Dynamics of complex ecological communities

Von den Fakultäten für Mathematik und Naturwissenschaften
der Carl von Ossietzky Universität Oldenburg
zur Erlangung des Grades und Titels einer
Doktorin der Naturwissenschaften (Dr. rer. nat.)
angenommene Dissertation

von

Frau Jana C. Maßing

geboren am 23.05.1995 in Münster

Gutachter: Prof. Dr. Thilo Gross

Weiterer Gutachter: Prof. Dr. Bernd Blasius

Tag der Disputation: 08.01.2024

Acknowledgements

It is a pleasure to thank the many people that have supported and inspired me throughout my PhD journey. First and foremost, I would like to thank my supervisor Thilo Gross. Thilo has introduced me to the fascinating world of modeling, nonlinear dynamics and networks and enabled me to work in this interdisciplinary context. Thilo always provided support and inspiration for this work. His lectures and talks have fascinated me and taught me so much. I have gained valuable insights not only in the realm of science but also in the art of delivering engaging presentations and writing interesting papers. Also, Thilo gave me the opportunity to present and discuss my research with many other scientists.

I would also like to thank Ashkaan Fahimipour who supported me a lot during this work. Even though we were quite distant in location and time, he was always available for a meeting and supported me during my first adventures into the realm of super-computing. Ashkaan also provided me with the opportunity to visit his lab in Florida, which was a great experience.

I have greatly profited from discussions with Carina Bunse, who has deep insights into the dynamics of the Baltic Sea microbiome and on microbes in general.

I would also like to thank Bernd Blasius and Helmut Hillebrand who always gave me new inspiration and helpful advice during my Thesis Advisory Committee meetings.

I have to thank the whole HIFMB family for the friendly and always supporting atmosphere. Special thanks to Ruth Krause, Christel Hoffmann and Rainer Pigge for always supporting us as young researchers in the academic system. I would also like to thank the Claudias from Polmar and Ferdinand from Oltech for their work in guiding us through this journey. A heartfelt thanks to Thilo's model masters, especially Pedro, Jordan, Rana and Laura for the emotional, funny, awesome lunch and coffee times and activities besides work that recharged my batteries. Many thanks to the AWIs4Future Team for their campaigns and enthusiasm in science communication, connecting science to society, fostering exchange, and addressing global challenges.

Special gratitude goes to my partner Mark for always being there for me, supporting and encouraging me so much and believing in me. My friends provided a lot of support during this time, especially Jonas and Kaveh. I would also like to thank my dad who always encouraged me in my interest in science. Lastly, I would like to thank Armin and Renate for being a source of light during challenging times, providing support, and helping me believe in myself.

Summary

Ecological communities not only frequently harbor an enormous species diversity but also, as a result of the multitude of interactions, regularly show complex nonlinear dynamics in community composition. While we have made progress in understanding communities of macroorganisms, the structure and dynamics of microbial communities poses a new challenge and calls existing theoretical concepts into question.

In the face of climate change and biodiversity loss and given the critical role that microbial communities play in sustaining life on Earth, one of the most important challenges of our time is to understand how anthropogenic influences are changing these communities and consequently impacting ecosystem functioning.

Achieving a mechanistic understanding of microbial community dynamics encounters obstacles in data-driven as well as model-led approaches. The curse of dimensionality complicates the extraction of important information from datasets, while uncertainties in functional forms and rate laws in microbial communities impede the formulation of mechanistic models.

In this thesis, we work towards a better understanding of complex microbial communities through a twofold approach. On the one hand we address the challenge of analyzing high-dimensional datasets and extracting key variables. On the other hand, we apply an alternative modeling approach to develop a mechanistic understanding of systems characterized by high uncertainties.

Applying a framework based on diffusion maps to a long-term bacterial time series from the Baltic Sea, we are able to identify key variables, i.e. metabolic niches in this dataset. This allows us to coarse-grain the bacterial communities in terms of their metabolic niches. Identifying these niches enables us to transform the species time series into potentially occupied metabolic niches over time. Thereby this approach reduces dimensionality by identifying key variables, that are nonlinear combinations of the measured variables and establishes a framework to link the data to a fundamental ecological concept, the niche.

In the second part of this thesis, we employ a generalized modeling approach to study a common motif of bacterial communities, i.e. mutualistic cross-feeding. While these mutualistic interactions are typically anticipated to be destabilizing, real-world observations of their common occurrence in bacterial communities appear to contradict these predictions. Applying the generalized modeling approach we can efficiently analyze a large range of plausible models and derive analytical results. We identify metabolic costs of trade metabolite production as an important factor that contributes to stability of these systems. In summary, the advancements made here in both, data analysis and modeling of microbial communities, contribute to develop an understanding of the mechanisms behind community composition, dynamics over time and the relation to ecosystem functioning.

Zusammenfassung

Ökologische Gemeinschaften beherbergen nicht nur häufig eine enorme Artenvielfalt, sondern zeigen aufgrund der Vielzahl von Interaktionen auch regelmäßig komplexe nichtlineare Dynamiken in der Gemeinschaftszusammensetzung. Während wir Fortschritte bei der Erforschung von Makroorganismen-Gemeinschaften gemacht haben, stellt die Struktur und Dynamik von mikrobiellen Gemeinschaften eine neue Herausforderung dar und stellt bestehende theoretische Konzepte in Frage.

Angesichts des Klimawandels und des Verlusts an Biodiversität sowie der entscheidenden Rolle, die mikrobielle Gemeinschaften für das Leben auf der Erde spielen, ist eine der wichtigsten Herausforderungen unserer Zeit, zu verstehen, wie anthropogene Einflüsse diese Gemeinschaften verändern und folglich die Funktionsweise von Ökosystemen beeinflussen.

Die Erlangung eines mechanistischen Verständnisses der Dynamik mikrobieller Gemeinschaften stößt sowohl bei datengestützten als auch bei modellgestützten Ansätzen auf Hindernisse. Der Fluch der Dimensionalität erschwert die Gewinnung wichtiger Informationen aus Datensätzen, während Unsicherheiten in funktionalen Formen und Ratengesetzen in mikrobiellen Gemeinschaften die Formulierung mechanistischer Modelle erschweren.

In dieser Arbeit streben wir ein besseres Verständnis komplexer mikrobieller Gemeinschaften durch einen zweigeteilten Ansatz an. Einerseits gehen wir die Herausforderung an, hochdimensionale Datensätze zu analysieren und Schlüsselvariablen zu extrahieren. Andererseits wenden wir einen alternativen Modellierungsansatz an, um ein mechanistisches Verständnis von Systemen zu entwickeln, die durch große Unsicherheiten gekennzeichnet sind

Durch die Anwendung einer auf Diffusion Maps basierenden Methode auf eine Langzeitdatenreihe von Bakteriengemeinschaften der Ostsee können wir Schlüsselvariablen, d.h. metabolische Nischen in diesem Datensatz, identifizieren. Dadurch können wir die bakteriellen Gemeinschaften im Hinblick auf ihre Stoffwechselnischen grob einteilen. Die Identifizierung dieser Nischen ermöglicht es uns, die Zeitreihen der Arten in potenziell besetzte metabolische Nischen im Laufe der Zeit zu übersetzen. Auf diese Weise reduziert diese Methode die Dimensionalität durch die Identifizierung von Schlüsselvariablen, die nichtlineare Kombinationen der gemessenen Variablen sind, und schafft einen Rahmen, um die Daten mit einem grundlegenden ökologischen Konzept, der Nische, zu verknüpfen.

Im zweiten Teil dieser Arbeit verwenden wir eine Methode zur Formulierung allgemeiner Modelle, um ein häufiges Motiv bakterieller Gemeinschaften zu untersuchen, nämlich die wechselseitige Nährstoffzufuhr. Während diese mutualistischen Interaktionen typischerweise als destabilisierend beurteilt werden, scheint das häufige Vorkommen dieser in bakteriellen Gemeinschaften dem zu widersprechen. Durch die Anwendung der Methode zur Formulierung allgemeiner Modelle können wir eine große Bandbreite plausibler Modelle effizient analysieren und analytische Ergebnisse ableiten. Wir konnten zeigen, dass die metabolischen Kosten der Produktion von Nährstoffen für den Partner wichtige Faktoren sind, die zur Stabilität dieser Systeme beitragen. Zusammenfassend lässt sich sagen, dass die hier erzielten Fortschritte sowohl bei der

Datenanalyse als auch bei der Modellierung mikrobieller Gemeinschaften dazu beitragen, die Mechanismen hinter der Zusammensetzung der Gemeinschaften, ihrer Dynamik im Laufe der Zeit und ihre Auswirkungen auf Ökosystemfunktionen besser zu verstehen.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Imagine an alien visiting Earth for the first time. Robert May pointed out that the first question of this visitor might be "How many distinct life forms - species - does your planet have?" [1]. Embarrassingly, we could still only provide a very rough estimate, for example that there are probably $\sim$ 8.7 million eukaryotic species [2]. More than 60 years ago, Hutchinson expressed his astonishment at the enormous diversity of animals and plants, asking how such a number of species can exist in the light of competitive exclusion [3]. A little later he formulated the paradox of the plankton, wondering how these organisms can be so diverse in the face of apparently limited resources [4].

Today, the estimated diversity of microbial species on Earth, which probably exceeds 1 trillion species, extends this apparent contradiction to microbes and makes us wonder if we can ever answer the question of how many life forms inhabit Earth [5]. Recent advancements in molecular methods, enabling the study of uncultivated and formerly unknown organisms [6], reveal the dominance of bacterial diversification in the tree of life, the Domain Bacteria encompasses more major lineages of organisms than any other Domain [7].

Not only are there an enormous number of organisms on this planet, these "endless forms most beautiful" are also entangled, interacting with each other in various ways, coming together in complex ecological communities [8]. Robert May opens our eyes for how special these communities are in showing that large random communities are almost always unstable (putting random organisms together is unlikely to lead to a stable community) [9]. Hence, the communities that we find in nature are these special combinations of a variety of organisms.

Through their diversity in genetic material, metabolic and physiological activity, microbial communities are capable of performing sophisticated and diverse chemical reactions, driving major biogeochemical cycles [10], and engaging in diverse interactions among each other and with macroorganisms e.g. [11, 12, 13]. Microbes inhibit almost every corner of Earth, encompassing the vast majority of the Earth's biological and evolutionary va-

riety [6]. Their diverse capabilities make them crucial for ecosystem functioning and ultimately critical for life on Earth.

Presently, the human population is drastically impacting the environment, causing the twin crisis of climate change and biodiversity loss. Due to the importance of microbial communities for life on Earth, two of today's most important challenges are to understand and predict how the anthropogenic influence alters microbial communities and how these changes in microbial community composition in turn affect ecosystem functioning [14, 15].

While we have made progress in understanding communities of macroorganisms, the structure and dynamics of microbial communities poses a new challenge and changes how we think about biology. For example, short generation times in microorganisms can lead to fast changes in microbial community composition and rapid evolution [16], horizontal gene transfer and asexual reproduction in microorganisms challenge the species concept [17] and the multitude and crucial role of interactions between microbes and multi-cellular organisms has led to the inception of the holobiont or more generally the metaorganism concept [18, 19].

Progress in molecular methods, especially the ability to study also non-culturable microorganisms, has given rise to the availability of large datasets of microbial species, gene and protein diversity and their (potential) functions [6, 20]. It has become possible to monitor microbial community composition over time, establishing short and long-term time series, enabling the recording of microbial diversity over time [21]. These revealed that in addition to their enormous diversity overall, microbial communities are dynamic, often undergoing complex dynamic fluctuations in abundances and species composition on the daily, monthly and annual scale [22, 23, 24].

The collection of extensive datasets documenting microbial diversity worldwide, such as the Earth microbiome project [25] and the Tara Oceans project [26], has played a crucial role in identifying significant environmental factors that shape overarching patterns of microbial composition. These factors include for example pH levels, nutrient availability, climate and light conditions [23, 27, 28, 29]. These approaches have revealed aspects of the interplay between microbial community composition and ecosystem functioning and response to environmental change, emphasizing the importance of changes in microbial community composition for ecosystem functions [30, 31, 32].

Experimental and field studies of microbial communities on a smaller scale shed light on the importance of interactions among microbes and with other organisms on driving community dynamics [22, 33]. These interactions range from resource competition and cross-feeding to viral lysis, grazing, toxin production and employment of mechanical

weaponry [23, 34, 35, 36]. Also, microbial interactions occur on multiple spatial scales, for example metabolite-based interactions may happen in neighboring cells or metabolites may be released, spreading as public goods that may eventually be taken up by other microbes in other locations [37].

Every endeavor to elucidate the underlying mechanisms of microbial community composition and its relation with ecosystem function is complicated by the complexity of these communities, i.e. their enormous species diversity, their complex nonlinear dynamics and their multitude of interactions. Descriptive studies still largely dominate the literature on microbial community ecology, providing species or gene sequence inventories or descriptions of the effect of environmental factors on the community composition [14]. Mechanistic understanding of microbial community composition and functions is still largely limited to small communities or specific functions studied in isolation [38].

To foster understanding of how complex microbial communities function, we need two essential ingredients: First conceptual frameworks to makes sense of the high-dimensional datasets that are available. Second a modeling framework that can deal with the complexities and the many uncertainties that characterize microbial communities.

Existing approaches reduce the dimensionality of large datasets of microbial communities by coarse-graining, for example at some taxonomic level [39], into functional groups derived by literature analysis [40] or along niche gradients obtained by linear dimensionality reduction methods [41]. In most cases these ways of coarse-graining constitute an approximation, i.e. information is probably lost and nonlinear relationships may be approximated by linear representations.

Commonly employed dimensionality reduction techniques like PCA (Principal Component Analysis) [42], PCoA (Principal coordinates analysis), MDS (multidimensional scaling), and t-SNE (t-Distributed Stochastic Neighbor Embedding) [43] have offered valuable insights. However, it is important to note their limitations [44, 45, 46]. While PCA detects certain global structures at the expense of local details [47], t-SNE finds some local structures at the cost of not reliably identifying global structures [44, 45]. MDS is susceptible to noise, often causing failure in generating insightful visualizations for complex, nonlinear data [44, 45]. The quality of the PCoA output suffers in particular if for example species turnover is high between sampled communities [46].

A recent study by Fahimipour and Gross [44] demonstrates an alternative: Diffusion maps [48], a de-facto parameter-free, nonlinear dimensionality reduction method yields new variables from bacterial metabolic networks, that can be interpreted as composite functional strategies of bacteria. In addition, Ryabov et al. [49] show that diffusion maps

can infer functional trait axes from monitoring data and thereby enable an estimation of functional diversity.

In this work, I extend the diffusion map approach to quantitatively organize genomic information into potentially occupied bacterial metabolic niches over time. I apply the method to reconstruct the dynamics of putatively occupied metabolic niches using a long-term bacterial time series from the Baltic Sea.

In terms of modeling frameworks, interesting insights have been gained by combining methods from statistical physics, nonlinear dynamics and complex systems theory. Combining a generalized Lotka-Volterra model framework with random matrix theory, Coyte et al. [50] study different kinds of microbial interactions and show that cooperation within microbial communities has a destabilizing effect that can be counteracted by competitive interactions that weaken the positive feedback loop between the interacting partners. Other approaches consider explicitly the resource-mediated interactions, building on the classic consumer-resource model by MacArthur [51, 52, 53]. Butler and O'Dwyer [54, 55] for example show that the incorporation of resources that mediate interactions can change the conclusions about system's stability. Integrating a coarse-grained metabolism into a consumer-resource model, Muscarella and O'Dwyer [56] demonstrate that metabolic rates and the distribution of resources can alter interactions between species and hence can change species dynamics.

While the generalized Lotka-Volterra modeling framework has the advantage of requiring only few parameters and being very flexible in incorporating a mixture of interactions, one of its limitations is that is lacks the ability to consider resource-mediated interactions that are prevalent in microbial communities [57]. Modeling approaches following the consumer-resource framework capture these interactions, however they often deal with a large number of unknown parameters [38]. Since the majority of bacteria can still not be cultured [58, 59], the precise rate laws and functional forms of ecological interactions in microbial communities are mostly unknown. Often Holling- [60] Hill- [61] and Monod-type [62] functions are applied in microbial community models [54, 55, 56, 63, 64, 65]. Although they are mathematical simple and satisfy some basic biological requirements, functions with similar shapes describe certain phenomena just as well, however the effects on stability may differ significantly [66]. In addition, estimating their parameters requires extensive time series and model-fitting procedures which aren't feasible even for simple microbial communities [67].

In this thesis I explore the generalized modeling (GM) approach [66, 68] for microbial communities. GM captures the structure of the system without restricting it to specific

functional forms, enabling us to analyze a whole class of systems in parallel. A special parameterization procedure yields easily interpretable parameters that can be robustly estimated from noisy empirical data [68, 69].

I apply this modeling approach to study a common motif in bacterial communities, cross-feeding of two types of bacteria that affect each other by producing and releasing chemical metabolites that the other uses for growth and metabolism. In particular, I ask which factors stabilize or destabilize these bacterial cross-feeding communities. Stability of mutualistic interactions is a hotly debated topic in microbial ecology. Despite that commensal and mutualistic interactions are predicted to be destabilizing [70] unless specific conditions are met [54, 55], we observe apparently stable [37, 71] environmental and host-associated microbial communities comprising opportunistic or obligate mutualists. Applying GM, I can study this bacterial community motif with a high degree of generality, allowing us to draw general conclusions about these systems.

I start in Chapt. 2 with introducing important concepts and tools that I use in this work. After I give a brief introduction to dynamics and bifurcations, I provide a short overview of networks and diffusion on networks. Then, I focus on diffusion maps in more detail as this constitutes one of the key methods that I apply in this work. In particular, I show how diffusion maps can overcome the so-called curse of dimensionality and identify major explanatory variables in high-dimensional datasets.

In Chapt. 3 I review the generalized modeling (GM) approach, providing a guide to the GM procedure as well as to the analysis of generalized models. Specifically, I show how the GM approach can deal nicely with uncertainties in the system, offering insights on a large class of systems. Additionally, I introduce the master stability function approach which extends the GM approach to spatial systems.

In Chapt. 4, I apply the diffusion map method to a long-term dataset of relative abundances of prokaryotic populations obtained from amplicon sequencing data of the Baltic Sea. The diffusion map reveals a broad range of metabolic strategies of the analyzed taxa and a distinct shape of the metabolic niche space. Using these newly identified variables, I extend the diffusion map approach such that it allows the translation of species time series into potentially occupied metabolic niches over time. Thereby, I can observe the potential niche occupation dynamics, which reveal interesting niche participation as well as patterns over time related to environmental conditions.

In Chapt. 5 I study a common cross-feeding motif of bacterial systems. I propose a simple generalized model of two cross-feeding types of bacteria, which enables me to study the stability properties for a large range of parameter values. Through the efficient analysis

procedure, I can for example analyze systems, where bacteria exchange costly, cheap or even both types of metabolites. I identify a key factor that contributes to stability in our model system.

In Chapt.6 I outline some future directions towards developing a mechanistic understanding of microbial community ecology. Using large datasets, I show the effectiveness of an aggregated diffusion map analysis leading to improved results and potentially a comprehensive global map of community characteristics and functional diversity. In addition, I introduce the concept of community maps to identify variables that influence changes in functional community composition. Finally, I address synthetic metabolic interaction networks that link reaction chemistry to bacterial substrate interactions and point the way to a potential niche model for bacteria.

Finally, in Chapt. 7 I conclude this work by summarizing the results and expressing my hope for future research. In particular, I envision the combination of diffusion maps and GM to develop a new way of modeling complex microbial communities.

Chapter 2

Concepts and Tools

The purpose of this work is to develop new approaches that capture the dynamics of complex communities by combining concepts and tools from dynamical systems theory and network science. In this chapter I introduce important concepts and tools. I begin in Sec. 2.1 with an introduction to dynamics and bifurcations. Then, I provide a short overview about networks and diffusion on networks (Sec. 2.2). In Sec. 2.3 I present the nonlinear dimensionality reduction method, diffusion maps, which is one of the key methods that I will come back to in the following chapters.

2.1 Dynamics and bifurcations

In this section we briefly introduce several fundamental concepts of the theory of dynamical systems without delving into mathematical rigor. These topics are fundamental and are commonly covered in most textbooks, e.g. Kuznetsov [72], Strogatz [73] and Guckenheimer and Holmes [74]. This section provides a short introduction to readers unfamiliar with the field. We start with an introduction to dynamical systems and steady states (Sec. 2.1.1). Thereafter, we explain the concept of local stability of steady states (Sec. 2.1.2) and finally outline the notion of bifurcations (Sec. 2.1.3).

2.1.1 Dynamical systems and steady states

Dynamical systems are used across many scientific disciplines to describe and predict the behaviour of systems that change in time. For example, in an ecosystem we could be interested in how the number of different species changes over time. To a certain extent we can predict the past and future state of some systems given the current state and prescriptions for the evolution of the system.

A dynamical system comprises a set of state variables and the prescriptions for the evolution of the state in time, with time serving as the independent variable [72]. The state variables can be interpreted as coordinates in the space of all possible states, the phase

space, and the prescriptions specify the trajectory of the system through phase space. Commonly, these prescriptions are formulated in the form of differential equations or discrete-time maps.

In this work, we focus on the study of dynamical systems that are described by ordinary differential equations (ODEs). The general form of an n-dimensional ODE can be written as

$$\dot{x}_n = f_n(x_1, ..., x_N, p_1, ..., p_M) \tag{2.1}$$

with $n = 1, ..., N$, where N is the total number of state variables. $x_1, ..., x_N$ is the set of state variables that change in time. The functions $f_1, ..., f_N$ describe the change of the state variables, i.e. dependent variables in time, depending on the parameters $p_1, ..., p_M$, where M is the total number of parameters. Sets of state variables for which the variables do not change in time are called steady states x^*,

$$f_n(x_n^*, p_m) = 0. \tag{2.2}$$

2.1.2 Stability of steady states

In natural systems, we are often interested in how the system would react to perturbations, for example how an ecosystem would react to a brief period of elevated temperature (pulse perturbation). Mathematically, we can determine the local stability of steady states. If a steady state is locally stable, then small perturbations will decay and the system variables will return to the steady state values.

Let's consider a system with multiple variables

$$\dot{\boldsymbol{x}} = f(\boldsymbol{x}) \tag{2.3}$$

that is subject to a small perturbation from the steady state

$$\boldsymbol{x} = \boldsymbol{x}^* + \boldsymbol{\delta}, \tag{2.4}$$

where we denote vectors in bold italic. Substituting into the differential equation system yields

$$\dot{\boldsymbol{x}} = \dot{\boldsymbol{\delta}} = f(\boldsymbol{x^*} + \boldsymbol{\delta}). \tag{2.5}$$

Using Taylor expansion, we obtain

$$\dot{\boldsymbol{\delta}} = f(\boldsymbol{x}^*) + \mathbf{J}\boldsymbol{\delta} + O(\boldsymbol{\delta}^2), \tag{2.6}$$

where $\mathbf{J} \in \mathbb{R}^{\mathrm{nxn}}$ denotes the *Jacobian matrix* with the elements

$$J_{i,j} = \frac{\partial}{\partial x_j}\dot{x}_i\bigg|_{\boldsymbol{x}=\boldsymbol{x}^*}, \tag{2.7}$$

where $i, j = 1, ..., N$ and $O(\boldsymbol{\delta}^2)$ denotes quadratically small terms in $\boldsymbol{\delta}$, that are negligible if $\mathbf{J}\boldsymbol{\delta} \neq \mathbf{0}$ and $f(\boldsymbol{x}^*) = \mathbf{0}$ since $\boldsymbol{x}^*$ is a steady state. Close to the steady state, we can hence approximate the system by the linear system

$$\dot{\boldsymbol{\delta}} = \mathbf{J}\boldsymbol{\delta}. \tag{2.8}$$

Since the matrix entangles different variables, the solution is to define new variables which do not get entangled when they pass through the matrix. Let v_n be eigenvectors of $\mathbf{J}$ with corresponding eigenvalues λ_n, such that

$$\mathbf{J}v_n = \lambda_n v_n. \tag{2.9}$$

We can then write the perturbation as a linear combination of eigenvectors (Seperationsansatz),

$$\boldsymbol{\delta} = \sum_n c_n v_n, \tag{2.10}$$

where the coefficients c_n act as a new set of variables. Unlike the variables x_i they do not get entangled when they pass through the matrix.

The solution of the linear system is then

$$\boldsymbol{\delta}(t) = \sum_{n=1}^{N} c_n e^{\lambda_n t} v_n, \tag{2.11}$$

where c_n denotes expansion coefficients determined by the initial conditions. Equation (Eq.) 2.11 shows that the perturbation will grow exponentially in time if there is at least one eigenvalue $Re(\lambda_n) > 0$. If all eigenvalues of $\mathbf{J}$ have negative real parts then the perturbation will decline exponentially, i.e. the system will return to the steady state. Hence, a steady state is stable if $\mathrm{Re}(\lambda_\mathrm{n}) < 0$ for all n. For the case $\mathrm{Re}(\lambda_\mathrm{n}) = 0$ higher orders of the expansion have to be considered.

In other cases, we are interested how a system will respond to a press perturbation, i.e. a permanent change of parameters, such as an increase in temperature or the invasion of a new species. A sufficiently small perturbation leads to a shift in the steady states, that can be described by

$$\boldsymbol{\delta} = -\mathbf{J}^{-1}\boldsymbol{p}, \tag{2.12}$$

where $\boldsymbol{\delta}$ is the shift in the steady state, $\mathbf{J}^{-1}$ is the inverse of the Jacobian matrix, and $\boldsymbol{p}$ is a vector describing the direct impacts of the perturbations on the individual equations [69]. We will discuss the impact of a press perturbation in more detail in Sec. 3.4.

2.1.3 Bifurcations

We saw above that dynamical systems usually depend on a set of parameters

$$\dot{x}_n = f_n(x_1, ..., x_N, p_1, ..., p_M). \tag{2.13}$$

In natural systems we are often interested in the impact of the change of one or more parameters on the system, for example an increase in temperature. It is therefore of interest to study families of dynamical systems with different parameter values. Take a system with the parameter set $\boldsymbol{p}$ and a second system with a slightly different parameter set $\boldsymbol{p'}$. Generally, we expect the two systems to show very similar dynamical behavior. There are however sets of parameter values where the dynamical behavior changes qualitatively, e.g. the number of steady states or their stability change. The parameter sets at which such changes occur are called bifurcation points and the transition itself is a bifurcation [74].

Dynamical systems often depend on a multitude of parameters and their time evolution is usually governed by smooth functions of these parameters. In this case, the bifurcation points are not isolated, but are located on manifolds, consisting of bifurcation points. The codimension of a bifurcation is the difference between the dimension of parameter space and the dimension of the manifold on which that bifurcation occurs [72]. So, the codimension defines the number of parameters that need to be varied to find the respective bifurcation. In this work, we focus on codimension-1 bifurcations.

Since the Jacobian matrix is generally a real matrix, its eigenvalues are real or form complex conjugate pairs. Therefore, we can define two fundamental types of bifurcations: (1) If real eigenvalues become positive, this is called a Saddle-node-type bifurcation. (2) If a conjugate pair becomes positive, we call this a Hopf bifurcation [73]. In a Saddle-node type bifurcation typically two steady states collide and annihilate. However, if a certain symmetry is present, the bifurcation can appear as a transcritical bifurcation in which

two states intersect and exchange their stability. The Hopf bifurcation can be further distinguished: If a complex conjugate eigenvalue pair becomes positive and as a result a stable limit cycle is born as the steady state loses stability, this is a supercritical Hopf bifurcation. If instead an unstable limit cycle vanishes, the respective bifurcation is a subcritical Hopf bifurcation.

After identifying the bifurcations within a system, one can represent them visually through bifurcation diagrams. Commonly three-parameter bifurcation diagrams are used to get an overview of the interactions of these parameters [75]. Each point in the volume spanned by the bifurcation diagram represents a particular steady state and the bifurcation points form surfaces that separate qualitatively different steady states. This approach is also useful to identify regions where different bifurcations meet and intersect, which can reveal other interesting dynamical features of the system (see for example Zumsande et al. [76]).

2.2 Networks

Networks can represent systems composed of many parts which interact in various ways, such as species in ecosystems, individuals in social groups or metabolites in a cell [77]. In a network each element is depicted as a node and interactions are illustrated as links between nodes, resulting in a reduction of the complexity of the system's elements, while preserving the complexity of the interaction structure [78]. The arrangement of interactions within a network is termed its topology. If the interactions are reciprocal, the network is classified as undirected, for example consider a network of protein-protein interactions, where the nodes represent proteins and the links represent physical interactions between the proteins. If protein A interacts with protein B, then protein B also interacts with protein A, hence the relationship is symmetric and the network undirected. In contrast, the network is directed when interactions lack reciprocity. For instance, in the context of an energy flow within a food chain, where each species is represented as a node, and the directional flow of energy between them is indicated as a link. In this scenario, energy transfers from the consumed species to the predator species, but the flow does generally not occur in the reverse direction. In addition, a network can be characterized solely by the presence or absence of links, resulting in an unweighted network. Or it may incorporate information about the attributes or weights associated with each link, giving rise to a weighted network. Considering the examples from above, the protein-protein interactions could for instance be represented by a binary "yes" or "no", resulting in an unweighted network. In the energy flow network, we may quantify the amount of transferred energy, leading to a weighted network.

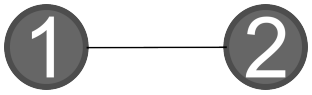

Figure 2.1: Example network of two nodes connected by one undirected link.

.

2.2.1 Network representation

To describe a network's topology, we can construct the adjacency matrix of the network [77, 79]. The adjacency matrix is a square matrix with dimensions corresponding to the size of the network. For every link in the network, e.g. from node i to node j the adjacency matrix has a nonzero entry at position $A_{i,j}$, whereas all other entries are set to zero, indicating the absence of a link. For an unweighted network, the entries are 1 (link present) or 0 (link absent), whereas for a weighted network, the adjacency entries represent the weights of the corresponding links. The adjacency matrix for the example network in Fig.2.1 is

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{2.14}$$

From the structure of a network we can also derive the Laplacian matrix [77, 80], which is closely related to the adjacency matrix. The Laplacian matrix is given by

$$\mathbf{L} = \mathbf{D} - \mathbf{A}, \tag{2.15}$$

where $\mathbf{D}$ is the degree matrix of the network. The degree matrix is a diagonal matrix that contains the sum over each row of $\mathbf{A}$ on its diagonal, i.e.

$$D_{i,i} = \sum_k A_{i,j}. \tag{2.16}$$

Considering again the example network in Fig.2.1, we obtain

$$\mathbf{D} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{L} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{2.17}$$

In this example, we can see that for undirected networks the degree matrix contains the information about the degree, i.e. the number of links connecting to the respective node.

2.2.2 Diffusion on networks

The Laplacian matrix is useful, if we investigate diffusion processes on a network. Let us again consider the example network in Fig. 2.1. Assuming we have a number of particles in node 1, denoted as x_1 and a number of particles in node 2, x_2, and these particles randomly jump from one node to the other at rate 1, we can write

$$\dot{x_1} = -x_1 + x_2 \tag{2.18}$$

$$\dot{x_2} = -x_2 + x_1. \tag{2.19}$$

In matrix notation this yields

$$\dot{\boldsymbol{x}} = -\mathbf{L}x, \tag{2.20}$$

where $x = \begin{pmatrix} x_1 & x_2 \end{pmatrix}^T$ and $\mathbf{L}$ is the Laplacian matrix from Eq. 2.17. Let us examine the scenario when the system is initiated by releasing a single particle at node 1. In this case, we can interpret x_1 and x_2 as the probabilities of finding the particles in node 1 or 2, respectively. We can use the eigenvalues of the Laplacian,

$$\lambda_1 = 0, \qquad \lambda_2 = 2 \tag{2.21}$$

and the corresponding eigenvectors

$$v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad v_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{2.22}$$

to write the initial state as

$$x(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{1}{2}v_1 + \frac{1}{2}v_2. \tag{2.23}$$

Applying the Separationsansatz, which we have introduced in the last chapter, the solution to the system is

$$x(t) = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{1}{2}e^{-2t}\begin{pmatrix} 1 \\ -1 \end{pmatrix}. \tag{2.24}$$

We can see that the differences between the two nodes become smaller and smaller in time and the second eigenvalue indicates the speed at which the system equilibrates.

If we consider a larger n-dimensional network, where particles travel between links at rate

1, the equation for particles in node i is

$$\dot{x_i} = -x_i \left(\sum A_{i,j} \right) + \left(\sum A_{i,j} x_j \right). \tag{2.25}$$

In matrix form this yields Eq. 2.20, where

$$L_{i,j} = \begin{cases} -A_{i,j} & \text{if } i \neq j \\ \sum A_{i,j} & \text{if } i = j \end{cases}. \tag{2.26}$$

We again solve the equation by eigendecomposition (see Sec. 2.1.2), obtaining a solution of the form

$$x(t) = \sum_n c_n(0) e^{-\lambda_n t} v_n, \tag{2.27}$$

where c_n are the expansion coefficients of an initial state $\boldsymbol{x}(0)$ in relation to the eigenvectors v_n of $\mathbf{L}$.

In a connected network as $t \to \infty$ all nodes approach the same state, since the Laplacian has exactly one eigenvalue $\lambda_1 = 0$ with the corresponding eigenvector $v_1 = (1, \ldots, 1)^T$ and all other eigenvalues are positive. The smallest non-zero eigenvalue determines the rate of equilibration and consequently corresponds to the most important eigenvector as the perturbations that excite the corresponding eigenvector will decay the slowest and hence have a long impact on the system. Differences in the entries of the most important eigenvectors therefore indicate a distance to each other. Due to these properties Laplacian matrices are employed in various applications, for example in spectral clustering and in nonlinear dimensionality reduction, as we will see in the next section.

2.3 Diffusion maps

While there is an increasing availability of high-quality ecological monitoring datasets and an increasing breadths of the information that we are able to capture, the complexity of these datasets poses new challenges to their analysis. Often, we are dealing with high-dimensional datasets, in which well-established data analysis methods fail due to the so-called curse of dimensionality. Diffusion mapping, a method introduced by Coifman et al. [48], is a powerful tool to reduce the dimensionality of these datasets via the detection of new explanatory variables.

We start in Sec. 2.3.1 with the concept of dimensionality reduction, followed by outlining the diffusion map procedure (Sec. 2.3.2). Thereafter, I provide examples of applications

of diffusion maps in ecological studies, explaining also the measure for functional diversity (Sec. 2.3.3). Finally, PHATE as a method to visualize high dimensional data is introduced (Sec. 2.3.4).

2.3.1 Dimensionality reduction

In ecological datasets, we often face a multitude of variables. Suppose we repeatedly sample a liter of seawater, each time we may record the population densities of several thousand bacterial species [81, 82]. In this case, the number of variables, i.e. the dimensionality of the data space equals the number of different variables that are recorded in the dataset, e.g. the number of detected species, hence several thousands. Advancements in molecular methods also enable the data collection of genomic data, gene expression data, protein structures etc. Take again the example of the liter of seawater, many of the detected species possess genomes containing probably at least several hundred genes [83]. Thus, we frequently find ourselves challenged by the high-dimensionality of those datasets.

In high-dimensional datasets , it is a challenge to make sense of the data, since for example, we cannot even consider all the extreme points anymore. Suppose we sample a bacterial community and record the relative population densities for each detected species over time. A system that is well described by one dimension is easy to understand, consider for instance that only one species would be enough to describe the system. In this case we could get a first impression of our dataset by considering the samples with the highest abundance of the species and the one with the lowest abundance. In the case of two species that matter, the number of extreme points would double, the dataspace has now four corners, the lowest value of species 1, the highest value of species 1, the lowest value of species 2 and the highest value of species 2. For every variable that we add the number or corners of the dataspace doubles. Let's say we find 4,000 different species, each sample would then map to a point in a 4,000 dimensional space. This space would have $2^{4,000}$ corners, which is greater than the estimated number of species on Earth of around 1-6 billion ($\sim 2^{30}$-2^{33}) [84]. As a consequence, we would not even be able to collect enough data to cover all the extreme points.

To overcome this so-called curse of dimensionality [85], we need to realize that the actual data points only cover a tiny part of the data space. We don't expect to find all the combinations for example of different species in a sample. Take again the continuous sampling of seawater, even though there are many different species, we will not see all possible combinations of species in the sampled communities. Instead species appearing together may be using similar resources, profit from certain interactions of benefit from the same environmental conditions. Due to these dependencies the data points may ap-

proximately trace a curve, a curved surface, or some other comparatively low-dimensional object within the data space. These underlying structures are summarized under the term manifold and the task of locating them is known as manifold learning.

Let's take a brief detour to understand this concept a bit more intuitively. As an illustrative example consider different types of houses, e.g. a single-family house, a villa or a skyscraper. Although numerous variables can be used to compare these different houses, most of the data space spanned by these variables is empty. Take for instance the variables number of floors and height of the building. It is likely that buildings with a large number of storeys are also characterized by a great height. Consequently, the datapoints along these two variables tend to show patterns, creating regions of empty space, such as those found at coordinates corresponding to low height and high number of floors or great height and low number of floors. With limited information like the number of floors we can therefore often accurately infer other properties, e.g. the approximate height of the building. The underlying structures of the data are manifolds and because the dimensionality of the manifold is lower than that of the embedding data space, manifold learning allows us to reduce the complexity of the data without losing information [86].

To describe our data in reduced dimensions we can hence define new explanatory variables that follow the main dimensions of the data manifolds. Take the abstract example shown in Fig. 2.2, the datapoints are characterized by two variables, but most of the dataspace is actually empty and the datapoints cluster around the spiral structure. Thus, we can reduce the number of variables, i.e. the dimensions, by describing the dataset just in terms of one new explanatory variable, that follows the main dimension of the data manifold (Fig. 2.2).

Principle component analysis (PCA), a de-facto manifold learning method, is widely applied to reduce the dimensionality of datasets [42]. PCA is a linear method that approximates the curved manifolds in the data as flat surfaces. These surfaces are constructed by a fitting procedure that is implemented as an algebraic operation. For this procedure it takes long-distance comparisons into account.

However, in high-dimensional datasets these long-distance comparisons become unreliable. We intuitively know that it is easier to compare similar things than it is to compare very dissimilar things. Comparing different types of houses is easy, comparing a house to a painting or to a cat is hard. This is because we do not know the traits that we should use to quantify very dissimilar things. To illustrate, consider the comparison of two species based on their gene similarity. If a species pair A,B is 99.9% similar in gene content, the 0.1% mismatch is a good quantification of the distance between the two species and we

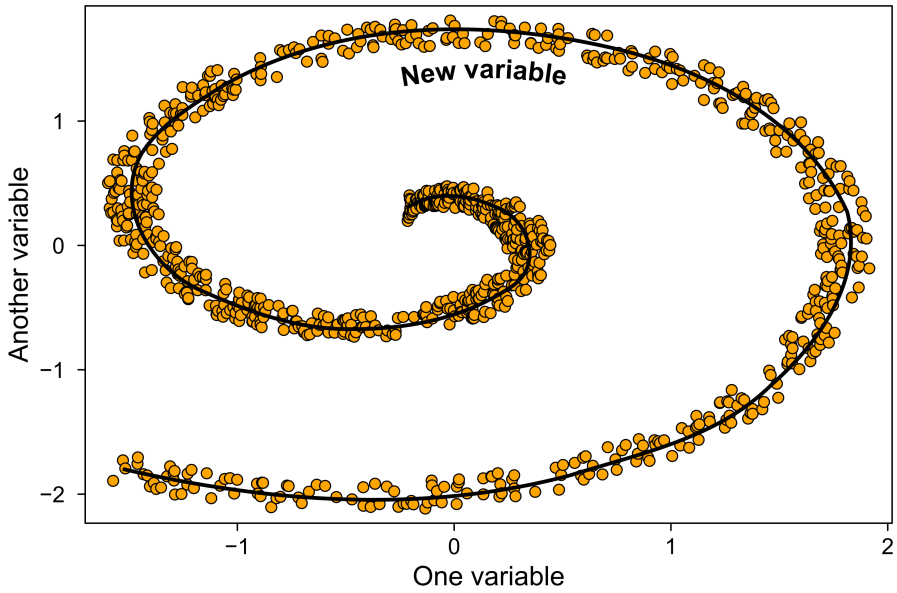

Figure 2.2: Abstract example of a data manifold. We can describe this dataset in terms of one variable and another variable. Since the datapoints cluster around the spiral structure, we can also describe the dataset in terms of a new variable, that follows the main dimension of the data manifold.

.

could say that species A is closer to species B than it is to another species C with which it has 0.2% mismatch. In contrast, if we have two species D,E that have an overlap of 50%, the distance across which we are now measuring is so large that we cannot confidently claim that species D is more similar to species E than it is to a species with which it is 55% similar.

From a mathematical perspective, this occurs because on the local scale the curvature of the data-manifolds is negligible, whereas on the large scale it needs to be taken into account. When comparing gene similarities of two species in a naive manner, we measure the distance between them along the shortest line. However, real-world systems have non-linear characteristics that cannot be captured by a line. Consequently, these long-distance comparisons, relying on the shortest line, lead us away from the manifold on which real genomes exist, thus departing from the biologically plausible space. As a consequence, such long-distance comparisons lack reliability unless deliberate efforts are made to conduct them along the actual data manifold. PCA thus eventually fails twice, because the long-distance comparisons leave the biologically plausible space and because it attempts to fit a flat surface to a structure that is generally curved.

An alternative approach is offered by diffusion maps [48]. This method rejects all long-distance comparisons between data points and instead constructs a network of short

reliable comparisons, where data points are linked to their k most similar neighbors. Establishing this network enables the comparison of distant data points along the network of trusted comparisons. Diffusion maps hence identify data manifolds and embed the data in a new space where long-distance measurements represent dissimilarity in the real-world plausible space.

In the next section, we will introduce the diffusion map method in detail and provide a simple introductory example.

2.3.2 Diffusion map method

The diffusion map method that we introduce here is a variation of the method presented in Coifman et al. [48]. It consists of the following six steps [47].

1. Standardize the data
2. Compute distances between all data points
3. Construct a similarity matrix
4. Threshold the similarity matrix
5. Define a Laplacian matrix
6. Eigendecomposition of the Laplacian provides new variables

Let's revisit the abstract example discussed in the previous section and examine each step of the diffusion map method. Our starting point is the data matrix $\mathbf{A}$ with the dimensions M x N, where M=1,000 is the number of datapoints and N=2 is the number of dimensions, i.e. the coordinates in the dataspace. First, we standardized the data such that each column has a mean of zero and a standard deviation of 1. This ensures that we consider the variables on the same scale. The components of the standardized data matrix $\hat{\mathbf{A}}$ are computed as

$$\hat{A}_{m,n} = \frac{A_{m,n} - \mu_n}{\sigma_n}, \tag{2.28}$$

with

$$\mu_n = \frac{\sum_m A_{m,n}}{M} \tag{2.29}$$

$$\sigma_n = \sqrt{\frac{\sum_m (A_{m,n} - \mu_n)^2}{M}} \tag{2.30}$$

being the mean and standard deviation of the nth column of $\mathbf{A}$.

Following step 2, we compute all Euclidean distances between all data points. Thereby we obtain an M x M distance matrix $\mathbf{D}$, where

$$D_{i,j} = \sqrt{\sum_n (\hat{A}_{i,n} - \hat{A}_{j,n})^2} \tag{2.31}$$

is the Euclidean distance between the datapoints i and j in the data space. This includes also many long-distance comparisons.

Next, we define similarities of two datapoints as the inverse of the Euclidean distance of the respective datapoint pair. As a result, we convert the distance matrix into a similarity matrix $\mathbf{C}$, where

$$C_{i,j} = 1/D_{i,j} \tag{2.32}$$

is the similarity of the datapoint pair i, j. The diagonal elements of the matrix $\mathbf{C}$, comparing each datapoint to itself, are set to zero. The selection of the distance and similarity metric depends on the dataset at hand. Recent papers used for example the re-scaled Spearman correlation coefficient as similarity measure for phytoplankton abundance data [49] and in Chapt. 4 we apply the hamming distance as distance measure to compare bacterial genomes using gene presence-absence data.

As discussed in the previous section, comparisons between distal points are highly unreliable and therefore a source of noise that rather hurts than helps our analysis. To eliminate all long-distance comparisons, we threshold the similarity matrix (step 3), keeping only the top-10 highest similarity entries for each datapoint and setting all other entries in the matrix to zero. Thus, an entry $C_{i,j}$ is kept if it is among the top-10 highest similarity scores for either datapoint i or for datapoint j or both. We thereby retain only the trusted comparisons in the dataset.

If we imagine the datapoints as nodes in a network, we can interpret the similarity matrix as the weight matrix of the network. Thus, we build a network from the comparisons that we trust. Establishing this network enables us to make comparisons between distant points by measuring the distance on the network of allowed comparisons. For example, in Fig. 2.3 we see that before thresholding, point A and C would be closer than A and B, whereas in the thresholded network, we measure the distances along the trusted comparisons, which makes A and B closer than A and C. Previous studies showed that the value of 10 is often a good threshold and the results are robust to the exact choice of the threshold [47, 49]. However, if the network becomes disconnected we need to apply a higher threshold.

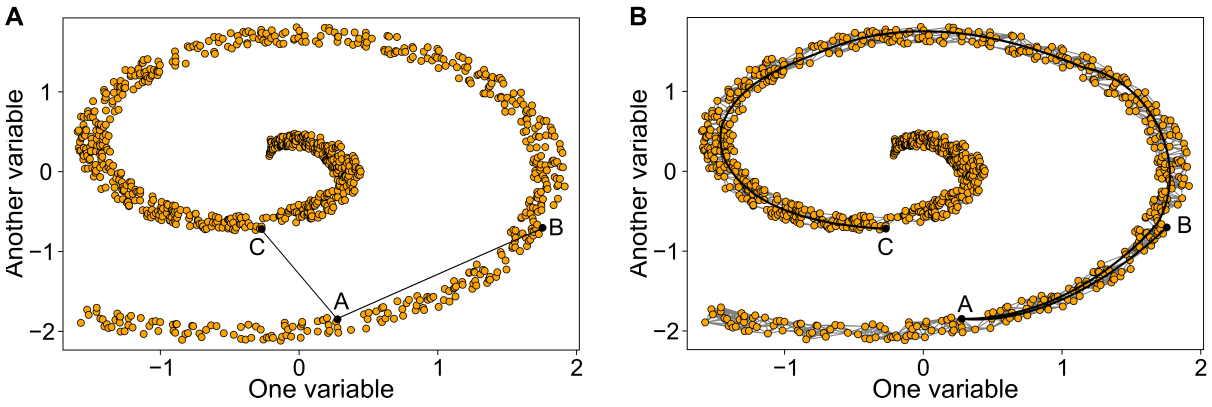

Figure 2.3: Measuring distances in a dataset. Allowing all comparisons in a dataset, we find that point A is closer to C than to B (A). If we trust only the comparisons to the 10-nearest neighbors of each point, we measure the distance along the data manifold. Point A is then closer to B than to C. The trusted connections are indicated in grey (B).

.

We have constructed a network of trusted comparisons and the idea is that we can now quantify the distance between dissimilar data points as the distance on this network. An intuitive way to measure the distances on the network would probably be to consider the shortest path distance. However, due to the dependence of this distance on the presence and absence of single links, the shortest path distance is very susceptible to noise. Diffusion maps use the concept of diffusion distance [48], which robustly quantifies the distance between data points, taking all the possible paths between the nodes in the network into account. The diffusion distances can be computed from the eigenvectors of the Laplacian matrix. Accordingly, from the thresholded similarity matrix we compute the corresponding M x M row-normalized Laplacian matrix (step 5), defined by

$$L_{i,j} = \begin{cases} 1 & \text{for } i = j, \\ -\frac{c_{i,j}}{\sum_n c_{nj}} & \text{otherwise.} \end{cases} . \tag{2.33}$$

In a final step, the eigenvectors and corresponding eigenvalues of this Laplacian matrix are computed. Due to the structure of the Laplacian, we obtain at least one zero eigenvalue. The number of zero eigenvalues indicates the number of components of the network [87]. Hence, if we obtain more than one zero eigenvalue, the network has become disconnected and we have to repeat our analysis applying a higher threshold. The eigenvectors have a dimensionality equal to the number of data points, thus each eigenvector assigns one value to every data point. The eigenvector corresponding to the zero eigenvalue does not contain any information and can be disregarded. The eigenvectors corresponding to the smallest non-zero eigenvalues are the most interesting as they identify the directions of the largest variation, i.e. the main dimensions of the data manifolds. In our example, we see

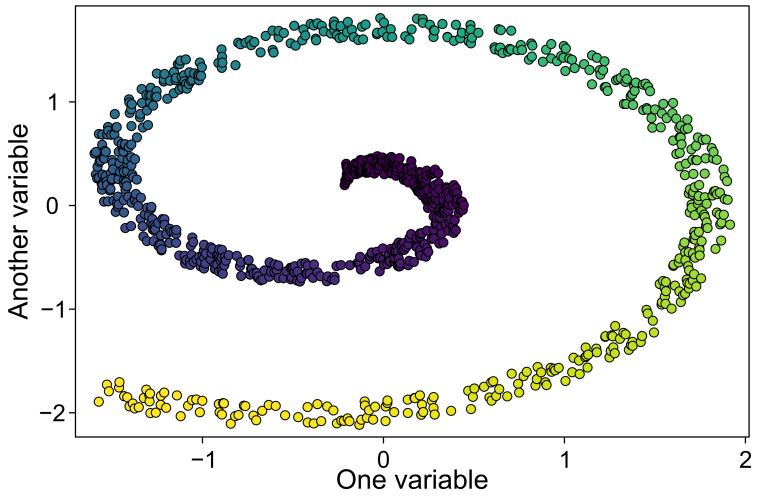

Figure 2.4: Diffusion map detects data manifold. The datapoints are colored according to their values in the first eigenvector of the diffusion map. The first eigenvector follows the main manifold of the data.

that the first eigenvector, i.e. the one corresponding to the smallest non-zero eigenvalue, follows the main dimension of the data manifold (Fig. 2.4). The eigenvectors represent new variables that are nonlinear combinations of the original variables and can be interpreted as coordinates in trait space [47, 48]. Hereby, we have moved from the measured space that is difficult to make sense of, because it is high dimensional and sparse, to a space in which we can confidently compare datapoints by measuring the Euclidean distances in this new trait space [48].

In conclusion the diffusion map is a deterministic and de-facto parameter-free method that represents a powerful tool to reduce the dimensionality of datasets by identifying the important dimensions of the data manifolds, which reflect nonlinear combinations of the measured variables.

2.3.3 Applications in Ecology

Recent papers demonstrate the power of diffusion maps in their application to ecological data [44, 46, 49, 88]. Diffusion mapping metabolic capabilities predicted from bacterial genomes, Fahimipour and Gross [44] identify a multitude of new variables representing interpretable metabolic strategies that span a functional coordinate system, i.e. the metabolic niche space. The mapping of this niche space enables coarse-graining of bacterial communities from different habitats in terms of their metabolic niches that may be filled.

Reiter et al. [88] combined diffusion maps with continuous differential expression analysis

to investigate asynchronous gene expression data during wine fermentation of *Saccharomyces cerevisiae*. They were able to identify site-specific differences and biologically relevant shifts in gene expression related to interactions with the environment.

Gault et al. [46] demonstrates the applicability of diffusion maps to compositional data. The diffusion map results in a dimensionality reduction that enables the calculation of meaningful distances between samples even when they share no species in common.

Applying diffusion maps to monitoring data of phytoplankton communities, i.e. species abundances or biomasses, Ryabov et al. [49] reconstructed the functional trait space of these communities. The reconstruction of the trait space enables them to quantify functional diversity across communities and time. To calculate the functional diversity, first they compute the distances in the reconstructed trait space for all species pairs. The distance between two species i, j is

$$d_{i,j} = \sqrt{\sum_k \left(\frac{v_{k,i} - v_{k,j}}{\lambda_k} \right)}, \tag{2.34}$$

where $v_{k,i}$ is the entry of eigenvector, i.e. trait, k for species i and λ_k is the corresponding eigenvalue [49]. These pairwise functional distances are then used to calculate the functional diversity of each sample as Rao index [89, 90]. The functional diversity (FD) for sample k is computed as

$$FD_k = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{i,j} p_k^{(i)} p_k^{(j)}, \tag{2.35}$$

where $p_k^{(i)} = a_k^{(i)} / \sum_j a_k^{(j)}$ is the relative biomass of species i in sample k [49]. The proposed procedure allows us to use existing monitoring data to measure changes in functional biodiversity.

Together, these applications highlight the diffusion map as a tool that yields new explanatory variables that represent composite functional strategies of the studied organisms. These new variables emerge as nonlinear functions of the input data. The diffusion map is hence able to unravel complex relationships in large datasets.

2.3.4 Visualizing high dimensional data using PHATE

Diffusion mapping high-dimensional datasets usually results in a still relatively high-dimensional representation. For quantitative analysis this is very useful as it enables to accurately quantify functional differences as we saw for example in the study by Ryabov

et al. [49]. However, if we want to gain an intuitive understanding of the structure or shape of our dataset, we need to create a 2-3-dimensional representation. This is the idea of PHATE (Potential of Heat-diffusion for Affinity-based Transition Embedding) [45]: Multi-dimensional scaling is used to find a good 2- or 3-dimensional embedding for the diffusion map results. Moon et al. [45] demonstrate that PHATE generates meaningful representations of high-dimensional data, making it a valuable tool for visualizing datasets with high dimensionality.

2.4 Summary

In this chapter I have reviewed concepts and tools for the analysis of dynamical systems, which I use in the following chapters to make sense of the dynamics of ecological communities. Also, I introduced diffusion maps as a powerful tool for nonlinear dimensionality reduction. This method will be among the key approaches in the following chapters to advance our understanding of complex ecological communities.

Chapter 3

Generalized modeling

Ecological systems are not only composed of many constituents as we have seen in the previous section, the different constituents also interact in diverse and complex ways. This poses a major challenge for understanding the dynamics of these systems and for predicting their future behavior. Often, we face limited data availability and can only approximate the specific functional forms and parameter values used in conventional dynamical models. It was shown that small changes in these functional forms can have large impacts on the qualitative behavior of the model [66, 91]. For example, such changes can lead to increasing supply of nutrients or prey being stabilizing in some foodchain models and destabilizing in other very similar models [66]. Additionally, conventional models frequently lack mathematical tractability and numerical explorations are impaired by both computational and data limitations. Generalized modeling [68, 92] (GM) is an alternative way of modeling that extracts insights using unspecified functions, instead of restricting the processes to specific functional forms. GM thereby offers a highly efficient analysis with insights into the dynamics and bifurcations of uncertain systems.

To make this method more accessible for a broad audience of researchers with different levels of experience in modeling, we summarized the state of the art of GM and provided a hands-on guide on how to use GM in [93]. The following sections on GM follow the lines of this paper. I begin in Sec. 3.1 with the idea of GM, followed by an introductory example of a generalized model in Sec. 3.2. Thereafter, I give a detailed description of the GM procedure (Sec. 3.3) and show different ways of analyzing a generalized model (Sec. 3.4). In Sec. 3.5 applications of GM in ecology are reviewed. Finally, I explain the master stability approach (Sec. 3.6), which can be employed to extend the generalized model analysis to spatial systems.

3.1 Idea of generalized modeling

Mathematics is a powerful tool especially because it can work with unknown objects. Consider the number π, no one knows its exact value, however this does not impact our

ability to use π in calculations or investigate its properties. Mathematics possesses the capability to operate with unknown entities beyond just numbers. In modeling, we can leverage this capacity by formulating models using unspecified functions. Working with unknown functions is commonplace in mathematics. When writing the definition of a derivative we use for instance the unknown function $x(t)$

$$\dot{x} = \frac{\mathrm{d}}{\mathrm{d}t}x(t) = \lim_{\delta \to 0} \frac{x(t+\delta) - x(t)}{\delta}. \tag{3.1}$$

Here, x represents an arbitrary variable and t is time.

If we leave variables or functions unspecified we are actually performing massively parallel computations, because we consider all possible values or forms these variables or functions could take. This is the idea of generalized modeling (GM): Capturing the structure of a system without restricting it to specific functional forms enables us to analyze a whole class of systems in parallel.

As a result of working with unspecified functional forms, we gain insights into how dynamical properties link to these unknown functions. To make sense of these results, it is therefore important that these properties are interpretable in the specific context. GM captures these properties in a set of parameters that have a clear and intuitive interpretation in the context of the model.

To see how this is achieved, let's compare GM to conventional modeling: The conventional modeling approach can be described as a 3-step process:

1. Parameterization: Restrict the model to equations that are specified except for a number of unknown parameters.

2. Steady states: Find steady states of the ordinary differential equations (ODEs).

3. Linearization: Compute the Jacobian matrix, which provides a linearization of the dynamics around the steady state.

Once we obtain the Jacobian, we can analyze the stability of steady states, find their bifurcations etc. In this conventional approach, we deal with different difficulties at each step. The first step involves the difficulty of finding the right model for a given phenomenon or process, which often requires experience. In the second step, we need to find the roots of an equation system, this is very difficult for all but the simplest systems. We often have to turn to numerics, but no algorithms with guaranteed convergence are known. In the last step we need to differentiate functions, which is generally easy.

The computation of steady states, step 2, clearly stands out in terms of technical difficulty. In many systems, we know the steady states already, hence the reward for dealing with these difficulties is often small. We therefore consider how to circumvent this step in the analysis. Random matrix models offer an alternative, they directly formulate a model for the Jacobian matrix, instead of deriving this matrix from ODEs [9, 70, 94]. While these models have provided powerful insights, e.g. that large random food webs are unlikely to be stable [9], they are relatively abstract and hence suffer from low interpretability, e.g. from the abstract model researchers gained little intuition what the features could be that stabilize large food webs.

GM combines the advantages of both approaches, being almost as interpretable as conventional models, while offering almost the efficiency of random matrix models. This is achieved by re-ordering and slightly modifying the three steps of the modeling procedure to:

1. Steady states: Consider a class of models that is general enough that steady states must exist in this class. Define symbols to denote the variables in these unknown steady states.

2. Linearization: Formally compute the derivative of the processes with respect to variables to compute the Jacobian.

3. Parameterization: Identify the quantities that appear in the Jacobian as parameters in the model.

In GM we do not restrict the processes in the model to specific functional forms. Therefore, we cannot meaningfully compute the steady states of the model. Hence, in the generalized model the steady states are unknown quantities. We can still formally linearize the dynamics around the steady states, which yields the Jacobian matrix. In the parameterization step we use a specific mathematical identify to give meaning to the quantities in the Jacobian matrix (explained in more detail in the following section). This specific way of parameterization sets generalized models apart from other models containing unknown functions. As a result of the GM procedure, we get the Jacobian matrix in a steady state as a function of a set of possibly unknown but interpretable parameters. The underlying ODEs provide guidance in the interpretation of this matrix.

3.2 An introductory example

Let us start with a simple example to illustrate how GM works. Consider a system where a variable, X, undergoes dynamic changes in response to gains and losses (Fig. 3.1). The variable X could for example represent the biomass or number of individuals of a species in an ecosystem. We can write the differential equation

$$\dot{X} = G(X) - L(X), \tag{3.2}$$

where the dot denotes a time derivative, G represents an unknown function of the gain terms and L represents an unknown function of the loss terms. The only assumption that we have made so far is that gain and loss are describable by mathematical functions.

Figure 3.1: Illustration of a system where a variable, X, undergoes dynamic changes in response to gains, $G(X)$, and losses, $L(X)$.

If we were following the conventional approach, we would now restrict the gain and loss processes to specific functional forms, compute steady states and then delve into deeper analysis, e.g. compute stability and bifurcations. GM draws on the insight that we do not need to restrict the system to specific functional forms to perform this deeper analysis of the system. Take for example the Jacobian matrix **J** that captures the stability of steady states. Its elements are defined as

$$J_{i,j} = \frac{\partial}{\partial X_j}\dot{X}_i\Big|_*, \tag{3.3}$$

where $|_*$ denotes that the expression is evaluated at the steady state under consideration. The Jacobian of the simple example system from above is

$$J_{11} = G'(X^*) - L'(X^*), \tag{3.4}$$

where X^* is the steady state under consideration and the dash denotes a partial derivative. We can see that we can derive the Jacobian of this system, however in this form it is not very informative. $G'(X^*)$ is the derivative of an unknown function at an unknown point. As we know that $G'(X^*)$ is a number, we can think of it as a parameter of the system. However, we do not have an intuitive interpretation of this parameter in the context of the application.

As mentioned in the previous section, in GM we parameterize the model in a way that provides an interpretable set of unknown parameters that captures the uncertainty about the system without restricting the processes in the model to a specific functional form. To achieve this, we need to make one more assumption: All variables and process rates have positive values. In many cases, this is very intuitive since for example the number of species does not take negative values. Process rates are by design non-negative, if we have a process that can run in two directions, we can define two processes that run antagonistically, which often leads to a better and more interpretable model.

We can however imagine that variables or processes become zero, for example a species could go extinct. In this case, we can make a model where the species is absent. In the scenario that we are interested in the transition where the extinction occurs, we can consider the model in which the species is present since it remains valid as we approach the point of extinction. Validity in this limit is sufficient to identify the transition in which the extinction of the species occurs.

Let's return to our example model from above (Eq. 3.2). Since this equation describes a whole class of models, positive steady states must exist. We use X^* as a placeholder for every positive steady state in the system and denote the rates of processes in the steady state as $L^* = L(X^*)$ and $G^* = G(X^*)$. We can formally normalize the equation with respect to X^*,

$$x = \frac{X}{X^*}, \tag{3.5}$$

hence $X = xX^*$. Also, we define normalized

$$g(x) = \frac{G(xX^*)}{G^*}, \tag{3.6}$$

$$l(x) = \frac{L(xX^*)}{L^*}. \tag{3.7}$$

A convention in GM is to use upper-case variables to define unnormalized quantities and lower-case variables to define normalized quantities. Writing the differential for the

normalized variable, we obtain

$$\begin{aligned} \dot{x} &= \frac{\mathrm{d}}{\mathrm{dt}}\frac{X}{X^*} && (3.8)\\ &= \frac{\dot{X}}{X^*} && (3.9)\\ &= \frac{G(X) - L(X)}{X^*} && (3.10)\\ &= \frac{G^*}{X^*}g(x) - \frac{L^*}{X^*}l(x). && (3.11) \end{aligned}$$

Note that for this normalized system, we defined the normalized quantities such that

$$\begin{aligned} x^* &= 1 && (3.12)\\ g^* &= 1 && (3.13)\\ l^* &= 1, && (3.14) \end{aligned}$$

i.e. in the normalized system the steady state is $x^* = 1$ and all the processes run at rate 1. Hence, by this normalization procedure, we have moved from a system in which we do not know the steady state to a system where we know it. In addition, the normalization procedure results in the appearance of the factors G^*/X^* and L^*/X^*. Since these factors are scalars we can interpret them as unknown parameters of the system. Note that these unknown parameters have an intuitive interpretation: They are per-unit turnover rates, G^*/X^* is the per-capita gain per X in the steady state and L^*/X^* is the per-capita loss per X in the steady state, respectively. If we take the example from above that X is the number of individuals of a species, G^*/X^* would represent the birth rate and L^*/X^* would be the per-capita death rate.

We can also see that that the per-capita gain must equal the per-capita loss, i.e. $G^* = L^*$ must hold in all steady states. It is important to incorporate this condition of identity into the model, since otherwise we might end up investigating steady states that cannot exist in the real world (i.e. where $G^* \neq L^*$). This allows us to define a scale parameter α such that

$$\alpha = \frac{G^*}{X^*} = \frac{L^*}{X^*}. \tag{3.15}$$

Rewriting the system yields

$$\dot{x} = \alpha(g(x) - l(x)). \tag{3.16}$$

Now let's return to stability analysis, the Jacobian of the normalized system with the steady state $x^* = 1$ is

$$\mathbf{J} = [\alpha(g'(1) - l'(1))]. \tag{3.17}$$

Since we did not constrain the functions g and l to specific functional forms, we cannot compute their derivatives, thus they represent two additional unknown parameters. We define

$$g_{\mathrm{x}} = g'(1) \tag{3.18}$$
$$l_{\mathrm{x}} = l'(1), \tag{3.19}$$

which are called *elasticities* or *exponent parameters* in the context of GM. Note that these parameters are the logarithmic derivatives of the original functions, e.g.

$$g_{\mathrm{x}} = \left.\frac{\mathrm{d}\ln G}{\mathrm{d}\ln X}\right|_*. \tag{3.20}$$

The exponent parameters measure the nonlinearity of the sensitivity of the functions to variations in the argument [68, 95]. Consider for example that the gain is a linear function,

$$G(X) = aX, \tag{3.21}$$

with $a > 0$. Normalization yields

$$g(x) = \frac{G(X)}{G^*} = \frac{aX}{aX^*} = \frac{X}{X^*}, \tag{3.22}$$

hence $g_{\mathrm{x}} = 1$ regardless of a. Consequently, every linear function results in a parameter value of 1. Let's see what happens in case of a quadratic relationship, e.g.

$$G(X) = aX^2. \tag{3.23}$$

Normalization in this case yields

$$g(x) = \frac{G(X)}{G^*} = \frac{aX^2}{a(X^*)^2} = \frac{X^2}{(X^*)^2} = x^2, \tag{3.24}$$

hence $g_{\mathrm{x}} = 2$.

In general, for any power law aX^p the corresponding exponent parameter is p. This also holds for decreasing functions, e.g. a/X^p for which the corresponding exponent parameter

is $-p$. In the case of more complex relationships, the exponent parameter can depend on the location of the steady state. Take for example the Holling type-II functional response [60], which is linear for low prey density and saturates for high prey density. The corresponding exponent parameters is approximately 1 where the functional response is linear and decreases to 0 as the functional response approaches the saturated regime [95]. Elasticities are used in a number of scientific disciplines [96, 97], because they offer an intuitive way to describe non-linearity and can be estimated conveniently from limited and noisy data.

Returning to our example system, we can now write the Jacobian that captures the dynamics around the steady states in all models of the form Eq. 3.2 as a function of the three parameters that we can interpret,

$$\mathbf{J} = [\alpha(g_{\mathrm{x}} - l_{\mathrm{x}})]. \tag{3.25}$$

In this case the Jacobian is a 1x1 matrix, so it has only one eigenvalue, that is

$$\lambda = \alpha(g_{\mathrm{x}} - l_{\mathrm{x}}). \tag{3.26}$$

Since a steady state is stable if all eigenvalues of the Jacobian have negative real parts, we can derive for our example system that a steady state under consideration is stable if

$$g_{\mathrm{x}} < l_{\mathrm{x}}. \tag{3.27}$$

Hence, in every system of the form of Eq. 3.2, every positive steady state is stable if the elasticity of the loss is greater than the elasticity of the gain. The turnover rate does not impact the stability directly. It may have an indirect effect, for example if the nonlinearity of the gains increases under high turnover rates. Overall, GM is able to extract concrete results even with the very little structural information we have provided in this simple example.

3.3 Generalized modeling procedure

We now discuss the procedure of formulating generalized models for more complex systems, taking the example of the predator-prey system from Yeakel et al. [95]. In this example system the growth of a predator population depends entirely on the prey. This example includes intra- as well as interspecific interactions.

(1) Identification of State Variables

In a first step, we need to identify the state variables that we want to describe. In our predator-prey example, we are for instance interested in the populations of the predator and prey species. Therefore, we define two state variables, X and Y that represent the population sizes of prey and predator, respectively. Often, it is easy to identify the state variables but it can get more complicated if the system involves more abstract parts playing a role such as human behavior. In general, it is a good idea to include additional state variables rather than leaving them out, since their cost is low and additional variables can increase the interpretability of the model.

(2) Identification of Processes

In the next step we have to identify the processes that drive the changes in our state variables. For each state variable there must be at least one gain and one loss process. Instead of formulating specific functional forms for all the interactions involved in the considered phenomena, we try to describe the basic structure of the system. E.g. for predator-prey interactions it is easy to say that predation depends on the number of predators and the number of prey. It is much harder to derive the exact functional form that quantifies this interaction [68].

Focusing on the structural information of our predator-prey system, we can for example write

$$\dot{X} = S(X) - F(X,Y) - L(X) \tag{3.28}$$

$$\dot{Y} = G(X,Y) - M(Y), \tag{3.29}$$

where $S(X)$ describes the reproduction of the prey, $F(X,Y)$ models the loss of prey due to predation, $L(X)$ is the loss of prey due to other causes, $G(X,Y)$ is the gain of predators through predation and $M(Y)$ describes the loss of predators.

One could argue that we could summarize the loss terms of X in one term, however separating each of the loss processes here makes the model more interpretable and the individual processes become tangible. Particularly, in the analysis of this predator-prey model we are probably interested in considering the loss terms by predation and by other causes separately. Also, in GM we generate insights by structural information on the system. Consequently, the more detailed the structural information we feed into the model, the more insights we can gain.

(3) Normalization

Once we have formulated our model, we proceed with the normalization procedure as

shown in the introductory example (Sec. 3.2). Accordingly, we define the normalized state variables

$$x = \frac{X}{X^*} \tag{3.30}$$

$$y = \frac{Y}{Y^*}, \tag{3.31}$$

where X^* and Y^* are placeholders for every positive steady state in the system. And we define normalized functions

$$s(x) := \frac{S(X^*x)}{S^*} \tag{3.32}$$

$$f(x,y) := \frac{F(X^*x, Y^*y)}{F^*} \tag{3.33}$$

$$l(x) := \frac{L(X^*x)}{L^*} \tag{3.34}$$

$$g(x,y) := \frac{G(X^*x, Y^*y)}{G^*} \tag{3.35}$$

$$m(y) := \frac{M(Y^*y)}{M^*}, \tag{3.36}$$

where

$$S^* := S(X^*) \tag{3.37}$$

$$F^* := F(X^*) \tag{3.38}$$

$$L^* := L(X^*) \tag{3.39}$$

$$G^* := G(X^*) \tag{3.40}$$

$$M^* := M(X^*) \tag{3.41}$$

are used as abbreviated notations.

Normalizing our model in this way, we have moved to a system in which we know the steady state:

$$x^* = y^* = s^* = f^* = l^* = g^* = m^* = 1. \tag{3.42}$$

We can write our normalized model as

$$\dot{x} = \frac{S^*}{X^*}s(x) - \frac{F^*}{X^*}f(x,y) - \frac{L^*}{X^*}l(x) \tag{3.43}$$

$$\dot{y} = \frac{G^*}{Y^*}g(x,y) - \frac{M^*}{Y^*}m(y). \tag{3.44}$$

Our goal is now to simplify the model through the introduction of new easily interpretable parameters, while considering the stationarity condition of the steady state. To start, we consider the system in the steady state, which yields

$$0 = \frac{S^*}{X^*} - \frac{F^*}{X^*} - \frac{L^*}{X^*} \tag{3.45}$$

$$0 = \frac{G^*}{Y^*} - \frac{M^*}{Y^*}, \tag{3.46}$$

so in the steady state the sum of the gain terms has to be equal to the sum of the loss terms for each variable, i.e.

$$\frac{S^*}{X^*} = \frac{F^*}{X^*} + \frac{L^*}{X^*} \tag{3.47}$$

$$\frac{G^*}{Y^*} = \frac{M^*}{Y^*}. \tag{3.48}$$

We now want to define scale parameters, which have been already introduced in Sec. 3.2. A convenient procedure is often to use one parameter for each variable to describe the total turnover and then to define additional parameters that capture the contributions of the individual loss and gain terms to the total turnover. Ergo, we define the scale parameters

$$\alpha_{\mathrm{x}} = \frac{S^*}{X^*} = \frac{F^*}{X^*} + \frac{L^*}{X^*} \tag{3.49}$$

$$\alpha_{\mathrm{y}} = \frac{G^*}{Y^*} = \frac{M^*}{Y^*}, \tag{3.50}$$

which describe the turnover of species X and Y, respectively. Re-writing the system using these scale parameters yields

$$\dot{x} = \alpha_{\mathrm{x}} \left(s(x) - \frac{1}{\alpha_{\mathrm{x}}} \frac{F^*}{X^*} f(x,y) - \frac{1}{\alpha_{\mathrm{x}}} \frac{L^*}{X^*} l(x) \right) \tag{3.51}$$

$$\dot{y} = \alpha_{\mathrm{y}} (g(x,y) - m(y)). \tag{3.52}$$

Due to the two loss terms of the prey population, i.e. loss through predation and loss through other causes, we are still left with prefactors in front of the f and l terms. However, these can be captured in additional parameters that have a convenient interpretation,

e.g. we use β to denote the prefactor of the f term and write

$$\begin{aligned}
\beta &= \frac{1}{\alpha_{\mathrm{x}}}\frac{F^*}{X^*} && (3.53)\\
&= \frac{1}{\left(\frac{F^*}{X^*}+\frac{L^*}{X^*}\right)}\frac{F^*}{X^*} && (3.54)\\
&= \frac{X^*}{(F^*+L^*)}\frac{F^*}{X^*} && (3.55)\\
&= \frac{F^*}{F^*+L^*}. && (3.56)
\end{aligned}$$

Thus, β quantifies the relative contribution of predation to prey's loss in the steady state. Correspondingly, we define

$$\begin{aligned}
\tilde{\beta} &= \frac{1}{\alpha_{\mathrm{x}}}\frac{L^*}{X^*} && (3.57)\\
&= \frac{L^*}{F^*+L^*}, && (3.58)
\end{aligned}$$

so $\tilde{\beta}$ describes the relative contribution of other causes to prey's loss in the steady state. These parameters that capture the relative contribution of the individual terms within the loss (or gain) terms are called *branching parameters* as they describe the branching or merging of flows in the system. The two branching parameters β and $\tilde{\beta}$ are not independent of each other as the losses of a particular variable have to add up to 1. We can show that

$$\beta+\tilde{\beta} = \frac{F^*}{F^*+L^*}+\frac{L^*}{F^*+L^*}=1. \quad (3.59)$$

We can now write our model as

$$\begin{aligned}
\dot{x} &= \alpha_{\mathrm{x}}(s(x)-\beta f(x,y)-\tilde{\beta}l(x)) && (3.60)\\
\dot{y} &= \alpha_{\mathrm{y}}(g(x,y)-m(y)) && (3.61)\\
\tilde{\beta} &= 1-\beta. && (3.62)
\end{aligned}$$

Once, we have familiarized ourselves with the normalization procedure, we can see that the resulting equations always follow the same pattern, for example the equation

$$\dot{Z} = A(Y)+B(Z)+C(Y,Z)-D(Y)-E(Z)-F(Y,Z) \quad (3.63)$$

normalizes to

$$\begin{aligned} \dot{z} &= \alpha(\gamma_{\mathrm{a}} a(y) + \gamma_{\mathrm{b}} b(z) + \gamma_{\mathrm{c}} c(y,z) - \rho_{\mathrm{d}} d(y) - \rho_{\mathrm{e}} e(z) - \rho_{\mathrm{f}} f(y,z)) && (3.64)\\ 1 &= \gamma_{\mathrm{a}} + \gamma_{\mathrm{b}} + \gamma_{\mathrm{c}} && (3.65)\\ 1 &= \rho_{\mathrm{d}} + \rho_{\mathrm{e}} + \rho_{\mathrm{f}}. && (3.66) \end{aligned}$$

(4) Timescale Normalization and Jacobian

One of the scale parameters can always be set to 1 by timescale normalization. For example, if we measure in terms of multiples of the turnover time of the prey $1/\alpha_{\mathrm{x}}$, rescaling both equations by this factor yields

$$\begin{aligned} \dot{x} &= s(x) - \beta f(x,y) - \bar{\beta} l(x) && (3.67)\\ \dot{y} &= \alpha(g(x,y) - m(y)) && (3.68)\\ \bar{\beta} &= 1 - \beta, && (3.69) \end{aligned}$$

where

$$\alpha = \frac{\alpha_{\mathrm{x}}}{\alpha_{\mathrm{y}}} \tag{3.70}$$

describes the relative rate of predator turnover to prey turnover. If predator and prey population are measured in terms of biomass this turnover rate is the metabolic rate of the prey divided by the metabolic rate of the predator, hence the ratio of metabolic rates. If we measure in terms of abundances the turnover rate is the prey life expectancy divided by the predator life expectancy.

In the next step, we calculate the Jacobian matrix, which elements are defined as

$$J_{i,j} = \left. \frac{\partial}{\partial X_j} \dot{X}_i \right|_*, \tag{3.71}$$

where $|_*$ denotes that the expression is evaluated at the steady state under consideration.

Defining the exponent parameters

$$
\begin{aligned}
s_{\mathrm{x}} &= \left.\frac{\partial}{\partial x} s(x)\right|_{*} && (3.72)\\
f_{\mathrm{x}} &= \left.\frac{\partial}{\partial x} f(x,y)\right|_{*} && (3.73)\\
f_{\mathrm{y}} &= \left.\frac{\partial}{\partial y} f(x,y)\right|_{*} && (3.74)\\
l_{\mathrm{x}} &= \left.\frac{\partial}{\partial x} l(x)\right|_{*} && (3.75)\\
g_{\mathrm{x}} &= \left.\frac{\partial}{\partial x} g(x,y)\right|_{*} && (3.76)\\
g_{\mathrm{y}} &= \left.\frac{\partial}{\partial y} g(x,y)\right|_{*} && (3.77)\\
m_{\mathrm{y}} &= \left.\frac{\partial}{\partial y} m(y)\right|_{*}, && (3.78)
\end{aligned}
$$

we obtain the derivatives

$$
\begin{aligned}
\left.\frac{\partial \dot{x}}{\partial x}\right|_{1} &= s_{\mathrm{x}} - \beta f_{\mathrm{x}} - (1-\beta) l_{\mathrm{x}} && (3.79)\\
\left.\frac{\partial \dot{x}}{\partial y}\right|_{1} &= -\beta f_{\mathrm{y}} && (3.80)\\
\left.\frac{\partial \dot{y}}{\partial x}\right|_{1} &= \alpha g_{\mathrm{x}} && (3.81)\\
\left.\frac{\partial \dot{y}}{\partial y}\right|_{1} &= \alpha (g_{\mathrm{y}} - m_{\mathrm{y}}) && (3.82)\\
& && (3.83)
\end{aligned}
$$

yielding the Jacobian matrix

$$
\mathbf{J} = \begin{pmatrix} s_{\mathrm{x}} - \beta f_{\mathrm{x}} - (1-\beta) l_{\mathrm{x}} & -\beta f_{\mathrm{y}} \\ \alpha g_{\mathrm{x}} & \alpha (g_{\mathrm{y}} - m_{\mathrm{y}}) \end{pmatrix}. \qquad (3.84)
$$

We arrived at a relatively simple matrix that captures the structural knowledge that we have about the system. For example, net growth is the difference between gains and losses, independent processes add up and the dependence of predator-prey interactions on the number of predators and the number of prey. We did not have to make any specifications that we are not certain about like the exact form of the predator-prey interaction.

(5) Additional Constraints and Auxiliary Variables Once we have a general model

like the predator-prey example model, we can always refine the model and add additional insights that we gained about the system in consideration. For example, in our predator-prey system, we know that the gain of the predator should be related to the loss of the prey, therefore the functions F and G are not independent. A simple way to incorporate this dependence in the model is to assume that the predator gain is a function of the prey's loss, for example,

$$G(X,Y) = H(F(X,Y)). \tag{3.85}$$

Incorporating this insight into the model even in this unspecified way is useful since all of the terms have an intuitive interpretation: G is the predation gain, F is the predation loss and H is the conversion efficiency. And including these elements in the equations makes them tangible, we can investigate their impact in the system.

Whenever we want to include such additional insights in our model, we need to check if this imposes additional constraints on the scale and exponent parameters. In this example, in the steady state we obtain $G^* = H^*$, thus this does not impose any additional constraints on the scale parameters. Next, we consider the constraints on the exponent parameters. We start by using the GM normalization procedure to normalize the new condition.

Defining

$$h(f) = \frac{H(F^*f, Y^*y)}{H^*}, \tag{3.86}$$

this yields

$$g(x,y) = \frac{G(X^*x, Y^*y)}{G^*} = \frac{H(F^*f, Y^*y)}{H^*} = h(f(x,y)). \tag{3.87}$$

Next, we compute the corresponding exponent parameters

$$g_{\mathrm{x}} = \left.\frac{\partial}{\partial x} h(f(x,y))\right|_1 = h_{\mathrm{f}} f_{\mathrm{x}} \tag{3.88}$$

$$g_{\mathrm{y}} = \left.\frac{\partial}{\partial y} h(f(x,y))\right|_1 = h_{\mathrm{f}} f_{\mathrm{y}}, \tag{3.89}$$

where h_{f} is a new parameter that captures the elasticity of the predator's gain with respect to prey's loss. Assuming a constant conversion efficiency like most models do, $h_{\mathrm{f}} = 1$ as H is a linear function. This is a very basic example of including additional constraints in our model. Gross and Feudel [68] for example show how to build more complex relationships into the model that result in realistic prey-switching behavior.

For illustration, we can build a slightly more complicated case by considering that the conversion efficiency should depend on the per-capita consumption of prey by predators. We could write this condition as

$$G(X, Y) = H(F, C), \tag{3.90}$$

where C is the per-capita consumption

$$C(F, Y) = \frac{F(X, Y)}{Y}. \tag{3.91}$$

In this example, in the steady state we obtain $G^* = H^*$, thus also this new form does not impose any additional constraints on the scale parameters. To find the implications of the condition for the exponent parameters we define

$$h(f, c) = \frac{H(F^* f, C^* c)}{H^*} \qquad c(f, y) = \frac{C(F^* f, Y^* y)}{C^*} \tag{3.92}$$

and then we can show that

$$g(x, y) = \frac{G(X^* x, Y^* y)}{G^*} = \frac{H(F^* f, C^* c)}{H^*} = h(f, c). \tag{3.93}$$

Computing the corresponding exponent parameters yields

$$g_{\mathrm{x}} = h_{\mathrm{f}} f_{\mathrm{x}} + h_{\mathrm{c}} c_{\mathrm{f}} f_{\mathrm{x}} \tag{3.94}$$

$$g_{\mathrm{y}} = h_{\mathrm{f}} f_{\mathrm{y}} + h_{\mathrm{c}} (c_{\mathrm{f}} f_{\mathrm{y}} + c_{\mathrm{y}}). \tag{3.95}$$

To gain an understanding of the new parameters, we need to think about their ecological interpretation. The parameter h_{f} is the partial derivative of h with respect to x at constant c, since c now appears as an explicit argument of h. Hence h_{f} denotes how the growth of the predator population is changing if more predators are feeding but the per-capita amount stays constant. We can assume a linear relationship with better confidence than in the previous example, so $h_{\mathrm{f}} = 1$.

The parameter h_{c} describes how the growth of the predator population changes with varying per-capita consumption. The conversion efficiency could increase or decrease in complex nonlinear ways depending on the ecological circumstances, thus we leave this parameter tunable, enabling us to investigate its behavior in the analysis of the generalized model.

The parameters c_{f} and c_y denote the elasticities of the per-capita consumption with respect to the prey's loss and with respect to the predator population, respectively. As we have defined C explicitly in Eq. 3.91, we can compute

$$c_{\mathrm{f}} = \frac{\partial}{\partial f} c(f,y)\Big|_1 = \frac{\partial}{\partial f} \frac{C(F^* f, Y^* y)}{C(F^*, Y^*)}\Big|_1 = \frac{\partial}{\partial f} \frac{F^* f}{Y^* y} \frac{Y^*}{F^*}\Big|_1 \tag{3.96}$$

$$= \frac{\partial}{\partial f} \frac{f}{y}\Big|_1 = 1 \tag{3.97}$$

$$c_{\mathrm{y}} = \frac{\partial}{\partial y} \frac{f}{y}\Big|_1 = -1. \tag{3.98}$$

We could have also guessed this relationship directly from the definition of C since we defined C to be linear in F and to be inversely related to Y.

This slightly more complex example illustrates the power of GM in enabling us to include structural information that we can confidently assume, e.g. $h_{\mathrm{f}} = 1$, and leaving the relationships, where we are not confident about the explicit structure, e.g. h_{c}, unspecified for further investigation. In the next step, we can write the Jacobian replacing g_{x} and g_{y} by the new parameters and then we can dive into the analysis of the generalized model.

(6) Conservation Laws, Derivative Conditions and Optimality

Additional constraints can also come in the form of conservation laws or derivative conditions. These are imposed on the generalized model in a similar way as described above. In the case of conservation laws, there are two possible procedures to impose them on our model. We could use the conservation law to reduce the number of variables and then proceed with the normalization or we could normalize the model and the conservation law and constrain the parameters with the conditions given by the conservation law. In most cases, the second option is advisable, since in this case we also allow for the investigation of perturbations that violate the conservation law. In many real-world examples this leads to a better representation of the system since perturbations from the outside are a realistic phenomenon. If we have to deal with a lot of constraints imposed by conservation laws, for example in metabolic models where the number of atoms is conserved throughout the metabolic reactions, we can manage the scale parameter for example by representing them as a linear combination of a set of fundamental flux modes that satisfy all constraints [98].

It is also possible to impose additional constraints in the form of derivative conditions. We can for example demand that the partial derivative of a process P with respect to

some state variable X is zero in the steady state, i.e.

$$\left.\frac{\partial P(X)}{\partial X}\right|_* = 0. \tag{3.99}$$

Normalization yields

$$0 = \left.\frac{\partial P(X)}{\partial X}\right|_* = \left.\frac{\partial P^* p(x)}{\partial x}\frac{\partial x}{\partial X}\right|_* = \frac{P^*}{X^*} p_{\mathrm{x}}, \tag{3.100}$$

so in this case the exponent parameter $p_x = 0$. The ability to impose constraints in the form of derivative conditions enables us to investigate phenomena, where a variable is in a local minimum or maximum of some function. We can for example study cooperation of governance models where each agent allocates their resources optimally. Or we can use an adaptive dynamics model to study biological evolution [99] where we force the species to stay in locally evolutionary stable states.

3.4 Analyzing generalized models

The GM procedure arrives at a Jacobian matrix, containing easily interpretable parameters in the context of the system. The different values the parameters in the Jacobian can take represent ensembles of possible realities that are consistent with the available structural knowledge. Analysis of the whole ensemble of possible worlds captured by the generalized model is more limited in comparison to the analysis of conventional models, since we cannot compute the steady state and we cannot simulate a generalized model.

However, we can provide some insights on the whole ensemble of possible worlds trough very efficient analyses of the generalized model. We can for example generate insights into the dynamics of the system and the stability of steady states, we can explore the response to different types of perturbations and we can identify important parameters and parameter regions.

Often a good step to start the analysis of a generalized model is the numerical stability analysis. We restrict the parameters to plausible ranges in the context of the model, draw random ensembles of parameter values and calculate the leading eigenvalue, i.e. the stability for each of the parameter ensembles. As this approach is computationally cheap, we can generate a high number of random parameter ensembles and calculate their stability efficiently. To get a first impression of the behavior of the system, we can then investigate the correlation of individual parameters with stability.

i.e. suppose we consider a generalized model with P different generalized parameters and

we draw M different sets of parameter values. We denote the m'th realization of parameter i as p_i^m where $i \in [1, P]$ and $m \in [1, M]$. And we denote the stability of the steady state by the parameter set m as s_m, where we define

$$s_m = \begin{cases} 1 & \mathrm{Re}(\lambda_0) < 0 \\ 0 & \text{otherwise} \end{cases}, \tag{3.101}$$

thus s_m is 1 if the parameter set m is stable and zero otherwise. After calculating the stability for all M parameter sets, we can compute

$$c_i = \mathrm{Cov}_m(p_i^m, s_m), \tag{3.102}$$

i.e. we estimate the impact of each individual parameter on stability. If a parameter is positively correlated to stability, that is high values of the parameter correspond to stability, c_i takes positive values up to the value of 1, if stability is completely explained by this particular parameter. If a parameter is negatively correlated to stability, that is high values of the parameter correspond to instability, c_i takes negative values up to the value of -1, if stability is completely explained by this particular parameter.

For the random sampling of parameter values it is important to restrict the range to reasonable values in the context of the model since the correlations are not independent of the sampling. Parameters from a wider range will for example show a higher correlation.

We can also explore the impact of press perturbations, i.e. permanent change of parameters, on the system. A sufficiently small perturbation leads to a shift in the steady states, that can be described by

$$\boldsymbol{\delta} = -\mathbf{J}^{-1}\boldsymbol{p}, \tag{3.103}$$

where $\boldsymbol{\delta}$ is the shift in the steady state, $\mathbf{J}^{-1}$ is the inverse of the Jacobian matrix, and $\boldsymbol{p}$ is a vector describing the direct impacts of the perturbations on the individual equations [69]. In the context of a generalized model, the vector $\boldsymbol{\delta}$ describes the impact on the steady state in the normalized variables and the vector $\boldsymbol{p}$ contains the direct impact on the individual equations in units of normalized turnover.

Take again the simple predator-prey example from Sec. 3.3 with the Jacobian matrix

$$\mathbf{J} = \begin{pmatrix} s_\mathrm{x} - \beta f_\mathrm{x} - (1-\beta) l_\mathrm{x} & -\beta f_\mathrm{y} \\ \alpha g_\mathrm{x} & \alpha(g_\mathrm{y} - m_\mathrm{y}) \end{pmatrix}. \tag{3.104}$$

We could for example ask how a small fraction of additional loss for the prey, e.g. by harvesting, would impact the steady state of the system. The direct impact on the

individual equations could then be described by

$$\boldsymbol{p} = \begin{pmatrix} -\rho \\ 0 \end{pmatrix}, \tag{3.105}$$

i.e. a small fraction ρ of the total turnover of the prey is harvested. Substituting into Eq. 3.103 yields

$$\boldsymbol{\delta} = -\frac{1}{\det \mathbf{J}} \begin{pmatrix} \alpha(g_\mathrm{y} - m_\mathrm{y}) & \beta f_\mathrm{y} \\ -\alpha g_\mathrm{x} & s_\mathrm{x} - \beta f_\mathrm{x} - (1-\beta) l_\mathrm{x} \end{pmatrix} \begin{pmatrix} -\rho \\ 0 \end{pmatrix}, \tag{3.106}$$

where

$$\det \mathbf{J} = \alpha((\mathrm{s_x} - \beta \mathrm{f_x} - (1-\beta)\mathrm{l_x})(\mathrm{g_y} - \mathrm{m_y}) + \mathrm{g_x}\beta \mathrm{f_y}). \tag{3.107}$$

Let's take a closer look at the impact on the steady state of prey and predator, described by

$$\delta_\mathrm{prey} = \frac{\alpha(g_\mathrm{y} - m_\mathrm{y})}{\det \mathbf{J}}\rho, \qquad \delta_\mathrm{pred} = -\frac{\alpha g_\mathrm{x}}{\det \mathbf{J}}\rho. \tag{3.108}$$

In the case that there is no social interaction or strong interference between predators, we can assume $g_\mathrm{y} = 1$, i.e. predation is linear in predator abundance. And if diseases and overcrowding are absent, we can assume linear mortality of the prey, hence $m_\mathrm{y} = 1$. In this scenario, we get

$$\delta_\mathrm{prey} = \frac{\alpha(g_\mathrm{y} - m_\mathrm{y})}{\det \mathbf{J}}\rho = 0, \tag{3.109}$$

hence there is no impact of the small amount of harvesting on the prey. The prey is still controlled by the predator, that is, an additional small amount of loss is compensated by reduced predation. In the equation for the predator, we can see that the predator population is negatively impacted by the loss of its prey. At low loss rates the predator population responds with a proportional loss. Doizy et al. [100] show that this impact analysis can also be applied to much larger generalized models containing more than 200 species, enabling us to predict the species that benefit and those that suffer from a bioinvasion and providing indicator species for the detection of bioinvasion.

If we are not interested in a specific perturbation but rather in the general response to perturbations, we can identify sensitive species, i.e. species that are easily perturbed and influential species, i.e. species that have a strong impact on other species. We can do so by computing the proposed measures by Aufderheide et al. [69]

$$\mathrm{Se_i} = \log\left(-\sum_n \frac{|v_i|^{(n)}}{\lambda_n}\right), \qquad \mathrm{In_i} = \log\left(-\sum_n \frac{|w_i|^{(n)}}{\lambda_n}\right), \tag{3.110}$$

where Se_i is the sensitivity of species i, In_i is the influence of species i, λ_n is the nth eigenvalue of $\mathbf{J}$, and $\boldsymbol{v_n}$ and $\boldsymbol{w_n}$ are the corresponding left and right eigenvectors. For example, $v_2|^{(1)}$ is the second element in the left eigenvector of $\mathbf{J}$ corresponding to the first eigenvalue. We can summarize the two measures in one variable describing the importance by taking the product of the sensitivity and the influence value. In general, these measures allow us to investigate not only species networks but also various other types of networks, such as supply networks. For these we can for example identify the sensitivity of firms and products [101].

We have seen in Sec. 2.1.3 that the Jacobian is commonly used for bifurcation analysis, i.e. to detect where the dynamical behavior of the system changes qualitatively. In generalized models there are two types of bifurcations that can occur, saddle-node-type bifurcations and Hopf bifurcations. Saddle-node-type bifurcations can be easily computed even in large generalized models. Since these bifurcations occurs if the Jacobian has a zero eigenvalue, we have to find the combinations of generalized parameters where the determinant becomes zero. To locate the parameter values that result in a Hopf bifurcation, the method in Guckenheimer et al. [102] is useful for systems with ca. 10 variables (see also [103]). In systems with more variables the equations become too complicated.

Typically during the analysis process, we gain more and more insights into our system of interest. We already saw in the previous section (Sec. 3.3) that we can add these additional information for example in the form of additional constraints and auxiliary variables. Other information may be used to restrict the plausible ranges for our parameter values in the generalized model. Often we can identify interesting parameter regions through the analysis of our generalized mode.

To further explore the dynamical behavior of the system in these regions, we may want to run numerical simulations. This means that we need to construct a conventional model that is consistent with the generalized parameter values of interest. An easy way to do this is to construct specific functions that are consistent with the normalization condition $F(1) = \alpha$, where α is the turnover rate. This results as before in a steady state of 1. Let's consider again the introductory example from the Sec. 3.2

$$\dot{X} = G(X) - L(X), \tag{3.111}$$

where the Jacobian of the normalized system is

$$\mathbf{J} = \alpha(g_{\mathrm{x}} - l_{\mathrm{x}}). \tag{3.112}$$

If we are interested in a steady state at the parameter values $\alpha = 1$, $g_{\mathrm{x}} = 2/3$ and $l_x = 2$, we have to find specific functions that obey the steady state condition and match the desired parameter values. One easy way to find those functions is to use power laws of the form $F(X) = X^p$, for the chosen parameter values one possible example model is

$$\dot{X} = X^{2/3} - X^2. \tag{3.113}$$

However, let's say the first term $X^{2/3}$ would be unrealistic in the context of our application. Instead, we want the function $G(X)$ in the form

$$G(X) = \frac{AX}{X + K}. \tag{3.114}$$

In this case, we first ensure that our function obeys the normalization condition $G(1) = 1$, by setting $A = 1 + K$. In the next step we choose K such that

$$p = \frac{2}{3} = \frac{\partial}{\partial X} \frac{(1 + K)X}{K + X}\bigg|_1 = \frac{K}{(K + 1)}, \tag{3.115}$$

resulting in $K = 2$. So, a specific model consistent with the chosen parameter values is

$$\dot{X} = \frac{3X}{2 + X} - X^2. \tag{3.116}$$

This approach commonly works well for simulation studies, however for the analysis of bifurcations one has to be careful, since the procedure results in degeneracy of certain bifurcations.

3.5 Applications of generalized modeling in ecology

Over the past 13 years, GM has been proven a powerful tool to advance our understanding of complex ecological systems. In the study of food webs, Gross et al. [66] show that minor changes in the shape of functions used in conventional food chain models can have a major impact on the stability of these systems. Through their efficient analysis, generalized models helped to study phenomena in food webs like the paradox of enrichment and helped to identify properties that lend food webs their stability [104, 105, 106]. Considering eigenvector localization in GM, Aufderheide et al. [107] showed that mesoscale symmetries in food webs can explain why certain food webs have a different structure but the same generalized bifurcation diagram.

As shown in the previous section, we can use GM analysis to identify species in the food

web that are most sensitive to perturbations and those that have the strongest impact on others [69, 100]. Yeakel et al. [108] applied this methodology to study a 6,000 years time span of Egyptian mammalian food web structure, reconstructed from Egyptian art history. Species identified as most vulnerable in the model, were indeed the first to go extinct. Lade and Gross [109] propose a new method for identifying early warning signals for critical transitions for example in fish stocks, integrating information trough GM. Generalized models were also applied in the study of delay-coupled networks of populations [110, 111]. Kühn et al. [112] extended the GM approach, using non local generalized models to study periodic predator-prey dynamics. Yeakel and Mangel used GM for the analysis of stock recruitment [113].

Generalized models were also applied to examine food webs in space. First, predator-prey systems were modeled in space trough partial differential equations [114]. This approach was used to study predator interference [43], the dynamics of ecoepidemic models [115] and the impact of nutrient content on predator-prey systems [116]. In more recent times, generalized models were also used to study meta-food webs, i.e. models in which food webs in different spatial patches are coupled by dispersal [117]: from the study of a single population on a spatial network [118] to food webs on two patches [119] up to the dynamics of complex food webs in large spatial networks [120]. Anderson and Fahimipour [121] used GM to study the effects of positive body size scaling of dispersal on the stability of heterogeneous metacommunities.

Another large field of application for generalized models is the study of metabolism. GM in this context is known as *structural kinetic modeling* [122]. Here, generalized models could for example help to identify metabolic components that impact stability in common metabolic pathways like glycolysis, pentose pathway and TCA cycle [98, 123, 124]. Generalized models proved also powerful in investigating the combined ecological and social effects in social-ecological models [125, 126, 127] and in understanding the stability of environmental governance [128].

3.6 Master Stability Approach

Ecological communities are not isolated entities, they inhabit specific spatial contexts, and this spatial dimension can significantly impact the dynamics of these communities. Initial models focused solely on a single species within a spatial framework, referred to as a metapopulation [118, 129, 130]. Following that, metacommunities, comprising competing, similar species and predator-prey systems were modeled, e.g. [119, 131, 132]. The complexity of these models often limits the analyses to either a small number of

species or a small number of habitats. The master stability function approach [120, 133, 134] offers an alternative: Diffusion-driven instabilities can be studied in complex meta-food webs on large spatial networks.

In this section, we explain the master stability function (MSF) approach providing an introductory example (Sec. 3.6.1) and outlining the algorithm (Sec. 3.6.2).

3.6.1 An introductory example

As an example, consider an ecological model featuring two species, A and B, with their population sizes governed by the following set of differential equations

$$\dot{A} = f(A, B) \tag{3.117}$$
$$\dot{B} = g(A, B). \tag{3.118}$$

Usually, we know the functions f and g, allowing us us to calculate the steady states, A^* and B^*

$$f(A^*, B^*) = 0 \tag{3.119}$$
$$g(A^*, B^*) = 0. \tag{3.120}$$

As discussed in Sec. 2.1.2, to analyze the stability of the steady state, we examine the eigenvalues of the Jacobian matrix $\mathbf{P}$

$$\mathbf{P} = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}, \tag{3.121}$$

with

$$P_{11} = \partial_A \dot{A}\Big|_* \tag{3.122}$$
$$P_{12} = \partial_B \dot{A}\Big|_* \tag{3.123}$$
$$P_{21} = \partial_A \dot{B}\Big|_* \tag{3.124}$$
$$P_{22} = \partial_B \dot{B}\Big|_*. \tag{3.125}$$

The steady state is considered stable if all eigenvalues of the Jacobian have negative real parts.

Let's now consider the scenario in which the two populations can disperse at rates r_A and

r_B between two habitat patches. We can write this system as

$$\dot{A}_1 = f(A_1, B_1) - r_a A_1 + r_a A_2 \tag{3.126}$$
$$\dot{B}_1 = g(A_1, B_1) - r_b B_1 + r_b B_2 \tag{3.127}$$
$$\dot{A}_2 = f(A_2, B_2) - r_a A_2 + r_a A_1 \tag{3.128}$$
$$\dot{B}_2 = g(A_2, B_2) - r_b B_2 + r_b B_1. \tag{3.129}$$

Since (A^*, B^*) is a steady state of the local system

$$A_1 = A_2 = A^* \tag{3.130}$$
$$B_1 = B_2 = B^* \tag{3.131}$$

is a steady state of the spatial system. We refer to these solutions as homogeneous states. Again, to investigate stability of the steady state, we consider the Jacobian matrix $\mathbf{J}$. J_{11} is computed as follows

$$J_{11} = \frac{\partial}{\partial A_1}\dot{A}_1\bigg|_* \tag{3.132}$$
$$= \frac{\partial}{\partial A_1}\Big[f(A_1, B_1) - r_a A_1 + r_a A_2\Big]\bigg|_* \tag{3.133}$$
$$= \frac{\partial}{\partial A_1} f(A_1, B_1)\bigg|_* - r_a \tag{3.134}$$
$$= \frac{\partial}{\partial A} f(A, B)\bigg|_* - r_a \tag{3.135}$$
$$= P_{11} - r_a. \tag{3.136}$$

In the same way we derive entries $J_{12} - J_{14}$ of the Jacobian

$$J_{12} = \frac{\partial}{\partial B_1}\Big[f(A_1, B_1) - r_a A_1 + r_a A_2\Big]\bigg|_* = P_{12} \tag{3.137}$$
$$J_{13} = \frac{\partial}{\partial A_2}\Big[f(A_1, B_1) - r_a A_1 + r_a A_2\Big]\bigg|_* = r_a \tag{3.138}$$
$$J_{14} = \frac{\partial}{\partial B_2}\Big[f(A_1, B_1) - r_a A_1 + r_a A_2\Big]\bigg|_* = 0. \tag{3.139}$$

Computing all the entries yields the Jacobian matrix for the spatial system

$$\mathbf{J} = \begin{pmatrix} P_{11} - r_a & P_{12} & r_a & 0 \\ P_{21} & P_{22} - r_b & 0 & r_b \\ r_a & 0 & P_{11} - r_a & P_{12} \\ 0 & r_b & P_{21} & P_{22} - r_b \end{pmatrix}. \tag{3.140}$$

Due to the specific structure of this Jacobian, we can re-write the matrix in block form

$$\mathbf{J} = \begin{pmatrix} \mathbf{P} - \mathbf{C} & \mathbf{C} \\ \mathbf{C} & \mathbf{P} - \mathbf{C} \end{pmatrix}, \tag{3.141}$$

where $\mathbf{P}$ is the Jacobian matrix of the local system and $\mathbf{C}$ is the coupling matrix of the form

$$\mathbf{C} = \begin{pmatrix} r_a & 0 \\ 0 & r_b \end{pmatrix}. \tag{3.142}$$

Using Kronecker products, we can re-write the Jacobian as

$$\mathbf{J} = \mathbf{I} \otimes \mathbf{P} - \mathbf{L} \otimes \mathbf{C}, \tag{3.143}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{L}$ is the Laplacian of the spatial network

$$\mathbf{L} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{3.144}$$

To gain insights into the stability of the spatial system our aim is now to compute the eigenvalues and eigenvectors. Since the Jacobian of the spatial system is the sum of the two Kronecker products we can assume that its eigenvectors can likewise be expressed in the form of Kronecker products [120], i.e.

$$\boldsymbol{w} = \boldsymbol{v} \otimes \boldsymbol{s}, \tag{3.145}$$

such that

$$\mathbf{J}\boldsymbol{w} = \lambda \boldsymbol{w} \tag{3.146}$$

$$\mathbf{J}(\boldsymbol{v} \otimes \boldsymbol{s}) = \lambda(\boldsymbol{v} \otimes \boldsymbol{s}). \tag{3.147}$$

Exploiting the properties of Kronecker products we can write

$$\begin{aligned}
\mathbf{J}\boldsymbol{w} &= (\mathbf{I}\otimes\mathbf{P}-\mathbf{L}\otimes\mathbf{C})(\boldsymbol{v}\otimes\boldsymbol{s}) && (3.148)\\
&= (\mathbf{I}\otimes\mathbf{P})(\boldsymbol{v}\otimes\boldsymbol{s})-(\mathbf{L}\otimes\mathbf{C})(\boldsymbol{v}\otimes\boldsymbol{s}) && (3.149)\\
&= (\mathbf{I}\boldsymbol{v}\otimes\mathbf{P}\boldsymbol{s})-(\mathbf{L}\boldsymbol{v}\otimes\mathbf{C}\boldsymbol{s}) && (3.150)\\
&= (\boldsymbol{v}\otimes\mathbf{P}\boldsymbol{s})-(\mathbf{L}\boldsymbol{v}\otimes\mathbf{C}\boldsymbol{s}). && (3.151)
\end{aligned}$$

We can make progress if $\boldsymbol{v}$ in an eigenvector of $\mathbf{L}$, such that $\mathbf{L}\boldsymbol{v}=\kappa\boldsymbol{v}$, where κ is the corresponding eigenvalue of $\mathbf{L}$. We can write

$$\mathbf{J}\boldsymbol{w}=(\boldsymbol{v}\otimes\mathbf{P}\boldsymbol{s})-(\kappa\boldsymbol{v}\otimes\mathbf{C}\boldsymbol{s}). \tag{3.152}$$

Now we need to bring the matrices together, i.e.

$$\begin{aligned}
\mathbf{J}\boldsymbol{w} &= (\boldsymbol{v}\otimes\mathbf{P}\boldsymbol{s})-\kappa(\boldsymbol{v}\otimes\mathbf{C}\boldsymbol{s}) && (3.153)\\
&= (\boldsymbol{v}\otimes\mathbf{P}\boldsymbol{s})-(\boldsymbol{v}\otimes\kappa\mathbf{C}\boldsymbol{s}) && (3.154)\\
&= \boldsymbol{v}\otimes(\mathbf{P}\boldsymbol{s}-\kappa\mathbf{C}\boldsymbol{s}) && (3.155)\\
&= \boldsymbol{v}\otimes(\mathbf{P}-\kappa\mathbf{C})\boldsymbol{s}. && (3.156)
\end{aligned}$$

Assuming that $\boldsymbol{s}$ is an eigenvector of $\mathbf{P}-\kappa\mathbf{C}$, so that $(\mathbf{P}-\kappa\mathbf{C})\boldsymbol{s}=\lambda\boldsymbol{s}$, where λ is the corresponding eigenvalue, we can write

$$\begin{aligned}
\mathbf{J}\boldsymbol{w} &= \boldsymbol{v}\otimes\lambda\boldsymbol{s} && (3.157)\\
&= \lambda(\boldsymbol{v}\otimes\boldsymbol{s}). && (3.158)
\end{aligned}$$

To recap, we have seen that if $\boldsymbol{v}$ is an eigenvector of $\mathbf{L}$ with eigenvalue κ and $\boldsymbol{s}$ is an eigenvector of $\mathbf{P}-\kappa\mathbf{C}$ with eigenvalue λ, then $\boldsymbol{v}\otimes\boldsymbol{s}$ is an eigenvector of $\mathbf{J}$ and the corresponding eigenvalue is also λ. Consequently all eigenvectors of $\mathbf{J}$ can be constructed in this way, revealing the complete set of eigenvalues of $\mathbf{J}$ [120]

$$\mathrm{Ev}(\mathbf{J})=\bigcup_{\mathrm{n}=1}^{\mathrm{N}}\mathrm{Ev}(\mathbf{P}-\kappa_{\mathrm{n}}\mathbf{C}). \tag{3.159}$$

Following this equation, we can define the so-called master stability function (MSF) [120]

$$M(\kappa)=\mathrm{Re}[\lambda_{\max}(\kappa)], \tag{3.160}$$

which is the real part of the leading eigenvalue λ_{max} of $\mathbf{P} - \kappa_n \mathbf{C}$ as a function of κ. Using this approach, the impact of the spatial topology on the dynamics is captured solely in the Laplacian eigenvalues. In conclusion stability of the homogeneous state is lost if any Laplacian eigenvalue κ falls into the range where $M(\kappa)$ is positive.

3.6.2 Master Stability Function

To construct the master stability function (MSF), we proceed as follows:

1. We compute the Jacobian $\mathbf{P}$ of the local system.
2. We construct the matrix $\mathbf{P} - \kappa_n \mathbf{C}$, handling κ as an unknown parameter.
3. We compute the leading eigenvalue of $\mathbf{P} - \kappa_n \mathbf{C}$ as a function of κ.
4. The real part of the leading eigenvalue as a function of κ yields the master stability function $M(\kappa)$.

Master stability functions offer the ability to efficiently explore instabilities in coupled identical systems. Also MSFs are useful for designing networks with specific stability properties by choosing κ according to the MSF such that the desired characteristics are met.

3.7 Summary

In this chapter, I have reviewed the GM approach, a powerful tool to capture the dynamics of uncertain systems. Additionally, I introduced the master stability approach that can be used to expand the GM analysis to spatial systems. In the upcoming chapters, GM will be a key tool to explore the underlying mechanisms of complex ecological communities.

Chapter 4

Bacterial niche occupancy dynamics

Advances in molecular methods enable us to monitor bacterial populations in time [21]. However, due to the tremendous diversity of bacteria, understanding their community dynamics and its links with ecosystem functioning remains challenging. We need conceptual frameworks that make sense of time series of taxonomically diverse bacterial communities in terms of their potential ecological function. In ecology, a key concept for such synthesis is the niche, the set of strategies that enable a population to survive and that shape its impacts on the surroundings.

In this chapter I present a framework based on diffusion maps [48] to coarse-grain taxonomically-rich bacterial communities in terms of their metabolic strategies and quantitatively organize genomic information in terms of potentially occupied metabolic niches over time. Applying this framework to a long-term bacterial time series from the Baltic Sea, I am able to reconstruct the dynamics of putatively occupied metabolic niches.

This chapter follows the lines of Massing et al. 2023 [135]. I start with a general introduction to the dynamics of marine bacterial communities (Sec. 4.1). In Sec. 4.2 I apply the diffusion map approach to a dataset of a Baltic Sea bacterial community time series. The method identifies important metabolic strategies and delineates the structure of the niche space. In Sec. 4.3 I translate the species time series to potentially occupied metabolic niches over time. In addition, I estimate the functional diversities of the Baltic Sea bacterial community over time (Sec. 4.4). Finally, I discuss the results in Sec. 4.5.

4.1 Dynamics of marine bacterial communities

Marine bacterial communities are a critical component of global element cycles [10], of the marine microbial food web [23] and in interacting with other micro- and macroorganisms (e.g. [11, 12, 136]). With over 40,000 marine microbial species detected so far [82, 137], marine bacterial communities are highly diverse. In addition to their high species diversity, these communities also display a high diversity in terms of their dynamics over time. As

a result of short generation times, in surface waters normally hours to days [22, 138], and their diverse interactions, marine bacterial communities show complex dynamics fluctuations in abundances and species composition on the daily, monthly and annual scale [22, 23].

Studies of short- and long-term time series of marine bacterial communities show that their composition changes in response to multiple forces acting over different time-scales [22]. In spite of strong variations on shorter scales, these communities show stability in their average community composition [22]. Biological interactions among bacteria themselves and between bacteria and other organisms, for example, phytoplankton, play an important role in shorter-scale community dynamics [34]. Driven by changes in multiple interacting environmental features, bacterial communities in surface waters in temperate and polar regions show strong seasonal patterns [23]. Despite the insights gained on these communities, their tremendous diversity remains a challenge for data analysis.

Progress in understanding bacterial community dynamics and its links with ecosystem functions requires abstracting from individual taxa and coarse-graining of these complex communities in terms of their occupied niches over time. To achieve this, we need to identify new variables that describe our data sets in reduced dimensions. We have seen in Chapt. 2.3 that the diffusion map [48] finds new dynamically relevant variables that describe the most important dimensions in a system. In the following we apply the diffusion map to reconstruct bacterial metabolic strategies from a monitoring dataset and subsequently use the identified metabolic strategies to convert our species time series into strategy time series.

4.2 Reconstructing the metabolic strategy space

4.2.1 Sampling data

Seawater samples from the Linnaeus Microbial Observatory (LMO) were collected approximately on a weekly basis from 2011 to 2013 and monthly from 2014 to 2019. The observatory is located at N 56°55.854', E 17°3.6420' in the Western Baltic Proper. Water sampling was conducted using 3- or 5-L Ruttner water samplers, extracting samples from a depth of 2 meters at approximately 9 a.m. during each sampling event. In the laboratory at Linnaeus University, seawater was processed, during which environmental parameters, i.e. temperature, salinity, chlorophyll a, dissolved organic carbon (DOC), nitrate and nitrite (together named as nitrate), phosphate, silicate, and ammonium were analyzed using methods previously outlined [139, 140, 141]. To assess microbial community composition,

seawater was filtered either directly onto 0.22 µm Sterivex cartridge filters (Millipore) or initially prefiltered onto 3 µm polycarbonate filters and then further processed through 0.22 µm Sterivex cartridges (collectively referred to as the 3-0.2 µm size fraction), using a peristaltic pump. The filters were preserved in Tris–EDTA (TE) buffer at -80°C until DNA extraction. The extraction process followed a phenol–chloroform method outlined by Boström et al. [142] and modified after Bunse et al. [143]. The V3V4 region of the 16S rRNA gene was amplified through PCRs, employing the primer pair 341f-805r [144, 145]. DNA concentrations were assessed using either a NanoDrop or Qubit 2.0 Fluorometer (Life Technologies) and gel electrophoresis was conducted to verify the specificity of the amplicons. Sample batches for sequencing were successively dispatched to the Science for Life Laboratory, Sweden, on the Illumina MiSeq platform, leading to 2 × 300 bp paired-end reads. Bioinformatic processing was conducted using the nf-core Ampliseq pipeline [146, 147] with the following software versions: nf-core/ampliseq = v1.2.0dev; Nextflow = v20.10.0; FastQC = v0.11.8; MultiQC = v1.9; Cutadapt = v2.8; and QIIME2 = v2019.10.0. DADA2 [148] implemented in QIIME2 [149] was employed and trimmed the sequences at forward 259 bp and reverse 199 bp before denoising. Among all samples from the LMO, we included all filter fractions in the niche space analysis. For abundance estimates, only the non-prefiltered 0.22 µm fraction was considered (refer to the detailed method description below).

4.2.2 Obtaining genomes and genes from ASV data

We obtained the GTDB reference database [150] of all 16S rRNA gene sequences from genomes that passed quality control criteria employed by GTDB [151]. Identification of sequences were performed with nhmmer [152] using the 16S rRNA model from the RFAM database [153]. Subsequently, we conducted a BLAST sequence similarity search using default parameters to match denoised sequence variants present in each LMO sample to the BLAST database and retain the top hits. In cases of multiple matches with equal similarity, we retained a randomly-chosen representative. Out of 48,098 Amplicon Sequence Variants (ASVs), 21,102 (so 44%) could be matched well (similarity greater than 95%) to a genome from the GTDB. In terms of abundances, a mean of 82% (column sums) were good matches. From the well-matched species we obtained 4,265 complete genomes from GTDB and NCBI (Refseq and GenBank). These complete genomes were annotated using Prokka [154].

4.2.3 Diffusion mapping the strategy space

Following the algorithm explained in Chapt. 2.3.2, we diffusion map the dataset of genomes and genes using the similarity in gene composition as similarity measure between the genomes. The starting point for the analysis is the data table of genomes (rows) and annotated genes (columns). If a certain gene is present in a genome the entry is 1 for the respective column (gene) and row (genome), if the gene is absent in the genome the table entry is 0. For the diffusion map analysis, we first need to establish some notion of similarity between the genomes. Due to the presence-absence type of data, we calculate the hamming distance between all genome pairs, counting their differences in gene composition. Taking the inverse of the hamming distances yields the similarity values. Next, we threshold the similarity matrix, keeping only the top-25 highest similarity entries for each datapoint. From the thresholded similarity matrix we calculate the row-normalized Laplacian matrix. Finally, we compute the eigenvectors and eigenvalues of the Laplacian matrix.

The result of the diffusion map analysis are new variables that are composites of metabolic capabilities of the analyzed taxa, i.e. metabolic strategies (Fig.4.1). The eigenvectors assign an entry to each genome for each new variable and the corresponding eigenvalues provide information on the importance of the identified variable. The most important eigenvector, hereafter called variable 1, corresponds to the smallest non-zero eigenvalue. The eigenvector corresponding to the second-smallest non-zero eigenvalue is the second most important, hereafter called variable 2, and so on.

Each new variable represents a dimension of the metabolic strategy space [44]. There are variables that trace a continuum in the strategy space, ranging from one pure strategy at one end to another pure strategy at the other end. In addition, we find variables that correspond to localized eigenvectors, i.e. there is a clear separation between non-zero and almost zero entries in the vector. The former type of variables describes a true continuum in the data, for example going from genomes that have the full set of genes for a certain strategy at one end, over partial sets to another strategy at the other end. The set of genes for this strategy is gradually completed as we approach the other end of the variable. Localized eigenvectors identify a separation in the data, i.e. the set of genes for a strategy is almost complete or almost completely absent with little middle ground. Each variable is divided into positive and negative side for interpretation.

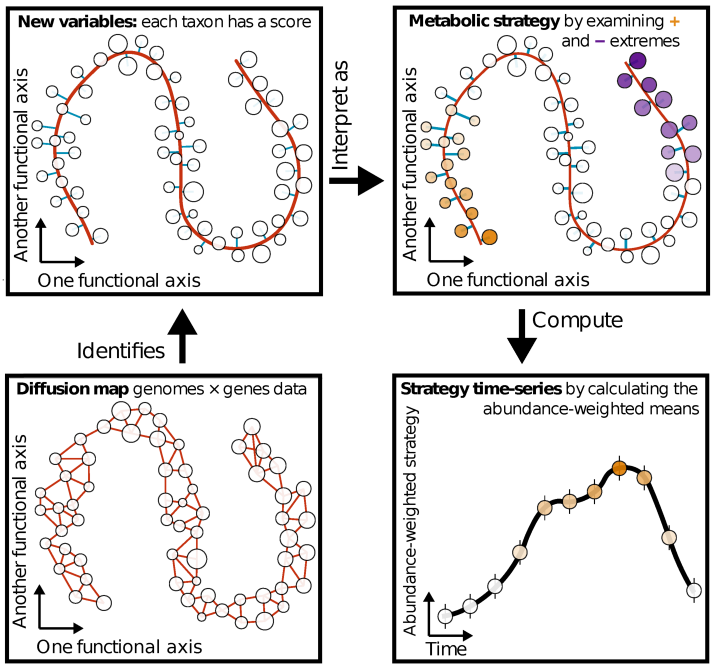

Figure 4.1: Overview of the procedure from diffusion mapping the dataset of genomes and genes to conversion of the species time series into strategy time series. The circles represent the genomes, which have different abundances over time (size). The two axes in the figure are just representatives for the many axes that describe the known gene content of the genomes in a high-dimensional space. Figure is adapted from Massing et al. [135] with modifications.

4.2.4 Identifying metabolic strategies

To identify the metabolic strategies that are described by the new variables, we examine the genomes that score extreme values in the respective variable and check the genes that are enriched in these genomes. For the sake of simplicity, we refer to the genomes with extreme variable entries as 'extreme taxa' in the following. To reveal overrepresented genes of the extreme taxa we use a permutational variant of the gene set enrichment analysis, GSEA [155]. Given an apriori defined set of genes S the goal of GSEA is to determine whether the members of S are randomly distributed throughout a ranked list (according to their appearance in genomes with high or low variable values) or if they are primarily found at the top or bottom. The orderings specified by each diffusion variable are used to rank the genomes. We perform gene set enrichment analysis with a Benjamini-Hochberg-adjusted [156] P value <0.01 used as a threshold for retaining genes corresponding to extreme taxa in the respective diffusion variable. For analysis we separate the negative and positive values for each diffusion variable and we refer to these sides as "variable x negative" and "variable x positive" respectively, where "x" corresponds to the variable number.

4.2.5 Important metabolic strategies in the Baltic Sea data

Variable 1, the most important variable identified by the diffusion map, separates a group of Gammaproteobacteria clearly from the rest of the taxa (Fig. 4.2A). The distinct group consists primarily of human- and animal-associated Enterobacteriaceae, for example *Cronobacter*, *Salmonella* and *Yersinia* (Table S2). While this group of a total of 92 genomes scores large negative values in variable 1, all other genomes are assigned values close to zero. This localized variable indicates that the separated taxa possess a cluster of unique capabilities. Overrepresented genes in these taxa include genes encoding machinery for iron acquisition common in Enterobacteriaceae, such as the Enterobactin synthase component F [157, 158], genes responsible for the flagellar formation [159] and genes associated to biofilm formation [160] (Table S3).

Given what is known about the ecological role of Enterobacteriaceae in marine ecosystems, this result may seem surprising at first. Despite the extremely low abundances of these taxa in the samples (mean relative abundances of 0.007) the diffusion map detects this variable as the most important one. Nevertheless, the distinct separation of this group from the rest of the taxa makes sense in the light of the well-known biases towards sequences of pathogenic taxa and genes associated with pathogenesis in global databases (e.g. [161]). As a result, these taxa stand out as significantly distinct from other bacterial species in the Baltic Sea community. This illustrates the power of the diffusion map in uncovering such differences and identifying biases within the dataset.

The diffusion map also reveals many strategies more relevant for the Baltic Sea bacterial community. For example, the localized variable 4 negative detects cyanobacterial photosynthesis. It separates all the Cyanobacteria (large negative values) from the rest of the taxa (positive or close-to-zero values) (Fig. 4.2D). Enriched genes encode the subunits of photosystem I and photosystem II as well as associated cytochrome components and cyanobacterial-specific light-harvesting antennae [162]. The localized character of this variable supports the findings that cyanobacterial photosynthesis is a yes-or-no strategy [44], indicating that the oxygenic photosynthetic lifestyle has wide ranging metabolic consequences. Indeed, this photosynthetic lifestyle is for example expensive in terms of avoiding and repairing photoinhibition and -damage. These costly adaptations mean that the invested energy cannot be spent for other metabolic pathways [163].

Variable 2 and variable 3 (Fig. 4.2B,C) are examples of variables that trace a continuum in the strategy space. At the positive extremum in variable 2 are marine host-associated Gammaproteobacteria, such as *Photobacterium*, *Shewanella* and *Vibrio*, whereas oligotrophic Gammaproteobacteria and Alphaproteobacteria are assigned values close to zero.

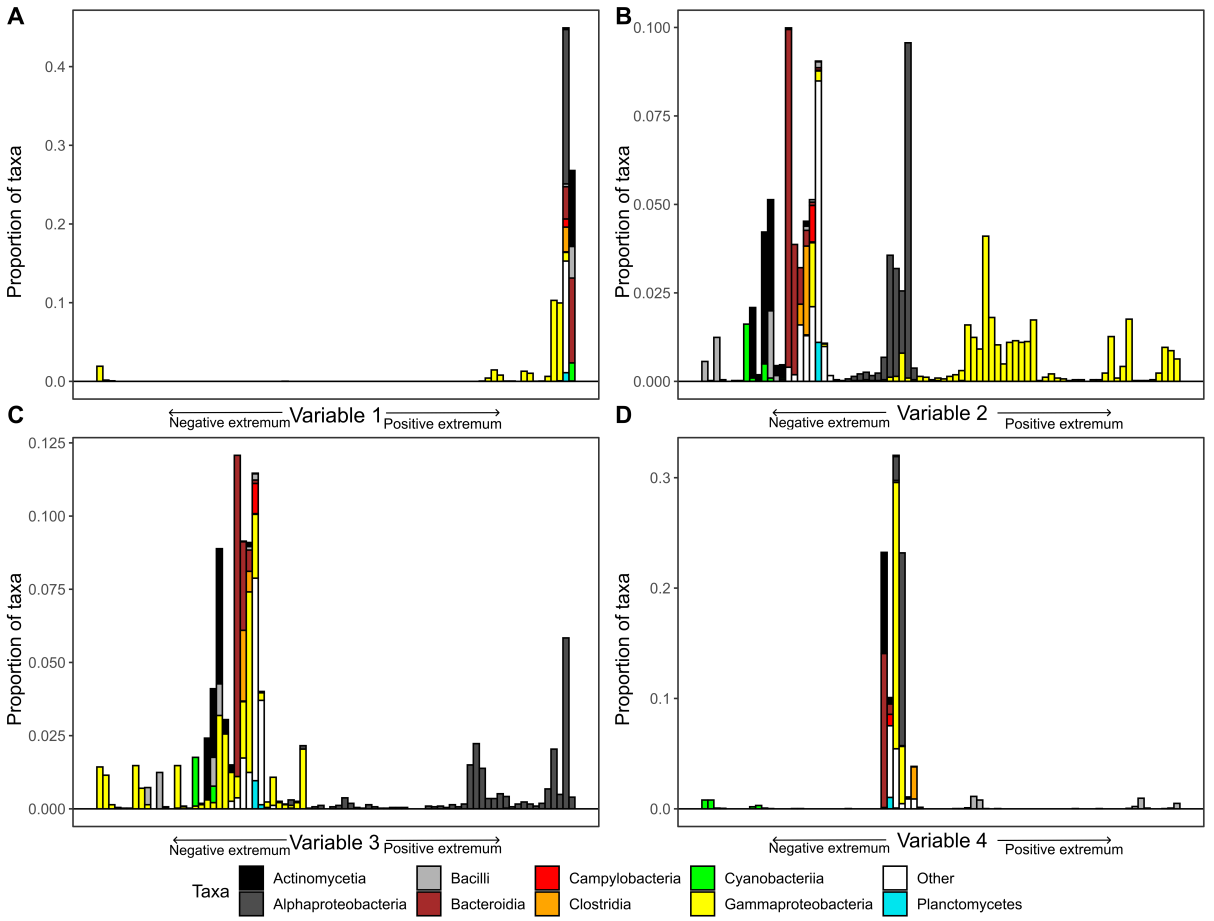

Figure 4.2: The ordering of taxa defined by variable entries 1-4 (A-D), from negative to positive (left to right). The taxonomic compositions corresponding to these variable entries are shown for each of 80 equally spaced bins. Figure is adapted from Massing et al. [135] with modifications.

Chemotaxis and response to various stressors are the most correlated capabilities for the taxa situated at the positive end of variable 2. Our interpretation is that variable 2 positive represents the metabolic strategy of marine host-associated Gammaproteobacteria. Variable 3 reveals different strategies of marine Alphaproteobacteria: Free-living Pelagibacterales and the obligate intracellular pathogens Rickettsiales, known for their streamlined genomes, score close-to-zero values, while Rhodobacteraceae and Rhizobiales, known for their capability of utilizing a variety of carbon sources [164, 165], are assigned large positive values. Since the overrepresented genes at the positive end of this variable encode machinery for the utilization of various dissolved organic compounds, e.g. phosphonate, acetate and urea that constitute exudates of phytoplankton [166], we interpret this variable as metabolic approach of Alphaproteobacteria that enables them to utilize a wide range of carbon sources.

Diffusion mapping also identifies strategies that subdivide groups that cluster together in other variables. For example, the Cyanobacteria that all scored large negative val-

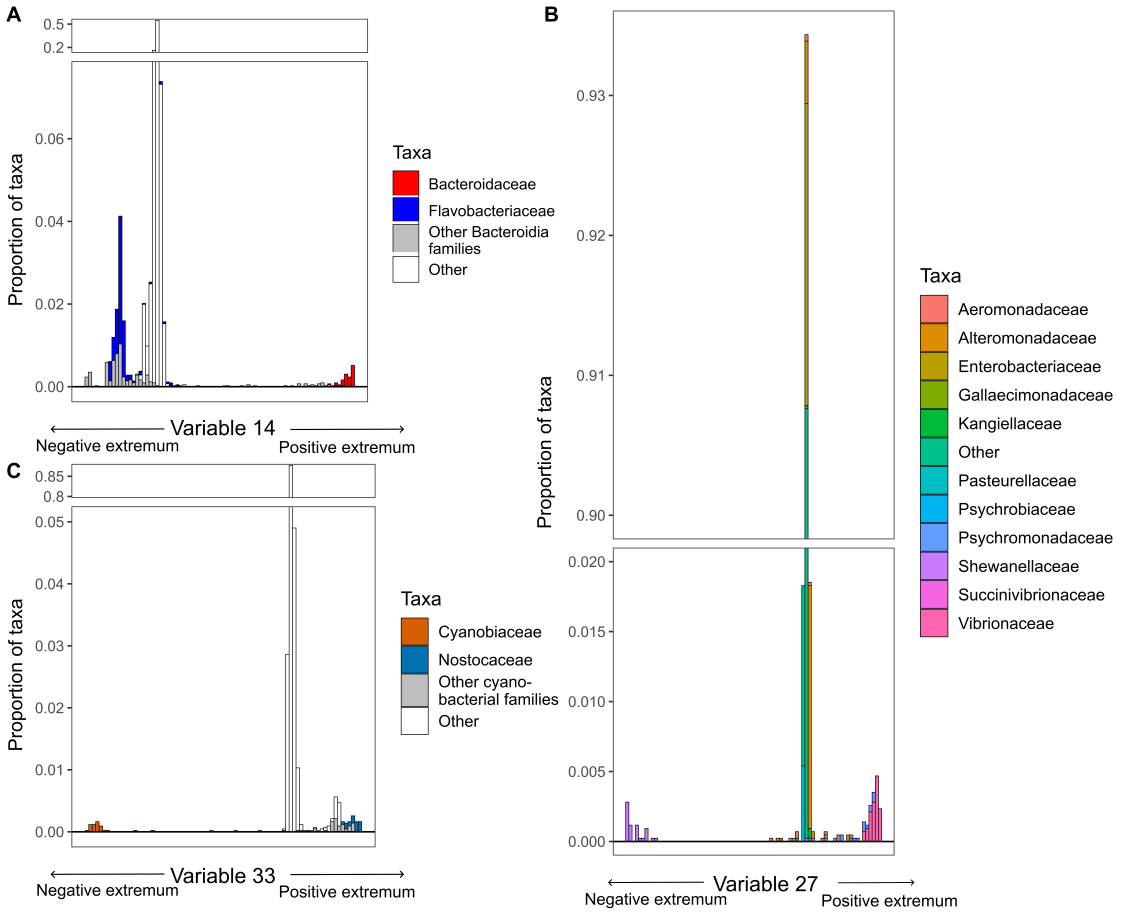

Figure 4.3: The ordering of taxa defined by variable entries 14, 27 and 33 (A-C), from negative to positive (left to right). The taxonomic compositions corresponding to these variable entries are shown for each of 80 equally spaced bins. Figure adapted from Massing et al. [135] with modifications.

ues in variable 4 are separated in variable 33 (Fig. 4.3C). While the Picocyanobacteria group towards the negative extremum, all other cyanobacterial genomes score positive values. Among these, the cyanobacterial family Nostocaceae, that is known to form heterocysts, scores highest values. Other examples are variable 27 (Fig. 4.3B) that separates the Enterobacterales of variable 2 positive into the family Shewanellaceae at the negative end and the other enterobacterial families like Vibrionaceae at the positive end as well as variable 14 (Fig. 4.3A) that divides the Bacteroidota into complex polysaccharide degraders e.g. *Flavobacterium* [167] on the negative side and anaerobic, intestinal Bacteroidota e.g. *Prevotella* [168] on the positive side. Genes for different CAZymes (carbohydrate-active enzymes) [169], responsible for the degradation of major plant cell wall components, are overrepresented in the genomes of the complex polysaccharide degraders.

Furthermore, there are variables that group genera from different taxonomic groups to-

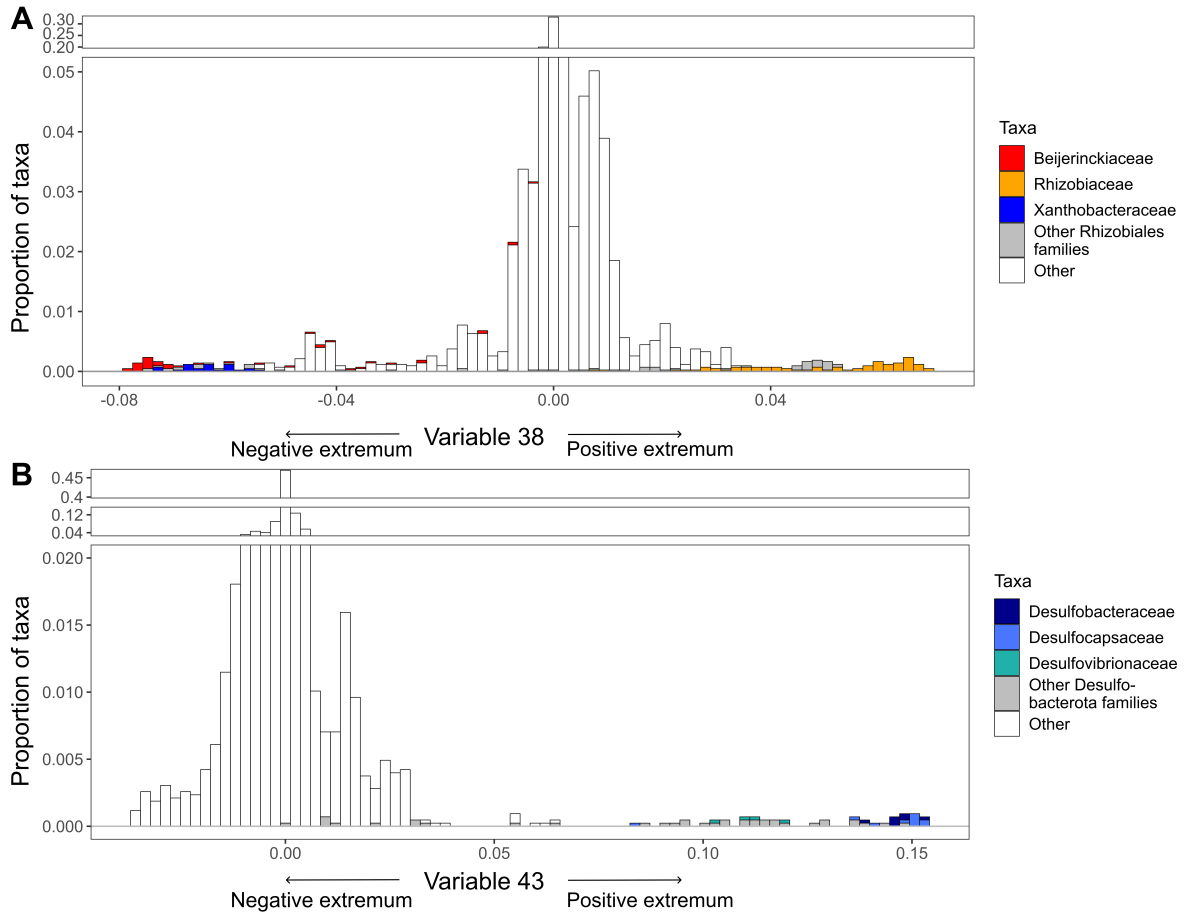

Figure 4.4: The ordering of taxa defined by variable entries 38 (A) and 43 (B), from negative to positive (left to right). The taxonomic compositions corresponding to these variable entries are shown for each of 80 equally spaced bins. Figure is adapted from Massing et al. [135] with modifications.

gether. For example, bacteria with the capability to oxidize methyl groups and C1 compounds (e.g. formaldehyde and methanol) belonging to different taxonomic families such as Acetobacteraceae, Beijerinckiaceae and Xanthobacteraceae are found at the negative end of variable 38 (Fig. 4.4A). Genes encoding machinery for formaldehyde and methanol degradation [170] are correlated to this strategy. Variable 43, associated with genes for sulfate respiration [171], clusters positive non-spore forming sulfate-reducing bacteria from the families Desulfocapsaceae, Desulfobacteraceae, Desulfurivibrionaceae and others together (Fig. 4.4B).

4.2.6 Inferred niche space

The examples explained above demonstrate that the diffusion variables provide possibly hundreds of meaningful coordinates that trace the space of bacterial metabolic strategies. Collectively, these strategies delineate the metabolic niche space of the community and al-

locate taxa to specific coordinates in a multidimensional space. Using PHATE [45], which we introduced in Chapt. 2.3.4, we combined the diffusion variables in a low-dimensional visualization of the strategy space. Despite the fact that low-dimensional embeddings should be interpreted with caution, the structure that we observe shows resemblance to the niche space of a previous study [44]: the strategy space of the Baltic Sea bacterial community is represented as a tree-like structure comprising clusters of taxa featuring localized traits and continuous branches (Fig. 4.5).

The coarse structure of the metabolic niche space roughly follows the taxa's phylogeny. The core of the branched structure is comprised of uncultured taxa and streamlined genomes, including Patescibacter [172], Pelagibacterales [173], and Rickettsiales [174]. Their genomes predominantly retain essential basic functions required for survival and reproduction, which they share with many other organisms and they lack many of the more 'specialized' genes [175]. The grouping of uncultured taxa towards the center occurs either because they also possess streamlined genomes or due to the limited knowledge about their genes and respective functions. Consequently, their position in the metabolic niche space may undergo changes as further knowledge is acquired.

Distinct clusters within the structure are formed by taxa characterized by localized variables such as the Cyanobacteria. Their separation could reflect that an intermediate strategy is not feasible due to certain trade-offs resulting from adopting the respective strategy, that applying the strategy demands complex and costly adaptations [176] and that the acquisition of necessary machinery is not easily achievable, like through horizontal gene transfer [177]

Bacterial taxa linked to human disease, such as species related to *Klebsiella*, *Mycobacterium*, *Staphylococcus*, and *Fusobacterium*, are positioned farthest from the central structure and manifest as peripheral clusters of dots. This spatial arrangement can probably be attributed to the bias in global databases, which primarily focus on human pathogens and their associated functional traits, as discussed earlier.

Another interesting feature of this mapping is the amount of white space, the geometric structure indicates significant unoccupied areas within the metabolic niche space. This suggests the presence of either strategies that are not viable within the specific ecosystem or in general, or strategies that have not yet been discovered [44].

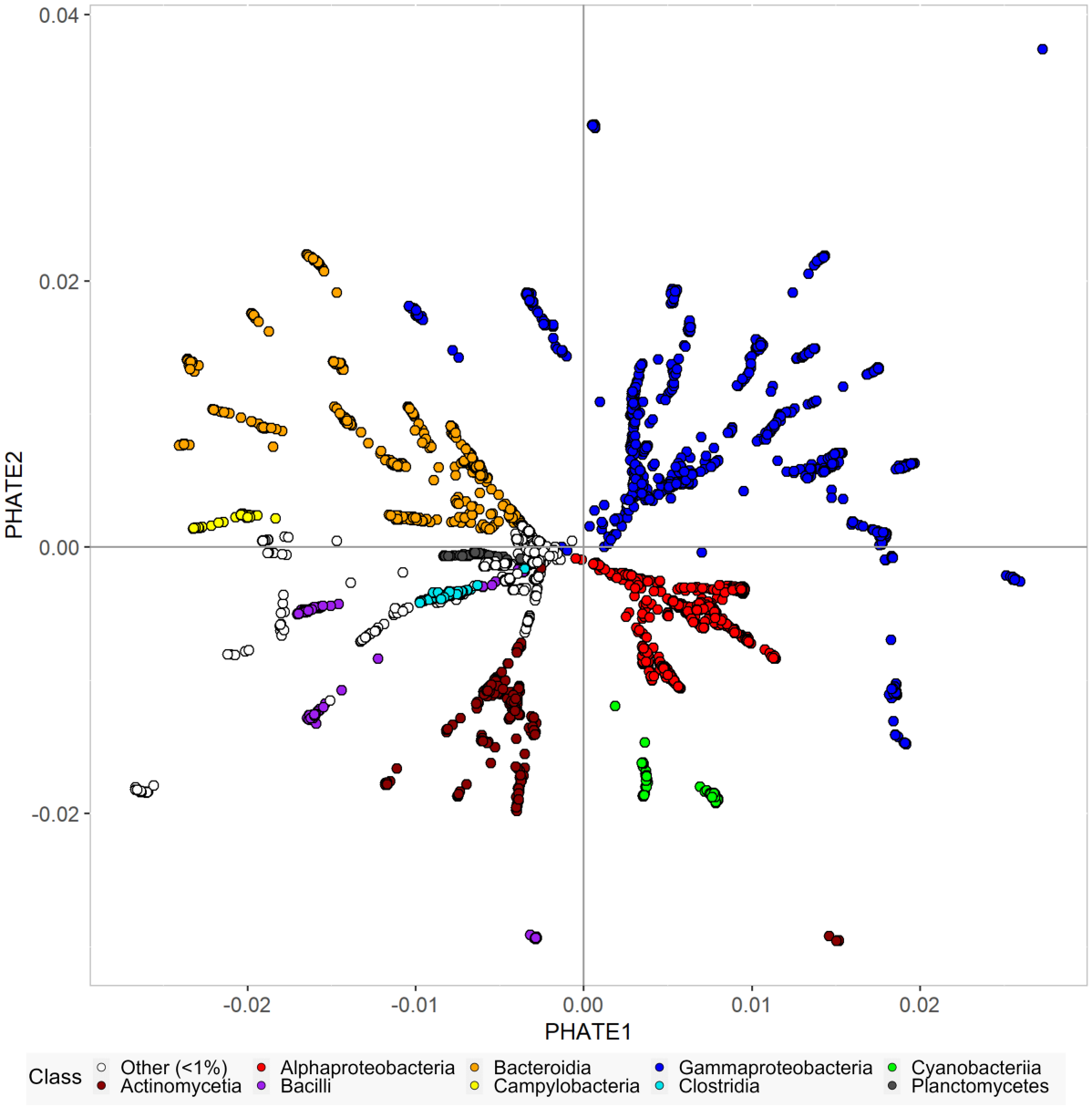

Figure 4.5: Two-dimensional embedding of diffusion variables created using the PHATE algorithm [45]. Each point represents an individual genome that is coloured by taxonomic class. Figure adapted from from Massing et al. [135] with modifications.

4.3 Dynamics of occupied metabolic niches

4.3.1 Translation into strategy time series

In the previous section we saw that the diffusion map approach identifies new variables from the dataset that represent interpretable metabolic bacterial strategies. Our next aim is to understand the bacterial communities in terms of their occupied metabolic strategies over time, i.e. we want to use the identified metabolic strategies to convert our species time series into strategy time series.

As previously mentioned, for our analysis, we divided the diffusion variables into negative and positive side. Considering the relative abundances of the amplicon sequence variants (ASVs), we calculated abundance-weighted mean values for each variable side. For a given variable side v and sample k, the abundance-weighted mean strategy value is defined as follows:

$$\text{mean strategy value}_{v,k} = \frac{\sum_{i=1}^{n} v_i p_{k,i}}{\sum_{i=1}^{n} p_{k,i}}, \tag{4.1}$$

where v_i represents the variable entry of variable side v assigned to genome i, $p_{k,i}$ denotes the relative abundance of the ASV corresponding to genome i in this particular sample, and n is the total number of genomes. To translate the species time series into a strategy time series, we computed the weighted means of each diffusion variable side, i.e. metabolic strategy for every sample, corresponding to each sampling time-point. As a result, this allows us to observe the dynamic changes in the occupation of metabolic strategies over time within the Baltic Sea bacterial community (Fig. 4.1).

4.3.2 Metabolic strategies over time in the Baltic Sea

The occupation of metabolic niches shows a variety of patterns over time. For example, the strategy of utilizing a variety of carbon sources (variable 3 positive) and the capability of metabolizing complex polysaccharides (variable 14 negative) both reach high abundance-weighted mean values in May (Fig. 4.6), following the peak of the phytoplankton spring bloom [139]. The occupation of the two strategies however differ in their evolution over time: The capability of metabolizing complex polysaccharides, primarily governed by members of the Flavobacteriaceae, shows a pronounced peak in May. In contrast, the strategy of utilizing a variety of carbon sources attains high values in May without displaying a distinct peak, instead it decreases more slowly, reaching a minimum mean value in September.

The peak of the strategy values of metabolizing complex polysaccharides is probably

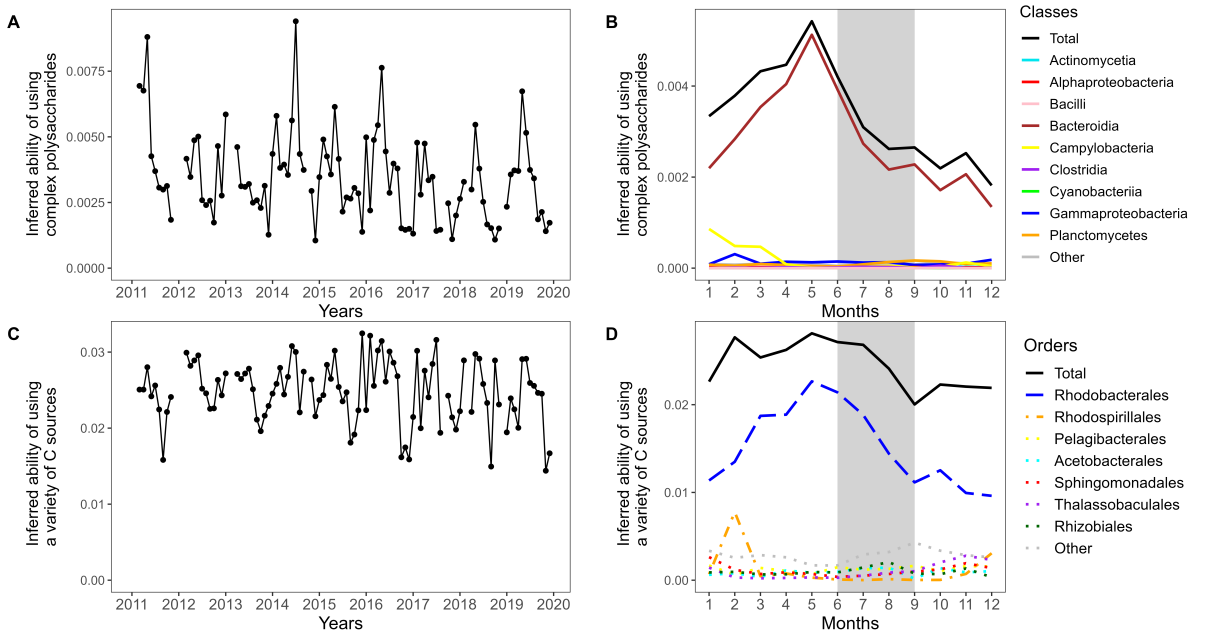

Figure 4.6: Abundance-weighted mean values of variable 14 negative, i.e. the inferred ability of using complex polysaccharides (A, B) and of variable 3 positive, i.e. inferred ability of using a variety of carbon sources (C, D) over the whole time period and over the yearly cycle. Summer months are indicated by a gray background. Taxonomic orders (B) and taxonomic classes (D) are color-coded. Figure is adapted from Massing et al. [135] with modifications.

closely linked to the phytoplankton bloom, since phytoplankton exudes significant quantities of photosynthetic products, primarily polysaccharides [178], and Flavobacteria have developed adaptations to effectively utilize these high-molecular-weight molecules [167]. During spring 2011, which coincided with the peak phytoplankton biomass [139], the occupation of this strategy shows particularly high levels. The strategy of utilizing a variety of carbon sources (variable 3 positive) is dominated by the marine Rhodobacteraceae, which play a crucial role in metabolizing low-molecular-weight phytoplankton-derived compounds and are characterized by their high trophic versatility [179, 180]. In winter 2015/16 and 2016/17 Rhodospirillales, especially of the genus *Thalassospira*, drive elevated values of these strategy values. *Thalassospira* is known for its ability to degrade polycyclic aromatic hydrocarbons (PAHs) [181], and its occurrence in the upper water column could be associated with Major Baltic Inflow events [182] and subsequent winter mixing.

Cyanobacterial photosynthesis (variable 4 negative) reaches its highest strategy values in summer (Fig. 4.7). This time pattern supports our interpretation of this strategy as cyanobacterial photosynthesis, since the occurrence of massive summer blooms caused by Cyanobacteria in the Baltic Sea is well-documented [183]. Distinguishing between various cyanobacterial families reveals different time patterns within these groups: an

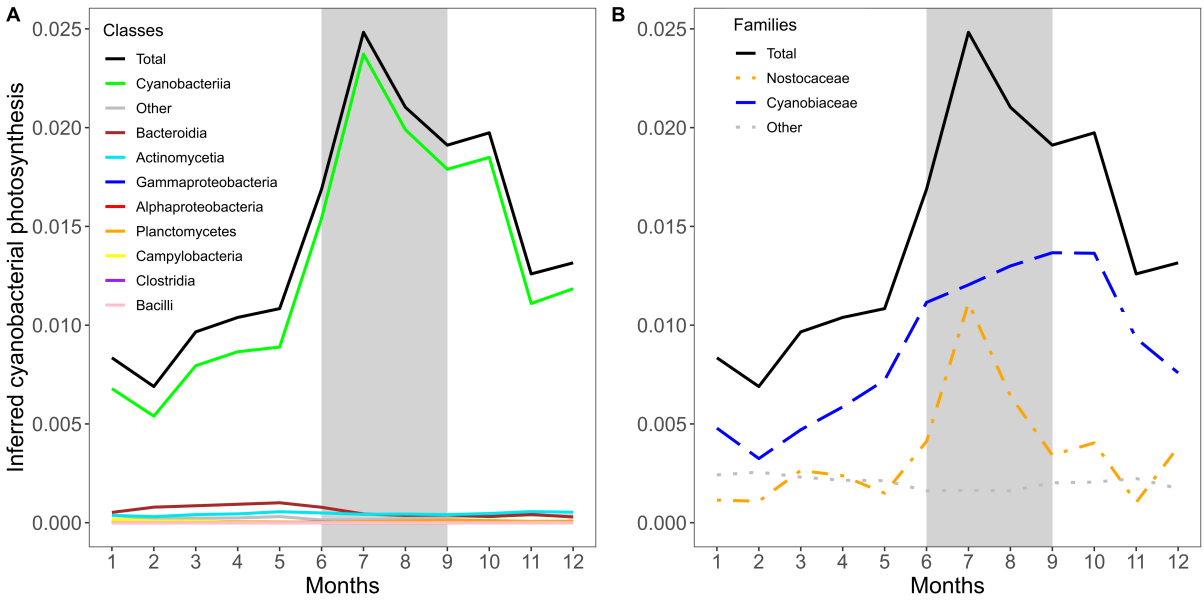

Figure 4.7: Strategy time series with different taxonomic resolutions: abundance-weighted mean values of variable 4 negative, i.e. inferred cyanobacterial photosynthesis, over the yearly cycle. Summer months are indicated by a gray background. Taxonomic class (A) and cyanobacterial taxonomic families (B) are color-coded. Figure adapted from Massing et al. [135] with modifications.

early summer peak attributed to the filamentous Nostocaceae and a sustained plateau of niche values for the unicellular Cyanobiaceae, gradually increasing until the beginning of autumn (Fig. 4.7B). The utilization of nutrients from filamentous Cyanobacteria could potentially fuel the metabolism of opportunistic picocyanobacteria [184].

There are also metabolic niches that reach their occupation minimum in May, right after the phytoplankton bloom, like the metabolic ability to oxidize methyl groups and C1 compounds (variable 38 negative), and the strategy of non-spore forming sulfate reducers (variable 38 negative) (Fig. 4.8). The former strategy is dominated by Alphaproteobacteria, especially *Pelagibacter* in winter and Planctomycetes in autumn in the Baltic Sea bacterial community. In the marine environment, diverse C1 and methylated compounds originate from dissolved organic carbon, where methanol constitutes a major fraction of oxygenated volatile organic chemicals, and formaldehyde is ubiquitous in seawater [170]. The capacity to utilize these compounds allows for energy production from relatively abundant substrates in the water, but this ability is outcompeted as concentrations of phytoplankton-derived substrates increase during spring.

Driven by Baltic Sea sulfate reducers of the phylum Desulfobacterota, predominantly found in sediments and oxygen-depleted waters [185, 186], variable 43 positive peaks in February (Fig. 4.8B). Strong winter mixing is probably the cause for the appearance of

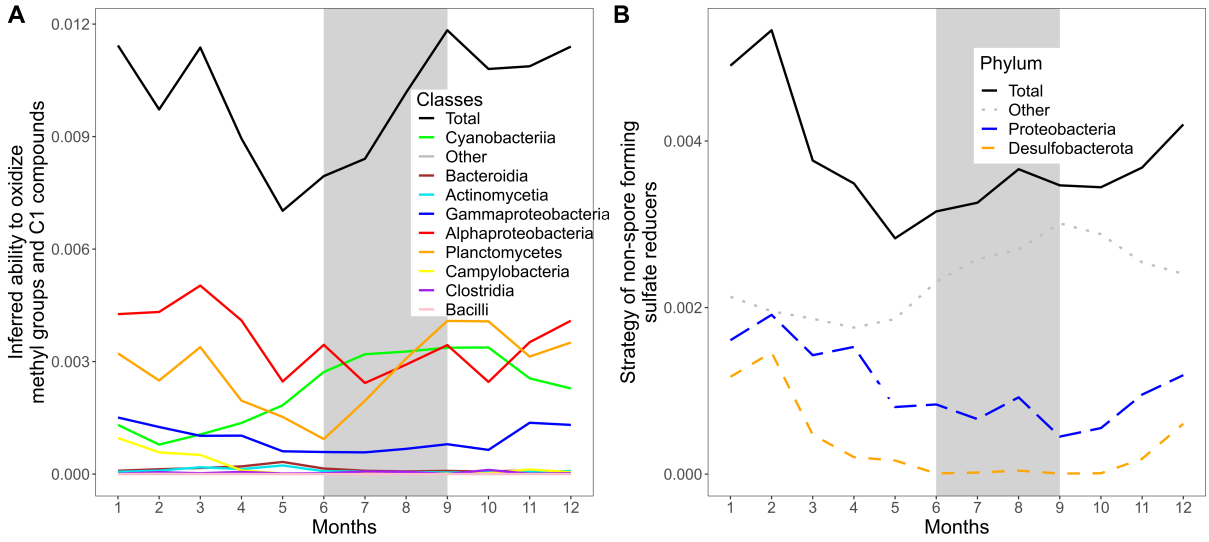

Figure 4.8: Abundance-weighted mean values of variable 38 negative, i.e. the inferred ability to oxidize methyl groups and C1 compounds (A) and of variable 43 positive, i.e. the strategy of non-spore forming sulfate reducers (B) over the yearly cycle. Taxonomic classes (A) and taxonomic phyla () are color-coded. Figure is adapted from Massing et al. [135] with modifications.

these sulfate reducers in the upper water column during this period. Overall, comparing the strategy time series with the environmental data obtained at LMO and examining the abundance-weighted mean values over the months, reveals a strong signal of seasonality. In Fig. 4.9A, the left-side variables correlate with higher nutrient concentrations and lower temperatures, indicative of winter conditions, while the right-side variables are associated with higher temperatures and chlorophyll a concentrations, reflecting summer conditions. Fig. 4.9B complements this observation, showing higher mean strategy values in summer and autumn for the variables on the left side, and higher values in winter and spring for the variables on the right side.

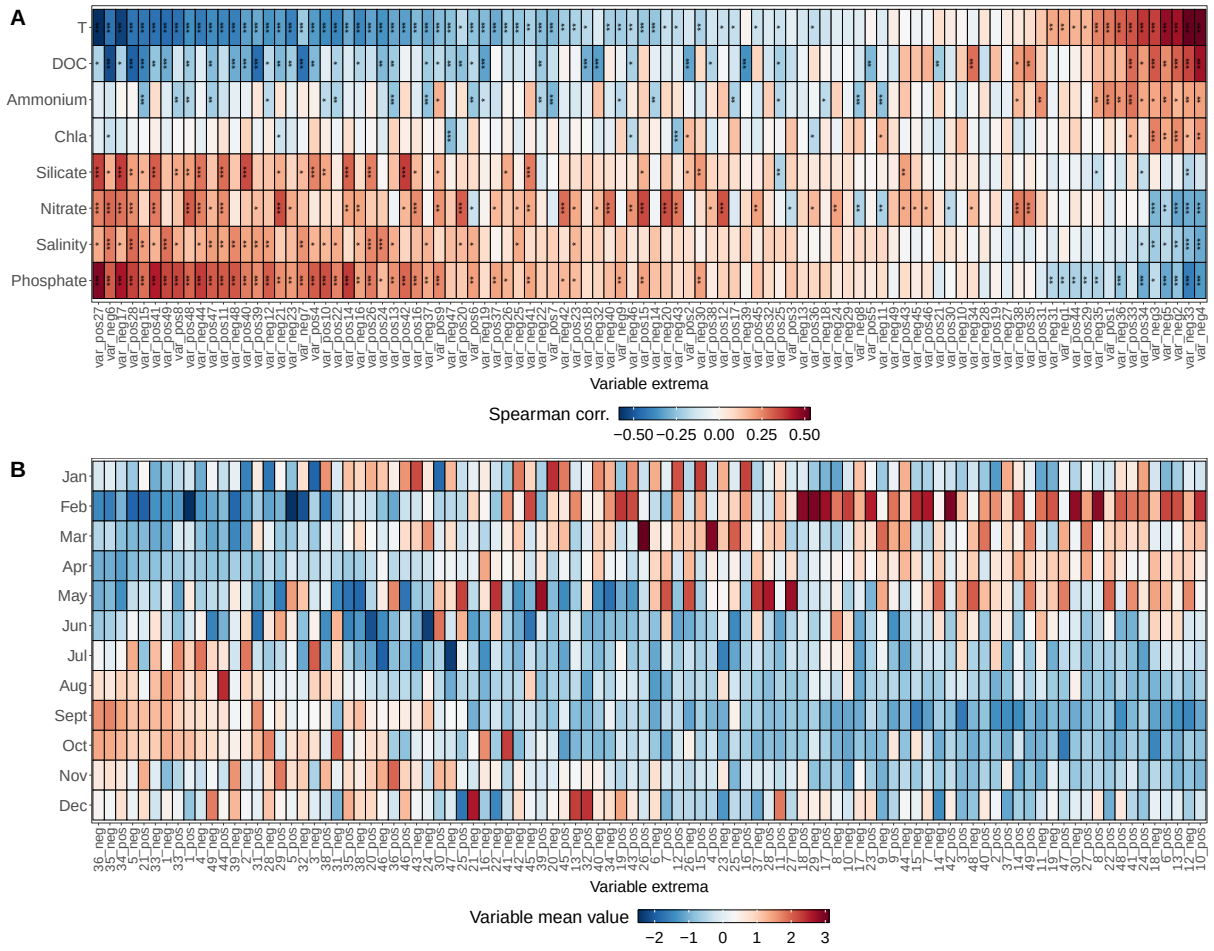

Figure 4.9: A: Heatmap of Spearman correlation coefficients (CV) between the first 49 variables, i.e. strategy time series and the environmental variables (Chla: chlorophyll a concentrations; DOC: dissolved organic carbon concentrations; T: temperature). All concentrations are in μM. Significance levels are expressed by asterisks (*** for p-value ≤ 0.001, ** for p-value ≤ 0.01, * for p-value ≤ 0.05). P-values are Benjamini-Hochberg-adjusted [156]. B: Heatmap of abundance-weighted variable mean values of the first 49 variables for each month over the whole sampling period, standardized to mean = 0 and standard deviation = 1 for each variable side. Figure adapted from Massing et al. [135].

4.4 Functional diversity

In Chapt. 2.3.3, we have seen that the diffusion distances in the reconstructed trait space can be used to robustly estimate functional diversities of communities [49]. This allows us to investigate changes in functional biodiversity from monitoring datasets. Adopting the approach of Ryabov et al. [49], we computed the diffusion distances between all pairs of species from the variables obtained via diffusion mapping. These distances were then used to quantify the functional diversity of each sample, calculated as Rao index [49]. Low functional diversity values imply that the sampled community is dominated by few

metabolic strategies, whereas higher values indicate that the community is more diverse in terms of metabolic strategies.

Summarizing all sampling years, the functional diversity mean values are highest in February and July and reach lowest values in May and October over the yearly cycle (Fig. 4.10). A likely explanation for the observed pattern is the absence of the thermocline in winter, that causes deeper mixing of the water layers [187], leading to bacterial communities from the former mesopelagic to be found also in surface waters [140, 188]. The presence of these communities in the surface waters during winter can lead to an increase in functional diversity, as their members may possess distinct strategies adapted to deeper water layer or sediment conditions [140]. Due to their taxonomic and functional diversity generally being higher than surface communities [82], their appearance in the surface layers contributes to the increased overall functional diversity during this period. Also, increased nutrient availability in winter leading to a higher resource heterogeneity benefits functional diversity [189, 190]. The decrease in functional diversity following the phytoplankton bloom is likely driven by the dominance of strategies that are associated to the utilization of phytoplankton-derived substrates during bloom phases as these compounds increase massively in relative abundance [167, 169, 191].

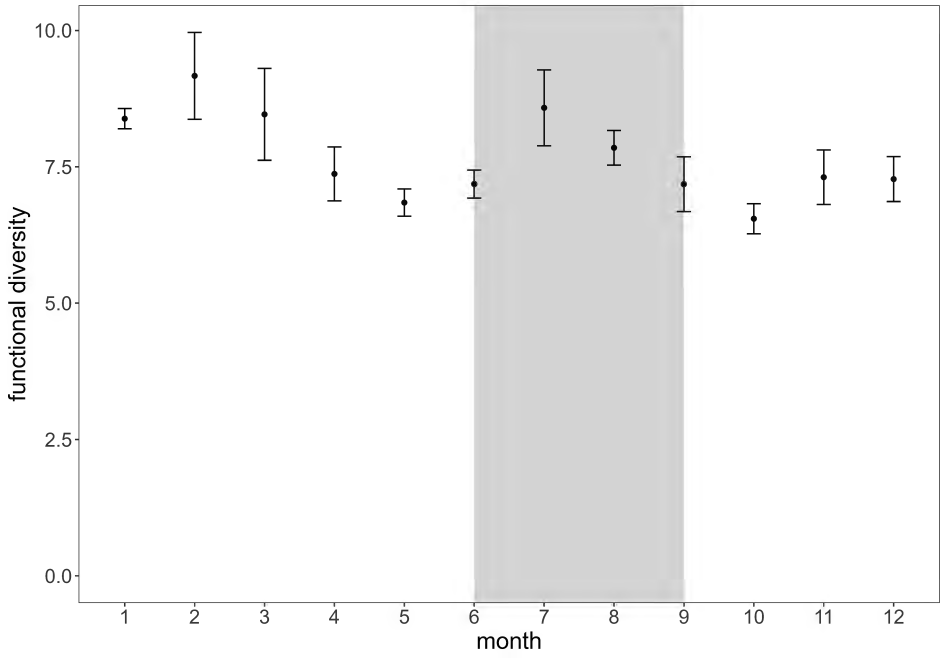

Figure 4.10: Mean and SE of functional diversity estimation calculated as Rao index [49] for the sampled Baltic Sea bacterial community over the yearly cycle summarizing the years 2011–2019. Gray background indicates summer months. Figure from Massing et al. [135].

4.5 Discussion

In this chapter, I showed that the diffusion map procedure is a powerful tool to coarse-grain complex bacterial communities in terms of their metabolic strategies over time. Diffusion mapping the bacterial capabilities uncovered a wide spectrum of interpretable metabolic strategies. These strategies ranged from localized ones, like cyanobacterial photosynthesis, to continua of strategies, such as the degree of association with marine hosts and the degree of trophic versatility. The analysis revealed strategies that align with phylogeny, strategies that differentiate closely-related taxa, and strategies that unite distantly-related taxa [44]. The observed similarities among distantly-related taxa in their metabolic strategies may be attributed to metabolic niche convergence [192] or horizontal gene transfer (e.g. [193]).

Systematizing the genomic information via our diffusion map approach enables us to translate the changes in bacterial species abundances to quantitative changes in potential occupation of metabolic niches over time. These abundance-weighted strategy values showcased a wide range of temporal patterns. These include seasonal dynamics, such as increasing trends in summer or increases following the phytoplankton bloom, as well as higher values related to winter mixing. Inter annual changes are also observed in some strategies. Moreover, specific events like a pronounced spring bloom or a Major Baltic Inflow event left discernible imprints in the strategy time series. Some functional strategies are strongly driven by a single bacterial group, whereas others were shared among multiple bacterial groups, with the distribution often varying depending on the season.

In the Baltic Sea, seasonality appears to play a significant role not only in shaping the phylogenetic composition of bacterial communities [140, 188] but also in influencing the occupation of bacterial metabolic niches and the functional diversity. This strong impact of seasonality on metabolic strategies is likely linked to the interplay between seasonal changes in substrate availabilities and fluctuations in abiotic parameters that influence metabolic activities [188, 194, 195].

The results also reveal the power of the diffusion map to objectively detect biases in the datasets, illustrated by the identification of the pathogen bias in our dataset. In the first variable, the diffusion map separated all the pathogenic members of the Enterobacteriaceae from all other taxa, revealing the bias that bacterial pathogenic taxa and genes involved in pathogenesis are overrepresented in global databases (e.g. [161]). This also highlights one of the major limitations of the approach: our knowledge of genes and their functions. As we acquire further knowledge, new strategies may be detected and the

metabolic strategy space might change. As a consequence, branches may be added to the tree-like structure, representing previously unknown taxa and strategies and positions of certain taxa and distances between taxa might change.

It is important to emphasize that genes merely represent the theoretical capabilities of a species [44], akin to the fundamental niche concept [196]. Nevertheless, recent findings reveal that functional genes can effectively predict a species' position along major niche gradients, surpassing predictions based solely on phylogenetic information [41]. Furthermore, Gralka et al. [197] have demonstrated that functional predictions derived from genomes exhibit remarkable accuracy in forecasting real metabolic niches. Gowda et al. [198] also demonstrated the feasibility of predicting community metabolic dynamics in the denitrification process based on the presence and absence of genes in metagenomes.

In contrast to their approach, our method involves the assignment of amplicon sequence variants (ASVs) to species and acquiring their complete genomes from available databases. Hence our method relies heavily on the quality of these databases and genome assemblages. The 95% similarity threshold we employed for matching ASVs to genomes provides a broad identification, and it is worth noting that substantial changes in genome content can occur at the sub-genus level [199]. As of now, to our knowledge, ASV data remains the only available data for our system, meeting the data requirements for our approach. Despite its relative simplicity, this crude tool has already provided evidence that our method yields a useful trait space.

In the future, diffusion map analysis of bacterial trait space stands to benefit from ongoing developments. One promising avenue is the integration of metagenome analysis, which can enrich the analysis by considering within-species genome variability, the accessory genome, and prophages' potential role in shaping organismal strategies and traits [200, 201, 202]. Furthermore, leveraging tools like PICRUSt 2 [203] in conjunction with diffusion mapping may enhance the prediction of metagenome functions.

In an ideal scenario, having access to the complete genomes of all taxa present in the sampled habitat would be highly beneficial. However, since this is not yet the case, we currently depend on a straightforward mapping scheme. Deep shotgun metagenomic sequencing and long read technologies will be crucial in gathering the data required to make our method even more powerful. Looking ahead, we also envision the potential of utilizing future transcriptomic data to investigate the strategies employed by organisms under specific environmental conditions [44].

In conclusion, the diffusion map approach outlined in this chapter allows us to coarse grain complex bacterial communities in terms of metabolic bacterial strategies and pro-

vides a quantitative framework to organize genomic information into potentially occupied metabolic niches over time. This approach advances our understanding of the consequences of changes in bacterial abundances and species composition for the dynamics of the potential occupation of metabolic niches. The results demonstrate the power of manifold learning approaches to shed light on the relationships between bacterial community composition and ecosystem functioning, facilitating analysis, monitoring, and future predictions.

Chapter 5

Generalized model of bacterial mutualism

In the previous chapters, we have developed tools to make sense of the large datasets that are available on bacterial communities. To gain insights into the underlying mechanisms of bacterial community dynamics and functioning we also need to be able to develop good models of this system and its phenomena and dynamics. The study of these complex communities requires an integrated approach that combines data analysis and modeling. This entails a reciprocal flow of information, where results from experimentation and data analysis inform the model, and conversely, insights derived from modeling guide experimentation and hypothesis generation.

As discussed above, the main challenge in modeling bacterial communities is their complexity and the many uncertainties associated with them. We have seen in Chapt. 3 that generalized modeling provides a framework to model systems with many uncertainties. In this chapter I apply this modeling approach to study a small common motif in bacterial communities, bacterial cross-feeding, i.e. two types of bacterial taxa that reciprocally produce and exchange limiting metabolites needed for growth [37, 204]. This exchange of metabolites is a crucial mechanism that impacts the growth and composition of bacterial communities [205]. Given that this mutualistic relationship forms a positive feedback loop, it should inherently be unstable. Therefore, here I aim to explore the factors that contribute to the stabilization of this mutualistic relationship.

I start in Sec. 5.1 with an introduction to metabolite-based bacterial interactions. Then I formulate the generalized model (Sec. 5.2), which I analyze in the subsequent sections. I investigate parameter correlations to stability (Sec. 5.3), response to parameter changes (Sec. 5.4) as well as bifurcations (Sec. 5.5). For an easier analysis, I reduce the 4-dimensional system to a 2-dimensional system in Sec. 5.6. Moreover, I study the 4-dimensional as well as the 2-dimensional system in space (Sec. 5.7). Finally, I discuss the findings in Sec. 5.8.

5.1 Metabolite-based bacterial interactions

Owing to their broad diversity of metabolic capabilities, bacteria release a variety of metabolites, that are frequently taken up and utilized by fellow bacteria. The prevalence of these metabolite-based interactions in bacteria plays a role in community composition, ecosystem function and stability [206, 207]. Outsourcing the production of essential metabolites such as amino acids, lipids, vitamins, cofactors and signaling molecules [37, 208, 209, 210], bacteria create metabolite trading networks that impact the dynamics of bacterial systems [54, 55, 211]. Around 98% of the bacterial genomes that have been sequenced apparently lack the capability to independently synthesize all essential amino acids, indicating their dependence on the production of metabolites by other bacteria [205]. This suggests that commensal or mutualistic interactions among microorganisms based on sharing metabolites, are likely widespread in ecosystems and macrobiotic hosts [204, 210, 212].

Since mutualistic interactions create positive feedback loops, i.e. each species benefits from the respective other, they are generally predicted to be destabilizing for ecological communities comprising interacting macro- or bacteria populations [50, 70], unless specific criteria are satisfied [54, 55]. These theoretical predictions appear to be in contrast to real-world observations of apparently stable environmental and host-associated bacterial communities, which include both opportunistic and obligate mutualists [37, 71]. To address this apparent paradox, we need to identify ecological factors that generally promote stability and coexistence within communities featuring beneficial bacterial interactions [54, 213].

Metabolite-based interactions present across a range of spatial scales [210]. From interactions between neighboring cells, for example between diverse phenotypes within bacterial biofilms, through cell-to-cell contact or along nanotubes [37, 212] to instances in which metabolites are released or leak into the surroundings and disperse as public goods that might eventually be taken up by microorganisms in distant areas [210, 214, 215]. The metabolites that are exchanged range from costly to produce and export [37] to waste or byproducts of the own metabolism [216].

Given that the majority of identified bacteria remain uncultured [58, 59], our knowledge on the precise rate laws and functional forms of these interactions is often very limited. This high uncertainty coupled with the diversity of interaction modes impedes our ability to predict dynamics of interacting bacterial populations across various scales [215, 217]. To circumvent these issues, in the next section I present a generalized model of bacterial cross-feeders. Without making specific assumptions about the functional forms or rate laws

underlying the bacterial metabolite-based interactions, the generalized model describes the stability properties of all stationary states for the entire class of systems.

5.2 Model formulation

To address the challenges of the diversity of interaction modes and the high uncertainty in rate laws and functional forms governing bacterial interactions, we apply a generalized modeling approach. We model the small bacterial motif of bacterial cross-feeding, i.e. two types of bacterial taxa, X and Y, that reciprocally produce and exchange essential growth-limiting metabolites (either A or B) [37, 204]. In this model, it is assumed that the additional metabolites needed for growth are sufficiently available.

The cross-feeding relationship is illustrated in Fig. 5.1 and represented by the 4-dimensional. system of equations

$$\dot{X} = M_x[G_x(X,B)] - L_x(X) \tag{5.1}$$
$$\dot{Y} = M_y[G_y(Y,A)] - L_y(Y) \tag{5.2}$$
$$\dot{A} = S_a + G_x(X,B) - D_a(A) - G_y(Y,A) \tag{5.3}$$
$$\dot{B} = S_b + G_y(Y,A) - D_b(B) - G_x(X,B). \tag{5.4}$$

The functions M_x and M_y represent the biomass growth of bacteria X and Y with respect to their metabolic throughput, G_x and G_y. Hence G_x and G_y model the metabolic throughput of bacterial cells, namely the transformation of metabolite type A to type B and vice versa by bacteria through uptake, metabolism, and release. The functions L_x and L_y model the losses of the bacteria to mortality. In the equations that describe the change of metabolites A and B, the constants S_a and S_b represent a constant exogenous supply of metabolites to the system. The functions D_a and D_b represent the losses of the metabolites to decay. Unlike phenomenological models of interspecific bacterial interactions, we consider the class of models with explicit metabolite dynamics, which can lead to qualitatively different behaviors compared to phenomenological models [54, 55, 218].

In conventional modeling, we would now proceed by assigning specific functional forms to each process, finding the steady states of the system and computing the Jacobian matrix, which provides a linearization of the dynamics around the steady state. As we have seen in Chapt. 3 generalized modeling (GM) however draws on the insight that computing steady states is technically difficult, whereas determining stability around the steady state is in comparison easier. For determining the stability around a given steady state, we only need to parameterize the Jacobian, which requires less information than restricting the

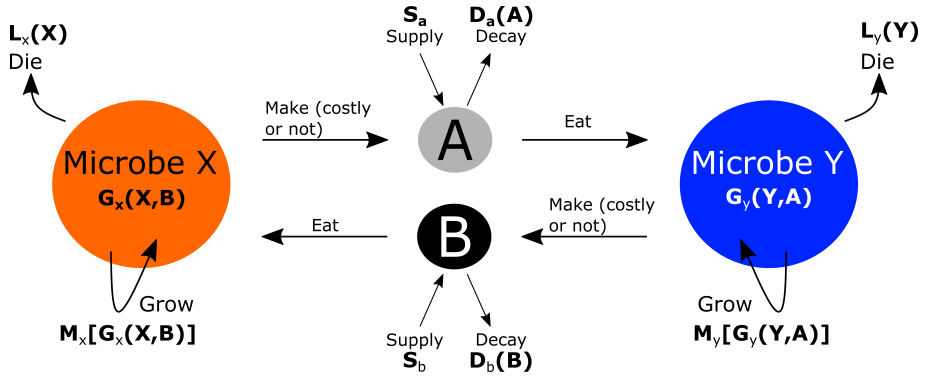

Figure 5.1: Community motif. Cross-feeding of two bacterial taxa, X and Y, on two types of metabolites, A and B. Gain and loss processes of bacteria and metabolites with corresponding unspecified functions and descriptions of the process they represent.

functions to specific functional forms.

Following the procedure explained in Chapt. 3.3, to parameterize the Jacobian in an interpretable way, we assume that all variables and process rates have positive values and the class of models we consider is general enough that positive steady states must exist. We denote these steady states as X^* and the rates of the processes in the steady state as $P^* = P(X^*)$ and normalize the equations with respect to X^*, i.e. for every variable X we define

$$x = \frac{X}{X^*} \tag{5.5}$$

and for every process $P(X)$ we define

$$p(x) = \frac{P(xX^*)}{P^*}. \tag{5.6}$$

This leads to the normalized differential equations

$$\dot{x} = \frac{M_x^*}{X^*} m_x[g_x(x,b)] - \frac{L_x^*}{X^*} l_x(x) \tag{5.7}$$

$$\dot{y} = \frac{M_y^*}{Y^*} m_y[g_y(y,a)] - \frac{L_y^*}{Y^*} l_y(y) \tag{5.8}$$

$$\dot{a} = \frac{S_a^*}{A^*} + \frac{G_x^*}{A^*} g_x(x,b) - \frac{D_a^*}{A^*} d_a(a) - \frac{G_y^*}{A^*} g_y(y,a) \tag{5.9}$$

$$\dot{b} = \frac{S_b^*}{B^*} + \frac{G_y^*}{B^*} g_y(y,a) - \frac{D_b^*}{B^*} d_b(b) - \frac{G_x^*}{B^*} g_x(x,b), \tag{5.10}$$

where normalized functions are indicated by lowercase letters. This normalization procedure maps the unknown steady state to a known location, i.e. $x^* = y^* = a^* = b^* = 1$, and in the steady state all processes run at rate 1. Because gains and losses are balanced in

the steady state, we can define scale parameters representing metabolites and bacterial turnover rates

$$\alpha_x := \frac{M_x^*}{X^*} = \frac{L^*}{X^*} \tag{5.11}$$

$$\alpha_y := \frac{M_y^*}{Y^*} = \frac{L_y^*}{Y^*} \tag{5.12}$$

$$\alpha_a := \frac{S_a^*}{A^*} + \frac{G_x^*}{A^*} = \frac{D_a^*}{A^*} + \frac{G_y^*}{A^*} \tag{5.13}$$

$$\alpha_b := \frac{S_b^*}{B^*} + \frac{G_y^*}{B^*} = \frac{D_b^*}{B^*} + \frac{G_x^*}{B^*}. \tag{5.14}$$

To quantify the relative contribution of each gain and loss process to the population turnovers we define branching parameters [68, 93]

$$\beta_a := \frac{1}{\alpha_a}\frac{D_a^*}{A^*} \tag{5.15}$$

$$\beta_b := \frac{1}{\alpha_b}\frac{D_b^*}{B^*} \tag{5.16}$$

$$\gamma_a := \frac{1}{\alpha_a}\frac{S_a^*}{A^*} \tag{5.17}$$

$$\gamma_b := \frac{1}{\alpha_b}\frac{S_b^*}{B^*}, \tag{5.18}$$

with $\tilde{\beta_a} = 1 - \beta_a$, $\tilde{\beta_b} = 1 - \beta_b$, $\tilde{\gamma_a} = 1 - \gamma_a$ and $\tilde{\gamma_b} = 1 - \gamma_b$. For instance, when $\tilde{\gamma_a} = 0$, it indicates that all of metabolite A originates from an external source, implying zero bacteria production of metabolite A. Conversely, if $\tilde{\gamma_a} = 1$ there is no external supply of metabolite A to the system, instead it is entirely produced by the bacteria. When $\tilde{\beta_a} = 0$, metabolite A solely undergoes decay, without any uptake by bacteria. Conversely, if metabolite A is exclusively removed from the environment through bacteria uptake, meaning there is no natural decay, then $\tilde{\beta_a} = 1$.

Descriptions and ranges for scale and branching parameters are provided in Table 5.1 and discussed by Gross et al. [68] and Yeakel et al. [95]. Substituting scale and branching parameters into Eqs. 5.7-5.10 yields

$$\dot{x} = \alpha_x[m_x(g_x(x,b)) - l_x(x)] \tag{5.19}$$

$$\dot{y} = \alpha_y[m_y(g_y(y,a)) - l_y(y)] \tag{5.20}$$

$$\dot{a} = \alpha_a[\gamma_a + \tilde{\gamma_a} g_x(x,b) - \beta_a d_a(a) - \tilde{\beta_a} g_y(y,a)] \tag{5.21}$$

$$\dot{b} = \alpha_b[\gamma_b + \tilde{\gamma_b} g_y(y,a) - \beta_b d_b(b) - \tilde{\beta_b} g_x(x,b)]. \tag{5.22}$$

Local stability of this system can be analyzed by calculating the corresponding Jacobian matrix $\mathbf{P}$, which determines the behavior of the system close to the steady state. It is defined by $P_{i,j} = \partial n_i / \partial n_j|_*$, $n \in \{x, y, a, b\}$. The Jacobian contains additional parameters called elasticities or exponent parameters [68, 93], where for instance

$$l'_{x,x} = \left. \frac{\partial \log L_x(X)}{\partial \log X} \right|_* \tag{5.23}$$

are logarithmic derivatives, and $|_*$ indicates that the derivatives are evaluated at the steady state. The double subscript "x, x" indicates that we are referring to bacteria denoted as X and that the function is being differentiated with respect to x. Hence, the parameter $l'_{x,x}$ is an indication of the sensitivity of bacteria X mortality to density of X. In general, these parameters are a measure of nonlinearity of the process at the steady state. Exponent parameters are defined in Table 5.1. We obtain the Jacobian

$$\mathbf{P} = \begin{pmatrix} \alpha_x & 0 & 0 & 0 \\ 0 & \alpha_y & 0 & 0 \\ 0 & 0 & \alpha_a & 0 \\ 0 & 0 & 0 & \alpha_b \end{pmatrix} \begin{pmatrix} m'_x g'_{x,x} - l'_{x,x} & 0 & 0 & m'_x g'_{x,b} \\ 0 & m'_y g'_{y,y} - l'_{y,y} & m'_y g'_{y,a} & 0 \\ \tilde{\gamma}_a g'_{x,x} & -\tilde{\beta}_a g'_{y,y} & -\beta_a d'_{a,a} - \tilde{\beta}_a g'_{y,a} & \tilde{\gamma}_a g'_{x,b} \\ -\tilde{\beta}_b g'_{x,x} & \tilde{\gamma}_b g'_{y,y} & \tilde{\gamma}_b g'_{y,a} & -\beta_b d'_{b,b} - \tilde{\beta}_b g'_{x,b} \end{pmatrix} \tag{5.24}$$

A steady state is stable if all eigenvalues of the system's Jacobian have negative real parts, and a loss of stability occurs if the real parts of one or more eigenvalues become positive.

Table 5.1: Parameter descriptions and ranges used for numerical analyses.

Parameter	Description	Range or Value
Scale parameters		
α_x	Turnover rate of bacterial species X and Y	0.1
Branching parameters		
β_a	Fraction of losses of metabolite A from natural decay	[0, 1]
$\tilde{\beta}_a$	Fraction of losses of metabolite A from uptake by Y	$1-\beta_a$
β_b	Fraction of losses of metabolite B from natural decay	[0, 1]
$\tilde{\beta}_b$	Fraction of losses of metabolite B from uptake by X	$1-\beta b$
$\tilde{\gamma}_a$	Fraction of gains of metabolite A from production by X	[0, 1]
$\tilde{\gamma}_b$	Fraction of gains of metabolite B from production by Y	[0, 1]
Elasticities		
$g'_{x,x}$	Sensitivity of metabolite A production to density of bacteria X	1
$g'_{x,b}$	Sensitivity of metabolite A production to concentration of B	[0.5, 1.5]
$g'_{y,y}$	Sensitivity of metabolite B production to density of bacteria Y	1
$g'_{y,a}$	Sensitivity of metabolite B production to concentration of A	[0.5, 1.5]
$l'_{x,x}$	Sensitivity of bacteria X mortality to density of X	[1, 2]
$l'_{y,y}$	Sensitivity of bacteria Y mortality to density of Y	[1, 2]
$d'_{a,a}$	Sensitivity of metabolite A decay to density of A	1
$d'_{b,b}$	Sensitivity of metabolite B decay to density of B	1
m'_x	Sensitivity of bacteria X biomass growth to metabolic throughput of bacteria X	[0.5, 1.5]
m'_y	Sensitivity of bacteria Y biomass growth to metabolic throughput of bacteria Y	[0.5, 1.5]

5.3 Parameter correlation with stability

Due to the relatively large model size and the many forms of cross-feeding relationships between bacteria, we do not want to restrict the model to few fixed parameter values, but instead we adopt an ensemble approach [100], in which various model realizations are explored. The generalized modeling procedure ensures the feasibility of all these states, meaning that for each state, we can construct a realistic cross-feeding model in which the bacterial populations and the metabolites' concentrations have positive values and are stationary.

To gain initial insights into the influence of the various model parameters on the stability of the cross-feeding motif, we perform a numerical stability analysis similar to the procedure explained in detail in Sec. 3.4. In short, we constrain the 17 parameters to biologically plausible ranges in the context of the model (Table 5.1), draw 10^7 random parameters sets and calculate the leading eigenvalue, i.e. the stability for each of the parameter sets.

We denote the m'th realization of parameter i as p_i^m where $i \in [1, 17]$ and $m \in [1, 10^7]$. And we denote the stability of the steady state by the parameter set m as s_m, where we define

$$s_m = \begin{cases} 1 & \mathrm{Re}(\lambda_0) < 0 \\ 0 & \text{otherwise} \end{cases}, \tag{5.25}$$

hence s_m is 1 if the parameter set m is stable and zero otherwise. We then employ a generalized linear model with binomial error and a logit link function to relate the system's stability s_m to the model parameters. For this procedure we standardize each parameter that was varied to a mean of zero and a standard deviation of one. The coefficients of the model estimate the effect of each parameter on the stability of the system.

We find that large fractions of metabolite uptake by bacteria ($\tilde{\beta}_m$) are strongly correlated to stability, whereas large fractions of metabolite production by bacteria ($\tilde{\gamma}_m$) are strongly correlated to instability (Fig. 5.2A). In Fig. 5.2B, we can also see that increasing the fraction of metabolite uptake by bacteria ($\tilde{\beta}_a$) and decreasing the fraction of metabolite production by bacteria ($\tilde{\gamma}_a$) result in the highest proportions of stable systems. The ratio of metabolite uptake to metabolite production of bacteria seems to be impacting the stability of the system and we will explore this phenomenon in the following sections in more detail.

We also find that the sensitivities of the bacteria's mortality to their own population size ($l'_{b,b}$) are positively correlated with stability (Fig. 5.2). The latter corresponds to our expectation that nonlinear, for example quadratic mortalities, promote stability [104]. Conversely, high sensitivities of bacteria's growth to metabolic throughput (m'_b) and high sensitivities of bacteria's metabolite production to the concentration of the consumed trade metabolites ($g'_{b,m}$) are negatively correlated to stability. These two observations correspond to positive feedback loops: When there is a high sensitivity of bacteria's growth to metabolic throughput, this means that bacteria's growth is profiting from the production of the trade metabolite. The bacterial population will grow more with an increasing production of the trade metabolite, which will in turn fuel the growth of the trade partner. High sensitivities of metabolite production to the concentration of the consumed trade metabolite imply that one metabolite promotes the production of the other metabolite hence also resulting in a reinforcing loop.

5.4 Response to parameter change

In Sec. 3.4, we showed how to analyze the response of the system to a permanent change of parameters, i.e. a press perturbation. We apply this approach here to analyze the

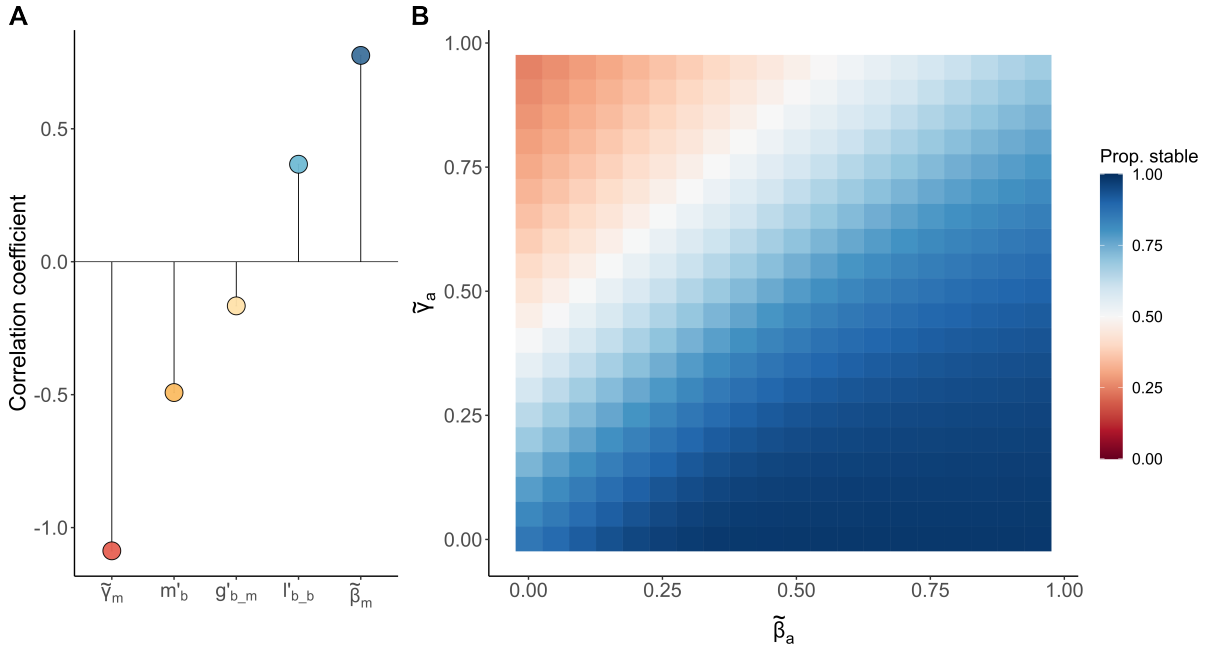

Figure 5.2: A: Correlations between model parameters and community stability, calculated as coefficients in a binomial GLM (generalized linear model) from 10^7 parameter sets (m= metabolite, b = bacteria). B: Proportion of stable systems as $\tilde{\beta}_a$ (fraction of metabolite uptake by bacteria A) and $\tilde{\gamma}_a$ (fraction of metabolite production by bacteria A) are varied, calculated from $4{\times}10^7$ parameter sets. All other parameters are randomly drawn from uniform distributions (Table 5.1).

4-dimensional system, writing

$$\boldsymbol{\delta} = -\mathbf{P}^{-1}\boldsymbol{k}, \tag{5.26}$$

where $\boldsymbol{\delta}$ is the shift in the steady state, $\mathbf{P}^{-1}$ is the inverse of the Jacobian matrix, and $\boldsymbol{k}$ is a vector describing the direct impacts of the perturbations on the individual equations [69]. The direct effect of decreasing the bacteria X population by 10% can for example be written as

$$\boldsymbol{k} = \begin{pmatrix} -0.1 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{5.27}$$

The impact vector, denoted as $\boldsymbol{\delta}$, and obtained by substituting the corresponding $\boldsymbol{k}$ into equation 5.26, reveals the indirect relative changes in bacteria or metabolite levels in the steady state once the perturbation has propagated within the motif. These changes are normalized with respect to the unperturbed steady state and are measured per unit of direct effect. These values should be interpreted as proxy values, where negative numbers represent losses, positive numbers represent gains and larger absolute values indicate more

substantial impacts. Hence the values provide insights into both the direction and the relative magnitude of the perturbation effects.

Specifically, we use the 10^7 parameter sets from Sec. 5.3 and eliminate the unstable states, resulting in a reduced number of parameter sets of $\sim 7\times10^6$. For the remaining parameter sets we then calculate the impact vector $\boldsymbol{\delta}$, using Eq. 5.26 for different perturbation scenarios (Fig. 5.3). Across the different impact scenarios, we observe a consistent pattern: the impact is always highest for at least one of the bacterial species. Notably, altering the density of bacteria has a much stronger effect (up to 7.5x) on the system in comparison to changing metabolite concentrations. A shift in metabolite levels primarily affects the bacterial species that consumes those metabolites, whereas a change in bacterial population density primarily influences the concerned bacterial population the most. This latter change has a comparatively smaller, opposing effect on the metabolites consumed by the affected bacterial species. Overall, we see that changing bacterial densities has a higher leverage on the mutualistic cross-feeding system compared to changing metabolite concentrations.

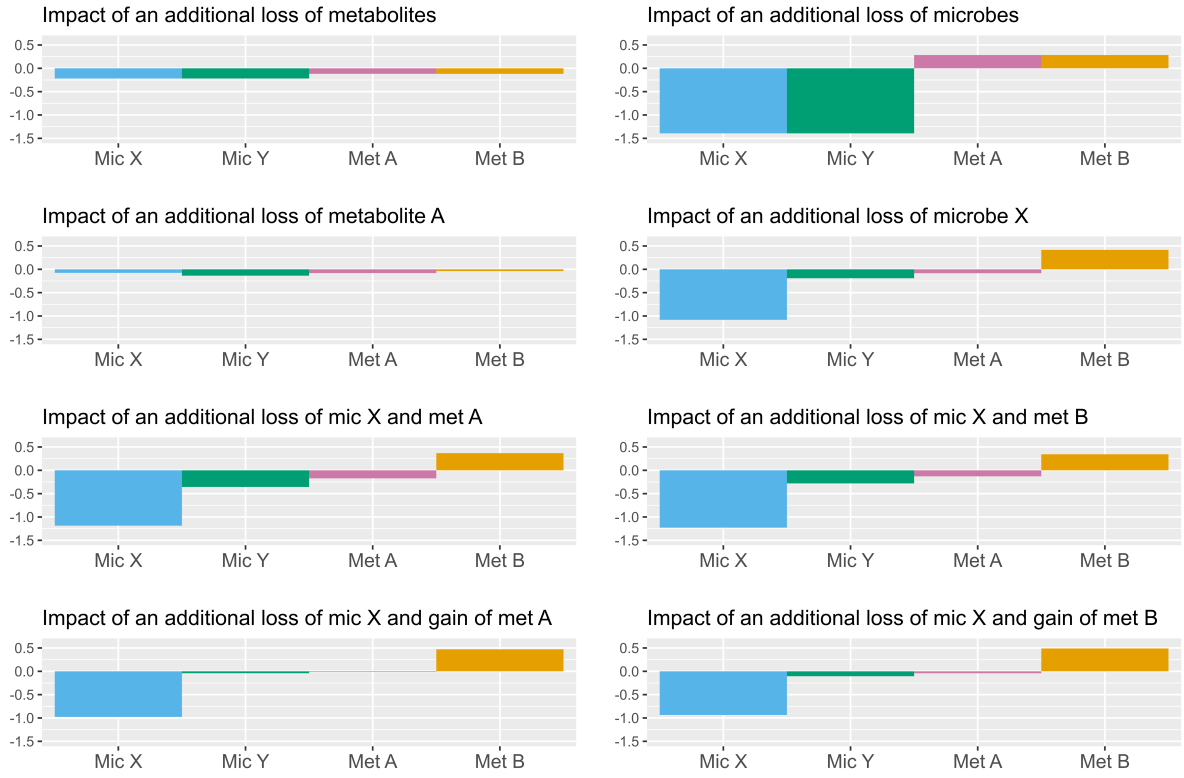

Figure 5.3: Impact of press perturbations, i.e. additional gain (+10%) and loss (-10%) terms for bacteria's and metabolites' density on the cross-feeding motif. The y-axis is the negative or positive impact of the change to the system. Each bar represents the median impact of each bacterium (mic) and metabolite (met) type. Calculated from $\sim 7\times10^6$ parameter sets.

5.5 Bifurcation analysis

Studying the transitions between dynamical regimes in the bacteria-metabolite system reveals the position of bifurcations in the parameter space, i.e. parameter sets at which the dynamical behavior of the system changes qualitatively. For the bacteria-metabolite system we located the saddle-node-type bifurcations and visualized them in bifurcation diagrams (Fig. 5.4). Each point in the plane/volume corresponds to a particular steady state. If steady states are located in the same plane/volume, they show qualitatively similar local dynamical properties. Steady states are stable in the lower part/volume of parameter space. The line/surface mark bifurcation points, where qualitative transitions take place [93]. Here, we focus on saddle-node type bifurcations, since they are especially a concern in mutualistic relationships. For example, the runoff of populations to infinity as well as changes that disrupt the symmetry within the mutualistic relationship may be caused by a saddle-node type bifurcation [219, 220]. In Sec. 2.3.1 we will also explore Hopf bifurcations.

As we have seen in Sec. 3.4, to locate the saddle-node type bifurcations, we have to find the combination of generalized parameters that results in a zero eigenvalue of the Jacobian. Since the determinant of a matrix is the product of its eigenvalues, the Jacobian has a zero eigenvalue only if its determinant is zero. The determinant of the Jacobian therefore serves as a test function for this bifurcation [75]. Solving $\det(\mathbf{P}) = 0$ for the different variables, we derive the following test functions:

$$g'_{x,b} = \frac{(\beta_b(l'_{x,x} - g'_{x,x}m'_x)(-g'_{y,a}l'_{y,y} + \beta_a((-1+g'_{y,a})l'_{y,y} + g'_{y,y}m'_y)))}{(l'_{x,x}(g'_{y,a}(-\beta_b + \gamma_a - (-1+\gamma_a)\gamma_b)l'_{y,y} + \beta_a(-1+\beta_b)((-1+g'_{y,a})l'_{y,y} + g'_{y,y}m'_y)))} \tag{5.28}$$

$$\begin{aligned} m'_x = {} & \frac{(l'_{x,x}(g'_{y,a}(\beta_b(-1+g'_{x,b}) + g'_{x,b}(-\gamma_a + (-1+\gamma_a)\gamma_b))l'_{y,y}}{(\beta_b g'_{x,x}(-g'_{y,a}l'_{y,y} + \beta_a((-1+g'_{y,a})l'_{y,y} + g'_{y,y}m'_y)))} - \\ & \frac{\beta_a(\beta_b(-1+g'_{x,b}) - g'_{x,b})((-1+g'_{y,a})l'_{y,y} + g'_{y,y}m'_y)))}{(\beta_b g'_{x,x}(-g'_{y,a}l'_{y,y} + \beta_a((-1+g'_{y,a})l'_{y,y} + g'_{y,y}m'_y)))} \end{aligned} \tag{5.29}$$

$$\beta_a = \frac{(g'_{y,a}l'_{y,y}(g'_{x,b}(\gamma_a(-1+\gamma_b) - \gamma_b)l'_{x,x} + \beta_b((-1+g'_{x,b})l'_{x,x} + g'_{x,x}m'_x)))}{((-g'_{x,b}l'_{x,x} + \beta_b((-1+g'_{x,b})l'_{x,x} + g'_{x,x}m'_x))((-1+g'_{y,a})l'_{y,y} + g'_{y,y}m'_y))} \tag{5.30}$$

$$\begin{aligned} \gamma_a = {} & \frac{(g'_{y,a}l'_{y,y}(g'_{x,b}\gamma_b l'_{x,x} + \beta_b(-((-1+g'_{x,b})l'_{x,x}) - g'_{x,x}m'_x))}{(g'_{x,b}g'_{y,a}(-1+\gamma_b)l'_{x,x}l'_{y,y})} + \\ & \frac{\beta_a(-g'_{x,b}l'_{x,x} + \beta_b((-1+g'_{x,b})l'_{x,x} + g'_{x,x}m'_x))((-1+g'_{y,a})l'_{y,y} + g'_{y,y}m'_y))}{(g'_{x,b}g'_{y,a}(-1+\gamma_b)l'_{x,x}l'_{y,y})} \end{aligned} \tag{5.31}$$

Let us first focus on the parameters m'_x and m'_y, that describe the sensitivity of the biomass growth of the two bacteria to their metabolic throughput. In the context of the

model, low values (< 1) of these parameters imply a costly production, i.e. the bacteria grow less than linearly in relation to their metabolic throughput. The production of the amino acid methionine for the trade partner is an example of a costly production, where energy is consumed during the process that cannot be invested into the own growth anymore [221]. High values (> 1) refer to a benefit of the metabolic throughput for the biomass growth of the bacteria. For instance, consider a scenario in which bacteria engage in the breakdown of organic matter, converting it into a rich energy source for their growth and releasing waste products. In this case, the more metabolic throughput they engage in (i.e. breaking down more organic matter), the faster they grow such that the growth can increase superlinearly in relation to the metabolic throughput.

We saw before, in Sec. 5.3, that the parameters m'_x and m'_y are negatively correlated to stability. Let us know examine the corresponding bifurcation diagram (Fig. 5.4A). In the bifurcation diagram we have assumed balanced metabolite supply and decay ($\beta_a = \beta_b = \gamma_a = \gamma_b = 0.5$), nonlinear mortalities ($l'_{x,x} = l'_{y,y} = 1.2$) and a linear relationship between metabolite concentrations and production of the trade metabolites ($g'_{x,b} = g'_{y,a} = 1.0$). The positive steady state is stable in the lower space of the parameter space (S) and unstable in the upper space (U). If m'_x and m'_y are increased, destabilization occurs in a saddle-node type bifurcation. For the steady state to be stable, if m'_y is increased, m'_x has to take lower values. In the context of the model, these results indicate that costly synthesis of metabolites is a stabilizing factor for the mutualistic cross-feeding relationship. Moreover, costly metabolite production by one of the bacterial species can offset the destabilizing effect of cheap synthesis by the trade partner. Costly metabolite synthesis serves as a stabilizing factor by attenuating the positive feedback between the two bacteria. Conversely, beneficial production amplifies the positive feedback, as an increased bacterial population also results in more substantial metabolite production.

The corresponding eigenvectors of the eigenvalues turning positive when crossing the bifurcation provide information about the relative changes in different variables of the system, i.e. on the direction of the perturbation that causes the system to move away from the equilibrium. For the parameter sets at the bifurcation in the m'_x and m'_y diagram, the eigenvectors corresponding to the largest eigenvalue indicate that the two bacteria and one of the metabolites change in the same direction, whereas the other metabolite changes in the opposite direction. The change of the bacterial population with the cheaper metabolite synthesis is more pronounced compared to the change of the trade partner's population. These results indicate that one bacterial species takes control over the system as its abundance increases in comparison to the trade partner. Let us now consider the influence of varying a third parameter, $g'_{x,b}$, the sensitivity of metabolite A

production to the concentration of metabolite B. This parameter describes the role of the consumed metabolite concentrations for the production of the trade metabolite. If the consumed metabolite is only required in small amounts (e.g. trace elements), increasing the consumed metabolite concentration will not result in an increase in the production of the trade metabolite once a concentration threshold is reached, hence $g'_{x,b} < 1$. The same holds if the uptake machinery or synthesis machinery of the bacteria is nearing saturation. Conversely, if the consumed metabolite enhances the production of the trade metabolite, $g'_{x,b}$ may be higher than 1. For example increasing concentrations of lactose can lead to an increase in the expression of enzymes for the degradation of lactose in *E. coli*, resulting in an increase of lactose metabolism and its products [222].

The bifurcation diagram including the parameter $g'_{x,b}$ (Fig. 5.4B) shows that if bacteria Y is performing costly trade metabolite synthesis, then high values of $g'_{x,b}$ are stabilizing. Conversely, when the synthesis is cheap, low values of $g'_{x,b}$ are stabilizing, counteracting the effect of m'_y in each case. Costly trade metabolite synthesis by bacteria Y is regulating the feedback loop, so in this case a superlinear sensitivity of the metabolite production on the concentration of B allows for higher m'_x since it strengthens the regulatory aspect. In contrast, if trade metabolite synthesis by bacteria Y is cheap, then the effect of the metabolite on the trade partner X can be kept small if $g'_{x,b}$ takes small values, weakening the positive feedback loop, since X cannot profit much from increased production of B. Overall, the effect of $g'_{x,b}$ on the stability of the system is much smaller compared to the effects of m'_x and m'_y. Considering the eigenvectors corresponding to the largest eigenvalue for the parameter sets at the bifurcation, indicates again that the two bacteria and one of the metabolites change in the same direction, whereas the other metabolite changes in the opposite direction. However, in the range where $g'_{x,b}$ stabilizes the system, the difference in the eigenvector entries for the two bacteria decreases.

Next, we want to analyze the role of the different branching parameters for the stability of the system. The fractions of metabolite supply to the system, γ_a and γ_b, and the sensitivity of bacteria X biomass growth to its metabolic throughput, m'_x, span the three-dimensional parameter space in Fig. 5.5. The bifurcation diagrams are plotted for three different parameter values for the fractions of natural decay of metabolites, β_a and β_b, from low fractions of 0.1 (A), to intermediate fractions of 0.5 (B) to high fractions of 0.9 (C). In general, increasing the portions of metabolites that are supplied externally to the system, γ_a and γ_b, has a positive effect on stability, whereas increasing the fractions that decay, β_a and β_b, has a destabilizing effect on the stability of the system.

This is intriguing because our initial assumption would likely be that low fractions of bacterial metabolite production, i.e. high values of γ_a and γ_b, and low portions of uptake

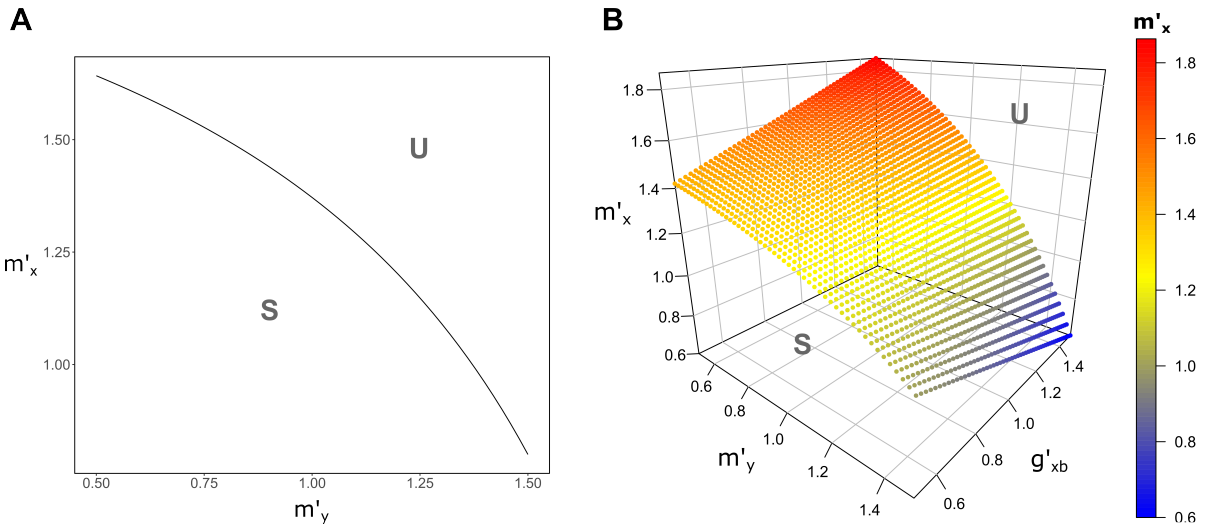

Figure 5.4: Bifurcation diagrams of bacteria-metabolite system, depending on the sensitivity of bacteria Y biomass growth to metabolic throughput of bacteria Y, m'_y, and the sensitivity of bacteria X biomass growth to metabolic throughput of bacteria X, m'_x, (A) and additionally on the sensitivity of metabolite A production to concentration of B, $g_{x,b}$ (B). Steady states are stable (S) in the lower part/volume of parameter space. Stability is lost (U) in a bifurcation of a saddle-node type, when crossing the line/surface (parameters: $\beta_a = \beta_b = \gamma_a = \gamma_b = 0.5$, $l'_{x,x} = l'_{y,y} = 1.2$, $g'_{x,b} = g'_{y,a} = 1.0$).

.

by the bacteria, i.e. high values of β_a and β_b, would weaken the positive feedback effect, due to less coupling between the bacteria. However, the ratio of metabolites consumed by the bacteria to the trade metabolite fraction produced by the bacteria is key in determining stability of the positive steady state. When the bacteria predominantly consume the partner's metabolite while making only a minor contribution to the partner's trade metabolite concentration, this contributes to the system's stability. The eigenvectors corresponding to the largest eigenvalue for the parameter sets at the bifurcation indicate that the bacterial population profiting the most from the mutualistic cross-feeding relationship experiences the largest changes.

Let us now examine what increasing γ and β at a constant ratio does to the system's stability (Fig. 5.6). As before, we see that in general higher fractions of metabolite supply, γ, compared to metabolite decay, β, are stabilizing (compare Y-axes of Fig. 5.6A vs. B). In the latter case, i.e. $\gamma > \beta$, increasing γ and β, while maintaining a constant ratio between them has a destabilizing effect (Fig. 5.6A). Conversely, if $\gamma < \beta$, increasing both at a constant ratio results in a minor stabilizing effect until large values of m_y (Fig. 5.6B). In systems, where $\gamma > \beta$ the fraction of bacterial uptake is larger than the fraction of bacterial production, i.e. the negative feedback loop that we identified before. If we decrease the fractions that are part of this negative feedback loop, this has a destabilizing

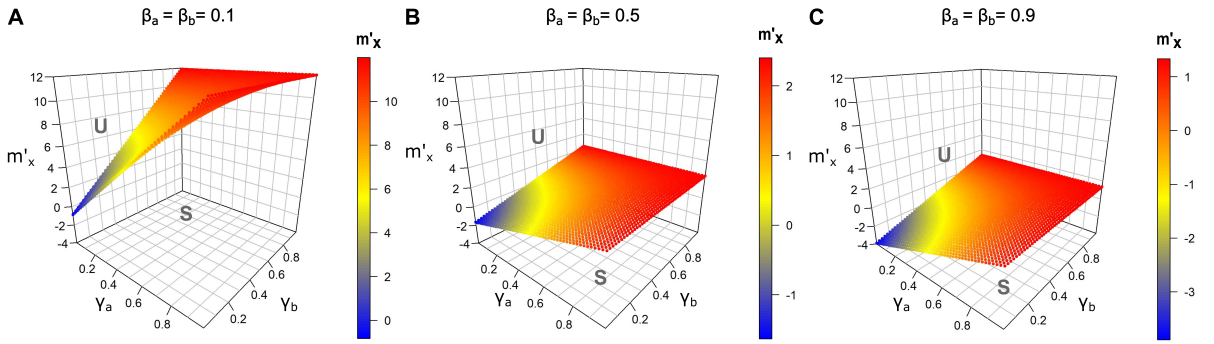

Figure 5.5: Bifurcation diagrams of bacteria-metabolite system, depending on the fraction of supply of metabolite A, γ_a, fraction of supply of metabolite B, γ_b, and sensitivity of bacteria X biomass growth to metabolic throughput of bacteria X, m'_x. Fraction of losses of metabolite A, β_a, and metabolite B, β_b are set to 0.1 (A), 0.5 (B) and 0.9 (C). Note the different color scales. Steady states are stable in the lower volume of parameter space. Stability is lost in a bifurcation of a saddle-node type, when crossing the line/surface (parameters: $m'_x = 1.5, l'_{x,x} = l'_{y,y} = 1.2$, $g'_{x,b} = g'_{y,a} = 1.0$).

.

effect since it weakens the impact of this inhibitory feedback loop. Conversely, for $\gamma < \beta$, the fraction of bacterial uptake is lower than the fraction of bacterial production, resulting in a positive feedback loop.

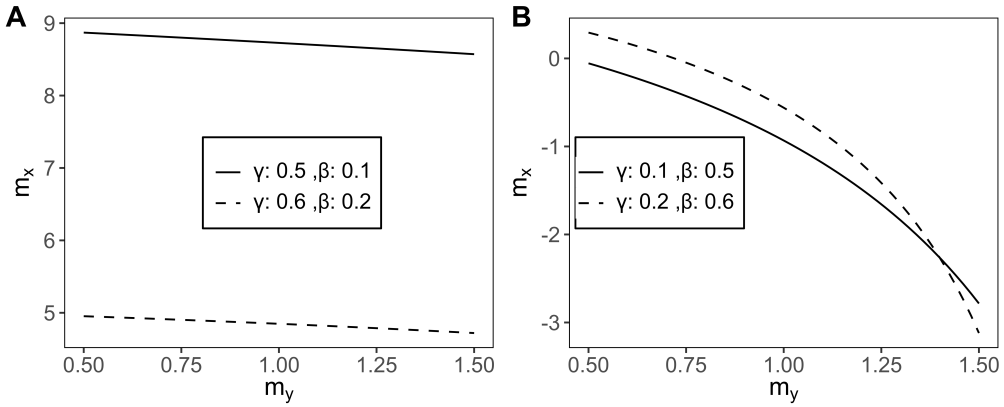

Figure 5.6: Bifurcation diagrams of bacteria-metabolite system, depending on the sensitivity of bacteria Y biomass growth to metabolic throughput of bacteria Y, m'_y, and the sensitivity of bacteria X biomass growth to metabolic throughput of bacteria X, m'_x for different combinations of fractions of gains of metabolites from supply, γ, and fractions of losses of metabolites to natural decay, β. For $\gamma > \beta$ increasing γ and β at a constant ratio results in a destabilization of the system (A), whereas when $\gamma < \beta$, increasing both at a constant ratio results in a stabilization until large values of m_y (B). Steady states are stable in the lower part of the parameter space. Stability is lost in a bifurcation of a saddle-node type, when crossing the line (parameters: $l'_{x,x} = l'_{y,y} = 1.2$, $g'_{x,b} = g'_{y,a} = 1.0$).

.

5.6 Reduction to a 2-dimensional system

Since the 4-dimensional system is still relatively complicated, it makes sense to reduce it to a 2-dimensional system to obtain further insights into the stability properties, enabling us to obtain analytical results. For the reduction to 2-dimensional we take advantage of the structure of the Jacobian of the 4-dimensional system. Assuming equal steady state values of the two bacterial populations and the two metabolite concentrations respectively as well as equal parameter pairs, e.g. $m'_x = m'_y$, $g'_{x,b} = g'_{y,a}$, we see that in this case, the matrix

$$\mathbf{P} = \begin{pmatrix} \alpha_x & 0 & 0 & 0 \\ 0 & \alpha_y & 0 & 0 \\ 0 & 0 & \alpha_a & 0 \\ 0 & 0 & 0 & \alpha_b \end{pmatrix} \begin{pmatrix} m'_x g'_{x,x} - l'_{x,x} & 0 & 0 & m'_x g'_{x,b} \\ 0 & m'_y g'_{y,y} - l'_{y,y} & m'_y g'_{y,a} & 0 \\ \tilde{\gamma}_a g'_{x,x} & -\tilde{\beta}_a g'_{y,y} & -\beta_a d'_{a,a} - \tilde{\beta}_a g'_{y,a} & \tilde{\gamma}_a g'_{x,b} \\ -\tilde{\beta}_b g'_{x,x} & \tilde{\gamma}_b g'_{y,y} & \tilde{\gamma}_b g'_{y,a} & -\beta_b d'_{b,b} - \tilde{\beta}_b g'_{x,b} \end{pmatrix} \tag{5.32}$$

is of the structure

$$\begin{pmatrix} A & 0 & 0 & B \\ 0 & A & B & 0 \\ C & D & E & F \\ D & C & F & E \end{pmatrix}. \tag{5.33}$$

And hence

$$\begin{pmatrix} A & 0 & 0 & B \\ 0 & A & B & 0 \\ C & D & E & F \\ D & C & F & E \end{pmatrix} \begin{pmatrix} a \\ a \\ b \\ b \end{pmatrix} = \begin{pmatrix} Aa + Bb \\ Aa + Bb \\ (C+D)a + (E+F)b \\ (C+D)a + (E+F)b \end{pmatrix}. \tag{5.34}$$

We can see that the vector $\begin{pmatrix} a & a & b & b \end{pmatrix}^T$ is an eigenvector of the matrix with corresponding eigenvalue λ if two conditions are met. Given that these conditions are met, we can represent the conditions in the matrix equation as a 2x2 matrix

$$\begin{pmatrix} A & B \\ (C+D) & (E+F) \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \end{pmatrix}. \tag{5.35}$$

Therefore, the 4x4 matrix has an eigenvalue λ if this 2x2 matrix has an eigenvalue λ.

Similarly, we can show that

$$\begin{pmatrix} A & 0 & 0 & B \\ 0 & A & B & 0 \\ C & D & E & F \\ D & C & F & E \end{pmatrix} \begin{pmatrix} a \\ -a \\ b \\ -b \end{pmatrix} = \begin{pmatrix} Aa - Bb \\ -Aa + Bb \\ (C-D)a + (E-F)b \\ (D-C))a + (F-E)b \end{pmatrix}. \tag{5.36}$$

In this case we get

$$\begin{pmatrix} A & -B \\ (C-D) & (E-F) \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \end{pmatrix}. \tag{5.37}$$

While in the first case the instabilities obey the symmetry, in this case they break the symmetry.

Taken together, we expect to get two eigenvectors from the symmetric case and two from the asymmetric case. Hence, we can find all eigenvectors of the 4x4 matrix in this way. Since there are no additional eigenvectors, λ can only be an eigenvalue of the 4x4 matrix if it is an eigenvalue of at least one of the 2x2 matrices. Consequently, assuming equal steady state values of the bacteria and the metabolites and equal parameter pairs, we can write the 4x4 Jacobian of the 4-dimensional system as two 2x2 Jacobians

$$\mathbf{Q} = \begin{pmatrix} \alpha & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} m'_z g'_{z,z} - l'_{z,z} & m'_z g'_{z,c} \\ \tilde{\gamma}_c g'_{z,z} - \tilde{\beta}_c g'_{z,z} & -\beta_c d'_{c,c} - \tilde{\beta}_c g'_{z,c} + \tilde{\gamma}_c g'_{z,c} \end{pmatrix} \tag{5.38}$$

$$\mathbf{R} = \begin{pmatrix} \alpha & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} m'_z g'_{z,z} - l'_{z,z} & -m'_z g'_{z,c} \\ \tilde{\gamma}_c g'_{z,z} + \tilde{\beta}_c g'_{z,z} & -\beta_c d'_{c,c} - \tilde{\beta}_c g'_{z,c} - \tilde{\gamma}_c g'_{z,c} \end{pmatrix}, \tag{5.39}$$

where z now represents the bacteria and c denotes the metabolites. In the following, we set $g'_{zz} = 1$ and $d'_{cc} = 1$ as before and use a simplified notation, i.e. $m'_z := m, l'_{zz} := l, g'_{z,c} := g, \tilde{\gamma}_c := 1 - \gamma, \beta_c := \beta, \tilde{\beta}_c := 1 - \beta$.

5.6.1 Stability conditions

The simplicity of the eigenvalue analysis for 2x2 matrices allows us to derive the stability conditions for the 2-dimensional system. In a 2-dimensional system, a steady state is considered stable, if the determinant of the Jacobian is positive and the trace is negative. Consequently, during the Hopf bifurcation, the trace becomes positive, while in the case of a saddle-node type bifurcation the determinant becomes negative. Therefore, to see if

the Hopf bifurcation can occur in the symmetric case, we consider

$$\mathbf{Q} = \begin{pmatrix} \alpha & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} m-l & mg \\ \beta-\gamma & -\beta+\beta g-\gamma g \end{pmatrix} \tag{5.40}$$

with

$$\begin{aligned} \det(\mathbf{Q}) &= \alpha(\beta l+\gamma g l-\beta g l-\beta m) > 0 && (5.41)\\ \mathrm{tr}(\mathbf{Q}) &= \alpha(m-l)-\beta+g(\beta-\gamma)=0. && (5.42) \end{aligned}$$

We know that $-\beta < 0$ and $g(\beta-\gamma) < 0$ if $\beta < \gamma$, hence β and γ have to take small values and $m > l$ for a Hopf bifurcation to occur.

For the Jacobian of the asymmetric case we consider

$$\mathbf{R} = \begin{pmatrix} \alpha & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} m-l & -mg \\ 2-\beta-\gamma & \beta g+\gamma g-\beta-2g \end{pmatrix} \tag{5.43}$$

with

$$\begin{aligned} \det(\mathbf{R}) &= \alpha(\beta l+2gl-\beta m-\beta g l-\gamma g l) > 0 && (5.44)\\ \mathrm{tr}(\mathbf{R}) &= \alpha(m-l)+\beta g+\gamma g-\beta-2g=0. && (5.45) \end{aligned}$$

A Hopf bifurcation within a realistic parameter range will not occur, since $\beta g+\gamma g < 2g$ the trace will always be negative.

Hence, a Hopf bifurcation can only occur in the symmetric case and only if the system is characterized by a tight cooperation between the bacteria. This means that we do not see a periodic cycle between the species, instead we expect a periodic cycle between the bacteria and the metabolites. The cycle is driven by resource exploitation. Since we do not see a Hopf bifurcation in the asymmetric case we do not expect cycles of shifting dominance between the two species.

Next, our aim is to identify the saddle-node type bifurcations. We consider

$$\begin{aligned} \det(\mathbf{Q}) &= \alpha(\beta l+\gamma g l-\beta g l-\beta m)=0 && (5.46)\\ \det(\mathbf{R}) &= \alpha(\beta l+2gl-\beta m-\beta g l-\gamma g l)=0, && (5.47) \end{aligned}$$

while keeping in mind that the trace values have to be negative for the saddle-node type bifurcation to occur in the 2-dimensional systems. To gain insights into the stabilizing or

destabilizing properties of each parameter, we differentiate the determinant with respect to each parameter.

First, we investigate the Jacobian $\mathbf{Q}$, i.e. the scenario where the instabilities obey the symmetric condition. We obtain

$$\frac{\partial \det(\mathbf{Q})}{\partial \gamma} = \alpha g l \tag{5.48}$$
$$\frac{\partial \det(\mathbf{Q})}{\partial m} = -\alpha\beta \tag{5.49}$$
$$\frac{\partial \det(\mathbf{Q})}{\partial g} = \alpha(\gamma l - \beta l) \tag{5.50}$$
$$\frac{\partial \det(\mathbf{Q})}{\partial l} = \alpha(\beta + \gamma g - \beta g) \tag{5.51}$$
$$\frac{\partial \det(\mathbf{Q})}{\partial \beta} = \alpha(l - gl - m). \tag{5.52}$$

As expected, we see that γ is always stabilizing, whereas m is always destabilizing. Also, we see again that the effect of g depends on the ratio of γ to β. When $\gamma > \beta$, higher values of g contribute to stabilization, and conversely, they lead to destabilization when $\gamma < \beta$. Interestingly, the part $g(\beta-\gamma)-\beta$ of the derivative with respect to l, equals a part of the trace. So, for l to be destabilizing, $g(\beta-\gamma)-\beta > 0$, hence $\beta > \gamma$ and $g > 1$ and for the trace to still be negative $l > m$. Therefore, increasing l can destabilize a system, where the bacteria strongly depend on the production of the partner's metabolite, while their profit from producing the trade metabolite is limited. However, these systems are not stable within the plausible parameter range. Increasing the parameter β is destabilizing if $m > 1$ or $g \geq 1$.

Considering the Jacobian $\mathbf{R}$, i.e. the scenario where the instabilities break the symmetric condition, we obtain

$$\frac{\partial \det(\mathbf{R})}{\partial \gamma} = \alpha(-gl) \tag{5.53}$$
$$\frac{\partial \det(\mathbf{R})}{\partial m} = -\alpha\beta \tag{5.54}$$
$$\frac{\partial \det(\mathbf{R})}{\partial g} = \alpha(2l - \beta l - \gamma l) \tag{5.55}$$
$$\frac{\partial \det(\mathbf{R})}{\partial l} = \alpha(\beta + 2g - \beta g - \gamma g) \tag{5.56}$$
$$\frac{\partial \det(\mathbf{R})}{\partial \beta} = \alpha(l - m - gl). \tag{5.57}$$

We can see that the conditions for m and β mirror those for the Jacobian $\mathbf{Q}$. Large values

of the parameters g and l are always stabilizing in the considered parameter range, since $2 > \gamma + \beta$. Interestingly, γ is always destabilizing in this scenario, probably because it decreases the dependence of the bacteria on the interaction and therefore could lead to diverging developments of the two bacterial populations.

5.6.2 Response to parameter change

In Sec. 3.4, we showed how to analyze the response of the system to a permanent change of parameters, i.e. a press perturbation. We apply this approach here to analyze the 2-dimensional system with respect to a reduction in bacterial density and a reduction in metabolite concentration. Let us first examine the scenario where the instabilities obey the symmetry. We can now ask, how an additional loss term for the bacteria impacts the steady state. Substituting the Jacobian into Eq. 3.103, we can write

$$\boldsymbol{\delta} = -\frac{1}{\det \mathbf{Q}} \begin{pmatrix} -\beta + \beta g - \gamma g & -\alpha m g \\ -\beta + \gamma & \alpha(m-l) \end{pmatrix} \begin{pmatrix} -\epsilon \\ 0 \end{pmatrix}, \tag{5.58}$$

where ϵ is the small fraction of the bacteria that is lost and

$$\det(\mathbf{Q}) \;=\; \alpha(\beta l + \gamma g l - \beta g l - \beta m). \tag{5.59}$$

We can examine the two components, i.e. for the bacteria and the metabolites respectively, of the resulting impact on the steady state

$$\delta_{\text{bac}} = \frac{-\beta + g(\beta - \gamma)}{\det \mathbf{Q}}\epsilon, \qquad \delta_{\text{met}} = \frac{\gamma - \beta}{\det \mathbf{Q}}\epsilon. \tag{5.60}$$

If the sensitivity of metabolite production to the concentration of the respective other metabolite, $g \leq 1$ a small additional loss of bacteria results in a reduction of bacterial density. Only if $g > 1$ and $\beta > \gamma$ then there could also be a positive effect of a small loss of bacteria on bacteria's density. However, these systems are unstable within the plausible parameter range. The impact term for the metabolites shows that if $\gamma > \beta$, i.e. bacteria take up a higher fraction of metabolites than they produce, then a small reduction in bacteria leads to an increase in metabolites, whereas if $\gamma < \beta$ metabolites are reduced by a loss of bacteria.

In the case that the system experiences a small addition of bacteria, we write

$$\boldsymbol{\delta} = -\frac{1}{\det \mathbf{Q}} \begin{pmatrix} -\beta + \beta g - \gamma g & -\alpha m g \\ -\beta + \gamma & \alpha(m-l) \end{pmatrix} \begin{pmatrix} \epsilon \\ 0 \end{pmatrix}, \tag{5.61}$$

resulting in

$$\delta_{\text{bac}} = \frac{\beta + g(\gamma - \beta)}{\det \mathbf{Q}}\epsilon, \qquad \delta_{\text{met}} = \frac{\beta - \gamma}{\det \mathbf{Q}}\epsilon. \tag{5.62}$$

Hence, if $\gamma > \beta$, a small addition of bacteria leads to an increase in bacterial density. Only in the case that $\gamma << \beta$ and $g > 1$ an addition of bacteria to the system could lead to a decrease in bacteria's density. However, we saw before that for system where $\gamma << \beta$, the system in unstable in the plausible parameter range. For the metabolites a gain in bacteria means an increase in metabolite concentrations if $\beta > \gamma$ and vice versa.

We can now also investigate the impact of a small additional loss of metabolites on the system's steady state, i.e.

$$\boldsymbol{\delta} = -\frac{1}{\det \mathbf{Q}} \begin{pmatrix} -\beta + \beta g - \gamma g & -\alpha m g \\ -\beta + \gamma & \alpha(m - l) \end{pmatrix} \begin{pmatrix} 0 \\ -\epsilon \end{pmatrix}, \tag{5.63}$$

resulting in

$$\delta_{\text{bac}} = \frac{-\alpha m g}{\det \mathbf{Q}}\epsilon, \qquad \delta_{\text{met}} = \frac{\alpha(m - l)}{\det \mathbf{Q}}\epsilon. \tag{5.64}$$

Hence, a loss of metabolites leads to a decrease in bacterial density. The response of the metabolite concentrations depends on the ratio of m/l. If $m > l$ then the loss of metabolites results in an increase of metabolite concentrations and vice versa. If we add an additional gain term of metabolites, we can write

$$\boldsymbol{\delta} = -\frac{1}{\det \mathbf{Q}} \begin{pmatrix} -\beta + \beta g - \gamma g & -\alpha m g \\ -\beta + \gamma & \alpha(m - l) \end{pmatrix} \begin{pmatrix} 0 \\ \epsilon \end{pmatrix}, \tag{5.65}$$

hence

$$\delta_{\text{bac}} = \frac{\alpha m g}{\det \mathbf{Q}}\epsilon, \qquad \delta_{\text{met}} = \frac{\alpha(l - m)}{\det \mathbf{Q}}\epsilon. \tag{5.66}$$

Consequently, an increase in metabolite concentrations results in an increase in bacterial density. Metabolite concentrations increases in response if $l > m$ and vice versa.

Let us now consider, how an additional loss term for the bacteria impacts the steady state of the Jacobian $\mathbf{R}$, i.e. the asymmetric case. Substituting the Jacobian into Eq. 3.103 yields

$$\boldsymbol{\delta} = -\frac{1}{\det \mathbf{R}} \begin{pmatrix} \beta g + \gamma g - \beta - 2g & \alpha m g \\ -2 + \beta + \gamma & \alpha(m - l) \end{pmatrix} \begin{pmatrix} -\epsilon \\ 0 \end{pmatrix}, \tag{5.67}$$

where ϵ is the small fraction of the bacteria that is lost and

$$\det(\mathbf{R}) \quad = \quad \alpha(\beta l + 2gl - \beta m - \beta gl - \gamma gl). \tag{5.68}$$

We can examine the two components, i.e. for the bacteria and the metabolites respectively, of the resulting impact on the steady state

$$\delta_{\text{bac}} = \frac{\beta g + \gamma g - \beta - 2g}{\det \mathbf{R}}\epsilon, \qquad \delta_{\text{met}} = \frac{-2 + \beta + \gamma}{\det \mathbf{R}}\epsilon. \tag{5.69}$$

We see that if a small fraction of bacteria is lost, bacterial populations as well as metabolite concentrations decrease.

In the case that the system experiences a small addition of bacteria, we write

$$\boldsymbol{\delta} = -\frac{1}{\det \mathbf{R}} \begin{pmatrix} \beta g + \gamma g - \beta - 2g & \alpha m g \\ -2 + \beta + \gamma & \alpha(m - l) \end{pmatrix} \begin{pmatrix} \epsilon \\ 0 \end{pmatrix}, \tag{5.70}$$

resulting in

$$\delta_{\text{bac}} = \frac{-\beta g - \gamma g + \beta + 2g}{\det \mathbf{R}}\epsilon, \qquad \delta_{\text{met}} = \frac{2 - \beta - \gamma}{\det \mathbf{R}}\epsilon. \tag{5.71}$$

In this case, both the bacterial populations as well as the metabolite concentrations increase.

Investigating the impact of a small additional loss of metabolites on the system's steady state,

$$\boldsymbol{\delta} = -\frac{1}{\det \mathbf{R}} \begin{pmatrix} \beta g + \gamma g - \beta - 2g & \alpha m g \\ -2 + \beta + \gamma & \alpha(m - l) \end{pmatrix} \begin{pmatrix} 0 \\ -\epsilon \end{pmatrix}, \tag{5.72}$$

results in

$$\delta_{\text{bac}} = \frac{\alpha m g}{\det \mathbf{R}}\epsilon, \qquad \delta_{\text{met}} = \frac{\alpha(m - l)}{\det \mathbf{R}}\epsilon. \tag{5.73}$$

Interestingly, a loss of metabolites leads to an increase in bacterial density. This may happen in the asymmetric case because decreasing the metabolite concentration increases the dependence on the trade partner's metabolite production and makes a shift of dominance of one bacterium less likely. The response of the metabolite concentration depends on the ratio of m/l. If $m > l$ then the loss in metabolite results in an increase of metabolite concentration and vice versa.

If we add an additional gain term of metabolites, we can write

$$\boldsymbol{\delta} = -\frac{1}{\det \mathbf{R}} \begin{pmatrix} \beta g + \gamma g - \beta - 2g & \alpha m g \\ -2 + \beta + \gamma & \alpha(m-l) \end{pmatrix} \begin{pmatrix} 0 \\ \epsilon \end{pmatrix}, \tag{5.74}$$

hence

$$\delta_{\text{bac}} = \frac{-\alpha m g}{\det \mathbf{Q}} \epsilon, \qquad \delta_{\text{met}} = \frac{\alpha(m-l)}{\det \mathbf{Q}} \epsilon. \tag{5.75}$$

Consequently, an increase in metabolite concentration results in decrease of bacterial density. In this case, an increase of metabolites could lead to the bacteria being less dependent on each other's metabolite production and hence a shift of dominance is more likely to occur. Metabolite concentrations increase in response if $m > l$ and vice versa.

To summarize, we have seen how different the effects of the parameters and changes in the system can be, depending on which instability we consider. While in the symmetric case the independence on the trading partner's production stabilizes the system, in the asymmetric case the opposite is true. This can lead to interesting observations, such as an increase of bacterial populations in response to a decrease of metabolites, a counter-intuitive response, also known as hydra effect [223].

5.7 Spatial system

The examined motif of bacterial cross-feeding is not confined to isolated occurrences in natural communities; instead, it is interwoven within a spatial trade network [211]. Consequently, the spatial aspect significantly influences the dynamics of these communities and, in turn, has broader implications for example for host health, ecosystem functioning and global nutrient cycling. In this section, we expand the small motif of bacterial cross-feeding to a spatial trade network. We explore the occurrence of diffusion-driven Turing instabilities [224] to investigate the emergence of complex spatial patterns. Interactions in systems coupled by diffusion can result in the spontaneous emergence of patterns, carrying significant implications [225, 226]. We begin with the 2-dimensional system and then move on to analyze the 4-dimensional system for the case that metabolites can diffuse in space, whereas bacteria cannot.

5.7.1 2D system

Let us start by considering the occurrence of Turing instabilities in the 2-dimensional system. As in the non-spatial system, we examine the two Jacobians $\mathbf{Q}$ and $\mathbf{R}$. From the Jacobian of the symmetric case, $\mathbf{Q}$, we derive the Jacobian of the diffusion-coupled

system [114], which is

$$\mathbf{J_Q} = \begin{pmatrix} \alpha(m-l) - \rho\kappa & \alpha m g \\ \beta - \gamma & -\beta + \beta g - \gamma g - \kappa \end{pmatrix}, \tag{5.76}$$

where ρ is the quotient of the diffusion coefficients of the bacteria and metabolites and κ is the wave number, i.e the eigenvalue of the Laplacian operator. The determinant for this matrix yields a quadratic polynomial with respect to κ

$$\det(\mathbf{J_Q}) = (\alpha(m-l) - \rho\kappa)(-\beta + \beta g - \gamma g - \kappa) - (\beta - \gamma)\alpha m g \tag{5.77}$$

$$= \rho\kappa^2 + \rho\kappa\beta - \rho\kappa\beta g + \rho\kappa\gamma g - \alpha\kappa m + \alpha\kappa l - \alpha\beta m + \alpha\beta l - \alpha\beta g l + \alpha\gamma g l. \tag{5.78}$$

For a Turing instability to occur in this system, presuming stability of the equilibrium in the local model, the extremum of the polynomial, κ_e, must be a minimum with $\kappa_e > 0$ and $\det(\mathbf{J_Q}(\kappa_\mathbf{e})) \leq 0$.

To find the minimum, κ_e, we differentiate Eq. 5.77 with respect to κ and set the resulting equation equal to zero

$$0 \;=\; \rho 2\kappa + \rho\beta - \rho\beta g + \rho\gamma g - \alpha m + \alpha l, \tag{5.79}$$

we obtain

$$\kappa_e \;=\; \frac{-\rho\beta + \rho\beta g - \rho\gamma g + \alpha m - \alpha l}{2\rho}. \tag{5.80}$$

To assure that $\det(\mathbf{J_Q}(\kappa_\mathbf{e})) \leq 0$, we substitute κ_e into Eq. 5.77

$$\det(\mathbf{J_Q}(\kappa_\mathbf{e})) \;=\; \frac{\rho(-\beta^2 - \beta^2 g^2 + 2\beta^2 g - 2\beta\gamma g + 2\beta\gamma g^2)}{4} + \frac{\alpha^2(-l^2 - m^2 + 2lm)}{4\rho}$$

$$+ \frac{\alpha(-\beta m + \beta l - \beta g m - \beta g l + \gamma g m + \gamma g l)}{2} \tag{5.81}$$

$$\leq\; 0. \tag{5.82}$$

For small ρ, i.e. the diffusion coefficient of metabolites is much larger than the diffusion coefficient of bacteria, the following term dominates

$$0 \;\geq\; \left(\frac{\alpha^2(-l^2 - m^2 + 2lm)}{4\rho}\right) \tag{5.83}$$

$$\geq\; \left(\frac{-\alpha^2(m-l)^2}{4\rho}\right) \tag{5.84}$$

This term takes only negative values, hence the system is always unstable at κ_e

$(\det(\mathbf{J_Q}(\kappa_\mathbf{e}) < 0))$ if ρ is small. For small ρ, in Eq. 5.80 the following term dominates

$$\kappa_e = \frac{\alpha(m-l)}{2\rho}. \tag{5.85}$$

As κ is the eigenvalue of the Laplacian operator, it is constrained to non-negative values. The zero eigenvalue has to be stable, since this describes the local system. Consequently, a necessary condition for a Turing bifurcation is $\kappa_e > 0$, hence finite positive wave numbers require $m > l$. This means that $\mathrm{J_{Q_{11}}} > 0$, hence $\mathrm{tr(J_Q)}$ requires stabilization by β or γg.

From the Jacobian of the asymmetric case, $\mathbf{R}$, we derive the Jacobian of the diffusion-coupled system, which is

$$\mathbf{J_R} = \begin{pmatrix} \alpha(m-l) - \rho\kappa & -\alpha m g \\ 2-\beta-\gamma & \beta g + \gamma g - \beta - 2g - \kappa \end{pmatrix}, \tag{5.86}$$

where ρ is the quotient of the diffusion coefficients of the bacteria and metabolites and κ is the eigenvalue of the Laplacian operator. The determinant for this matrix yields a quadratic polynomial with respect to κ

$$\begin{aligned} \det(\mathbf{J_R}) &= (\alpha(m-l)-\rho\kappa)(\beta g + \gamma g - \beta - 2g - \kappa) - (2-\beta-\gamma)(-\alpha m g) && (5.87)\\ &= \rho\kappa^2 - \alpha m\kappa + \alpha l\kappa - \beta\rho g\kappa - \gamma\rho g\kappa + \beta\rho\kappa + 2\rho g\kappa - \alpha\beta m - \alpha\beta g l \\ &\quad -\alpha\gamma g l + \alpha\beta l + 2\alpha g l && (5.88)\end{aligned}$$

Differentiating with respect to κ and setting the resulting equation equal to zero

$$0 = 2\rho\kappa - \alpha m + \alpha l - \beta\rho g - \gamma\rho g + \beta\rho + 2\rho g, \tag{5.89}$$

we obtain

$$\kappa_e = \frac{\alpha m - \alpha l + \beta\rho g + \gamma\rho g - \beta\rho - 2\rho g}{2\rho}. \tag{5.90}$$

Substituting κ_e into Eq. 5.87 yields

$$\begin{aligned} \det(\mathbf{J_R}(\kappa_\mathbf{e})) &= \rho\left(\frac{\alpha m - \alpha l + \beta\rho g + \gamma\rho g - \beta\rho - 2\rho g}{2\rho}\right)^2 \\ &\quad -\left(\frac{(\alpha m - \alpha l + \beta\rho g + \gamma\rho g - \beta\rho - 2\rho g)^2}{2\rho}\right) \\ &\quad -\alpha\beta m - \alpha\beta g l - \alpha\gamma g l + \alpha\beta l + 2\alpha g l. && (5.91)\end{aligned}$$

For small ρ, i.e. the diffusion coefficient for metabolites are much larger than the diffusion coefficient for bacteria, the following term dominates

$$0 \geq \frac{-(\alpha m - \alpha l + \beta\rho g + \gamma\rho g - \beta\rho - 2\rho g)^2}{2\rho}. \tag{5.92}$$

This term takes only negative values, hence the system is always unstable at κ_e if ρ is small. For small ρ, the minimum is at

$$\kappa_e = \frac{\alpha(m-l)}{2\rho}. \tag{5.93}$$

Since one condition is $\kappa_e > 0$, finite positive wave numbers require $m > l$. This means that $\mathrm{J}_{\mathrm{Q}_{11}} > 0$, however this is stabilized by β and g.

Overall, in the case that only metabolites diffuse, we can identify the following factor as promoter for diffusion-induced instabilities: The sensitivity of the bacterial biomass to metabolic throughput, m, is higher than the sensitivity of bacterial mortality to bacterial density, l.

5.7.2 4D system

We now analyze the 4-dimensional system, applying the Master Stability Function Approach introduced in Sec. 3.6. Similarly to the symmetric reduction to the 2-dimensional system that we performed before in Sec. 5.6, also the master stability function approach exploits a symmetry in the system. In this case, it is the symmetry that the local patches all contain the same local community motif.

Eq. 5.24 can be straightforwardly extended to the case of longer range metabolic interactions using a master stability function approach [120, 133, 134]. Namely, we are interested in extending Eq. 5.24 to the case of metabolites that spread on an unspecified spatial network with nodes that represent habitat patches. These patches contain copies of the community motif, and are linked by metabolite diffusion. The eigenvalues of the meta-Jacobian $\mathbf{J}$ with patches linked by density-independent per capita dispersal can be calculated as $\mathrm{Ev}(\mathbf{P} - \kappa\mathbf{C})$. The eigenvalues of this matrix will equal the eigenvalues of $\mathbf{J}$ for a spatial topology represented by a Laplacian matrix with eigenvalues κ [120].

$\mathbf{C}$ is the coupling matrix with diagonal entries representing diffusion rates r_a and r_b. We consider the case of metabolites that spread on much faster timescales than bacteria. We

can therefore rewrite Eq. 5.24 for the general spatial case as

$$\mathbf{J} = \mathbf{E}\begin{pmatrix} g'_{x,x} - \beta_x l'_{x,x} - \tilde{\beta}_x t'_{x,x} & 0 & -\tilde{\beta}_x t'_{x,a} & g'_{x,b} \\ 0 & g'_{y,y} - \beta_y l'_{y,y} - \tilde{\beta}_y t'_{y,y} & g'_{y,a} & -\tilde{\beta}_y t'_{y,b} \\ g'_{x,x} & -\tilde{\beta}_a g'_{y,y} & -\beta_a d'_{a,a} - \tilde{\beta}_a g'_{y,a} - \kappa r_a & g'_{x,b} \\ -\tilde{\beta}_b g'_{x,x} & g'_{y,y} & g'_{y,a} & -\beta_b d'_{b,b} - \tilde{\beta}_b g'_{x,b} - \kappa r_b \end{pmatrix}, \tag{5.94}$$

with

$$\mathbf{E} = \begin{pmatrix} \alpha_x & 0 & 0 & 0 \\ 0 & \alpha_y & 0 & 0 \\ 0 & 0 & \alpha_a & 0 \\ 0 & 0 & 0 & \alpha_b \end{pmatrix}. \tag{5.95}$$

When $\kappa = 0$, the system is equivalent to the nonspatial model. The spatial system is stable on any spatial topology if all eigenvalues for all $\kappa \geq 0$ are negative. Since the Laplacian is a positive semidefinite matrix, it will have at least one zero eigenvalue. This means that the multi-patch system cannot be stable on any spatial topology if the nonspatial system is unstable. That is, spatial flows of metabolites have the potential to destabilize already stable systems, but not the reverse.

In Fig. 5.7 we see an example of a diffusion-driven Turing instability, occurring due to diffusion of metabolites across the network. The master stability function relates the Jacobian eigenvalue of the spatial system to the Laplacian eigenvalue of the spatial network [120] (Fig. 5.7A). Here it reveals that stronger coupling strength (bubbles in Fig. 5.7A) can destabilize a previously stable system by expanding the range of the Laplacian spectrum. The Turing instability here is very localized (Fig. 5.7B and C), affecting only few nodes the network. This instability is the result of a positive feedback loop within the system, as bacterial growth is enhanced by higher metabolite concentrations.

The identified instability is only meaningful in the context of a network, as in continuous space, the parameter κ can become infinitely large. Consequently, if the system exhibits instability for high values of κ in continuous space, it implies an unrealistic scenario where everything converges to a single point in space. On a network, however, large values of κ indicate concentration on a node, which represents a plausible scenario.

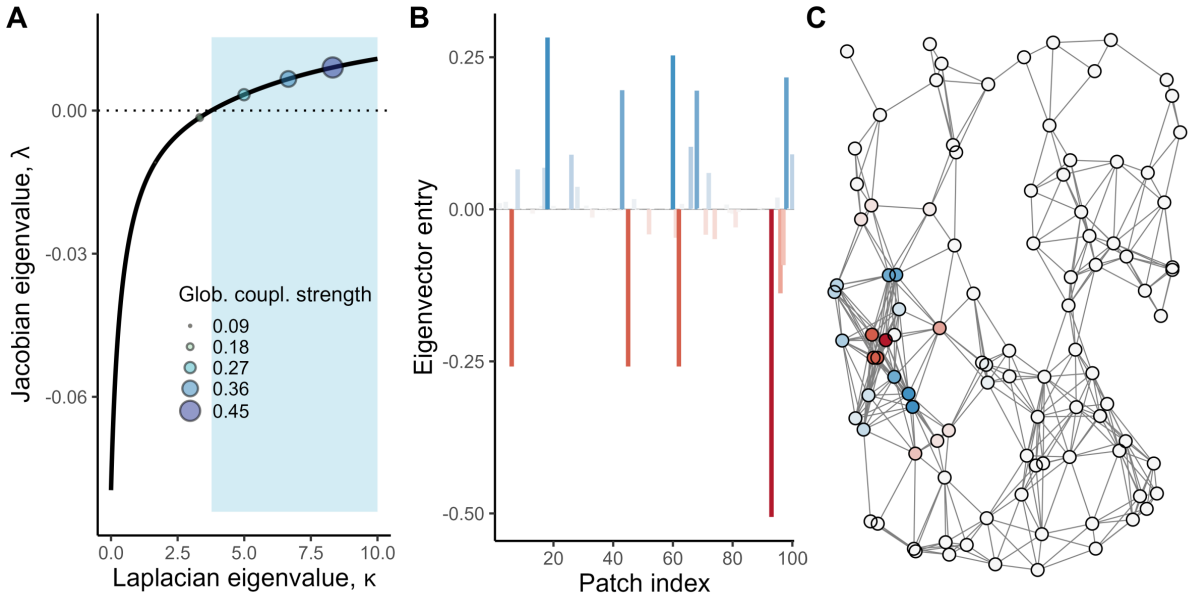

Figure 5.7: Example of a diffusion-driven Turing instability that occurs when metabolites diffuse. The master stability function (MSF) (A) links the Jacobian eigenvalue of the spatial system to the Laplacian eigenvalue of the spatial network. The spatial system is stable if none of the Laplacian eigenvalues lie within the interval where the MSF takes on positive values (blue shaded area). Increasing global coupling strength (Glob. coupl. strength) can destabilize the system by expanding the range of the Laplacian spectrum. Panel B shows the eigenvector entries of the different patches and panel B shows the example of a fully connected random geometric graph with a 100 nodes and a radius of 0.32, where the patches are colored according to their eigenvector entries.

5.8 Discussion

In this chapter I applied the generalized modeling approach to study a small common cross-feeding motif of bacterial systems. Specifying only the general structure of the system, with this simple model we were able to explore a wide range of ensembles of models with different parameter values. This includes qualitatively different cross-feeding systems, e.g. from the release of cheap to the production of costly metabolites and from the bacteria being the major source of the metabolites to the bacteria only contributing to a minority of the metabolites' concentration.

The results highlight the role of the bacteria's cost of metabolite production for the stability of the system. While exchange of cheap metabolites tends to destabilize the system, costly production of metabolites is a strong factor to promote system's stability. Also, systems with a cheap production of one metabolite can be stabilized by a costly production of the other metabolite involved in the cross-feeding relationship. An example from the literature is the interaction between acetate-releasing *E.coli*, where production

is cheap and methionine-releasing *S.enterica*, a process incurring metabolic costs [221].

Mee et al. [205] showed that amino acids with high bio-synthetic costs tend to favour stronger cooperative interactions than amino acids that are less costly. Analysing $> 6,000$ sequenced bacteria from various environments, they also revealed that only few bacteria produce the costliest amino acids [205]. If the bacterial community depends on the production of these costly metabolites by a few species, it could imply that the community is regulated by this costly production, similar to a limiting nutrient in the environment that regulates organism's growth.

From an evolutionary perspective, benefits in exchanging metabolites lie in the division of metabolic labor [227] and in increasing resource efficiency by separating conflicting metabolic pathways [228], creating a fitness advantage for cross-feeding bacteria [229, 230]. This is particularly meaningful when metabolites are expensive to produce, such as in cases where the metabolic machinery incurs significant costs. This suggests that from an evolutionary perspective, there is a selection for cross-feeding of costly metabolites. Also, bidirectional costly cross-feeding can emerge from initially costless exchange [231].

Another factor influencing stability in our model is the ratio of the fraction of uptake of trade metabolites by bacteria to the production of trade metabolites by bacteria. Ratios > 1 promote stability in our model, probably by weakening the positive feedback loop as bacteria take up more of the metabolite they use for metabolism than they contribute to the metabolite of the partner. In real world systems, the availability of external nutrient supply to the systems can actually change the mode of interaction, e.g. from mutualism to competition and to competitive exclusion [219]. In their study, Hoek et al. [219] showed that low nutrient concentrations favored cooperation as the partners were more dependent on each other. This comparison underscores the significance of recognizing that bacteria engage in not just a single interaction but a myriad of interactions, which can influence each other and exhibit variations across conditions, space, and time [232].

Investigating the way in which instabilities occur, we observed a consistent pattern where one bacterial species consistently exhibited stronger reactions than its trade partner, indicating that one species might take control over the system by becoming more abundant. Due to differences in external nutrient supply rates, uptake rates by bacteria and metabolic costs, the instability is not trivial, i.e. both species decrease or increase the same, instead we see a change in the same direction, but with different magnitudes. This result highlights the role of the different parameters on the dynamics of the system.

Previous models on mutualistic relationships primarily predicted destabilizing effects of these interactions on the system, given their inherent nature as positive feedback

loops [50, 70]. More recent approaches, explicitly incorporating resource dynamics, show that mutualistic interactions are unstable unless specific criteria are met, such as weak interaction strengths or exact reciprocation by the partner [54, 55]. In contrast, our results demonstrate that metabolic costs associated to the production of trade metabolites contribute to the stabilization of these mutualistic cross-feeding interactions. In the previous section, we also showed that the exchange of costly metabolites may be evolutionary attractive.

In addition to identifying stabilizing and destabilizing conditions for the stability of steady states in this cross-feeding system, we explored the response of the system components to perturbations. In general, we found that changes in bacterial density had a higher influence on the entire system and all types of perturbations investigated had a stronger effect on the bacteria compared to the metabolites' concentration. These findings could have important implications for example for the manipulation of bacterial systems, e.g. through pro- or prebiotics [24].

Furthermore, we extended our analysis to a spatial system of homogeneous patches. Whereas a simple bacterial motif can be stable on its own, the same motif can lead to pattern formation in space by diffusion of metabolites. On a network, the homogeneous state becomes more prone to instability when the diffusive coupling is stronger, since this extends the Laplacian spectrum [120]. Even though in our model this instability did not occur in continuous space, this kind of analysis could be very interesting when investigating future bacterial models. Due to these cross-feeding motifs in bacteria, bacteria may form large trade networks over space, making metabolism an attribute of the community instead of the single cell [37]. Pattern formations play an important role in these networks of cooperating organisms [233], their survival [234], evolution [235] and community properties [236].

In summary, the formulation of a simple generalized model of bacterial mutualism generated a multitude of insights regarding the dynamics of the system, the conditions for the stability of steady states, the response to perturbations, susceptible components of the system and the system's behavior over spatial scales. This is the strength of GM, being able to analyze a multitude of realistic functional forms and parameter ranges in a short time frame, we can generate a quick intuition for the system and its key features. In addition, owing to the simplicity of the model we can explore the system using pen and paper math, enabling a deep understanding of the underlying mechanism driving specific outcomes.

Chapter 6

Towards microbial community ecology

In the previous chapters, we have seen that diffusion maps identify key variables in high-dimensional datasets, while generalized modeling allows for the formulation of models encompassing uncertainties in microbial systems. The paper on applying diffusion maps to reconstruct bacterial niche occupancy dynamics is published in Massing et al. [135], while the work on the generalized model of bacterial mutualism is in preparation. My hope for the future is the integration of these two approaches to devise a new way of modeling microbial communities and thereby develop a mechanistic understanding of community composition, dynamics and the relations to ecosystem functioning. In this chapter I want to outline some routes towards achieving this goal.

6.1 Combining datasets

While we have demonstrated the application of diffusion maps on a single large dataset in Chapt. 4, it is also possible to merge multiple datasets for the analysis [237]. This aggregation of datasets is generally a nontrivial task, as variations in the observation of taxa and taxonomic names may arise due to discrepancies in the equipment, personnel, and procedures employed by different sampling teams. Consequently, the absence of e.g. a species in one of the datasets may indicate either its actual absence or that the species was not identified or assigned a different name [238, 239].

Hence, when aggregating different datasets, we cannot just merge two datasets into one and follow the diffusion map analysis. Instead, we need to first derive and threshold the similarity matrices for both datasets independently. As a next step, we merge these similarity matrices by considering all pairs of the units that we are comparing, for instance species. We apply the following rules [237]:

- If a species is identified in both datasets, we average the values of their similarities.

- If a species is only identified in one of two datasets, we take the values from this matrix.
- If one species exists only in one matrix, whereas the other exists only in the other matrix, we set the similarity to zero.

We show in Carrasco de la Cruz et al. [237] that this approach works well for merging two datasets of phytoplankton communities. Since the characteristics of phytoplankton and bacterial data are similar, e.g. high diversity of species, different analysis procedures can differ in their results etc. we can also apply this approach in the future to combine datasets of bacteria. Merging datasets, hence including a higher number of observations, should increase the accuracy in the diffusion map results [44, 47] as well as downstream analyses such as the functional diversity estimation [49]. In addition, dataset aggregation on a large scale could enable comparisons of bacterial communities from very different habitats in their key variables [44] and their functional diversity [237].

6.2 Community maps

In Chapt. 4 we have applied the diffusion map approach to identify key metabolic strategies of bacterial species. To get a grasp of community dynamics, a crucial step is the identification of system variables that describe the major dimensions along which the functional community changes. To illustrate the idea of system variables, let us draw an analogy to the global job market. Similar to the diversity of jobs in different places, one could imagine the functional capabilities of the different bacteria in a community. The dimensionality of the jobs is high since they are characterized by numerous features that vary among them, similar to the functional capabilities of bacteria. However, when considering for instance the dimensionality of industries across different locations, one might anticipate a comparatively smaller dimensionality. For example, a specific type of industry comprises a diverse but also quite distinct set of jobs. In the context of functional microbial communities, we could therefore ask, what we would expect to see when we observe a community over time. To how many dimensions could we map the change of the functional community in time? Our idea is to use diffusion maps as manifold learning method to find the dimensionality of the functional community composition.

To be able to compare different functional communities to each other, we first need to derive a new distance metric. The functional diversity measure by Ryabov et al. [49] allows us to quantify the functional diversity of single samples. However, to employ the diffusion map as embedding method, we need a certain understanding of similarity between these single communities. Here, we therefore take the functional diversity measure as a basis to

derive a distance measure that compares different communities in terms of their species abundances and positions in the functional coordinate system.

Reconstructing the trait space through diffusion mapping the species co-occurrences, we derive the positions of the species in the trait space and their distances to each other. The distance between two species i, j is

$$d_{i,j} = \sqrt{\sum_b \left(\frac{v_{b,i} - v_{b,j}}{\lambda_b} \right)}, \tag{6.1}$$

where $v_{b,i}$ is the entry of eigenvector (i.e. trait) b for species i, $v_{b,j}$ is the entry of trait b for species j and λ_b is the corresponding eigenvalue.

Next, we employ these distances in the reconstructed trait space in combination with the abundances of the species in the samples to calculate the functional diversities of each sample as Rao index following the procedure explained in Sec. 2.3.3. The functional diversity (FD) for sample k is computed as

$$FD_k = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{i,j} p_k^{(i)} p_k^{(j)}, \tag{6.2}$$

where $p_k^{(i)} = a_k^{(i)} / \sum_j a_k^{(j)}$ is the relative biomass of species i in sample k [49]. The Rao index represents the weighted average functional distance between all pairs of species in each sample.

Then, we calculate the functional diversity of the sample pair by taking the mean abundances of the species in the two samples into account

$$FD_{kl} \;=\; \sum_{i_1}^{n-1} \sum_{j=i+1}^{n} d_{i,j} \left(\frac{p_k^{(i)} + p_l^{(i)}}{2} \right) \left(\frac{p_k^{(j)} + p_l^{(j)}}{2} \right). \tag{6.3}$$

The result is a weighted average functional distance between all pairs of species i and j, where the weighting factor is the mean probability of the two samples that one of two randomly selected individuals belongs to species i and the other to species j.

Finally, we subtract the mean functional diversities of the two samples, here k and l, from the FD of the sample pair, FD_{kl},

$$FDist_{kl} \;=\; FD_{kl} - \left(\frac{FD_k + FD_l}{2} \right). \tag{6.4}$$

This functional distinctness measure allows us to establish a notion of similarity between

two samples in terms of their difference in functional community composition. The functional distinctness of a sample pair equals zero if the communities are identical in species composition and abundances and increases with communities being more different to each other.

To test our procedure for deriving system variables, we generated synthetic communities by sampling from generated bivariate species response surfaces to two environmental variables. Following the procedure of Minchin [240], also applied in Gault et al. [46] we proceed as follows:

1. We randomly select a response surface for each of a total of 1,000 species modeled as bivariate Gaussian dome with random mean, variance and covariance in relation to two environmental variables. In ecological context the mean equals the environmental optimal conditions for the respective species, hence the niche centroid. The shape of the Gaussian dome, i.e. the squared length of the principal axes and the orientation, namely the rotation angle is controlled by the variance and covariance values. We adjust the peak height of each Gaussian dome by the standard prefactor of a multivariate normal distribution, i.e. Gaussian density. This results in narrower domes having higher peaks than broader domes and simulates the trade-off between specialist and generalist species.

2. We choose 100 sampling locations along the environmental gradients. We employ a regular 10x10 grid, which is a simplification, since real-world environmental factors are often correlated and seasonally variant. Reinterpreting the two environmental factors as coordinates along the first two principal axes derived from a PCA of a higher-dimensional environmental space helps to reduce this discrepancy.

3. At each sampling location, we evaluate the joint probability density function for each species and scale these values by a sampling "space" to obtain the expected number of individuals of each species in the sampling space. We then take the expected values of the number of individuals for each species and generate random numbers, i.e. Poisson variates to represent the actual observed numbers of individuals in a single sample. Since the sampling space is constant, a smaller number of species will be sampled when species densities are low.

In detail, the bivariate Gaussian response surface is defined in a way that the expected number of individuals n is a function of $\boldsymbol{x}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, i.e.

$$n(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}} \exp[-(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})]. \tag{6.5}$$

$|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$, $\boldsymbol{x}$ denotes a vector that specifies the locations of the sample along the environmental gradients

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \tag{6.6}$$

$\boldsymbol{\mu}$ is the vector defining the location of the niche centroid along the environmental gradients

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \tag{6.7}$$

and $\mathbf{\Sigma}$ denotes the symmetric variance-covariance matrix with respect to the environmental gradients x_1 and x_2

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}. \tag{6.8}$$

For a non-degenerate covariance ellipsoid with $|\mathbf{\Sigma}| > 0$ the inverse $\mathbf{\Sigma}^{-1}$ exists. Via a principal axis transformation

$$\boldsymbol{x} - \boldsymbol{\mu} = \mathbf{R}\boldsymbol{y} \quad \Leftrightarrow \quad \boldsymbol{y} = \mathbf{R}^T(\boldsymbol{x} - \boldsymbol{\mu}) \tag{6.9}$$

where $\mathbf{R}$ is the orthogonal matrix and its two columns contain the two orthonormal eigenvectors of the matrix $\mathbf{\Sigma}$, we can show that

$$(\boldsymbol{x} - \boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = \boldsymbol{y}^T\mathbf{R}^T\mathbf{\Sigma}^{-1}\mathbf{R}\boldsymbol{y} = \boldsymbol{y}^T\mathbf{\Lambda}^{-1}\boldsymbol{y}, \tag{6.10}$$

where the diagonal matrix $\mathbf{\Lambda}$ is

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1^2 & 0 \\ 0 & \lambda_2^2 \end{pmatrix}, \tag{6.11}$$

with λ_1^2 and λ_2^2 being related eigenvalues of $\mathbf{\Sigma}$. Hence, the matrices $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$ are connected trough the relations

$$\mathbf{\Lambda} = \mathbf{R}^T\mathbf{\Sigma}\mathbf{R} \quad \Leftrightarrow \quad \mathbf{\Sigma} = \mathbf{R}\mathbf{\Lambda}\mathbf{R}^T. \tag{6.12}$$

Due to orthonormality $\mathbf{R}$ is a rotation matrix

$$\boldsymbol{R} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{6.13}$$

and $|\mathbf{R}| = |\mathbf{R}^T| = 1$, which leads to $|\mathbf{\Sigma}| = |\mathbf{\Lambda}|$. Consequently, we can write

$$\mathbf{\Sigma} = \begin{pmatrix} \lambda_1^2 \cos^2\theta + \lambda_2^2 \sin^2\theta & \lambda_1^2 \cos^2\theta \sin\theta - \lambda_2^2 \cos\theta \sin\theta \\ \lambda_1^2 \cos\theta \sin\theta - \lambda_2^2 \cos\theta \sin\theta & \lambda_1^2 \sin^2\theta + \lambda_2^2 \cos^2\theta \end{pmatrix}. \tag{6.14}$$

Thus, selecting values for the two scaling factors λ_1^2 and λ_2^2 and for the angle of rotation θ suffices to generate the variance-covariance matrix $\mathbf{\Sigma}$. Therefore, to define each species' response surface we only need to specify the niche centroid $\boldsymbol{\mu}$, the length of the principal axes λ_1 and λ_2 and the angle of rotation θ.

Assuming that our environmental variables are restricted to values between 0 and 1, we drew the values of $\boldsymbol{\mu}$ from a uniform distribution between 0 and 1. The values for the angle of rotation were generated by randomly drawing the value for $\cos(\theta)$ from a uniform distribution spanning the interval from -1 to 1 and the value for $\sin(\theta)$ as $\pm\sqrt{1 - \cos^2(\theta)}$, while randomly choosing the sign. For the simulation we distinguished two scenarios, following Gault et al. [46]. For the high turnover scenario, we selected the values of λ_1^2 and λ_2^2 from a normal distribution, constrained within 0.005-0.01, while for the low turnover scenario we chose them from a normal distribution between 0.01-0.09.

Diffusion mapping the functional distinctness values of the 100 generated synthetic communities successfully recovers the two main dimensions of variation in the synthetic communities of the low and high turnover scenarios (Fig. 6.1).

Moving forward, we additionally evaluated the approach using real-world datasets. First, we applied the approach to Dutch and German phytoplankton data from the Southern North Sea, aggregated and diffusion mapped in Carrasco et al. [237]. The diffusion map of functional distinctness values identifies the single-species blooms as the most important variables in the dataset (Fig. 6.2). While this analysis successfully identifies important features, specifically phytoplankton blooms, it falls short of achieving our primary goal of extracting community traits from the dataset.

The application of this method to the bacterial data from the Baltic Sea that we have focused on in Chapt. 4, reveals one important variable that is mainly driven by species appearing in very high relative abundances in the samples (Fig. 6.3). The higher variables seem to be harmonic eigenvectors of the first.

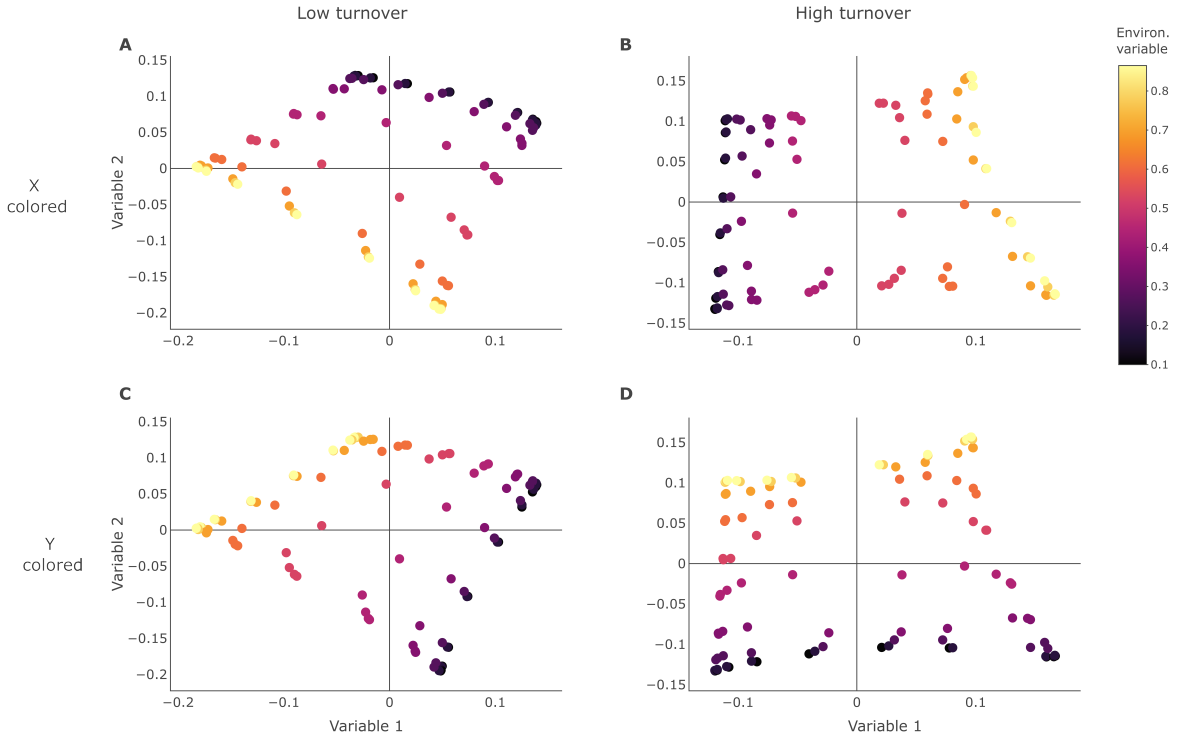

Figure 6.1: Diffusion maps of functional distinctness values for low (A,C) and high turnover scenarios. Values of environmental variables (Environ. variable) X (A,B) and Y (C,D) are color-coded.

.

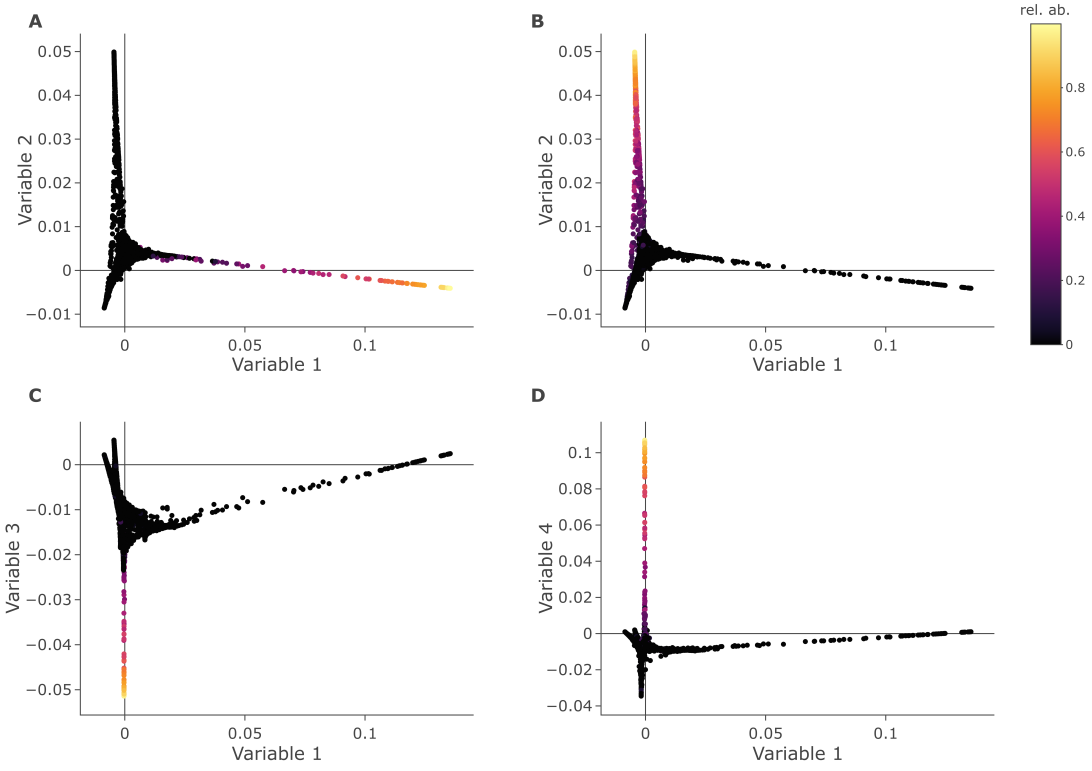

Figure 6.2: Diffusion map results of functional distinctness values of phytoplankton data from the Southern North Sea [237]. Values of relative abundances (rel. ab.) of the bloom species, i.e. *Phaeocystis globosa* (A), *Micromonas pusilla* (B) and *Planktothrix agardhii* (C, D) are color-coded.

.

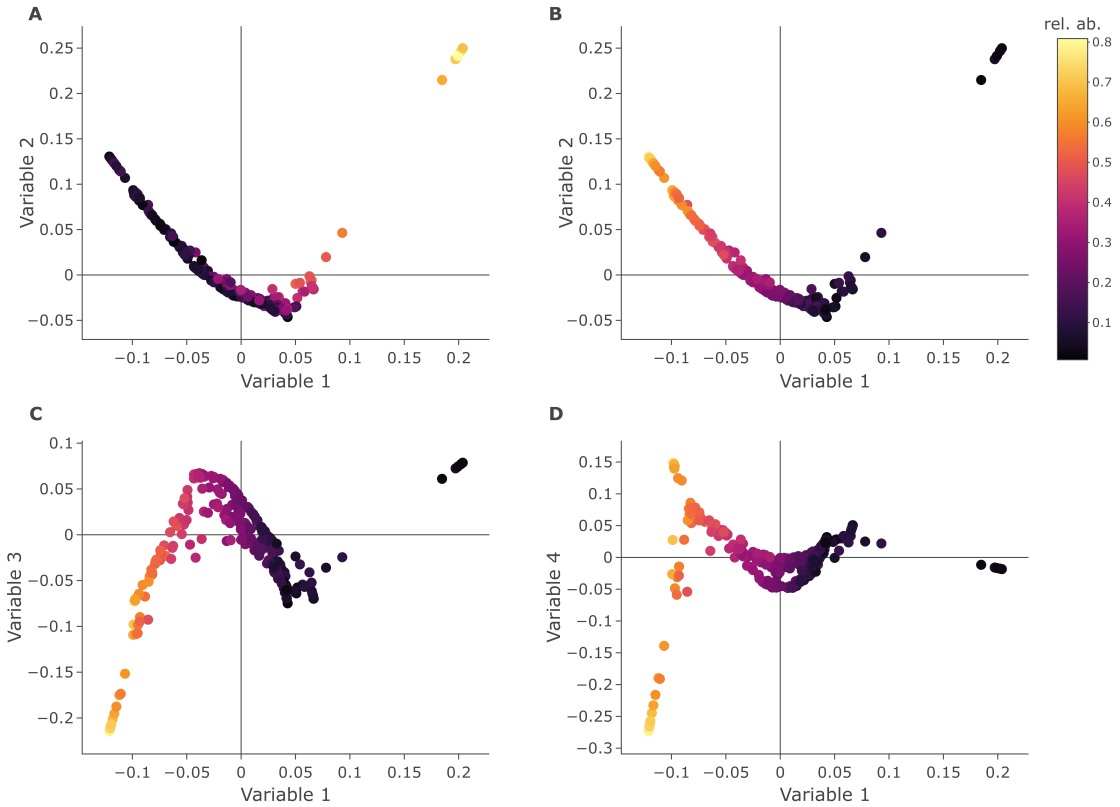

Figure 6.3: Diffusion map results of functional distinctness values of bacterial data from the Baltic Sea [135]. Relative abundances (rel. ab.) of species belonging to the Class Bacteroidia (A) and the Class Cyanobacteria (B, C, D) are color-coded.

.

From these results, we can see that the functional distinctness measure is strongly driven by variations in abundances, which is problematic if our data shows large variation in relative abundances. Since this is a prevalent feature in microbial communities, a future direction is to derive an alternative similarity measure that can handle such data characteristics.

6.3 Synthetic metabolic interaction networks

We have seen that the prevalence of metabolite-based interactions in bacteria contributes to shaping community composition, ecosystem function and stability [206, 207, 211]. An intriguing question is if we can derive overarching principles governing bacterial metabolic interactions. For example, a rule for how many species can co-exist on a single substrate or a principle specifying the most effective metabolic strategy based on a given interaction structure. More generally, we could explore how the structure of metabolic reactions corresponds to co-existence, i.e. diversity of bacteria and their metabolic strategies.

Mentges et al. [241] showed that a simple mechanistic model of microbe-substrate interactions already provides interesting insights. In their modeling approach they represent microbe-substrate interactions as uptake and release matrices, that are governed by simple rules that ignore the specific chemistry of reactions. Our hope is to expand this approach by integrating insights from chemical reaction systems, such as the conservation of the number of atoms in reactions, as well as drawing on principles from thermodynamics, e.g. the concept of increasing entropy. One route towards this aim is to create an artificial chemistry that follows general chemical principles such as the stoichiometric balance. An approach involves devising an algorithm capable of predicting all potential reactions based on a given set of chemical compounds and their elemental composition.

An an example, let us consider that we have the four chemical compounds methane (CH_4), oxygen (O_2), carbon dioxide (CO_2) and water (H_2O) and we want to derive possible chemical reactions between these that obey stoichiometric balance. We know the elemental composition of the compounds (Table 6.1).

Table 6.1: Number of elements in chemical compounds

	Compounds			
Elements	Methane	Oxygen	Carbon dioxide	Water
C	1	0	1	0
H	4	0	0	2
O	0	2	2	1

Writing the elemental composition in matrix notation yields

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 4 & 0 & 0 & 2 \\ 0 & 2 & 2 & 1 \end{pmatrix}. \tag{6.15}$$

Now we want to find all possible chemical reactions that can take place between these molecules, while ensuring stoichiometric balance. To achieve this, let us define a column vector of educts, denoted as $\boldsymbol{v}$, and a column vector of products, denoted as $\boldsymbol{w}$, with $\boldsymbol{v}, \boldsymbol{w} \in \{(n_{\mathrm{CH_4}}, n_{\mathrm{O_2}}, n_{\mathrm{CO_2}}, n_{\mathrm{H_2O}})\}$, where $n_{\mathrm{CH_4}}, n_{\mathrm{O_2}}, n_{\mathrm{CO_2}}, n_{\mathrm{H_2O}}$ are the numbers of methane, oxygen, carbon dioxide and water molecules. Our objective is to determine vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ that satisfy

$$\mathbf{M}\boldsymbol{v} \;=\; \mathbf{M}\boldsymbol{w}. \tag{6.16}$$

This equation ensures stoichiometric balance, since the matrix $\mathbf{M}$ converts the chemical compounds into their atom numbers. Rewriting Eq. 6.16, we obtain

$$\mathbf{M}(\boldsymbol{v} - \boldsymbol{w}) \quad = \quad \mathbf{0}. \tag{6.17}$$

Hence, we see that the vector $(\boldsymbol{v} - \boldsymbol{w})$ has to be (part of) the nullspace of the matrix $\mathbf{M}$. The nullspace (or kernel) of a matrix represents the set of all vectors that, when multiplied by the matrix, result in the zero vector.

For our example system, we can show that

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 4 & 0 & 0 & 2 \\ 0 & 2 & 2 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ -1 \\ -2 \end{pmatrix} \quad = \quad \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{6.18}$$

Translating into possible chemical reactions yields

$$\mathrm{CH_4} + 2\,\mathrm{O_2} \quad \xrightarrow{\mathrm{r}} \quad \mathrm{CO_2} + 2\,\mathrm{H_2O} \tag{6.19}$$

$$\mathrm{CO_2} + 2\,\mathrm{H_2O} \quad \xrightarrow{\mathrm{s}} \quad \mathrm{CH_4} + 2\,\mathrm{O_2}, \tag{6.20}$$

where r and s represent the reaction rates. Therefore, determining the nullspace of the matrix reveals potential chemical reactions among the chemical compounds, ensuring stoichiometric balance. Using mass action laws we can write the following ODE system

$$[\dot{\mathrm{CH_4}}] \quad = \quad s[\mathrm{CO_2}][\mathrm{H_2O}]^2 - \mathrm{r}[\mathrm{CH_4}][\mathrm{O_2}]^2 \tag{6.21}$$

$$[\dot{\mathrm{O_2}}] \quad = \quad 2s[\mathrm{CO_2}][\mathrm{H_2O}]^2 - 2\mathrm{r}[\mathrm{CH_4}][\mathrm{O_2}]^2 \tag{6.22}$$

$$[\dot{\mathrm{CO_2}}] \quad = \quad r[\mathrm{CH_4}][\mathrm{O_2}]^2 - \mathrm{s}[\mathrm{CO_2}][\mathrm{H_2O}]^2 \tag{6.23}$$

$$[\dot{\mathrm{H_2O}}] \quad = \quad 2r[\mathrm{CH_4}][\mathrm{O_2}]^2 - 2\mathrm{s}[\mathrm{CO_2}][\mathrm{H_2O}]^2, \tag{6.24}$$

where the square brackets indicate the concentrations of the chemical compounds.

The next step could be the development of a method to generate a potential interaction network among bacteria from the possible chemical reactions. Our hope is to create synthetic metabolic interaction networks automatically from given random matrices that capture the element composition of certain chemical compounds. Here, we showed a first step towards such a model, i.e. identifying possible chemical reactions from a matrix of chemical compounds and their elemental composition.

Being able to create synthetic metabolic interaction networks from simple elemental com-

position matrices could enable us to build realistic topologies of complex interaction networks of bacteria. This approach could be considered a niche model for bacteria, akin to the niche model by Williams and Martinez [242] that yields complex food webs based on a simple set of rules. Such a model has the potential to unveil the mechanisms behind the structure of bacterial communities, addressing also long-standing mysteries such as the enormous diversity of bacteria as a whole and of rare species in particular [243].

6.4 Summary

In this chapter I outlined some future directions towards developing a mechanistic understanding of microbial community ecology. I showed that the increasing availability of large datasets can be used for an aggregated diffusion map analysis that improves the overall results and could eventually provide a global map of communities, their traits and functional diversities. Also, I explored the concept of community maps as an approach to derive the variables along which the functional community composition changes. Finally, I discussed synthetic metabolic interaction networks that link the chemistry of reactions to bacterial substrate interactions and could serve as a niche model for microbes.

Chapter 7

Conclusions

In this work, I explored novel approaches for the analysis and modeling of the dynamics of complex ecological communities with a focus on bacterial communities. These approaches have enabled us to make sense of large datasets of bacterial communities on the one hand and to study the dynamics of interacting bacteria with a high degree of generality on the other hand. I have analyzed a large dataset of bacterial community composition from the Baltic Sea, coarse-graining the complex bacterial communities into potentially occupied metabolic niches over time. Subsequently, I have focused on the question how a mutualistic cross-feeding relationship between two bacterial species that inherently constitutes a positive feedback loop can be stabilized. Let us now discuss the results of this work and outline some future research directions.

Diffusion map approach to make sense of large datasets

In Chapt. 4 we have shown that the diffusion map approach enables us to make sense of high-dimensional datasets of bacterial communities by coarse-graining the over 4,000 bacterial taxa in terms of their metabolic strategies. The identification of these metabolic strategies allows us to translate the species time series into potentially occupied metabolic niches over time. Not only does this approach achieve a dimensionality reduction in terms of the most important variables, it also provides a framework to connect the data to an important theoretical concept from ecology, i.e. the niche. These conceptual frameworks are important to advance our understanding of bacterial communities alongside the production of data on these [14]. Findings through these frameworks can help the formulation of hypotheses-driven research, enabling the development and testing of general principles [14].

The diffusion map approach of the Baltic Sea dataset identified a wealth of metabolic strategies, representing localized as well as continuous strategies, strategies that group related taxa together and those that distinguish them, forming together the metabolic strategy space of the community. Our results also highlight the power of the diffusion

map to objectively detect biases in the dataset. If there are data points that are isolated from the rest, the diffusion map identifies those separations. Organizing the genomic information into potentially occupied metabolic niches over times allowed us to observe the metabolic niche occupation dynamics over time. These revealed the driving taxa of these strategies as well as seasonal and environmental impacts on the dynamics.

One of the limiting factors of diffusion maps is the data that we have available on the system we want to investigate. One factor to consider is that we need a certain amount of data to ensure the applicability of the diffusion map, while another aspect is that the analysis is limited by the type of data. The diffusion map can be applied to almost any dataset where we have a certain understanding of similarity between the data points. The data usually consists of a tabular format that records a set of properties for a specific number of samples. Because the insights by diffusion mapping diminish if the diffusion map network diameter remains low after thresholding, this method should not be applied to datasets with less than ca. 30 samples. In the case of very large datasets, the diffusion map is able to even analyze for example millions of samples in a comparably short time frame. The limiting factor lies in the sample comparisons, which scales as $O(N^2)$. In the case of exceptionally large data sets, complexity can be mitigated by employing heuristics that decrease the number of necessary comparisons.

The type of data plays a major role in the insights we can gain using diffusion maps, especially as bacteria show genetic diversity for example among individuals of the same species [201, 244, 245]. In our case, assigning amplicon sequence variants data to species and obtaining the complete genomes from databases is a rather crude tool, that cannot fully capture the whole genomic diversity and possible adaptations specific to the bacteria in those communities. Also, when working with genomic data, we can only gain insights into potentially employed strategies. To investigate which strategies are actually applied by the bacteria, we would need other types of data, for example transcriptomic data.

Thanks to technological progress and ongoing long-term monitoring efforts, we are witnessing an increasing availability of extensive datasets that meet the prerequisites for the diffusion map method. This expansion for example also encompasses datasets of metagenomes and transcriptomic data, that could create a more detailed picture of the specific metabolic strategies employed in the respective system under specific conditions.

Diffusion maps provide a ranking of the importance of the detected variables, however they do not provide us with an interpretation of these variables. Therefore, it is essential to possess some knowledge of the system, and additional information can prove valuable in formulating these interpretations. As a result, the process of interpreting the new

variables can be time-consuming.

Overall, we believe that the diffusion map approach has great potential to help us make sense of the increasing wealth of datasets of ecological communities, especially bacterial communities. The coarse-graining of species into their metabolic niches provides a first step to bridge the current gap between data and theory. Our approach to translate the species time series into strategy time series provides new insights into the dynamics of these communities and possible relation to environmental factors as well as ecosystem functioning. The results illustrate the power of manifold learning approaches to advance our understanding of the links between community composition and functioning.

Generalized modeling to gain insights into uncertain systems

In Chapt. 3 we have reviewed the generalized modeling (GM) approach and provided a hands-on guide to apply it to various kinds of systems. The many insights gained through GM in the past demonstrate the power of this approach to extract valuable information from a wide range of systems [93]. GM is especially useful for systems where we have limited information available and hence are uncertain about specific functional forms and precise rate laws.

In Chapt. 4 we have applied the GM approach to gain insights into a common motif in bacterial communities, i.e. mutual cross-feeding between two types of bacteria. The GM approach of specifying the structure but not the specific functional forms has enabled us to explore a whole range of plausible cross-feeding scenarios, while making few assumptions. We not only acquired efficient numerical results through the analysis of 10^7 parameter sets but also derived analytical findings for this relatively complex four-variable model. Moreover, the general model allowed for the analysis of the response to various perturbations and the behaviour of multiple motifs within a spatial context.

While GM enables efficient analysis, the applicable tools are limited, constraining the analysis to local dynamics close to equilibrium [246]. It is thus not useful to study behaviours of systems that are far from equilibrium or to explore transient behaviors. Also, the results of GM cannot be verified by data directly. However, bifurcations of specific models can be compared to bifurcations in general models, that describe those specific models [103].

Current modeling approaches of microbial communities have provided valuable insights into the dynamics of microbial communities [54, 55, 56, 57, 64, 70, 247]. However, these approaches are often limited either by ignoring features that characterize microbial sys-

tems [70], like metabolite-mediated interactions, or by dealing with a large number of parameters [64, 247], which often restricts the analysis to simulations. In other cases the models only explore a small range of plausible models [54, 55, 56], e.g. assuming the release of costless metabolites or linear mortalities. GM is different in that it can integrate structural information, like the interaction through metabolite exchange, while incorporating the uncertainty about the system in easily interpretable parameters. Thereby, GM allows us to explore the whole range of plausible models and study their dynamical implications.

Overall, we believe that GM provides a good addition to current modeling approaches of microbial communities. GM can deal to a certain extent with the complexity of microbial communities and their interactions by being able to analyze a whole ensemble of different models in a relatively short time frame. We have also shown that we can integrate new information that we gain on the system iteratively into a generalized model, allowing us to refine the model efficiently. Moreover, to enable simulations, we can construct conventional models from generalized models, allowing us to study specific parameter regions in more detail. Future investigations may reveal additional factors that impact stability of mutualistic interactions between microbes, e.g. removal of inhibitory molecules could play an important role [248]. GM could also be used to compare conventional models with different dynamics [232], as the general parameters span a natural coordinate system that can be used to compare specific models [103]. It would also be interesting to extend our analysis to interactions of > 2 bacterial species and multiple metabolites or different types of interactions, including for example competition for resources [56].

Towards a better understanding of the dynamics of complex ecological communities

Our aim in this work has been the development of new tools and frameworks to deal with the complexity of ecological community dynamics, especially among microbes. We have worked on extending both frontiers, the frontier of analyzing high-dimensional datasets and the frontier of incorporating the many uncertainties that exist in complex communities. Our vision for the future is the combination of the two approaches, diffusion maps and generalized modeling.

We have seen that diffusion maps detect important metabolic strategies of the bacterial species. As mentioned above we envision that this method can also be used to identify system variables that describe the major dimensions along which the functional community changes. These system variables could then be used in a GM framework to explore the dynamics of these communities, their stability properties and to study their response

to perturbations, like a change in temperature or a decrease in oxygen concentrations. Our hope is therefore to combine diffusion maps and the generalized modeling framework to develop a new way of modeling bacterial communities and thereby develop an understanding of the mechanisms behind community composition, dynamics over time and the relation to ecosystem functioning.

Bibliography

[1] May, R. M. Tropical arthropod species, more or less? *Science* **329**, 41–42 (2010).

[2] Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. & Worm, B. How many species are there on earth and in the ocean? *PLoS biology* **9**, e1001127 (2011).

[3] Hutchinson, G. E. Homage to santa rosalia or why are there so many kinds of animals? *The American Naturalist* **93**, 145–159 (1959).

[4] Hutchinson, G. E. The paradox of the plankton. *The American Naturalist* **95**, 137–145 (1961).

[5] Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences* **113**, 5970–5975 (2016).

[6] Zimmerman, N., Izard, J., Klatt, C., Zhou, J. & Aronson, E. The unseen world: environmental microbial sequencing and identification methods for ecologists. *Frontiers in Ecology and the Environment* **12**, 224–231 (2014).

[7] Hug, L. A. *et al.* A new view of the tree of life. *Nature microbiology* **1**, 1–6 (2016).

[8] Darwin, C. *On the Origin of Species by Means of Natural Selection* (Murray, 1859).

[9] May, R. M. Will a large complex system be stable? *Nature* **238**, 413–414 (1972).

[10] Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive earth's biogeochemical cycles. *science* **320**, 1034–1039 (2008).

[11] Hadfield, M. G. Biofilms and marine invertebrate larvae: what bacteria produce that larvae use to choose settlement sites. *Annual review of marine science* **3**, 453–470 (2011).

[12] Bourne, D. G., Morrow, K. M. & Webster, N. S. Insights into the coral microbiome: underpinning the health and resilience of reef ecosystems. *Annual Review of Microbiology* **70**, 317–340 (2016).

[13] Weiland-Bräuer, N. Friends or foes—microbial interactions in nature. *Biology* **10**, 496 (2021).

[14] Prosser, J. I. & Martiny, J. B. Conceptual challenges in microbial community ecology (2020).

[15] Cavicchioli, R. *et al.* Scientists' warning to humanity: microorganisms and climate change. *Nat. Rev. Microbiol.* **17**, 569–586 (2019).

[16] Bernhardt, J. R. *et al.* The evolution of competitive ability for essential resources. *Philosophical Transactions of the Royal Society B* **375**, 20190247 (2020).

[17] Arnold, B. J., Huang, I.-T. & Hanage, W. P. Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews Microbiology* **20**, 206–218 (2022).

[18] Lederberg, J. & McCray, A. T. Ome sweetomics–a genealogical treasury of words. *The scientist* **15**, 8–8 (2001).

[19] Bosch, T. C. & McFall-Ngai, M. J. Metaorganisms as the new frontier. *Zoology* **114**, 185–190 (2011).

[20] Bharti, R. & Grimm, D. G. Current challenges and best-practice protocols for microbiome analysis. *Briefings in bioinformatics* **22**, 178–193 (2021).

[21] Buttigieg, P. L. *et al.* Marine microbes in 4d—using time series observation to assess the dynamics of the ocean microbiome and its links to ocean health. *Current opinion in microbiology* **43**, 169–185 (2018).

[22] Fuhrman, J. A., Cram, J. A. & Needham, D. M. Marine microbial community dynamics and their ecological interpretation. *Nature Reviews Microbiology* **13**, 133–146 (2015).

[23] Bunse, C. & Pinhassi, J. Marine bacterioplankton seasonal succession dynamics. *Trends in microbiology* **25**, 494–505 (2017).

[24] Faust, K., Lahti, L., Gonze, D., De Vos, W. M. & Raes, J. Metagenomics meets time series analysis: unraveling microbial community dynamics. *Current opinion in microbiology* **25**, 56–66 (2015).

[25] Thompson, L. R. *et al.* A communal catalogue reveals earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).

[26] Sunagawa, S. *et al.* Tara oceans: towards global ocean ecosystems biology. *Nature Reviews Microbiology* **18**, 428–445 (2020).

[27] Griffiths, R. I. *et al.* The bacterial biogeography of british soils. *Environmental microbiology* **13**, 1642–1654 (2011).

[28] Cui, J. *et al.* Nutrient availability is a dominant predictor of soil bacterial and fungal community composition after nitrogen addition in subtropical acidic forests. *PLoS One* **16**, e0246263 (2021).

[29] Waldrop, M. & Firestone, M. Response of microbial community composition and function to soil climate change. *Microbial ecology* **52**, 716–724 (2006).

[30] Auffret, M. D. *et al.* The role of microbial community composition in controlling soil respiration responses to temperature. *PLoS One* **11**, e0165448 (2016).

[31] Glassman, S. I. *et al.* Decomposition responses to climate depend on microbial community composition. *Proceedings of the National Academy of Sciences* **115**, 11994–11999 (2018).

[32] Galand, P. E., Pereira, O., Hochart, C., Auguet, J. C. & Debroas, D. A strong link between marine microbial community composition and function challenges the idea of functional redundancy. *The ISME journal* **12**, 2470–2478 (2018).

[33] Goldford, J. E. *et al.* Emergent simplicity in microbial community assembly. *Science* **361**, 469–474 (2018).

[34] Needham, D. M. & Fuhrman, J. A. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nature microbiology* **1**, 1–7 (2016).

[35] Granato, E. T., Meiller-Legrand, T. A. & Foster, K. R. The evolution and ecology of bacterial warfare. *Current biology* **29**, R521–R537 (2019).

[36] Friedman, J., Higgins, L. M. & Gore, J. Community structure follows simple assembly rules in microbial microcosms. *Nature ecology & evolution* **1**, 0109 (2017).

[37] D'Souza, G. *et al.* Ecology and evolution of metabolic cross-feeding interactions in bacteria. *Natural Product Reports* **35**, 455–488 (2018).

[38] Qian, Y., Lan, F. & Venturelli, O. S. Towards a deeper understanding of microbial communities: integrating experimental data with dynamic models. *Current opinion in microbiology* **62**, 84–92 (2021).

[39] Dal Bello, M., Lee, H., Goyal, A. & Gore, J. Resource–diversity relationships in bacterial communities reflect the network structure of microbial metabolism. *Nature Ecology & Evolution* **5**, 1424–1434 (2021).

[40] Louca, S., Parfrey, L. W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 1272–1277 (2016).

[41] Alneberg, J. *et al.* Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. *Communications biology* **3**, 119 (2020).

[42] Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **374**, 20150202 (2016).

[43] van Voorn, G. A. *et al.* Stabilization due to predator interference: comparison of different analysis approaches. *Mathematical Biosciences and Engineering* **5**, 567–583 (2008).

[44] Fahimipour, A. K. & Gross, T. Mapping the bacterial metabolic niche space. *Nature communications* **11**, 4887 (2020).

[45] Moon, K. R. *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology* **37**, 1482–1492 (2019).

[46] Gault, J. A., Freund, J. A., Hillebrand, H. & Gross, T. Dissimilarity analysis based on diffusion maps. *Oikos* e10249 (2023).

[47] Barter, E. & Gross, T. Manifold cities: social variables of urban areas in the uk. *Proceedings of the Royal Society A* **475**, 20180615 (2019).

[48] Coifman, R. R. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the national academy of sciences* **102**, 7426–7431 (2005).

[49] Ryabov, A., Blasius, B., Hillebrand, H., Olenina, I. & Gross, T. Estimation of functional diversity and species traits from ecological monitoring data. *Proceedings of the National Academy of Sciences* **119**, e2118156119 (2022).

[50] Coyte, K. Z., Schluter, J. & Foster, K. R. The ecology of the microbiome: networks, competition, and stability. *Science* **350**, 663–666 (2015).

[51] MacArthur, R. Species packing, and what competition minimizes. *Proceedings of the National Academy of Sciences* **64**, 1369–1371 (1969).

[52] MacArthur, R. Species packing and competitive equilibrium for many species. *Theoretical population biology* **1**, 1–11 (1970).

[53] Chesson, P. Macarthur's consumer-resource model. *Theoretical Population Biology* **37**, 26–38 (1990).

[54] Butler, S. & O'Dwyer, J. P. Stability criteria for complex microbial communities. *Nature communications* **9**, 2970 (2018).

[55] Butler, S. & O'Dwyer, J. P. Cooperation and stability for complex systems in resource-limited environments. *Theoretical Ecology* **13**, 239–250 (2020).

[56] Muscarella, M. E. & O'Dwyer, J. P. Species dynamics and interactions via metabolically informed consumer-resource models. *Theoretical Ecology* **13**, 503–518 (2020).

[57] van den Berg, N. I. *et al.* Ecological modelling approaches for predicting emergent properties in microbial communities. *Nature ecology & evolution* **6**, 855–865 (2022).

[58] Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J. & Crosby, L. Phylogenetically novel uncultured microbial cells dominate earth microbiomes. *MSystems* **3**, 10–1128 (2018).

[59] Steen, A. D. *et al.* High proportions of bacteria and archaea across most biomes remain uncultured. *The ISME journal* **13**, 3126–3130 (2019).

[60] Holling, C. S. Some characteristics of simple types of predation and parasitism1. *The canadian entomologist* **91**, 385–398 (1959).

[61] Hill, A. V. The combinations of haemoglobin with oxygen and with carbon monoxide. i. *Biochemical Journal* **7**, 471 (1913).

[62] Monod, J. *Recherches sur la croissance des cultures bactériennes* (Hermann et Cie, 1942).

[63] Wade, M. J. *et al.* Perspectives in mathematical modelling for microbial ecology. *Ecological Modelling* **321**, 64–74 (2016).

[64] Marsland III, R. *et al.* Available energy fluxes drive a transition in the diversity, stability, and functional structure of microbial communities. *PLoS computational biology* **15**, e1006793 (2019).

[65] Marsland, R., Cui, W., Goldford, J. & Mehta, P. The community simulator: A python package for microbial ecology. *Plos one* **15**, e0230430 (2020).

[66] Gross, T., Ebenhöh, W. & Feudel, U. Enrichment and foodchain stability: the impact of different forms of predator–prey interaction. *Journal of theoretical biology* **227**, 349–358 (2004).

[67] Barner, A. K., Coblentz, K. E., Hacker, S. D. & Menge, B. A. Fundamental contradictions among observational and experimental estimates of non-trophic species interactions. *Ecology* **99**, 557–566 (2018).

[68] Gross, T. & Feudel, U. Generalized models as a universal approach to the analysis of nonlinear dynamical systems. *Physical Review E* **73**, 016205 (2006).

[69] Aufderheide, H., Rudolf, L., Gross, T. & Lafferty, K. D. How to predict community responses to perturbations in the face of imperfect knowledge and network complexity. *Proceedings of the Royal Society B: Biological Sciences* **280**, 20132355 (2013).

[70] Allesina, S. & Tang, S. Stability criteria for complex ecosystems. *Nature* **483**, 205–208 (2012).

[71] Helling, R. B., Vargas, C. N. & Adams, J. Evolution of escherichia coli during growth in a constant environment. *Genetics* **116**, 349–358 (1987).

[72] Kuznetsov, Y. A., Kuznetsov, I. A. & Kuznetsov, Y. *Elements of applied bifurcation theory*, vol. 112 (Springer, 1998).

[73] Strogatz, S., Friedman, M., Mallinckrodt, A. J. & McKay, S. Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering. *Computers in Physics* **8**, 532–532 (1994).

[74] Guckenheimer, J. & Holmes, P. *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*, vol. 42 (Springer Science & Business Media, 2013).

[75] Stiefs, D., Gross, T., Steuer, R. & Feudel, U. Computation and visualization of bifurcation surfaces. *International Journal of Bifurcation and chaos* **18**, 2191–2206 (2008).

[76] Zumsande, M. & Gross, T. Bifurcations and chaos in the mapk signaling cascade. *Journal of theoretical biology* **265**, 481–491 (2010).

[77] Newman, M. *Networks: An Introduction* (Oxford University Press, 2010). URL `https://doi.org/10.1093/acprof:oso/9780199206650.001.0001`.

[78] Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics* **74**, 47 (2002).

[79] Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D.-U. Complex networks: Structure and dynamics. *Physics reports* **424**, 175–308 (2006).

[80] Mohar, B., Alavi, Y., Chartrand, G. & Oellermann, O. The laplacian spectrum of graphs. *Graph theory, combinatorics, and applications* **2**, 12 (1991).

[81] Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences* **103**, 12115–12120 (2006).

[82] Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

[83] Glass, J. I. *et al.* Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences* **103**, 425–430 (2006).

[84] Larsen, B. B., Miller, E. C., Rhodes, M. K. & Wiens, J. J. Inordinate fondness multiplied and redistributed: the number of species on earth and the new pie of life. *The Quarterly Review of Biology* **92**, 229–265 (2017).

[85] Bellman, R. Dynamic programming. *Princeton, USA: Princeton University Press* **1**, 3 (1957).

[86] Ma, Y. & Fu, Y. *Manifold learning theory and applications*, vol. 434 (CRC press Boca Raton, 2012).

[87] Fiedler, M. Algebraic connectivity of graphs. *Czechoslovak mathematical journal* **23**, 298–305 (1973).

[88] Reiter, T., Montpetit, R., Runnebaum, R., Brown, C. T. & Montpetit, B. Charting shifts in saccharomyces cerevisiae gene expression across asynchronous time trajectories with diffusion maps. *Mbio* **12**, e02345–21 (2021).

[89] Rao, C. R. Diversity and dissimilarity coefficients: a unified approach. *Theoretical population biology* **21**, 24–43 (1982).

[90] Ricotta, C. & Moretti, M. Cwm and rao's quadratic diversity: a unified framework for functional ecology. *Oecologia* **167**, 181–188 (2011).

[91] Fussmann, G. F. & Blasius, B. Community response to enrichment is highly sensitive to model structure. *Biology letters* **1**, 9–12 (2005).

[92] Groß, T. *Population dynamics: General results from local analysis.* Ph.D. thesis, Universität Oldenburg (2004).

[93] Massing, J. C. & Gross, T. Generalized structural kinetic modeling: A survey and guide. *Frontiers in Molecular Biosciences* **9** (2022).

[94] Wigner, E. P. Characteristic vectors of bordered matrices with infinite dimensions i. *The Collected Works of Eugene Paul Wigner: Part A: The Scientific Papers* 524–540 (1993).

[95] Yeakel, J. D., Stiefs, D., Novak, M. & Gross, T. Generalized modeling of ecological population dynamics. *Theoretical Ecology* **4**, 179–194 (2011).

[96] Reilly, E. The use of the elasticity concept in economic theory. *Canadian Journal of Economics and Political Science/Revue canadienne de economiques et science politique* **6**, 39–55 (1940).

[97] Fell, D. A. & Sauro, H. M. Metabolic control and its analysis: additional relationships between elasticities and control coefficients. *European Journal of Biochemistry* **148**, 555–561 (1985).

[98] Steuer, R. *et al.* From structure to dynamics of metabolic pathways: application to the plant mitochondrial tca cycle. *Bioinformatics* **23**, 1378–1385 (2007).

[99] Allen, B., Nowak, M. A. & Dieckmann, U. Adaptive dynamics with interaction structure. *The American Naturalist* **181**, E139–E163 (2013).

[100] Doizy, A., Barter, E., Memmott, J., Varnham, K. & Gross, T. Impact of cyber-invasive species on a large ecological network. *Scientific reports* **8**, 13245 (2018).

[101] Demirel, G., MacCarthy, B. L., Ritterskamp, D., Champneys, A. R. & Gross, T. Identifying dynamical instabilities in supply networks using generalized modeling. *Journal of Operations Management* **65**, 136–159 (2019).

[102] Guckenheimer, J., Myers, M. & Sturmfels, B. Computing hopf bifurcations i. *SIAM Journal on Numerical Analysis* **34**, 1–21 (1997).

[103] Gross, T. & Feudel, U. Analytical search for bifurcation surfaces in parameter space. *Physica D: Nonlinear Phenomena* **195**, 292–302 (2004).

[104] Gross, T., Rudolf, L., Levin, S. A. & Dieckmann, U. Generalized models reveal stabilizing factors in food webs. *Science* **325**, 747–750 (2009).

[105] Plitzko, S. J., Drossel, B. & Guill, C. Complexity–stability relations in generalized food-web models with realistic parameters. *Journal of theoretical biology* **306**, 7–14 (2012).

[106] Awender, S., Wackerbauer, R. & Breed, G. A. Stability of generalized ecological-network models. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **31**, 023106 (2021).

[107] Aufderheide, H., Rudolf, L. & Gross, T. Mesoscale symmetries explain dynamical equivalence of food webs. *New Journal of Physics* **14**, 105014 (2012).

[108] Yeakel, J. D. *et al.* Collapse of an ecological network in ancient egypt. *Proceedings of the National Academy of Sciences* **111**, 14472–14477 (2014).

[109] Lade, S. J. & Gross, T. Early warning signals for critical transitions: a generalized modeling approach. *PLoS computational biology* **8**, e1002360 (2012).

[110] Höfener, J. M., Sethia, G. C. & Gross, T. Stability of networks of delay-coupled delay oscillators. *Europhysics Letters* **95**, 40002 (2011).

[111] Höfener, J. M., Sethia, G. C. & Gross, T. Amplitude death in networks of delay-coupled delay oscillators. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371**, 20120462 (2013).

[112] Kuehn, C. & Gross, T. Nonlocal generalized models of predator-prey systems. *Discrete and Continuous Dynamical Systems - Series B* **18**, 693–720 (2013).

[113] Yeakel, J. D. & Mangel, M. A generalized perturbation approach for exploring stock recruitment relationships. *Theoretical ecology* **8**, 1–13 (2015).

[114] Baurmann, M., Gross, T. & Feudel, U. Instabilities in spatially extended predator–prey systems: Spatio-temporal patterns in the neighborhood of turing–hopf bifurcations. *Journal of Theoretical Biology* **245**, 220–229 (2007).

[115] Stiefs, D., Venturino, E., Feudel, U. *et al.* Evidence of chaos in eco-epidemic models. *Math. Biosci. Eng* **6**, 855–871 (2009).

[116] Stiefs, D., Van Voorn, G. A., Kooi, B. W., Feudel, U. & Gross, T. Food quality in producer-grazer models: a generalized analysis. *The American Naturalist* **176**, 367–380 (2010).

[117] Leibold, M. A. *et al.* The metacommunity concept: a framework for multi-scale community ecology. *Ecology letters* **7**, 601–613 (2004).

[118] Tromeur, E., Rudolf, L. & Gross, T. Impact of dispersal on the stability of metapopulations. *Journal of theoretical biology* **392**, 1–11 (2016).

[119] Gramlich, P., Plitzko, S. J., Rudolf, L., Drossel, B. & Gross, T. The influence of dispersal on a predator–prey system with two habitats. *Journal of Theoretical Biology* **398**, 150–161 (2016).

[120] Brechtel, A., Gramlich, P., Ritterskamp, D., Drossel, B. & Gross, T. Master stability functions reveal diffusion-driven pattern formation in networks. *Physical Review E* **97**, 032307 (2018).

[121] Anderson, K. E. & Fahimipour, A. K. Body size dependent dispersal influences stability in heterogeneous metacommunities. *Scientific reports* **11**, 17410 (2021).

[122] Steuer, R., Gross, T., Selbig, J. & Blasius, B. Structural kinetic modeling of metabolic networks. *Proceedings of the National Academy of Sciences* **103**, 11868–11873 (2006).

[123] Carbonaro, N. J. & Thorpe, I. F. Using structural kinetic modeling to identify key determinants of stability in reaction networks. *The Journal of Physical Chemistry A* **121**, 4982–4992 (2017).

[124] Gehrmann, E. *et al.* Robustness of glycolysis in yeast to internal and external noise. *Physical Review E* **84**, 021913 (2011).

[125] Lade, S. J., Tavoni, A., Levin, S. A. & Schlüter, M. Regime shifts in a social-ecological system. *Theoretical ecology* **6**, 359–372 (2013).

[126] Lade, S. J. *et al.* An empirical model of the baltic sea reveals the importance of social dynamics for ecological regime shifts. *Proceedings of the National Academy of Sciences* **112**, 11120–11125 (2015).

[127] Lade, S. J. & Niiranen, S. Generalized modeling of empirical social-ecological systems. *Natural Resource Modeling* **30**, e12129 (2017).

[128] Molla, N., DeIonno, J., Gross, T. & Herman, J. Governing change: a dynamical systems approach to understanding the stability of environmental governance. *Earth System Dynamics* **13**, 1677–1688 (2022).

[129] Levins, R. Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bulletin of the ESA* **15**, 237–240 (1969).

[130] Yeakel, J. D., Moore, J. W., Guimarães Jr, P. R. & de Aguiar, M. A. Synchronisation and stability in river metapopulation networks. *Ecology letters* **17**, 273–283 (2014).

[131] Pillai, P., Loreau, M. & Gonzalez, A. A patch-dynamic framework for food web metacommunities. *Theoretical Ecology* **3**, 223–237 (2010).

[132] Mougi, A. & Kondoh, M. Food-web complexity, meta-community complexity and community stability. *Scientific reports* **6**, 24478 (2016).

[133] Nakao, H. & Mikhailov, A. S. Turing patterns in network-organized activator–inhibitor systems. *Nature Physics* **6**, 544–550 (2010).

[134] Segel, L. A. & Levin, S. A. Application of nonlinear stability theory to the study of the effects of diffusion on predator-prey interactions. In *AIP conference proceedings*, vol. 27, 123–152 (American Institute of Physics, 1976).

[135] Massing, J. C., Fahimipour, A. K., Bunse, C., Pinhassi, J. & Gross, T. Quantification of metabolic niche occupancy dynamics in a baltic sea bacterial community. *Msystems* e00028–23 (2023).

[136] Egan, S. *et al.* The seaweed holobiont: understanding seaweed–bacteria interactions. *FEMS Microbiology Reviews* **37**, 462–476 (2013).

[137] Zhang, W. *et al.* Marine biofilms constitute a bank of hidden microbial diversity and functional potential. *Nature communications* **10**, 517 (2019).

[138] Kirchman, D. L. Growth rates of microbes in the oceans. *Annual review of marine science* **8**, 285–309 (2016).

[139] Bunse, C. *et al.* High frequency multi-year variability in baltic sea microbial plankton stocks and activities. *Frontiers in microbiology* **9**, 3296 (2019).

[140] Lindh, M. V. *et al.* Disentangling seasonal bacterioplankton population dynamics by high-frequency sampling. *Environmental microbiology* **17**, 2459–2476 (2015).

[141] Fridolfsson, E. *et al.* Multiyear analysis uncovers coordinated seasonality in stocks and composition of the planktonic food web in the baltic sea proper. *Scientific Reports* **13**, 11865 (2023).

[142] Boström, K. H., Simu, K., Hagström, Å. & Riemann, L. Optimization of dna extraction for quantitative marine bacterioplankton community analysis. *Limnology and Oceanography: Methods* **2**, 365–373 (2004).

[143] Bunse, C. *et al.* Spatio-temporal interdependence of bacteria and phytoplankton during a baltic sea spring bloom. *Frontiers in microbiology* **7**, 517 (2016).

[144] Herlemann, D. P. *et al.* Transitions in bacterial communities along the 2000 km salinity gradient of the baltic sea. *The ISME journal* **5**, 1571–1579 (2011).

[145] Hugerth, L. W. *et al.* Degeprime, a program for degenerate primer design for broad-taxonomic-range pcr in microbial ecology studies. *Applied and environmental microbiology* **80**, 5116–5123 (2014).

[146] Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nature biotechnology* **38**, 276–278 (2020).

[147] Straub, D. *et al.* Interpretations of environmental microbial community studies are biased by the selected 16s rrna (gene) amplicon sequencing pipeline. *Frontiers in Microbiology* **11**, 550420 (2020).

[148] Callahan, B. J. *et al.* Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods* **13**, 581–583 (2016).

[149] Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature biotechnology* **37**, 852–857 (2019).

[150] GTDB, release 95 database. `https://data.gtdb.ecogenomic.org/`. Accessed: 2021-05-11.

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

[152] Wheeler, T. J. & Eddy, S. R. nhmmer: Dna homology search with profile hmms. *Bioinformatics* **29**, 2487–2489 (2013).

[153] Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: an rna family database. *Nucleic acids research* **31**, 439–441 (2003).

[154] Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

[155] Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).

[156] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995).

[157] Pollack, J. R. & Neilands, J. Enterobactin, an iron transport compound from salmonella typhimurium. *Biochemical and biophysical research communications* **38**, 989–992 (1970).

[158] O'brien, I. & Gibson, F. The structure of enterochelin and related 2, 3-dihydroxy-n-benzoyne conjugates from eschericha coli. *Biochimica et Biophysica Acta (BBA)-General Subjects* **215**, 393–402 (1970).

[159] Aldridge, P. & Hughes, K. T. Regulation of flagellar assembly. *Current opinion in microbiology* **5**, 160–165 (2002).

[160] Domka, J., Lee, J. & Wood, T. K. Ylih (bssr) and ycep (bsss) regulate escherichia coli k-12 biofilm formation by influencing cell signaling. *Applied and environmental microbiology* **72**, 2449–2459 (2006).

[161] Blackwell, G. A. *et al.* Exploring bacterial diversity via a curated and searchable snapshot of archived dna sequences. *PLoS biology* **19**, e3001421 (2021).

[162] Hallenbeck, P. C. *Modern topics in the phototrophic prokaryotes: Metabolism, bioenergetics, and omics* (Springer, 2017).

[163] Raven, J. A. The cost of photoinhibition. *Physiologia Plantarum* **142**, 87–104 (2011).

[164] Luo, H. & Moran, M. A. How do divergent ecological strategies emerge among marine bacterioplankton lineages? *Trends in microbiology* **23**, 577–584 (2015).

[165] Alves, L. M. C., de Souza, J. A. M., de Mello Varani, A. & de Macedo Lemos, E. The family rhizobiaceae. *The Prokaryotes* 419–437 (2014).

[166] Zecher, K., Hayes, K. R. & Philipp, B. Evidence of interdomain ammonium cross-feeding from methylamine-and glycine betaine-degrading rhodobacteraceae to diatoms as a widespread interaction in the marine phycosphere. *Frontiers in microbiology* **11** (2020).

[167] Buchan, A., LeCleir, G. R., Gulvik, C. A. & González, J. M. Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nature Reviews Microbiology* **12**, 686–698 (2014).

[168] Tett, A., Pasolli, E., Masetti, G., Ercolini, D. & Segata, N. Prevotella diversity, niches and interactions with the human host. *Nature Reviews Microbiology* **19**, 585–599 (2021).

[169] Teeling, H. *et al.* Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science* **336**, 608–611 (2012).

[170] Sun, J. *et al.* One carbon metabolism in sar11 pelagic marine bacteria. *PloS one* **6**, e23973 (2011).

[171] Venceslau, S. S., Lino, R. R. & Pereira, I. A. The qrc membrane complex, related to the alternative complex iii, is a menaquinone reductase involved in sulfate respiration. *Journal of Biological Chemistry* **285**, 22774–22783 (2010).

[172] Tian, R. *et al.* Small and mighty: adaptation of superphylum patescibacteria to groundwater environment drives their genome simplicity. *Microbiome* **8**, 1–15 (2020).

[173] Grote, J. *et al.* Streamlining and core genome conservation among highly divergent members of the sar11 clade. *MBio* **3**, e00252–12 (2012).

[174] Darby, A. C., Cho, N.-H., Fuxelius, H.-H., Westberg, J. & Andersson, S. G. Intracellular pathogens go extreme: genome evolution in the rickettsiales. *TRENDS in Genetics* **23**, 511–520 (2007).

[175] Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory for microbial ecology. *The ISME journal* **8**, 1553–1565 (2014).

[176] Ferenci, T. Trade-off mechanisms shaping the diversity of bacteria. *Trends in microbiology* **24**, 209–223 (2016).

[177] Blankenship, R. E. Early evolution of photosynthesis. *Plant physiology* **154**, 434–438 (2010).

[178] Mühlenbruch, M., Grossart, H.-P., Eigemann, F. & Voss, M. Mini-review: Phytoplankton-derived polysaccharides in the marine environment and their interactions with heterotrophic bacteria. *Environmental microbiology* **20**, 2671–2685 (2018).

[179] Ferrer-González, F. X. *et al.* Resource partitioning of phytoplankton metabolites that support bacterial heterotrophy. *The ISME journal* **15**, 762–773 (2021).

[180] Newton, R. J. *et al.* Genome characteristics of a generalist marine bacterial lineage. *The ISME journal* **4**, 784–798 (2010).

[181] Tsubouchi, T. *et al.* Thalassospira alkalitolerans sp. nov. and thalassospira mesophila sp. nov., isolated from a decaying bamboo sunken in the marine environment, and emended description of the genus thalassospira. *International journal of systematic and evolutionary microbiology* **64**, 107–115 (2014).

[182] Bergen, B. *et al.* Impact of a major inflow event on the composition and distribution of bacterioplankton communities in the baltic sea. *Frontiers in Marine Science* 383 (2018).

[183] Lindh, M. V. & Pinhassi, J. Sensitivity of bacterioplankton to environmental disturbance: a review of baltic sea field studies and experiments. *Frontiers in Marine Science* **5**, 361 (2018).

[184] Bertos-Fortis, M. *et al.* Unscrambling cyanobacteria community dynamics related to environmental factors. *Frontiers in microbiology* **7**, 625 (2016).

[185] Vetterli, A. *et al.* Seasonal patterns of bacterial communities in the coastal brackish sediments of the gulf of finland, baltic sea. *Estuarine, Coastal and Shelf Science* **165**, 86–96 (2015).

[186] Korneeva, V., Pimenov, N., Krek, A., Tourova, T. & Bryukhanov, A. Sulfate-reducing bacterial communities in the water column of the gdansk deep (baltic sea). *Microbiology* **84**, 268–277 (2015).

[187] Reissmann, J. H. *et al.* Vertical mixing in the baltic sea and consequences for eutrophication–a review. *Progress in Oceanography* **82**, 47–80 (2009).

[188] Herlemann, D. P., Lundin, D., Andersson, A. F., Labrenz, M. & Jürgens, K. Phylogenetic signals of salinity and season in bacterial community composition across the salinity gradient of the baltic sea. *Frontiers in Microbiology* **7**, 1883 (2016).

[189] Rieck, A., Herlemann, D. P., Jürgens, K. & Grossart, H.-P. Particle-associated differ from free-living bacteria in surface waters of the baltic sea. *Frontiers in Microbiology* **6**, 1297 (2015).

[190] Raes, E. *et al.* Seasonal bacterial niche structures and chemolithoautotrophic ecotypes in a north atlantic fjord (2022). Doi:doi.org/10.21203/rs.3.rs-1469763/v1.

[191] Pernthaler, J. Competition and niche separation of pelagic bacteria in freshwater habitats. *Environmental microbiology* **19**, 2133–2150 (2017).

[192] Pianka, E. R., Vitt, L. J., Pelegrin, N., Fitzgerald, D. B. & Winemiller, K. O. Toward a periodic table of niches, or exploring the lizard niche hypervolume. *The American Naturalist* **190**, 601–616 (2017).

[193] Gogarten, J. P. & Townsend, J. P. Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology* **3**, 679–687 (2005).

[194] Ward, C. S. *et al.* Annual community patterns are driven by seasonal switching between closely related marine bacteria. *The ISME journal* **11**, 1412–1422 (2017).

[195] Seidel, M. *et al.* Composition and transformation of dissolved organic matter in the baltic sea. *Frontiers in Earth Science* **5**, 31 (2017).

[196] Hutchinson, G. E. Cold spring harbor symposium on quantitative biology. *Concluding remarks* **22**, 415–427 (1957).

[197] Gralka, M., Pollak, S. & Cordero, O. X. Fundamental metabolic strategies of heterotrophic bacteria. *bioRxiv* (2022).

[198] Gowda, K., Ping, D., Mani, M. & Kuehn, S. Genomic structure predicts metabolite dynamics in microbial communities. *Cell* (2022).

[199] Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D. & Konstantinidis, K. T. Strengths and limitations of 16s rrna gene amplicon sequencing in revealing temporal microbial community dynamics. *PloS one* **9**, e93827 (2014).

[200] García-García, N., Tamames, J., Linz, A. M., Pedrós-Alió, C. & Puente-Sánchez, F. Microdiversity ensures the maintenance of functional microbial communities under changing environmental conditions. *The ISME journal* **13**, 2969–2983 (2019).

[201] Zhu, A., Sunagawa, S., Mende, D. R. & Bork, P. Inter-individual differences in the gene content of human gut bacterial species. *Genome biology* **16**, 1–13 (2015).

[202] Forcone, K., Coutinho, F. H., Cavalcanti, G. S. & Silveira, C. B. Prophage genomics and ecology in the family rhodobacteraceae. *Microorganisms* **9**, 1115 (2021).

[203] Douglas, G. M. *et al.* Picrust2 for prediction of metagenome functions. *Nature biotechnology* **38**, 685–688 (2020).

[204] Germerodt, S. *et al.* Pervasive selection for cooperative cross-feeding in bacterial communities. *PLoS computational biology* **12**, e1004986 (2016).

[205] Mee, M. T., Collins, J. J., Church, G. M. & Wang, H. H. Syntrophic exchange in synthetic microbial communities. *Proceedings of the National Academy of Sciences* **111**, E2149–E2156 (2014).

[206] Douglas, A. E. The microbial exometabolome: ecological resource and architect of microbial communities. *Philosophical Transactions of the Royal Society B* **375**, 20190250 (2020).

[207] Kost, C., Patil, K. R., Friedman, J., Garcia, S. L. & Ralser, M. Metabolic exchanges are ubiquitous in natural microbial communities. *Nat. Microbiol.* (2023).

[208] Morris, B. E., Henneberger, R., Huber, H. & Moissl-Eichinger, C. Microbial syntrophy: interaction for the common good. *FEMS microbiology reviews* **37**, 384–406 (2013).

[209] D'Souza, G. *et al.* Less is more: selective advantages can explain the prevalent loss of biosynthetic genes in bacteria. *Evolution* **68**, 2559–2570 (2014).

[210] Zengler, K. & Zaramela, L. S. The social network of microorganisms—how auxotrophies shape complex communities. *Nature Reviews Microbiology* 1 (2018).

[211] Embree, M., Liu, J. K., Al-Bassam, M. M. & Zengler, K. Networks of energetic and metabolic interactions define dynamics in microbial communities. *Proceedings of the National Academy of Sciences* **112**, 15450–15455 (2015).

[212] Kost, C. *et al.* Metabolic cross-feeding via intercellular nanotubes among bacteria. *Nature Communications* **6** (2015).

[213] Kallus, Y., Miller, J. H. & Libby, E. Paradoxes in leaky microbial trade. *Nature communications* **8**, 1361 (2017).

[214] Marleau, J. N., Guichard, F. & Loreau, M. Meta-ecosystem dynamics and functioning on finite spatial networks. *Proceedings of the Royal Society B: Biological Sciences* **281**, 20132094 (2014).

[215] Harcombe, W. R. *et al.* Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell reports* **7**, 1104–1115 (2014).

[216] Pacheco, A. R., Moel, M. & Segre, D. Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nature communications* **10**, 103 (2019).

[217] Peaudecerf, F. J. *et al.* Microbial mutualism at a distance: The role of geometry in diffusive exchanges. *Physical Review E* **97**, 022411 (2018).

[218] Muscarella, M. E. & O'Dwyer, J. P. Species dynamics and interactions via metabolically informed consumer-resource models. *BioRxiv* 518449 (2019).

[219] Hoek, T. A. *et al.* Resource availability modulates the cooperative and competitive nature of a microbial cross-feeding mutualism. *PLoS biology* **14**, e1002540 (2016).

[220] Vet, S., Gelens, L. & Gonze, D. Mutualistic cross-feeding in microbial systems generates bistability via an allee effect. *Scientific reports* **10**, 7763 (2020).

[221] Harcombe, W. Novel cooperation experimentally evolved between species. *Evolution* **64**, 2166–2172 (2010).

[222] Gottschalk, G. & Gottschalk, G. *Regulation of bacterial metabolism* (Springer, 1986).

[223] Cortez, M. H. & Abrams, P. A. Hydra effects in stable communities and their implications for system dynamics. *Ecology* **97**, 1135–1145 (2016).

[224] Turing, A. The chemical theory of morphogenesis. *Phil. Trans. Roy. Soc* **13** (1952).

[225] Baurmann, M., Ebenhöh, W. & Feudel, U. Turing instabilities and pattern formation in a benthic nutrient-microorganism system. *Mathematical Biosciences & Engineering* **1**, 111–130 (2004).

[226] Ruuskanen, M. O., Sommeria-Klein, G., Havulinna, A. S., Niiranen, T. J. & Lahti, L. Modelling spatial patterns in host-associated microbial communities. *Environmental Microbiology* **23**, 2374–2388 (2021).

[227] Pande, S. *et al.* Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. *The ISME journal* **8**, 953–962 (2014).

[228] Lilja, E. E. & Johnson, D. R. Segregating metabolic processes into different microbial cells accelerates the consumption of inhibitory substrates. *The ISME journal* **10**, 1568–1578 (2016).

[229] Hillesland, K. L. & Stahl, D. A. Rapid evolution of stability and productivity at the origin of a microbial mutualism. *Proceedings of the National Academy of Sciences* **107**, 2124–2129 (2010).

[230] D'Souza, G. & Kost, C. Experimental evolution of metabolic dependency in bacteria. *PLoS genetics* **12**, e1006364 (2016).

[231] Harcombe, W. R., Chacón, J. M., Adamowicz, E. M., Chubiz, L. M. & Marx, C. J. Evolution of bidirectional costly mutualism from byproduct consumption. *Proceedings of the National Academy of Sciences* **115**, 12000–12004 (2018).

[232] Coyte, K. Z. & Rakoff-Nahoum, S. Understanding competition and cooperation within the mammalian gut microbiome. *Current Biology* **29**, R538–R544 (2019).

[233] Momeni, B., Waite, A. J. & Shou, W. Spatial self-organization favors heterotypic cooperation over cheating. *Elife* **2**, e00960 (2013).

[234] Ratzke, C. & Gore, J. Self-organized patchiness facilitates survival in a cooperatively growing bacillus subtilis population. *Nature microbiology* **1**, 1–5 (2016).

[235] Kayser, J., Schreck, C. F., Yu, Q., Gralka, M. & Hallatschek, O. Emergence of evolutionary driving forces in pattern-forming microbial populations. *Philosophical Transactions of the Royal Society B: Biological Sciences* **373**, 20170106 (2018).

[236] Goldschmidt, F., Caduff, L. & Johnson, D. R. Causes and consequences of pattern diversification in a spatially self-organizing microbial community. *The ISME Journal* **15**, 2415–2426 (2021).

[237] Carrasco De La Cruz, P., Antonucci Di Carvalho, J., Massing, J. C. & Gross, T. Aggregation of monitoring datasets for functional diversity estimation. *Frontiers in Ecology and Evolution* In review.

[238] Petchey, O. L., Hector, A. & Gaston, K. J. How do different measures of functional diversity perform? *Ecology* **85**, 847–857 (2004).

[239] Legras, G., Loiseau, N., Gaertner, J.-C., Poggiale, J.-C. & Gaertner-Mazouni, N. Assessing functional diversity: the influence of the number of the functional traits. *Theoretical Ecology* **13**, 117–126 (2020).

[240] Minchin, P. R. Simulation of multidimensional community patterns: towards a comprehensive model. *Vegetatio* **71**, 145–156 (1987).

[241] Mentges, A., Feenders, C., Deutsch, C., Blasius, B. & Dittmar, T. Long-term stability of marine dissolved organic carbon emerges from a neutral network of compounds and microbes. *Scientific reports* **9**, 17780 (2019).

[242] Williams, R. J. & Martinez, N. D. Simple rules yield complex food webs. *Nature* **404**, 180–183 (2000).

[243] Lynch, M. D. & Neufeld, J. D. Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology* **13**, 217–229 (2015).

[244] Leimbach, A., Hacker, J. & Dobrindt, U. E. coli as an all-rounder: the thin line between commensalism and pathogenicity. *Between pathogenicity and commensalism* 3–32 (2013).

[245] Van Rossum, T., Ferretti, P., Maistrenko, O. M. & Bork, P. Diversity within species: interpreting strains in microbiomes. *Nature Reviews Microbiology* **18**, 491–506 (2020).

[246] Gross, T. Generalised modelling in systems biology. *New Frontiers of Network Analysis in Systems Biology* 59–75 (2012).

[247] Gibbs, T., Zhang, Y., Miller, Z. R. & O'Dwyer, J. P. Stability criteria for the consumption and exchange of essential resources. *PLoS computational biology* **18**, e1010521 (2022).

[248] Boza, G., Barabas, G., Scheuring, I. & Zachar, I. Eco-evolutionary modelling of microbial syntrophy indicates the robustness of cross-feeding over cross-facilitation. *Scientific Reports* **13**, 907 (2023).

Appendix A

Supplementary Tables

Table S1: Genomes that map to the 100 most abundant ASVs obtained from amplicon sequencing data in terms of relative mean abundance over the whole sampling period. Genome, taxonomic information, mean and maximum abundance (ab) over the whole sampling period are provided.

Genome	Class	Family	Species	Mean ab	Max ab
GCF_003011885.1	Cyanobacteriia	Cyanobiaceae	Cyanobium_A usitatum	0.1150	0.7222
GCF_002252665.1	Cyanobacteriia	Cyanobiaceae	Cyanobium_A sp002252665	0.0533	0.2779
GCA_003569125.1	Acidimicrobiia	Ilumatobacteraceae	BACL27 sp003569125	0.0488	0.3656
GCA_001593825.1	Cyanobacteriia	Nostocaceae	Aphanizomenon_B flosaquae	0.0400	0.5437
GCF_000173115.1	Bacteroidia	Flavobacteriaceae	MAG-120531 sp000173115	0.0397	0.5443
GCA_002358295.1	Gammaproteobacteria	D2472	D2472 sp002358345	0.0287	0.2754
GCA_003569145.1	Actinomycetia	Nanopelagicaceae	MAG-120802 sp003569145	0.0286	0.2757
GCA_001438235.1	Alphaproteobacteria	Rhodobacteraceae	UBA10365 sp003536295	0.0235	0.1846
GCA_002405515.1	Planctomycetes	UBA1268	UBA4655 sp002405515	0.0226	0.1789
GCA_007280255.1	Planctomycetes	UBA1268	QWOQ01 sp003669585	0.0213	0.2010
GCA_001437765.1	Acidimicrobiia	Ilumatobacteraceae	UBA3006 sp002367695	0.0208	0.2445
GCA_002325485.1	Bacteroidia	Flavobacteriaceae	BACL21 sp002694465	0.0201	0.3039
GCA_002340585.1	Gammaproteobacteria	Porticoccaceae	HTCC2207 sp001438605	0.0191	0.2454
GCF_001485105.1	Actinomycetia	Streptomycetaceae	Streptomyces acidiscabies	0.0163	0.3385
GCA_002711735.1	Acidimicrobiia	Ilumatobacteraceae	Ilumatobacter_A sp002711735	0.0142	0.2166
GCF_000496475.1	Gammaproteobacteria	Burkholderiaceae	RS62 sp000496475	0.0139	0.1293
GCF_000257665.1	Actinomycetia	Microbacteriaceae	Aquiluna sp000257665	0.0129	0.2494
GCF_000312705.1	Cyanobacteriia	Nostocaceae	LE011-02 sp000312705	0.0122	0.5288
GCF_002287885.2	Actinomycetia	Nanopelagicaceae	Nanopelagicus limnes	0.0114	0.0733
GCF_000242915.1	Campylobacteria	Sulfurimonadaceae	Sulfurimonas gotlandica	0.0107	0.6154
GCA_002746305.1	Bacteroidia	UBA9320	UBA9320 sp002746305	0.0100	0.0928
GCA_002430225.1	Actinomycetia	Microbacteriaceae	Pontimonas sp001438965	0.0093	0.2095
GCF_002252705.1	Cyanobacteriia	Cyanobiaceae	Vulcanococcus limneticus	0.0092	0.1075
GCA_000750175.1	Alphaproteobacteria	Pelagibacteraceae	IMCC9063 sp000750175	0.0090	0.1093
GCA_002340845.1	Gammaproteobacteria	Methylophilaceae	BACL14 sp002384685	0.0079	0.0679
GCA_001438645.1	Gammaproteobacteria	Methylophilaceae	BACL14 sp002384685	0.0076	0.0700
GCA_001438145.1	Gammaproteobacteria	Pseudohongiellaceae	OM182 sp001438145	0.0074	0.0646
GCF_002284895.1	Actinomycetia	Nanopelagicaceae	Planktophila sp002284895	0.0074	0.0575
GCA_000485495.1	Actinomycetia	Nanopelagicaceae	AAA044-D11 sp000485495	0.0074	0.0642
GCA_002170165.1	Bacteroidia	BACL11	TMED123 sp002170165	0.0072	0.0501
GCF_900129545.1	Bacteroidia	Flavobacteriaceae	Flavobacterium fluvii	0.0060	0.2432
GCF_001983935.1	Planctomycetes	Planctomycetaceae	Fuerstia marisgermanicae	0.0058	0.1676
GCA_001438305.1	Bacteroidia	Schleiferiaceae	TMED14 sp001438205	0.0056	0.0693
GCF_002252635.1	Cyanobacteriia	Cyanobiaceae	WH-5701 sp002252635	0.0055	0.0868
GCA_004292795.1	Bacteroidia	Microscillaceae	RDXI01 sp004292795	0.0054	0.1049
GCF_006491595.1	Bacteroidia	Flavobacteriaceae	Flavobacterium jejuense	0.0052	0.0682
GCF_002943715.1	Bacteroidia	Flavobacteriaceae	Polaribacter filamentus	0.0052	0.1530
GCF_000299115.1	Alphaproteobacteria	HIMB59	HIMB59 sp000299115	0.0051	0.0616
GCF_900114485.1	Alphaproteobacteria	Rhodobacteraceae	Loktanella salsilacus	0.0049	0.0452
GCA_004379135.1	Acidimicrobiia	Ilumatobacteraceae	Casp-actino8 sp004379135	0.0048	0.0220
GCA_003249095.1	Cyanobacteriia	Microcystaceae	Snowella sp003249095	0.0047	0.1042
GCF_002631185.1	Alphaproteobacteria	Acetobacteraceae	Roseomonas rhizosphaerae	0.0047	0.3294
GCA_000738435.1	Alphaproteobacteria	Rhodobacteraceae	Planktomarina temperata	0.0046	0.1321
GCF_003096315.1	Gammaproteobacteria	Burkholderiaceae	Achromobacter insuavis	0.0046	0.4719
GCA_003284275.1	Alphaproteobacteria	Pelagibacteraceae	Pelagibacter_A sp003284275	0.0045	0.0547
GCF_002252675.1	Cyanobacteriia	Cyanobiaceae	Cyanobium sp002252675	0.0044	0.0434
GCF_002954645.1	Bacteroidia	Flavobacteriaceae	Polaribacter gangjinensis	0.0043	0.0744
GCF_002101315.1	Alphaproteobacteria	Pelagibacteraceae	Pelagibacter sp002101315	0.0040	0.0398

GCF_002115755.1	Alphaproteobacteria	Thalassospiraceae	Thalassospira mesophila	0.0039	0.2429
GCA_002733565.1	Gammaproteobacteria	Psychromonadaceae	Moritella sp000170855	0.0039	0.0718
GCF_002288225.1	Actinomycetia	Nanopelagicaceae	Planktophila dulcis	0.0039	0.0286
GCA_002428815.1	Gammaproteobacteria	Porticoccaceae	HTCC2207 sp001438605	0.0038	0.0499
GCF_000590925.1	Alphaproteobacteria	Rhodobacteraceae	Roseicyclus elongatum	0.0037	0.0555
GCA_002346275.1	Gammaproteobacteria	Halieaceae	IMCC3088 sp003520285	0.0037	0.0802
GCF_002940745.1	Bacteroidia	Flavobacteriaceae	Hanstruepera crassostreae	0.0036	0.0623
GCF_003335085.1	Bacteroidia	Flavobacteriaceae	Polaribacter sp003335085	0.0036	0.1667
GCF_000699505.1	Actinomycetia	Microbacteriaceae	Rhodoluna lacicola	0.0035	0.0265
GCA_003149555.1	Actinomycetia	Microbacteriaceae	Aquiluna sp003149555	0.0034	0.0827
GCA_004379115.1	Actinomycetia	S36-B12	Mxb001 sp004379115	0.0034	0.0716
GCA_001438005.1	Verrucomicrobiae	UBA3015	UBA3015 sp001438005	0.0032	0.0291
GCA_002346225.1	Bacteroidia	BACL12	UBA11426 sp002346225	0.0031	0.1118
GCF_003003055.1	Gammaproteobacteria	Burkholderiaceae	SCGC-AAA027-K21 sp003003055	0.0031	0.0317
GCF_002284855.1	Actinomycetia	Nanopelagicaceae	Planktophila sp002284855	0.0029	0.0379
GCA_002863125.1	Bacteroidia	UA16	UA16 sp002863125	0.0029	0.0250
GCF_002284915.1	Actinomycetia	Nanopelagicaceae	IMCC26077 sp002284915	0.0029	0.0279
GCA_001438165.1	Bacteroidia	Schleiferiaceae	TMED14 sp002381225	0.0029	0.0274
GCF_001457835.1	Clostridia	Ezakiellaceae	Fenollaria timonensis	0.0029	0.1789
GCF_000152785.1	Alphaproteobacteria	Rhodobacteraceae	Yoonia vestfoldensis_A	0.0028	0.0290
GCA_002292365.1	Bacteroidia	Cyclobacteriaceae	UBA4465 sp002292365	0.0028	0.0194
GCF_001439695.1	Gammaproteobacteria	Pseudomonadaceae	Pseudomonas_E veronii	0.0028	0.1510
GCA_000421325.1	Alphaproteobacteria	AAA536-G10	AAA536-G10 sp000421325	0.0027	0.0406
GCA_003208775.1	Cyanobacteriia	Cyanobiaceae	Synechococcus_C sp002500205	0.0026	0.1927
GCA_003671255.1	Planctomycetes	Gemmataceae	UBA969 sp003671255	0.0025	0.0379
GCA_007093895.1	Gammaproteobacteria	Enterobacteriaceae	Salmonella enterica	0.0024	0.0168
GCF_003011125.1	Cyanobacteriia	Cyanobiaceae	Synechococcus_D lacustris	0.0024	0.0380
GCF_004337435.1	Actinomycetia	Streptomycetaceae	Streptomyces sp004337435	0.0024	0.0120
GCA_002167745.1	Gammaproteobacteria	SG8-40	UBA3031 sp002167745	0.0023	0.0320
GCA_900618205.1	Gammaproteobacteria	Burkholderiaceae	Bordetella trematum	0.0023	0.1770
GCF_000173095.1	Bacteroidia	Flavobacteriaceae	MS024-2A sp000173095	0.0023	0.0428
GCF_000797465.1	Bacteroidia	Flavobacteriaceae	Psychroserpens jangbogonensis	0.0021	0.0377
GCA_002690755.1	Phycisphaerae	SM1A02	UBA12014 sp002690755	0.0021	0.0255
GCA_002480055.1	Gammaproteobacteria	Porticoccaceae	HTCC2207 sp002335945	0.0020	0.0215
GCF_000143825.1	Actinomycetia	Mycobacteriaceae	Corynebacterium genitalium_A	0.0020	0.0926
GCF_006385135.1	Alphaproteobacteria	Emcibacteraceae	Emcibacter_A congregatus	0.0020	0.0176
GCF_002284875.1	Actinomycetia	Nanopelagicaceae	Planktophila sp002284875	0.0019	0.0111
GCA_002697205.1	Gammaproteobacteria	HTCC2089	GCA-2697205 sp002697205	0.0019	0.0172
GCA_003045825.1	Bacteroidia	Schleiferiaceae	UBA10364 sp003045825	0.0019	0.0476
GCA_000762985.1	Actinomycetia	Mycobacteriaceae	Mycobacterium rufum	0.0018	0.0245
GCA_002282055.1	Bacteroidia	Sphingobacteriaceae	Daejeonella sp002257025	0.0018	0.0238
GCA_002499015.1	Poseidoniia	Poseidoniaceae	MGIIa-L1 sp002499015	0.0018	0.0959
GCF_000176015.1	Alphaproteobacteria	Rhodobacteraceae	Pseudorhodobacter_B sp000176015	0.0018	0.0476
GCF_006937785.1	Cyanobacteriia	Pseudanabaenaceae	Pseudanabaena sp006937785	0.0017	0.1092
GCF_001623485.1	Cyanobacteriia	Nostocaceae	Nodularia spumigena	0.0016	0.0348
GCF_000171835.1	Alphaproteobacteria	Thalassobaculaceae	BAL199 sp000171835	0.0015	0.0138
GCF_000156155.1	Gammaproteobacteria	Methylophilaceae	BACL14 sp000156155	0.0015	0.0159
GCA_002733945.1	Campylobacteria	Sulfurimonadaceae	Sulfurimonas sp002733945	0.0015	0.0763
GCF_003856375.1	Bacteroidia	Crocinitomicaceae	Fluviicola sp003856375	0.0015	0.0449
GCF_900110395.1	Alphaproteobacteria	Reyranellaceae	Reyranella sp900110395	0.0015	0.0222
GCF_002368115.1	Cyanobacteriia	Nostocaceae	Dolichospermum_A compactum	0.0014	0.0579
GCF_900100865.1	Actinomycetia	Microbacteriaceae	Aquiluna sp900100865	0.0014	0.0188

Table S2: Species that map to the 100 ASVs scoring most negative values in variable 1.

Species	**Value**	**Genome**	**Class**	**Family**
Salmonella enterica	-0.09769	GCA_007094035.1	Gammaproteobacteria	Enterobacteriaceae
Salmonella enterica	-0.09769	GCA_007093895.1	Gammaproteobacteria	Enterobacteriaceae
Salmonella enterica	-0.09763	GCA_007093765.1	Gammaproteobacteria	Enterobacteriaceae
Salmonella enterica	-0.09756	GCF_003548795.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.09754	GCF_002094495.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.09754	GCF_002094665.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.09754	GCF_002977865.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter turicensis	-0.09754	GCF_002976545.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.09753	GCA_002094675.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.09753	GCF_002094645.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.09752	GCF_002977315.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter malonaticus	-0.09752	GCF_002978245.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter malonaticus	-0.09752	GCF_002978235.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.09751	GCF_002094575.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.09751	GCF_002976775.1	Gammaproteobacteria	Enterobacteriaceae

Cronobacter malonaticus	-0.09751	GCF_002978375.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.09751	GCF_002094585.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.09751	GCA_002976965.2	Gammaproteobacteria	Enterobacteriaceae
Cronobacter dublinensis	-0.09751	GCF_002979155.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter malonaticus	-0.09751	GCF_002978545.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter malonaticus	-0.09751	GCF_002978185.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.09751	GCF_002977405.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.09751	GCF_002094475.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter malonaticus	-0.0975	GCF_002978535.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.0975	GCF_002976735.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.0975	GCF_002976795.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.0975	GCA_002977005.2	Gammaproteobacteria	Enterobacteriaceae
Escherichia coli	-0.09749	GCA_002078275.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.09749	GCF_002977155.1	Gammaproteobacteria	Enterobacteriaceae
Escherichia coli	-0.09749	GCF_001268585.1	Gammaproteobacteria	Enterobacteriaceae
Escherichia coli	-0.09749	GCF_001269185.1	Gammaproteobacteria	Enterobacteriaceae
Escherichia coli	-0.09748	GCF_004523105.1	Gammaproteobacteria	Enterobacteriaceae
Escherichia coli	-0.09748	GCF_005889645.1	Gammaproteobacteria	Enterobacteriaceae
Escherichia coli	-0.09748	GCF_001268685.1	Gammaproteobacteria	Enterobacteriaceae
Escherichia coli	-0.09748	GCF_002007165.1	Gammaproteobacteria	Enterobacteriaceae
Escherichia coli	-0.09748	GCF_002959275.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.09748	GCF_002978035.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.09747	GCF_002978105.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter dublinensis	-0.09747	GCA_002978875.2	Gammaproteobacteria	Enterobacteriaceae
Cronobacter dublinensis	-0.09747	GCF_002978655.1	Gammaproteobacteria	Enterobacteriaceae
Enterobacter sp.	-0.09746	GCF_000534395.1	Gammaproteobacteria	Enterobacteriaceae
Scandinavium goeteborgense	-0.09744	GCF_004361715.1	Gammaproteobacteria	Enterobacteriaceae
Scandinavium goeteborgense	-0.09743	GCA_003935895.2	Gammaproteobacteria	Enterobacteriaceae
Klebsiella pneumoniae	-0.09742	GCF_003967395.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter dublinensis	-0.09742	GCF_002978855.1	Gammaproteobacteria	Enterobacteriaceae
Klebsiella quasivariicola	-0.09742	GCF_002269255.1	Gammaproteobacteria	Enterobacteriaceae
Klebsiella variicola	-0.09742	GCF_001033575.1	Gammaproteobacteria	Enterobacteriaceae
Klebsiella quasipneumoniae	-0.09742	GCF_002853635.1	Gammaproteobacteria	Enterobacteriaceae
Klebsiella pneumoniae	-0.09742	GCF_004127885.1	Gammaproteobacteria	Enterobacteriaceae
Citrobacter freundii	-0.09741	GCA_001686345.1	Gammaproteobacteria	Enterobacteriaceae
Enterobacter cloacae	-0.09741	GCF_001562175.1	Gammaproteobacteria	Enterobacteriaceae
Enterobacter sp.	-0.0974	GCF_000493015.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter sakazakii	-0.0974	GCF_002977115.1	Gammaproteobacteria	Enterobacteriaceae
Citrobacter koseri	-0.09739	GCF_002393245.1	Gammaproteobacteria	Enterobacteriaceae
Klebsiella pneumoniae	-0.09739	GCF_003227185.1	Gammaproteobacteria	Enterobacteriaceae
Klebsiella pneumoniae	-0.09739	GCF_004127515.1	Gammaproteobacteria	Enterobacteriaceae
Pantoea sp.	-0.09737	GCF_002920175.1	Gammaproteobacteria	Enterobacteriaceae
Klebsiella grimontii	-0.09736	GCA_902159485.1	Gammaproteobacteria	Enterobacteriaceae
Enterobacter sp.	-0.09736	GCA_007035975.1	Gammaproteobacteria	Enterobacteriaceae
Pantoea sp.	-0.09736	GCF_002313185.2	Gammaproteobacteria	Enterobacteriaceae
Pantoea sp.	-0.09736	GCF_003813865.1	Gammaproteobacteria	Enterobacteriaceae
Buttiauxella izardii	-0.09735	GCF_003601925.1	Gammaproteobacteria	Enterobacteriaceae
Klebsiella grimontii	-0.09735	GCA_902158675.1	Gammaproteobacteria	Enterobacteriaceae
Lelliottia nimipressuralis	-0.09734	GCF_004402045.1	Gammaproteobacteria	Enterobacteriaceae
Pantoea stewartii	-0.09732	GCF_001310295.1	Gammaproteobacteria	Enterobacteriaceae
Pantoea sp.	-0.09729	GCF_000963985.1	Gammaproteobacteria	Enterobacteriaceae
Rahnella aquatilis	-0.09729	GCF_000735505.1	Gammaproteobacteria	Enterobacteriaceae
Buttiauxella sp.	-0.09728	GCF_003675305.1	Gammaproteobacteria	Enterobacteriaceae
Cronobacter dublinensis	-0.09727	GCF_002978705.2	Gammaproteobacteria	Enterobacteriaceae
Erwinia sp.	-0.09726	GCF_002752575.1	Gammaproteobacteria	Enterobacteriaceae
Rouxiella chamberiensis	-0.09724	GCF_000951135.1	Gammaproteobacteria	Enterobacteriaceae
Erwinia sp.	-0.09724	GCF_004551645.1	Gammaproteobacteria	Enterobacteriaceae
Enterobacter sp.	-0.09721	GCF_000277545.1	Gammaproteobacteria	Enterobacteriaceae
Gamma proteobacterium	-0.09715	GCA_000335795.1	Gammaproteobacteria	Enterobacteriaceae
Rahnella aquatilis	-0.09711	GCA_003956145.2	Gammaproteobacteria	Enterobacteriaceae
Rahnella woolbedingensis	-0.09709	GCF_003602095.1	Gammaproteobacteria	Enterobacteriaceae
Ewingella americana	-0.09702	GCF_006438725.1	Gammaproteobacteria	Enterobacteriaceae
Ewingella americana	-0.09696	GCF_900451015.1	Gammaproteobacteria	Enterobacteriaceae
Serratia quinivorans	-0.09689	GCF_900457075.1	Gammaproteobacteria	Enterobacteriaceae
Atlantibacter hermannii	-0.09685	GCF_900635495.1	Gammaproteobacteria	Enterobacteriaceae
Serratia proteamaculans	-0.09685	GCF_004684015.1	Gammaproteobacteria	Enterobacteriaceae
Serratia sp.	-0.09684	GCF_002607755.1	Gammaproteobacteria	Enterobacteriaceae
Pseudescherichia vulneris	-0.09681	GCF_900450975.1	Gammaproteobacteria	Enterobacteriaceae
Buttiauxella sp.	-0.09678	GCF_006376615.1	Gammaproteobacteria	Enterobacteriaceae
Pantoea vagans	-0.09644	GCF_001506165.1	Gammaproteobacteria	Enterobacteriaceae
Ewingella americana	-0.09632	GCF_000735345.1	Gammaproteobacteria	Enterobacteriaceae
Pectobacterium carotovorum	-0.096	GCF_002250215.1	Gammaproteobacteria	Enterobacteriaceae

Serratia fonticola	-0.09576	GCF_006714955.1	Gammaproteobacteria	Enterobacteriaceae
Serratia sp.	-0.0957	GCF_003668775.1	Gammaproteobacteria	Enterobacteriaceae
Yersinia enterocolitica	-0.09544	GCF_002082245.2	Gammaproteobacteria	Enterobacteriaceae
Yersinia enterocolitica	-0.09536	GCF_002083285.2	Gammaproteobacteria	Enterobacteriaceae
Yersinia kristensenii	-0.09526	GCF_002188895.1	Gammaproteobacteria	Enterobacteriaceae
Morganella morganii	-0.05621	GCF_003287815.1	Gammaproteobacteria	Enterobacteriaceae
Plesiomonas sp.	-0.01279	GCF_000800945.1	Gammaproteobacteria	Enterobacteriaceae
Plesiomonas shigelloides	-0.01279	GCF_002093895.1	Gammaproteobacteria	Enterobacteriaceae
Aeromonas jandaei	-0.01248	GCF_000708125.1	Gammaproteobacteria	Aeromonadaceae
Aeromonas veronii	-0.01245	GCF_000298015.1	Gammaproteobacteria	Aeromonadaceae
Photobacterium kishitanii	-0.01235	GCF_003025945.1	Gammaproteobacteria	Vibrionaceae
Photobacterium phosphoreum	-0.01235	GCF_003025815.1	Gammaproteobacteria	Vibrionaceae
Aeromonas popoffii	-0.01224	GCF_000820025.1	Gammaproteobacteria	Aeromonadaceae

Table S3: Top 100 over-represented annotated genes in the genomes of the taxa that receive the most negative entries on variable 1. The *NES* and *FDR-Adj. P* columns show the normalized 'Enrichment score' and FDR-adjusted [156] *P*-value from the enrichment analysis [155].

Gene	**FDR-Adj. P**	**NES**
Major outer membrane lipoprotein Lpp 1	-11.844	0.00025
Phage shock protein G	-11.798	0.00025
Primosomal replication protein N"	-11.753	0.00025
DNA damage-inducible protein I	-11.694	0.00025
Outer membrane porin C	-11.614	0.00025
Chaperone protein YcdY	-11.562	0.00025
USG-1 protein	-11.506	0.00025
HTH-type transcriptional regulator cbl	-11.357	0.00025
Inner membrane protein YghB	-11.34	0.00025
Cytochrome c-type protein NrfB	-11.321	0.00025
Plasmid partition protein A	-11.288	0.00025
Protein rof	-11.241	0.00025
putative inner membrane protein Smp	-11.208	0.00025
putative protein YbjN	-11.199	0.00025
putative lipoprotein YbaY	-11.157	0.00025
putative HTH-type transcriptional regulator YbdO	-11.112	0.00025
Inner membrane protein YqjE	-11.063	0.00025
putative protein YfeY	-11.051	0.00025
Secretion monitor	-11.041	0.00025
putative ECA polymerase	-11.035	0.00025
Protein Sxy	-10.994	0.00025
Kdo(2)-lipid A phosphoethanolamine 7"-transferase	-10.986	0.00025
Intermembrane phospholipid transport system binding protein MlaB	-10.98	0.00025
putative cyclic di-GMP phosphodiesterase PdeD	-10.94	0.00025
Phosphatidylglycerophosphatase C	-10.93	0.00025
Inner membrane protein YlaC	-10.922	0.00025
Multiple stress resistance protein BhsA	-10.922	0.00025
Protein YdgH	-10.922	0.00025
Protein PhoH	-10.898	0.00025
Outer membrane protein X	-10.882	0.00025
Sensor protein BasS	-10.881	0.00025
Inner membrane protein YgbE	-10.872	0.00025
putative protein YjjI	-10.872	0.00025
Biofilm regulator BssS	-10.87	0.00025
DNA polymerase III subunit theta	-10.87	0.00025
Multidrug efflux pump accessory protein AcrZ	-10.87	0.00025
putative protein YejG	-10.87	0.00025

putative protein YhcO	-10.87	0.00025
Protein YhjJ	-10.864	0.00025
Constitutive lysine decarboxylase	-10.863	0.00025
Pirin-like protein YhaK	-10.861	0.00025
Constitutive ornithine decarboxylase	-10.843	0.00025
Inner membrane protein YdgK	-10.817	0.00025
putative cyclic di-GMP phosphodiesterase PdeK	-10.815	0.00025
Flagellar regulator flk	-10.81	0.00025
Lipoprotein BsmA	-10.81	0.00025
Modulator protein MzrA	-10.81	0.00025
Inner membrane protein YgfX	-10.808	0.00025
Protein DsrB	-10.808	0.00025
Phosphoethanolamine transferase OpgE	-10.764	0.00025
Transcriptional regulatory protein RcsA	-10.748	0.00025
Osmotically-inducible lipoprotein B	-10.748	0.00025
Cyclic di-GMP phosphodiesterase PdeH	-10.735	0.00025
Hha toxicity modulator TomB	-10.732	0.00025
Type II secretion system protein H	-10.728	0.00025
putative lipoprotein YajI	-10.69	0.00025
putative protein YebV	-10.682	0.00025
Inner membrane protein YbjO	-10.682	0.00025
putative lipoprotein YbjP	-10.682	0.00025
putative protein YccJ	-10.663	0.00025
Inner membrane protein YfeZ	-10.655	0.00025
Flagella synthesis protein FlgN	-10.649	0.00025
Periplasmic chaperone Spy	-10.637	0.00025
putative ferredoxin-like protein YdhX	-10.629	0.00025
Regulatory protein SoxS	-10.485	0.00025
Protein TonB	-10.47	0.00025
Cyclic di-GMP binding protein BcsE	-10.384	0.00025
Quorum-sensing regulator protein G	-10.381	0.00025
Signal transduction histidine-protein kinase/phosphatase UhpB	-10.368	0.00025
Regulator of sigma S factor FliZ	-10.328	0.00025
Inner membrane protein YbjM	-10.323	0.00025
Trimethylamine-N-oxide reductase	-10.318	0.00025
Ferric iron reductase protein FhuF	-10.31	0.00025
putative protein YaiA	-10.29	0.00025
Negative regulator of flagellin synthesis	-10.282	0.00025
Anti-adapter protein IraP	-10.27	0.00025
Ferric enterobactin transport protein FepE	-10.235	0.00025
Intracellular growth attenuator protein igaA	-10.225	0.00025
Outer membrane porin N	-10.133	0.00025
Flagellar protein FlhE	-10.095	0.00025
Inner membrane protein YebE	-10.087	0.00025
HTH-type transcriptional repressor BluR	-10.05	0.00025
Enterobactin synthase component F	-10.035	0.00025
Sec-independent protein translocase protein TatE	-9.992	0.00025
Alternative ribosome-rescue factor A	-9.987	0.00025
Primosomal protein 1	-9.919	0.00025
putative csgAB operon transcriptional regulatory protein	-9.909	0.00025
putative protein YgaM	-9.887	0.00025
Multiple antibiotic resistance protein MarA	-9.846	0.00025
Inner membrane protein YbhQ	-9.832	0.00025
Inner membrane protein YjiG	-9.829	0.00025

Putative selenoprotein YdfZ	-9.827	0.00025
HTH-type transcriptional regulator MlrA	-9.802	0.00025
Tryptophanase	-9.8	0.00025
Universal stress protein C	-9.798	0.00025
Hemolysin expression-modulating protein Hha	-9.771	0.00025
6-phospho-beta-glucosidase BglB	-9.762	0.00025
Ribonucleoside-diphosphate reductase 2 subunit alpha	-9.711	0.00025
Transcriptional regulator SutA	-9.71	0.00025
Cyclic di-GMP binding protein	-9.705	0.00025

Table S4: Top 100 over-represented annotated genes in the genomes of the taxa that receive the most positive entries on variable 2. The column descriptions are provided with Supplementary Table 3.

Gene	**FDR-Adj. P**	**NES**
Outer-membrane lipoprotein LolB	15.094	0.0004
DNA-binding protein Fis	15.02	0.0004
LPS-assembly lipoprotein LptE	15.004	0.0004
Ammonia monooxygenase gamma subunit	14.873	0.0004
Thiol:disulfide interchange protein DsbA	14.848	0.0004
DNA polymerase III subunit delta	14.719	0.0004
Phosphate regulon sensor protein PhoR	14.697	0.0004
Pyruvate kinase II	14.344	0.0004
1,6-anhydro-N-acetylmuramyl-L-alanine amidase AmpD	14.232	0.0004
Cell division protein FtsN	14.223	0.0004
Stringent starvation protein A	14.185	0.0004
Lipopolysaccharide export system permease protein LptF	14.002	0.0004
Succinate dehydrogenase hydrophobic membrane anchor subunit	13.978	0.0004
Recombination-associated protein RdgC	13.944	0.0004
Modulator of FtsH protease YccA	13.94	0.0004
Cell division protein ZipA	13.785	0.0004
Cbb3-type cytochrome c oxidase subunit CcoN1	13.784	0.0004
Cytochrome c4	13.777	0.0004
2-octaprenylphenol hydroxylase	13.656	0.0004
2-octaprenyl-6-methoxyphenol hydroxylase	13.643	0.0004
Phosphoenolpyruvate synthase regulatory protein	13.401	0.0004
HTH-type transcriptional regulator CysB	13.346	0.0004
2-methyl-aconitate isomerase	13.193	0.0004
Ferredoxin 1	13.12	0.0004
Intermembrane phospholipid transport system ATP-binding protein MlaF	13.091	0.0004
Large ribosomal RNA subunit accumulation protein YceD	13.051	0.0004
Cytoskeleton protein RodZ	13.043	0.0004
Chorismate pyruvate-lyase	12.981	0.0004
Soluble lytic murein transglycosylase	12.981	0.0004
Intermembrane phospholipid transport system binding protein MlaD	12.951	0.0004
2Fe-2S ferredoxin	12.898	0.0004
Cell division protein DedD	12.872	0.0004
Protein YcgL	12.871	0.0004
Uroporphyrinogen-III synthase	12.748	0.0004
High frequency lysogenization protein HflD	12.712	0.0004
putative protein YaeQ	12.679	0.0004
Exodeoxyribonuclease I	12.653	0.0004
Co-chaperone protein HscB	12.641	0.0004

Phosphatase NudJ	12.606	0.0004
Glutamate-pyruvate aminotransferase AlaA	12.599	0.0004
Ribonuclease T	12.541	0.0004
Intermembrane phospholipid transport system binding protein MlaC	12.528	0.0004
Colicin V production protein	12.514	0.0004
5-amino-6-(5-phospho-D-ribitylamino)uracil phosphatase YigB	12.51	0.0004
Fimbrial protein	12.493	0.0004
Inner membrane transport protein YajR	12.469	0.0004
50S ribosomal protein L16 3-hydroxylase	12.462	0.0004
Protein phosphatase CheZ	12.411	0.0004
FKBP-type 16 kDa peptidyl-prolyl cis-trans isomerase	12.376	0.0004
putative protein YibN	12.342	0.0004
Molybdopterin-synthase adenylyltransferase	12.331	0.0004
Inner membrane protein YpjD	12.312	0.0004
Iron-sulfur cluster assembly protein CyaY	12.218	0.0004
Sigma factor AlgU regulatory protein MucB	12.217	0.0004
Protein-glutamine gamma-glutamyltransferase	12.205	0.0004
Pyrimidine/purine nucleotide 5'-monophosphate nucleosidase	12.141	0.0004
Bifunctional (p)ppGpp synthase/hydrolase SpoT	12.134	0.0004
Putative glutamine amidotransferase YafJ	12.126	0.0004
Penicillin-binding protein 1B	12.077	0.0004
CDP-6-deoxy-L-threo-D-glycero-4-hexulose-3-dehydrase reductase	12.042	0.0004
Peptidoglycan hydrolase FlgJ	12.011	0.0004
Methylmalonate-semialdehyde dehydrogenase [acylating]	11.998	0.0004
Acid stress protein IbaG	11.995	0.0004
Ribosome modulation factor	11.907	0.0004
Cbb3-type cytochrome c oxidase subunit CcoP2	11.904	0.0004
ATP synthase subunit beta 1	11.902	0.0004
3-deoxy-D-manno-octulosonic acid kinase	11.79	0.0004
ADP compounds hydrolase NudE	11.777	0.0004
HTH-type transcriptional regulator YhaJ	11.776	0.0004
Nucleoid-associated protein YejK	11.714	0.0004
Sensor protein QseC	11.713	0.0004
UTP pyrophosphatase	11.706	0.0004
ATP-dependent DNA helicase DinG	11.685	0.0004
Riboflavin transporter	11.672	0.0004
Membrane-bound lytic murein transglycosylase B	11.611	0.0004
putative acyltransferase YihG	11.598	0.0004
Protein ImuB	11.561	0.0004
HTH-type transcriptional regulator HmrR	11.528	0.0004
DNA polymerase III subunit chi	11.502	0.0004
Flagellar basal-body rod protein FlgF	11.498	0.0004
Outer membrane protein assembly factor BamC	11.498	0.0004
Aerotaxis receptor	11.478	0.0004
Protein-glutamate methylesterase/protein-glutamine glutaminase 1	11.468	0.0004
Dihydroorotase-like protein	11.451	0.0004
Protein YciI	11.451	0.0004
Cell division protein ZapD	11.45	0.0004
Peptidyl-prolyl cis-trans isomerase cyp18	11.442	0.0004
FAD assembly factor SdhE	11.418	0.0004
D-erythrose-4-phosphate dehydrogenase	11.393	0.0004
putative DNA endonuclease SmrA	11.37	0.0004
Protein Smg	11.35	0.0004
Type II secretion system protein K	11.347	0.0004

Cysteine synthase A	11.339	0.0004
Exoribonuclease 2	11.332	0.0004
Chaperone protein HscA	11.319	0.0004
tRNA/tmRNA (uracil-C(5))-methyltransferase	11.319	0.0004
Murein hydrolase activator NlpD	11.28	0.0004
Methyl-accepting chemotaxis protein McpP	11.26	0.0004
Glutathione S-transferase GST-6.0	11.243	0.0004
Methyl-accepting chemotaxis protein McpH	11.228	0.0004

Table S5: Top 100 over-represented annotated genes in the genomes of the taxa that receive the most positive entries on variable 3. The column descriptions are provided with Supplementary Table 3.

Gene	**FDR-Adj. P**	**NES**
GTP pyrophosphokinase rsh	12.285	0.00054
putative peptidoglycan D,D-transpeptidase FtsI	12.172	0.00054
Chromosome-partitioning protein ParB	12.152	0.00054
5-aminolevulinate synthase	12.02	0.00054
FtsZ-localized protein C	11.923	0.00054
6,7-dimethyl-8-ribityllumazine synthase 1	11.812	0.00054
FtsZ-localized protein A	11.775	0.00054
Aerobic cobaltochelatase subunit CobT	11.742	0.00054
Phyllosphere-induced regulator PhyR	11.705	0.00054
NADH-quinone oxidoreductase chain 1	11.649	0.00054
Ubiquinone hydroxylase UbiL	11.525	0.00054
Cell cycle response regulator CtrA	11.398	0.00054
Protein phosphotransferase ChpT	11.394	0.00054
Heat shock protein HspQ	11.378	0.00054
Aerobic cobaltochelatase subunit CobS	11.303	0.00054
Ferredoxin-2	11.301	0.00054
Dihydrolipoyl dehydrogenase 3	11.181	0.00054
flagellum biosynthesis repressor protein FlbT	11.152	0.00054
Cytochrome c oxidase subunit 1 , bacteroid	11.079	0.00054
Thiol:disulfide interchange protein CycY	11.068	0.00054
putative protein RP812	10.703	0.00054
Polyphosphate:NDP phosphotransferase 3	10.652	0.00054
Serine hydroxymethyltransferase 2	10.633	0.00054
Transcriptional regulatory protein ros	10.535	0.00054
Ferredoxin-6	10.444	0.00054
Glutamate–cysteine ligase EgtA	10.444	0.00054
Propionyl-CoA carboxylase regulator	10.367	0.00054
RNA polymerase sigma-54 factor 2	10.354	0.00054
Bifunctional enzyme IspD/IspF	10.297	0.00054
Cytochrome c1	10.28	0.00054
Cold shock protein CspA	10.25	0.00054
Thiol:disulfide interchange protein TlpA	10.212	0.00054
Nitrogen fixation regulation protein FixK	10.079	0.00054
Cytochrome c oxidase subunit 1-beta	9.815	0.00054
NADH-quinone oxidoreductase chain 5	9.719	0.00054
(3S)-malyl-CoA thioesterase	9.564	0.00054
UDP-2,3-diacylglucosamine pyrophosphatase LpxI	9.371	0.00054
ATP synthase protein I	9.331	0.00054
Hypotaurine/taurine–pyruvate aminotransferase	9.33	0.00054

Blue-light-activated histidine kinase	9.329	0.00054
HTH-type transcriptional regulator RamB	9.327	0.00054
Porin	9.325	0.00054
Penicillin-insensitive murein endopeptidase	9.315	0.00054
Glycine betaine methyltransferase	9.294	0.00054
L-arabinose 1-dehydrogenase (NAD(P)(+))	9.267	0.00054
Putative metal-sulfur cluster biosynthesis proteins YuaD	9.264	0.00054
ATP synthase subunit b'	9.25	0.00054
N-acetylmuramoyl-L-alanine amidase AmiD	9.212	0.00054
Periplasmic alpha-galactoside-binding protein	9.195	0.00054
10 kDa chaperonin 1	9.186	0.00054
Urease subunit gamma 1	9.152	0.00054
(2S)-methylsuccinyl-CoA dehydrogenase	9.097	0.00054
Urease subunit alpha 1	9.066	0.00054
Hemolysin C	9.047	0.00054
Urease accessory protein UreE 1	8.894	0.00054
Lysine/ornithine decarboxylase	8.874	0.00054
Dicamba O-demethylase 1, ferredoxin reductase component	8.873	0.00054
Serine–glyoxylate aminotransferase	8.855	0.00054
Precorrin-3B C(17)-methyltransferase	8.849	0.00054
Nicotinate phosphoribosyltransferase	8.819	0.00054
Glycogen synthase 1	8.812	0.00054
3-hydroxybenzoate 6-hydroxylase 1	8.768	0.00054
nicotinate-nucleotide adenylyltransferase	8.741	0.00054
Molybdenum cofactor insertion chaperone PaoD	8.735	0.00054
Arginine–pyruvate transaminase AruH	8.718	0.00054
D-hydantoinase/dihydropyrimidinase	8.695	0.00054
putative 3-hydroxyisobutyrate dehydrogenase	8.592	0.00054
Response regulator receiver protein CpdR	8.56	0.00054
60 kDa chaperonin 5	8.547	0.00054
Lysophospholipase L2	8.545	0.00054
Type I secretion system ATP-binding protein PrsD	8.544	0.00054
Phosphatidylcholine synthase	8.531	0.00054
Carbonic anhydrase 1	8.507	0.00054
Bifunctional coenzyme PQQ synthesis protein C/D	8.503	0.00054
NAD-dependent dihydropyrimidine dehydrogenase subunit PreT	8.474	0.00054
Anti-sigma-F factor NrsF	8.456	0.00054
FAD-dependent catabolic D-arginine dehydrogenase DauA	8.437	0.00054
Nopaline-binding periplasmic protein	8.432	0.00054
Mesaconyl-CoA hydratase	8.431	0.00054
putative riboflavin import permease protein RfuD	8.419	0.00054
Sulfite dehydrogenase subunit C	8.391	0.00054
Crotonyl-CoA carboxylase/reductase	8.381	0.00054
Precorrin-2 C(20)-methyltransferase	8.38	0.00054
Acyl carrier protein AcpXL	8.376	0.00054
Polysialic acid transport ATP-binding protein KpsT	8.372	0.00054
Alkane 1-monooxygenase 2	8.352	0.00054
HTH-type transcriptional regulator RafR	8.331	0.00054
S-formylglutathione hydrolase	8.326	0.00054
Ethylmalonyl-CoA mutase	8.282	0.00054
Outer membrane protein	8.273	0.00054
Cytochrome c oxidase subunit 4	8.252	0.00054
Cytochrome c-556	8.237	0.00054
Alpha-D-ribose 1-methylphosphonate 5-triphosphate synthase subunit PhnG	8.236	0.00054

Alpha-D-ribose 1-methylphosphonate 5-triphosphate diphosphatase	8.223	0.00054
Sulfite dehydrogenase subunit A	8.218	0.00054
Glutathione-specific gamma-glutamylcyclotransferase	8.17	0.00054
(S)-ureidoglycine aminohydrolase	8.165	0.00054
Hydrogenobyrinate a,c-diamide synthase	8.101	0.00054
Alanine racemase, biosynthetic	8.068	0.00054
Alpha-D-ribose 1-methylphosphonate 5-triphosphate synthase subunit PhnH	8.045	0.00054

Table S6: Top 100 over-represented annotated genes in the genomes of the taxa that receive the most negative entries on variable 4. The column descriptions are provided with Supplementary Table 3.

Gene	**FDR-Adj. P**	**NES**
Cytochrome f	-6.889	0.00031
Photosystem II manganese-stabilizing polypeptide	-6.876	0.00031
Protein Thf1	-6.874	0.00031
Photosystem II CP47 reaction center protein	-6.859	0.00031
NAD(P)H-quinone oxidoreductase subunit N	-6.857	0.00031
Photosystem I assembly protein Ycf4	-6.857	0.00031
Photosystem I reaction center subunit III	-6.857	0.00031
Phycocyanobilin:ferredoxin oxidoreductase	-6.85	0.00031
Photosystem II reaction center Psb28 protein	-6.836	0.00031
Photosystem II lipoprotein Psb27	-6.835	0.00031
NAD(P)H-quinone oxidoreductase subunit O	-6.833	0.00031
Photosystem I P700 chlorophyll a apoprotein A1	-6.833	0.00031
Photosystem II reaction center protein K	-6.829	0.00031
Cytochrome b559 subunit alpha	-6.826	0.00031
Photosystem II CP43 reaction center protein	-6.826	0.00031
Photosystem I reaction center subunit IV	-6.824	0.00031
30S ribosomal protein S21 A	-6.812	0.00031
Pentapeptide repeat protein Rfr32	-6.807	0.00031
Photosystem II reaction center protein H	-6.804	0.00031
Ycf54-like protein	-6.803	0.00031
Ferredoxin-thioredoxin reductase, catalytic chain	-6.802	0.00031
NAD(P)H-quinone oxidoreductase subunit M	-6.795	0.00031
Long-chain acyl-[acyl-carrier-protein] reductase	-6.795	0.00031
Bifunctional pantoate ligase/cytidylate kinase	-6.792	0.00031
RNA polymerase sigma factor SigA2	-6.78	0.00031
Photosystem I reaction center subunit II	-6.779	0.00031
Photosystem I reaction center subunit XI	-6.779	0.00031
Phycobiliprotein beta chain	-6.773	0.00031
Phycobilisome 7.8 kDa linker polypeptide, allophycocyanin-associated, core	-6.773	0.00031
Photosystem I reaction center subunit XII	-6.769	0.00031
Aldehyde decarbonylase	-6.752	0.00031
Photosystem II reaction center protein Z	-6.75	0.00031
Ferredoxin-thioredoxin reductase, variable chain	-6.739	0.00031
Photosystem II 12 kDa extrinsic protein	-6.73	0.00031
Photosystem II protein Y	-6.723	0.00031
Protein PsbN	-6.723	0.00031
Proton extrusion protein PcxA	-6.719	0.00031
Cytochrome b6-f complex subunit 7	-6.698	0.00031
Phycocyanobilin lyase subunit alpha	-6.692	0.00031
Orange carotenoid-binding protein	-6.684	0.00031

Cytochrome b559 subunit beta	-6.665	0.00031
Photosystem I iron-sulfur center	-6.638	0.00031
Vitamin K epoxide reductase	-6.628	0.00031
Photosystem I reaction center subunit IX	-6.626	0.00031
ATP-dependent zinc metalloprotease FtsH 2	-6.612	0.00031
Putative diflavin flavoprotein A 3	-6.577	0.00031
putative glutaredoxin	-6.556	0.00031
Monoglucosyldiacylglycerol epimerase	-6.546	0.00031
Photosystem II protein D1 2	-6.546	0.00031
NAD(P)H-quinone oxidoreductase subunit L	-6.543	0.00031
Lipoyl synthase 2	-6.517	0.00031
Phycocyanobilin lyase CpcT	-6.495	0.00031
2-methyl-6-phytyl-1,4-hydroquinone methyltransferase	-6.492	0.00031
Photosystem II reaction center X protein	-6.491	0.00031
Photosystem II D2 protein	-6.47	0.00031
Transcription regulator LexA	-6.463	0.00031
Ycf53-like protein	-6.463	0.00031
D-fructose 1,6-bisphosphatase class 2/sedoheptulose 1,7-bisphosphatase	-6.447	0.00031
putative arabinosyltransferase C	-6.445	0.00031
Phycobiliprotein ApcE	-6.435	0.00031
Photosystem II reaction center protein T	-6.434	0.00031
NAD(P)H-quinone oxidoreductase subunit K 1	-6.424	0.00031
Sensor protein SphS	-6.419	0.00031
Allophycocyanin beta chain	-6.412	0.00031
Photosystem II reaction center protein M	-6.396	0.00031
Photosystem II reaction center protein Ycf12	-6.39	0.00031
Circadian clock protein KaiA	-6.388	0.00031
Photosystem I reaction center subunit VIII	-6.371	0.00031
High-affinity Na(+)/H(+) antiporter NhaS3	-6.34	0.00031
putative 30S ribosomal protein PSRP-3	-6.308	0.00031
Isoaspartyl peptidase/L-asparaginase	-6.297	0.00031
Phytol kinase	-6.261	0.00031
Regulatory protein CysR	-6.248	0.00031
Photosystem II reaction center protein I	-6.243	0.00031
Serine/threonine-protein kinase B	-6.237	0.00031
Ferredoxin-dependent glutamate synthase 2	-6.229	0.00031
Phycobilisome rod-core linker polypeptide CpcG	-6.226	0.00031
Galactan 5-O-arabinofuranosyltransferase	-6.217	0.00031
NAD(P)H-quinone oxidoreductase subunit J	-6.213	0.00031
Putative serine protease HhoA	-6.174	0.00031
Phycocyanobilin lyase subunit CpcS	-6.151	0.00031
Photosystem II reaction center protein J	-6.149	0.00031
Putative acetyl-coenzyme A carboxylase carboxyl transferase subunit beta	-6.134	0.00031
Phycocyanobilin lyase subunit beta	-6.125	0.00031
NAD(P)H-quinone oxidoreductase subunit I	-6.123	0.00031
Phosphoribulokinase	-6.118	0.00031
Putative isochorismate synthase MenF	-6.103	0.00031
Allophycocyanin subunit alpha-B	-6.097	0.00031
Putative cytochrome P450 120	-6.075	0.00031
Putative diflavin flavoprotein A 5	-6.071	0.00031
Bicarbonate-binding protein CmpA	-6.071	0.00031
C-phycocyanin beta chain	-6.068	0.00031
Chromophore lyase CpcS/CpeS	-6.065	0.00031
4-hydroxybenzoate solanesyltransferase	-6.034	0.00031

Hydrolase	-6.024	0.00031
L,D-transpeptidase 2	-6.004	0.00031
putative ferredoxin/ferredoxin–NADP reductase	-5.991	0.00031
Photosystem II reaction center protein L	-5.972	0.00031
Alpha-(1-¿3)-arabinofuranosyltransferase	-5.963	0.00031
Serine/threonine-protein kinase F	-5.943	0.00031

Table S7: Top 100 over-represented annotated genes in the genomes of the taxa that receive the most negative entries on variable 14. The column descriptions are provided with Supplementary Table 3.

Gene	**FDR-Adj. P**	**NES**
Thymidylate synthase 1	-5.522	0.00055
DNA-binding protein Bv3F	-5.446	0.00055
mupirocin-resistant isoleucine–tRNA ligase MupA	-5.294	0.00055
Cobalt-precorrin-7 C(5)-methyltransferase	-5.114	0.00055
Cobalt-zinc-cadmium resistance protein CzcI	-5.103	0.00055
IS200/IS605 family transposase ISCth10	-4.897	0.00055
Na(+)-translocating ferredoxin:NAD(+) oxidoreductase complex subunit C	-4.843	0.00055
Anaerobic sulfite reductase subunit C	-4.814	0.00055
(R)-2-hydroxyglutaryl-CoA dehydratase activating ATPase	-4.812	0.00055
Salicylate 5-hydroxylase, large oxygenase component	-4.778	0.00055
Tyrosine aminotransferase	-4.765	0.00055
putative protein YgcP	-4.703	0.00055
Alkaline phosphatase PhoK	-4.685	0.00055
Salicylate 5-hydroxylase, small oxygenase component	-4.684	0.00055
Propanediol utilization protein PduU	-4.64	0.00055
Putative superoxide reductase	-4.596	0.00055
Sortase B	-4.552	0.00055
Propanediol utilization protein PduV	-4.424	0.00055
Elongation factor G, mitochondrial	-4.343	0.00055
Germination protease	-4.315	0.00055
Stage IV sporulation protein A	-4.315	0.00055
Stage V sporulation protein AD	-4.315	0.00055
putative N-glycosylase/DNA lyase	-4.309	0.00055
(R)-phenyllactyl-CoA dehydratase alpha subunit	-4.297	0.00055
Nickel-cobalt-cadmium resistance protein NccX	-4.295	0.00055
(R)-2-hydroxyglutaryl-CoA dehydratase, subunit beta	-4.293	0.00055
Putative transport protein YbjL	-4.292	0.00055
RNA polymerase sigma-G factor	-4.286	0.00055
Translocation-enhancing protein TepA	-4.271	0.00055
Stage V sporulation protein T	-4.261	0.00055
Light-activated DNA-binding protein EL222	-4.254	0.00055
RNA polymerase sigma-28 factor	-4.245	0.00055
putative anti-sigma-F factor NrsF	-4.242	0.00055
Oxalate-binding protein	-4.225	0.00055
IS256 family transposase ISCth4	-4.222	0.00055
Propanediol utilization protein PduB	-4.202	0.00055
Stage III sporulation protein D	-4.195	0.00055
Spore protein YabP	-4.194	0.00055
3,4-dehydroadipyl-CoA semialdehyde dehydrogenase	-4.18	0.00055
Nickel and cobalt resistance protein CnrR	-4.118	0.00055
Neopullulanase 1	-4.117	0.00055

Small, acid-soluble spore protein C2	-4.111	0.00055
Mini-ribonuclease 3-like protein	-4.103	0.00055
L-threonine kinase	-4.09	0.00055
IS3 family transposase ISStma17	-4.039	0.00055
Outer membrane protein 40	-4.031	0.00055
Methionine-rich peptide X	-3.993	0.00055
IS5 family transposase ISBmu20	-3.97	0.00055
Reverse rubrerythrin-1	-3.967	0.00055
Histidine racemase	-3.951	0.00055
Spore germination protein B1	-3.947	0.00055
2-pyrone-4,6-dicarboxylate hydrolase	-3.936	0.00055
Glycine/sarcosine/betaine reductase complex component C subunit alpha	-3.931	0.00055
Tryptophanase 1	-3.919	0.00055
Nickel and cobalt resistance protein CnrC	-3.918	0.00055
putative sporulation protein YlmC	-3.91	0.00055
mupirocin-resistant isoleucine–tRNA ligase MupB	-3.907	0.00055
putative deoxyuridine 5'-triphosphate nucleotidohydrolase YncF	-3.89	0.00055
IS110 family transposase ISCaa14	-3.886	0.00055
SpoIVB peptidase	-3.885	0.00055
Glycine reductase complex component B subunit gamma	-3.874	0.00055
Propionate catabolism operon regulatory protein	-3.87	0.00055
RNA polymerase sigma-35 factor	-3.865	0.00055
Propanediol dehydratase medium subunit	-3.852	0.00055
Chloroacetanilide N-alkylformylase, ferredoxin reductase component	-3.835	0.00055
Ribulose bisphosphate carboxylase large chain, chromosomal	-3.835	0.00055
IS66 family transposase ISBcen14	-3.824	0.00055
putative tryptophan transport protein	-3.82	0.00055
IS66 family transposase ISBcen19	-3.809	0.00055
Peptidoglycan-N-acetylmuramic acid deacetylase PdaA	-3.808	0.00055
Light-harvesting protein B-870 beta chain	-3.798	0.00055
PEP-dependent dihydroxyacetone kinase 2, phosphoryl donor subunit DhaM	-3.796	0.00055
Glycine reductase complex component B subunits alpha and beta	-3.793	0.00055
IS1182 family transposase ISCpe5	-3.79	0.00055
Accessory gene regulator protein B	-3.788	0.00055
CRISPR-associated endoribonuclease Cas6	-3.787	0.00055
Phosphoglycolate phosphatase, plasmid	-3.786	0.00055
Glycerol dehydratase large subunit	-3.781	0.00055
Glycine/sarcosine/betaine reductase complex component C subunit beta	-3.773	0.00055
Diol dehydratase-reactivating factor alpha subunit	-3.759	0.00055
IS3 family transposase ISElsp1	-3.754	0.00055
Stage II sporulation protein E	-3.754	0.00055
IS110 family transposase ISCaa7	-3.754	0.00055
Propanediol dehydratase small subunit	-3.748	0.00055
Reverse rubrerythrin-2	-3.747	0.00055
Cytochrome c-type protein SHP	-3.744	0.00055
Antigen TpF1	-3.731	0.00055
Serine/threonine-protein kinase CtkA	-3.731	0.00055
Diadenosine hexaphosphate hydrolase	-3.73	0.00055
Glycine/sarcosine/betaine reductase complex component A	-3.724	0.00055
Outer membrane protein 41	-3.721	0.00055
Stage III sporulation protein AE	-3.703	0.00055
N-acetylmuramoyl-L-alanine amidase	-3.7	0.00055
Catechol 1,2-dioxygenase 2	-3.696	0.00055
Glycine/sarcosine/betaine reductase complex component A1	-3.678	0.00055

D-proline reductase proprotein PrdA	-3.654	0.00055
Catechol 1,2-dioxygenase 1	-3.633	0.00055
Phthalate 4,5-dioxygenase oxygenase reductase subunit	-3.63	0.00055
Iron hydrogenase 1	-3.623	0.00055
Metal-staphylopine import system ATP-binding protein CntD	-3.618	0.00055

Table S8: Top 100 over-represented annotated genes in the genomes of the taxa that receive the most negative entries on variable 38. The column descriptions are provided with Supplementary Table 3.

Gene	**FDR-Adj. P**	**NES**
Bifunctional protein MdtA	-5.039	0.00064
Flagellar assembly protein FliX	-4.91	0.00064
Presqualene diphosphate synthase	-4.54	0.00064
mupirocin-resistant isoleucine–tRNA ligase MupA	-4.537	0.00064
Formyltransferase/hydrolase complex subunit D	-4.476	0.00064
Formyltransferase/hydrolase complex Fhc subunit A	-4.455	0.00064
Methenyltetrahydromethanopterin cyclohydrolase	-4.406	0.00064
3',5'-cyclic-nucleotide phosphodiesterase	-4.329	0.00064
Formyltransferase/hydrolase complex Fhc subunit C	-4.301	0.00064
Methylmalonyl-CoA mutase small subunit	-4.277	0.00064
Bifunctional dihydropteroate synthase/dihydropteroate reductase	-4.27	0.00064
Plasminogen-binding protein PgbB	-4.237	0.00064
Oxygen-independent coproporphyrinogen-III oxidase-like protein HemZ	-4.205	0.00064
GTP cyclohydrolase 1 type 2	-4.195	0.00064
5,6,7,8-tetrahydromethanopterin hydro-lyase	-4.19	0.00064
Methanol dehydrogenase [cytochrome c] subunit 2	-4.181	0.00064
Sensor protein DivL	-4.173	0.00064
Hydroxycarboxylate dehydrogenase B	-4.148	0.00064
Sortase B	-4.142	0.00064
2-amino-5-chloromuconate deaminase	-4.127	0.00064
L-hydantoinase	-4.024	0.00064
Cytochrome c-L	-4.022	0.00064
Flagellar FliL protein	-4.012	0.00064
Putative ATP-dependent DNA helicase YjcD	-3.986	0.00064
Beta-methylmalyl-CoA dehydratase	-3.968	0.00064
Cytochrome c-553	-3.959	0.00064
(2R)-sulfolactate sulfo-lyase subunit alpha	-3.955	0.00064
Malyl-CoA/beta-methylmalyl-CoA/citramalyl-CoA lyase	-3.953	0.00064
Oxalate:formate antiporter	-3.934	0.00064
Inducible ornithine decarboxylase	-3.921	0.00064
Dihydromethanopterin reductase	-3.891	0.00064
10 kDa chaperonin 2	-3.884	0.00064
Na(+)-translocating ferredoxin:NAD(+) oxidoreductase complex subunit G	-3.882	0.00064
Lipoprotein NlpI	-3.873	0.00064
Bifunctional DNA-directed RNA polymerase subunit beta-beta'	-3.83	0.00064
Inner membrane protein YabI	-3.823	0.00064
Cbb3-type cytochrome c oxidase subunit FixP	-3.797	0.00064
Bifunctional coenzyme PQQ synthesis protein C/D	-3.774	0.00064
Opacity-associated protein OapA	-3.769	0.00064
Surface-adhesin protein E	-3.769	0.00064
Accessory gene regulator protein B	-3.752	0.00064
Blue-light absorbing proteorhodopsin	-3.709	0.00064

Beta-(1–¿2)glucan export ATP-binding/permease protein NdvA	-3.708	0.00064
Chaperone protein YcdY	-3.702	0.00064
Ribulose bisphosphate carboxylase large chain 2	-3.684	0.00064
Outer membrane protein P5	-3.68	0.00064
Chloramphenicol resistance protein CraA	-3.654	0.00064
Na(+)-translocating ferredoxin:NAD(+) oxidoreductase complex subunit C	-3.644	0.00064
Neopullulanase 1	-3.64	0.00064
DNA transformation protein TfoX	-3.632	0.00064
Beta-carotene 15,15'-dioxygenase	-3.631	0.00064
Redox-sensing transcriptional repressor Rex 1	-3.631	0.00064
Metallopeptidase AprA	-3.629	0.00064
Rubrerythrin-1	-3.628	0.00064
Hydrogenase/urease maturation factor HypB	-3.615	0.00064
Small, acid-soluble spore protein C2	-3.598	0.00064
Ribulose bisphosphate carboxylase small chain 2	-3.595	0.00064
Translocation-enhancing protein TepA	-3.591	0.00064
Gamma-glutamyl-L-1-hydroxyisopropylamide hydrolase	-3.587	0.00064
putative protein YgcP	-3.566	0.00064
RNA polymerase sigma-28 factor	-3.526	0.00064
PTS system N-acetylglucosamine-specific EIIB component	-3.52	0.00064
Na(+)-translocating ferredoxin:NAD(+) oxidoreductase complex subunit D	-3.513	0.00064
Germination protease	-3.511	0.00064
Stage IV sporulation protein A	-3.511	0.00064
Squalene–hopene cyclase	-3.508	0.00064
putative cobalt-factor III C(17)-methyltransferase	-3.491	0.00064
USG-1 protein	-3.485	0.00064
Valine dehydrogenase	-3.482	0.00064
Molybdenum storage protein subunit alpha	-3.479	0.00064
Stage V sporulation protein AD	-3.476	0.00064
RNA polymerase sigma-G factor	-3.474	0.00064
5-(methylthio)ribulose-1-phosphate aldolase	-3.473	0.00064
Translational regulator CsrA2	-3.469	0.00064
Spore protein YabP	-3.459	0.00064
Stage V sporulation protein T	-3.446	0.00064
Oxalyl-CoA decarboxylase	-3.444	0.00064
Quinone-reactive Ni/Fe-hydrogenase large chain	-3.442	0.00064
Translational regulator CsrA1	-3.44	0.00064
L-proline trans-4-hydroxylase	-3.437	0.00064
Undecaprenyl-diphosphooligosaccharide–protein glycotransferase	-3.428	0.00064
D(-)-tartrate dehydratase	-3.423	0.00064
Resuscitation-promoting factor Rpf	-3.42	0.00064
Nucleoid-associated protein Lsr2	-3.405	0.00064
Elongation factor G, mitochondrial	-3.399	0.00064
Potassium/sodium uptake protein NtpJ	-3.398	0.00064
Malate synthase	-3.396	0.00064
Formyltransferase/hydrolase complex Fhc subunit B	-3.394	0.00064
Stage III sporulation protein D	-3.381	0.00064
Putative septation protein SpoVG	-3.369	0.00064
Hemolysin C	-3.365	0.00064
Tyrosine recombinase XerH	-3.355	0.00064
NAD(P)-dependent methylenetetrahydromethanopterin dehydrogenase	-3.355	0.00064
Oxalate decarboxylase OxdD	-3.353	0.00064
Cobalt-dependent inorganic pyrophosphatase	-3.352	0.00064
60 kDa chaperonin 3	-3.351	0.00064

Protein PhoH	-3.35	0.00064
Glutathione amide-dependent peroxidase	-3.342	0.00064
putative quinol monooxygenase YgiN	-3.326	0.00064
DNA-binding protein HB1	-3.326	0.00064

Table S9: Top 100 over-represented annotated genes in the genomes of the taxa that receive the most positive entries on variable 43. The column descriptions are provided with Supplementary Table 3.

Gene	**FDR-Adj. P**	**NES**
Sirohydrochlorin cobaltochelatase CbiKP	5.05	0.00064
Bifunctional protein MdtA	4.982	0.00064
Cytochrome c-L	4.843	0.00064
Methanol dehydrogenase [cytochrome c] subunit 2	4.798	0.00064
Carbon monoxide dehydrogenase 1	4.71	0.00064
NAD(+)–dinitrogen-reductase ADP-D-ribosyltransferase	4.631	0.00064
Carbon monoxide dehydrogenase/acetyl-CoA synthase subunit alpha	4.487	0.00064
Corrinoid/iron-sulfur protein large subunit	4.47	0.00064
Hydrogenase-2 large chain	4.455	0.00064
Molybdenum storage protein subunit beta	4.406	0.00064
Sulfite reductase, dissimilatory-type subunit gamma	4.308	0.00064
Acetolactate synthase isozyme 1 small subunit	4.265	0.00064
Protein DsvD	4.254	0.00064
mupirocin-resistant isoleucine–tRNA ligase MupA	4.214	0.00064
Hopanoid C-3 methylase	4.162	0.00064
Menaquinone reductase, iron-sulfur cluster-binding subunit	4.161	0.00064
Metal-binding protein SmbP	4.145	0.00064
Menaquinone reductase, molybdopterin-binding-like subunit	4.065	0.00064
Reverse rubrerythrin-1	4.054	0.00064
Rubredoxin 3	4.026	0.00064
Hydrogenase-2 small chain	3.991	0.00064
Ribulose bisphosphate carboxylase small chain 2	3.969	0.00064
Periplasmic [NiFe] hydrogenase large subunit	3.965	0.00064
Ribulose bisphosphate carboxylase large chain 2	3.918	0.00064
Menaquinone reductase, multiheme cytochrome c subunit	3.913	0.00064
(R)-2-hydroxyisocaproyl-CoA dehydratase beta subunit	3.895	0.00064
Formyltransferase/hydrolase complex Fhc subunit B	3.822	0.00064
Toluene-4-monooxygenase system, ferredoxin component	3.812	0.00064
Hydroxylamine oxidoreductase	3.79	0.00064
Menaquinone reductase, integral membrane subunit	3.777	0.00064
Accessory gene regulator protein B	3.775	0.00064
Resuscitation-promoting factor Rpf	3.76	0.00064
Split-Soret cytochrome c	3.752	0.00064
Carbon monoxide dehydrogenase 2	3.747	0.00064
Dihydromethanopterin reductase	3.726	0.00064
Protein FeSII	3.712	0.00064
PTS system N-acetylglucosamine-specific EIIB component	3.705	0.00064
Nitrogen fixation regulatory protein	3.626	0.00064
Na(+)-translocating ferredoxin:NAD(+) oxidoreductase complex subunit C	3.608	0.00064
Alpha-amylase 1	3.584	0.00064
Elongation factor G, mitochondrial	3.576	0.00064
Neopullulanase 1	3.574	0.00064
Hydrogenase-4 component G	3.525	0.00064

Ribulose bisphosphate carboxylase	3.525	0.00064
D-xylonate dehydratase YagF	3.51	0.00064
Sporulation-specific cell division protein SsgB	3.506	0.00064
EtfAB:quinone oxidoreductase	3.505	0.00064
Lipoprotein NlpI	3.494	0.00064
CRISPR-associated endonuclease Cas6	3.461	0.00064
Cytochrome c-type protein ImcH	3.421	0.00064
Opacity-associated protein OapA	3.417	0.00064
Surface-adhesin protein E	3.417	0.00064
Fused nickel transport protein NikMN	3.416	0.00064
Molybdenum storage protein subunit alpha	3.414	0.00064
Corrinoid/iron-sulfur protein small subunit	3.399	0.00064
Small, acid-soluble spore protein C2	3.399	0.00064
Cytochrome c”	3.384	0.00064
IS1182 family transposase ISRssp12	3.375	0.00064
Mannosylglucosyl-3-phosphoglycerate synthase	3.347	0.00064
Flagellar FliL protein	3.344	0.00064
Valine dehydrogenase	3.333	0.00064
Sensor protein CseC	3.319	0.00064
Tyrosine-protein kinase CpsD	3.312	0.00064
Rubredoxin-oxygen oxidoreductase	3.306	0.00064
putative nitrate/nitrite transporter NarK2	3.298	0.00064
5-hydroxybenzimidazole synthase BzaA	3.297	0.00064
Enoyl-[acyl-carrier-protein] reductase [NADPH] FabI	3.295	0.00064
Putative sulfur carrier protein YeeD	3.292	0.00064
Translocation-enhancing protein TepA	3.28	0.00064
IS66 family transposase ISSwo2	3.258	0.00064
Citrate (Re)-synthase	3.257	0.00064
putative protein YgcP	3.257	0.00064
PTS system N-acetylglucosamine-specific EIIC component	3.256	0.00064
Putative superoxide reductase	3.256	0.00064
Cyanuric acid amidohydrolase	3.247	0.00064
IS91 family transposase ISCARN110	3.24	0.00064
IS5 family transposase ISPso2	3.234	0.00064
IS1595 family transposase ISMpo2	3.232	0.00064
60 kDa chaperonin 3	3.229	0.00064
Particulate methane monooxygenase beta subunit	3.206	0.00064
Oxalate oxidoreductase subunit beta	3.205	0.00064
10 kDa chaperonin 2	3.2	0.00064
Sucrose synthase	3.199	0.00064
putative sporulation protein YlmC	3.187	0.00064
IS66 family transposase ISDpr4	3.186	0.00064
DNA transformation protein TfoX	3.185	0.00064
(R)-2-hydroxyisocaproyl-CoA dehydratase alpha subunit	3.17	0.00064
Outer membrane protein P5	3.159	0.00064
putative secretion system apparatus ATP synthase SsaN	3.15	0.00064
IS1182 family transposase ISClbu1	3.15	0.00064
2-amino-5-chloromuconate deaminase	3.147	0.00064
Type A flavoprotein fprA	3.145	0.00064
Benzylsuccinate synthase activating enzyme	3.144	0.00064
(R)-phenyllactate dehydratase activator	3.142	0.00064
Particulate methane monooxygenase alpha subunit	3.133	0.00064
(R)-phenyllactyl-CoA dehydratase alpha subunit	3.126	0.00064
Barbiturase 1	3.122	0.00064

CRISPR system Cascade subunit CasE	3.112	0.00064
NADH-dependent phenylglyoxylate dehydrogenase subunit gamma	3.106	0.00064
ECF RNA polymerase sigma factor ShbA	3.103	0.00064

Eigenständigkeitserklärung

Name:

Matrikelnummer:

Ich versichere, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die allgemeinen Prinzipien wissenschaftlicher Arbeit und Veröffentlichungen, wie sie in den Leitlinien guter wissenschaftlicher Praxis der Carl von Ossietzky Universität Oldenburg festgelegt sind, befolgt habe.

Datum:

Unterschrift: