate denitrification and anammox rates across the water column, independently of the gene-centric model and the rate measurements described in Appendix F.2.5. In short, a steady state diffusion model was used to estimate the net metabolite production (or consumption) rates that “best” explained the observed depth profiles. This so called inverse linear transport modeling (ILTM) approach is widespread in oceanography and atmospheric sciences, were known global distributions of compounds such as trace gases are used to estimate unknown sources and sinks (303, 446).
In the following, we explain our procedure for estimating the net production profile, $\rho(z)$, for a particular metabolite with a given concentration profile, $\hat{C}(z)$. All calculations were performed in Mathworks MATLAB®. Each profile $\hat{C}(z)$ was obtained through Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) interpolation of the actual measured concentrations. ILTM was applied separately to concentration profiles from cruises 47 and 48, as well as to the chemical profiles used for model calibration (cruises 41–44, Appendix F.3.8) after averaging across replicates at each depth.

Our starting point is the diffusive transport model

$$0 = \rho + \frac{\partial}{\partial z} \left[ K(z) \cdot \frac{\partial C}{\partial z} \right], \quad (F.5.1)$$

which describes the steady-state distribution $C(z)$ across depth $z$, given a particular net production profile $\rho(z)$ and eddy diffusion coefficient $K(z)$. The eddy diffusion coefficient was calculated as described in Appendix F.3.7. Our goal is to determine the appropriate $\rho(z)$ that “best” explains the observed steady state profile $\hat{C}(z)$, through the following steps:

1. Calculate the discretized Green’s function (394) of the above partial differential equation (PDE) with zero Dirichlet boundary conditions: Let $G_{nm}$ be an approximation for $G(z_n, z_m)$, where $G$ solves the time-independent PDE

$$0 = \frac{\partial}{\partial x} \left[ K(x) \frac{\partial}{\partial x} G(x, y) \right] + \delta(x - y) \quad (F.5.2)$$

on the domain $\Omega := [\text{top}, \text{bottom}]$, with boundary conditions

$$G(x, y) \bigg|_{x \in \partial \Omega} = 0. \quad (F.5.3)$$

In practice, $G_{nm}$ can be set to $dz_m \cdot G(z_n, z_m)$, where $G$ is the solution to the PDE system

$$0 = \frac{\partial}{\partial x} \left[ K(x) \frac{\partial}{\partial x} G(x, z_m) \right] + H(x - z_m + dz_m/2)H(z_m + dz_m/2 - x)/dz_m, \quad (F.5.4)$$

$$G(x, z_m) \bigg|_{x \in \partial \Omega} = 0.$$

Here, $H$ is the Heaviside step function and $dz_m$ is the grid’s step at $z_m$, assumed to be chosen small enough ($dz = 2$ m in our case).
2. Note that for any candidate net production profile \( \rho(x) \), the sum

\[
\sum_m G_{nm} \cdot \rho(z_m)
\]

becomes an approximation for \( C^o(z_n) \), where \( C^o \) is a solution to the following steady-state transport problem with zero Dirichlet boundary conditions:

\[
0 = \frac{\partial}{\partial x} \left[ D(x) \frac{\partial C^o}{\partial x} \right] + \rho(x), \quad C^o(x) \bigg|_{x \in \partial \Omega} = 0.
\]

3. For the given measured concentrations \( \hat{C}(x) \) at the domain boundary \( x \in \{\text{top, bottom}\} \), calculate the particular solution \( C^p \) to the transport problem with given boundary values but no sources:

\[
0 = \frac{\partial}{\partial x} \left[ K(x) \frac{\partial C^p}{\partial x} \right], \quad C^p(x) \bigg|_{x \in \partial \Omega} = \hat{C}(x).
\]

After solving for \( C^p \), evaluate \( C^p \) on the grid, i.e., set \( C^p_n = C^p(z_n) \).

4. Note that for any candidate net production profile \( \rho(x) \), the sum \( C := C^o + C^p \) is a solution to the full PDE problem

\[
0 = \frac{\partial}{\partial x} \left[ K(x) \frac{\partial C}{\partial x} \right] + \rho(x), \quad C(x) \bigg|_{x \in \partial \Omega} = \hat{C}(x).
\]

Similarly, the sum

\[
C^p_n + \sum_m G_{nm} \cdot \rho(z_m)
\]

is an approximation for \( C(z_n) \).

5. Note that \( C^p \) corresponds to the hypothetical steady-state profile that would result purely from transport across the domain boundary, in the absence of any sources or sinks in its interior. Similarly, the difference \( B = \hat{C} - C^p \) is the part that cannot be explained by transport across boundaries, but must rather be attributed to production and consumption inside \( \Omega \). Hence, using the particular discretized solution \( C^p_n \), the discretized profile \( \hat{C}_n = \hat{C}(z_n) \) and the discretized steady-state transport kernel \( G_{nm} \), one could in principle estimate \( \rho_m = \rho(z_m) \) by minimizing the sum of squared residuals (SSR)

\[
SSR = \sum_n \left| \sum_m G_{nm} \cdot \rho_m - B_n \right|^2.
\]
where $B_n = \hat{C}_n - C_n^p$. The above problem is a classical linear least-squares problem if one considers $G_{nm}$ as a matrix ($\mathbb{G}$) and $\rho_m$, $\hat{C}_n$, $C_n^p$ as vectors ($\boldsymbol{\rho} \in \mathbb{R}^M$, $\hat{\mathbf{C}} \in \mathbb{R}^N$ and $\mathbf{C}^p \in \mathbb{R}^N$):

$$\text{SSR} = \|\mathbb{G} \cdot \boldsymbol{\rho} - \mathbf{B}\|^2.$$  \hspace{1cm} (F.5.11)

The minimum SSR is then obtained for

$$\boldsymbol{\rho} = \mathbb{G}^* (\hat{\mathbf{C}} - \mathbf{C}^p),$$ \hspace{1cm} (F.5.12)

where $\mathbb{G}$ is the Moore-Penrose pseudoinverse of $\mathbb{G}$. Put simply, the so estimated $\boldsymbol{\rho}$ is the net production profile that “best” explains the observed steady-state concentration profile $\hat{\mathbf{C}}$, after subtracting the part $\mathbf{C}^p$ explained by transport across the domain boundaries.

6. The least-squares estimator in Eq. (F.5.12) becomes unstable if the reference profile $\hat{\mathbf{C}}$ stretches linearly (or almost linearly) across large depth intervals, leading to spurious oscillations in the estimated profile $\boldsymbol{\rho}$. To address this problem, we “penalized” strong oscillations in the estimated net production profile by instead minimizing the modified SSR

$$\text{SSR}^* = \|\mathbb{G} \cdot \boldsymbol{\rho} - \mathbf{B}\|^2 + M^{-2} \|\xi \boldsymbol{\rho}\|^2,$$ \hspace{1cm} (F.5.13)

where $\xi$ is an appropriately chosen regularization parameter (36) that quantifies the penalty imposed on large $|\boldsymbol{\rho}|$. The above regularization method is known as Tikhonov regularization. A larger Tikhonov factor $\xi$ will typically result in a smoother $\boldsymbol{\rho}$ but also a poorer overall fit, since goodness of fit is sacrificed in favor of small $\boldsymbol{\rho}$. We manually chose $\xi$ as large as possible but still small enough such that the residual $\|\mathbb{G} \cdot \boldsymbol{\rho} - \mathbf{B}\|$ remained much smaller than $\|\mathbf{B}\|$.

7. Assuming that $\text{H}_2\text{S}$ is mostly consumed by denitrification (PDNO and nosZ) according to the stoichiometry given in Appendix F.3.3, one mol of consumed $\text{H}_2\text{S}$ corresponds to $8 \cdot (1 - L_{\text{PDNO}})/(5 - 3L_{\text{PDNO}})$ mol N released as $\text{N}_2$. Similarly, one mol of consumed $\text{NH}_4^+$ by anammox corresponds to 2 mol N released as $\text{N}_2$, however nitrification likely also contributes to $\text{NH}_4^+$ consumption in the more oxygenated layers. Hence, whenever the net $\text{NO}_3^-$ production was positive, the net $\text{NO}_3^-$ production rate was subtracted from the net $\text{NH}_4^+$ consumption rate, yielding an estimate for $\text{NH}_4^+$ consumption purely by anammox.
Appendix G

Chapter 8: Supplemental material

G.1 Methods

G.1.1 Details on example 1 (batch-fed incubator)

Parameterization

In the model, temperature was held constant at 20°C and pH was held constant at 5, in accordance with the original incubation experiment (94). Ammonia and ammonium were assumed to be at dissociation equilibrium, determined by the pH and the standard ammonium dissociation constant $5.69 \times 10^{-10}$ M (73). The dissociation constant was corrected for the lower temperature in the experiment using the Van ’t Hoff equation (20).

The initial urea capacity $M_{\text{ure}}^o$ was estimated from the derivative of the urea time series, assuming that the initial urea kinetics were saturated by high substrate concentration. Time series derivatives were estimated via 4th order Savitzky-Golay smoothening with a sliding window span of 10 days (222), followed by centered finite differences. The initial urea, $\text{NH}_4^+$ and $\text{NO}_3^-$ concentrations were set to 1.12 mM, 124 $\mu$M and 49.8 $\mu$M, respectively, according to the first sampling point in the measured time series. The initial $\text{NO}_2^-$ concentration was assumed to be zero. The parameters $K_{\text{ure}}$, $K_{\text{nzx}}$, $A_{\text{amo}}$ and $A_{\text{nzx}}$ were taken from existing literature on Nitrosospira and Nitrobacter (Table G.1).

The remaining free parameters $K_{\text{amo}}$, $M_{\text{nzx}}^o$, $A_{\text{ure}}$, $A_{\text{amo,ure}}$, $\lambda_{\text{AOB}}$ and $\rho_{\text{ure}}$ were simultaneously calibrated to the urea, $\text{NH}_4^+$ and $\text{NO}_3^-$ time series via maximum-likelihood estimation (113). This approach estimates unknown parameters by maximizing the likelihood of observing the available data given a particular candidate choice of parameter values. Maximum likelihood estimation is widely used in statistical inference such as multilinear regression and physics (291). In our case, the likelihood of the data was calculated on the basis of a mixed deterministic-stochastic structure, in which the deterministic part is given by the reaction-centric model and errors are assumed to be normally distributed on a logarithmic scale.
The likelihood was maximized using the SBPLX optimization algorithm (214), which uses repeated simulations and gradual exploration of parameter space. To reduce the possibility of only reaching a local maximum, fitting was repeated 100 times using random initial parameter values and the best fit among all 100 runs was used. Parameter confidence intervals were calculated using the inverse observed Fisher information, which is an estimator of the parameter covariance matrix (92). Fitted parameter values, their confidence intervals and a comparison to available literature are given in Table G.1.

Assessing the importance of ure-amo cross-amplification

To test the suitability of a model variant without ure-amo cross-amplification as outlined in the main text, we treated ure and amo as independent reactions performed by separate cell populations. Hence, we assumed $A_{amo,ure} = A_{ure,amo} = 0$ and $\rho_{ure} = 0$, and replaced the maintenance rate $\lambda_{AOB}$ with two independent rates $\lambda_{amo}$ and $\lambda_{ure}$. Furthermore, the initial capacities $M_0^{ure}$ and $M_0^{amo}$ were treated as independent parameters. The new set of free parameters thus comprised $K_{amo}$, $M_0^{amo}$, $M_0^{urn}$, $A_{ure}$, $A_{amo}$, $\lambda_{ure}$ and $\lambda_{amo}$, while the remaining parameters were fixed as described above. Fitting was performed as with the original model and yielded multiple local optima, none of which matched the data as well as the original model (Supplemental Fig. G.2).
Table G.1: Fixed and fitted model parameters for the batch bioreactor incubated with *Nitrosospira* sp. and *Nitrobacter* sp. Parameters marked with an asterisk (*) were unknown and were thus fitted to the time series; approximately comparable literature values are provided where available. SE refers to the standard error of the fitted value, in the same units. The initial metabolite concentrations $C_{\text{urea}}^0$, $C_{\text{NH}_4^+}^0$, and $C_{\text{NO}_2^-}^0$ were taken from the chemical time series on day 1. The initial reaction capacity $M_{\text{ure}}^0$ was estimated from the slope of the time series at time zero. The remaining parameter values were taken from the indicated literature.

<table>
<thead>
<tr>
<th>param.</th>
<th>value</th>
<th>SE</th>
<th>comparison</th>
<th>group</th>
<th>literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{\text{ure}}$</td>
<td>670 µM urea</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td><em>Nitrosospira L115</em></td>
</tr>
<tr>
<td>$K_{\text{amo}}$</td>
<td>*4.59 µM NH$_3$</td>
<td>±0.27</td>
<td>6–11</td>
<td><em>Nitrosospira spp.</em></td>
<td>(208)</td>
</tr>
<tr>
<td>$K_{\text{nxr}}$</td>
<td>27.2 µM NO$_2^-$</td>
<td>–</td>
<td>–</td>
<td><em>Nitrobacter spp.</em></td>
<td>(37)</td>
</tr>
<tr>
<td>$A_{\text{ure}}$</td>
<td>*1.11 d$^{-1}$</td>
<td>±0.004</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$A_{\text{amo}}$</td>
<td>1.2 d$^{-1}$</td>
<td>–</td>
<td>–</td>
<td><em>Nitrosospira AV2</em></td>
<td>(30)</td>
</tr>
<tr>
<td>$A_{\text{nxr}}$</td>
<td>1.03 d$^{-1}$</td>
<td>–</td>
<td>–</td>
<td><em>Nitrobacter sp.</em></td>
<td>(227)</td>
</tr>
<tr>
<td>$A_{\text{amo,ure}}$</td>
<td>*12.8 d$^{-1}$</td>
<td>±0.54</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$\lambda_{\text{AOB}}$</td>
<td>*0.0055 d$^{-1}$</td>
<td>±0.0005</td>
<td>0.027 d$^{-1}$</td>
<td><em>N. europaea</em></td>
<td>(467)</td>
</tr>
<tr>
<td>$\rho_{\text{ure}}$</td>
<td>*0.26</td>
<td>±0.018</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$C_{\text{urea}}^0$</td>
<td>1.12 mM</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$C_{\text{NH}_4^+}^0$</td>
<td>124 µM</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$C_{\text{NO}_2^-}^0$</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$C_{\text{NO}_3^-}^0$</td>
<td>49.8 µM</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$M_{\text{ure}}^0$</td>
<td>773 nM/d</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$M_{\text{nxr}}^0$</td>
<td>*35.8 mM/d</td>
<td>±1.4</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
G.1.2 Details on example 2 (flow-through bioreactor)

Assimilation of time series

Experimental time series of $\text{NH}_4^+$ and $\text{NO}_3^-$ concentrations were noise-filtered using 4th order Savitzky-Golay smoothening with a sliding window time span of 30 days (222). Derivatives of concentration profiles were estimated by applying a centered finite differences scheme to the noise-filtered profiles. $amo$ and $nxr$ rates were estimated from the derivatives of the $\text{NH}_4^+$ and $\text{NO}_3^-$ concentration profiles, respectively, after accounting for substrate input and dilution. Estimated $amo$ and $nxr$ rates were then used in the growth model for the reaction capacities, Eqs. (8.4.9,8.4.10), as described in the main text.

Parameterization

In the experiment, pH was maintained around 7 by the automatic addition of an alkaline solution, and the bioreactor was maximally ventilated to ensure sufficient oxygenation (109). In our model we thus assumed pH = 7 and ignored oxygen limitation in the reaction kinetics. Temperature was assumed to be 30°C until day 181 and 25°C afterwards, in accordance with the original experiment. Bioreactor dilution rates and input substrate concentrations were obtained from the authors of the original experiment upon personal correspondence. $\text{NH}_3$ concentration was calculated from $\text{NH}_4^+$ concentration by assuming that the two are at acid-dissociation equilibrium, similarly to the first example.

The initial $amo$ capacity, $M_{amo}^0$, was estimated from the $\text{NH}_4^+$ time series but had negligible effects on the simulations. The initial $nxr$ capacity was set to zero based on the absence of $\text{NO}_3^-$ accumulation. The $amo$ and $nxr$ half-saturation constants and the self-amplification factors $A_{amo}$ and $A_{nxr}$ were calibrated to the $\text{NH}_4^+$, $\text{NO}_2^-$ and $\text{NO}_3^-$ time series by maximizing the mean coefficient of determination ($R^2$) across all three data sets, which is analogous to weighted least-squares fitting in the univariate case. Only data from days 1–250 were used for the calibration. The mean $R^2$ was maximized using the SBPLX algorithm (214). To reduce the possibility of only reaching a local maximum, fitting was repeated 100 times using random initial parameter values and the best fit among all 100 runs was used. Fitted parameter values and a comparison to available literature are given in Table G.2.
Table G.2: Fixed and fitted model parameters for the flow-through bioreactor (109). Parameters marked with an asterisk (⋆) were calibrated using data from days 1–250 and are compared to literature values. The initial metabolite concentrations $C^{o}_{\text{NH}_4^+}$, $C^{o}_{\text{NO}_2^-}$ and $C^{o}_{\text{NO}_3^-}$ were taken from the chemical time series on day 1. The initial reaction capacities $M^{o}_{\text{amo}}$ and $M^{o}_{\text{nxr}}$ were estimated from the slopes of the chemical time series on day 1. The parameters $C^{\text{in}}_{\text{NH}_4^+}$, $\mu$, pH and temperature were controlled throughout the experiment.

<table>
<thead>
<tr>
<th>param.</th>
<th>value</th>
<th>comparison</th>
<th>group</th>
<th>literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{\text{amo}}$</td>
<td>⋆ 3.21 $\mu$M NH$_3$</td>
<td>1.2–23</td>
<td>AOB</td>
<td>(463, 507)</td>
</tr>
<tr>
<td>$K_{\text{nxr}}$</td>
<td>⋆ 1.32 mM NO$_2^-$</td>
<td>0.01–1.68</td>
<td>NOB</td>
<td>(198, 294)</td>
</tr>
<tr>
<td>$A_{\text{amo}}$</td>
<td>⋆ 0.145 d$^{-1}$</td>
<td>0.32–2.1</td>
<td>AOB</td>
<td>(202, 375)</td>
</tr>
<tr>
<td>$A_{\text{nxr}}$</td>
<td>⋆ 0.176 d$^{-1}$</td>
<td>0.17–1.4</td>
<td>NOB</td>
<td>(31, 375)</td>
</tr>
<tr>
<td>$C^{o}_{\text{NH}_4^+}$</td>
<td>26.7 mM</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$C^{o}_{\text{NO}_2^-}$</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$C^{o}_{\text{NO}_3^-}$</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$M^{o}_{\text{amo}}$</td>
<td>17.1 mM/d</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$M^{o}_{\text{nxr}}$</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$C^{\text{in}}_{\text{NH}_4^+}$</td>
<td>35.7 – 143 mM</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$\mu$</td>
<td>0 – 0.46 d$^{-1}$</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>pH</td>
<td>7</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>temperature</td>
<td>30$^\circ$C – 25$^\circ$C</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
G.1.3 Computational methods: Using MCM for reaction-centric models

The biochemical models described in the main text were constructed using MCM (Chapter 4; (284)). The framework allows the specification of microbial and abiotic reaction networks within an environmental context resembling, for example, a bioreactor. All considered metabolites (e.g., $\text{NO}_2^-$ and $\text{NH}_4^+$), any reactions between them (e.g., amo and nxr) and any environmental variables (such as temperature and dilution rate), are specified in special configuration files using high-level code. For example, the specification of metabolites may look as follows:

```
NO2
  environmental_dynamics: initial 0 flux -NO2*dilution_rate
                           and environmental_production

NH4
  environmental_dynamics: initial 0 flux (input_NH4 - NH4)*dilution_rate
                           and environmental_production
```

Notice that we specify the initial concentration (i.e., at the beginning of the experiment) of both metabolites as zero. Furthermore, both metabolites are subject to biochemical fluxes (indicated by the keyword `environmental_production`), as well as depletion at a rate proportional to the bioreactor’s dilution rate, represented by a separate model variable `dilution_rate`. In addition, $\text{NH}_4^+$ is subject to repletion at a rate proportional to the input substrate concentration, represented by the model variable `input_NH4`. Both environmental variables `dilution_rate` and `input_NH4`, in turn, are explicitly specified using time series that are linearly interpolated between time points:

```
dilution_rate
  dynamics: value interpolation of "data/dilution_rate.txt"
  units: 1/day

input_NH4
  dynamics: value interpolation of "data/input_NH4_concentration.txt"
  units: mol/L
```
Since we assume NH$_3$ and NH$_4^+$ to be at dissociation equilibrium, we define NH$_3$ as a non-dynamical variable explicitly depending on NH$_4$:

```
NH3
    environmental_dynamics: base_of_acid NH4 5.62e-10
```

Observe that we provide the ammonium standard dissociation constant, which is used to calculate NH$_3$ at any given time depending on NH$_4$, pH and temperature. The latter two are, in turn, specified as additional environmental variables, e.g., in the continuous-flow bioreactor model as follows:

```
pH
    dynamics: value 7

temperature
    dynamics: value piecewise2(t,181,30,25)
    units: C
```

The definition of each reaction requires a chemical equation and a specification of the reaction’s rate, as demonstrated below:

```
amo
    equation: NH4 + 1.5*O2 -> NO2 + H2O + 2H
    environmental_rate: amo_capacity * NH3/(NH3 + $K_{half_amo}$)

nxr
    equation: NO2 + 0.5*O2 -> NO3
    environmental_rate: nxr_capacity * NO2/(NO2 + $K_{half_nxr}$)
```
Observe that while in the model *amo* consumes \( \text{NH}_4 \), its rate is limited by \( \text{NH}_3 \) in accordance with suggestions by Suzuki et al. (463) that ammonia, and not ammonium, is the limiting substrate. The rates of both reactions are proportional to the bioreactors reaction capacities, \texttt{amo_capacity} and \texttt{nxr_capacity}, modeled as separate dynamic variables (see below). The half-saturation constants, \texttt{Khalf\_amo} and \texttt{Khalf\_nxr}, are enclosed in dollar signs indicating that these are so called symbolic model parameters. Symbolic parameters allow a high-level analysis of the model, and can be automatically calibrated to available data. Symbolic model parameters are themselves specified in appropriate configuration files, using a syntax similar to the following:

```
Khalf\_amo
  default: 8.5e-6
  minimum: 1e-10
  maximum: 1e-1
  units: mol/L
  fixed: no

Khalf\_nxr
  default: 2.29e-4
  minimum: 1e-10
  maximum: 1e-1
  units: mol/L
  fixed: no
```

Observe that we specified both symbolic parameters as non-\texttt{fixed}, which tells MCM to calibrate them whenever possible. Model calibration is an iterative process, which begins with the specified \texttt{default} values, and gradually explores the parameter space within the constraints specified by \texttt{minimum} and \texttt{maximum}. The *amo* and *nxr* capacities are defined as dynamic environmental variables:

```
amo\_capacity
  dynamics: initial 0.017 rate $A_{amo}\ast data\_rate\_amo - amo\_capacity\ast dilution\_rate$
  units: mol/L/d
```
nxr_capacity

dynamics: initial 0 rate \( A_{nxr} \) \(*\ data\_rate\_nxr - nxr\_capacity \* dilution\_rate \)
units: mol/L/d

Observe that the growth of \( amo\_capacity \) and \( nxr\_capacity \) is driven by the reaction rates estimated from the time series, \( data\_rate\_amo \) and \( data\_rate\_nxr \), respectively. The latter are, in turn, calculated directly from the available time series as well as the known dilution rate and input substrate concentration:

data\_NH4

dynamics: value smoothening\_SG4 30 of "data/NH4.txt"
units: mol/L

data\_NO3

dynamics: value smoothening\_SG4 30 of "data/NO3.txt"
units: mol/L

data\_NH4\_rate\_of\_change

dynamics: value derivative\_CFD of smoothening\_SG4 30 of "data/NH4.txt"
units: mol/L/d

data\_NO3\_rate\_of\_change

dynamics: value derivative\_CFD of smoothening\_SG4 30 of "data/NO3.txt"
units: mol/L/d

data\_rate\_amo

dynamics: value (input\_NH4-data\_NH4) \* dilution\_rate - data\_NH4\_rate\_of\_change
constraints: positive
units: mol/L/d

data\_rate\_nxr

dynamics: value data\_NO3 \* dilution\_rate + data\_NO3\_rate\_of\_change
constraints: positive
units: mol/L/d
The above code is parsed by MCM, which sets up and numerically solves the corresponding differential equations for the bioreactor’s reaction capacities, the actual reaction rates as well as the metabolite concentrations. When provided with time series data corresponding to any of the model’s predictions (e.g., NH4 concentration), MCM calibrates unknown model parameters (e.g., Khalf_amo) using maximum-likelihood estimation (113). The likelihood is optimized using an iterative optimization algorithm involving step-wise parameter adjustments and repeated simulations. Other fitting objectives are also available, such as maximization of the average coefficient of determination ($R^2$), which is analogous to weighted least-squares fitting.

MCM itself is controlled through custom scripts, i.e., text files containing a sequence of special commands, such as for running simulations or fitting parameters. For example, the following four commands specify the output directory, the model configuration files, the total simulation time (in days), and subsequently invoke a simulation of the model:

```plaintext
setod simulation_output/nitrifier
set model models/nitrifier
set maxSimulationTime 525
runMCM
```

The full incubator and bioreactor models, as well as all necessary MCM scripts, are available at: http://www.zoology.ubc.ca/MCM

### G.2 Mathematical proofs

#### G.2.1 On specific maintenance rates

The effects of maintenance requirements of individual cells on their population is typically represented by exponential decay that acts against flux-driven biosynthesis (210). However, sometimes it may be desirable to account for a stagnation of metabolism and growth if energy harvest per cell falls below a certain threshold. For example, in constraint-based metabolic cell models (354) this occurs automatically when the solution space becomes empty due to a fixed ATP flux representing cell maintenance requirements (538). Here we show how maintenance requirements with similar thresholds can be incorporated into the reaction-based model framework introduced in the main text.
We assume that maintenance requirements impose a constant cost on a cell’s growth and metabolism that can be represented by a specific maintenance rate $\lambda_s$, where $s$ denotes a particular cell species (210). Hence, the net population growth is given by the expression

$$\frac{dN_s}{dt} = \sum_{r \in s} Y_r H_r - \lambda_s N_s,$$

as long as this expression is positive, where $N_s$ is the cell density and “$r \in s$” indicates that the sum only covers the reactions performed by the particular species. Note that the decay rate $\lambda_s$ only represents an offset in biosynthetic yield due to maintenance requirements and does not account for cell lysis or washout from a bioreactor. When the right hand side in Eq. (G.2.1) becomes negative, i.e., when maintenance requirements exceed yield, dissimilatory metabolism is assumed to halt.

Recall that $H_r = N_s V_r h_r$, where $h_r$ are normalized reaction kinetics. Hence, Eq. (G.2.1) is positive exactly when

$$\sum_{r \in s} N_s V_r Y_r h_r > \lambda_s N_s.$$  \hspace{1cm} (G.2.2)

Also recall that $V_r Y_r = A_r$ is the self-amplification factor for reaction $r$, so that condition (G.2.2) translates to

$$\sum_{r \in s} A_r h_r > \lambda_s.$$  \hspace{1cm} (G.2.3)

For example, if $ure$ and $amo$ are performed by the same AOB cells, condition (G.2.3) becomes

$$A_{ure} h_{ure} + A_{amo} h_{amo} > \lambda_{AOB},$$  \hspace{1cm} (G.2.4)

as used in Eq. (8.4.3) in the main text. In the special case where species $s$ only performs one reaction $r$, condition (G.2.3) simplifies to $A_r h_r > \lambda_r$.

### G.2.2 On the concentration of organic components

Here we derive the general formula for the concentration ($X$) of a particular biomass component (e.g., organic N or cell wall proteins; in short, “compound”) within the context of the reaction-centric model described in the main text (Eq. (8.3.6)). We assume that the amount
of compound per cell only depends on the cell species, but is otherwise constant. Hence,

\[ X = \sum_r \frac{N_r \phi_r}{m_r}, \]  

(G.2.5)

where \( N_r \) is the concentration of cells performing reaction \( r \), \( \phi_r \) is the amount of focal compound per cell and \( m_r \) is the total number of reactions performed by cells performing reaction \( r \). Note that \( N_r, \phi_r \) and \( m_r \) are the same for any reactions performed by the same cells. We therefore divide each term in Eq. (G.2.5) by \( m_r \) to correct for multiple counting of the same cell species. Next, recall that \( M_r = N_r V_r \) and that \( A_{\rho r} = V_{\rho r} Y_r \) whenever the reactions \( r \) and \( \rho \) are performed by the same cell species. As is shown in detail in Section G.2.3 below, this means that there exists at least one vector \( w \) such that \( Aw = M \), or formally "\( w = A^{-1}M \)". Note that \( w \) is not always uniquely defined. Specifically, if multiple reactions are performed by the same cells then the matrix \( A \) is not invertible and there may exist multiple solutions \( w \) to the equation \( Aw = M \). However, it can be shown that any such vector \( w \) satisfies

\[ \sum_r \frac{N_r \phi_r}{m_r} = T_r w_r = T^T w, \]  

(G.2.6)

were we defined \( T_r = \phi_r Y_r \) (see Eq. (G.2.13) in Section G.2.3). Note that Eq. (G.2.6) holds regardless of the exact choice of \( w \). Hence, we can write in matrix notation

\[ X = T^T A^{-1} M, \]  

(G.2.7)

as mentioned in the main text. The coefficient \( T_r \) can be interpreted as the assimilation factor for reaction \( r \), i.e., as the amount of compound assimilated or synthesized per reaction flux (e.g., mol compound per mol reaction flux). For example, the stoichiometry of dissimilatory and assimilatory N-metabolism of *Nitrosomonas europaea* is conventionally summarized by the following equation:

\[ 55\text{NH}_4^+ + 76\text{O}_2 + 109\text{HCO}_3^- \rightarrow \text{C}_5\text{H}_7\text{NO}_2 + 54\text{NO}_2^- + 57\text{H}_2\text{O} + 104\text{H}_2\text{CO}_3, \]  

(G.2.8)

where \( \text{C}_5\text{H}_7\text{NO}_2 \) represents biomass (520). Hence \( T_{amo} = 1 : 55 \approx 0.018 \) for organic N, where \( amo \) represents the dissimilatory reaction

\[ \text{NH}_4^+ + 1.5\text{O}_2 \rightarrow \text{NO}_2^- + \text{H}_2\text{O} + 2\text{H}^+. \]  

(G.2.9)
Note that if two reactions \( r \) and \( \rho \) are performed by the same cells, \( T_r \) and \( T_\rho \) must satisfy the consistency condition

\[
\frac{T_r}{T_\rho} = \frac{A_{rr}}{A_{r\rho}},
\]

stemming from the fact that \( \phi_r = \phi_\rho \) and \( A_{r\rho} = V_rY_\rho \).

### G.2.3 Properties of the amplification matrix

This section is a reference summary of technical details on the amplification matrix \( A \) of the reaction-centric model. It is only provided for the mathematical completeness of calculations presented elsewhere in the Appendix.

We denote by \( V \) and \( Y \) the column vectors containing the maximum cell-specific reaction rates \( V_r \) and the cell yield factors \( Y_r \). We denote by \( A^o \) the matrix whose entries \( A^o_{r\rho} \) are 1 if reactions \( r \) and \( \rho \) are performed by the same cell, and 0 otherwise. We also denote by \( V \) and \( Y \) the diagonal matrices whose diagonals are given by \( V \) and \( Y \), respectively. Hence, \( A \) can be written as \( A = VA^oY \). The following assertions hold:

1. \( A^o \) and \( A \) are invertible if and only if no two reactions are performed by the same cell (i.e., \( A^o \) is diagonal).
   **Proof:** If two reactions are hosted by the same cell, then at least two rows of \( A^o \) are identical. Hence, \( \det(A^o) = 0 \), so \( A^o \) and \( A \) are not invertible.

2. Let \( \phi \) and \( \nu \) be two vectors such that \( A^o\nu = 0 \), and such that \( \phi_r = \phi_\rho \) whenever the reactions \( r \) and \( \rho \) are performed by the same cell (i.e., \( \phi_r = \phi_\rho \) whenever \( A^o_{r\rho} = 1 \)). Then \( \phi^T\nu = 0 \).
   **Proof:** From standard matrix theory we know that \( \mathcal{N}(A^o)^\perp = \mathcal{R}(A^{oT}) \), where \( \mathcal{N} \) and \( \mathcal{R} \) denote the null space and image space (a.k.a. column space), respectively, and \( \perp \) denotes the orthogonal complement space. Note that \( A^o = A^{oT} \), so that \( \phi \in \mathcal{N}(A^o)^\perp \) whenever \( \phi \in \mathcal{R}(A^o) \). Finally, note that \( \mathcal{R}(A^o) \) is spanned by precisely those vectors \( \phi \) satisfying \( \phi_r = \phi_\rho \) whenever \( A^o_{r\rho} \neq 0 \).

3. Let \( \phi, y \in \mathcal{R}(A^o) \). Let \( z \) be any vector satisfying \( A^o z = y \). Then
   \[
   \phi^Tz = \sum_r \frac{\phi_r y_r}{m_r},
   \]
   where \( m_r = \sum_\rho A^o_{r\rho} \) is the number of reactions performed by cells performing reaction
Proof: Choose $x$ in the following way: $x_r := y_r/m_r$. Then $A^o x = y$. Hence, $z = x + \nu$ for some $\nu \in N(A^o)$. By the previous assertion one has $\phi^T \nu = 0$, hence

$$\phi^T z = \phi^T x + \phi^T \nu = \phi^T x = \sum_r \phi_r x_r = \sum_r \phi_r y_r/m_r,$$

(G.2.12)

as claimed.

4. Let $\phi, N \in R(A^o)$. Let $M_r = V_r N_r$ and $T_r = \phi_r Y_r$. Then $M \in R(A)$. Moreover, for any vector satisfying $A w = M$ one has

$$\sum_r \frac{\phi_r N_r}{m_r} = T^T w,$$

(G.2.13)

regardless of the exact choice of $w$.

Proof: Recall that $A = V A^o Y$. Also note that since $N$ is in $R(A^o)$, $M$ is in $R(V A^o)$. Since $Y$ is invertible, one has $R(V A^o) = R(V A^o Y) = R(A)$, hence $M$ is in $R(A)$ as claimed. Define $z_r = Y_r w_r$. Then

$$A^o z = V^{-1} A Y^{-1} z = V^{-1} A w = V^{-1} M = N.$$

(G.2.14)

Eqs. (G.2.11) and (G.2.14) imply that

$$\sum_r \frac{\phi_r N_r}{m_r} = \phi^T z = \sum_r \phi_r Y_r w_r = \sum_r T_r w_r = T^T w,$$

(G.2.15)

as claimed.
Figure G.1: Comparison of models for the incubation experiment involving urea hydrolysis and nitrification, with (top row) and without (bottom row) explicitly accounting for partial oxidation of NH$_4^+$ produced by urea hydrolysis in the same cells (NH$_4^+$ “recycling”). Continuous curves show model predictions for urea (left column), NH$_4^+$ (center column) and NO$_3^-$ concentration (right column), compared to experimental data (circles; 94). While both models predict urea concentrations with similar accuracy, the 2nd model fails to explain the early nitrification of NH$_4^+$ to NO$_3^-$ (sub-figures E,F). On the other hand, including partial NH$_4^+$ recycling improves the model’s agreement with the NO$_3^-$ time series (sub-figures B,C; log-likelihood = 41.3 with recycling vs 21.2 without). Despite the additional complexity of the first model (6 fitted parameters instead of 5), statistical model selection criteria show a clear preference for the inclusion of partial recycling (AIC = −70.7 and BIC = −62.3 with recycling vs AIC = −32.4 and BIC = −25.4 without; 246).
Figure G.2: Comparison of models for the incubation experiment involving urea hydrolysis and nitrification, with (top row) and without (bottom row) ure-amo cross-amplification. Curves show model predictions for urea (left column), NH$_4^+$ (center column) and NO$_3^-$ concentration (right column), compared to experimental data (circles; 94). Continuous and dashed curves in (D–F) show two typical alternative model fitting outcomes (local and global fitting optima, respectively), obtained in the absence of ure-amo cross-amplification. Fitting attempts starting at random parameter values repeatedly converged to one of the two optima, which either completely fail to predict nitrification (continuous curve) or yield unrealistically high estimates (i.e. > 10 d$^{-1}$) for the amo self-amplification factor (dashed curve). All model calibrations without ure-amo cross-amplification achieved a substantially lower match to the data (best log-likelihood = 24.6, dashed curve in bottom row) than the model with ure-amo cross-amplification (log-likelihood = 41.3, top row).