Universal laws in Ecology

A thesis submitted for the degree of
Doctor of Philosophy

Jody Fisher (McKerral)
B.Mus. (Hons/Hons), M.Mus., B.Sc. (Hons)

College of Science & Engineering

Flinders University

30 September 2021

Contents

List of Tables

List of Figures

Summary

Multispecies communities are inherently complex, with a myriad of processes influencing their interacting parts. Yet, these communities often exhibit highly conserved emergent behaviour, which is suggestive of unifying organisational principles operating within ecological settings. Through investigating the distributions and scaling of organisms, their abundances, and metabolic diversity, this doctoral research explores potential mechanisms behind emergent behaviour at scales of both microbes and ecosystems. The association between organism physiology and ecosystem structure captured by size-abundance scaling laws is probed through an allometric setting of the Rosenzweig-Macarthur differential equations. Through extensions to the model motivated by empirical biological research and classical biophysics, it is shown that terrestrial and marine biospheres are fundamentally different, with turbulence restructuring the dynamics of oceanic ecosystems by imposing additional energetic costs on large organisms. The macro stability observed in size-abundance scaling laws is mirrored by a ubiquitous feature of functional stability within the microbial communities which sit at the base of the food web. To interrogate plausible assembly rules that may give rise to this behaviour, a network-based framework is used to link taxa and function, resolving a fundamental challenge in probing taxa-metabolism relationships in microbial ecology. Analysis across real-world microbial communities spanning major environmental and host microbiomes reveals a universal taxa-function structure, which would facilitate horizontal gene transfer and thus strengthen community stability and resilience.

Declaration

I certify that this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Jody Fisher (McKerral)

Acknowledgements

I am expressly grateful for the support, generosity and mentorship shown by my supervisory team of Jim Mitchell, Jerzy Filar and Nima Dehmamy, without whom this thesis would not have been possible. Jim, you showed me the best of science through inspiring and creative research, and created many opportunities to help me on a path toward it. Jerzy, thankyou for grounding me in attention to detail, showing me the rigor and beauty of mathematics, and for always emphasising the question *but how?* until I asked it for myself. Finally, thankyou Nima for impeccable advice and a rich perspective on linking data with theory, and your steady assistance in the final stages of my PhD.

I also must thank the many other researchers who have contributed time and resources toward supporting me during my candidature. Firstly, Albert-László Barabási, and the Barabási lab, for providing me with an intellectually rich environment and always being available for discussion, critique, and advice during my time in Boston. Rob Edwards, for your generosity in providing bioinformatics expertise and extensive computational resources, as well as Michael Haythorpe, Kieran Clancy, Maria Kleshnina, and the mathematics group at Flinders University for ongoing assistance and advice. Alex Newcombe: thankyou for your generosity in providing this thesis template!

Multiple funding bodies have supported me during my PhD, which was undertaken on an Australian Government Research Training Program Scholarship. This was further supported by the Australian-American Fulbright Commission, the Australian and New Zealand Industrial and Applied Mathematics Society's AF Pillow award, the Playford Trust, and the Australian Federation of University Women.

Finally, I give my everlasting thanks to my family and my partner Chris for their understanding and patience.

Chapter 1

Introduction

"I think the next century will be the century of complexity."

Stephen Hawking, January 2000

1.1 Background

The notion of a universal law, especially one which may be captured by a parsimonious equation, is central to physics. However, truly universal properties in biological systems seem likely to remain constrained to principles or descriptors. In 1999, a year before Hawking's prediction about the direction of science in the 21st Century, it was claimed that it was unlikely constructive or useful generalisations may be made about multispecies assemblages, and doubts were raised that the concept of a scientific law was realistic or possibly even relevant within community ecology [1]. This is, in many ways, understandable. Multispecies communities have nonlinear and chaotic dynamics. They exhibit feedback with and influence over surrounding environments, and adapt according to changes. Furthermore, the emergent properties of a community often bear little relation to the behaviours or apparent rules governing its interacting parts. These are, of course, the hallmarks of a complex system; indeed, serendipitously 1999 also heralded the emergence of network science - a central focus of which is complex systems - as a discipline in its own right with the discovery of scale-free networks [2].

The defeatist conclusions within [1] sparked debate - still ongoing - about the respectability of ecology itself as a scientific discipline and what (if anything) may pass as a 'law' within the field [3]. Yet, 1999 was also the year West, Brown and Enquist developed their model for metabolic scaling, a quarter of a century after the founding of the complex systems institute at Santa Fe where it was produced [4]. This model is now generally accepted as the mechanistic principle behind allometric laws, extending beyond physiology to explain a range of allometric scaling behaviours and linking organisms to ecosystems [5]. Detractors to the framework have ample counter-examples, but - as is now a common theme within complex systems research - it is now widely accepted that ecology is a discipline of scales [6]. The question of whether modelling aims should prioritise local predictive power or be generalisable has contributed to the vociferous arguments on either side of the ecological laws debate [7]. Whilst a 'law' implies generalisation, it is unlikely such a model in ecology would simultaneously have strong predictive capability for individual organisms within noisy ecological data. Indeed, amongst this noisy data and astronomical numbers of variables, it can be difficult to identify ubiquitous phenomena even before attempting to find an explanation for the observed pattern. However, there can be multiple levels of explanation, and it seems likely that top down and bottom up approaches may be necessary to deepen our understanding of ecological systems.

1.2 Thesis aims and structure

In this thesis, I aim to find unifying organisational principles in multispecies communities. Through examining the distributions and scaling of organisms, their abundances, and metabolic diversity, this dissertation investigates potential mechanisms behind emergent behaviour at scales of both microbes and ecosystems.

Chapters 2 and 3 build on the metabolism-allometry relationship described in [4] by linking metabolic theory with classical population dynamics. With the goal of reproducing ecosystem-wide size-abundance scaling laws through a minimal model, allometric settings of population dynamics ODEs are examined from the perspectives of the mathematical and empirical biological literature. I uncover disparities between the disciplines, examine how they would impact dynamical behaviour when applied to large scale domains, before assessing how well my paramaterisation reproduces the behaviour of data gathered from the terrestrial biosphere. Next, size-abundance scaling distributions in the ocean are investigated. As far as I am aware, an analysis of the global marine size-abundance distribution has not been undertaken for over two decades, and never at this scale: 15,000 data points are collated across over 2000 species[1] spanning viruses to blue whales. All organisms greater than 10cm exhibit reduced abundance compared to ecological prediction, with a structural break in the distribution corresponding to the exponent of the scaling law changing from -0.75 to -1.9. To explain this, the hypothesis that turbulence fundamentally restructures marine ecological scaling dynamics by imposing larger energetic costs on organisms living within the oceanic environment is assessed. My minimal model is extended to incorporate the metabolic demands of living in turbulence, derived from classical biophysics models of swimming organisms, to consider the impact of ocean physics in reshaping the scaling laws that define marine ecosystem structure and function.

In Chapters 4, 5, and 6, I transition from ecosystem level models to considering the empirical distributions of functional diversity across prokaryotes. Microbes form complex, diverse communities of thousands of species, yet a ubiquitous characteristic of these communities is functional stability. This stability may be a crucial foundation of the scaling laws discussed in Chapters 2 and 3, as microbes sit at the base of the food web and are the principal drivers of global biogeochemical cycles. Whilst there is consensus that the emergent property of community stability is likely due to metabolic assembly rules, a fundamental obstacle to quantifying these rules has been how to link taxa and function in microbial ecology. This long standing issue is resolved by utilising bipartite networks to quantitatively explore the taxa-function relationship in microbial systems, allowing me to examine the distribution of metabolic pathways across the prokaryotic tree of life, revealing critical structures in the metabolic organisation of the biosphere. I then examine the networks of 248 real-world microbial communities across multiple environments at a planetary scale, exploring the hypothesis that there is a universal functional redundancy structure within microbiomes that would facilitate horizontal gene transfer and thus promote stability and resilience.

[1] Microbes excluded from diversity counts

Chapter 2

Universal parameterisation of a predator-prey system

Allometric settings of population dynamics models are appealing due to their parsimonious nature and broad utility when studying system level effects. Here, I parameterise the size-scaled Rosenzweig-MacArthur ODEs to eliminate prey-mass dependency. I define the functional response term to match experiments, and examine where metabolic theory derivations and observation diverge. Dynamics are produced which are consistent with observation. My parameterisation of the Rosenzweig-MacArthur system is an accurate minimal model across 15+ orders of mass magnitude.

Refer to Appendix A for a link to the code used in this chapter.

2.1 Background

Allometric scaling relationships have been the subject of scrutiny and debate since the connection between organism size and its metabolic rate was first defined by Rubner in 1883 [4, 8–11]. These models, which link some characteristic y to the size x of an organism via the power law $y = ax^b$ (where a, b are scalar constants), are appealing due to their capacity to capture a multitude of relationships despite their simplicity. Scaling laws have been used to express a variety of biological rate measures, such as metabolism, consumption, and birth or death rates [12–15]. Allometry is also utilised in modelling behavioural traits and bioenergetic characteristics, such as movement behaviour or locomotory costs [14, 16–19]. At broad scales, such laws have been applied to ecosystem-level properties, including predictions of organism population density and carrying capacity [20–22]. However, despite scaling laws' wide utility and intensive study, there has been a limited exploration of the properties of minimally constructed, size-generalised predator-prey models [23–25].

An extensive literature examines the empirical relationship between organism and population sizes [20, 21, 26–29]. Reported exponents fall between -1 and $-1/4$, depending on factors such as taxonomy or environment. The classical $-3/4$ value describing the global size-density relationship is the direct inverse of Kleiber's $3/4$ law for metabolic scaling [9, 21]. This has led the 'energetic equivalence' hypothesis: that is, the net energy contained within each size class is invariant [26, 28]. This conjecture has been widely debated, particularly with respect to whether this invariance is cause or effect of other bioenergetic drivers [30, 31]. However, despite disagreement regarding the specific mechanisms, there is general consensus that the observed consistency of size-density scaling within empirical data is likely reflective of fundamental physical constraints [12, 31]. To examine what may be driving the limitations in macro-scaling behaviour, it is possible to use dynamical size-based models which incorporate organism traits that scale across the size range [13, 23, 24]. This approach allows for the investigation of critical breaks in ecosystem-level scaling laws within a global framework. It also provides scope to explore potential impacts from changes that may affect a large proportion of organisms in a similar way - for example, warming temperatures or emergent hypoxia in the oceans [32–34]. However, perturbing parameters across 15+ orders of magnitude in size poses challenges. For example, the coexistence regions of size-generalised predator-prey models are strongly dominated by scaling exponents [24].

Previous work has attempted to resolve this in several ways. It is more straightforward to keep the global model behaviour stable by using the 4-parameter Lotka-Volterra system, but that setting is unsatisfactorily simple for most applications [13, 25]. Other approaches use series of models solved piece-wise for different sizes, yet this means that they are not truly generalised. In the most comprehensive study to date, a size-based parameterisation of the Rosenzweig-MacArthur system places restrictions on the relationships between the exponents of each parameter [24]. However, this in turn limits the types of perturbations that may be applied or investigated. Finally, there are discrepancies in the treatment of the functional response term between the mathematical and biological literature. The mathematical literature broadly assumes that the maximal consumption limit ties predator production to prey production and thus scales negatively, however, there is evidence in the empirical biological literature for positive scaling.

Here, I present an alternate approach of parameterising size-based predator-prey interactions for the classical Rosenzweig-MacArthur system of equations. Under this framework, I examine the parameter sensitivity and required ranges for species coexistence within the context of real-world observations. I am able to show that, despite the number of assumptions inherent within this style of modelling, the mathematical restrictions are closely related to biological observations. Finally, I describe the conditions required to create an entirely size-invariant model, and how using the most commonly observed empirical values for the parameters generates a size-abundance distribution matching real-world observations.

2.2 Parameterisation of the model

I begin with the Rosenzweig-MacArthur ODEs and Holling II functional response,

$$\begin{aligned}\frac{dR}{dt} &= rR(1-\frac{R}{K}) - \frac{bR}{1+hbR}C \\ \frac{dC}{dt} &= \epsilon\frac{bR}{1+hbR}C - \delta C.\end{aligned} \tag{2.1}$$

I designate variables R for resources and C for consumers. The parameters r and δ are birth and death rates respectively. Carrying capacity is given by K, interaction rate b, handling time h and the conversion efficiency ϵ. To investigate the system across the full size range I scale the parameters by mass. Organism size (in g) is given by S_R for resources and S_C for consumers. I depart from [24] by parameterising the functional response term after empirical settings [13, 35, 36]. The global parameters, which are all strictly positive, are expressed as

$$\begin{aligned} r &= r_0 S_R^{\sigma_r} \\ K &= K_0 S_R^{\sigma_K} \\ b &= b_0 S_C^{\sigma_b} \\ h &= h_0 S_R^{\sigma_{hR}} S_C^{\sigma_{hC}} \\ \delta &= d_0 S_C^{\sigma_\delta}, \end{aligned} \tag{2.2}$$

where for each parameter i, the coefficients i_0 may be standardised to a boundary value, and scaling exponents are given by σ_i. Next, the prey-predator mass ratio is denoted as ρ, where $\rho > 0$, allowing me to explore the effects of nonfixed size differences between predator and prey.

$$\begin{aligned} \hat{r} &= r_0(\rho S_C)^{\sigma_r} = r_0 \rho^{\sigma_r} S_C^{\sigma_r} \\ \hat{h} &= h_0(\rho S_C)^{\sigma_{hR}} S_C^{\sigma_{hC}} = h_0 \rho^{\sigma_{hR}} S_C^{\sigma_{hR}+\sigma_{hC}} = h_0 \rho^{\sigma_{hR}} S_C^{\sigma_h} \\ \hat{K} &= K_0(\rho S_C)^{\sigma_K} = K_0 \rho^{\sigma_K} S_C^{\sigma_K}. \end{aligned} \tag{2.3}$$

With this approach I extend the results of [24] by placing no restrictions on the exponents, allowing h to be an independent term which may be matched to empirical observations. I now also follow standard practice by setting $\epsilon \propto \rho$, that is, the conversion efficiency is proportional to the prey-predator mass ratio [24]. Next, standard rescaling of the Rosenzweig-MacArthur system is used to reduce the number of parameters and simplify the following analyses. I define $u = R/\tilde{R}$ and $v = C/\tilde{C}$. Next, I set $\tilde{R} = 1/(b\hat{h})$, $\tilde{C} = \epsilon/(b\hat{h})$, $\mu = \hat{K}b\hat{h}$. After scaling time by $\hat{r}$ such that $t/\hat{r} = s$, and defining $\gamma = \epsilon/(\hat{h}\hat{r})$ and $\omega = \delta/\hat{r}$, I arrive to the new system

$$\begin{aligned} \frac{du}{ds} &= u(1 - \frac{u}{\mu}) - \frac{\gamma uv}{1+u} \\ \frac{dv}{ds} &= \frac{\gamma uv}{1+u} - \omega v, \end{aligned} \tag{2.4}$$

which has the same dynamical behaviour as (2.1). The parameters in (2.4) are also all strictly positive and scaling across the size range as for (2.2)-(2.3). For clarity, I provide the explicit relationship between the old and new parameters below in Table

2.1, and the system of equations with substituted terms is

$$\begin{aligned}\frac{du}{ds} &= u(1-\frac{S_C^{-\sigma_h-\sigma_b-\sigma_K}}{K_0h_0b_0\rho^{\sigma_K+\sigma_{hR}}}u)-\frac{\epsilon S_C^{-\sigma_h-\sigma_r}}{h_0r_0\rho^{\sigma_r+\sigma_{hR}}}\frac{uv}{1+u}\\ \frac{dv}{ds} &= \frac{\epsilon S_C^{-\sigma_h-\sigma_r}}{h_0r_0\rho^{\sigma_r+\sigma_{hR}}}\frac{uv}{1+u}-\frac{\delta_0 S_C^{\sigma_\delta-\sigma_r}}{r_0\rho^{\sigma_r}}v.\end{aligned} \tag{2.5}$$

The expression of size-scaling of S_R in terms of S_C facilitates interpretability in downstream analyses. All of the exponent terms may be collected within S_C, which is now referred to as S. This provides greatly simplified expressions within Table 2.1 and (2.5), yet preserves the ability to examine the impacts of perturbations to any one parameter. Table 2.2 provides a summary of these values taken from prior research.

Table 2.1: Relationship between parameters in original and rescaled Rosenzweig-MacArthur system. Here, $\sigma_h = \sigma_{hR} + \sigma_{hC}$.

	Definition	Coefficient	Exponent
μ	$\hat{K}\hat{h}b$	$\mu_0 = K_0h_0b_0\rho^{\sigma_K+\sigma_{hR}}$	$\sigma_\mu = \sigma_K + \sigma_h + \sigma_b$
γ	$\epsilon/\hat{h}\hat{r}$	$\gamma_0 = \epsilon/(r_0h_0\rho^{\sigma_r+\sigma_{hR}})$	$\sigma_\gamma = -\sigma_h - \sigma_r$
ω	$\delta/\hat{r}$	$\omega_0 = \delta_0/(r_0\rho^{\sigma_r})$	$\sigma_\omega = \sigma_\delta - \sigma_r$

Table 2.2: Literature bounds on parameter values. The top portion of the table outlines scalars[a]. The middle section provides the scaling exponents, and the bottom the possible exponent range for the rescaled system. I provide the limits to these values for completeness. However, the general consensus is that σ_r, $\sigma_d \simeq -1/4$, which has been verified in a substantial recent review [37]. Similarly, despite the potential range for σ_b, typically $1/2 \leq \sigma_b \leq 1$, which significantly constrains the exponent ranges in the rescaled system.

Symbol	Parameter	Minimum	Maximum	References
ρ	Predator-prey mass ratio	1E-4	1E2	
S	(Consumer) mass, g	1E-10	1E7	
ϵ	Conversion efficiency	0	ρ	
σ_r	Birth rate	-0.81	-0.25	[13, 23, 24, 32, 37–42]
σ_δ	Death rate	-0.35	-0.22	[13, 23, 24, 32, 37–42]
σ_b	Interaction rate	-0.25	1.58	[13, 25, 35, 36, 41]
σ_K	Carrying capacity	-0.88	-0.74	[13, 24, 42]
σ_{hR}	Handling time (resource)	0	1	[23, 35, 36]
σ_{hC}	Handling time (consumer)	-1.1	0	[13, 23, 35, 36]
σ_μ		-2.2	1.84	
σ_γ		-0.75	1.92	
σ_ω		-0.1	0.59	

[a] Note that viruses are not included in the model due to the non-classical role they may play in interactions with their 'prey', e.g. phage may be beneficial to host bacteria, and eukaryotic viruses are tens of orders of magnitude smaller than their hosts, leading to substantially different dynamics.

2.3 Results

2.3.1 Coexistence & Sensitivity

The non-trivial equilibrium of interest (coexistence) is

$$\begin{aligned} u^* &= \frac{\omega}{\gamma - \omega} \\ v^* &= \frac{(\mu\gamma - \mu\omega - \omega)}{\mu(\gamma - \omega)^2}. \end{aligned} \tag{2.6}$$

For there to be non-negative values for (u^*, v^*), I require that

$$\gamma > \omega \text{ or } \frac{\gamma}{\omega} > 1, \text{ and} \tag{2.7a}$$

$$\mu > \frac{\omega}{\gamma - \omega}. \tag{2.7b}$$

The condition det > 0, where det is the determinant, is also fulfilled by (2.7b). I may express (2.7b) as $\gamma/\omega > 1 + 1/\mu$. As all parameters are strictly positive, if (2.7b) is satisfied, it immediately follows that (2.7a) is satisfied also. The inequality

$$\frac{\gamma}{\omega} < \frac{\mu + 1}{\mu - 1} \tag{2.8}$$

determines the sign of the trace of the Jacobian, which dictates whether the system converges to a point or to a stable limit cycle. At equality, there is a Hopf bifurcation. The dynamical characteristics of the Rosenzweig-MacArthur system bave been explored in depth elsewhere (e.g. [43–46] and references within). My focus is on the interplay between biological and mathematical constraints. These inequalities are now discussed in the context of empirical observations.

Handling time

The condition from (2.7a) simplifies to

$$\frac{\epsilon}{h_0 \delta_0 S^{\sigma_h + \sigma_\delta}} > 1. \tag{2.9}$$

Examining the scaling exponents, if $\sigma_\delta + \sigma_h < 0$, the condition may fail for small organisms. As σ_δ is tightly constrained (Table 2.2), I examine the behaviour of the system under different scaling values of the handling time.

The classical null model from Yodzis & Innes [23] based on metabolic theory is determined in [35][1] to be equivalent to $h \propto S_R^1 S_C^{-3/4}$, or $\hat{h} \propto S^{1/4}$. It results in a maximal consumption rate of $c \propto S^{-1/4}$. This matches the setting of [24], where the birth rate of the predator in the presence of unlimited resources is assumed to scale with the birth rate of prey. However, more recent functional response research suggests a more nuanced picture, where physiological traits may affect these exponents. Attacking, killing, and then physically eating and digesting prey all impact handling time [36] and there is consensus that it is vital to consider the prey's contribution to this process [35, 36, 47]. The term has since been recast to the more biologically representative form used in this study: $h \propto S_R^{\sigma_{hR}} S_C^{\sigma_{hC}}$, where typically $0 \leq \sigma_{hR} \leq 1$ and $-1 \leq \sigma_{hC} \leq 0$ [35, 36]. The exponents σ_{hR}, σ_{hC} have been empirically determined in several reviews and display considerable variability [13, 35, 36]. I now discuss the implications this variability has for coexistence under the inequalities in (2.7).

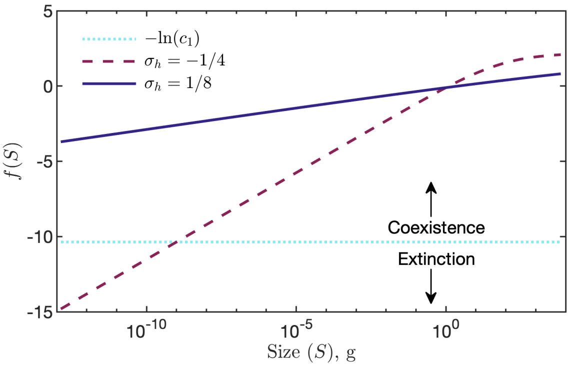


Figure 2.3: Graphical representation of coexistence conditions in (2.7). Rearranging (2.7b): $\ln(S^{-\sigma_h-\sigma_\delta}) - \ln(1 + c_2 S^{-\sigma_K-\sigma_h-\sigma_b}) > -\ln(c_1)$, where c_1 and c_2 are constants derived from the coefficients. I denote the left side of the inequality as $f(S)$; if $f(S) > -ln(c_1)$, there is coexistence. Under the feasible values outlined in Table 2.2, σ_K and σ_b have less effect than σ_h; here I assign $\sigma_K = -3/4$ and $\sigma_b = 1/2$.

Two of the three major reviews listed above conclude that the prey-predator components of handling time scale more gently (whether positive, or negative) than null models predict. However, the resultant exponent for $\hat{h}$ is positive ($\simeq 1/3$) in [35], which reviews arthropod functional responses, and negative ($\simeq -1/8$) in the broader taxonomic review within [36]. Most of the organisms in [36] display negative scaling for $\hat{h}$ in taxa-specific breakdowns due to gentler scaling of the resource exponent. Only σ_{hC}

[1] This derivation assumes predator consumption (the inverse of handling time) scales with metabolic demand ($S_C^{3/4}$), and the per-prey metabolic demand is therefore $S_C^{3/4} S_R^{-1}$. This matches the assumptions of [24, 25], however it should be noted that other interpretations of the same model either do not normalise against prey mass (e.g. [42]) or do so implicitly (e.g. [13]).

is assessed in [13]. This study estimates σ_{hC} for 2D and 3D environments but does not find a significant difference between them with the resultant empirical $\sigma_{hC} \simeq -1.1$. The authors account for the steeper scaling relative to metabolic expectation by noting that feeding is an active process which scales with maximal rather than basal metabolic cost. This review contains a larger mass range than in [36]. To calculate the scaling of $\hat{h}$ from the empirical assessment in [13], I use the assumption $\sigma_{hR} = 1$, which is the upper limit for the parameter. This implies the exponent $\sigma_h \leq -0.1$, and again suggests $\hat{h}$ scales far below the value of $1/4$ assumed by previous theoretical work on the model. Conceptually, this indicates that the parameter may be constrained by physical processes rather than a bioenergetic flux balance. Note that despite the phenomenological formulation of the functional response predator production is still implicitly constrained by the prey density.

Taking this into consideration for coexistence conditions (2.7), if $\sigma_h < -1/5$ across the full size range, smaller organisms may violate this condition (Figure 2.3). If $\sigma_h \simeq -\sigma_\delta$ then this condition will easily be fulfilled across the size range. There appears to be empirical support for these observations. The taxonomic group breakdowns in [36] indicate that smaller taxa may display positive scaling for $\hat{h}$ as concluded in [35] and the strongly negative scaling is observed in macro-organisms, particularly vertebrates. The notable exception of unicellular marine organisms ($\sigma_h \simeq -1/3$) has a very small sample size. It is possible further experimental investigations may reach an alternate conclusion.

Carrying capacity and scaling of population cycling

An alternate expression of (2.7b) is

$$\mu_0 S^{\sigma_\mu} = \mu_0 S^{\sigma_K + \sigma_h + \sigma_b} > u^*, \tag{2.10}$$

where $u^* \propto S^{\sigma_\delta + \sigma_h}$. If I combine (2.7b) and (2.8), I obtain

$$\mu_0 S^{\sigma_\mu} = \mu_0 S^{\sigma_K + \sigma_h + \sigma_b} > 1. \tag{2.11}$$

Together, (2.10) and (2.11) indicate the need for a sufficiently high carrying capacity for a sustainable prey population, and that consumer attack rates must be high enough to compensate for mortality across the size range. Assuming reasonable values for σ_K, σ_b such as those given in Table 2.2, these inequalities will generally hold. Under the original system (2.1) and assuming coexistence, resource equilibria will scale with size as $R^* \propto S^{\sigma_\delta - \sigma_b}$ and consumer equilibria will scale as $C^* \propto S^{\sigma_r - \sigma_b}$. Despite their importance for the coexistence domain, the carrying capacity and half saturation do not meaningfully impact equilibria abundances in allometric parameterisations. However, they do impact some properties of the limit cycle.

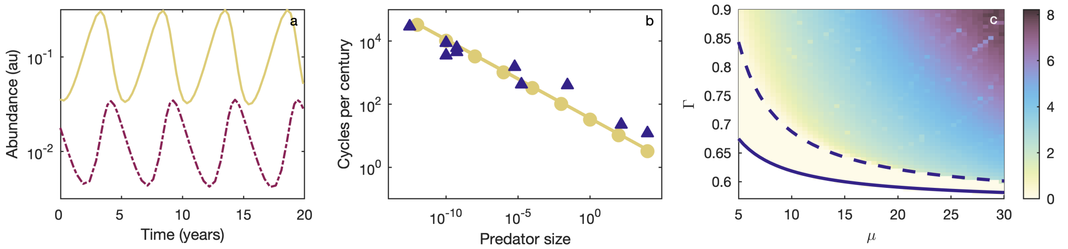


Figure 2.4: Properties of the limit cycle: (a) Predator-prey oscillations for a predator of size 10g (b) Scaling of the period of the limit cycle for numerical (circles), empirical (triangles) and analytic (solid line) results. Only predators are shown. Data is from predator-prey population time series from [25, 48–55]; periods were calculated by either using supplied data files or software to extract data from figures [56], and then averaging the time between peaks. (a)-(b) use empirical scaling of $\hat{h}$, where $\sigma_h = -1/8$. (c) Dynamics of the rescaled system associated with different parameter values. Here, $\Gamma = \gamma/\omega$, which is plotted against μ. The region below the solid line indicates no coexistence, between the solid and dashed lines denotes a sink to the equilibrium point, and above the dashed line a stable limit cycle. Colour indicates the difference between the (log) maximum and minimum abundances attained for the predator.

Allometric settings of the Rosenzweig-MacArthur system usually result in oscillating solutions due to size-scaled parameter values relative to the constraints in (2.8) [40]. I find stronger empirical support than in previous work for the period τ to have a discernible size scaling signal [40]. Indeed, the theoretical scaling $\tau \propto S^{\sigma_\delta}$ (derived in both [24, 40], with a similar result in [23]) agrees well with observed values (Figure 2.4b). A previous review has found that the ratio of maximum to minimum densities is size-invariant [40]. Practically speaking, this means that the oscillation amplitude decreases with increasing size. Further qualitative support that this mathematical behaviour is aligned with features of real-world biological systems is given by the fact that log-transformed size-density relationships demonstrate near-constant variance across 15+ orders of magnitude [37, 57]. Under my parameterisation the log-scaled oscillations are relatively sinusoidal and may be constrained to 1-3 orders of magnitude (Figure 2.4a). Hence, they are more realistic than allometric Lotka-Volterra dynamics which push predator populations to unreasonably low levels with fluctuations exceeding 15 orders of magnitude [25]. Unfortunately early efforts to find analytic approximations of the oscillation amplitude of the Rosenzweig-MacArthur system have not been generalised [58]. More recent results have only been derived for specific - and highly restricted - parameter values [59]. However, the rescaled system (2.4) does provide scope for me to examine the effects of perturbations in a simplified manner. In Figure 2.4c, I show through simulation that perturbations to all parameters may impact the magnitude of the fluctuation of the limit cycle. However, unless these perturbations are applied to σ_r or σ_δ, the oscillation amplitude will remain (nearly) invariant with respect to the mean population density. The system's robustness to different forms of perturbation may be assessed via a sensitivity analysis which is outlined below.

Sensitivity

A local sensitivity analysis allows me to obtain a first-order approximation of the relative impact of changing parameters on the solutions of (2.4). I adhered to the methodology described in [60]. In order to check how sensitive the system, $\dot{x}$, is to small changes in parameters, λ_i, I construct a sensitivity function, $S(t)$, such that

$$S(t) = \frac{\partial}{\partial \lambda} x(t, \lambda) \tag{2.12}$$

and $x(t, \lambda)$ is a solution of $\dot{x}$. Next, I characterise the solution to the sensitivity equation given by

$$\dot{S}(t) = A(t, \lambda_0) S(t) + B(t, \lambda_0),\ S(t_0) = 0. \tag{2.13}$$

Applying (2.13) to (2.4), A is the Jacobian of (2.4) with respect to variables u and v, and B is the Jacobian of (2.4) with respect to parameters μ, γ, and ω, both of which are evaluated at nominal parameter values. After setting initial conditions u_0 and v_0, I obtain numerical solutions for (2.13). Figure 2.5 shows the trajectories for two initial conditions. Firstly, for u_0 and v_0 in the neighborhood of u^*, v^* respectively, and secondly for u_0, v_0 an order of magnitude greater/smaller than u^*, v^* respectively. The qualitative behaviour remains the same in both cases. For an initial state near the equilibrium (Figure 2.5(a-b)), there is monotonic behaviour as the system converges to the limit cycle. In the case of Figure 2.5(c-d), the limit cycle will emerge after some critical time t_c. For each, the system is least sensitive to μ, providing further support that perturbations to r and δ have the largest impacts on the system. The qualitative behaviour remains similar for other nominal values of the parameters, provided they are not set on the other side of the bifurcation boundary. The exponents with the least empirical variance - by a significant margin - are σ_r and σ_δ, which aligns well with the mathematical constraints discussed in this section. That is, the dynamical behaviour is relatively robust to perturbing the functional response parameters which display the highest empirical variance.

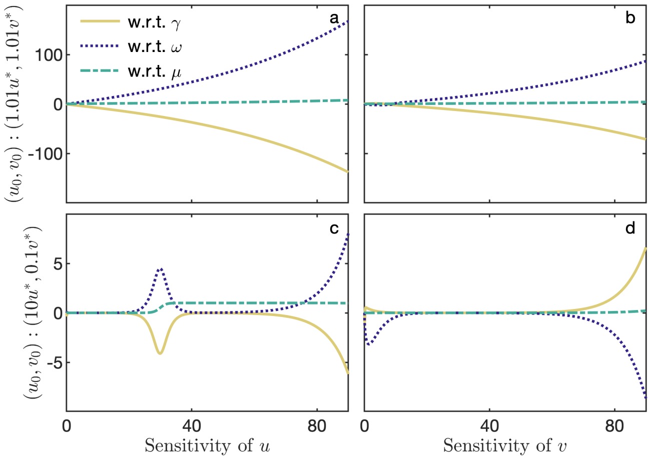


Figure 2.5: Sensitivity of rescaled system. x-axis denotes time (au). (a-b): Sensitivity of the solutions of (5) to perturbations to each of the parameters under an initial condition near the point (u^*, v^*). (c-d): As above, except under an initial condition $(10u^*, 0.1v^*)$; the trajectory also converges to the limit cycle.

2.3.2 Applications

By scrutinising the rescaled parameter definitions in Table 2.1, it is possible to determine an entirely size-invariant system. This is clearly desirable as it is straightforward to set coexistence for an arbitrary size range, and facilitates analytic study of the equations. By the definition of $\omega = \delta/\hat{r}$, the size scaling of r must match δ, both of which consistently display an exponent of $-1/4$. It immediately follows that $\sigma_h = 1/4$ as $\gamma = \epsilon/\hat{h}\hat{r}$. Assuming that carrying capacity scales to a $-3/4$ power law, and using the definition $\mu = \hat{K}\hat{h}b$ it therefore follows that $\sigma_b = 1/2$. As the dynamics of (2.4) are identical to (2.1), eliminating the size-dependency in (2.4) is certainly mathematically expedient and I note that the majority of allometric studies to date have used this approach. However, given the tenuous empirical support for $\hat{h} \propto S^{1/4}$, it may be more biologically suitable to assign a value where $\sigma_h \leq 0$. It is still feasible to generate a full size-abundance distribution and match empirical scaling of the parameters. An extensive recent analysis of size-density scaling suggests that the relationship may scale closer to $N^* \propto S^{-1}$ than the classical $-3/4$ exponent (where N^* is population density or abundance). Indeed, using empirical scaling of b results in a distribution which aligns with that result (Figure 2.7). Whilst the lower bound of $\sigma_b \simeq 1/2$, estimates of the 'universal' value suggest $0.6 < \sigma_b < 0.9$. A limitation of my treatment of b is no restrictions are placed on the interactions between a predator and any arbitrary-sized prey. Interaction processes are a pivotal component of predator-prey systems and increasing evidence suggests they follow a 'hump-shaped' curve with the prey-predator mass ratio, ρ [35]. A natural extension to my model would be to introduce a term reliant on ρ to the parameter b. This would imply $b = f(\rho, b_0)S^{\sigma_b}$, where $f(\rho, b_0)$ may be any function which assigns a probability of prey capture based on the prey-predator size ratio. Whilst a theoretical form for $f(\rho)$ has been proposed [24], it would be feasible to use a function that could encode of a broader range of life history traits for the tradeoff of introducing additional parameters. Processes to consider may include habitat effects on foraging, prey refuges, and optimal size ratios. This would impose further constraints on the capacity for predator coexistence.

My final consideration is the contribution of ρ to the conversion efficiency ϵ. The empirical distribution of ρ is approximately lognormal with its peak at $\simeq 0.02$ [41]. Equilibria population ratios do not follow a 1:1 relationship with the prey-predator mass ratio when using the relationship $\epsilon = \rho$ (solid white line, Figure 2.6). For a fixed predator size and increasing prey size, organisms become less efficient at converting biomass. However, this result does not align with observed data.

A review of 15,000+ predator-prey pairs concludes that size differences between predator and prey has an upper limit, potentially due to inefficiencies when the size discrepancy becomes too extreme [61]. Furthermore, the larger the predator, the more generalist its feeding strategies [35]; increases in prey biomass - which could indicate predators feeding on smaller prey - do not translate to a proportionate increase in

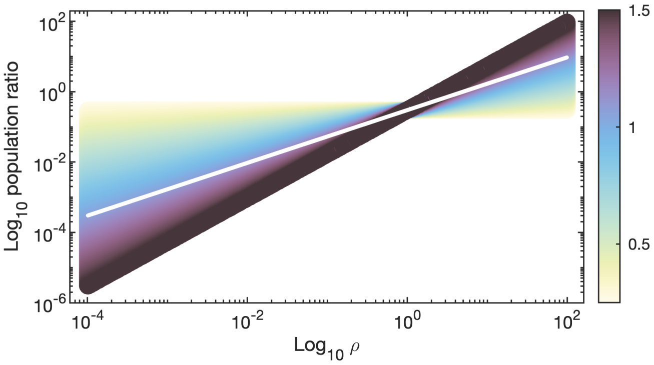


Figure 2.6: Impact of perturbing conversion efficiency ϵ by setting a function on ρ. Here, $\epsilon = \rho^{\psi}$, where $-1/2 < \psi < 3/2$. Colours map to the value of ψ, and the white line depicts $\psi = 1$. Y-axis shows the ratio of the predator-prey equilibria populations. Result is size invariant.

predator biomass [62]. I therefore apply the assumption that energetic reward (and biomass conversion) for predator effort declines as the size difference increases, and that the scaling of ρ with equilibria population ratios is superlinear. I can implicitly capture the result by assigning a function $\epsilon = a\rho^{\psi}$, where ψ is a scalar. For simplicity, I set a to 1, and in Figure 2.6, assess the predator-prey population ratios for varying ψ. A value of $\psi > 1$ increases the difference between the predator-prey populations; $\psi < 1$ reduces it. More sophisticated functional forms may include favourable size ratios or introduce a size dependency to the value of ϵ. However, there is limited empirical research on scaling properties of ϵ [21, 27, 39]. A theoretical investigation of optimal predator-prey size ratios together with more complex functional response formulations reflecting alternate foraging/feeding strategies may yield interesting results, and I leave this question open for future work.

To generate the full size-abundance distribution shown in Figure 2.7, I use a value of $\psi \simeq 1.3$, together with scaling values of $\sigma_r = \sigma_\delta = -1/4$, $\sigma_h = -0.1$, $\sigma_K = -3/4$ and $\sigma_b = 2/3$. The model's full distribution scales to -0.92, close to the -0.95 value determined in [37]. The inset shows the predator-prey density relationship, which is 0.76, close to the findings of $\simeq 3/4$ in [62]. Note that to generate Damuth's law, a value of $\psi = 1.3$ and $\sigma_b = 1/2$ results in exponents of $-3/4$ and $3/4$ for the size-abundance and predator-prey density scaling respectively.

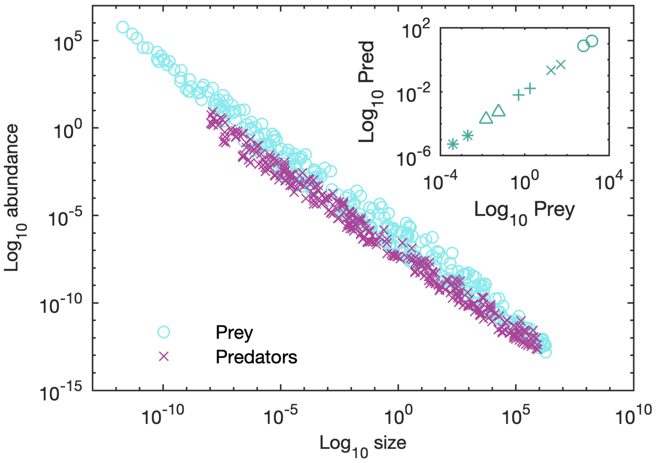


Figure 2.7: Main: Size-abundance data generated from the model. Circles depict prey abundances, and crosses predators. Prey-predator mass ratios were randomly selected within the interval of 1E-4 and 1E2. The full size-abundance distribution from the model scales to -0.91, close to the empirical distribution in [37] of -0.95. Inset: each pair of points are the maximum and minimum abundances attained by the predator/prey during limit cycle oscillations. I show examples of five predator sizes: 1E-6 ($\circ$), 1E-4($\times$), 1E-2 ($+$), 1E0 ($\triangle$), and 1E2g ($*$). For each $\rho = 0.02$. The slope within each symbol group $\simeq 0.76$. That is, increasing prey density does not result in a 1:1 increase in predator density. The scaling relationship is sublinear, matching the observations of [62].

2.4 Conclusions

Here, I investigate the links between empirical and theoretical allometric literature. By explicitly encoding the prey-predator mass ratio, ρ, I remove the resource size dependency from the system. This simplifies analyses and provides a more parsimonious base for customising the equations, which may be useful for food web or trophic modelling. I find that the constraints of the model complement empirical observation. Contrary to most previous studies, I use an empirically determined parameterisation of the functional response term. My results suggest that the standard approach of setting all parameters based on metabolic theory may need to be reassessed. The handling time parameter shows the strongest departure from those assumptions, and the highest variance, which is consistent with the massive trait variation found in hunting and feeding strategies and may suggest that organisms are adaptable to changing conditions. Nevertheless, I find that results generated from an empirical setting agree well with results in recent reviews of size-abundance scaling. This work may be extended in several ways. Firstly, it would be feasible to incorporate temperature effects, e.g. after [32, 63, 64], which may have the effect of further stabilising the model by reducing the interaction strengths [39]. Secondly, it would be valuable to have more empirical data pertaining to the functional response at the extreme ends of the size range to more accurately define the scaling of $\hat{h}$ and b. It may also be useful to investigate Type I, Type III, or generalised functional responses, although I note that previous work has found that the system behaviour usually remains qualitatively similar when considering a large size ranges [13, 42], potentially indicating the importance of average or maximal feeding rates rather than the specific shape of the functional response curve.

The broad caveat of allometric modelling is that a generalist strategy can be a poor predictor of taxon-specific outcomes. Challenges to the framework arise not only from biological differences but physical or spatial processes, such as prey patchiness or heterogeneous habitat distribution [65, 66]. Thus, care over interpretation and the applicability of results must be taken, particularly at the size limits in either direction. For example, prokaryotic reproduction rates fall between minutes and millenia [67, 68]. Furthermore, large organisms such as whales play a critical role in nutrient recycling; assuming a single species may be defined as a resource or consumer alone does not account for the intrinsic complexities within natural environments [69]. However, it can also be said that when investigating macro properties of a system, allometric approaches have been found to outperform those which attempt to explicitly encode organisms' individual and life-history traits [35]. Classical population dynamics models remain a powerful tool in ecology, and the consistency across many allometric laws suggest self-organising processes we are yet to unravel. I propose that systematically assessing where theoretical and empirical properties of allometric modelling diverge may assist in identifying plausible mechanisms governing these phenomena.

Chapter 3

Synergy of turbulence and fishing reduce aquatic biomass

A universal scaling relationship exists between organism abundance and body size. Within ocean habitats this relationship deviates from that generally observed in terrestrial systems, where marine macro-fauna display steeper size-abundance scaling than expected. This is indicative of a fundamental shift in food-web organization, yet a conclusive mechanism for this pattern has remained elusive. I demonstrate that while fishing has partially contributed to the reduced abundance of larger organisms, a larger effect comes from ocean turbulence: the energetic cost of movement within a turbulent environment induces additional biomass losses among the nekton. These results identify turbulence as a novel mechanism governing the marine size-abundance distribution, highlighting the complex interplay of biophysical forces that must be considered alongside anthropogenic impacts in processes governing marine ecosystems.

3.1 Motivation

As introduced in Chapter 2, a fundamental scaling relationship exists between organism abundance and body size, where

$$N \propto S^{\alpha} \tag{3.1}$$

and the exponent α typically approximates $-3/4$. This universal rule derives from resource acquisition as a function of body size [12], which is a barometer for ecosystem health that simplifies interactions in complex food webs and may direct fisheries management [70]. However, within marine ecosystems, the exponent for this relationship often differs from that in terrestrial ecosystems [57]. Life-history, trophic strategies, and altered productivity proposed to alter the scaling slopes of terrestrial size-abundance distributions, and these as well as fisheries posited to impact the slope of the univariate size spectra studied in marine systems [11, 57, 71, 72]. Here, I quantify, empirically and with an independent model, how fishing and ocean turbulence cause qualitatively distinct breaks in the global marine size-abundance distribution.

3.2 Data Analysis

3.2.1 Raw data sources and pooling

To assess the size-abundance scaling relationship, I examined data for over 2179 species ranging from viruses to blue whales. These encompassed over 800 genera.[1] For quality purposes, I undertook analysis with two datasets. The first was manually curated from over 200 articles to ensure there was not systematic bias within database sources, and consists of 1719 size-abundance pairs across 700+ species (Appendix A). The second dataset expands on the first via the inclusion of a further 13,455 entries predominantly sourced from online databases, for a total of 15,174 data points (Appendix A). Five databases were used: IMOS (flow cytometry and zooplankton) [73], Tara Oceans (flow cytometry) [74], Phytobase [75] for phytoplankton, a global diatom database [76], and a reef fish dataset [77]. Size data was taken from the same source as the abundance data, or if it was not included, I assigned the average adult size for that taxon referenced from WoRMS [78], fishbase [79], or [76] for diatoms. All entries which dated pre-2000 were removed to reduce the chance of methodological or quality control problems being introduced from older data. For Phytobase entries, any data with the flags ‘unrealistic day or year’ and ‘presumably sedimentary’ were deleted; I note this particular database is otherwise well suited to this application as capturing local diversity patterns is not critical for global size density analyses [57]. For the flow cytometry data, any entries which had not undergone or passed quality control checks were removed. Next, I outline

[1] Due to their astonishing diversity, bacteria and viruses were excluded from these diversity counts to ensure that the counts provided are a legitimate reflection of the species diversity studied across the entire size range, rather than being an artefact of microbial diversity alone.

pooling information for taxonomic and sampling groups.

For most nekton, abundance estimates were given at the species level, with the exception of hard-to-differentiate taxa, e.g. striped and common dolphins. Unless the data had been provided that way by the primary source, no averaging or grouping was undertaken. For bacterial and viral data, I elected to use flow cytometric data rather than DNA-based methods, as the high variance in copy numbers of marker genes in prokaryotes precludes reliable abundance estimates. In addition, defining 'species' grouping is inherently problematic for microbes. No manually curated data was aggregated unless that was its original format. For the databases, I pooled according to the following principles. Firstly, I took taxa abundance averages by year and location. A single location was taken to be one station, or the same degree of latitude or longitude. We averaged at the lowest available taxonomic level (usually genus for organisms $< 5\mathrm{E}{-4}$ m, and species for anything larger), and selected taxa which, together, provided $> 90\%$ of the total abundance of that sample to avoid skewing with singletons. The exceptions to this pooling rule were for targeted flow cytometry counts of abundant cyanobacteria (*Prochlorococcus, Synechococcus*), which was included as is.[2] Figure 3.1 from the consistent variance across the size range and relatively narrow confidence intervals in the laminar data fits in Table 3.3. Abundance data is localised, hence spatial and temporal variation across local snapshots captures natural variability of populations across space and time. Therefore, the inclusion of data from different environments, e.g. tropical and temperate, or low and high biomass regions, or different species with different life history strategies, or across different sampling efforts, is suitable – and even desirable – as the goal is to build the universal distribution, which should ideally contain a broad spread of data [57]. Given the similarity between the manually curated and complete database results, and the generally well-behaved nature of the model statistics (Table 3.3, Appendix B), I elected not to transform or apply other corrections to the data. There is certainly variance introduced from differences in species trait differences [80, 81], and potentially from inconsistencies from

[2] Because microbial abundance data is a heavy tailed distribution, the 20 most abundant species typically make up over 90% of the total abundance of a sample. The tailed distribution also means that the abundance of the most prolific species are usually within 1 order of magnitude of each other as well as the gross abundance across all species. This means that either pooling all species together or including the densities of some individual, highly abundant, organisms will have negligible impact across a global size-abundance distribution such that is studied here as it will be within any noise factor of the data. Indeed, this can be observed in

underlying experimental methods. However, these impacts remain with noise factor of this dataset. Furthermore, whilst more targeted studies can be sensitive to this variance due to scaling size range and data limitations (e.g. bony fish, at 3 orders of mass magnitude) [57], fitting the scaling exponent over 23 orders magnitude, with this quantity of aggregated data, drastically mitigates the effect of any one source of error. Notably, the noise was sufficiently low for a strong statistical signal without the need for any manipulation, which could introduce other errors or biases, and reduce transparency of the result.

3.2.2 Standardisation and units

Due to the large mass range ($>$ 23 orders of magnitude) and measuring uncertainty in the body mass of microorganisms, I used body length, l (m), as the measure of organism size. To accurately compare data sets where abundance measurements were presented either as species numbers per unit volume or per unit area, and to account for organism behaviour, I calculated the separation distance, d (m), between organisms as a proxy measurement for abundance. To calculate separation distances, it was assumed the spatial distribution of organisms followed a Poisson distribution. Thus, the separation distance for organisms where abundance was measured per unit area was given by $d = C^{-1/2}$, and per unit volume, $d = C^{-1/3}$. We now discuss the raw data and the potential errors that may have arisen due to this standarisation. Plankton data was near universally presented by volume; I note that plankton distributions are by definition patchy and this variance far exceeds that of methodological error. Volume-based measurements in the reef fish dataset were based on study areas <30m deep and already undergone significant quality controls for accuracy; I did not undertake any further corrections. We assumed volume-based data for small nekton in the manually curated literature data did not require further adjustments. We acknowledge some small amount of error may have been introduced under this assumption in the event that depths were incorrectly measured, but note that (a) in the context of incorrectly measured depths, the cube root transformation reduces the impact of that error and (b) the data covers approximately 0.5 of an order of (length) magnitude, meaning that impacts on the full distribution would be minimal, particularly after log-transformation. For marine megafauna, only studies using standard methodologies according to transect and aerial surveys were included. Note that as the transformation of both axes is the same, the standardisation to length does not change the empirical scaling values, but ensures consistency with units in the physics-based processes and derivations used in the corrections and model.

3.2.3 Statistical model fitting and preliminary results

To determine the scaling relationship across the dataset, organism length was plotted against the inverse of the separation distance $^1/_d$ (m^{-1}) on a logarithmic scale, so that $d \propto l^{-\tau}$, where τ is a scaling exponent. Note that I consider a global, bivariate, size-abundance distribution more commonly applied in terrestrial settings, and not the univariate size distribution often studied in aquatic environments [57]. As previously observed within individual size spectra [11], nonlinearity was apparent in the log-transformed global size-abundance plot, where it appeared there was a break at $l \simeq 0.1$ m (Figure 3.1).

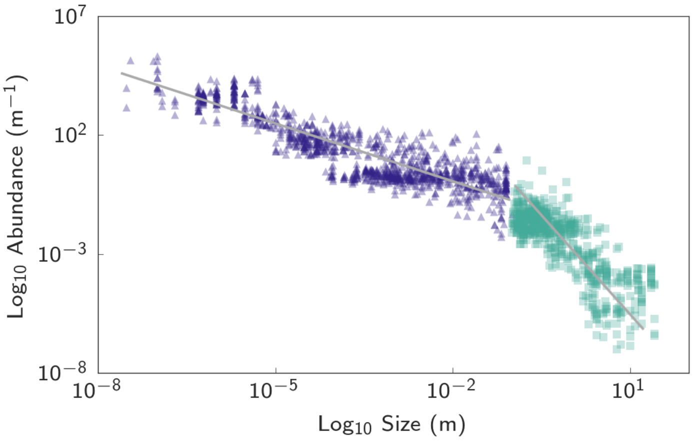


Figure 3.1: Size versus abundance for viruses to blue whales. There is a break in the scaling relationship at $\simeq 0.1$ m. Blue triangles represent plankton ranging from viruses to zooplankton, and benthic invertebrates. Green squares are fished nekton, ranging from small fish to whales.

In considering model fitting methods, my data is bivariate, meaning that methods developed for univariate distribution fits are not directly applicable [62]. Regression methods with log-transformed axes are standard for the bivariate case and may be used provided the dependent variable contains higher measurement error than the independent variable [37]. Therefore, models were fitted using ordinary least squares on log-transformed data. To determine the precise location where the distribution break occurred, I used used MATLAB's `fminbnd` function to find the segmented regression breakpoint which minimised mean square error. The breakpoint was bootstrapped for a percentile-based confidence interval on subsampled data, where the subsampling methods are as specified in the next paragraph. This revealed a break in the scaling value at the plankton-nekton transition of $l \simeq 0.1$ m ($l = 0.08$ m, 95% CI $(0.06, 0.1)$). Following an assessment of the residuals (Appendix B), I then fit linear models to determine the exponent within each size range of interest using the following process.

Firstly, a balanced subsampling routine was used to ensure an even spread of data across the distribution and improve fit quality [82]. We did not use a naive with-replacement bootstrapping routine as this would simply bias the sampling towards whichever data (taxa and/or sizes) were most frequent in my database. The data was stratified by organism sizes, and by taxa. We then randomly sampled m data points (without replacement) such that the quantity of data per (log)bin was uniform across the relevant size range and balanced the probabilities of sampling from different taxonomic groups. The optimal subsampling size m may be estimated by $m = kn^{\kappa}$, where n is the size of the dataset being drawn from, $k = 3$, and $\kappa = 0.5$ [82, 83]. We then generated $10,000$ parameter estimates from subsampled data. Percentile confidence intervals (95%) were created from the bootstrapped statistics (histograms are included in B).

Following model fits, it was found that marine virus to marine invertebrate slope at $\alpha = -0.77$ is comparable to terrestrial slopes [20, 37]. However, for organisms ≥ 0.1 m α was -1.9 (Figure 1a, Table 3.3), representing a significant negative perturbation in the slope. A shift in biomass would only translate the line downward (i.e. change the intercept via a step break), but the large slope break evidenced by these two exponent values (Figure 1) is indicative of a more fundamental alteration in the mechanistic processes shaping the species size-abundance distribution and ecosystem structure.

3.2.4 Correction for Fishing

To find the cause of the break in the marine size-abundance relationship, I note that fishing has reduced the abundance of fish, pinnipeds, sea turtles and marine mammals by up to 99% [84]. To investigate the impact of fishing on the observed scaling relationships, organisms were assigned to groups of impacted large marine animals according to standard conventions [84]. These included species ≥ 0.08 m, including fish, sharks, pinnipeds, whales, sea turtles and sea birds. Separation distances were corrected for each group to reflect theoretical historical abundance values, assuming losses ranging between 50 and 99.7% [84, 85]. Where no specific loss estimate was available, the mean decline for all large marine species (89%) was allocated [84].

We corrected for this by adjusting the abundances of impacted populations to pre-human impact estimates [84]. This caused an upward translation of the scaling line, removing the step break in the dataset and corroborating earlier findings [70]. However, whilst the translation is indicative of a decreased abundance of animals larger than 0.1 m, correcting for fishing did not result in a change in exponent, rather just a vertical shift in the data (Figure 3.2, Table 3.3).

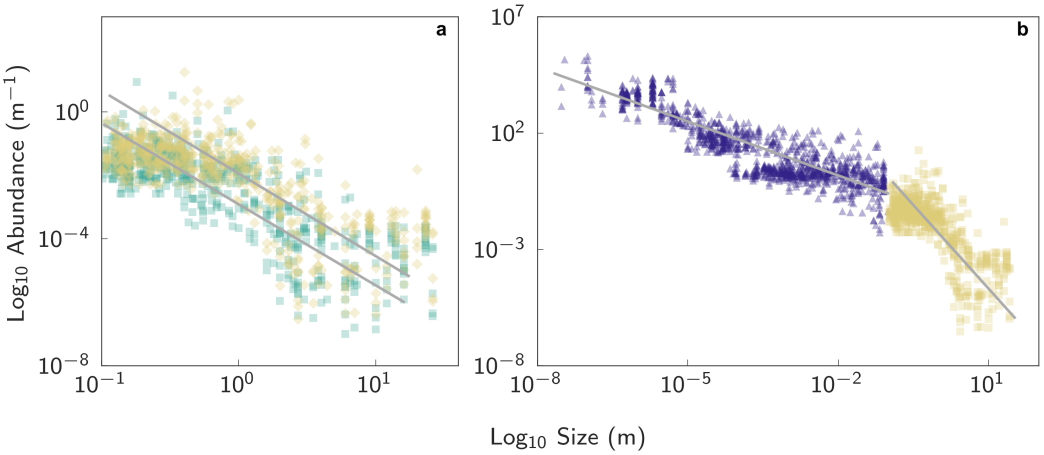


Figure 3.2: Size versus abundance for fishing adjusted data. (a) Shows corrected abundance for removal by fishing, with green squares the raw data, and yellow diamonds showing the fishing-corrected values. (b) Shows the fishing-corrected data substituted back into the full distribution.

In considering the drivers of this phenomenon, it is to be noted that the size-abundance distribution may be interpreted as an average or upper bound on local population densities[57]. The slope change is thus indicative of a constraint limiting nekton abundances which is not present in planktonic or terrestrial systems. To probe for a mechanistic explanation of the exponent change, I note that many aquatic organism scaling laws break at $\simeq 0.1$ m [14, 71, 86]; this size corresponds to the laminar-turbulent transition, where the change in the physical fluid environment causally affects the biology [71, 86]. We subsequently tested the hypothesis that the change in scaling value is due to implicit and explicit costs associated with turbulence: that is, nekton must expend energy actively moving to match planktonic prey distributions, and that this expenditure propagates through higher trophic levels.

3.2.5 Correction for Turbulence

Aquatic predators and grazers are challenged by the chaotic nature of turbulence. As absolute abundances of resources scale similarly in three-dimensional aquatic and two-dimensional terrestrial environments [13], their statistical distribution is scarcer in the three-dimensional ocean. Plankton live within patches created by an interplay of physical and biological processes [87]. Within these resource hotspots, plankton foraging and movement is localised and constrained within the patch, allowing them to use hunting strategies such as chemotaxis or rheotaxis to maximise their food acquisition [88, 89]; that is, plankton move passively with the turbulence that creates the aggregations. Beyond several millimetres and up to ten centimetres is a transition zone where eddies play an increasingly important role. Whilst they are below the swimming speeds of most fish, eddies on the scale of tens to hundreds of metres cause bulk transport and dispersal. Mesoscale eddies reach hundreds of kilometres in diameter and can move organisms hundreds or thousands of kilometres [90]. Food may not be transported, or it may be consumed and not replaced due to low light, low temperature or other unfavourable conditions [91]. Thus, nekton must migrate between patches to feed, which are continually and unpredictably dispersed, meaning they have resource encounter rates that typically cannot be bettered using local information [16]. Nekton live at a scale where the foraging landscape is highly fragmented and disordered due to these physical processes, and operate on biological timescales which are significantly longer than eddy lifespans [91, 92]. As they are trophically linked to the plankton, they must actively work to overcome the dispersal, ultimately increasing their locomotory costs, which also grow with prey size [93]. Short distance dispersal within or just beyond local habitats is difficult to quantify. However, at a global scale, physical dispersal – and consequently the spatial distribution of plankton – follows the Kolmogorov power law for the turbulent energy cascade [87]. The overall effect is that dispersal, encoded here as the separation distance, is a key factor in nekton survival. We propose that resource acquisition forces nekton movement to follow the turbulence-driven distribution of plankton, increasing energy expenditure [94], and consequently reducing available energy for growth and reproduction, which decreases abundances. The positioning of the break in the scaling relationship at the laminar-turbulent transition is consistent with this reasoning.

Table 3.3: Estimates of the scaling exponent (α) with 95% confidence intervals for the empirical data (raw and adjusted) and the model simulated data, all calculated from 10,000 bootstrapped values.

Regime	Manually curated data	Full dataset	Model
Laminar	$-0.74\ (-0.79, -0.69)$	$-0.77\ (-0.81, -0.73)$	$-0.73\ (-0.76, -0.71)$
Turbulent (raw)	$-2.5\ (-2.7, -2.3)$	$-1.9\ (-2.0, -1.8)$	-
Turbulent (fishing adjusted)	-2.5 $(-2.6, -2.2)$	$-1.7\ (-1.8, -1.6)$	$-2.1\ (-2.2, -2.0)$
Turbulent (adjusted)	$-0.94\ (-1.1, -0.74)$	$-0.56\ (-0.69, -0.43)$	-
Full spectrum (turbulence adjusted)	$-0.83\ (-0.88, -0.79)$	$-0.73\ (-0.76, -0.69)$	$-0.71\ (-0.72, -0.71)$

The influence of turbulence on the scaling relationship for netkon was addressed by applying a phenomenological correction for the $-5/3$ relationship arising from the Kolmogorov power law of the inertial subrange of the energy spectrum[3] [95]. The spectral energy density, a proxy of the variance of the variable under consideration, i.e. turbulent velocity fluctuations in the framework of fully developed turbulence, is given by $E(k) = C_k \epsilon^{2/3} k^{-2/3}$, where C_k is the Kolmogorov constant ($\simeq 1.5$) , ϵ is the turbulent kinetic energy dissipation rate and k is the wave-number (2π/(eddy diameter), rad.m^{-1}) [95, 96]. Here I approximate this relationship as $E(k) \propto k^{-5/3}$, providing a dimension of m^{-1}. The spatial distribution of plankton has been observed to follow the same power law [87, 97], and the separation distance d as a function of size (both units in m) may therefore be considered as an implicit measure of the effect of dispersion due to turbulence. Thus, by considering $d \propto k^{-5/3}$ I undertook a phenomenological correction for the abundances of nekton, whose foraging effort is impacted by the turbulent dispersal of plankton, by subtracting the Kolmogorov power law, intersecting at $l = 0.1$m, and calculated an adjusted scaling value for the entire data range.

[3] Note that whilst the inertial subrange, at a scale of approximately 0.1 to 1m-5m, scales to $-5/3$, as length scales extend beyond 10m, turbulence may begin to exhibit anisotropic properties with the exponent approaching -2. However, Langmuir circulations (i.e. Langmuir turbulence) that are on the scale of metres to kilometres are known to have properties of variabilities that are similar to small scale turbulence. Large eddies are also acknowledged as a source of energy supply to turbulent energy cascade. Hence the larger eddies ($>$ 1m) have significant contributions to develop the purely isotropic turbulence which shows the inertial subrange of the energy spectrum. Therefore, I apply an approximation where scaling behaviour follows that of the inertial subrange across the nekton size range.

Testing the hypothesis that turbulence increased the nekton slope by adjusting for the Kolmogorov power law, which affected small fish the least and large pelagics the most, removed the structural break in the distribution and resulted in a near-canonical exponent of $\alpha = -0.73$ for the entire spectrum (Figure 3.4, Table 3.3).

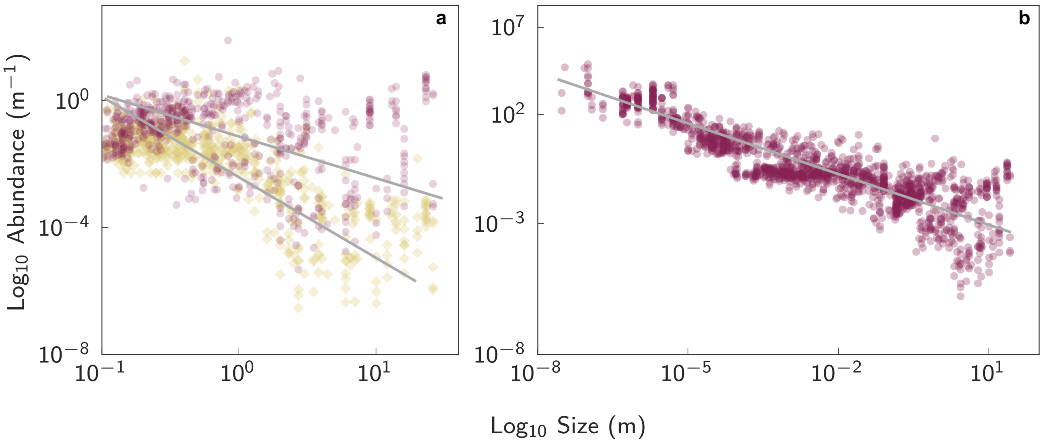


Figure 3.4: Size versus abundance for turbulent adjusted data. (a) Shows corrected abundance for removal by fishing and turbulence with yellow diamonds for fishing adjusted data, and red circles showing the turbulence-corrected values. (b) Shows the turbulence-corrected data substituted back into the full distribution. After both corrections all points fall along a line with a slope of –0.73.

3.3 A mechanistic model

3.3.1 Motivation

To build a minimal model which captures this phenomenon, I note other scaling breaks for aquatic organisms [71] also occur at 0.1 m due to movement changes at the laminar-turbulent transition [86]. The classical assumption that swimming is more energetically efficient than running [10] does not consider drag, which increases with the square of velocity and carries extreme metabolic cost [98, 99]. Research examining cost of swimming may also underestimate real-world metabolic effort for nekton as it frequently uses theoretically 'optimal' size-speed scaling [14] rather than utilising empirical values which are steeper [71]. Finally, relative consumption rates are higher in oceanic than terrestrial environments, yet a steeper inverse scaling of nekton abundances in marine systems exists even at high resource densities [13]. This discrepancy has not been resolved but indicates there must be a significant energetic cost associated with living and feeding in oceanic environments that has not been considered. I incorporated classical formulations of swimming cost for organisms living in laminar and turbulent environments, together with foraging effort, into a size-dependent predator-prey model to assess these effects. In short, I expand the trophic transfer efficiency parameter, ε, in the classical Rosenzweig-MacArthur predator-prey model to account for energy diversion toward locomotion. The following section, I incorporate classical formulations of swimming cost for organisms living in laminar and turbulent environments, together with foraging effort, into a size-dependent predator-prey model to assess these effects.

3.3.2 Predator-prey model

We used the classical Rosenzweig-MacArthur model to investigate the effect of turbulence on population dynamics and size-abundance relationships for consumer and resource pairs, from phytoplankton to whales. To maintain consistency in units across empirical data, model, and adjustments, size was given by length l (m) and abundance was defined as organism separation distance (n.m^{-1}), rather than size (g) and biomass (density, g.m^{-3}). The base ordinary differential equation contains strictly positive parameters and is described by

$$\begin{aligned} \frac{dR}{dt} &= rR(1-\frac{R}{K}) - \frac{\psi R}{H+R}C \\ \frac{dC}{dt} &= \varepsilon\frac{\psi R}{H+R}C - \delta C \end{aligned} \tag{3.2}$$

where R and C are resource and consumer abundances, respectively. The parameter H denotes the half saturation constant for a Holling Type II functional response, whereas K is the carrying capacity, r and δ are birth and death rates, ε the conversion efficiency, and ψ the consumption rate for the consumer. Throughout the model definitions and derivations I use subscript v to denote the viscous or laminar regime and subscript t for the turbulent regime. Here, I use an alternate expression of the Holling II term than in Chapter 2. This is because the assumptions applied in the biophysics section below are distinct to those in Chapter 2 which is predominantly focused on terrestrial settings. Regardless, it is possible to convert between the two versions of the system using the relationships $\psi = 1/h$ and $H = 1/(bh)$. As introduced in Chapter 2, each of the parameters in (3.2) follows scaling models according to the size of the organism. A 25°C standard temperature was assumed, as whilst temperature has some impact on metabolism, the variance in the data was sufficiently low to indicate that it was reasonable to not incorporate temperature effects; indeed perturbations to basal metabolic scaling properties across the livable temperature range in the ocean (approximately 30°C) are likely to have less impact on metabolic cost than an organism's locomotory strategies and behaviours [94, 99, 100]. Resource-consumer size ratios were varied between 0.01 and 0.5 (corresponding to approximate prey predator mass ratios of 1E-6 and 0.13 respectively). Exponents were given by representative values from previous research, which was typically specialised on deriving empirical scaling for that specific parameter. As my dataset ranges over more than 23 orders of mass magnitude, where there was some variability across literature scaling models, my study used the most "universal" exponents. Values chosen were (i) frequently reported with consensus (r, δ, ψ_v, K), (ii) mid-range (H) or (iii) specifically calculated for aquatic vertebrates (ψ_t). Here, the size of the resource (l_R) or consumer (l_C) and parameters which scale in the laminar regime are given as

$$\begin{aligned} r &= r_0 l_R^{-3/4} \\ K &= K_0 l_R^{-3/4} \\ H &= H_0 l_C^{-3/4} \\ \psi_v &= \frac{\psi_0}{\varepsilon_v} l_C^{-3/4} \\ \varepsilon_v &= \varepsilon_{v0} l_C^{1/8}. \end{aligned} \tag{3.3}$$

Note that length scaling values of $-3/4$ and $1/8$ are equivalent to mass scaling values of $-1/4$ and 0.04 respectively. As described in Sections 2.3.1 and 2.3.2, this paramaterisation is close to the null model of the allometric Rosenweig-Macarthur system. The scaling values for two parameters change between the viscous and turbulent regime (organism length > 0.1m): $\varepsilon_t = \varepsilon_{t0} l_C^{-1.3}$ and $\psi_t = \frac{\psi_0}{\varepsilon_t} l_C^{-3/4}$. Under this parameterisation, there is a switch to a positive consumption rate in the turbulent regime, whilst half-saturation H remains fixed. This occurs because a greater amount of resource is required to support a consumer population without translating to new biomass. As described in Section 2.3.1, the resultant length scaling exponent ($\simeq 0.55$) is reflective of observed empirical values for macroscopic fauna in aquatic environments.

The derivations of ε are outlined in Section 3.3.3, and a table summarising scaling values is provided in Table 3.5. Coefficients were standardised against phytoplankton/-zooplankton models to ensure the boundary value for primary producers was feasible. The smallest primary producer (i.e. 0.7μm in length) was assumed to be the cyanobacterium *Prochlorococcus* [101]. For coefficients, biomass was divided by species mass to obtain the number of organisms. Model equilibria were calculated using analytical solutions.

Table 3.5: Base parameters for the Rosenzweig-MacArthur model exponents taken from prior research (refer to Table 2.2 for references). For biomass-to-abundance conversions, the smallest primary producer was assumed to be Prochlorococchus with a mass of 100fg [101].*Denotes effective scaling in the laminar and turbulent regimes respectively under my parameterisation.

	Type	Mass scaling	Model value	Length scaling
r	Birth rate	$-1/4$	$-1/4$	$-3/4$
δ	Death rate	$-1/4$ to 1	$-1/4$	$-3/4$
ψ	Max. Consumption	$-1/3$ to 1	$-1/4\ {}^*(-0.29, 0.18)$	$-3/4\ {}^*(-0.875, 0.55)$
K	Carrying capacity	$-3/4$ to $-1/4$	$-3/4$	$-3/4$
h	Half saturation	-1 to $1/4$	$-1/4$	$-3/4$
ε	Conversion efficiency	-0.9 to 0	0.04 (-0.43 turb.)	0.125 (-1.3 turb.)

3.3.3 Locomotion cost: biophysics derivations for the model

To remain consistent with the literature, throughout this section, I use scaling of mass unless otherwise specified. To account for movement cost in the Rosenzweig-MacArthur system, I consider locomotion energy budgets across the whole size range (bacteria to whales). If movement energy usage scales equivalently to basal metabolic processes, its impacts would not be noticeable. However, if it scales differently, some of the energy previously used to create new biomass would instead be diverted to locomotion. Alternately, if locomotion were to become more efficient, additional energy could be provided for biomass. This can be seen by examining the gross metabolic power of an organism

$$P_{gross} = P_{basal} + P_{locomotion} \propto S^b + S^{loc}.$$

Normalising by P_{basal} results in

$$\frac{P_{gross}}{P_{basal}} = 1 + \frac{P_{locomotion}}{P_{basal}} \propto 1 + S^{loc-b}.$$

If there is a discrepancy between the power exponents, the (relative) locomotory power consumption will change across the size spectrum.

This deviation can be captured within the parameter for biomass transfer efficiency, summarised in an infographic in Figure 3.6. To achieve this, I use a classical ecological relation, which links basal and locomotory metabolic cost to abundance [17, 102], given as

$$N \propto S^{-b-c+q(F-D)}.$$

In this master equation, N is the number of individuals, and c is the relative transport cost scaling. I have $c := \Upsilon - b$, where b is basal metabolic scaling, and Υ is the scaling of transport cost (T_C) defined below. The term $q(F-D)$ describes search effort, including q, swimming speed scaling, and the parameters F and D, which describe density/fragmentation and dimensionality of the resource space. Note that if the term $-c+q(F-D)$ equates to zero, classical population dynamics apply. That is, the standard Rosenzweig-MacArthur system, with a typical value of ε e.g. the predator-prey mass ratio. However, when it is non-zero, it captures the shift in locomotion energy allocation across the size spectrum. This provides the relationship

$$\varepsilon \propto S^{-c+q(F-D)}. \tag{3.4}$$

In the subsequent derivations for the exponents of ε, I use empirical swimming speed scaling results from the review of marine scaling laws in [71]: 1/4 and 1/6 for viscous and inertial swimmers respectively. This is important because it implies the scaling of real-world nekton swimming speed is steeper than what would be theoretically derived for maximum efficiency. 'Optimal' speed scaling would be given as 5/24 and 1/12 for viscous and inertial regimes (calculated according to methods in [14] Supplementary Information, under the assumption of a 3/4 basal metabolic law).

Search effort scaling ($q(F-D)$)

The parameter q is simply the scaling of swimming speed. The dimensionality of the space, D, is taken as 3 for the turbulent regime. In the laminar/viscous regime, I consider $D = D' = 2.4$, to account for the patch constraint and the fact that organisms can use local information to optimise their hunting strategies [88, 89, 102]. We set

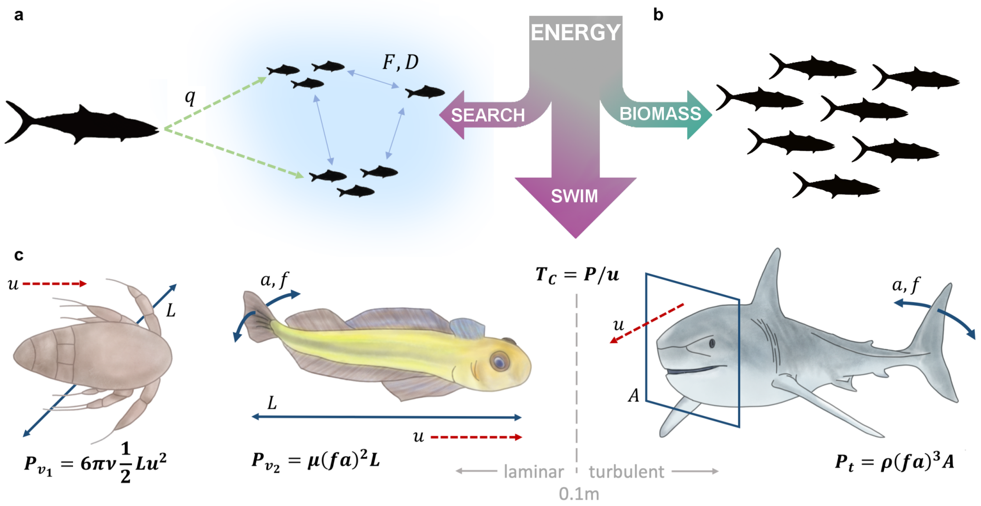


Figure 3.6: Energy partitioning - organisms have a finite energy budget which is split between movement and creation of new biomass. (a) A search effort term $q(F-D)$ is described by the scaling of swimming speed (q), as well as parameters F and D , which denote resources' fractal dimension (space-filling amount) and the physical dimension of the search space respectively. (b) Energy not spent on locomotion is utilised in reproduction and creation of new biomass. (c) Transport/swim cost (T_C) is defined as power, P, divided by speed u. In the laminar regime, power for viscous paddlers, such as copepods, is described by length (diameter) l, speed, and viscosity μ. Viscous undulatory swimmer power (i.e. larvae or small fish) is given by kick frequency f, kick amplitude a, length l, and viscosity μ. In the turbulent regime power is described by kick frequency and amplitude, frontal area A and fluid density ρ. I use these formulae to calculate size scaling exponents for swimming cost. The values can then be used in the master equation (Equation 3.3.3) to capture changes in energy partitioning across sizes.

the fractal dimension of the space, F, to a mid-range value of 1.9 [102]. Whilst this expression has some sensitivity at the extreme ends of the parameter ranges, I note that the multiplier q makes it a slow parameter. Therefore, standard values for F (between 1 and 2) and D (between 2 and 3) as outlined in [102] provide sensibly bounded solutions (Appendix A). Search effort is summarised in Figure 3.6a.

Transport cost scaling (Υ, c)

In this section, μ and ρ denote the viscosity and density of the liquid respectively. For the purposes of this study, I assume fluid properties are a constant value with negligible changes due to pressure or salinity and a temperature of 25°C. A conceptual summary of the equations used in this section may be seen in Figure 3.6.

Transport cost is defined as $T_C = P/\mu$ where P is power and u is swimming speed [14]. The master equations for the power of swimmers in the viscous regime are given by $P_{vu} = \mu(fa)^2 l$ for undulatory swimmers [14] and $P_{vp} = 6\pi\mu\frac{1}{2}lu^2$ for paddlers [19]. Here, f, a and l are the kick frequency, kick amplitude and body length respectively. By using the classical equality [86, 103]

$$f \propto \frac{u}{l} \tag{3.5}$$

and following standard convention by assuming changes in the length measurements a, l are scaling approximately proportional to $S^{1/3}$, I have

$$\begin{aligned} P_{vu} &= \mu(fa)^2 l = \mu(ua/l)^2 l = \mu u^2 a^2/l \\ &\propto M^{1/2} M^{2/3} M^{-1/3} = M^{5/6} \text{ and} \end{aligned} \tag{3.6}$$

$$\begin{aligned} P_{vp} &= 6\pi\mu\frac{1}{2}lu^2 \\ &\propto M^{1/2} M^{1/3} = M^{5/6}. \end{aligned} \tag{3.7}$$

That is, the power cost scales equivalently for paddlers and undulatory swimmers in the viscous regime.

For the turbulent regime, the power of inertial swimmers is given by $P_t = \rho(fa)^3 A$, where A is the frontal area of the organism (scaling as $S^{2/3}$ accordingly) [14, 86]. Once again, I use Equation 3.5 and substitute in mass scaling values to obtain $P_t \propto S^{7/6}$. Using the definition of transport cost, I obtain $T_{C_v} \propto S^{7/12}$ for organisms in the viscous environment and $T_{C_t} \propto S$ for the turbulent environment. As the units for T_C are J/m, it is possible to non-dimensionalise via multiplying by ρ/μ^2, which is simply a constant. This means that: $c_v = \Upsilon_v - b = \frac{7}{12} - \frac{3}{4} = -\frac{1}{6}$, and $c_t = \Upsilon_t - b = 1 - \frac{3}{4} = \frac{1}{4}$. With the values for c, q, F and D, I apply (3.4) to derive the scaling for ε in the viscous (3.8) and turbulent (3.9) regimes and convert to length scaling via

$$\begin{aligned} \varepsilon_v &\propto S^{-c_v + q_v(F_v - D')} = S^{\frac{1}{6}+\frac{1}{4}(-\frac{1}{2})} = S^{\frac{1}{24}} \\ &\propto l^{1/8}, \text{ and} \end{aligned} \tag{3.8}$$

$$\begin{aligned} \varepsilon_t &\propto S^{-c_t+q_t(F_t-D)} = S^{\frac{1}{4}+\frac{11}{10}} = S^{-\frac{13}{30}} \\ &\propto l^{-1.3}. \end{aligned} \tag{3.9}$$

We switch between the parameterisations at the length of 0.1 m, corresponding to the transition from laminar/mixed fluid regime to a fully turbulent flow of $Re > 1000$. Finally, the normalising constants $\varepsilon_{t_0} \simeq 1/100$ and $\varepsilon_{v_0} \simeq 9.5$ set initial values. The resultant mean, maximum and minimum conversion efficiencies are 0.09, 0.2 and 4E$-$3 respectively, which are within expected literature values [104].

For the $\alpha-$estimates generated from the Rosenzweig-MacArthur simulated data, I randomly generated m datapoints (matching the empirical subsample sizes) for the laminar, turbulent, and full size ranges. Confidence intervals were then generated from least-squares regressions on $10,000$ sets of the log-transformed model equilibria.

3.3.4 Rosenzweig-MacArthur Model results

Including locomotion cost for simulated predator-prey combinations from primary producers to blue whales reproduced the empirical results. Calculating the slope for model equilibria abundances in the turbulent regime resulted in a value of -2.1, consistent with the data (Figure 3.7a, Table 3.3). For the laminar model, and the turbulence-corrected predator-prey formulation across the entire data set, the slopes were -0.73 and -0.71 respectively, matching the empirical results (Figure 3.7b, Table 3.3). In my model, living in a turbulent fluid regime impacts the system by translating the predator abundances downward. This means prey support fewer predators in a turbulent environment than they would in viscous or terrestrial regimes because of the increased energetic costs of foraging in turbulence. Increasing locomotion energy budgets decreases biomass transfer to higher trophic levels where reduced prey availability places even more restrictions on energetic resources [93], pushing large marine organism abundances closer to an unviable population threshold where natural population fluctuations also render them more vulnerable to extinction [105].[4]

As my model includes a parameter for resource density, direct impacts of overfishing may also be incorporated. We find that whilst heavy fishing could theoretically perturb

[4] Whilst the model could theoretically provide predictions around the transition zone at 1-10cm, it is unlikely that a scaling-based model would provide accurate results for small size ranges or specific organisms.

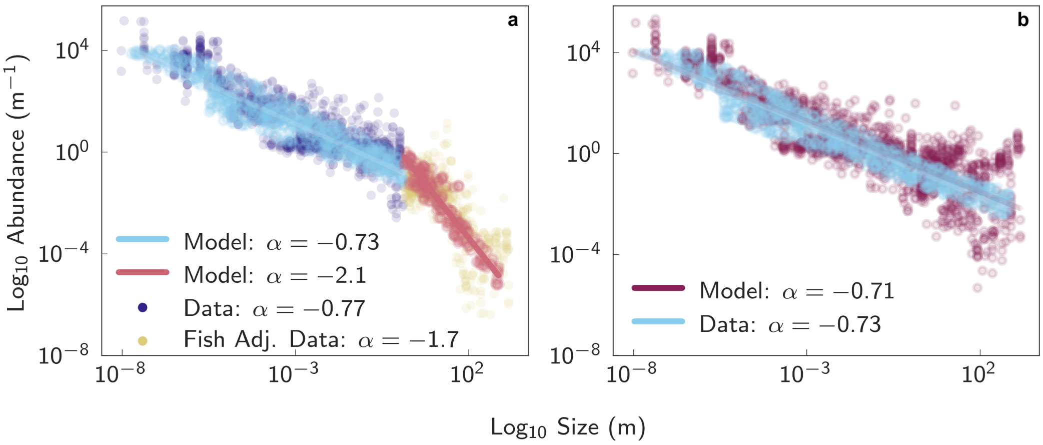


Figure 3.7: Rosenzweig-MacArthur model results. (a) Plankton (dark blue) with fishing corrected empirical data (yellow), the laminar model simulated data ($l < 0.1$ m, pale blue) and turbulent model simulated data ($l \geq 0.1$ m, red). (b) Fishing and turbulence corrected data (purple circles), are shown with the model simulated data (pale blue), whereby the laminar model is applied across the full size range, superimposed over the data-fitted regression line. Simulated data consists of prey-predator mass ratios between 1E-6 and 0.13.

the size-abundance scaling value by decreasing resource saturation and consequently reduce the parameter F, which denotes the resource's space-filling amount,, the search effort multiplier q is $\simeq 0.17$ (relative to mass). This means it is a slow parameter, which also reaches an asymptotic value as $F \to 0$. Hence, whilst fishing removes biomass, my integrated model indicates it could only perturb the scaling law by $\simeq -0.2$ before the asymptote is reached. This is an order of magnitude less impact than turbulence effects, and entirely consistent with the data (Table 3.3).

3.4 Accounting for fishing induced evolution

A complicating factor with my analysis is that organisms and biomes are not fixed physical or chemical variables. Their characteristics can change in response to environmental pressures. Ecosystem-wide size shifts in size-abundance relationships may be exacerbated by compensatory genetic changes, particularly when they have occurred under strong selection pressures such as fishing. Such a fisheries-induced evolution (FIE) causes further size reduction and earlier maturation age [106], which could alter the scaling relationship. To assess the relative impact of FIE, I extracted data from 113 time series for 10 commercially exploited species of fish, and assessed global changes in size and age at maturation. In some cases, this was provided as probability norms of weight or length at 50% maturity (Wp50 or Lp50). Time-series with large gaps or fewer than 20 measured time points were excluded. Data was manually extracted using WebPlotDigitizer (v 3.12) and visually verified by replotting and super-positioning over the original. For plots without discrete data points (i.e. smooth line graphs), one data point per year was used. Each time series was normalised and then split in two halves, for which mean values were calculated for the first/second half of study period. This was imported into a data structure consisting of the mean values, data type (size or age at maturity, 50% maturity, Wp50 or Lp50), gender, species, and length of study. For testing the difference in means between the first and second halves of a study period, data was firstly assessed for normality by using a 2-sided Kolmogorov-Smirnov test ($n = 113$, critical value=0.1262, observed values 0.0774 and 0.0958 for pre- and post- respectively, MATLAB R2016b, Mathworks). A paired t-test (SPSS 24.0.0.0, 2017) indicated a 10.6% shift in mean value in the second half of the study period (df=112, 95% CI $(9.4, 11.9)$, 2-tailed, t-statistic -17.374, p< 0.001). That is, there was a mean decline of 11% in size or age at maturity, when accounting for gender, species, and length of study. The results from 10 of the 14 studies led to the conclusion that these changes were attributable to fishing pressure [106]. In considering FIE's contribution to universal size-abundance scaling, the breadth and size of my dataset gives insight into the signal-to-noise ratio for this problem. It would be extremely challenging to detect shifts in a global scaling law over the restricted size range of 0.1 to 2 m used for FIE impacts. While prior research suggests that FIE can perturb local scaling properties[107], I argue an 11% impact (or even significantly greater) would not be enough to shift the global size-abundance scaling value of nekton by -1 or more. We conclude that scaling alterations occurring due to FIE would be small relative to the turbulence effect explored in this chapter.

3.5 Conclusions

Global size-abundance laws provide a different form of ecological insight to that given by local scaling behaviour, as they capture macroscale, aggregate processes rather than examining small-scale drivers such as inter- and intra-specific trait variation [57]. In this context, I introduce turbulence, and its impact on energy and movement cost for large organisms, showing through empirical and modelling approaches that it is a novel but important process to consider for the large scale organisation of ocean ecosystems. Climate change impacts have the potential to exacerbate these costs, as current and predicted increases in ocean surface energy [108] will increase nekton locomotion costs [109], forcing increased movement cost and potentially decreasing energy available for reproduction. These losses may be further exacerbated as warming temperatures increase respiration rates, reduce global primary productivity [110], and cause greater resource patchiness [111], effecting higher foraging effort. Turbulence may thus reduce the capacity of nekton to withstand fishing pressure as I begin to observe oceanic anthropogenic impacts classically associated with terrestrial systems, including loss of large apex predators, shifts to smaller size, and a faster onset of sexual maturity. We propose that a deeper understanding of the role physical mechanistic processes play in structuring marine ecosystems will be necessary when formulating strategies to preserve biodiversity and retain the productivity of ocean resources in future.

Chapter 4

Consistency and stability despite complexity and chaos in the microbial biosphere

Chapters 2 and 3 examined the global size-abundance scaling distribution of organisms spanning all domains of life. Here, I shift from all size scales to focus on prokaryotes, the smallest living organisms within the distribution, to investigate organising principles within microbial communities. These communities anchor scaling laws across terrestrial, freshwater, and saltwater systems and thus play a key role in determining ecosystem structure. Given the ubiquity of global scaling laws, I investigate whether there may also be consistency within the community structure of the microbes that underpin these macro distributions. Microbiomes across diverse environments display remarkably similar emergent behaviour, implying communities may be governed by universal organising principles. Here, I outline some of the challenges behind studying these communities, existing conjectures behind their ecological properties, and a path forward for studying the distributions and scaling of taxonomic and genetic diversity within microbial systems.

4.1 Background

Prokaryotes are ubiquitous and of critical importance for the biosphere. At a global scale bacteria and archaea control biogeochemical processes, playing a central role in regulating the carbon, nitrogen, and sulfur cycles [112]. In natural systems prokaryotes rarely exist as monocultures. Rather, microbes form complex multispecies communities as they undertake a variety of functions required to survive and reproduce. Yet, despite their importance, there are many gaps in our understanding of the rules under which these communities operate. The combinatoric complexity alone poses major challenges for studying these systems. Global species diversity estimates are in excess of 10^{12} [113–115]. 'Low' alpha diversity systems such as host-mediated microbiomes generally possess several hundred species [116]. However, environmental biomes may have thousands - or even tens of thousands - of species, and microscale temporal or spatial shifts in sampling may present an entirely different community to study [115, 117, 118]. Further adding to the complexity is that the vast majority of microbes are unable to be cultured in a laboratory [119–122]. Even were it feasible to do so, interaction-based experiments are a combinatoric impossibility due to the number of species in natural systems. It is also unknown whether the dynamics of idealised assembly or interaction based experiments scale up and are genuinely reflective of the rules governing real-world communities, although it seems likely that accounting for function, as well as taxonomy, is a necessity [123–125].

Further complications in the study of microbial communities stem from extreme variance within most of the system parameters. Reproduction times may vary between under 10 minutes and thousands of years as cells may lie dormant under unfavourable conditions, yet retain potential for reactivating their metabolic pathways for tens or hundreds of millions of years [67, 68, 126–128]. Bacterial genomes are fluid, having the capacity to hypermutate or rapidly assimilate DNA from the environment [129–132]. Population density is also variable with species abundance distributions being power law or otherwise heavy tailed and 5+ orders of magnitude or more separating the most abundant and most rare species [117]. Additionally, identical initial conditions may give rise to differences in dominant species over time as the dynamics are nonlinear and chaotic on top of underlying processes which are stochastic in nature [133–135].

Instantaneous growth rates are not necessarily correlated to abundances and population fluctuations over short timescales may span many orders of magnitude [133, 136].

In light of these factors it is surprising that microbial communities *in situ* are resistant to changes and resilient to perturbation [137–141]. Indeed, this stability appears to be a feature which is common across all biomes - that is, a universal characteristic of microbial systems. I propose that this stability is an emergent property of the way in which metabolic pathways are distributed across taxa: these complex communities are able to respond to disturbances by leveraging a specific redundancy structure and gene sharing processes unique to prokaryotes, a signature which should be detectable through a ubiquitous genetic structure in these systems.

As for many other biological problems, the resolution that the microbiome is measured at appears to impact the level of stability observed in the community: scale does matter [123, 133, 142]. Nevertheless, locally sampled communities - whether in a host or environmental biome - are usually more similar to themselves than to samples taken elsewhere. This principle appears to apply at multiple levels of resolution. For instance, the human microbiota can be considered as a ‘fingerprint’ from person to person due to its individual specificity and asymptotically stable behaviour [140, 143, 144]. That is, the community may be perturbed under varying conditions but will return to its original composition when the environment shifts back to its initial state. At a larger scale, it has been noted that there are national or continental wide patterns in gut microbiomes, such that geography - and corresponding lifestyle - is predictive of community composition [145]. In turn, human microbiomes will generally be more similar to other human microbiomes than the bacterial communities associated with oceans, soils, lakes, or other environments [115]. This nested self-similarity is also applicable in other biomes, both in terms of taxonomic and functional diversity, with the caveat that physical/chemical partitioning of geography is usually a stronger predictor than distance alone [146–148]. Given the unpredictability and chaotic behaviour of the individual components within bacterial communities, stability seems so unlikely that one author has commented that there appears to be an ‘invisible hand’ behind the processes, which in turn implies consistent organising principles within microbiomes [141].

One proposed mechanism for the observed stability is through the biodiversity-stability relationship, which has been investigated in ecological systems for over 70 years [149–151]. The core principle is that that biodiversity improves stability by providing functional diversity, allowing for the system to manage a variety of perturbations, and functional redundancy, which mitigates against functional loss following the extinction of any particular species. Microbial systems classically carry high species diversity and high functional redundancy [152, 153]. In many natural systems, the diversity may be so high that it appears to be in excess of what may be expected given the presence of opposing processes such as competitive exclusion. However, it is debated as to what extent purely competitive interactions dominate in real-world communities [154]. Despite the fact that exclusion is a common outcome in laboratory settings there are also conflicting results, particularly within more complex environments [124]. Several physical and biological factors are proposed to contribute to this higher than expected, but empirically evident, community species diversity. Spatial heterogeneity mitigates against the capacity for a single species to have maximal fitness across all local conditions, decreasing the likelihood that particular organism can dominate across all microenvironments [152, 155–157]. Spatial processes can contribute to niche differentiation, which is enhanced both by prokaryotes' capacity to (a) utilise multiple different metabolic pathways or cross-feeding for hard-to-obtain nutrients and (b) specialise on different compounds for key resources [158]. These processes appear to take place even in oceanic or aquatic environments where fluid movement generates microscale nutrient gradients [157, 159]. Migration and dispersal may then provide the potential for organisms to spread, invade, or re-invade, as the resource landscape changes over time [160, 161]. It has been shown that phylogenetically diverse communities are more resistant to invasion, potentially due to the lower probability of a particular niche being vacant [162]. The interplay between resource patchiness and shared metabolism may be particularly important for oligotrophic environments, where maximal nutrient affinity and absolute reproduction rate may only be temporarily, and locally, beneficial, providing scope for more variable interactions than competition alone. Additional complexity is then provided by potential trade-offs via allelopathy, phage or other resistances, and buffering against other environmental stressors [152]. Given the extreme fitness cost of carrying a large genome, in this context, there may be higher levels of diversity associated with a finite set of functions than would otherwise be predicted [163, 164]. The functional overlap between the species then provides redundancy across the system, which in turn confers stability.

The notion of functional redundancy in microbial systems appears to be more complex than in areas of macroecology. The classical view on the taxonomy-function link in microbial systems is that diversity and growth and adaptation rates are so high that it is merely the "environment that selects" [165, 166]. That is, there is always sufficient metabolic capacity in a community for rapid optimisation on current conditions. However, it is now clear that a more nuanced perspective applies. Whilst it is certainly the case that functional profiles are generally conserved within similar environments, taxonomy is not necessarily so [123, 148]. In one of the largest investigations of community assembly to date, researchers were able to show that despite the emergence of consistent functional profiles, imposing one environment on different communities will lead to significantly different species assemblages, whereby the finest phylogenetic resolution of an attractor is at the family level [123]. Furthermore, it has been shown that such differences are likely to persist over time [167]. Another issue is that redundancy is typically dependent on which function is quantified. Common functions, such as production, do not appear to be strongly linked to taxonomic diversity, and the community may lose a large number of species without any discernible impact on the level of redundancy in the system [153, 167]. More specialised functions, however, are far more sensitive to species loss [153]. In turn, when considering multiple specialised functions, community redundancy of *all* functions is more reliant on the presence of broad phylogenetic diversity [153]. Timescale has also been shown to be important. In the short term, diversity losses may not lead to functional losses or affect the stability of the system, but with each successive generation, the larger the negative impact is likely to be [167]. The means to probe the interplay between these effects has improved over the last decade as the cost of shotgun whole genome sequencing (WGS) has decreased, allowing for large scale quantitative approaches to the problem. However, many challenges are also present within empirical methods, ranging from missing data to biased or incomplete databases which we outline fully in Section 4.2. Hence, whilst functional redundancy appears to be an important contributor to community stability, many open questions remain.

A second mechanism which has been proposed to effect stability in microbial systems is horizontal - or lateral - gene transfer (HGT). Prokaryotes are unique in their capacity to share DNA through HGT, which allows them to obtain novel genes from other cells in the environment. Of the various types of mobile genetic elements, phage and plasmids dominate HGT processes in bacteria [132, 168]. HGT is ubiquitous in microbial communities, with an estimated 20% of genes being recently acquired [169, 170]. It has been hypothesised that microbiomes are connected at a cross-continent global scale through HGT processes [171]. Genes which have been acquired through HGT are typically present in specific regions of the genome, in so-called 'gene islands', rather than spread evenly throughout. As metabolic functions usually work in blocks of genes, this is thought to prevent damaging critical cellular processes such as division [169]. It also appears that both phylogeny and ecology restricts who shares with who, with HGT being more probable amongst those occupying similar habitats and between closely related species [170–172]. Additionally, genes involved in HGT are usually associated with secondary metabolism and membrane transport, often being associated with cell defence [169]. Indeed, one of the ways to increase the rates of HGT in a community is to perturb the environment with a particularly strong driver being the imposition of a stressor which induces a SOS response in a cell [168]. The proposition that HGT improves stability in microbial systems has been examined from multiple perspectives, with a wealth of supporting evidence from experiments, data-driven models, and purely theoretical models [173, 174]. In essence, HGT stabilises communities by providing a method for vulnerable populations to acquire survival related genes, whilst also allowing 'streamlining', thereby improving an individual cell's chance of success [174].

Streamlining theory proposes that reduction of cell complexity, typically equating to a smaller genome through gene loss or smaller cell sizes, confers a fitness advantage. This is primarily due to the reduction in cost of replication, but may also be due to physical factors including the increased surface area/volume ratio of a smaller cell, or biological factors such as the facilitation of alternate metabolic pathways [163]. This advantage may be particularly strong in resource limited landscapes. There is an expectation that streamlining effects are likely underestimated in nature due to the difficulty of culturing most environmental bacteria [163]. The persistence of large genome sizes across multiple biomes may be explained by the fact that certain niches require or benefit from more complex functional pathways - it is still possible for cells to undergo streamlining within the bounds of the restrictions and needs of a particular niche. In the context of system stability, there is increasing evidence that horizontal gene transfer is especially prevalent amongst streamlined cells [163]. This would provide both the efficiency advantages of the small sized, small genome organism, but also allow for the capacity to uptake genes if and when they become advantageous to hold. Provided that genetic memory persists somewhere within the full community, the cost to individual cells may be minimised whilst the system as a whole retains the capacity to manage the full spectrum of stressors.

The hypothesis motivating the remainder of the thesis is that it is not just functional redundancy, but how the redundancy is distributed, which is conferring this stability. We propose that there is a universal genetic structure in microbial communities which allows for a rapid propagation of genes through the whole system via HGT. In this way, it would be possible for species to avoid the expense of carrying every stress or resistance related gene - which would be extremely metabolically expensive - but simultaneously be able to gain access through closely related organisms should it be required. In this way, the community as a whole could be robust to all but the most rapid, extreme perturbations. We shall now investigate this problem by creating an analysis framework which allows me to assess the paired distribution of taxa and genes, and whether there are common topological features across all biomes.

4.2 Empirical challenges: taxonomic and functional classification

Our first consideration is the use of shotgun WGS data to reconstruct the taxonomic and functional structure of microbial systems. WGS classification software requires the use of reference databases, even for alignment-free methods. In 2016, Hug *et al.* undertook an extensive analysis of unannotated database sequences to examine the spread of genetic diversity in nature [175]. The results of that analysis near doubled the known tree of life, with a large portion of diversity belonging to microbes that had never been seen in a laboratory setting [175]. However, with many of these genomes belonging in the category of metagenomics assembled genomes (MAGs) they are not included in reference genome sets, with the standard being RefSeq from the National Centre for Biotechnology (NCBI) [176]. This can be problematic when undertaking a cross-biome survey. Environmental biomes, with their high proportions of microbial 'dark matter', may have a significant majority of data which cannot be classified within standard databases such as RefSeq [119, 177]. A further complication is given by the fact that most databases are biased toward human microbiome data and species which can be cultured in the laboratory [176, 178]. For example, RefSeq contains approximately 15,000 genomes and over 3000 species or strains of *E. Coli.* This can create biases within taxonomic classifiers, where there is a methodological tradeoff between precision and recall [179]. One instance of this may be seen in k-mer based software, which is extremely fast but there can be a high probability of mismatches between strains; the more uneven a database, the more serious the potential bias. Other packages use methods which match to unique marker genes, such as MetaPhlaN, which uses NCBI's non-redundant database to accurately detect the presence of individual species [180]. However, whilst MetaPhlaN's results can be interpreted with a high degree of confidence, a significant amount of data may remain unclassified, especially within non-human biomes.

Given the focus of this project on cross-biome analysis, it is critical that any reference genome set is reflective of the full spread of prokaryotic diversity, including MAGs. Additionally, it should not be biased toward species which are more commonly studied. The Genome Taxonomy Database (GTDB) fulfills both of these criteria [119]. The number of species selected for each leaf of the phylogenetic tree reflects the proportion of genetic diversity that clade contributes to the full tree of life, and contains 28,439 prokaryotic genomes. To compare with NCBI's RefSeq, in GTDB, there are just 4 species of *E. Coli* - they make up less than than 0.02% of the species in the database as opposed to RefSeq's 20%+. GTDB is designed as a phylogenetically balanced and standardised reference set representative of cross-biome genetic diversity, making it suited for our application.

Functional profiling also has fundamental challenges. Protein annotation traditionally relies on translated searches to find amino acid (AA) sequence matches in reference databases, e.g. by using local alignments or hidden Markov models [181, 182]. However, due to exponentially growing sizes of these reference sets, this approach is increasingly computationally demanding [183]. Furthermore, as for taxonomic classification, most tools rely on preexisting database sequences to return a match. However, unlike phylogenetic assignments for unknown genomes, which may be determined through software-based methods, functional classification of unknown proteins usually requires manual experimental assessment. This means that the rate of growth of unknown protein sequences in databases has far exceeded the rate at which their functions can be assigned. Further complications are introduced by ontology design. Commonly used ontologies such as KEGG were developed with a human or eukaryotic focus and labelling is biased toward those applications; to the best of our knowledge, the only widely used ontology designed for prokaryotes is SEED-subsystems [184].

As our goal is to study the underlying structure of microbiomes, i.e. how genes are distributed amongst species, it is desirable to maximise coverage not only for taxonomic but also functional classification. Coverage of bacterial genomes when annotated with KEGG is on average just 50%; coverage may be even worse when annotating genes in environmental samples [177]. The annotation method with the best coverage is Pfam (protein families), which captures domains within a coding region (CDS). The issue with using Pfam is that a CDS may have multiple Pfam domains; different combinations of Pfam domains may result in different functions. An alternate approach to maximising coverage is to use software which attempts to gain the best of both worlds. Eggnog uses a support vector machine trained on a non-redundant sequence set to match proteins to homologs. These may match to Pfam domains but will return significant matches to more complex functions. However, Eggnog (like Pfam) is not organised into an ontology meaning that there is no way to group the tens of thousands of unique labels into high-level categories for interpretation. These issues motivate our choice to use a range of annotation methods in our pipeline. This allows us to maximise the information content, ensures robustness of results, and permits us to place our results into a biological context through use of an appropriate ontology.

4.3 Network representations of microbial communities

The second consideration in tackling this question is how to structure the data and perform the analysis. There are multiple works in the literature titled ‘The structure and function of ... microbiome.' [148, 177, 185–190]. However, the overwhelming majority of these papers - and papers within microbial ecology - study the taxonomic and functional profiles of samples in isolation. Classical analyses apply measures such as richness, α- and β- diversity, SIMPER and PERMANOVA [191]; any associations between taxa and function have either only applied to the dominant taxa and functions or have been achieved through inference by using reference genomes, correlation methods, or linear or statistical models [153, 187, 190, 192–194]. We propose that a network representation of the data may allow us to formally elucidate the underlying structure of microbial communities by explicitly linking taxa and genes, and allowing us to study their joint distribution through network properties and statistics.

Previous attempts to capture the network structure of microbial communities have largely been driven by various forms of correlation or similarity analysis [195, 196]. Samples are taken in the form of a spatial or time series, and statistical or information theory-based tests are applied in a pairwise fashion to determine whether two species appear to be linked [197]. Unfortunately, there are major problems with this approach, which also neglects to account for genome function. Most statistical methods have inherent assumptions regarding the form of the data or the data's underlying dynamics. For instance, the application of any correlation test is problematic due to the non-paramteric distributions in the data. The microbial species abundance distribution is tailed [117]; making matters worse is that the dominant and rare species may fluctuate in unpredictable ways. Secondly, the sparse nature of microbial data means that the number of zeros in the datasets tend to skew the results, giving falsely low p-values and artificially inflated R^2 values on any classical correlation metric. Attempts to rectify these issues have been only partially successful and there are doubts as to whether correlation networks may meaningfully infer any properties of microbial systems [197].

An alternate approach which does not need to rely on inference is through applying network methods to sequence information directly. Networks constructed from k-mer or genome similarity have been studied since mid-2000, often in the context of studying HGT, and often in prokaryotes [198–200]. Whilst methods of constructing these networks vary, the common principle is that taxa (or genomes) - represented by nodes - with more shared genes have a higher probability of being linked together, or that link may have a higher weight [201]. However, as has been identified as an important consideration within the social network literature, these unipartite networks are actually projections of an underlying bipartite graph [202]. The underlying bipartite graph is defined by the top set of nodes being taxa, and the bottom set genes. Links indicate whether a particular taxon has that gene in its genome. It is preferable to study the bipartite graph rather than its projection. Projections cause a loss of information; whilst features of the bipartite network do drive the statistics of the projections, it is often not possible to disentangle which specific properties from the original network are causing which specific effects within metrics of interest [203]. Secondly, if there is a tailed degree distribution in one or both of the node sets, the projections become dense and require thresholding, which loses even more information.

Existing literature using native bipartite networks in microbial systems usually constructs them by using a large database of non-redundant protein sequences (the bottom set), and tools such as blastp to identify homologous matches within viral, plasmid or prokaryotic genomes (the top set) [204–206]. These networks are simplified by merging redundant nodes (a set of proteins which share the same two or more genomes), deemed ‘twins’, resulting in a smaller network for downstream analysis. These networks are used specifically to identify HGT genes in reference databases. Although this method has the benefit of being able to match genes across the phylogenetic tree, there are some limitations. Firstly, it requires the use of a secondary plasmid database to construct the network and is built using alignment based methods, which limits scalability. Secondly, examination of the networks has often been descriptive, using overlap counts and majority rule analyses rather than using graph-based approaches, especially when larger databases have been used [207]. Otherwise, the number of genomes has been limited on the order of hundreds up to approximately one thousand [205, 206].

An alternate approach is to construct the network is by using a taxa-gene function bipartite network. This network may be used to examine the joint distribution between taxa and functions. For example, in [192], the authors investigate the human microbiome, arguing that this network is nested, which in turn increases the functional redundancy of the community. However, there are limitations with the work. Whilst nestedness is a classical ecological metric on bipartite networks, its significance has recently been questioned [208]. Nestedness is extremely common in bipartite networks as it can emerge as a result of other mechanisms within the graph and is strongly correlated to node degree. Indeed, in [192], once the graph's degree distribution is controlled for, the p-value for the main result increases to be just within the margin of significance. Furthermore, the power law species abundance distribution in the data may affect the functional redundancy metric proposed in the work, due to the sensitivity of any redundancy score associating reference genomes to abundance distributions [193]. Finally, whilst biology is clearly of fundamental importance within microbial communities, meaning that different types of genes may display different network properties, this was not accounted for.

In the following chapters, we construct and then analyse taxa-gene function bipartite networks, firstly for the complete GTDB tree of life, and then within multiple biomes. Through moving beyond mean-field metrics such as nestedness and breaking down the behaviours of different gene classes, we are able to determine key topological features within microbial communities, and identify whether there is a universal gene-taxa structure which exists across all environments.

Chapter 5

A graph of the prokaryotic universe

Bacteria and archaea carry the majority of the world's genetic diversity, and the metabolic organisation of the prokaryotic tree of life is critical to our understanding of the composition of the biosphere. Exponential growth in sequencing databases has led to a need for new and scalable analytical methods to examine distributions of functional diversity amongst taxa. Here, I show network based methods can reproduce analytic results grounded in phylogeny, and identify statistical properties in the tree of life's taxa-function network to reveal macro scale patterns in metabolic diversity.

5.1 Motivation

The pursuit of deep sequencing to explore the microbial biosphere has led to ever increasing catalogues of taxonomic and functional diversity [175]. Debate is ongoing over how to best organise and manage the information, from processing the sequencing data, to analysing the resulting outputs, and even how to define fundamental concepts such as species [119, 191, 209, 210]. However, the trend is clear: with increased diversity comes increased complexity and a need for new analytical approaches [115]. Prior work on the distributions of metabolic pathways in the tree of life is primarily informed by an evolutionary approach [205, 206, 209], yet here, I consider the problem from the perspective of ecosystem functionality. Through creating a binary network encoding of taxa-function relationships in the microbial tree of life, I aim to capture the organisational principles of functional diversity across the prokaryotic universe.

5.2 Core genome annotation

To create a network representation of the system, it was necessary to develop a database of annotated reference genomes. To avoid skewing results through database biases, the reference set was taken from GTDB's dereplicated taxonomy; this reference set is also designed to be an balanced representation of functional diversity across the phylogenetic tree [119]. Whilst the majority of the 28,439 available genomes were sourced through the NCBI's Genbank [211], 300 archaeal genomes were downloaded from GTDB site. At the time of this analysis (November 2018), annotations for the GTDB database were not available.[1] Futhermore, translated sequences were not necessarily available from Genbank for metagenome assembled genomes. I therefore downloaded all genomes in nucleic acid format, and then used the software Prokka (v1.14.5) [212] to translate the sequences, generate the CDS regions and construct files in genbank format (required in a later step of the pipeline). All amino acid sequences were then annotated. To ensure that different annotation systems did not substantively alter my results, I ran my analysis on three different annotation systems - Eggnog, KEGG and Subsystems - each of which has different drawbacks.

Across all genomes, there were a total of 91,083,952 genes. To maximise the amount of information within the network, I endeavoured to maximise the annotation coverage; to achieve this, I used the software Eggnog v4.5 with default thresholds [213]. The Eggnog software also returns any significant hits to unique KO identifiers from the KEGG ontology [214]. For the Eggnog-based gene node labels, I compiled a list of all of the unique functions returned, totalling over 60,000. Many of these 'unique' labels were duplicates with alternate capitalisation or other trivial typographical differences; I manually dereplicated them to ultimately obtain 30,208 unique labels (refer to code package for mapping files). For the Subsystems annotations, all available sequences with a functional annotation were downloaded from PATRIC - over 301 million in total [178]. These were then dereplicated to form a non-redundant database of 16,475,282

[1] KEGG, COG and Pfam annotations have since been provided through [209].

sequences. Annotations were then obtained for the GTDB reference genome sequences by using blastp against this database, with an e-value threshold of 1E^{-}5 and taking the top hit.

In Table 5.1, I show the coverage obtained for each annotation system. Although Eggnog has the greatest coverage at 79%, which likely captures almost all of the coding genes, the lack of a hierarchy (ontology) attached to over 30,000 unique labels means it is extremely difficult to group genes into categories and thus interpret results. I found the least coverage is given by KEGG at 56%, which is consistent with findings from recent research [215]. Furthermore, although KEGG has a labelled ontology, many of the named categories are human or eukaryotic associated, such as 'cardiovascular disease', making their interpretability in a prokaryotic setting problematic. It also has relatively few hierarchical levels, jumping from 4 classes to 300+, limiting its usefulness for my application. At 67%, Subsystems had less coverage than Eggnog, but more than KEGG. As the upper levels of the SEED ontology move from 11 to 30 to 132 classes, they enable me to summarise the structural behaviours of different functional categories and thus achieve my goal of quantifying their distributions within the network.

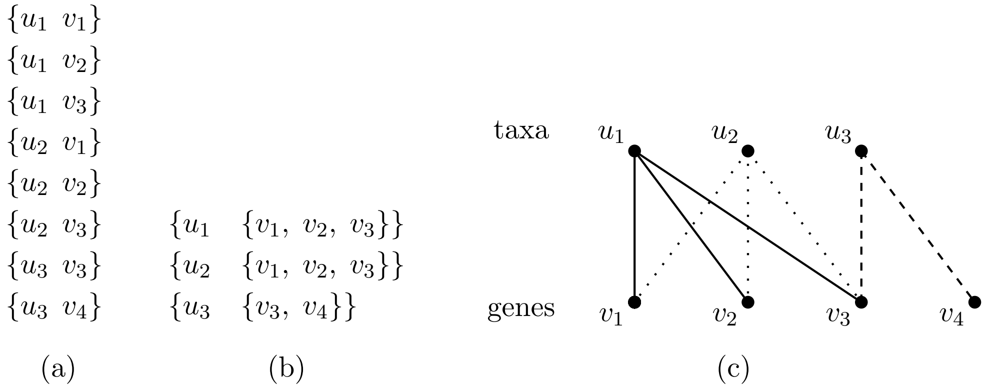
Table 5.1: Annotation hits and percent coverage out of 91,083,952 genes for different systems or ontologies.

Annotation	Number of genes annotated	Coverage
Eggnog	72, 039, 807	79.1%
Subsystems	61, 150, 266	67.1%
KEGG	50, 514, 144	55.5%

5.3 Network construction

To construct the networks, each CDS region within a genome was labelled with its taxonomic and functional annotation. These were then compiled into a unweighted and undirected edge list to create the bipartite graph, shown in Figure 5.2. For brevity, I refer to the bottom node set as 'genes' or 'functions' rather than 'gene functional annotations' throughout. However, they represent the functional assignments. In the case of Eggnog, the bottom nodes are associated with labels from multiple annotation systems (e.g. NOG, COG, Pfam and more). For KEGG, they are the unique KO identifiers, and for Subsystems, they represent the 'product' assignments within the

SEED-Subsystems ontology.



$\{u_1\ v_1\}$
$\{u_1\ v_2\}$
$\{u_1\ v_3\}$
$\{u_2\ v_1\}$
$\{u_2\ v_2\}$
$\{u_2\ v_3\}$
$\{u_3\ v_3\}$
$\{u_3\ v_4\}$

(a)

$\{u_1\ \ \{v_1,\ v_2,\ v_3\}\}$
$\{u_2\ \ \{v_1,\ v_2,\ v_3\}\}$
$\{u_3\ \ \{v_3,\ v_4\}\}$

(b)

(c)

Figure 5.2: Building a taxa-function bipartite network. I create hypothetical taxa u_1, u_2 and u_3. Let me define that u_1 has genes with functional annotations v_1, v_2, and v_3. Taxon u_2 has the same annotations as taxon u_1. Taxon u_3 has annotations v_3 and v_4. Three representations of the same graph: (a) edge list format (taxon-function labels from every CDS across all genomes), (b) adjacency list format (every genome and its functional annotation labels), and (c) the bipartite graph itself.

For multilabel sequences, meaning a sequence from a single CDS which was assigned to more than one function, an edge was created between the taxon node and each of the gene functions (Figure 5.3). Eggnog had no multilabel hits, whilst the KEGG and Subsystems networks had 13,168,921 and 297,145 respectively; in the case of KEGG, multilabel sequences could include hits to 3 or even 4 functional tags per sequence.

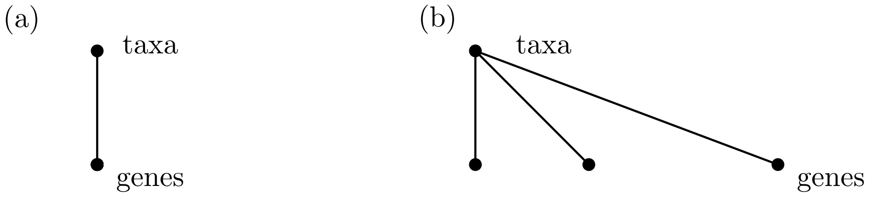


Figure 5.3: The effect of multilabel genes: each plot shows the bipartite graph associated with a single CDS. (a) depicts Eggnog, which has no multilabel sequences; each CDS is associated with a taxon and a single gene. (b) shows an example from KEGG, for which one CDS may create hits to one or more functional labels.

Whilst I obtained the edge weight information, indicative of gene copy numbers, all analyses were undertaken on the binarised (unweighted) network. This is for several reasons: gene copy numbers in prokaryotes are unlikely to be consistent when considering the same species in different locations or at different times [216]; conclusions drawn from an analysis which is sensitive to weighting may therefore be spurious; and results are less likely to be generalisable than the presence-absence case. Furthermore, it is often algorithmically simpler or computationally less intensive to study the unweighted case. With the large dimensionality of the reference and empirical networks, this becomes a reasonable consideration given that there is not biological value added by studying the weighted system. Hence, I introduce an additional representation for the bipartite graph with the $m \times n$ biadjacency matrix B, where m taxa are assigned to rows and n genes to columns. As I am working with presence-absence of genes rather than copy numbers, this is a binary matrix where the presence of an edge between a taxon and a gene is indicated by ones as follows,

$$B = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix},$$

and B shows the same ‘taxa’ and ‘genes’ as shown in Figure 5.2.

In Table 5.4, the basic network statistics for the different annotation systems may be seen. Note that although KEGG covers significantly fewer genes across the taxa than the other annotation systems, there are more links in the network. This is because of the larger proportion of multilabel sequences with KEGG, which results in a denser network despite there being fewer matches from the original gene set. There were also some genomes which returned no functional hits in the subsystems-built network, leading to 27,263 rather than 28,439 taxa for the top node set.

Table 5.4: Basic network statistics for different annotation systems (unweighted network).

Annotation system	Number of taxa	Number of genes	Number of links
Eggnog	28,439	30,208	$41,730,591$
Subsystems	27,263	12,479	$33,506,568$
KEGG	28,439	12,712	$45,203,803$

Figure 5.5 shows the degree distributions are broadly similar for each of the networks. The taxa degree distribution - or number of unique gene annotations per taxon - correlates to the genome size distribution [217]. However, my use of unweighted networks means that the total gene counts are slightly below other values in the literature which have been studied in other contexts [217]. The fact that fewer genes were able

to be annotated with KEGG is offset by the high number of multilabel sequences, which shifts the center of the distribution rightward. The taxa degree distributions are approximately Gaussian. Although the distributions display some spikes, they are unimodal rather than bimodal as has observed within genome size distributions taken from highly redundant databases, implying that the previously observed bimodality may be an artefact of biased databass [218]. However, the gene distributions are heavy tailed. This indicates that the most ‘common’ type of gene is associated with rare proteins found in fewer than 30 taxa within the database, with 27.9%, 23.0% and 26.5% of functions falling into this category for the Eggnog, Subsystems and KEGG networks respectively. Conversely, there are comparatively few functions which are highly prevalent, with 1.2%, 3.5% and 4.5% of functions present in 70% or more of taxa for the Eggnog, Subsystems and KEGG networks respectively. The high degree nodes are for ubiquitous core functions such as those necessary for cellular replication, and they create large hubs with high connectivity within the network. If this gene degree distribution is also observed in the networks of environmental microbiomes, which prior work suggests is possible [192], it may be that graph based methods are particularly suited toward studying their compositions. This is because network-based analyses are able to distinguish ubiquity - or node degree - from importance. I may thus ascertain how the significant diversity contained within less common functions are distributed amongst taxa and begin to elucidate the associations between taxonomic diversity and ecosystem function [186].

5.4 Network structure

5.4.1 Network heatmaps

In Figure 5.6, I present exploratory visualisations of the biadjacency matrices to assess whether there may be large level organisational features that are immediately evident within the networks. These were constructed by sorting the rows (taxa) in phylogenetic order, and sorting the columns (genes) automatically by calculating their pairwise Jaccard distances and creating a dendrogram with Ward clustering. Again, the Eggnog, KEGG and Subsystems based networks are qualitatively similar.

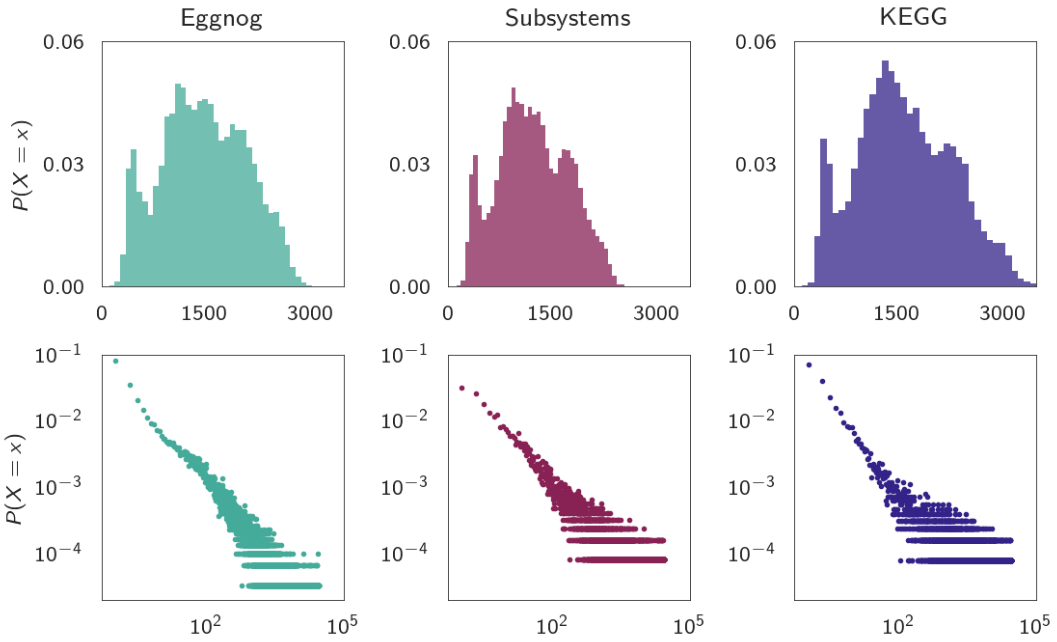
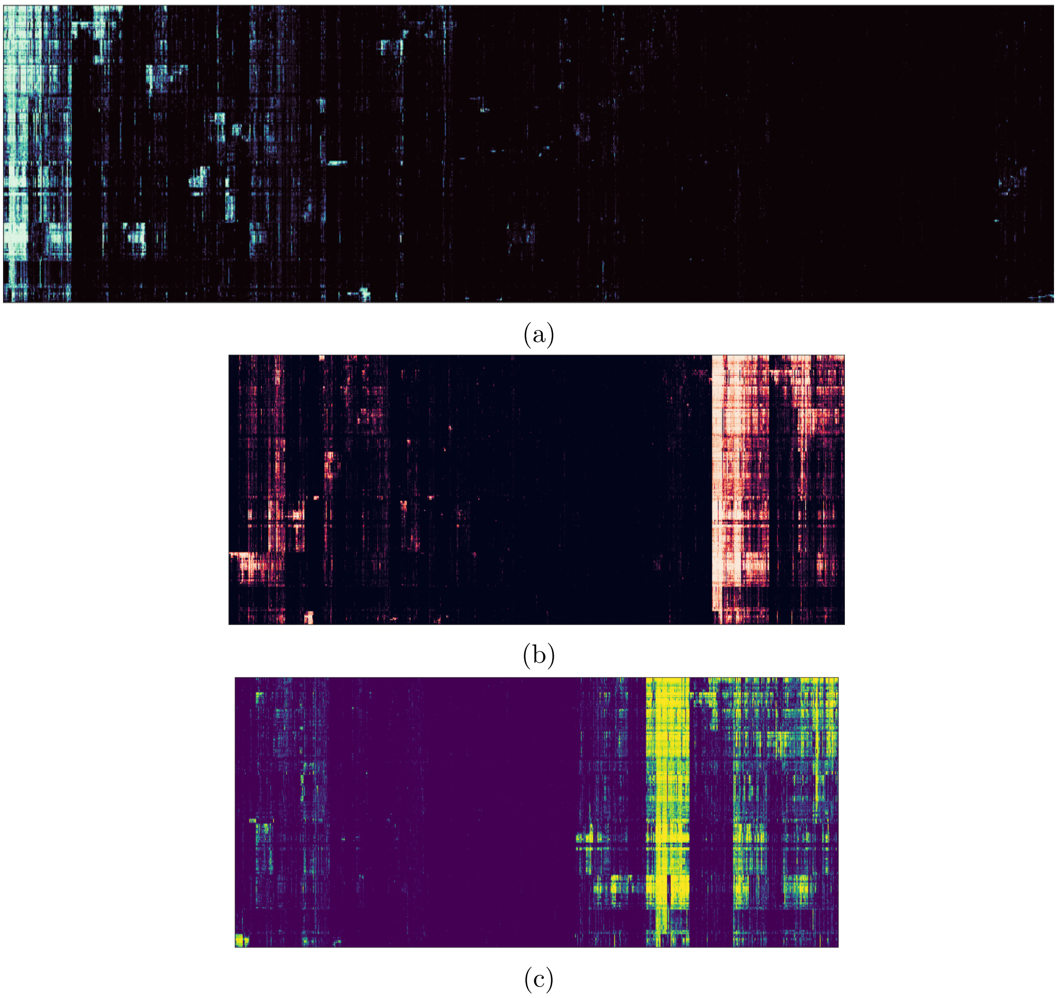


Figure 5.5: Degree distributions for reference network constructed from different annotation systems. Top row shows taxa distributions, bottom row gene distributions.

Blocks of 'common' functions cover approximately 10% of the columns, and there are clear modular elements indicated by the blocks which are only present in certain phyla or groups of taxa. This shows the presence of local structures within the network potentially relating to niche functionality. In turn, this indicates a need for at least one analytic strategy to identify the presence of small scale structures which, in a biological context, indicate a group of redundant functions shared between metabolically similar taxa.

The Eggnog-built network has a larger proportion of rare genes - present in < 30 taxa - than the other annotation systems. It should be noted that my threshold for defining rare (or housekeeping) genes is arbitrary, however altering the values does not qualitatively impact any of the results discussed in this chapter. In Figure 5.7, I break down the Subsystems biadjacency to see whether genes grouped together at the top level of the ontology ('Superclass') appear to have similar proportions of housekeeping or niche genes. Whilst there are differences in the number of genes in each of the superclasses, there are not obvious differences between the categories; all have genes which are ubiquitous, rare, and contain blocks or patches indicative of genes unique to particular phyla or taxonomic group. However, there may be differences at lower



(a)

(b)

(c)

Figure 5.6: Biadjacency matrices for the three different annotation systems. The rows (taxa) are sorted in phylogenetic order, and the columns (genes) were sorted using Ward clustering based on pairwise Jaccard distances between the genes. (a) shows the Eggnog network, (b) shows the Subsystems network, and (c) shows KEGG.

levels of the protein ontology. I next move beyond descriptive methods to examine this question in more detail, with the aim of disentangling the contributions of different gene types to quantify how metabolic niches are distributed through the network and relate to large scale organisation within the system.

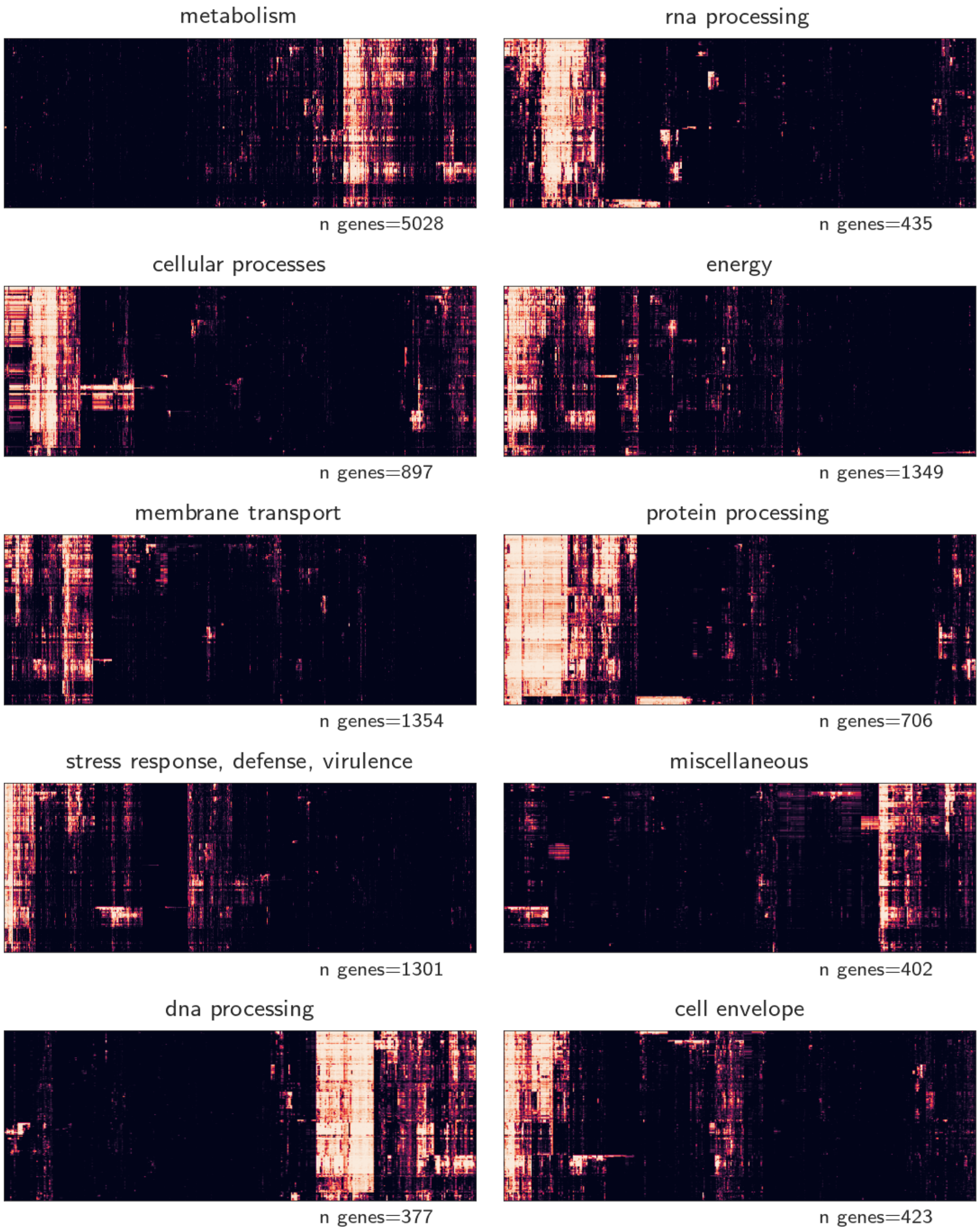


Figure 5.7: Heatmaps of the subsystems biadjacency matrix, broken down by Superclass gene categories.

5.4.2 Spectral properties of the reference network

To interrogate some of the structural features I qualitatively described in the biadjacency heatmaps (Figures 5.6 and 5.7), I examine the spectral properties of the network. This is achieved by applying singular value decomposition (SVD) to the three biadjacency matrices from the different annotation systems. To prevent the housekeeping genes from masking contributions from other rarer genes, I take biadjacency B, transform each column (gene) to have zero mean and unit variance, and then factor the matrix using SVD via

$$B^* = U\Sigma V^T, \tag{5.1}$$

where B^* is the standardised biadjacency, Σ is a diagonal matrix with singular values σ along the diagonal, and U and V contain the left and right singular vectors respectively.

Spectral gaps are evident in the histograms of the first 1,000 singular values of the biadjacency matrices of the Eggnog, KEGG and Subsystems built networks (Figure 5.8). The bulk of the singular values are trending toward zero, but for each of the annotation systems, approximately 15 are significantly larger. This aligns with well-known graph theoretic results used in spectral clustering, where the bulk of the spectrum lies within a semicircle around zero, and the magnitude of one or more eigenvalues associated with topologically important community structures is significantly greater and sits outside of the semicircle [219].

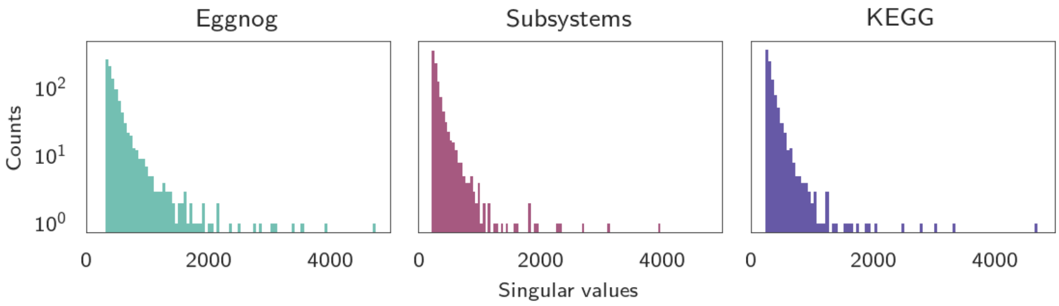
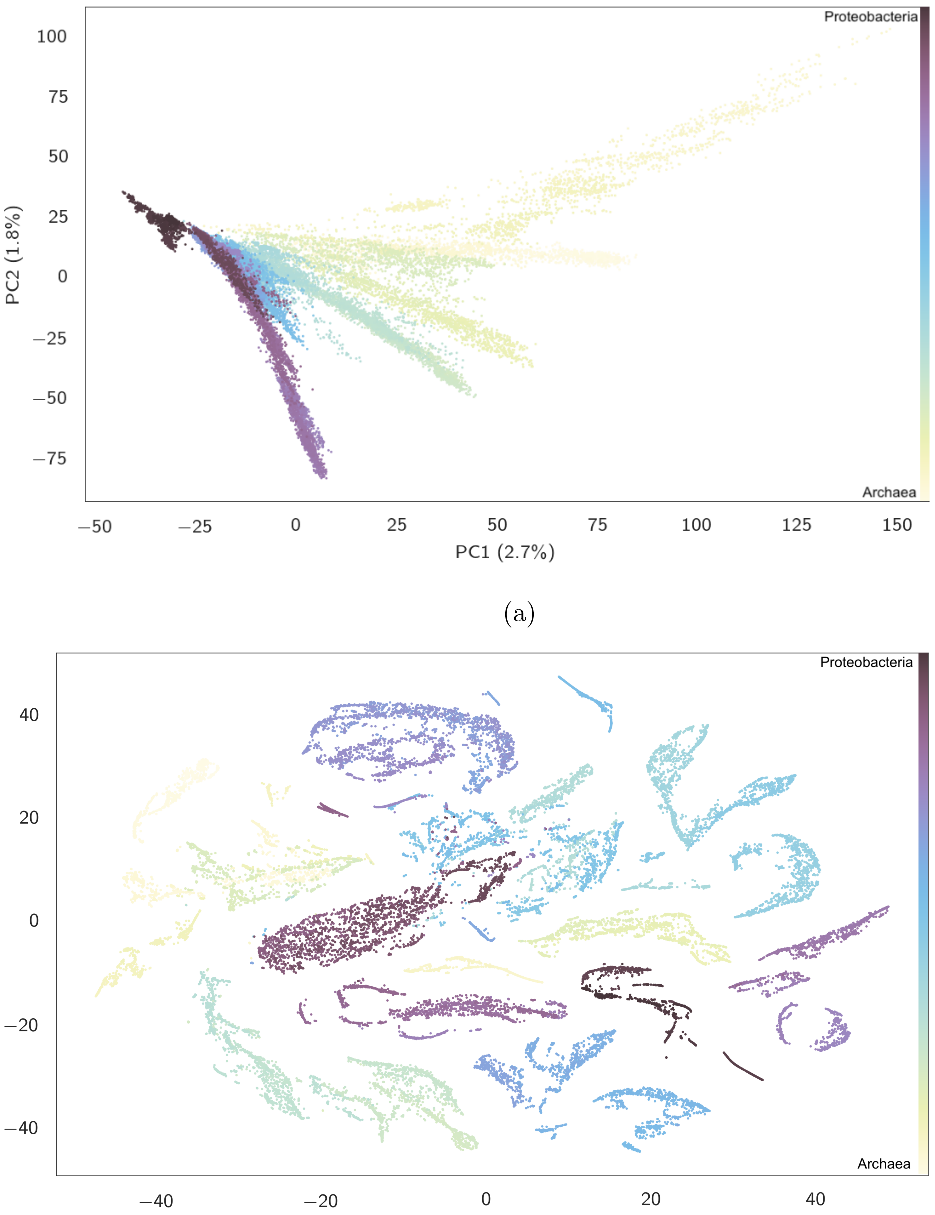


Figure 5.8: The first 1,000 singular values for each network built from different annotation methods.

The singular value distributions imply that principal components (PCs) which are larger than the bulk of the spectra - being the first 15 PCs in this instance - may capture the macro structures in the network I wish to identify. Using the Eggnog network, as it has the best sequence coverage, I test whether there is large level community organisation by doing a PCA on the taxa and plotting the first two components (Figure 5.9a). The grouping of similar colours (and therefore species, see figure caption) indicates that although they capture just 4.5% of the variance, there does appear to be a signal grouping the taxa which correlates to their phylogeny. I next apply tSNE (t-distributed stochastic neighbor embedding) to the first 15 principal components (Figure 5.9b), where clearer clustering and delineation of the different phylogenetic groups may be observed. It is to be emphasised that the goal of this analysis was not to do a dimension reduction, but to test the hypothesis that the top 15 PCs on the genes would generate clearly defined communities as suggested by the singular value distributions; furthermore, whether these may cluster by taxonomy as implied by the modular elements in the heatmaps, or another underlying property of the networks. The near-perfect partitioning across phylogenetic groups indicates that niche or taxa-group specific genes are a critical structural component of the network, despite the fact they make up a relatively small proportion of nodes.

Whilst this is an intuitive result, to the best of my knowledge, this is the first time that large scale metabolic organisation across the prokaryotic phylogenetic tree has been quantified in this fashion. Genome-gene family networks use homology alone, which will track with phylogeny as an immediate result due to its grounding in sequence similarity [200, 206]. Alternately, researchers have used a proxy such as growth rate to partition a community under the assumption it is reflective of overall metabolic similarity between organisms [220]; other approaches in phylogenomics use pairwise sequence similarity between marker genes (not all genes), which are then clustered using dendrograms [221]. These methods are all significantly more involved than matrix decomposition, either from the perspective of experimental investment or computational cost. Furthermore, whilst homology-driven approaches will recover a strong phylogenetic signal, non-homologous genes may share an identical functional annotation [209]. Whilst it is true that homology and function are correlated, as my focus is ecosystem functionality, confirming that - for example - metabolic generalists can clearly be distinguished is an informative result: indeed Figure 5.9b does reveal a very small proportion of taxa sitting in different taxonomic groups to expectation. However, the strong clustering signal overall indicates the presence of well defined communities within the taxa-gene network, captured by relatively few PCs, which is reflective of metabolic organisation in the tree of life mirroring the phylogeny. The Subsystems and KEGG derived networks produce equivalent results - PCA and tSNE plots undertaken on those networks may may be seen in Appendix C.

Next, I assess whether the top principal components are weighted toward specific types of functions, or whether the weights are distributed evenly amongst all functional categories. This allows me to determine which types of genes contribute to the high



(a)

(b)

Figure 5.9: Taxa after clustering on up to 15 PCs of the biadjacency matrix. Colour is set using continuous colourmap applied to all 28,439 taxa in phylogenetic order; that is, closely related taxa will have similar colours. (a)PCA applied to the taxa. (b) tSNE applied to the taxa's first 15 principal components, using a perplexity value of 80.

level organisation in the network, and thus drive the large scale variance within the prokaryotic tree of life. The lack of a hierarchy in Eggnog means I am unable to group the genes in that network in any way. The KEGG hierarchy has 4 labels for the first level and >300 labels for the second. These are few and too many categories to form useful summaries respectively; with 4 categories, there is not enough definition to separate the functional groups, but with >300 there are too many to identify macro patterns in the data. This issue is compounded by a highly uneven distribution of genes between the labels, meaning that it would be difficult to interpret results from the KEGG-derived network. However, the Subsytems network is ideal to use in this setting. The Class level of the ontology has 33 labels, allowing the 12,479 genes to be grouped into these higher level bins for assessment.

Figure 5.10 shows the distribution of the scores within the eigenvectors of the first 15 principal components. As the x and y axes have the same limits, it is possible to see that whilst each eigenvector contains a Laplacian-type distribution around the centre (at 0), there are still substantial differences between their distributions. For example, the eigenvectors for PC3 and PC13 show skewness in the opposite direction to those of PC9 and PC10. The distribution associated with PC4 has two groups situated on either side of a central peak, a characteristic feature of stochastic block models. Similar, albeit of smaller magnitude, signals may be seen in the eigenvectors of most of the top PCs. The tails and gaps in these histograms are consistent with the features I would expect in the distributions given the clear clustering within Figure 5.9. To interrogate this further and identify which types of genes from each PC drive the large scale network structure, I extracted the highest scoring genes (top 100 and bottom 100) from eigenvectors 1-8; these were aggregated by class and are plotted in Figures 5.11 and 5.12.

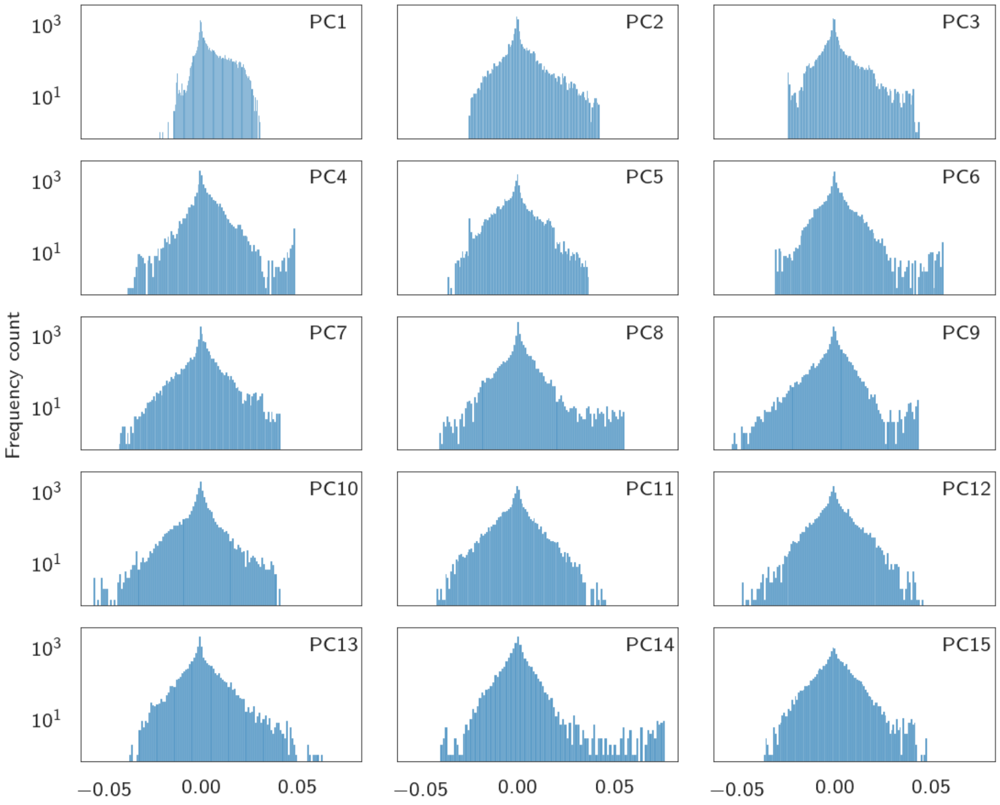


Figure 5.10: Distribution of component scores from the first 15 principal components (subsystems network).

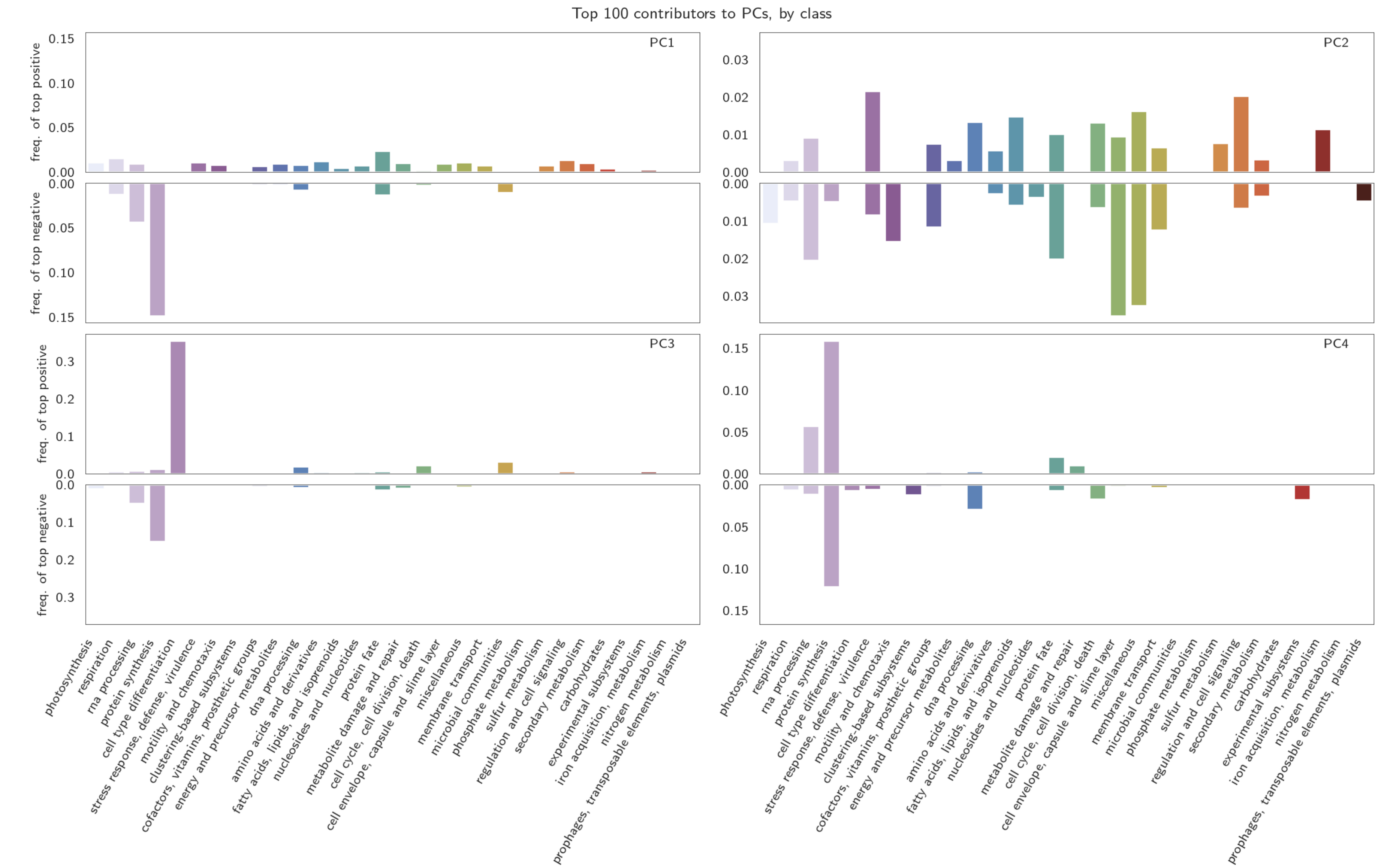


Figure 5.11: Top contributors (by class) taken from eigenvectors 1-4 in the subsystems network. Barplots show the frequency the top 100 positive and top 100 negative (by magnitude) weights appear in certain classes. Frequencies are normalised against their abundances in the full networks.

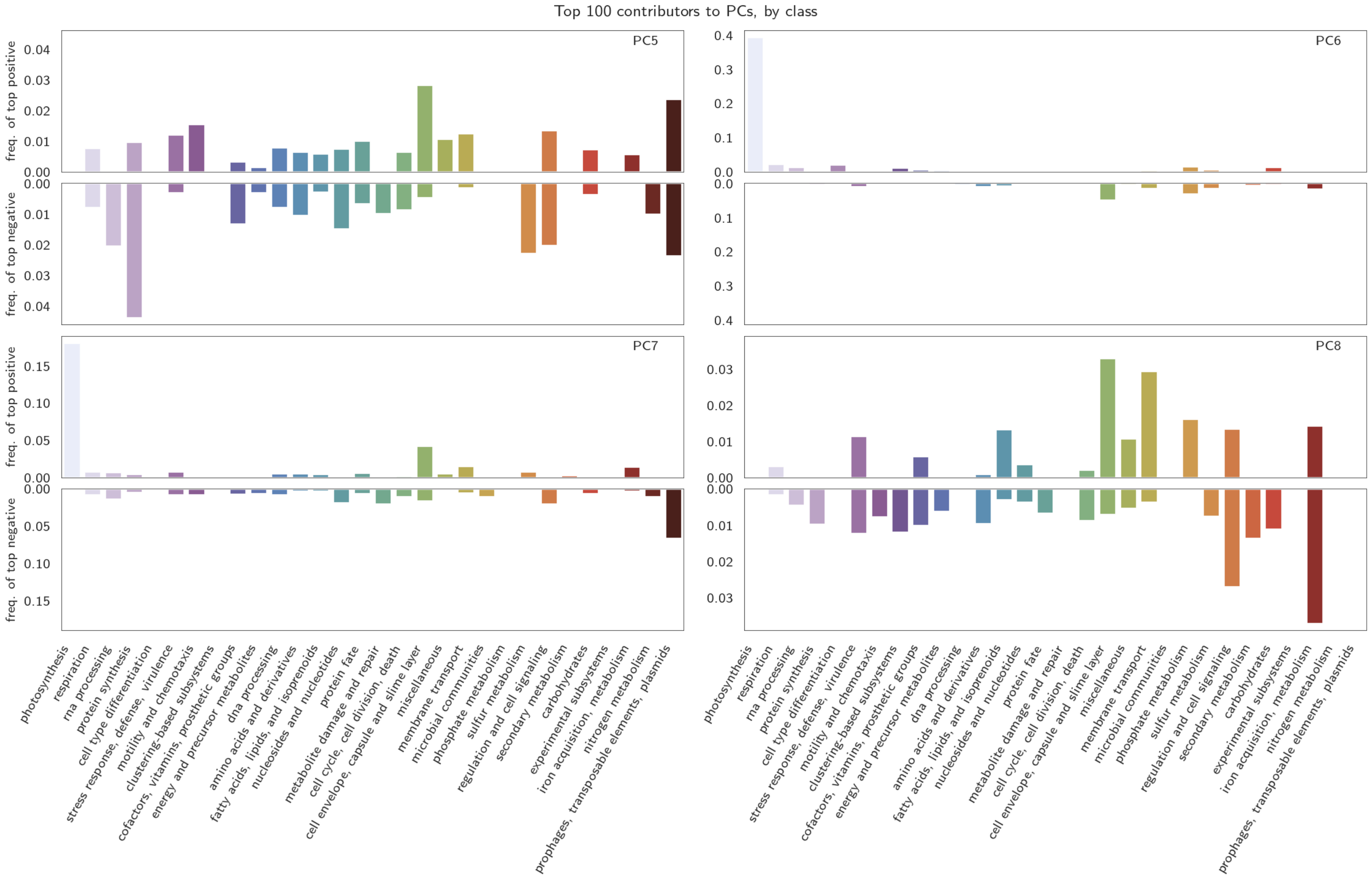


Figure 5.12: Top contributors (by class) taken from eigenvectors 5-8 in the subsystems network. Barplots show the frequency the top 100 positive and top 100 negative (by magnitude) weights appear in certain classes. Frequencies are normalised against their abundances in the full networks.

The breakdown of these eigenvectors (Figures 5.11, 5.12) shows that certain groups are heavily over- (or under-) represented within the top scoring genes[2]. PC1's eigenvector is heavily over-represented by protein synthesis, which encompasses the universal functions associated with ribosomes, tRNA synethases, and translation. However, there are low quantities of almost all other classes associated with PC1. To check whether PC1 would not only be weighted towards protein synthesis, but also the most abundant (and housekeeping) genes, I constructed a binary vector **a**, where each gene j was assigned 1 or 0 according to the following rule:

$$a_j = \begin{cases} 1, & \text{if } d_j > 10905 \\ 0, & \text{otherwise,} \end{cases} \tag{5.2}$$

and d_j is the degree of that gene's node. By taking the dot product of **a** with the top principal components, I may determine which PC is associated with the housekeeping or highly abundant genes. Irrespective of the (reasonable) threshold I assigned to define an abundant function - beginning at 10905 in Equation 5.2 corresponding to a saturation of 40% and rising to saturation thresholds of 70% - PC1 was the top scoring component (Table 5.13).

Table 5.13: Dot product between a binary vector indicating the presence of a 'housekeeping' or highly abundant gene and the top 8 eigenvectors, under different thresholds for defining a housekeeping gene (total saturation amongst taxa). Higher values indicate that the eigenvector associated with that principal component is more strongly weighted towards the abundant genes.

Saturation Threshold	40%	50%	60%	70%
PC1	15.16	11.24	8.32	6.11
PC2	8.55	5.93	4.25	2.84
PC3	10.14	7.99	6.38	5.05
PC4	8.85	7.14	6.07	5.12
PC5	8.27	6.30	4.76	3.39
PC6	8.41	6.60	5.08	3.66
PC7	5.10	3.49	2.33	1.37
PC8	4.15	2.67	1.77	1.03

The vector for PC2 displays a larger mix of strongly contributing classes, yet those for PC3 and PC4 also both had an over-representation of protein synthesis. Positive weights for cell type differentiation - predominantly sporulation related genes - and negative weights on protein synthesis were evident in the loadings for PC4. The niche-

[2] I note for the reader that the classes in Figures 5.11 and 5.12 have been coloured and sorted by their average entropy, which I cover in the next section.

specific genes associated with photosynthesis were strong contributors to PC6 and PC7 despite being comparatively rare (0.8%). I note that ubiquitous functions such as those involved in membrane transport (11.1% of functions) were not strong contributors to the top principal components. Indeed, the top contributors to the PCs were not correlated to the prevalence of those functional classes in the network, nor to the density of the genes amongst the taxa, indicating that certain types of genes contain more information about taxonomy than others. This is not surprising from a biological standpoint, but the fact that such a large amount of information is contained in comparatively few PCs, and that certain functional groups are heavily over represented in the top principal components suggest that critical elements of network organisation may be dominated by a small minority of functions.

To assess which gene classes were driving the community partitions seen in Figure 5.9b's tSNE, I performed k-means clustering on all of the genes within the top 15 PCs (Figure 5.14). Whilst the bar plots in Figures 5.11 and 5.12 provide a general overview of how the first 8 eigenvectors weight the genes in order to partition taxa along its own axis, applying k-means reveals how the linear combinations of these vectors are realised in the full vector space. By clustering the matrix formed by the first 15 vectors of V, it is possible to determine which genes are located together in the embedding space, informing me which gene categories drive the clustering effects seen in the PCA and tSNE plots in Figure 5.9. This consequently reveals which elements of functional diversity are the strongest contributors to taxonomic community structure in the network, and may reveal elements in the reference network which would drive environmental filtering in community assembly processes [194]. These results are shown in Figure 5.14.

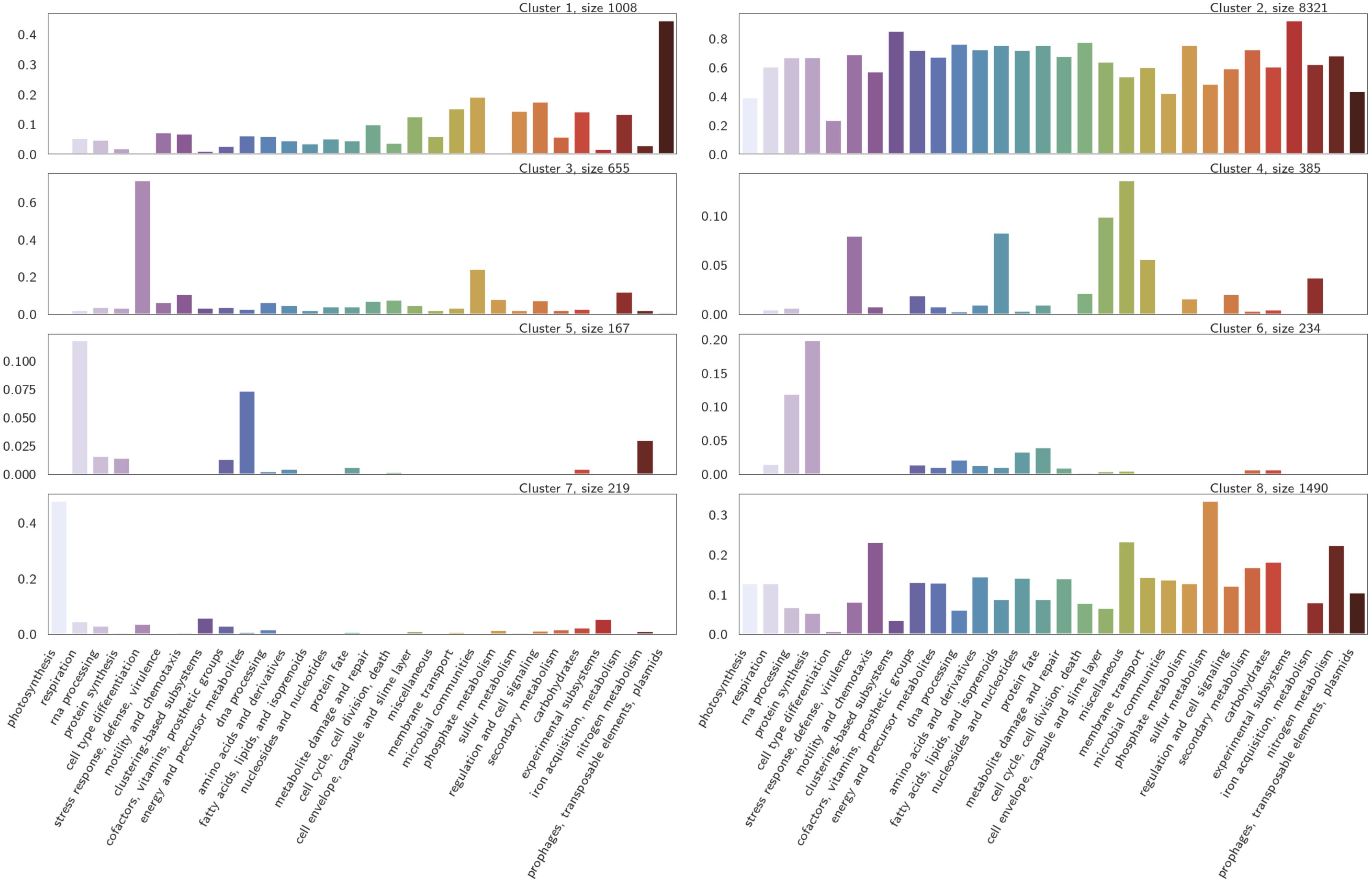


Figure 5.14: Frequency (normalised against the frequency in the network) of different classes in clusters, which were automatically defined with k-means on the top 15 PCs. Cluster order (e.g. cluster 1, 2) is arbitrary.

It is evident that the majority of genes are grouped within a large cluster (Cluster 2, with 8,321 genes), which is also relatively uniformly distributed across the different gene categories. Similar to the patterns seen in the top gene weights from the eigenvectors, photosynthesis (Cluster 7), respiration (Cluster 5), protein synthesis (Cluster 6), RNA processing (Cluster 6), and cell type differentiation (Cluster 3) are dominant. However, from this analysis it is evident there is also a strong signal resulting from from the class of prophages, plasmids and transposable elements (Cluster 1). It appears that a small number of classes - just 6 to 8 out of 33 - are key to the separation of taxa in the embedding space. Furthermore, these categories are over-represented in isolation within half of the clusters rather than being present in combination with the other dominant classes. This indicates that suites of genes involved in highly specific functional processes, such as photosynthesis, play a key role in splitting the taxonomic groups (see Clusters 3, 5, 6, and 7) as well as more complex combinations of different classes (see Clusters 1,4 and 8).

Next, I examine whether it is the common or rare genes which are driving these effects. This is achieved by constructing 8 binary vectors of length $12,479$, $\{\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_8}\}$, to represent the presence (or absence) of genes in each of the 8 clusters. For example, if a gene belongs to Cluster 1, the value at that gene's index would be 1 within the first vector, and 0 in the other 7 vectors. Generalising this principle, I may assign entry v_j for gene j in vector $\mathbf{v_n}$ as follows:

$$v_j = \begin{cases} 1, & \text{if } j \in \text{cluster } n \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}$$

Note that k-means places each gene in exactly one cluster, meaning that the non-zero indices from each $\mathbf{v_n}$ form disjoint sets. Using these vectors, I may assess which clusters the ubiquitous genes fall into by taking their dot products with another vector $\mathbf{a}$ which encodes the abundant genes. I create this vector by assigning a value of 1 for all genes j which are present in 10905 ($> 40\%$) of taxa, and 0 for those which are not, so

$$a_j = \begin{cases} 1, & \text{if } d_j > 10,905 \\ 0, & \text{otherwise,} \end{cases} \tag{5.4}$$

where d_j is the degree of gene j. I may also use the same method to determine which clusters the very rare genes (present in < 30 taxa) were assigned to. In this case, I construct vector $\mathbf{r}$ by applying my binarising threshold to have a value of 1 for genes with a degree of less than 30, and 0 otherwise,

$$r_j = \begin{cases} 1, & \text{if } d_j < 30 \\ 0, & \text{otherwise.} \end{cases} \tag{5.5}$$

These results may be seen in Table 5.15.

Table 5.15: Ubiquity and rarity of genes in clusters. Common genes are defined as being present in 10,905 ($> 40\%$) of species, and rare genes present in < 30 species.

Cluster	1	2	3	4	5	6	7	8
Common	0	778	3	0	0	0	0	204
Rare	36	2780	5	28	14	0	0	2

The largest cluster, 2, contained the greatest number of both the rarest and most common genes; even when accounting for the size of the cluster, over a third of the genes have an intersection with $\mathbf{r}$ or $\mathbf{a}$, a far larger proportion than for the other clusters. Nearly all other remaining genes from the abundant category were found in Cluster 8 (the second largest), which also displayed a more even representation and distribution of classes than the other clusters. Crucially, Cluster 2 is the group centred closest to zero, meaning that that abundant or rare genes do not strongly contribute to the distinguishing features of the taxa in the embedding space. It is also possible to interpret Table 5.15 in conjunction with the histograms in Figure 5.10. The peaks which are situated around zero contain the bulk of the abundant and rare genes (seen in Cluster 2, and to some extent Cluster 8). Next, I consider the left- and rightmost extremes of the distributions, especially where there is a high variance, or there are gaps between the main distribution and small groups of genes; these contain the biggest contributors to clusters 3 through to 7 and drive the separation of taxa in the embedded space. This is not only biologically reasonable - after all, if a gene is present everywhere, it is not taxonomically informative - but also indicates that the low to medium saturation genes are partitioning the taxa. This result may be of interest to those researching gene-gene family networks, their projections, or their similarity matrices, to assist with feature selection to resolve challenges with combinatorics and thresholding [201, 204]. Similarly, the fact that the taxonomic signal was maximised by strong weights on a small proportion of functions may assist in simplifying statistical or modelling methods in microbial ecology by reducing the number of variables necessary to capture the majority of system variance [147, 190]. Without the paired taxa-function distribution, this cannot be quantified without relying on simulation [194]. Finally, this result implies that mean-field network metrics, such as nestedness, may need to be treated with caution in microbial taxa-function networks, as these low or medium saturation functions would make only a small contribution to the statistic and it may miss key biological detail [186, 222]. I now quantify whether these 'informative' (or other) functional categories are grouped tightly within phylogenetic bins, scattered across the complete graph, or a combination of both.

5.4.3 Node Entropy

A prevailing challenge in unravelling taxa-function relationships is that the same level of functional redundancy in a natural system may arise from contributions from closely or distantly related microbes [194, 223]. Network methods are ideally suited to resolve these difficulties as they are able to define abundance (degree) as well as network distribution (relatedness) [204]. To examine how strongly different types of genes are associated with specific communities within the network, I calculate their entropy. Entropy on graphs is classically defined to measure the overall complexity of the network, and is notoriously challenging to implement for large networks as it is an NP-hard problem [224]. In this setting, my focus is not to find the overall complexity measure for the network itself, but to establish whether certain nodes (genes) are decentralised and non-taxon specific, or are strongly associated with a community. Local entropy measures defined for vertices often assign the probability distribution based on node degree [224]. The difficulty with that approach is that unless the node degree is very high, it does not indicate how widely a particular function is distributed amongst the taxa. To resolve this issue in the context of my phylogeny, I define a form of Shannon entropy for the nodes of bipartite graph B, where the entropy of a node v_j in the bottom (gene) set is given by

$$S_j = -\sum_i p_i \log(p_i). \tag{5.6}$$

In Equation 5.6, if I consider node v_j as a column from the biadjacency matrix, each p_i is calculated within a bin (a series of rows) defined by the GTDB phylogeny. I may then calculate entropy of a gene at different phylogenetic levels, from phylum down to order (at the genus level there are too few taxa per group for the metric to be meaningful).

To illustrate the method, I have a column vector representing a hypothetical gene in the biadjacency matrix,

$$[\,1,0,1,1,0,0,1,0,1,1,0,0,1\,]^T,$$

with 3 phylogenetic bins assigned as

$$[\,1,0,1 \mid 1,0,0,1 \mid 0,1,1,0,0,1\,]^T. \tag{5.7}$$

I may apply (5.6) to (5.7) by taking the sum of the counts in each bin and converting to probabilities:

$$[\,{}^2\!/\!{}_7 \mid {}^2\!/\!{}_7 \mid {}^3\!/\!{}_7\,]^T. \tag{5.8}$$

The gene entropy in this case would be $S = -{}^2\!/\!{}_7 \cdot \log({}^2\!/\!{}_7) - {}^2\!/\!{}_7 \cdot \log({}^2\!/\!{}_7) - {}^3\!/\!{}_7 \cdot \log({}^3\!/\!{}_7)$.

Before applying Equation 5.6 to the network, there are several considerations to take into account. Firstly, I exclude genes which are present in $> 70\%$ of the taxa, and set a minimum degree of 8 for the genes. This is because genes present in almost all taxa or extremely rare genes are not informative regarding community structure, having very high and very low entropy respectively by default. For each gene remaining, I calculate a randomised entropy S_r by shuffling the gene rows and normalise the entropy S against its random baseline: $\dot{S} = S/S_r$. This allows me to establish the deviations from a random network; that is, which classes are more tightly clustered within a community than would be expected by chance. It is also necessary to correct for the gene degree - higher degree would bias the scores upward. One strategy to manage that bias may be to apply a subsampling bootstrap where I select n rows per gene. This would mitigate the discrepancy between high saturation and low saturation genes. However, this process was extremely computationally intensive and required a minimum degree threshold that was higher than desirable. I found that applying a weighted mean per class performed equivalently to a subsampling boostrap whilst being orders of magnitude faster and less restricted by minimum thresholds. The weighted mean is given by

$$\bar{S} = \frac{\frac{\dot{S}_1}{d_1} + \frac{\dot{S}_2}{d_2} + ... + \frac{\dot{S}_n}{d_n}}{\frac{1}{d_1} + \frac{1}{d_2} + ... + \frac{1}{d_n}}, \tag{5.9}$$

where d_j is the degree of gene node j, and the n genes are aggregated by their ontology labels.

The final correction I made was to account for the imbalances in the size of the phylogenetic bins. For example, the largest phylum is Proteobacteria, with 8,882 species; however there are multiple phyla with fewer than 5 species. This bias needs to be addressed, otherwise genes falling within Proteobacteria will always have a larger discrepancy from the random baseline and skew the results. To resolve this issue, I use a subsampling bootstrap. Taxonomic bins with fewer than 3 species were excluded from the analysis, and I randomly sampled 3 taxa, or rows of the biadjacency, from each phylogenetic bin for the bootstrap. I then calculated the average entropy per class for 1000 bootstrapped networks, producing a distribution of means.

To confirm that the entropy scores were not trivially correlated to the node degrees I plot the entropy score against the node degree (Figure 5.16). Although an upper and lower bound of values is apparent, corresponding to the minimum and maximum possible number of values falling within bins across the network, there is a wide spread of datapoints within the envelope.

A boxplot of the entropy distributions at the order level may be seen in Figure 5.17 (refer to Appendix D for plots at different phylogenetic levels). Consistent with prior work, I observe the lowest entropy ratios (highest specificity) in niche related genes such as photosynthesis and photosynthesis-related genes (respiration), as well as protein synthesis (encompassing ribosomes[3]). In addition, the highest entropy ratios - which fall close to randomly distributed - are seen in genes associated with ‘experimental subsystems’ (CRISPR), as well as transposable elements, phages and plasmids.

[3] Note that species specificity at the lowest functional level does not preclude the class level of the ontology from being ubiquitous.

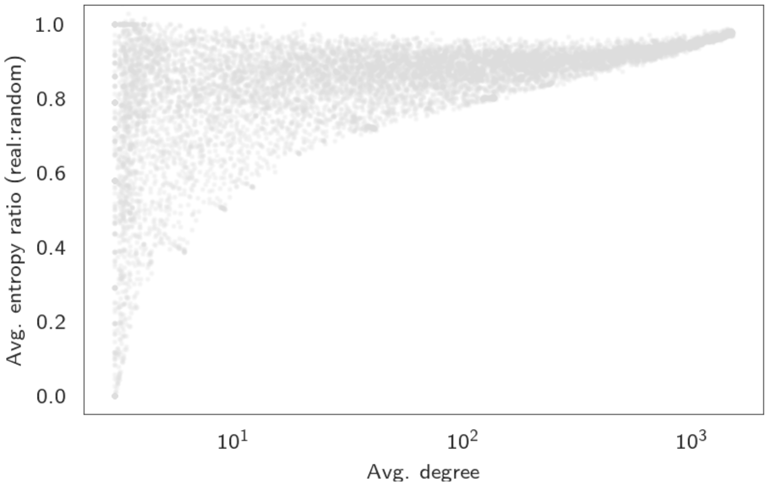


Figure 5.16: Entropy vs. degree for 500 bootstrapped networks. The average entropy versus the average number of node degrees for each taxa. The distribution indicates that there is a minimal possible entropy ratio close to the upper limit of 1. The hyperbolic lower envelope of the points shows the function for the lower limit of the relationship, i.e. how low the entropy ratio can go as a degrees increase. The concentration of points toward the upper right corner indicates that saturation of a function is so high that shuffling has negligible impact. If the entropy ratio was a simple correlation to node degree the points would be distributed along the bottom portion of the triangle.

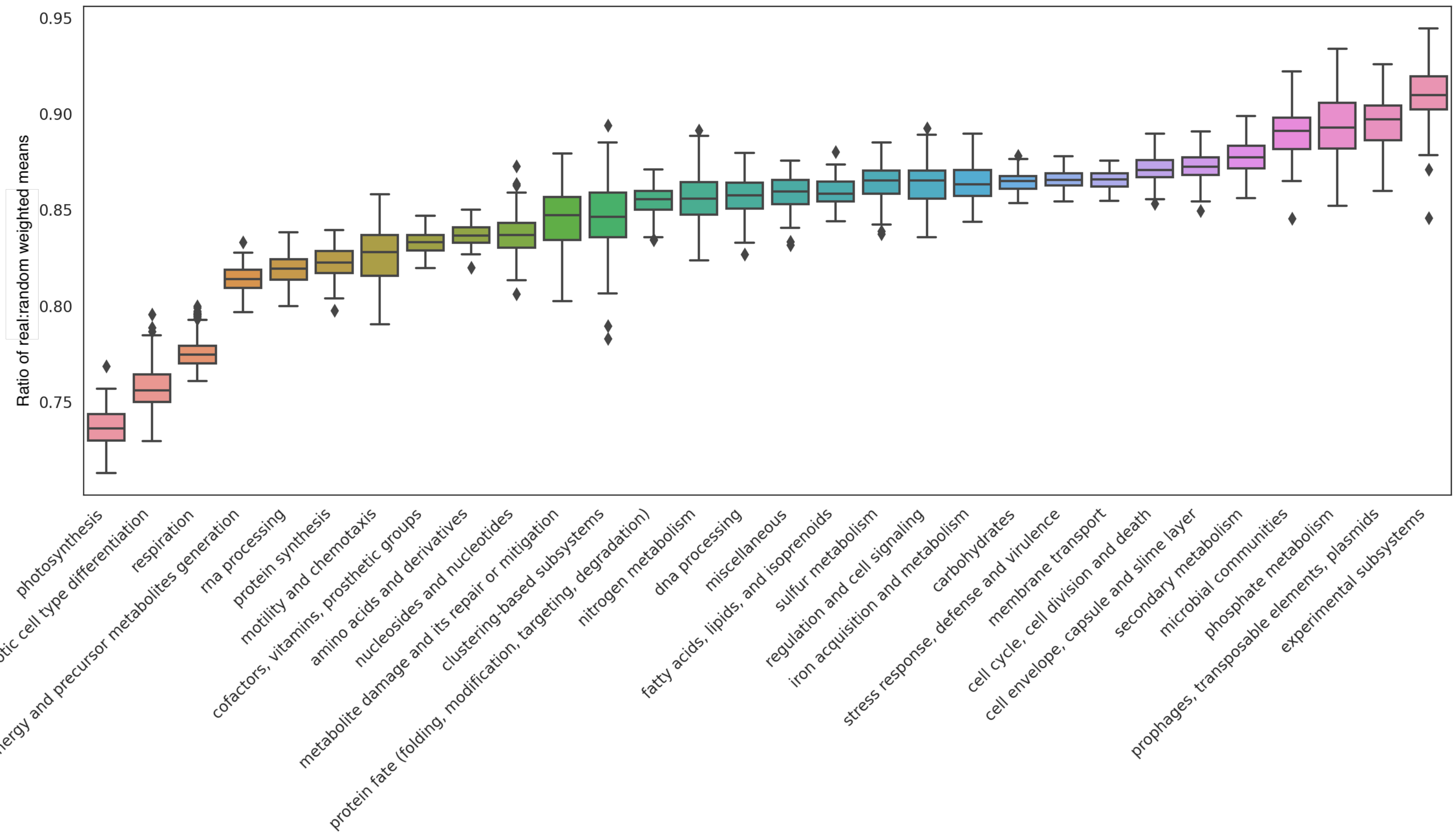


Figure 5.17: Mean entropy score, relative to random baseline, per subsystems Class at the order level. The $y-$axis shows the ratio of the real to randomised weighted mean, as given in Equation 5.9. The boxplots are generated based on values from 1000 bootstrapped networks.

My entropy results reproduce previous findings from phylogenetic-based approaches. The groups of genes with the lowest entropy scores, in network terms considered as strongly modular and in biological terms niche associated, match those identified as having low homoplasy (or ‘phylogenetic patchiness’) in [209], such as photosynthesis and ribsomes. This is further supported by the general correlation of low entropy genes with those which play the biggest role in separating the taxa in the embedding space in the top PCs. Conversely, high homoplasy results were seen in [209] for viral (phage) proteins for which I observed extremely high entropy, and almost no contributions to the ‘informative’ taxa in the top principal components of the network. The other high entropy gene groups, including CRISPR, plasmids, and secondary metabolism, are associated with HGT, and the sporadic distribution of them throughout the phylogenetic tree is attributed to them being shared through lateral transfer processes rather than being vertically transmitted [169, 200, 209]. These results, in conjunction with the SVD analysis, reveal that large scale metabolic organisation across the prokaryotic tree of life occurs through a relatively small proportion of low saturation, strongly niche associated functions.

5.5 Conclusions

By creating a binary network encoding of the prokaryotic tree of life, I provide a representation of the system which is less information-dense than nucleic or amino acid sequences. This allows me to include all genes, rather than marker genes alone, to examine the distribution of metabolism across taxa [190, 221]. My results show that a significant amount of biological detail is recoverable and it is possible to capture the large scale distributions of how functions are shared amongst taxa. Whilst some spectral methods, such as PCA, are routine in microbial ecology, to the best of my knowledge, they have never been utilised in the study of bipartite networks to interrogate their structures. In assessing the properties of the first 15 eigenvectors in the context of a gene ontology, it was surprising to find that individual functional classes were so clearly delineated in the embedding space. In the top 15 eigenvectors, 8,321 genes - equivalent to 66.7% - were close to zero in the vector space, and most of the remainder clustered specifically within a single Class in their PC within the gene ontology. This indicates that the taxonomic signal was maximised by strong weights on a small proportion of genes; building upon observations from prior ecological research, the result highlights the importance of metabolic niches in the graph structure [186].

Network-based methods examining gene sharing between organisms - whether based on a bipartite or sequence similarity alone - has generally used descriptive or similarity based methods, centrality statistics or community detection algorithms [200, 206, 207]. However, I found it was possible to reproduce results from the largest phylogeny-metabolism study to date through a combination of linear algebra and Shannon entropy alone [209]. The large scale structure in the network, which created a strong phylogenetic clustering signal, had minimal contribution from HGT associated groups; rather, the contributions appeared to be from functional categories which would involve suites of genes contributing to ecological (or phylogenetic) niche associated metabolic pathways. Previous network-based methods to identify HGT genes have either relied on constructing the network with both prokaryotic genomes and mobile element genomes explicitly and then measuring overlap or finding communities [205, 225, 226], or using an observational approach in combination with bioinformatics-based validation [207]. It is encouraging that my gene-taxa network topology alone is able to reproduce the results, although I provide some caveats to how generalisable this approach may be. Firstly, as far as I am aware, this network is the largest of this type ever assembled across extremely diverse genomes, allowing me to create a large number of bins for the calculations; it may otherwise be difficult obtain a random baseline without modifying the method. Secondly, I was not aiming to identify the likelihood an extremely rare gene may be associated with HGT; my binning strategy relies on a gene being abundant enough for its presence or absence in a bin to provide a meaningful signal. Finally, whilst it would not necessarily preclude replication, my use of a dereplicated taxonomy with significant numbers of genomes allowed for an information-rich bootstrap which I caution may otherwise simply reinforce whichever biases are present in a starting database. Despite these caveats, the entropy and SVD results indicate that different gene types display fundamentally different distributions throughout the graph and thus the tree of life, highlighting that it is unlikely to be reasonable to treat each node in taxa-function networks interchangeably.

In summary, through interrogation of a paired taxa-function network distribution, I have uncovered key topological properties of the prokaryotic tree of life. These include a heavy tailed functional degree distribution combined with local structures generated by suites of niche-associated genes, which in turn drive critical large scale organisation in the system. Functions which are highly diffuse and spread throughout the graph are found to be uninformative with respect to taxonomy, but are identified as belonging to HGT-associated processes. These results highlight the utility of network methods in examining paired taxa-function relationships in microbial systems, and show that there is - for the first time - quantified taxonomic organisation with respect to ecosystem function for the most diverse living organisms in the biosphere.

Chapter 6

A universal functional topology in microbial systems

A fundamental challenge facing microbial ecology is unraveling the assembly processes and mechanisms which drive community structure and stability. Here, I introduce a novel network-based framework to examine the shared distribution of functions amongst taxa, allowing the investigation of the redundancy properties of different metabolic processes in microbial communities. I demonstrate that there is a universal taxa-function structure across real-world microbiomes which would facilitate horizontal gene transfer and thus strengthen community stability and resilience. My findings provide new insight into the relationship between taxonomic diversity and ecosystem function through a novel quantification of redundancy structures within microbial systems.

6.1 Motivation

Microbiomes are highly dynamic and undergo continual species turnover, yet at a community level, there are highly conserved functional profiles [123, 148]. Although there is broad consensus that this stability is likely indicative of universal assembly rules, mechanisms have remained elusive, in part due to a lack of an analysis framework that explicitly links taxonomic and functional profiles [123, 153, 227]. The high diversity of microbiomes means that organisms are subject to continual competitive pressure, and how functionality is shared across the community is likely to play a profound role for species and community level success and survival. For individual microbes, a smaller cell size frequently confers a fitness advantage, with decreased costs of cellular replication leading to increased competitive success [163]; the active reduction of a microorganism's cellular size achieved through the pruning of extraneous genes and metabolic pathways is termed streamlining. The low-cost replication benefits of streamlining are offset by the need to mitigate against transient stressors, ranging from phage attack, allelopathic warfare, anthropogenically induced antibiotic pressure, or fluctuations in the physical or chemical environment such as temperature or pH [138, 228]. Thus, at a community level, tradeoffs between retaining survival-oriented metabolic potential and reducing cell size could theoretically be optimised through exploiting the plastic nature of prokaryotic genomes. Horizontal gene transfer (HGT) confers the ability for microbes to share DNA, and typically occurs between closely related organisms or those within an ecological niche [170–172]. In the event of environmental disruption, the community floods the environment with extracellular DNA, allowing organisms lacking a particular stress response gene to source it from peers and integrate it into their genomes, increasing the likelihood of survival and stabilising the community [168]. Indeed, a function may not need to be highly redundant for the community to be resilient against an associated stressor: provided it is present across a suitably diverse cross section of the community, HGT provides the potential for a gene to be readily accessible should it be required. Here, through analysing sequencing data from natural communities with network-based methods, I uncover the taxa-function landscape and reveal quantitative differences between the distributions of functions associated with metabolic or ecological niches and those with genome editing, phages and extracellular DNA.

6.2 Data methods

6.2.1 Data sourcing

To investigate joint taxa-function distributions in real-world microbial communities, I sourced shotgun WGS data from 5 of the earth's major biomes, including three free-living environments, encompassing soil, open water marine, and freshwater, and two host-associated biomes with data from the human gut and rhizosphere. I sourced

20-80 samples per biome, distributed around the globe, to ensure a spread of data inclusive of global diversity patterns. The data were downloaded from NCBI's sequence read archive (SRA) [229]. Where possible, I used sequencing data from high quality consortium studies, including Tara Oceans [148], the human microbiome project [230], and global soil surveys [177, 185]. Otherwise, I used NCBI project summary abstracts to determine the environmental source of the data. I chose samples with more than 1 million raw paired reads and downloaded the data using fastq-dump.2.9.2, discarding technical and the first 10,000 reads for quality control reasons. A maximum of 5 million reads per sample were taken to keep a consistent sequencing depth. A summary of the sample data may be seen in Table 6.3, and a map showing the locations of the samples and their associated biomes may be seen in Figure 6.1. A list of SRA Accession IDs and their associated biomes may be seen in Appendix E.

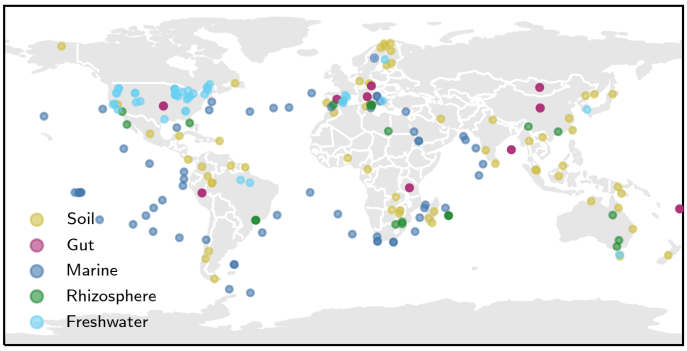


Figure 6.1: Global map showing locations of samples and their associated biomes.

6.2.2 Bioinformatics and network construction

As within Chapter 5, a network representation of the community is used to study the taxa-function structure of microbiomes. However, constructing a bipartite network from shotgun WGS data poses bioinformatics challenges. As soon as sequences are assembled for functional annotation, they require alignment against some form of reference database for classification. If certain species or clades are over-represented in the chosen database, there is a chance for multiple misassignments to closely related species which would in turn bias the network statistics. It is unfortunately not possible to completely eliminate these biases. However, they can be mitigated by the use of a dereplicated database such as the taxonomically balanced GTDB reference genome set. This ensures that any biases present will be minimised, and scale consistently and proportionally with diversity patterns in the samples, allowing a fair comparison of results across biomes when searching for common patterns [119, 209].

To create per-read taxa and function labels for the shotgun WGS data, I used the software *k*-SLAM, which permits the use of a custom reference genome database, for which I use GTDB [119, 231]. To place the data in a *k*-SLAM compatible format, I created custom `names.dmp` and `nodes.dmp` files, equivalent to those found in NCBI's genbank summaries, to pass taxa IDs and phylogeny from the GTDB taxonomy to *k*-SLAM [119, 231]. It was not possible to use NCBI taxa accessions as some assignments differed from GTDB [119]; for example, taxa designated as two species in GTDB may have been categorised as a single species within the NCBI taxonomy or vice-versa. Finally, I provided the genbank (.gbk) format files generated by Prokka as described in Section 5.2 [212], along with custom taxonomy files, to *k*-SLAM's database build function, which creates a serialised database to use for WGS sample classification.

For sequence classification, *k*-SLAM uses a heuristic to decrease the computational cost of sequence assembly. It firstly uses *k*-mers for lowest common ancestor taxonomic assignment. Because this step bins the reads, assembly occurs within a smaller search space and therefore at reduced computational expense. These assembled reads may then be then aligned. Due to the assembly, it is possible to get accurate species-level assignments for a significantly higher portion of the data than would usually be possible using *k*-mer based methods [231]. Furthermore, because the alignment data to gene loci is returned, it is possible to map the assembled reads back to their functional annotations, giving me the taxonomic and functional labels for each read, which in turn form the basis for network construction. This moves beyond previous research by explicitly pairing the taxa and functional data in the sample instead of inferring genes and functions based on taxonomy alone, or being constrained to examining the most abundant taxa and functions in isolation [153, 192–194].

6.3 Network methods and results

The WGS samples were processed using cloud computing (Amazon EC2) and cluster computing servers with the default parameter settings in k-SLAM, and then cross-referenced hits to gene loci against Eggnog, Subsystems and KEGG gene annotations from Section 5.2 to create network edgelists of taxa ID and function ID.

6.3.1 Overview

To ensure there are not obvious artefacts or errors arising from the bioinformatics pipeline that may lead to biases in the resulting networks, I examine summaries of the sequencing statistics. Figure 6.2 shows the distributions for the proportion of reads mapped to species-level taxonomy and proportion which received functional assignments (i.e are included in the network edge list) for the 248 WGS samples. Table 6.3 provides averages for these classifications broken down by biome and annotation system.

Table 6.3: Summary of WGS samples and classification information by biome.

Biome	No. samples	Avg. sequences classified	Avg. prop. at species level	Avg. prop. to networks: Eggnog/Subsystems/KEGG
Human gut	46	3.1E6 ± 1.2E6	96.5	61.6 / 51.2 / 43.2
Soil	83	4.4E5 ± 3.3E5	76.8	43.7 / 36.9 / 33.4
Rhizosphere	23	9.16E5 ± 2.5E5	70.0	48.2 / 41.6 / 38. 0
Marine	60	1.5E6 ± 2.9E5	91.4	63.6 / 56.1 / 51.5
Freshwater	34	9.0E5 ± 6.1E5	84.1	52.4/ 46.1 / 41.2

The bimodality in the proportion of reads mapped to species assigments (Figure 6.2) is due to lower proportions of classified reads in soil data than in other biomes (Table 6.3). The gut samples have the highest share of reads mapped to taxa and func-

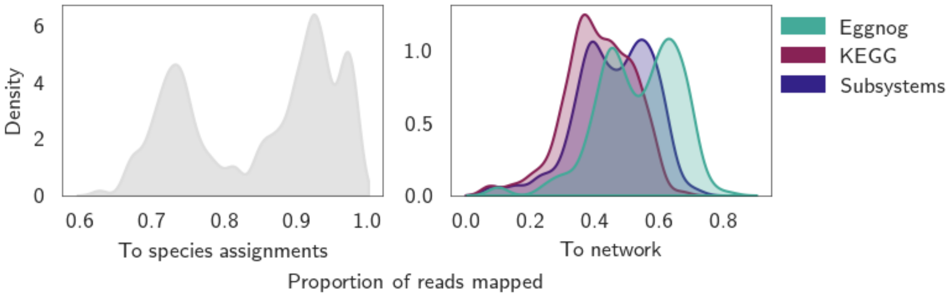


Figure 6.2: Proportion of reads mapped to network under different annotation schemes.

tion, followed by those taken from marine systems. This may be explained by historic sequencing practices and the diversity of the samples. Species-level (and functional) assignments rely upon assembly, and the probabilistic capacity for this to occur decreases with higher diversity, an issue which would be further exacerbated by high numbers of novel sequences. Multiple years of deep sequencing effort within gut and marine biomes, together with lower diversity, means that it is likely most of the functional diversity in those environments has been identified [148, 186]. Conversely, the highly diverse soil and rhizosphere biomes, which also have the lowest proportion of mapped reads, have only been recently prioritised for global deep sequencing exploration, increasing the probability that any given sequence in a sample is unknown [177, 185]. This discrepancy will resolve itself over time as additional sequencing is undertaken in these biomes and coverage improves in the databases. The differences in network coverage between Eggnog, KEGG and Subsystems annotations is consistent with the results seen in the reference genome network (Table 5.1). This suggests that the software is returning functional annotation hits in a comparatively unbiased fashion, matching the functional degree distributions in Figure 5.5, where Eggnog has the highest proportion of hits, followed by Subsystems, and the KEGG ontology returning the fewest.

I now probe the relationships between taxonomic and functional diversity across ecosystems by quantifying key topological properties of the WGS built graphs. To assess results in the context of inference-based prior work (e.g. [153, 187, 190, 192–194]) - which has predicted functional profiles based on taxonomy alone - I also run analyses on a 'predicted' network for each sample, which is constructed by taking the taxonomic profile and including the entirety of each taxon's reference genome (rows of the biadjacency matrices in Section 5.3), whilst noting that the empirical networks are representative of the real-world environment and thus the key focus of this study. As for the reference networks discussed in Chapter 5, all networks and analyses are unweighted.

6.3.2 Summary distributions

I firstly assess overall diversity patterns in the joint taxa-function distributions across the communities by plotting the number of unique genes per sample against the number of taxa (Figure 6.4). This corresponds to the number of columns and rows in the biadjacency respectively. From inspection of Figure 6.4, it is apparent that within each biome, there is a positive correlation between the number of taxa and the number of genes which reaches a horizontal asymptote. This is consistent with observations from previous studies where functional richness saturates at an upper bound with increasing species richness [194]. The asymptotic value observed in the empirical networks is likely a consequence of the sequencing depth. The empirical soil samples sit below the other biomes with respect to the overall functional richness in the WGS samples. This is consistent with the reduced numbers of sequences classified in the soil biomes, the likely cause of which I identified in Section 6.3 as being high taxonomic richness in combination with novel sequences not yet being incorporated into reference databases.

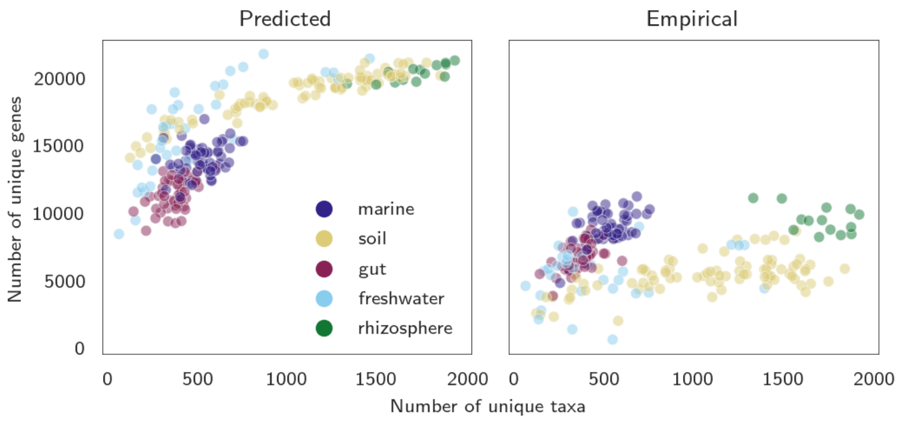


Figure 6.4: Number of unique taxa (rows in the biadjacency) and genes (columns in the biadjacency) per sample

Degree distributions

The core topology of the taxa-function networks may be examined by analysing their degree distributions. Degree distributions drive or are correlated to many statistical properties of complex networks, and in other applications have been used to investigate everything from resiliency of a system to its capacity for control [232, 233]. Here, they provide a snapshot of the macroscale behaviour of the top and bottom node sets, and thus summarise how taxa are shared amongst functions, and vice-versa, across communities. Distributions from representative samples in each biome may be seen in Figure 6.5. Distribution summaries from all samples and annotations (including

Subsystems and KEGG) are qualitatively consistent and may be seen in Appendices F.2 and F.1.

The taxa abundance distributions (top panel for each sample) were calculated by tallying the number of reads assigned to each species, weighting those totals by genome size, and converting to relative abundances. All taxa hits of fewer than $5\text{E}^{-}5$ relative abundance were discarded to reduce the likelihood of false positives. Distributions across all samples show characteristic power-law tendencies common to prokaryotic species abundance distributions [117]. However, the networks' taxa degree distributions - i.e. how many unique functions each taxon has - differ between the empirical and predicted networks. This is because networks built from empirical WGS data is driven by the empirical taxa abundance distribution. Therefore, it shows exponential decay, as few functions were detected in the rarest species. Conversely, with the predicted network, every function from those 'rare' taxa are included in the network, meaning that the predicted network taxa degree distribution is more similar to the entire reference network itself.

The probability mass function for the empirical and predicted network gene distributions indicate heavy tailed behaviour. This is consistent with prior work in genome-gene family viral networks as well as genome-function networks in the human gut microbiome [192, 206]. A slight peak in the tail in function degree distributions for the predicted network is observed due to a binning effect from the housekeeping genes, where there is a maximum degree corresponding to the number of taxa in the samples. Whilst prior work has attempted to fit power laws to these distributions through regression on the probability mass function, recent state of the art methods detail fundamental theoretical flaws with such approaches and I therefore do not attempt to do so [206, 234]. From the complementary cumulative distribution functions in Figure 6.5 it is clear that these are not pure power laws. However, they do display heavy tail properties, which is the informative feature in a biological context. The tailed behaviour indicates that a block of core functions across all taxa make up a small but significant proportion of the functional diversity in the system. Therefore, from the perspective of multifunctional redundancy, how the 'uncommon majority' of functions are distributed amongst taxa is of high importance, especially as the results indicate that this tailed function degree distribution is ubiquitous across all biomes [152, 153, 186]. This indicates that the first statistical property of the WGS networks - degree distributions - is consistent across both samples and all of the biomes.

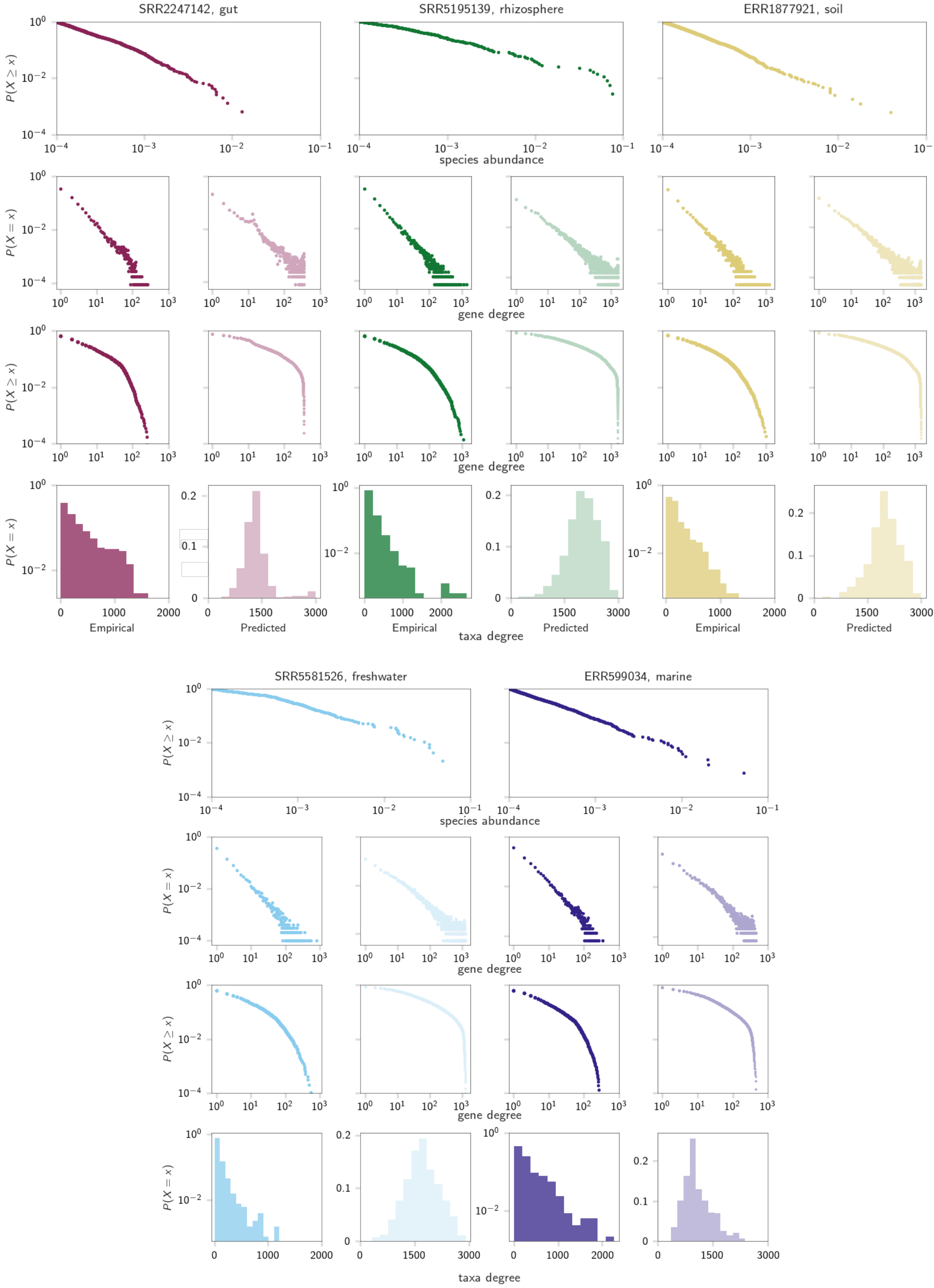


Figure 6.5: Degree distributions in empirical networks annotated with Eggnog, showing representative samples from different biomes.

6.3.3 Network analyses

Overview

To quantify redundancy structures within the microbial community networks, and examine the network distributions of different functional categories, I examine the singular value distributions and run three statistical measures on the graphs. In short, to understand the behaviours of different functional groups in microbial communities, I use statistics which are informative about specific structural elements within bipartite networks by using two mean field metrics and one which is able to capture local structure. These measures were run on empirical and predicted networks (for each of the annotation methods) for each of the WGS samples. Furthermore, the metrics were also run on the degree-preserved randomisations of each network. As many statistical properties of networks are correlated to node degree, comparing against these random baselines allows the separation of which characteristics of the system are driven by the degree distributions, and those which may be indicators of other underlying structure and require further investigation. Between the different network construction methods, annotation systems, and randomisations, I generated an ensemble of over 3000 networks to analyse.

The Curveball algorithm was used to randomise the networks [235, 236], with the Python implementation from [237]. This algorithm was designed for bipartite networks, and works by swapping individual entries between rows in an adjacency list (whilst preserving the total number of entries per row). It is fast, unbiased, and quickly converges to the maximal possible perturbation from the original system [235]. The conserved degree sequences in combination with the tailed gene distribution mean that there are some links which do not change between the original network and its randomised counterpart. However, it is possible to check that the system has been shuffled as much as possible within the degree distribution constraint by plotting the proportion of altered links versus the number of iterations (Figure 6.6). Once the perturbation score reaches its asymptote, the randomisation may stop. For the randomisations, I ran triple the iterations expected for each network to be sure asymptote was reached (perturbation vectors from the randomisations are available in Appendix A).

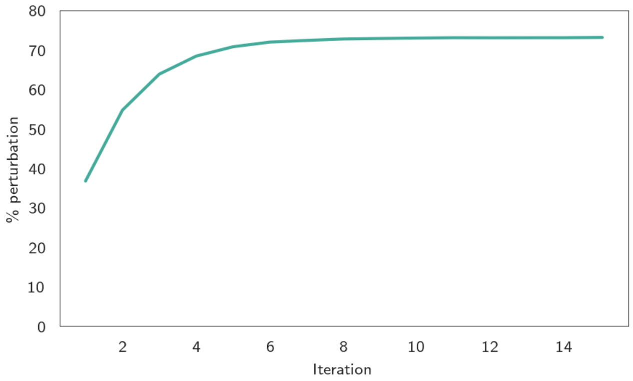
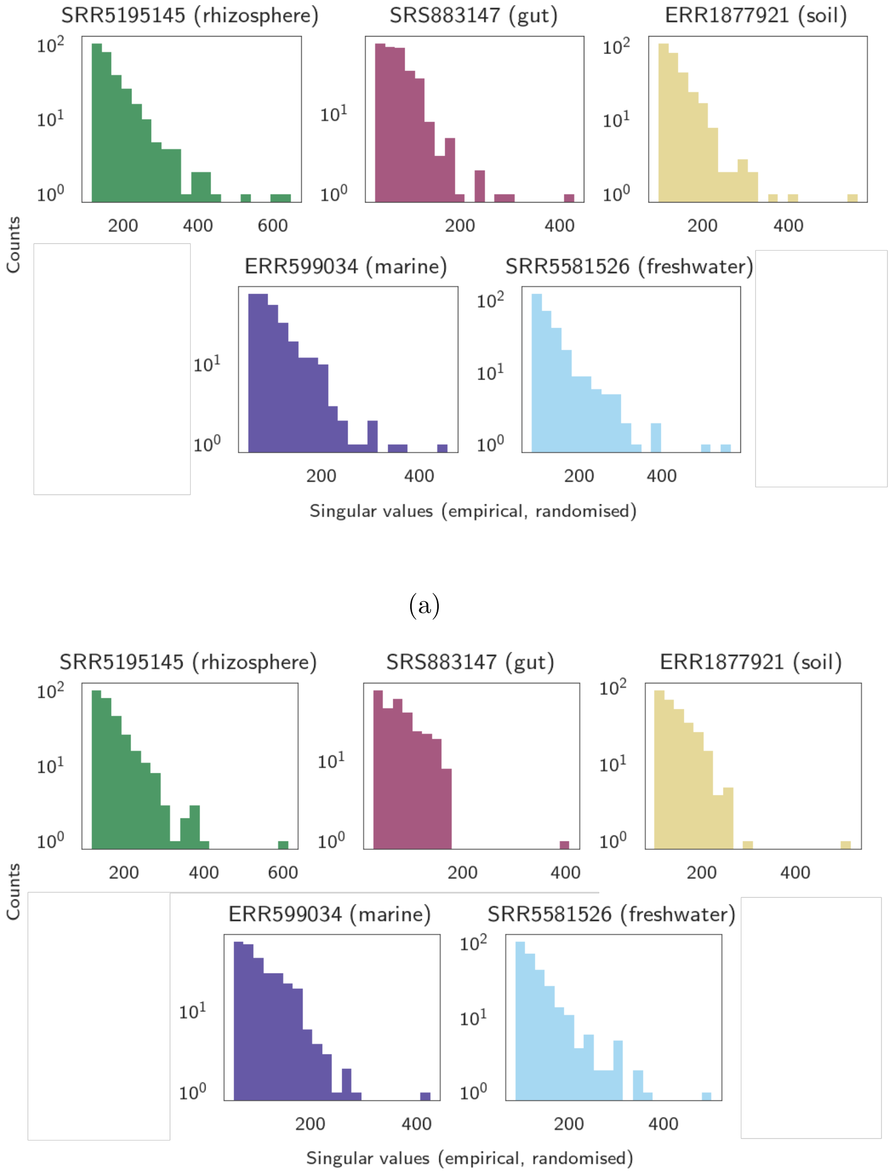


Figure 6.6: Perturbation of a network during successive iterations of Curveball algorithm's randomisation scheme.

Singular value distributions

As introduced in Chapter 5, spectral decomposition of a biadjacency matrix may quickly identify network community structure should its singular value distribution reveal a series of large values sitting outside of the bulk of the spectrum. In the natural communities under investigation, it would indicate that organisms may be grouped into coarse-grained metabolic niches due to the presence of suites of redundant functions. Community detection algorithms or modularity scores were not used to identify these redundancy blocks for multiple reasons. Firstly, community detection is a NP-hard problem [238]. Whilst there are a large number of heuristics available, their propensity for success is application specific. Furthermore, heuristics find local minima, or require maximisation of other parameters, meaning solutions require bootstrapping, making the computational cost too great in this setting, where we would need to apply them to several thousand large networks. Given many algorithms rely on spectral methods to identify network communities, examining singular value distributions directly can qualitatively confirm the presence (or absence) of community organisation within the network [219].

Figure 6.7 depicts representative distributions for empirical networks and their randomised counterparts, obtained by applying SVD to the biadjacency matrices (methods outlined in Section 5.4.2). Singular value distributions for predicted networks and their randomisations may be seen in Figure 6.8.



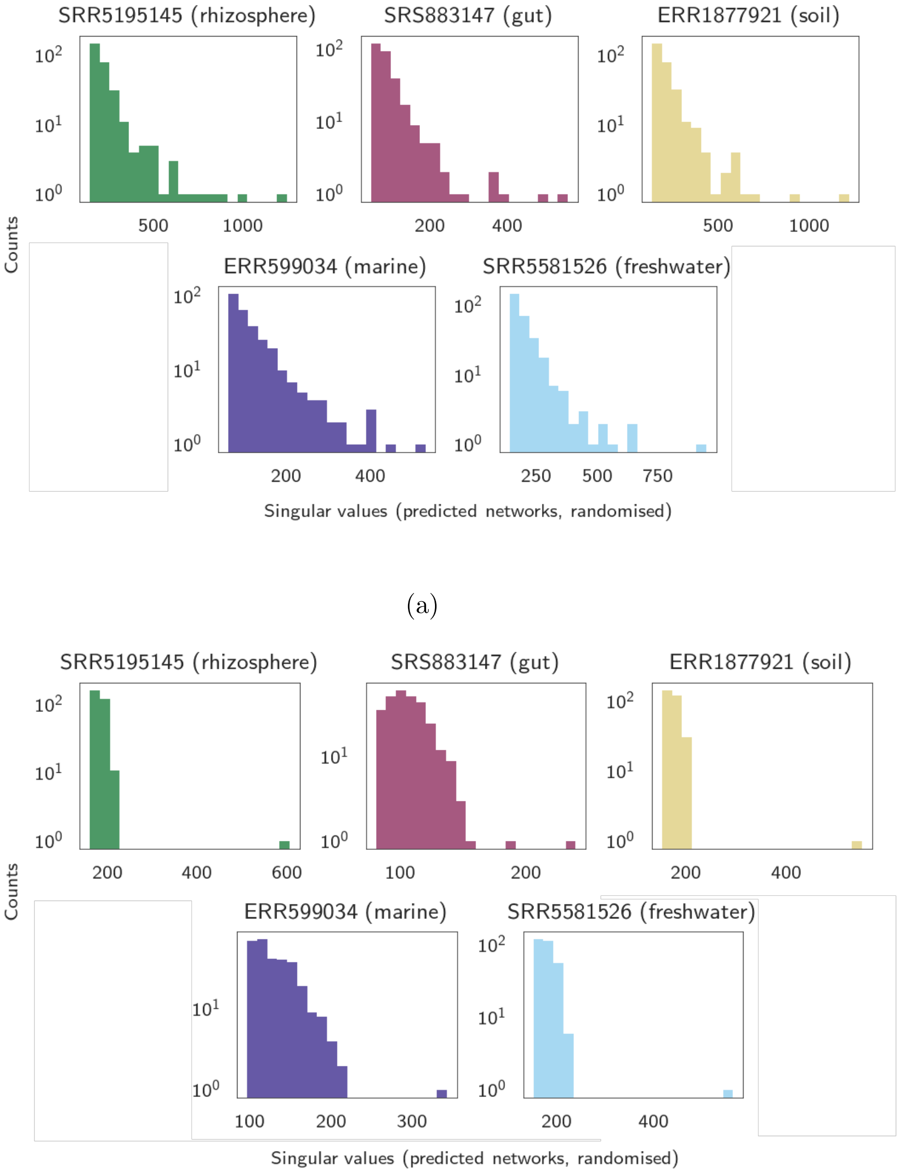
(a)

(b)

Figure 6.7: Distributions of top 300 singular values from (a) empirical networks and (b) randomised empirical networks. Representative samples are consistent with those listed in Figure 6.5.

Similar to the patterns observed within the complete reference network, the empirical and predicted WGS networks yield a series of larger singular values which sit apart from the bulk of the distributions. It would be expected that the predicted networks - being samples of complete rows from the reference biadjacency - may display this behaviour. However, the networks constructed from WGS sequencing data alone also reveal similar patterns. These distributions are indicative not simply of high saturation for a few functions (as would be the case for housekeeping genes), but an organisational principle where suites of functions are segregated across different communities in the network.

In the randomised predicted networks, there is one large singular value separated from the bulk of the distribution (Figure 6.8b). That is, the modular elements and other fine-grained structures in the real-world networks which gave rise to multiple large singular values have been destroyed by the randomisation. In the randomised empirical networks (Figure 6.7b), there is a less dramatic difference between the real-world and randomised networks. This can be explained by the tailed taxa abundance distribution, as the majority of taxa are found in the rare biosphere, and they therefore contribute only one or two links in the network. The small number of dominant taxa are also likely to be more metabolically similar than the taxa within the rare biosphere, meaning that whilst the modular elements evident in the predicted networks are present, they are at lower levels of resolution in the empirical WGS data. However, the qualitative behaviour is consistent, providing evidence that the WGS constructed networks have detectable community structure.



(a)

(b)

Figure 6.8: Distributions of top 300 singular values from (a) predicted networks and (b) randomised predicted networks. Representative samples are consistent with those listed in Figure 6.5

To examine these distributions in aggregate, the first 30 singular values from each network decomposition were normalised to a maximal value of 1. For each sample, I examined the pairwise relative distribution of the real-world network and its randomisation's top 30 singular values [239]. Noting that measures such as Kullback-Liebler divergence reveal if not where distributions differ, the use of relative distributions provide a scale-invariant, semi-quantitative method to reveal the location of disparities between non-parametric distributions [239, 240]. Here, they reveal whether there are consistently greater numbers of large singular values in the real-world networks compared to the randomised networks. Relative distributions may be visualised in multiple ways; here, we plot the ratio of the real to randomised singular value against its rank. Figure 6.9 shows the relative distributions for the empirical samples, and predicted samples, which are qualitatively the same, are shown in Appendix F.

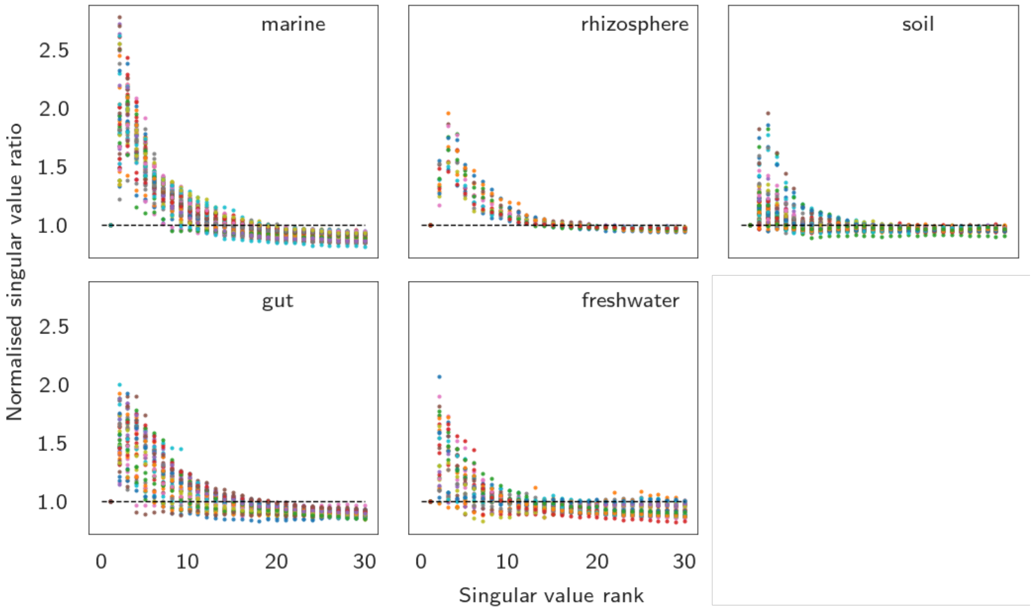


Figure 6.9: Relative distribution ratios of the top 30 singular values across biomes. For each network, the top 30 singular values were normalised to a maximum value of 1. For each sample, pairwise ratios of real-world and randomised singular values are plotted against their rank. The dotted line indicates a value of 1; if the distributions are the same, they should fall along or near the line.

If the real-world and randomised singular value distributions were the same, their relative distribution ratio would be constant around an approximate value of 1. However, instead, there is consistently a large peak in the first few ranks before this ratio drops back to unity: this indicates that across all biomes and samples, real-world networks have more large singular values sitting outside of the bulk distribution compared to their randomised counterparts. From the network perspective, this is reflective of real-world networks having detectable community structure, and from a biological perspective indicates that all biomes appear to display metabolic niche redundancy. That is, the microbial community is able to be coarsely partitioned by taxa metabolism - most likely reflecting the major ecological niches within the biome. Whilst this has been an assumed property of microbiomes, to my knowledge this is the first time it has been explicitly shown from WGS data, rather than inferred from reference genomes or models [153, 192]. I next quantify gross structural features in the networks by assessing whether the networks have higher levels of nestedness than their random baselines, or if it is necessary to examine alternate explanations for functional redundancy in the community.

Nestedness

Nestedness is a classical metric for bipartite networks in ecology. Initially used as a measure to describe species' spatial patterns, it has since been applied to a wide range of organism interaction networks, such as plant-pollinator communities, and in a variety of economic settings, for example, to study trade networks [241]. It is a measure of self-similar structures within the network, and in this context measures the extent to which smaller genomes are subsets of larger genomes. Nestedness has been proposed to either increase or decrease the stability of networks depending on whether the objective is to preserve specialist or generalist roles; furthermore, it has been proposed to increase functional redundancy in microbiomes [192, 241].

Here, I assess whether nestedness appears as a consequence of other topological features of the network, or is itself a mechanistic feature of the joint taxa-function distribution within microbial systems as has been proposed in prior work [192]. Whilst there are several methods to measure nestedness, due to its low levels of bias, I use the method of 'overlap and decreasing fill': NODF [241–243]. It is one of the only nestedness measures which is computationally tractable for networks of the dimensionality used in this study, as it is possible to vectorise the equation in a way that is robust to changes in the order of rows or columns in the biadjacency.

To calculate NODF, I consider two rows i, j in the biadjacency associated with vertices of degrees k_i and k_j. For every pair of vertices within the R rows, a value S_{ij} is defined by

$$S_{ij} = \begin{cases} 0, & \text{if } k_i = k_j \\ \frac{I_{ij}}{min(k_i,k_j)}, & \text{otherwise} \end{cases}, \tag{6.1}$$

where I_{ij} is the total number of edges in common between them. Equation 6.1 may also be applied to determine S_{ij} across the pairs of vertices within the C columns. NODF is then defined by

$$\mathrm{NODF} = \frac{\sum_{i<j}^{R} S_{ij} + \sum_{i<j}^{C} S_{ij}}{\frac{R(R-1)}{2} + \frac{C(C-1)}{2}}, \tag{6.2}$$

and I implement the vectorised algorithm developed by [244] for the calculations.

There are differences in the NODF scores between the annotation systems, the biomes, and also the empirical and predicted networks (Figure 6.10). The empirical networks display lower nestedness than their predicted counterparts. This a result of the exponential taxa degree distribution, where the rare taxa have only a few genes, increasing the sparsity of the matrix and therefore decreasing the NODF score [243]. The highest nestedness appears in the gut biome, whereas the lowest is apparent in the rhizosphere. Other biomes lie between, with the NODF scores appearing to correlate to the dimensionality of the networks (Figure 6.4). Furthermore, the higher nestedness observed in the KEGG network over the Subsystems network is likely an artefact of the multilabel nature of KEGG annotations, which causes a higher network density [243]. Finally, the similarity in nestedness scores between the randomised and real-world networks (Figure 6.10) implies that the degree distributions may be the main driver of NODF values.

To determine whether real-world taxa-function networks display higher nestedness than would be expected by chance, within each biome, I undertook Mann-Whitney U tests on the Eggnog NODF scores across real and randomised network ensembles (Table 6.11; refer to Appendix H for qualitatively similar results in other annotation systems). For the NODF scores, at a threshold of $\alpha = 0.05$, there were not significant differences between the empirical networks and their degree preserved randomisations. For the predicted networks, the randomised networks had significantly higher NODF scores for the soil and rhizosphere (2.1% and 6.4% respectively) and significantly lower scores for the marine biome (5.6%). Through simulation, it was possible to infer that the higher scores in the randomised soil-associated networks arose from modularity in the network. Thus, whilst some of the predicted networks display minor differences between the real-world data and random baselines, the effect sizes are small and not in consistent directions; furthermore, no differences were observed when using the WGS constructed empirical networks. This differs from the outcomes of the analysis in [192], likely due

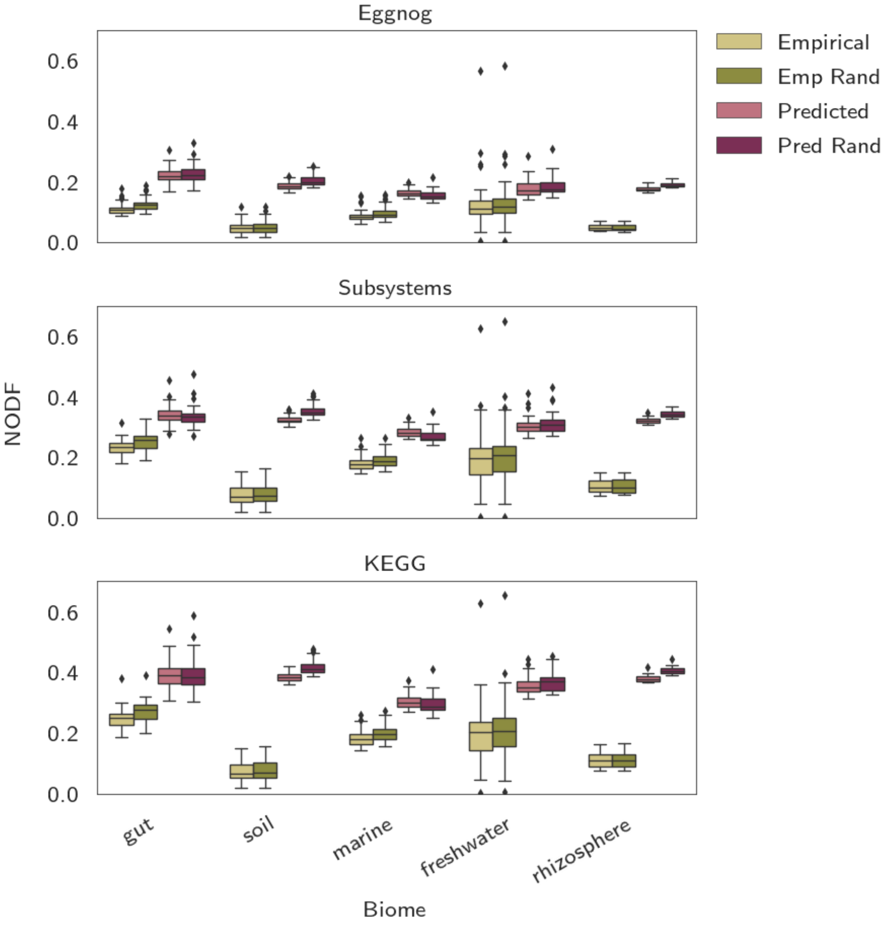


Figure 6.10: NODF in the empirical and predicted networks, as well as each network's degree preserved randomisation.

to my use of a dereplicated taxonomy and database, together with a randomisation algorithm which preserves the unweighted degree distribution. My results are consistent with theoretical findings in prior work detailing the problematic nature of interpreting nestedness in ecological networks, and particularly the extent of its explanatory power [208, 222]. Indeed, here the degree distribution is the largest contributor to nestedness scores in taxa-function relationships in microbial communities, likely due to the statistic being dominated by the power law and exponential behaviour of the bottom and top node sets. Deviations from random baseline scores are more likely to arise from other small scale structures in the network, which I explore in later sections.

Table 6.11: Statistical results: identification of differences between Eggnog NODF scores in the WGS and randomised networks (Mann-Whitney U test, two-sided). Here (E) denotes empirical networks and (P) denotes predicted networks. Significant p-values following a Šídák multiple test correction are noted with an asterisk. †The first value in these columns is for the real-world networks, the second for the randomised.

Biome	n	Median† (E)	U (E)	p-value (E)	Median† (P)	U (P)	p-value (P)
Gut	46	0.22 / 0.22	545	0.29	0.11 / 0.12	948	0.20
Marine	60	0.16 / 0.15	1338	6.6E-03	0.084 / 0.089	2533	5.1E-05*
Freshwater	35	0.17 / 0.17	567	0.29	0.11 / 0.12	529	0.17
Soil	84	0.18 / 0.20	3358	0.26	0.047 / 0.048	1169	1.6E-11*
Rhizosphere	23	0.18 / 0.19	280	0.37	0.047 / 0.044	59	3.3E-06*

Functional redundancy (FR)

As a key project aim is to explain redundancy properties in microbial systems through topological features of taxa-function networks, I endeavoured to reproduce and extend recent results presented on functional redundancy in the human microbiome [192]. The authors in [192] define functional redundancy as the taxonomic diversity which is unexplained by the functional diversity. They consider genome distances across the community weighted by relative abundance to propose a redundancy measure for a taxa-function bipartite graph, given by

$$FR = 1 - \sum_i p_i^2 - \sum_i \sum_j d_{ij} p_i p_j, \tag{6.3}$$

where the total taxonomic diversity is given by the Gini-Simpson index $1 - \sum_i p_i^2$, and total genetic diversity is given by Rao's quadratic entropy, $\sum_i \sum_j d_{ij} p_i p_j$ [192]. The p_i's are the relative abundances for taxa i, and d_{ij} is the Jaccard distance between the genomes of species i and j, which can be calculated from the bipartite graph. I note that the definition of Jaccard distance,

$$J(A,B) = \frac{A \cap B}{A \cup B}, \tag{6.4}$$

allows for an extremely fast implementation for a graph in adjacency list format with the use of list and set operations; for networks of the size and sparsity used in this study, this algorithm was needed to make analyses computationally feasible (Appendix A).

For calculating the score, I only use the predicted network. This is because using the empirical networks - for which the function degree distribution is correlated to the taxa abundance distribution - would bias the result given the taxa abundance distribution is explicitly encoded into the metric. I also depart from [192] by using an unweighted graph, for the reasons outlined in Chapter 5. My results broadly mirror the patterns observed for the NODF scores. Indeed, when I break down the redundancy values to check the correlation to species richness, we are able to see a natural 'envelope' in which microbial communities appear to sit (Figure 6.12), consistent with the taxa-function richness saturation curve [152]. This indicates there is an upper and lower bound of redundancy values, with similar scores across most biomes when considering the taxa richness. As for NODF, I next test whether this result is driven by unique topological features within the networks, or alternately whether it is largely driven by the degree distributions of the network.

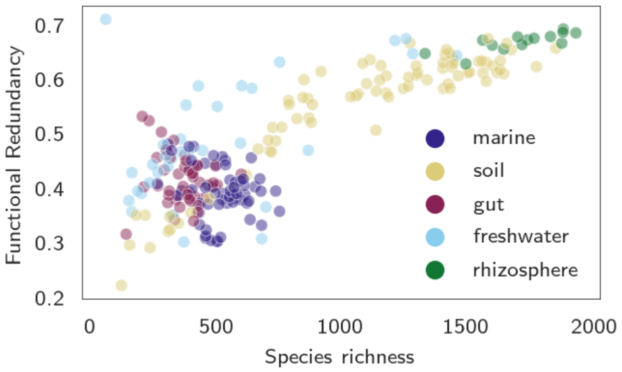


Figure 6.12: Functional redundancy score vs. taxonomic diversity: communities sit within an envelope of values. Calculations undertaken with Eggnog-annotated networks (refer to Appendix G for KEGG and Subsystems figures).

The FR equation links microbial abundances to a network-based Jaccard distance, necessitating an additional step in the randomisation. In real-world microbial systems, tailed degree distributions mean that a small number of taxa dominate the community. As the p_i's in Equation 6.3 follow a power law distribution, and are effectively squared, the FR value is weighted toward the genome distances between the most abundant organisms. Naiive degree-preserved randomisation destroys this structure, i.e. there is no longer a signal which correlates the dominant taxa with those which are - on balance of probabilities - more closely related than any two random taxa from the rare biosphere. A coarse approximation for this behaviour may be achieved by sorting the rows and columns of the biadjacency by node degree. I also sort the taxa abundances to correlate with the degrees of the randomised network. The ratio of real to randomised networks for both the unsorted and sorted case may be seen in Figure 6.14.

Table 6.13: Statistical results: identification of differences between Eggnog FR scores in the WGS and randomised networks (Mann-Whitney U test, two-sided). Here (E) denotes the sample results, (R) the randomised network and (RS) the sorted randomised network. Significant p-values following a Šídák multiple test correction are noted with an asterisk.

Biome	n	Median (E/R/RS)	U (E-R)	p-value (E-R)	U (E-RS)	p-value (E-RS)
Gut	46	0.40 / 0.37 / 0.39	1559	9.3E-5*	1271	0.10
Marine	60	0.39 / 0.34 / 0.38	2831	6.3E-8*	2008	0.28
Freshwater	35	0.47 / 0.42 / 0.46	673	0.25	568	0.91
Soil	84	0.59 / 0.61 / 0.63	2693	0.01	2179	4.4E-5*
Rhizosphere	23	0.67 / 0.67 / 0.68	101	0.91	68	0.18

Sorting corrects for the loss of the phylogenetic signal and eliminates what may have initially appeared to be differences between functional redundancy scores within real and randomised networks. Similarly, if I test these findings statistically (Table 6.13), what was initially a significantly higher FR score in real-world gut and marine biomes becomes insignificant. Furthermore, the soil biosphere displays significantly higher redundancy in the randomised sample by 3.3%, and when the randomised network is sorted, this increases to a difference of 6.3%. This may be understood by the taxa-function richness curve: at high species diversity, few new functions are introduced via the introduction of new organism, and increasingly large numbers of more common functions overlap [194], which serves to inflate the FR score. I conclude that microbial communities from the human gut or other biomes do not display higher levels of redundancy than their random baselines using the FR metric proposed in [192], and that the redundancy levels are consistent across communities for a given level of diversity. However, it is be noted that the tailed species abundance (and functional) distribution makes it challenging to develop a numerical formulation of redundancy which is not skewed toward the most abundant organisms and functions (an issue also considered in [153, 193]). This is conceptually problematic as the majority of taxonomic and functional diversity in microbial communities is contained within the rare biosphere. Furthermore, typically rare genes, e.g. forms of antibiotic resistance, may rapidly propagate through the community if they become beneficial, and dominant species change with boom-bust dynamics where up to 40% of species may lie dormant for long periods of time [245]. Indeed, the sensitivity of functional redundancy measures to abundance fluctuations, labeled by the authors as 'robustness', is examined in detail within [193]. Therefore, I next consider the network distribution of different types of functions independent of their, or an organism's, abundances. This enables an examination of redundancy from the perspective of how metabolism is organised across communities, capturing the underlying functional organisation whilst avoiding biases arising from transient abundances and genome plasticity.

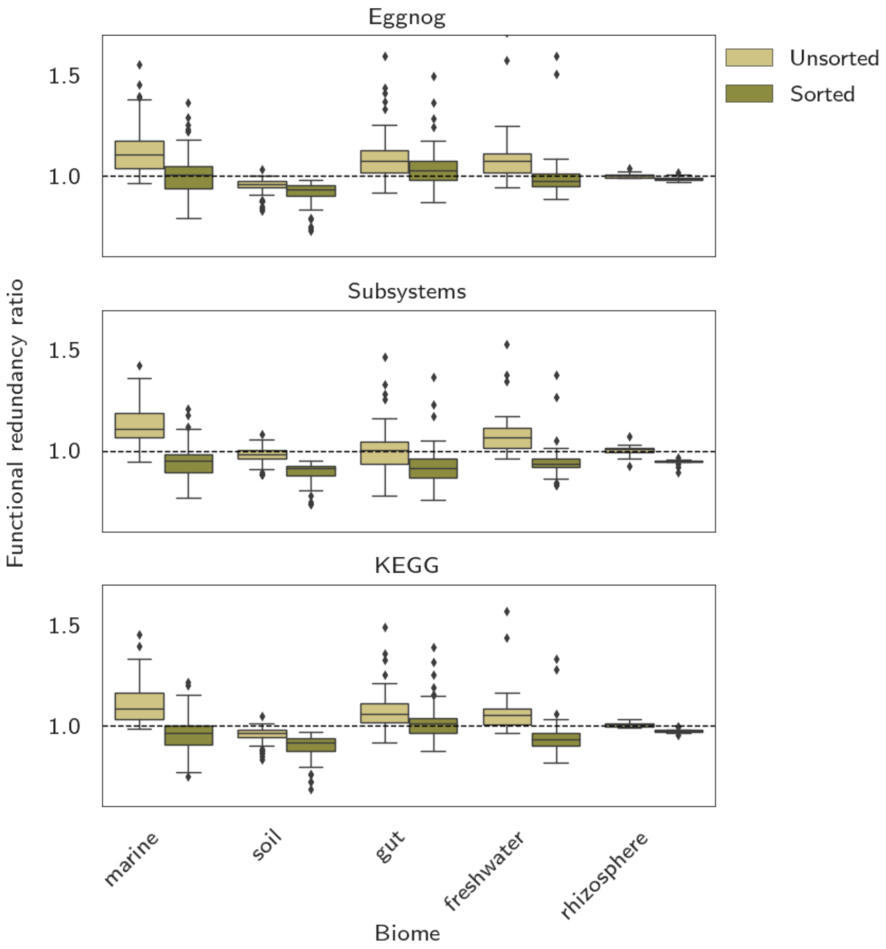


Figure 6.14: Ratio between real and randomised network functional redundancy scores across biomes. The first value indicates an unsorted randomised network, the second value indicates a sorted randomised network and sorted abundance profile to mimic the presence of phylogenetic structure.

Clustering

As an alternate and novel approach to quantifying functional redundancy in microbial communities, I examine global and local clustering behaviour. This allows for an examination of local graph structure and whether different types of functions are distributed differently across the network, irrespective of their saturation. This network-based method allows certain functions are grouped within metabolically related organisms, or spread widely across the community.

The bipartite clustering coefficient is the two-mode generalisation of the classical triadic closure principle of unipartite networks [246]. In bipartite graphs, triadic closure needs to be generalised to neighbors of neighbors. Each node's local clustering coefficient (Equation 6.5) indicates a node's propensity to form a 4-cycle. That is, it captures the likelihood that in general if taxa u_1 has function v_1 and v_2, and taxa u_2 has function v_1, it also has v_2 (Figure 6.15).

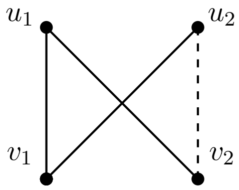


Figure 6.15: Schematic of the concept of the local clustering coefficient, which captures a node's propensity to form 4-cycles with other nodes in its set (top or bottom).

The bipartite clustering coefficient of a node, also known as the Latapy clustering coefficient, is formally defined by

$$c_u = \frac{\sum_{v \in N(N(v))} c_{uv}}{|N(N(u))|} \tag{6.5}$$

where $N(v)$ are the second order neighbor nodes of u in bipartite graph G, and c_{uv} is the pairwise clustering between two nodes given by $c_{uv} = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$ [203]. A high clustering coefficient indicates a node is grouped within a suite of similar taxa and functions: that is, within a metabolic niche. A low clustering value for a function means that it is spread across random (i.e. metabolically dissimilar) taxa in the network. The network's clustering coefficient is the average of all of the vertices' individual clustering coefficients.

In the WGS-built networks, there are similar trends in the network's average clustering coefficient as there are for the NODF score, with correlations to the dimensions of the network (Figure 6.16). The sparser networks generated by Eggnog annotations have lower clustering values than those within Subsytsems or KEGG networks. KEGG-built networks have on average higher clustering scores than those derived from Subsystems. Testing whether these signals arise purely from the degree distributions or are indicative of fundamental assembly principles between taxa and functions across natural communities (Table 6.17), it is evident that unlike the scores for NODF or functional redundancy which revealed no or weak differences, there are significant differences between average clustering real-world and randomised networks across all biomes. The effect is always in the same direction, with the average clustering coefficient in the Eggnog empirical and predicted networks being 23.5% and 20.7% larger than their randomised counterparts respectively; refer to Appendix I for qualitatively similar results in other annotation systems. The same modular elements which contribute to the behaviour of the singular value distributions would also increase the network's average clustering coefficient, quantitatively supporting the conclusions drawn from the relative distribution analysis and providing evidence that communities can be clearly partitioned into metabolic niches across all of the biomes.

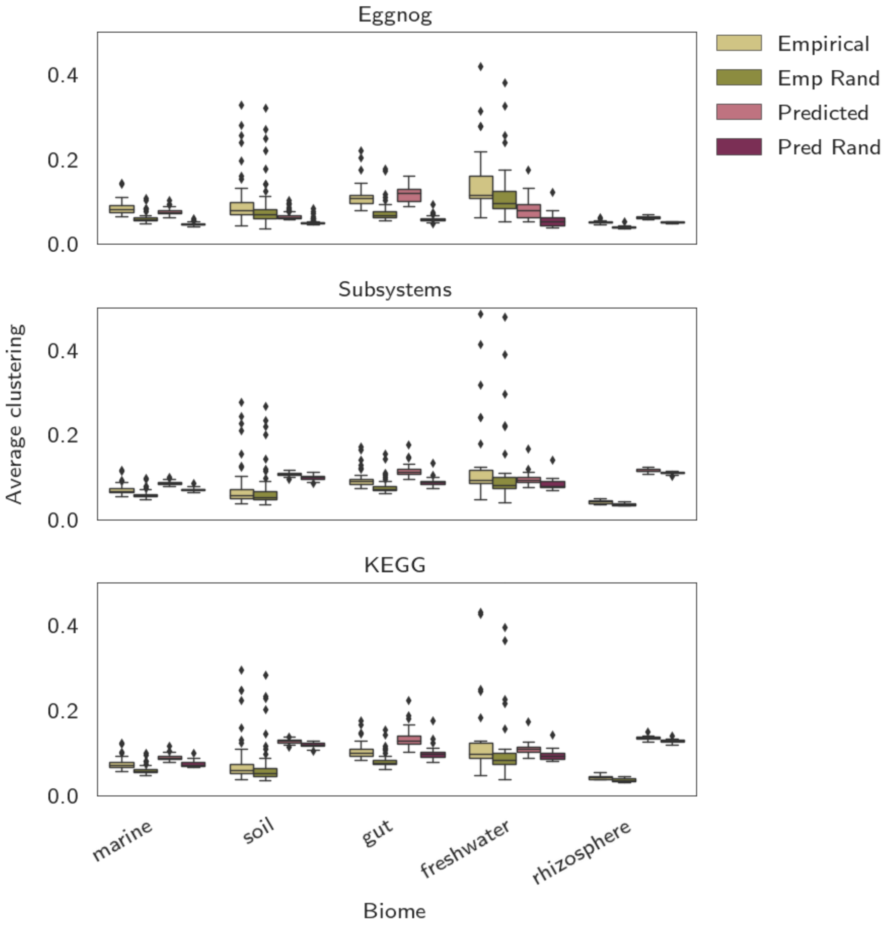


Figure 6.16: Clustering coefficients for empirical, predicted and degree preserved randomisations for each sample network.

To break down where these differences occurred and identify which functions were randomly spread through the community versus associated with metabolic or taxonomic niches, I examine the local clustering coefficients of the functional vertex set from the Subsystems networks. This allowed me to analyse clustering scores for different functional categories. This analysis was applied to all fully connected empirical networks, as it was otherwise not possible to undertake node-wise comparisons between

Table 6.17: Statistical results: identification of differences between Eggnog average clustering scores in the WGS and randomised networks (Mann-Whitney U test, two-sided). Here (E) denotes the empirical networks, (P) the predicted networks. All p-values were significant following a Šídák multiple test correction. †The first value in these columns is for the real-world networks, the second for the randomised.

Biome	n	Median† (E)	U (E)	p-value (E)	Median† (P)	U (P)	p-value (P)
Gut	46	9.7E-2/7.3E-2	1917	2.0E-11	0.12/8.5E-2	2114	1.7E-16
Marine	60	7.3E-2/5.8E-2	3371	1.7E-16	8.4E-2/6.8E-2	3600	3.6E-21
Freshwater	33	0.11/8.6E-2	752	7.9E-3	9.5E-2/8.0E-2	912	2.5E-06
Soil	84	6.4E-2/5.7E-2	4354	3.3E-3	0.11/9.9E-2	6495	6.8E-23
Rhizosphere	23	4.5E-2/3.8E-2	186	5.8E-5	0.12/0.11	196	7.5E-06

the real-world networks and their random baselines. To analyse the network scores, nodes were removed for ubiquitous functions which were 70% saturation or greater in the reference network. Next, I removed nodes that had fewer than 2 links; as I aimed to capture the distribution of functions across the full network, this required the function be associated with at least two taxa. I then transformed the clustering scores for the remaining nodes to a standard normal distribution using quantile transformation, and aggregated the mean scores (per sample and functional class) for each biome. This allowed assessment of the relative ranks (above or below the mean clustering per network) within each biome, revealing the tendency for each gene class to cluster within a community, or alternately, be shared amongst taxa which otherwise have few similar genes (Figure 6.18).

There are evident differences in the clustering behaviour of different functional classes, which is a surprising finding given that the housekeeping genes would have a significant smoothing effect on the statistic, speaking to a more dramatic real-world effect. Here, a low value indicates a functional class is shared amongst taxa that have fewer genes in common (on average). Across all biomes, low scores were ubiquitous for prophages, transposable elements and plasmids, along with CRISPR, whereas motility genes were consistently above the mean.

There was a high amount of variability amongst other functions, which may be indicative of shifting core functionality in the context of changing biochemical environments [147]. Photosynthesis related genes were especially variable, appearing to be spread randomly through the network in soil, rhizosphere and gut, and being strongly niche associated in aquatic environments. However, the aquatic biomes would argu-

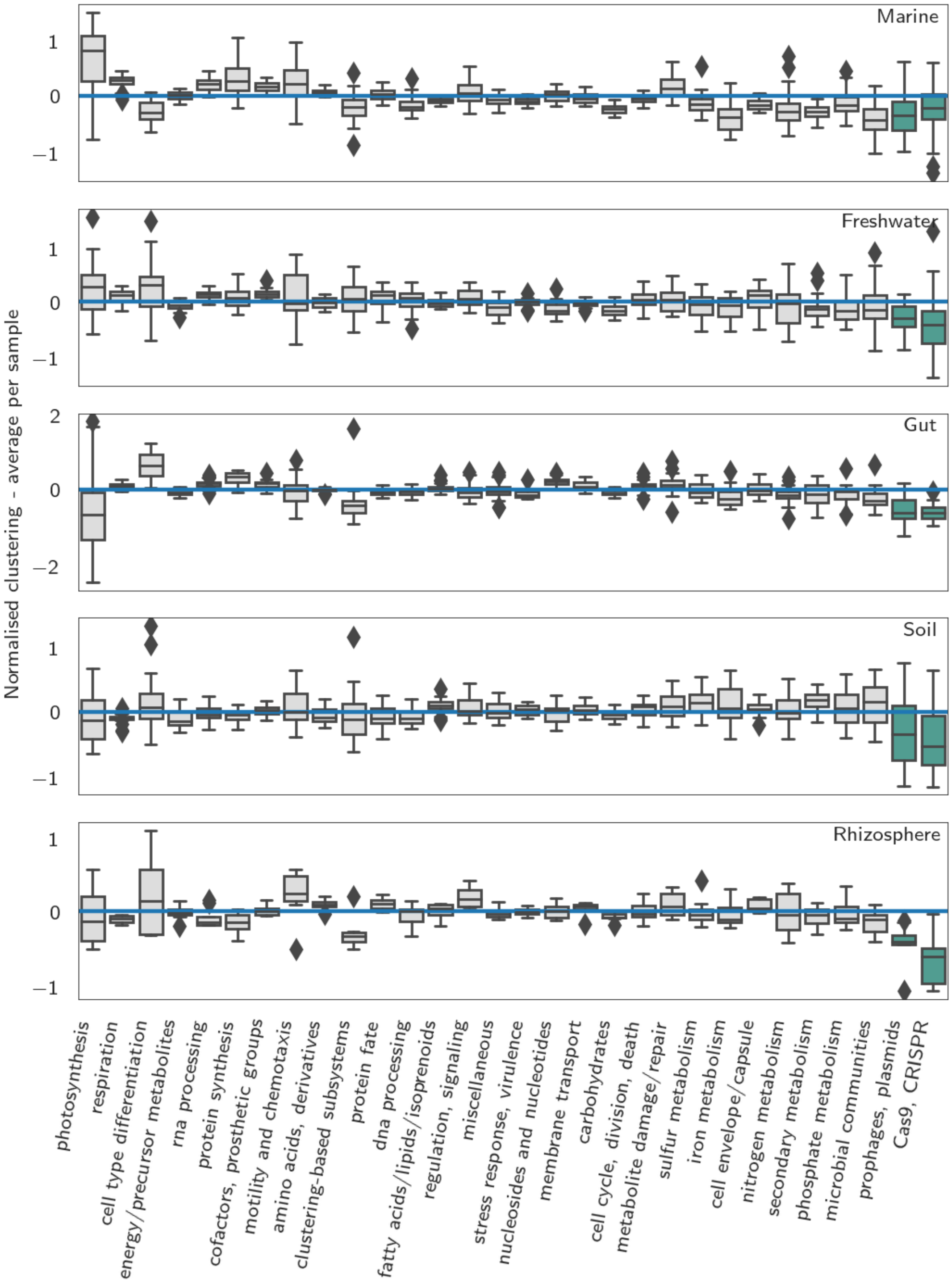


Figure 6.18: Average clustering of different functional categories across biomes. Node clustering scores were transformed to a standard normal distribution for each sample, and then aggregated within each biome to reveal whether different functional categories are tightly clustered or spread widely across the network.

ably be the only environments which would have those metabolic pathways in full; there were just 2.7 hits on average within the gut biome samples, where it is clear no photosynthesis occurs and it is likely that the functions categorised as belonging to photosynthesis are being utilised in alternate metabolic pathways.

To assess whether there are statistically significant differences between the distribution of different functional groups across the community, there was an additional step to correct for potential degree correlation in the clustering scores. Prior to the normal transformation, I took the ratio of each node's local clustering value with that of its random baseline; approximately 100 samples were excluded at this step as fully connected networks were necessary for a meaningful real-random nodewise comparison. Next, to confirm that there was no degree correlation, the average clustering ratio against the average degree of the class was plotted (Figure 6.19), and I verified that the distribution was uncorrelated by undertaking a linear regression (revealing a slope coefficient p-value of 0.52). As the effect size of the randomisation is reduced due to the smoothing impact of housekeeping genes on the clustering values, and a large quantity of data was excluded, samples were aggregated across the biomes to increase statistical power (Figure 6.20). Finally, multiple comparison tests (Tukey HSD at a significance and false discovery threshold of 0.05 [247]) were undertaken to assess which functional classes were significantly different from each other (Figure 6.21).

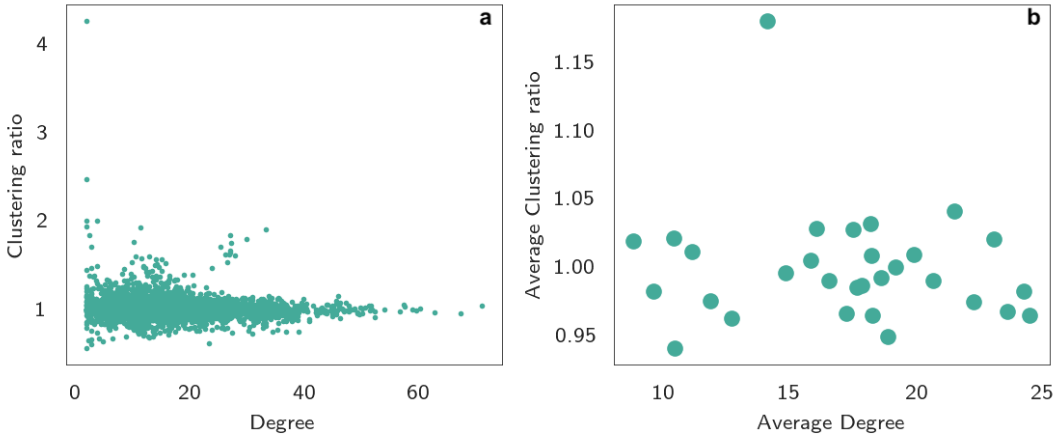


Figure 6.19: Assessing potential for degree correlation in the clustering ratios, across all data points (left) and the their averages for each of the 31 Class level functional categories (right).

Whilst Figure 6.20 shows that many groups are distributed around the mean, the categories showing high entropy in Chapter 5 (prophages, plasmids and CRISPR) had the lowest real-random clustering ratios in the natural communities. Conversely, the

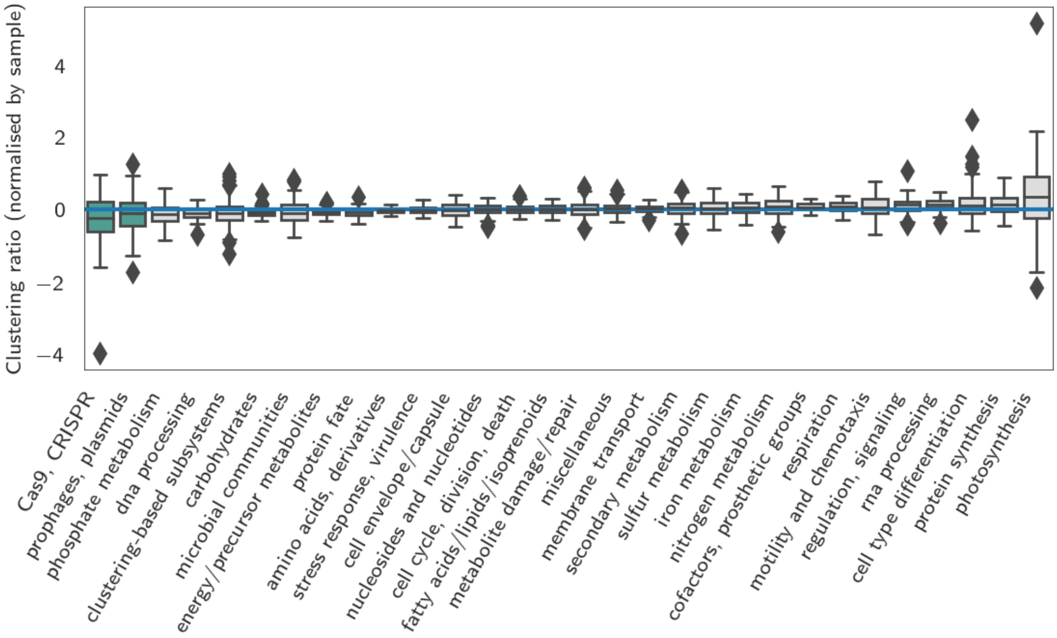


Figure 6.20: Local clustering value of vertices across the biomes ($n = 154$ samples), broken down by class and sorted by median clustering ratio.

low entropy groups such as respiration, photosynthesis, RNA processing, and protein synthesis had higher ratios, suggesting stronger tendencies to associate within a niche. The high variance in the photosynthesis functional group may be explained by the aggregation of data across biomes in conjunction with the fact that three of the biomes (gut, soil and rhizosphere) were unlikely to have organisms undertaking photosynthetic processes. Regardless, photosynthesis displayed significantly higher clustering values than every other functional group (outlier in Figure 6.19b, Figure 6.21). This may be understood by the fact that samples which had photosynthesising taxa would have suites of functions across similar taxa leading to high clustering values, whereas samples without photosynthetic organisms would have very few photosynthesis functional hits, leading to high average clustering values overall. This corroborates recent research showing photosynthesis occurs in metabolically specialised, and highly specific, phylogenetic groups [209]. Whilst this study was not designed to identify HGT genes from the environment, the functional categories associated with HGT (extracellular DNA groups such as plasmids, phages and CRISPR) had significantly lower clustering ratios than almost every other functional group (although not compared to each other). This indicates that these functions are spread in a diffuse pattern throughout the networks, meaning that an organism's overall metabolism is not predictive of whether it has genes in these categories or not. When taken in conjunction with Figure 6.16, this shows that the functional redundancy structure of microbiomes follows universal behaviours, where HGT associated functions are broadly spread through communities, and functions known to be specific to an organism's ecological niche or phylogeny are clustered within metabolically similar taxa. As there is a higher probability for genes to be shared between closely related taxa or those within an ecological niche [132, 171], such a network distribution would facilitate rapid uptake of HGT-linked genes amongst the full community (even if present at low saturation), providing a buffer against stressors and thus promoting community survival and stability.

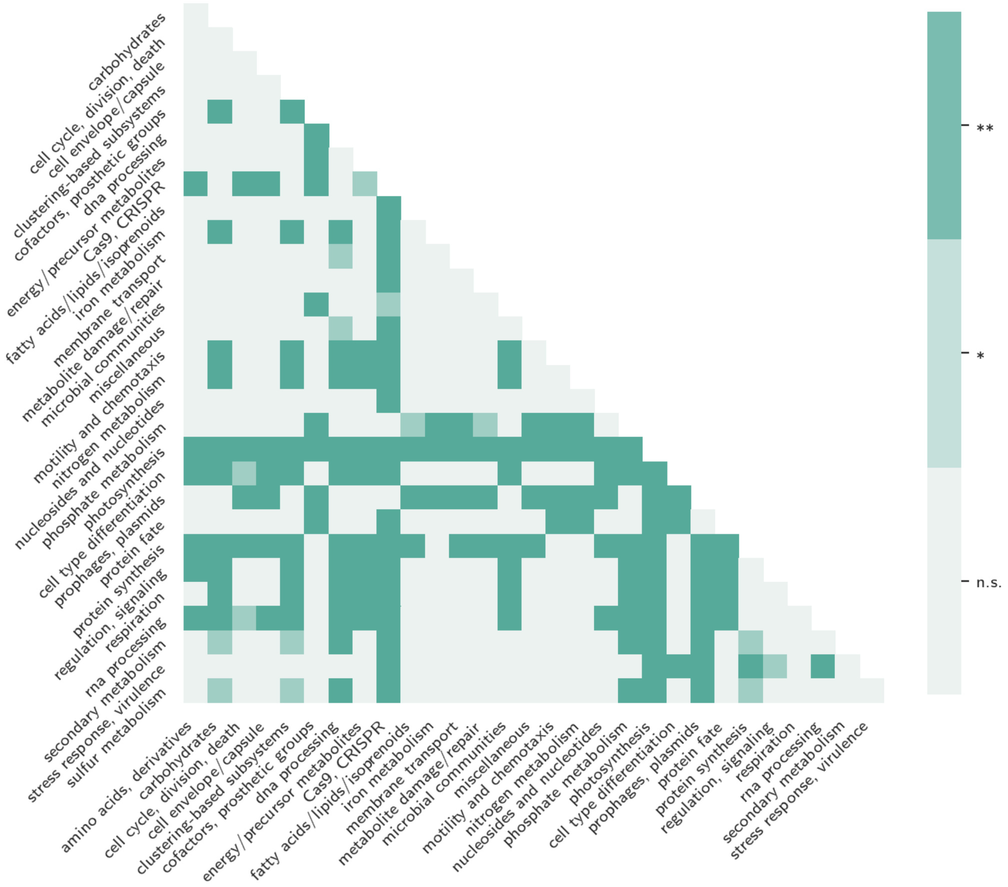


Figure 6.21: Heatmap showing significant pairwise p-values of differences in local node clustering values for different functional classes. Here, 'n.s.' denotes not significant, and '*', '**' denote significance at a threshold of $\alpha = 0.05$ and 0.01 respectively. Results failing the FDR q-value test are assigned as 'n.s.'.

6.4 Conclusions

Here, through the use of network-based methods, I analyse the joint-taxa function distributions of natural microbial communities and identify unifying principles which define how functional diversity is shared amongst taxa. Multiple levels of metabolic organisation in the community are revealed through macro and local network properties. Highly consistent degree distributions are observed across all biomes, with exponential behaviour in the taxa degree distribution and heavy tailed behaviour being observed for the functional degree distribution. At the highest level of community organisation, the 'tail' portion of the functional degree distribution captures common functions associated with housekeeping genes. At the next level, biological communities can be partitioned into ecological niches based upon emergent community structure in the networks, evidenced both by the singular value distributions and average clustering values; such communities indicate metabolic niche redundancy, with large groups of taxa sharing similar sets of functions. At the lowest level of organisation, functions associated with HGT are found to be randomly scattered amongst metabolically diverse taxa, which would increase the speed at which those functions may be taken up by the community should it undergo a disturbance. This underlying pattern appears to scale proportionally with the diversity of the biome [115, 148, 177, 185, 186], and is a universal property across all communities and environments.

With the added complexity that genome plasticity creates when studying these systems, network methods present an unparalleled opportunity to uncover new ecological insights, a conclusion also drawn in prior work [200, 204]. However, challenges remain. The null results for the mean-field scores of NODF and FR, which could ultimately be explained by the degree distributions alone, highlight the difficulty of identifying potential sources of error in such analyses. For example, degree (or other) correlations in network metrics make it notoriously difficult to identify small scale topological features driving macro scale statistics [202]. Furthermore, managing potential bias in the data and bioinformatics methods used to construct the networks is nontrivial; efforts to systematically address biases in prokaryotic taxonomy are a recent development and the quality of reference datasets will continue to improve over time [119, 209]. Despite these caveats, the nature of network methods chosen here mean that my main results would be robust to the most likely sources of data-driven error, such as taxonomic misclassification to closely related species: as the functions would still sit within the same network neighborhood, the analyses would have qualitatively similar results. However, within extreme environments where HGT processes are rare, such as hot springs, it may be possible for this otherwise ubiquitous taxa-gene distribution to break down, and I leave exploration of this question open to future work.

My findings provide a new perspective on redundancy structures in microbial communities, and quantitative support for the hypothesis that phage-infected bacteria may play a beneficial role in the community [248]. Shotgun metagenomics has increasingly shown the importance of phages as reservoirs of prokaryotic genes and their more subtle role in microbial dynamics than as predators or parasites alone [147]. With as many as 60% of cells being infected by lysogenic phage at any one time, viruses are central in shaping the genetic fate of these communities [248]. Cells infected with lysogenic phage vertically transmit genetic material, and if there is a temperate-lytic switch, one such cell may result in thousands of viruses carrying packets of genetic information being sent into the environment for uptake. An estimated 85% of HGT events occur through viral transduction and, whilst the majority of phage are in a lysogenic state the majority of the time, disturbed communities undergo significant increases in lytic viral activity [132]. It therefore follows that the network topology revealed in this chapter not only supports the proposition that there are optimal distributions across the community to effect rapid uptake of HGT genes, but also supports the conjecture that viruses are one key mechanism driving the emergent stability of natural communities due to their central role in HGT processes [132, 249, 250]. Indeed, despite general consensus that microbial assemblages likely operate under consistent assembly rules, quantifying those rules has remained a long lasting challenge [123, 227, 251]. It seems likely that bottom-up and top-down conceptual approaches, in conjunction with a combination of data-driven and modelling methods, may be needed to generate hypotheses for further experimental work. Here, using network methods to interrogate paired taxa-function behaviour in these communities not only reveals new insight, but provides a rich and flexible framework for exploring the fundamental processes which govern microbial dynamics in future.

Chapter 7

Conclusions and future work

In this thesis, I explored the mechanistic drivers behind emergent properties of multispecies communities to identify universal principles governing their structure. I began in Chapter 2 by linking metabolic theory with the classical Rosenzweig-MacArthur differential equations, and produced a more parsimonious allometric setting than in prior work by eliminating the prey size-scaling dependency. Through paramaterising the system with empirical values, it was shown that the model dynamics and equilibria aligned closely to biological observations. Counter to previously held assumptions regarding the explanatory power of minimal allometric models, it was found that scaling of the period and amplitude of population cycling, along with size-abundance scaling, were an excellent match to distributions found in large scale terrestrial surveys.

The results concerning the amplitude of the limit cycles were dependent on simulation, leading to a natural question, *what is the analytic solution for the cycling amplitude in the Rosenzweig-MacArthur ODEs*? Whilst some existing work on limit cycle amplitude in Lokta-Volterra type systems exists, firstly from 1975 in [58] and more recently in [59], there are restrictions on the relationships between parameters. A derivation for an assumption-free general case will be required to fully understand the scaling behaviour of the limit cycle amplitude and form a more complete picture of the strengths and weaknesses of the model, a task left for future work. Despite this caveat, I propose that similar minimal model approaches as taken in Chapter 2 may be useful in food web or trophic modelling by helping reduce the number of parameters and thus assist in managing overfitting.

In Chapter 3, I shifted from size-abundance distributions in terrestrial to marine ecosystems, and examined scaling across 15,000 data points ranging in size from viruses to blue whales. It was demonstrated that a structural break in the exponent at 0.1m could not be explained by anthropogenic pressure, but was the result of turbulent dispersal increasing metabolic demand on large organisms, which in turn reduced abundances. Whilst the effects of the physical parameter of temperature is commonly considered in shaping ecosystem level properties, the role of turbulence has largely been restricted to microscale, localised processes [22, 159, 252]. Through extending my minimal allometric model to include the cost of locomotion and foraging in turbulent environments, it was shown how the physics of fluids are constraining biological systems at the scale of the global biosphere.

Following an exploration of how abundance diversity is structured by size, I investigated how metabolic diversity is structured by taxonomy in microbial systems. Cell size places a hard physical constraint on the number of genes single microbe can carry, leading to an evolutionary tradeoff between survival and reproductive cost; how this functional diversity is distributed amongst taxa in turn impacts the resilience of the microbial community. Through using a network representation of taxa and their gene functions, in Chapter 5 large scale metabolic organisation was identified across the prokaryotic tree of life which mirrored organism phylogeny and their ecological niche. Unlike current phylogenomics methods, which are limited to marker genes from approximately 10,000 genomes due to computational expense [253], my network-based analysis would easily scale to hundreds of thousands of genomes. Furthermore, the network framework introduced may allow for future research to define precise definitions for presently qualitative descriptors, such as, *what is a niche gene?*, or *which prokaryotes share ecological niches?*.

The network approach was then extended to examine the joint taxa-function distributions of real-world microbial communities. With functional and taxonomic profiles of WGS data usually being studied in isolation, my novel analysis methods explicitly linking taxa and function resolved a long standing difficulty in microbial ecology. Furthermore, as my networks were generated from empirical WGS data, they reflect the ground truth of community structure. This marked a significant departure from previous work attempting to reconcile taxa-function behaviour: whilst multiple analysis methodologies have been utilised, all have instead relied on predicting functions based on reference genomes [153, 187, 190, 192–194]. Through breaking down the network distributions of different functional classes, it was shown that HGT associated genes were spread across diverse taxa, effecting a redundancy structure which would promote community survival in the event of disturbance. This metabolic organisation was universal across 248 metagenomes sourced from environmental and host-associated biomes across the globe, and could thus be a key mechanism behind the emergent stability of microbiomes.

This hypothesis could be tested experimentally in future work. Through taking WGS samples of a disturbed community over time - for example, by adding antibiotics to mesocosms - it should be possible to recover the mobile elements and track their distribution through the system. It would also be feasible to explore this through modelling. A natural way to encode this would be as a Markov process. The goal would be to identify the configurations of the system and different functions which provide the optimal balance between streamlining, i.e. minimising the cost to an organism, and maximising the public good, i.e. ensuring a gene is present and easily accessible by the community. It would be possible to provide a series of rules linking gene type, saturation, and competitive processes within the community, and examine the genes' diffusion capacity on the network, to probe whether there is emergent community stability under deletions, insertions, or disturbance.

Whilst multispecies communities have many moving parts and complex interactions, they also display emergent properties. I argue that modelling and data driven approaches are crucial to identify and explore plausible mechanisms for ubiquitous behaviours; mechanisms which may then be interrogated through experimental work. Whether we seek to predict ecosystem tipping points, manipulate a microbiome to improve host health, or achieve one of a myriad of other outcomes reliant on community ecology 'rules', it seems that an understanding of the drivers behind universal ecological phenomena will be vital to our future success.

Bibliography

[1] Lawton, J.H. Are there general laws in ecology? *Oikos*, pages 177–192, 1999.

[2] Barabási, A.L. and Albert, R. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[3] Lange, M. Ecological laws: what would they be and why would they matter? *Oikos*, 110(2):394–403, 2005.

[4] West, G.B., Brown, J.H. and Enquist, B.J. The fourth dimension of life: fractal geometry and allometric scaling of organisms. *Science*, 284(5420):1677–1679, 1999.

[5] Currie, D.J. Where newton might have taken ecology. *Global Ecology and Biogeography*, 28(1):18–27, 2019.

[6] Chave, J. The problem of pattern and scale in ecology: what have we learned in 20 years? *Ecology letters*, 16:4–16, 2013.

[7] Linquist, S., Gregory, T.R., Elliott, T.A. et al. Yes! there are resilient generalizations (or "laws") in ecology. *The Quarterly review of biology*, 91(2):119–131, 2016.

[8] Rubner, M. Ueber den einfluss der korpergrosse auf stoffund kaftwechsel. *Zeitschrift fur Biologie*, 19:535–562, 1883.

[9] Kleiber, M. Body size and metabolism. *ENE*, 1:E9, 1932.

[10] Schmidt-Nielsen, K. *Scaling: why is animal size so important?* Cambridge University Press, 1984.

[11] Blanchard, J.L., Heneghan, R.F., Everett, J.D. et al. From bacteria to whales: Using functional size spectra to model marine ecosystems. *Trends in Ecology & Evolution*, 2017.

[12] West, G.B. and Brown, J.H. The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization. *Journal of experimental biology*, 208(9):1575–1592, 2005.

[13] Pawar, S., Dell, A.I. and Savage, V.M. Dimensionality of consumer search space drives trophic interaction strengths. *Nature*, 486(7404):485–489, 2012.

[14] Bale, R., Hao, M., Bhalla, A.P.S. et al. Energy efficiency and allometry of movement of swimming and flying animals. *Proceedings of the National Academy of Sciences*, 111(21):7517–7521, 2014.

[15] Reiss, M.J. The allometry of reproduction: why larger species invest relatively less in their offspring. *Journal of Theoretical Biology*, 1985.

[16] Sims, D.W., Southall, E.J., Humphries, N.E. et al. Scaling laws of marine predator search behaviour. *Nature*, 451(7182):1098–1102, 2008.

[17] Haskell, J.P., Ritchie, M.E. and Olff, H. Fractal geometry predicts varying body size scaling relationships for mammal and bird home ranges. *Nature*, 418(6897): 527–530, August 2002.

[18] Hedenström, A. Scaling migration speed in animals that run, swim and fly. *Journal of Zoology*, 259(2):155–160, February 2003.

[19] Mitchell, J.G. The energetics and scaling of search strategies in bacteria. *The American Naturalist*, 160(6):727–740, 2002.

[20] Damuth, J. Home range, home range overlap, and species energy use among herbivorous mammals. *Biological Journal of the Linnean Society*, 15(3):185–193, 1981.

[21] White, E.P., Ernest, S.M., Kerkhoff, A.J. et al. Relationships between body size and abundance in ecology. *Trends in ecology & evolution*, 22(6):323–330, 2007.

[22] Bernhardt, J.R., Sunday, J.M. and O'Connor, M.I. Metabolic theory and the temperature-size rule explain the temperature dependence of population carrying capacity. *The American Naturalist*, 192(6):687–697, 2018.

[23] Yodzis, P. and Innes, S. Body size and consumer-resource dynamics. *The American Naturalist*, 139(6):1151–1175, 1992.

[24] Weitz, J.S. and Levin, S.A. Size and scaling of predator–prey dynamics. *Ecology letters*, 9(5):548–557, 2006.

[25] Eilersen, A. and Sneppen, K. Applying allometric scaling to predator-prey systems. *Physical Review E*, 99(2):022405, 2019.

[26] Allen, A.P., Brown, J.H. and Gillooly, J.F. Global biodiversity, biochemical kinetics, and the energetic-equivalence rule. *Science*, 297(5586):1545–1548, 2002.

[27] DeLong, J.P. and Vasseur, D.A. Size-density scaling in protists and the links between consumer-resource interaction parameters. *Journal of Animal Ecology*, 81(6):1193–1201, July 2012.

[28] Damuth, J. A macroevolutionary explanation for energy equivalence in the scaling of body size and population density. *The American Naturalist*, 169(5):621–631, 2007.

[29] Malerba, M.E. and Marshall, D.J. Size-abundance rules? Evolution changes scaling relationships between size, metabolism and demography. *Ecology Letters*, 22(9):1407–1416, July 2019.

[30] Marquet, P.A., Navarrete, S.A. and Castilla, J.C. Body size, population density, and the energetic equivalence rule. *Journal of Animal Ecology*, pages 325–332, 1995.

[31] Isaac, N.J., Storch, D. and Carbone, C. The paradox of energy equivalence. *Global Ecology and Biogeography*, 22(1):1–5, 2013.

[32] Savage, V.M., Gillooly, J.F., Woodruff, W.H. et al. The predominance of quarter-power scaling in biology. *Functional Ecology*, 18(2):257–282, April 2004.

[33] Clarke, A. and Johnston, N.M. Scaling of metabolic rate with body mass and temperature in teleost fish. *Journal of Animal Ecology*, 68(5):893–905, 1999.

[34] Yool, A., Martin, A.P., Anderson, T.R. et al. Big in the benthos: Future change of seafloor community biomass in a global, body size-resolved model. *Global Change Biology*, 2017.

[35] Kalinkat, G., Schneider, F.D., Digel, C. et al. Body masses, functional responses and predator-prey stability. *Ecology letters*, 16(9):1126–1134, 2013.

[36] Rall, B.C., Brose, U., Hartvig, M. et al. Universal temperature and body-mass scaling of feeding rates. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1605):2923–2934, 2012.

[37] Hatton, I.A., Dobson, A.P., Storch, D. et al. Linking scaling laws across eukaryotes. *Proceedings of the National Academy of Sciences*, 116(43):21616–21622, 2019.

[38] Lewis, N.D., Breckels, M.N., Archer, S.D. et al. Grazing-induced production of dms can stabilize food-web dynamics and promote the formation of phytoplankton blooms in a multitrophic plankton model. *Biogeochemistry*, 110(1-3): 303–313, 2012.

[39] Osmond, M.M., Barbour, M.A., Bernhardt, J.R. et al. Warming-induced changes to body size stabilize consumer-resource dynamics. *The American Naturalist*, 189 (6):718–725, 2017.

[40] Hendriks, A.J. and Mulder, C. Delayed logistic and rosenzweig–macarthur models with allometric parameter setting estimate population cycles at lower trophic levels well. *Ecological complexity*, 9:43–54, 2012.

[41] Brose, U., Williams, R.J. and Martinez, N.D. Allometric scaling enhances stability in complex food webs. *Ecology letters*, 9(11):1228–1236, 2006.

[42] DeLong, J.P., Gilbert, B., Shurin, J.B. et al. The body size dependence of trophic cascades. *The American Naturalist*, 185(3):354–366, 2015.

[43] Rosenzweig, M.L. and MacArthur, R.H. Graphical representation and stability conditions of predator-prey interactions. *The American Naturalist*, 97(895):209–223, 1963.

[44] Cheng, K.S., Hsu, S.B. and Lin, S.S. Some results on global stability of a predator-prey system. *Journal of Mathematical Biology*, 12(1):115–126, 1982.

[45] Hsu, S.B., Hwang, T.W. and Kuang, Y. Global analysis of the michaelis–menten-type ratio-dependent predator-prey system. *Journal of mathematical biology*, 42 (6):489–506, 2001.

[46] Hsu, S.B. and Shi, J. Relaxation oscillation profile of limit cycle in predator-prey system. *Discrete & Continuous Dynamical Systems-B*, 11(4):893, 2009.

[47] Vucic-Pestic, O., Rall, B.C., Kalinkat, G. et al. Allometric functional response model: body masses constrain interaction strengths. *Journal of Animal Ecology*, 79(1):249–256, 2010.

[48] Huffaker, C. et al. Experimental studies on predation: dispersion factors and predator-prey oscillations. *Hilgardia*, 27(14):343–383, 1958.

[49] Luckinbill, L.S. Coexistence in laboratory populations of paramecium aurelia and its predator didinium nasutum. *Ecology*, 54(6):1320–1327, 1973.

[50] Balagaddé, F.K., Song, H., Ozaki, J. et al. A synthetic escherichia coli predator–prey ecosystem. *Molecular systems biology*, 4(1):187, 2008.

[51] Blasius, B., Rudolf, L., Weithoff, G. et al. Long-term cyclic persistence in an experimental predator–prey system. *Nature*, 577(7789):226–230, 2020.

[52] Trpis, M. Interaction between the predator toxorhynchites brevipalpis and its prey aedes aegypti. *Bulletin of the World Health Organization*, 49(4):359, 1973.

[53] Utida, S. Cyclic fluctuations of population density intrinsic to the host-parasite system. *Ecology*, 38(3):442–449, 1957.

[54] Goulden, C.E. and Hornig, L.L. Population oscillations and energy reserves in planktonic cladocera and their consequences to competition. *Proceedings of the National Academy of Sciences*, 77(3):1716–1720, 1980.

[55] Sudo, R., Kobayashi, K. and Aiba, S. Some experiments and analysis of a predator-prey model: Interaction between colpidium campylum and alcaligenes faecalis in continuous and mixed culture. *Biotechnology and Bioengineering*, 17 (2):167–184, 1975.

[56] Rohatgi, A. Webplotdigitizer, 2017.

[57] White, E.P., Ernest, S.M., Kerkhoff, A.J. et al. Relationships between body size and abundance in ecology. *Trends in ecology & evolution*, 22(6):323–330, 2007.

[58] De Angelis, D.L. Estimates of predator-prey limit cycles. *Bulletin of Mathematical Biology*, 37:291–299, 1975.

[59] Lundström, N.L. and Söderbacka, G. Estimates of size of cycle in a predator-prey system. *Differential Equations and Dynamical Systems*, pages 1–29, 2018.

[60] Khalil, H.K. *Noninear Systems*. Prentice-Hall, New Jersey, 1996.

[61] Brose, U., Jonsson, T., Berlow, E.L. et al. Consumer–resource body-size relationships in natural food webs. *Ecology*, 87(10):2411–2417, 2006.

[62] Hatton, I.A., McCann, K.S., Fryxell, J.M. et al. The predator-prey power law: Biomass scaling across terrestrial and aquatic biomes. *Science*, 349(6252), 2015.

[63] Gilbert, B., Tunney, T.D., McCann, K.S. et al. A bioenergetic framework for the temperature dependence of trophic interactions. *Ecology Letters*, 17(8):902–914, 2014.

[64] Santini, L. and Isaac, N.J. Rapid anthropocene realignment of allometric scaling rules. *Ecology Letters*, 24(7):1318–1327, 2021.

[65] Seymour, J.R., Seuront, L. and Mitchell, J.G. Microscale and small-scale temporal dynamics of a coastal planktonic microbial community. *Marine Ecology Progress Series*, 300:21–37, 2005.

[66] Waters, R.L., Mitchell, J.G. and Seymour, J. Geostatistical characterisation of centimetre-scale spatial structure of in vivo fluorescence. *Marine Ecology Progress Series*, 251:49–58, 2003.

[67] Eagon, R.G. Pseudomonas natriegens, a marine bacterium with a generation time of less than 10 minutes. *Journal of bacteriology*, 83(4):736–737, 1962.

[68] Lowenstein, T.K., Schubert, B.A. and Timofeeff, M.N. Microbial communities in fluid inclusions and long-term survival in halite. *GSA Today*, 21(1):4–9, 2011.

[69] Lavery, T.J., Roudnew, B., Seymour, J. et al. Whales sustain fisheries: blue whales stimulate primary production in the southern ocean. *Marine Mammal Science*, 30(3):888–904, 2014.

[70] Garcia, S., Kolding, J., Rice, J. et al. Reconsidering the consequences of selective fisheries. *Science*, 335(6072):1045–1047, 2012.

[71] Andersen, K.H., Berge, T., Gonçalves, R. et al. Characteristic sizes of life in the oceans, from bacteria to whales. *Annual review of marine science*, 8:217–241, 2016.

[72] Graham, N., Dulvy, N., Jennings, S. et al. Size-spectra as indicators of the effects of fishing on coral reef fish assemblages. *Coral Reefs*, 24(1):118–124, 2005.

[73] IMOS, 2020. URL `http://www.marine.csiro.au/marq/edd_search.Browse_Citation?txtSession=9012`.

[74] Ibarbalz, F.M., Henry, N., Brandão, M.C. et al. Global trends in marine plankton diversity across kingdoms of life. *Cell*, 179(5):1084–1097. e21, 2019. ISSN 0092-8674.

[75] 2019. URL `https://doi.org/10.1594/PANGAEA.904397`.

[76] 2012. URL `https://doi.org/10.1594/PANGAEA.777384`.

[77] Barneche, D., Kulbicki, M., Floeter, S.R. et al. Energetic and ecological constraints on population density of reef fishes. *Proceedings of the Royal Society B: Biological Sciences*, 283(1823):20152186, 2016. ISSN 0962-8452.

[78] WoRMS, May 2020. URL `https://www.marinespecies.org`.

[79] Pauly, D., May 2019. URL `www.fishbase.org`.

[80] Barrios-O'Neill, D., Kelly, R. and Emmerson, M.C. Biomass encounter rates limit the size scaling of feeding interactions. *Ecology letters*, 22(11):1870–1878, 2019.

[81] Brose, U., Archambault, P., Barnes, A.D. et al. Predator traits determine food-web architecture across ecosystems. *Nature ecology & evolution*, 3(6):919–927, 2019.

[82] Politis, D.N., Romano, J.P. and Wolf, M. *Subsampling*. Springer Science Business Media, 1999. ISBN 0387988548.

[83] Romano, J.P. and Wolf, M. Subsampling intervals in autoregressive models with linear time trend. *Econometrica*, 69(5):1283–1314, 2001. ISSN 0012-9682.

[84] Lotze, H.K. and Worm, B. Historical baselines for large marine animals. *Trends in ecology & evolution*, 24(5):254–262, 2009.

[85] Williams, I.D., Baum, J.K., Heenan, A. et al. Human, oceanographic and habitat drivers of central and western pacific coral reef fish assemblages. *PLoS One*, 10 (4), 2015.

[86] Gazzola, M., Argentina, M. and Mahadevan, L. Scaling macroscopic aquatic locomotion. *Nature Physics*, 10(10):758–761, 2014.

[87] Strutton, P.G., Mitchell, J.G., Parslow, J.S. et al. Phytoplankton patchiness: quantifying the biological contribution using fast repetition rate fluorometry. *Journal of Plankton Research*, 19(9):1265–1274, 1997.

[88] Smriga, S., Fernandez, V.I., Mitchell, J.G. et al. Chemotaxis toward phytoplankton drives organic matter partitioning among marine bacteria. *Proceedings of the National Academy of Sciences*, 113(6):1576–1581, 2016.

[89] Jiang, H. and Kiørboe, T. The fluid dynamics of swimming by jumping in copepods. *Journal of the Royal Society Interface*, 8(61):1090–1103, 2011. ISSN 1742-5689.

[90] Lehahn, Y., d'Ovidio, F., Lévy, M. et al. Long range transport of a quasi isolated chlorophyll patch by an agulhas ring. *Geophysical Research Letters*, 38(16), 2011. ISSN 0094-8276.

[91] Condie, S. and Condie, R. Retention of plankton within ocean eddies. *Global Ecology and Biogeography*, 25(10):1264–1277, 2016. ISSN 1466-822X.

[92] Vortmeyer-Kley, R., Lünsmann, B., Berthold, M. et al. Eddies: Fluid dynamical niches or transporters?–a case study in the western baltic sea. *Frontiers in Marine Science*, 6:118, 2019.

[93] Goldbogen, J.A., Cade, D.E., Wisniewska, D.M. et al. Why whales are big but not bigger: Physiological drivers and ecological limits in the age of ocean giants. *Science*, 366(6471):1367–1372, December 2019.

[94] Speers-Roesch, B., Norin, T. and Driedzic, W.R. The benefit of being still: energy savings during winter dormancy in fish come from inactivity and the cold, not from metabolic rate depression. *Proceedings of the Royal Society B: Biological Sciences*, 285(1886):20181593–10, September 2018.

[95] Kolmogorov, A.N. The local structure of turbulence in incompressible viscous fluid for very large reynolds numbers [in russian]. In *Dokl. Akad. Nauk SSSR*, volume 30, pages 299–303, 1941.

[96] Tennekes, H. and Lumley, J.L. *A first course in turbulence.* MIT press, 2018.

[97] Yamazaki, H., Mitchell, J.G., Seuront, L. et al. Phytoplankton microstructure in fully developed oceanic turbulence. *Geophysical research letters*, 33(1), 2006.

[98] Bejan, A. and Marden, J.H. Unifying constructal theory for scale effects in running, swimming and flying. *Journal of Experimental Biology*, 209(2):238–248, 2006.

[99] Goldbogen, J.A., Calambokidis, J., Croll, D.A. et al. Scaling of lunge-feeding performance in rorqual whales: mass-specific energy expenditure increases with body size and progressively limits diving capacity. *Functional Ecology*, 26(1): 216–226, 2012.

[100] Savage, V.M., Gillooly, J.F., American, J.B.T. et al. Effects of body size and temperature on population growth. *journals.uchicago.edu*, 163(3):429–441, March 2004.

[101] Partensky, F., Hess, W.R. and Vaulot, D. Prochlorococcus, a marine photosynthetic prokaryote of global significance. *Microbiology and molecular biology reviews*, 63(1):106–127, 1999.

[102] Nilsen, E.B., Finstad, A.G., Næsje, T.F. et al. Using mass scaling of movement cost and resource encounter rate to predict animal body size–Population density relationships. *Theoretical Population Biology*, 86:23–28, June 2013.

[103] Bainbridge, R. The speed of swimming of fish as related to size and to the frequency and amplitude of the tail beat. *Journal of experimental biology*, 35(1): 109–133, 1958.

[104] Parsons, T.R. and Chen, Y. Estimates of trophic efficiency, based on the size distribution of phytoplankton and fish in different environments. *Zool. Stud*, 33: 296–301, 1994.

[105] Olden, J.D., Hogan, Z.S. and Zanden, M.J.V. Small fish, big fish, red fish, blue fish: size-biased extinction risk of the world's freshwater and marine fishes. *Global Ecology and Biogeography*, 16(6):694–701, November 2007.

[106] Heino, M., Díaz Pauli, B. and Dieckmann, U. Fisheries-Induced Evolution. *Annual Review of Ecology, Evolution, and Systematics*, 46(1):461–480, December 2015.

[107] Jennings, S. and Blanchard, J.L. Fish abundance with no fishing: predictions based on macroecological theory. *Journal of Animal Ecology*, 73(4):632–642, 2004.

[108] Bhatia, K., Vecchi, G., Murakami, H. et al. Projected response of tropical cyclone intensity and intensification in a global climate model. *Journal of Climate*, 31 (20):8281–8303, 2018.

[109] Silva, A.T., Katopodis, C., Santos, J.M. et al. Cyprinid swimming behaviour in response to turbulent flow. *Ecological Engineering*, 44:314–328, 2012.

[110] Cabré, A., Marinov, I. and Leung, S. Consistent global responses of marine ecosystems to future climate change across the ipcc ar5 earth system models. *Climate Dynamics*, 45(5-6):1253–1280, 2015.

[111] McHenry, J., Welch, H., Lester, S.E. et al. Projecting marine species range shifts from only temperature can mask climate vulnerability. *Global change biology*, 2019.

[112] Falkowski, P.G., Fenchel, T. and Delong, E.F. The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, 320(5879):1034–1039, May 2008.

[113] Locey, K.J. and Lennon, J.T. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, 113(21):5970–5975, 2016.

[114] Bolyen, E., Rideout, J.R., Dillon, M.R. et al. Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature biotechnology*, 37 (8):852–857, 2019.

[115] Thompson, L.R., Sanders, J.G., McDonald, D. et al. A communal catalogue reveals earth's multiscale microbial diversity. *Nature*, 551(7681):457–463, 2017.

[116] Lloyd-Price, J., Abu-Ali, G. and Huttenhower, C. The healthy human microbiome. *Genome medicine*, 8(1):1–11, 2016.

[117] Dann, L.M., McKerral, J.C., Smith, R.J. et al. Microbial micropatches within microbial hotspots. *PloS one*, 13(5), 2018.

[118] Dann, L.M., Smith, R.J., Tobe, S.S. et al. Microscale distributions of freshwater planktonic viruses and prokaryotes are patchy and taxonomically distinct. *Aquatic Microbial Ecology*, 77(2):65–77, 2016.

[119] Parks, D.H., Chuvochina, M., Waite, D.W. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnology*, 2018.

[120] Sharma, R., Ranjan, R., Kapardar, R.K. et al. Unculturable bacterial diversity: An untapped resource. *Current Science*, pages 72–77, 2005.

[121] Pande, S. and Kost, C. Bacterial unculturability and the formation of intercellular metabolic networks. *Trends in microbiology*, 25(5):349–361, 2017.

[122] Hug, L.A. Sizing up the uncultured microbial majority. *MSystems*, 3(5), 2018.

[123] Goldford, J.E., Lu, N., Bajić, D. et al. Emergent simplicity in microbial community assembly. *Science*, 361(6401):469–474, August 2018.

[124] Kamneva, O.K. Genome composition and phylogeny of microbes predict their co-occurrence in the environment. *PLOS Computational Biology*, 13(2):e1005366–20, February 2017.

[125] Le Roux, X., Recous, S. and Attard, E. 17 soil microbial diversity in grasslands and its importance for grassland functioning. *Grassland Productivity and Ecosystem Services*, page 158, 2011.

[126] Cano, R.J. and Borucki, M.K. Revival and identification of bacterial spores in 25-to 40-million-year-old dominican amber. *Science*, 268(5213):1060–1064, 1995.

[127] Vreeland, R.H., Rosenzweig, W.D. and Powers, D.W. Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. *Nature*, 407(6806): 897–900, 2000.

[128] Morono, Y., Ito, M., Hoshino, T. et al. Aerobic microbial life persists in oxic marine sediment as old as 101.5 million years. *Nature Communications*, 11(1): 1–9, 2020.

[129] Dubnau, D. and Blokesch, M. Mechanisms of dna uptake by naturally competent bacteria. *Annual review of genetics*, 53:217–237, 2019.

[130] Weinert, L.A. and Welch, J.J. Why might bacterial pathogens have small genomes? *Trends in ecology & evolution*, 32(12):936–947, 2017.

[131] Sousa, S.A., Feliciano, J.R., Pita, T. et al. Burkholderia cepacia complex regulation of virulence gene expression: a review. *Genes*, 8(1):43, 2017.

[132] Chen, J., Quiles-Puchalt, N., Chiang, Y.N. et al. Genome hypermobility by lateral transduction. *Science*, 362(6):207–212, October 2018.

[133] Gilbert, J.A., Steele, J.A., Caporaso, J.G. et al. Defining seasonal marine microbial community dynamics. *The ISME Journal*, 6(2):298–308, August 2011.

[134] De Vrieze, J., De Mulder, T., Matassa, S. et al. Stochasticity in microbiology: managing unpredictability to reach the sustainable development goals. *Microbial Biotechnology*, 2020.

[135] Maynard, D.S., Serván, C.A., Capitán, J.A. et al. Phenotypic variability promotes diversity and stability in competitive communities. *Ecology letters*, 22(11): 1776–1786, 2019.

[136] Kurm, V., van der Putten, W.H., de Boer, W. et al. Low abundant soil bacteria can be metabolically versatile and fast growing. *Ecology*, 98(2):555–564, February 2017.

[137] Botton, S., Van Heusden, M., Parsons, J. et al. Resilience of microbial systems towards disturbances. *Critical reviews in microbiology*, 32(2):101–112, 2006.

[138] Shade, A., Peter, H., Allison, S.D. et al. Fundamentals of microbial community resistance and resilience. *Frontiers in microbiology*, 3:417, 2012.

[139] Greenhalgh, K., Meyer, K.M., Aagaard, K.M. et al. The human gut microbiome in health: establishment and resilience of microbiota over a lifetime. *Environmental microbiology*, 18(7):2103–2116, 2016.

[140] Mehta, R.S., Abu-Ali, G.S., Drew, D.A. et al. Stability of the human faecal microbiome in a cohort of adult men. *Nature Microbiology*, pages 1–12, February 2018.

[141] Fuhrman, J.A., Cram, J.A. and Needham, D.M. Marine microbial community dynamics and their ecological interpretation. *Nature Publishing Group*, 13(3): 133–146, February 2015.

[142] Schindler, D.E., Armstrong, J.B. and Reed, T.E. The portfolio concept in ecology and evolution. *Frontiers in Ecology and the Environment*, 13(5):257–263, 2015.

[143] Consortium, T.H.M.P. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, June 2012.

[144] Franzosa, E.A., Huang, K., Meadow, J.F. et al. Identifying personal microbiomes using metagenomic codes. *Proceedings of the National Academy of Sciences*, 112 (22):E2930–E2938, 2015.

[145] Yatsunenko, T., Rey, F.E., Manary, M.J. et al. Human gut microbiome viewed across age and geography. *nature*, 486(7402):222–227, 2012.

[146] Fierer, N. and Jackson, R.B. The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences*, 103(3):626–631, 2006.

[147] Dinsdale, E.A., Edwards, R.A., Hall, D. et al. Functional metagenomic profiling of nine biomes. *Nature*, 452(7187):629–632, 2008.

[148] Sunagawa, S., Coelho, L.P., Chaffron, S. et al. Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359–1261359, May 2015.

[149] De Boeck, H.J., Bloor, J.M., Kreyling, J. et al. Patterns and drivers of biodiversity–stability relationships under climate extremes. *Journal of Ecology*, 106(3):890–902, 2018.

[150] Bardgett, R.D. and Caruso, T. Soil microbial community responses to climate extremes: resistance, resilience and transitions to alternative states. *Philosophical Transactions of the Royal Society B*, 375(1794):20190112, 2020.

[151] Yang, G., Wagg, C., Veresoglou, S.D. et al. How soil biota drive ecosystem stability. *Trends in plant science*, 23(12):1057–1067, 2018.

[152] Louca, S., Polz, M.F., Mazel, F. et al. Function and functional redundancy in microbial systems. *Nature Ecology & Evolution*, pages 1–8, April 2018.

[153] Miki, T., Yokokawa, T. and Matsui, K. Biodiversity and multifunctionality in a microbial community: a novel theoretical approach to quantify functional redundancy. *Proceedings of the Royal Society B: Biological Sciences*, 281(1776): 20132498–20132498, December 2013.

[154] Dworkin, J. and Shah, I.M. Exit from dormancy in microbial organisms. *Nature reviews microbiology*, 8(12):890–896, 2010.

[155] Franklin, R.B. and Mills, A.L. Multi-scale variation in spatial heterogeneity for microbial community structure in an eastern virginia agricultural field. *FEMS microbiology ecology*, 44(3):335–346, 2003.

[156] Porter, S.S. and Rice, K.J. Trade-offs, spatial heterogeneity, and the maintenance of microbial diversity. *Evolution: International Journal of Organic Evolution*, 67 (2):599–608, 2013.

[157] Stocker, R. and Seymour, J.R. Ecology and physics of bacterial chemotaxis in the ocean. *Microbiol. Mol. Biol. Rev.*, 76(4):792–812, 2012.

[158] Smith, N.W., Shorten, P.R., Altermann, E. et al. The classification and evolution of bacterial cross-feeding. *Frontiers in Ecology and Evolution*, 7:153, 2019.

[159] Stocker, R. Marine microbes see a sea of gradients. *Science*, 338(6107):628–633, 2012.

[160] Nemergut, D.R., Schmidt, S.K., Fukami, T. et al. Patterns and processes of microbial community assembly. *Microbiol. Mol. Biol. Rev.*, 77(3):342–356, 2013.

[161] Lindström, E.S. and Östman, Ö. The importance of dispersal for bacterial community composition and functioning. *PloS one*, 6(10), 2011.

[162] Vila, J.C., Jones, M.L., Patel, M. et al. Uncovering the rules of microbial community invasions. *Nature ecology & evolution*, 3(8):1162–1171, 2019.

[163] Giovannoni, S.J., Thrash, J.C. and Temperton, B. Implications of streamlining theory for microbial ecology. *The ISME journal*, 8(8):1553–1565, 2014.

[164] Gamfeldt, L. and Roger, F. Revisiting the biodiversity–ecosystem multifunctionality relationship. *Nature Ecology & Evolution*, 1:1–7, June 2017.

[165] Whitfield, J. Biogeography: is everything everywhere? researchers have dug up some surprising evidence casting doubt on the long-held belief that microbes are impervious to geographic constraints. *Science*, 310(5750):960–962, 2005.

[166] Thompson, L.R., Haroon, M.F., Shibl, A.A. et al. Red sea sar11 and prochlorococcus single-cell genomes reflect globally distributed pangenomes. *Applied and environmental microbiology*, 85(13), 2019.

[167] Jurburg, S.D. and Salles, J.F. Functional Redundancy and Ecosystem Function — The Soil Microbiota as a Case Study. In *Biodiversity in Ecosystems - Linking Structure and Function*, pages 1–22. InTech, April 2015.

[168] Gillings, M.R. Lateral gene transfer, bacterial genome evolution, and the Anthropocene. *Annals of the New York Academy of Sciences*, 1389(1):20–36, October 2016.

[169] Polz, M.F., Alm, E.J. and Hanage, W.P. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in Genetics*, 29(3): 170–175, March 2013.

[170] Popa, O., Hazkani-Covo, E., Landan, G. et al. Directed networks reveal genomic barriers and dna repair bypasses to lateral gene transfer among prokaryotes. *Genome research*, 21(4):599–609, 2011.

[171] Smillie, C.S., Smith, M.B., Friedman, J. et al. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*, 480(7376):241–244, November 2011.

[172] Bolotin, E. and Hershberg, R. Horizontally Acquired Genes Are Often Shared between Closely Related Bacterial Species. *Frontiers in Microbiology*, 8:650–10, August 2017.

[173] Fan, Y., Xiao, Y., Momeni, B. et al. Horizontal gene transfer can help maintain the equilibrium of microbial communities. *Journal of theoretical biology*, 454: 53–59, 2018.

[174] Soucy, S.M., Huang, J. and Gogarten, J.P. Horizontal gene transfer: building the web of life. *Nature Publishing Group*, 16(8):472–482, August 2015.

[175] Hug, L.A., Baker, B.J., Anantharaman, K. et al. A new view of the tree of life. *Nature microbiology*, 1(5):16048, 2016.

[176] O'Leary, N.A., Wright, M.W., Brister, J.R. et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733–D745, 2016.

[177] Bahram, M., Hildebrand, F., Forslund, S.K. et al. Structure and function of the global topsoil microbiome. *Nature*, 560(7717):233–237, 2018.

[178] Wattam, A.R., Davis, J.J., Assaf, R. et al. Improvements to patric, the all-bacterial bioinformatics database and analysis resource center. *Nucleic acids research*, 45(D1):D535–D542, 2016.

[179] Sczyrba, A., Hofmann, P., Belmann, P. et al. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nature methods*, 14(11): 1063–1071, 2017.

[180] Truong, D.T., Franzosa, E.A., Tickle, T.L. et al. Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature methods*, 12(10):902–903, 2015.

[181] Altschul, S.F., Gish, W., Miller, W. et al. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[182] Eddy, S.R. et al. Multiple alignment using hidden markov models. In *Ismb*, volume 3, pages 114–120, 1995.

[183] Franzosa, E.A., McIver, L.J., Rahnavard, G. et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nature methods*, 15(11):962, 2018.

[184] Overbeek, R., Olson, R., Pusch, G.D. et al. The seed and the rapid annotation of microbial genomes using subsystems technology (rast). *Nucleic acids research*, 42(D1):D206–D214, 2013.

[185] Xu, J., Zhang, Y., Zhang, P. et al. The structure and function of the global citrus rhizosphere microbiome. *Nature communications*, 9(1):1–10, 2018.

[186] Huttenhower, C., Gevers, D., Knight, R. et al. Structure, function and diversity of the healthy human microbiome. *nature*, 486(7402):207, 2012.

[187] Shan, X. and Cordero, O.X. A fundamental structure-function mapping in the ocean microbiome. *bioRxiv*, 2020.

[188] Nemergut, D.R., Costello, E.K., Meyer, A.F. et al. Structure and function of alpine and arctic soil microbial communities. *Research in microbiology*, 156(7): 775–784, 2005.

[189] Balmonte, J.P., Teske, A. and Arnosti, C. Structure and function of high arctic pelagic, particle-associated and benthic bacterial communities. *Environmental microbiology*, 20(8):2941–2954, 2018.

[190] Zhang, W., Cao, S., Ding, W. et al. Structure and function of the arctic and antarctic marine microbiota as revealed by metagenomics. *Microbiome*, 8:1–12, 2020.

[191] Dinsdale, E.A., Edwards, R.A., Bailey, B. et al. Multivariate analysis of functional metagenomes. *Frontiers in Genetics*, 4:41, 2013.

[192] Tian, L., Wang, X.W., Wu, A.K. et al. Deciphering functional redundancy in the human microbiome. *Nature communications*, 11(1):1–11, 2020.

[193] Eng, A. and Borenstein, E. Taxa-function robustness in microbial communities. *Microbiome*, 6(1):1–19, 2018.

[194] Louca, S., Parfrey, L.W. and Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277, 2016.

[195] Faust, K., Lima-Mendez, G., Lerat, J.S. et al. Cross-biome comparison of microbial association networks. *Frontiers in Microbiology*, 6(219):55–13, October 2015.

[196] Friedman, J. and Alm, E.J. Inferring Correlation Networks from Genomic Survey Data. *PLOS Computational Biology*, 8(9):e1002687, September 2012.

[197] Weiss, S., Van Treuren, W., Lozupone, C. et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME journal*, 10(7):1669–1681, 2016.

[198] Bernard, G., Greenfield, P., Ragan, M.A. et al. K-mer similarity, networks of microbial genomes, and taxonomic rank. *Msystems*, 3(6), 2018.

[199] Acman, M., van Dorp, L., Santini, J.M. et al. Large-scale network analysis captures biological features of bacterial plasmids. *Nature communications*, 11(1): 1–11, 2020.

[200] Halary, S., Leigh, J.W., Cheaib, B. et al. Network analyses structure genetic diversity in independent genetic worlds. *Proceedings of the National Academy of Sciences*, 107(1):127–132, 2010.

[201] Watson, A.K., Lannes, R., Pathmanathan, J.S. et al. The methodology behind network thinking: Graphs to analyze microbial complexity and evolution. In *Evolutionary Genomics*, pages 271–308. Springer, 2019.

[202] Vasques Filho, D. and O'Neale, D.R. Transitivity and degree assortativity explained: The bipartite structure of social networks. *Physical Review E*, 101(5): 052305, 2020.

[203] Latapy, M., Magnien, C. and Del Vecchio, N. Basic notions for the analysis of large two-mode networks. *Social networks*, 30(1):31–48, 2008.

[204] Corel, E., Lopez, P., Méheust, R. et al. Network-thinking: graphs to analyze microbial complexity and evolution. *Trends in Microbiology*, 24(3):224–237, 2016.

[205] Corel, E., Méheust, R., Watson, A.K. et al. Bipartite network analysis of gene sharings in the microbial world. *Molecular biology and evolution*, 35(4):899–913, 2018.

[206] Iranzo, J., Krupovic, M. and Koonin, E.V. The double-stranded dna virosphere as a modular hierarchical network of gene sharing. *MBio*, 7(4), 2016.

[207] Jaffe, A.L., Corel, E., Pathmanathan, J.S. et al. Bipartite graph analyses reveal interdomain lgt involving ultrasmall prokaryotes and their divergent, membrane-related proteins. *Environmental microbiology*, 18(12):5072–5081, 2016.

[208] Payrató-Borras, C., Hernández, L. and Moreno, Y. Breaking the spell of nestedness: The entropic origin of nestedness in mutualistic systems. *Physical Review X*, 9(3):031024, 2019.

[209] Mendler, K., Chen, H., Parks, D.H. et al. Annotree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic acids research*, 47(9): 4442–4448, 2019.

[210] Bharti, R. and Grimm, D.G. Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics*, 2019.

[211] Benson, D.A., Cavanaugh, M., Clark, K. et al. Genbank. *Nucleic acids research*, 41(D1):D36–D42, 2012.

[212] Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30 (14):2068–2069, 2014.

[213] Huerta-Cepas, J., Szklarczyk, D., Forslund, K. et al. eggnog 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic acids research*, 44(D1):D286–D293, 2016.

[214] Kanehisa, M., Araki, M., Goto, S. et al. Kegg for linking genomes to life and the environment. *Nucleic acids research*, 36(suppl_1):D480–D484, 2007.

[215] Lobb, B., Tremblay, B.J.M., Moreno-Hagelsieb, G. et al. An assessment of genome annotation coverage across the bacterial tree of life. *Microbial Genomics*, 6 (3), 2020.

[216] Greenblum, S., Carr, R. and Borenstein, E. Extensive strain-level copy-number variation across human gut microbiome species. *Cell*, 160(4):583–594, 2015.

[217] Kellner, S., Spang, A., Offre, P. et al. Genome size evolution in the archaea. *Emerging Topics in Life Sciences*, 2(4):595–605, 2018.

[218] Gweon, H.S., Bailey, M.J. and Read, D.S. Assessment of the bimodality in the distribution of bacterial genome sizes. *The ISME journal*, 11(3):821–824, 2017.

[219] Krzakala, F., Moore, C., Mossel, E. et al. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

[220] Morrissey, E.M., Mau, R.L., Schwartz, E. et al. Phylogenetic organization of bacterial activity. *The ISME journal*, 10(9):2336–2340, 2016.

[221] Young, A.D. and Gillung, J.P. Phylogenomics—principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology*, 45(2):225–247, 2020.

[222] Staniczenko, P.P., Kopp, J.C. and Allesina, S. The ghost of nestedness in ecological networks. *Nature communications*, 4(1):1–6, 2013.

[223] Martiny, J.B., Jones, S.E., Lennon, J.T. et al. Microbiomes in light of traits: a phylogenetic perspective. *Science*, 350(6261), 2015.

[224] Minello, G., Rossi, L. and Torsello, A. On the von neumann entropy of graphs. *Journal of Complex Networks*, 7(4):491–514, 2019.

[225] Tamminen, M., Virta, M., Fani, R. et al. Large-scale analysis of plasmid relationships through gene-sharing networks. *Molecular biology and evolution*, 29(4): 1225–1240, 2012.

[226] Iranzo, J., Koonin, E.V., Prangishvili, D. et al. Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsidless mobile elements. *Journal of virology*, 90(24):11043–11055, 2016.

[227] Friedman, J., Higgins, L.M. and Gore, J. Community structure follows simple assembly rules in microbial microcosms. *Nature Ecology & Evolution*, 1:1–7, March 2017.

[228] Nogales, B., Lanfranconi, M.P., Piña-Villalonga, J.M. et al. Anthropogenic perturbations in marine microbial communities. *FEMS Microbiology reviews*, 35(2): 275–298, 2011.

[229] Leinonen, R., Sugawara, H., Shumway, M. et al. The sequence read archive. *Nucleic acids research*, 39(suppl_1):D19–D21, 2010.

[230] Lloyd-Price, J., Mahurkar, A., Rahnavard, G. et al. Strains, functions and dynamics in the expanded human microbiome project. *Nature*, 550(7674):61–66, 2017.

[231] Ainsworth, D., Sternberg, M.J., Raczy, C. et al. k-slam: accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. *Nucleic acids research*, 45(4):1649–1656, 2017.

[232] Paul, G., Sreenivasan, S. and Stanley, H.E. Resilience of complex networks to random breakdown. *Physical Review E*, 72(5):056130, 2005.

[233] Liu, Y.Y., Slotine, J.J. and Barabási, A.L. Controllability of complex networks. *nature*, 473(7346):167–173, 2011.

[234] Voitalov, I., van der Hoorn, P., van der Hofstad, R. et al. Scale-free networks well done. *Physical Review Research*, 1(3):033034, 2019.

[235] Carstens, C.J. Proof of uniform sampling of binary matrices with fixed row sums and column sums for the fast curveball algorithm. *Physical Review E*, 91(4): 042812, 2015.

[236] Carstens, C.J., Berger, A. and Strona, G. Curveball: a new generation of sampling algorithms for graphs with fixed degree sequence. *arXiv preprint arXiv:1609.05137*, 2016.

[237] Curveball. `https://git.cs.uni-kl.de/siebert/curveball`. Accessed: 2020-04-16.

[238] Fortunato, S. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[239] Handcock, M.S. and Morris, M. *Relative distribution methods in the social sciences*. Springer Science & Business Media, 2006.

[240] Handcock, M.S. and Morris, M. Relative distribution methods. *Sociological Methodology*, 28(1):53–97, 1998.

[241] Mariani, M.S., Ren, Z.M., Bascompte, J. et al. Nestedness in complex networks: observation, emergence, and implications. *Physics Reports*, 813:1–90, 2019.

[242] Almeida-Neto, M., Guimaraes, P., Guimaraes Jr, P.R. et al. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos*, 117(8):1227–1239, 2008.

[243] Payrato-Borras, C., Hernandez, L. and Moreno, Y. Measuring nestedness: A comparative study of the performance of different metrics. *arXiv preprint arXiv:2002.00534*, 2020.

[244] Flores, C.O., Poisot, T., Valverde, S. et al. Bimat: a matlab package to facilitate the analysis of bipartite networks. *Methods in Ecology and Evolution*, 7(1):127–132, 2016.

[245] Jones, S.E. and Lennon, J.T. Dormancy contributes to the maintenance of microbial diversity. *Proceedings of the National Academy of Sciences*, 107(13):5881–5886, 2010.

[246] Bianconi, G., Darst, R.K., Iacovacci, J. et al. Triadic closure as a basic generating mechanism of communities in complex networks. *Physical Review E*, 90(4): 042806, 2014.

[247] Storey, J.D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[248] Forterre, P. The virocell concept and environmental microbiology. *The ISME journal*, 7(2):233–236, 2013.

[249] Sausset, R., Petit, M., Gaboriau-Routhiau, V. et al. New insights into intestinal phages. *Mucosal immunology*, 13(2):205–215, 2020.

[250] Rohwer, F. and Barott, K. Viral information. *Biology & Philosophy*, 28(2): 283–297, 2013.

[251] Sanchez-Gorostiaga, A., Bajić, D., Osborne, M.L. et al. High-order interactions distort the functional landscape of microbial consortia. *PLoS biology*, 17(12): e3000550, 2019.

[252] Durham, W.M., Climent, E., Barry, M. et al. Turbulence drives microscale patches of motile phytoplankton. *Nature communications*, 4, 2013.

[253] Zhu, Q., Mai, U., Pfeiffer, W. et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains bacteria and archaea. *Nature communications*, 10(1):1–14, 2019.

Appendix A

Online code and data

Data and code required to reproduce the analyses and figures in this thesis is available on github at https://github.com/jcmckerral.

Appendix B

Turbulence model statistics

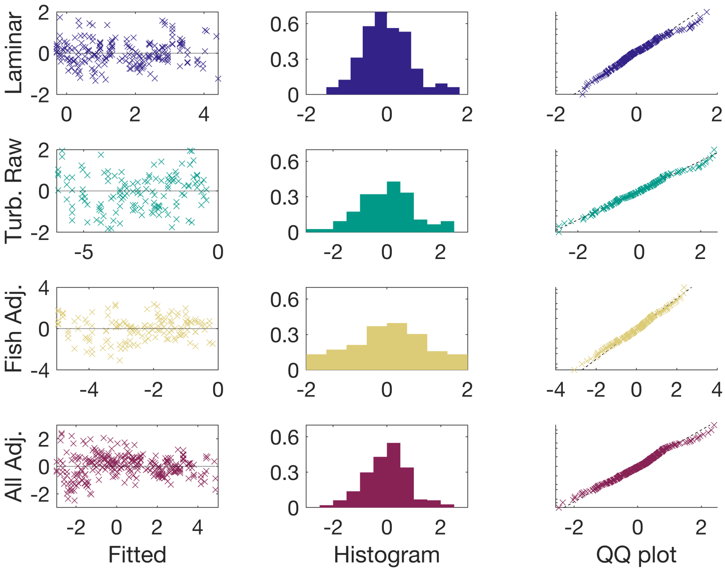


Figure B.1: Model statistics plots for the manually curated data fitted with OLS following log- transformation. L-R: fitted raw residuals, histogram of raw residuals, and raw residuals QQ plot. Top to bottom: plankton (m = 108), raw nekton (m = 61), fishing corrected nekton (m = 61), and full (corrected) spectrum (m = 124). (Sub)sample size given by m.

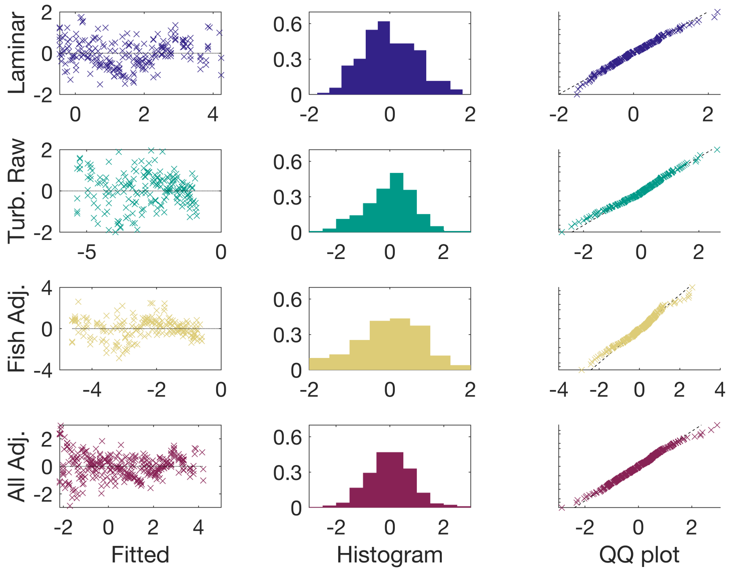


Figure B.2: Residuals plots for a representative subsample from the complete dataset, fitted with OLS following log-transformation. L-R: fitted residuals, histogram, and QQ plot. Top to bottom: plankton ($m = 284$), raw nekton ($m = 36$), fishing corrected nekton ($m == 236$), and full (corrected) spectrum ($m = 70$). (Sub)sample size is denoted by m.

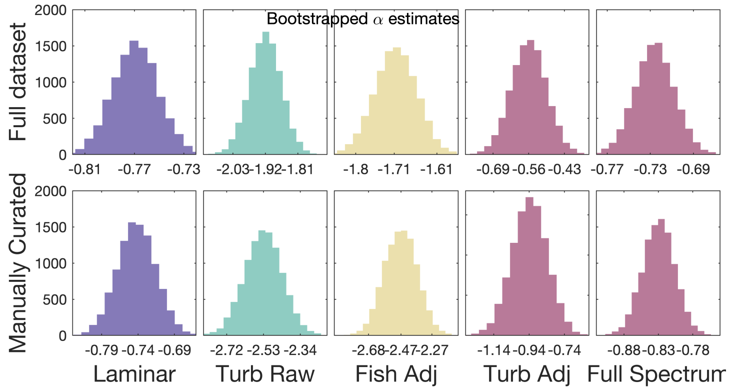
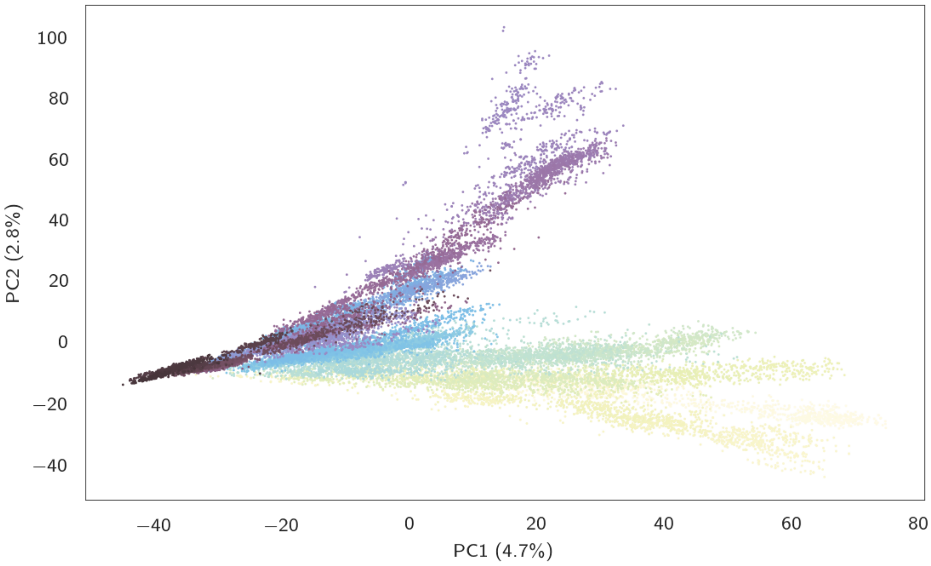


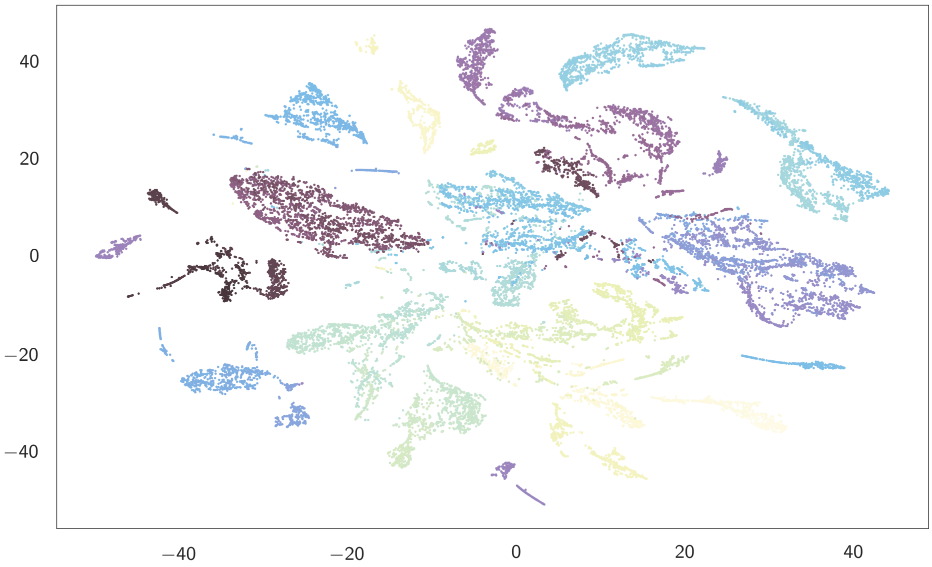
Figure B.3: Histograms of 10,000 bootstrapped alpha (slope) estimates for each model for the values shown in Table 3.3.

Appendix C

PCA and tSNE across annotation systems

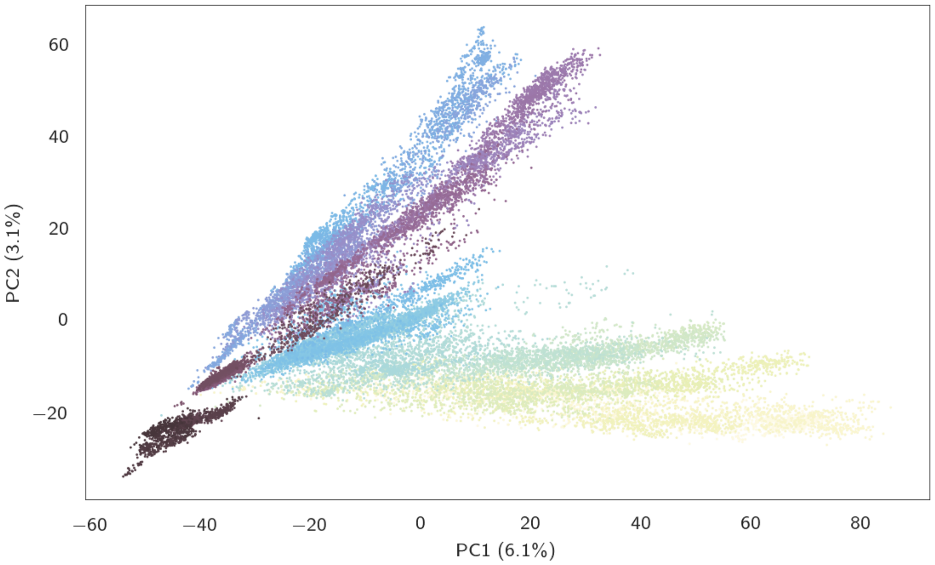


(a)

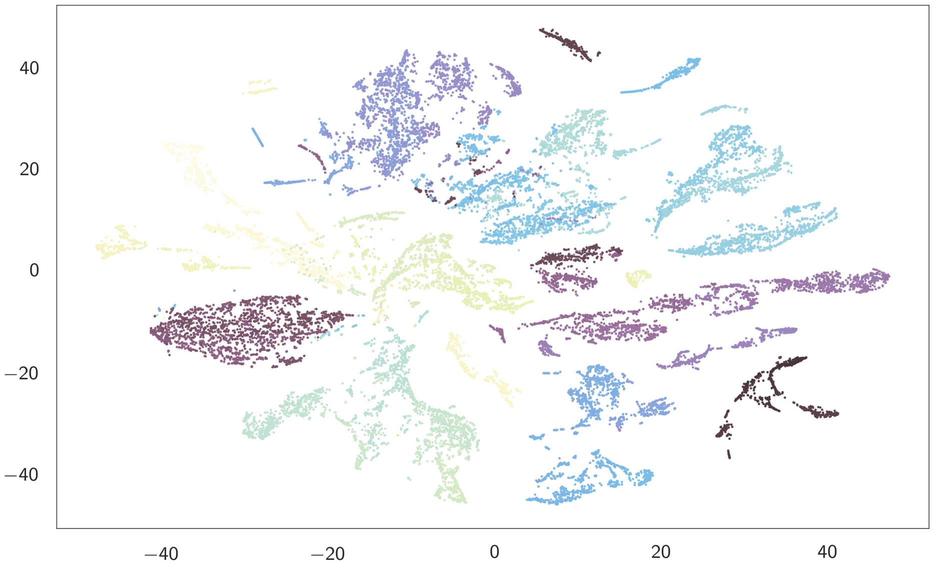


(b)

Figure C.1: Taxa after clustering on up to 15 PCs of the Subsystems biadjacency matrix. Colour is set using continuous colourmap applied to all taxa in phylogenetic order; that is, closely related taxa will have similar colours. (a)PCA applied to the taxa. (b) tSNE applied to the taxa's first 15 principal components, using a perplexity value of 80.



(a)



(b)

Figure C.2: Taxa after clustering on up to 15 PCs of the KEGG biadjacency matrix. Colour is set using continuous colourmap applied to all 28,439 taxa in phylogenetic order; that is, closely related taxa will have similar colours. (a)PCA applied to the taxa. (b) tSNE applied to the taxa's first 15 principal components, using a perplexity value of 80.

Appendix D

Entropy across phylogenetic levels

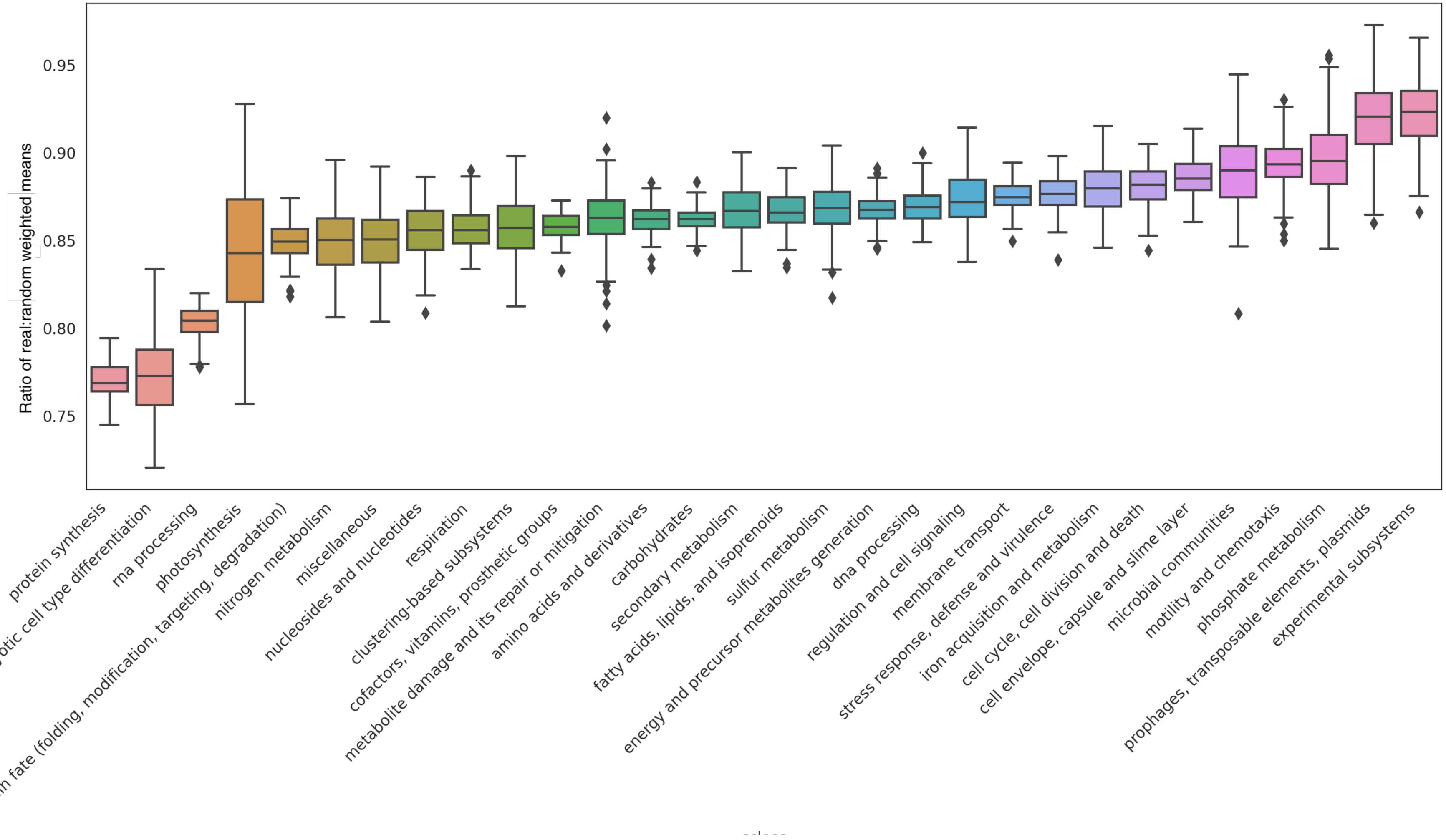


Figure D.1: Mean entropy score, relative to random baseline, per subsystems Class at the phyla level. The $y-$axis shows the ratio of the real to randomised weighted mean, as given in Equation 5.9. The boxplots are generated based on values from 1000 bootstrapped networks.

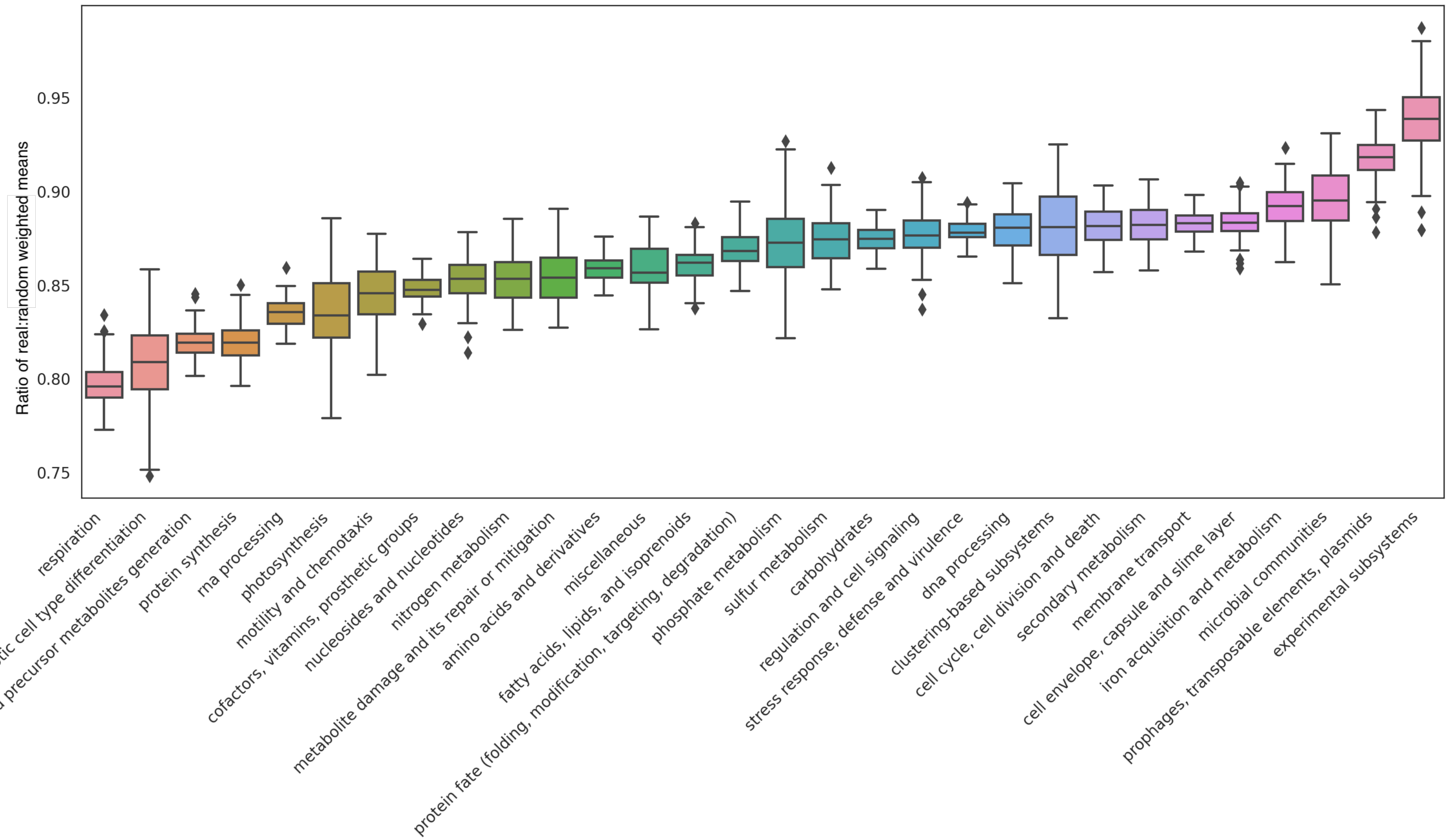


Figure D.2: Mean entropy score, relative to random baseline, per subsystems Class at the (taxonomic) class level. The $y-$axis shows the ratio of the real to randomised weighted mean, as given in Equation 5.9. The boxplots are generated based on values from 1000 bootstrapped networks.

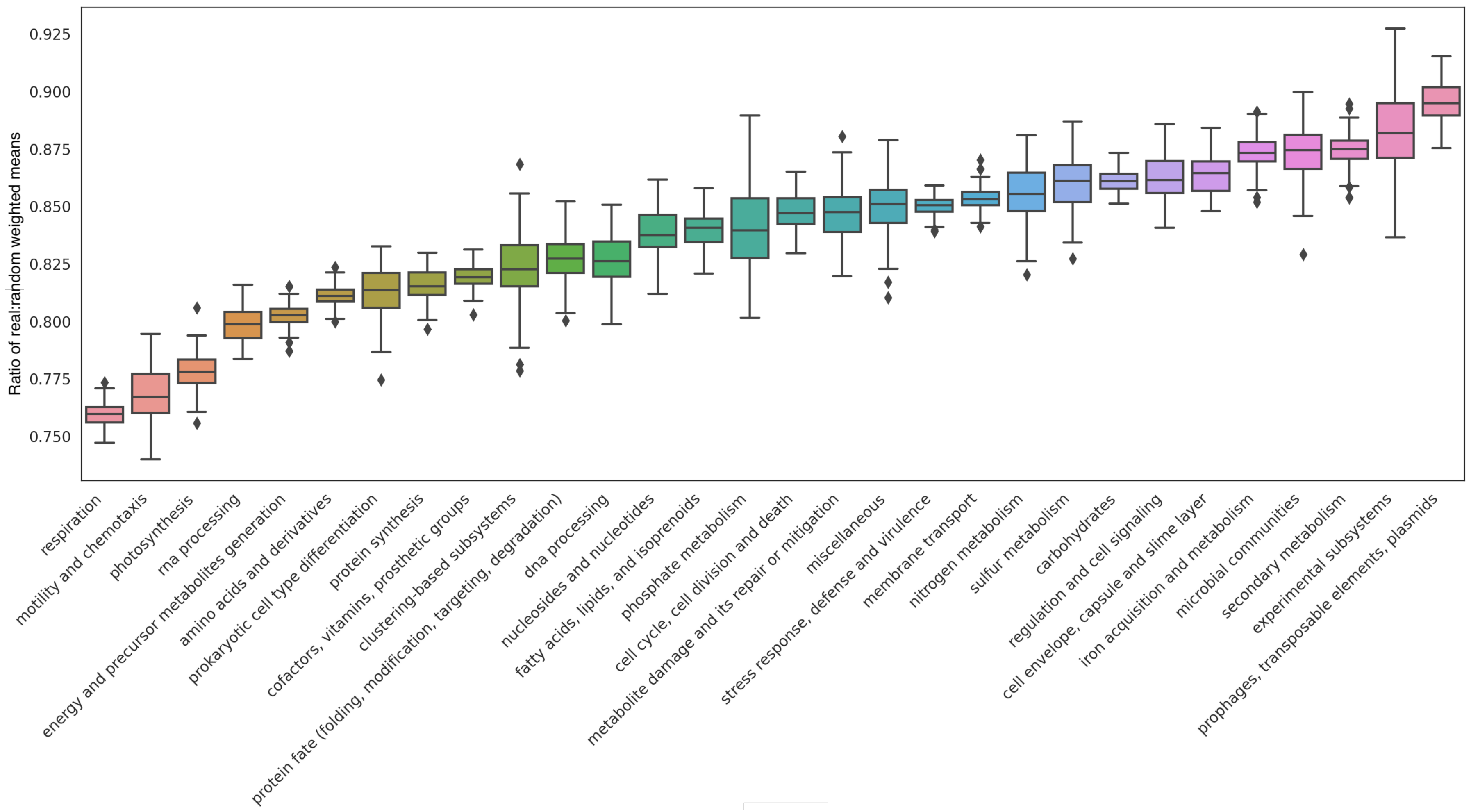


Figure D.3: Mean entropy score, relative to random baseline, per subsystems Class at the family level. The $y-$axis shows the ratio of the real to randomised weighted mean, as given in Equation 5.9. The boxplots are generated based on values from 1000 bootstrapped networks.

Appendix E

SRA sample run identifiers

Table E.1: Sequence read archive run IDs and associated biome for WGS samples analysed in Chapter 6.

Run ID	Biome	Run ID	Biome	Run ID	Biome	Run ID	Biome
ERR1544006	freshwater	SRS018656	gut	ERR599010	marine	ERR1877666	soil
ERR1750013	freshwater	SRS019267	gut	ERR599012	marine	ERR1877677	soil
SRR3306837	freshwater	SRS019397	gut	ERR599018	marine	ERR1877682	soil
SRR3989312	freshwater	SRS019582	gut	ERR599020	marine	ERR1877690	soil
SRR5208983	freshwater	SRS078177	gut	ERR599023	marine	ERR1877692	soil
SRR5209621	freshwater	SRS1170700	gut	ERR599024	marine	ERR1877695	soil
SRR5246518	freshwater	SRS1170723	gut	ERR599031	marine	ERR1877702	soil
SRR5754828	freshwater	SRS1170767	gut	ERR599034	marine	ERR1877709	soil
SRR6050958	freshwater	SRS1170770	gut	ERR599035	marine	ERR1877710	soil
ERR1457091	freshwater	SRS1170816	gut	ERR599042	marine	ERR1877712	soil
ERR1815064	freshwater	SRS1170825	gut	ERR599045	marine	ERR1877714	soil
ERR472738	freshwater	SRS1170847	gut	ERR599050	marine	ERR1877716	soil
SRP100355	freshwater	SRS1170891	gut	ERR599053	marine	ERR1877721	soil
SRR077313	freshwater	SRS147022	gut	ERR599056	marine	ERR1877730	soil
SRR1107072	freshwater	SRS147039	gut	ERR599057	marine	ERR1877733	soil
SRR1173821	freshwater	SRS148424	gut	ERR599058	marine	ERR1877738	soil
SRR167723	freshwater	SRS148721	gut	ERR599069	marine	ERR1877739	soil
SRR3098756	freshwater	SRS1596771	gut	ERR599092	marine	ERR1877746	soil
SRR3184732	freshwater	SRS1596811	gut	ERR599094	marine	ERR1877747	soil
SRR3568916	freshwater	SRS1596815	gut	ERR599096	marine	ERR1877749	soil
SRR3568916	freshwater	SRS1596816	gut	ERR599107	marine	ERR1877750	soil
SRR3986827	freshwater	SRS1596876	gut	ERR599111	marine	ERR1877757	soil
SRR3987495	freshwater	SRS1596877	gut	ERR599115	marine	ERR1877759	soil
SRR3987657	freshwater	SRS2320639	gut	ERR599119	marine	ERR1877764	soil
SRR3987663	freshwater	SRS2320642	gut	ERR599122	marine	ERR1877765	soil
SRR4029415	freshwater	SRS475931	gut	ERR599130	marine	ERR1877775	soil
SRR4198666	freshwater	SRS475962	gut	ERR599135	marine	ERR1877778	soil
SRR5211153	freshwater	SRS476034	gut	ERR599142	marine	ERR1877786	soil
SRR5214089	freshwater	SRS476101	gut	ERR599144	marine	ERR1877789	soil
SRR5216661	freshwater	SRS476119	gut	ERR599155	marine	ERR1877814	soil
SRR5246785	freshwater	SRS883031	gut	ERR599157	marine	ERR1877847	soil
SRR5260362	freshwater	SRS883037	gut	ERR599159	marine	ERR1877848	soil
SRR5260654	freshwater	SRS883066	gut	ERR599166	marine	ERR1877849	soil
SRR5260685	freshwater	SRS883067	gut	ERR599168	marine	ERR1877855	soil
SRR526911	freshwater	SRS883113	gut	ERR599170	marine	ERR1877856	soil
SRR5273324	freshwater	SRS883147	gut	ERR599176	marine	ERR1877858	soil
SRR5277061	freshwater	ERR315860	marine	SRR5195106	rhizosphere	ERR1877859	soil
SRR5298537	freshwater	ERR315861	marine	SRR5195108	rhizosphere	ERR1877863	soil
SRR5468366	freshwater	ERR315863	marine	SRR5195110	rhizosphere	ERR1877865	soil
SRR5468414	freshwater	ERR598944	marine	SRR5195112	rhizosphere	ERR1877868	soil
SRR5581337	freshwater	ERR598945	marine	SRR5195114	rhizosphere	ERR1877869	soil
SRR5581526	freshwater	ERR598948	marine	SRR5195116	rhizosphere	ERR1877870	soil
SRR5818193	freshwater	ERR598949	marine	SRR5195117	rhizosphere	ERR1877878	soil
SRR5818249	freshwater	ERR598952	marine	SRR5195118	rhizosphere	ERR1877881	soil
SRR6048557	freshwater	ERR598953	marine	SRR5195119	rhizosphere	ERR1877887	soil
ERS235535	gut	ERR598954	marine	SRR5195121	rhizosphere	ERR1877888	soil
ERS235587	gut	ERR598961	marine	SRR5195123	rhizosphere	ERR1877893	soil
ERS235598	gut	ERR598964	marine	SRR5195125	rhizosphere	ERR1877912	soil
ERS396405	gut	ERR598965	marine	SRR5195127	rhizosphere	ERR1877914	soil
ERS396472	gut	ERR598967	marine	SRR5195129	rhizosphere	ERR1877915	soil
ERS396473	gut	ERR598968	marine	SRR5195131	rhizosphere	ERR1877916	soil
ERS537325	gut	ERR598973	marine	SRR5195133	rhizosphere	ERR1877921	soil
ERS537340	gut	ERR598984	marine	SRR5195135	rhizosphere	ERR1877926	soil
ERS537384	gut	ERR598985	marine	SRR5195137	rhizosphere	ERR1877933	soil
ERS537387	gut	ERR598986	marine	SRR5195139	rhizosphere	ERR1877936	soil
ERS537410	gut	ERR598987	marine	SRR5195141	rhizosphere	ERR1877937	soil
ERS608499	gut	ERR598991	marine	SRR5195143	rhizosphere	ERR1877661	soil
ERS608510	gut	ERR599002	marine	SRR5195145	rhizosphere	ERR1877663	soil
ERS608530	gut	ERR599003	marine	SRR5195147	rhizosphere	ERR1877657	soil
ERS631833	gut	ERR599004	marine	ERR1877650	soil		

Appendix F

Degree distributions supplement

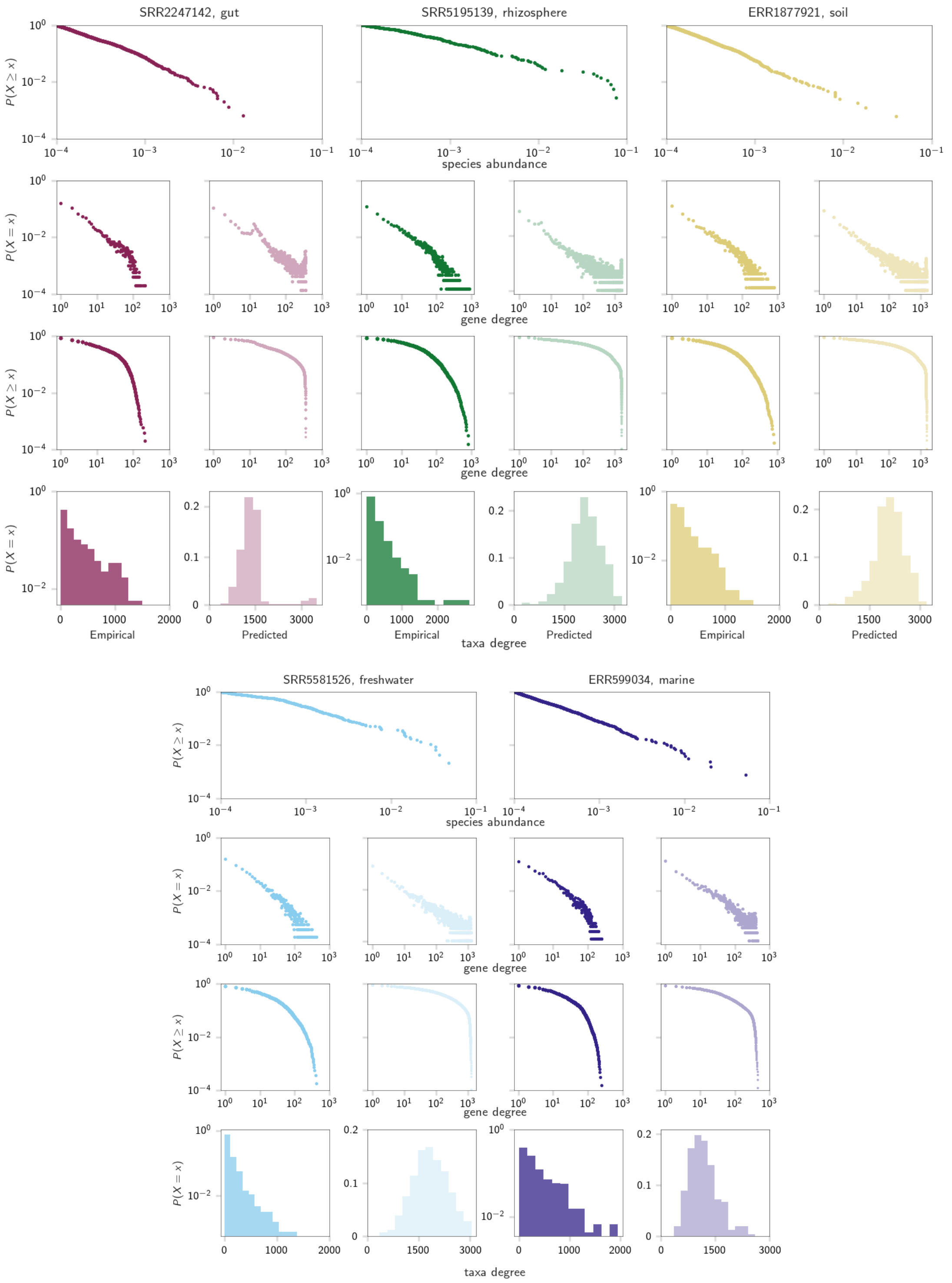


Figure F.1: Degree distributions in empirical networks annotated with KEGG, showing representative samples from different biomes.

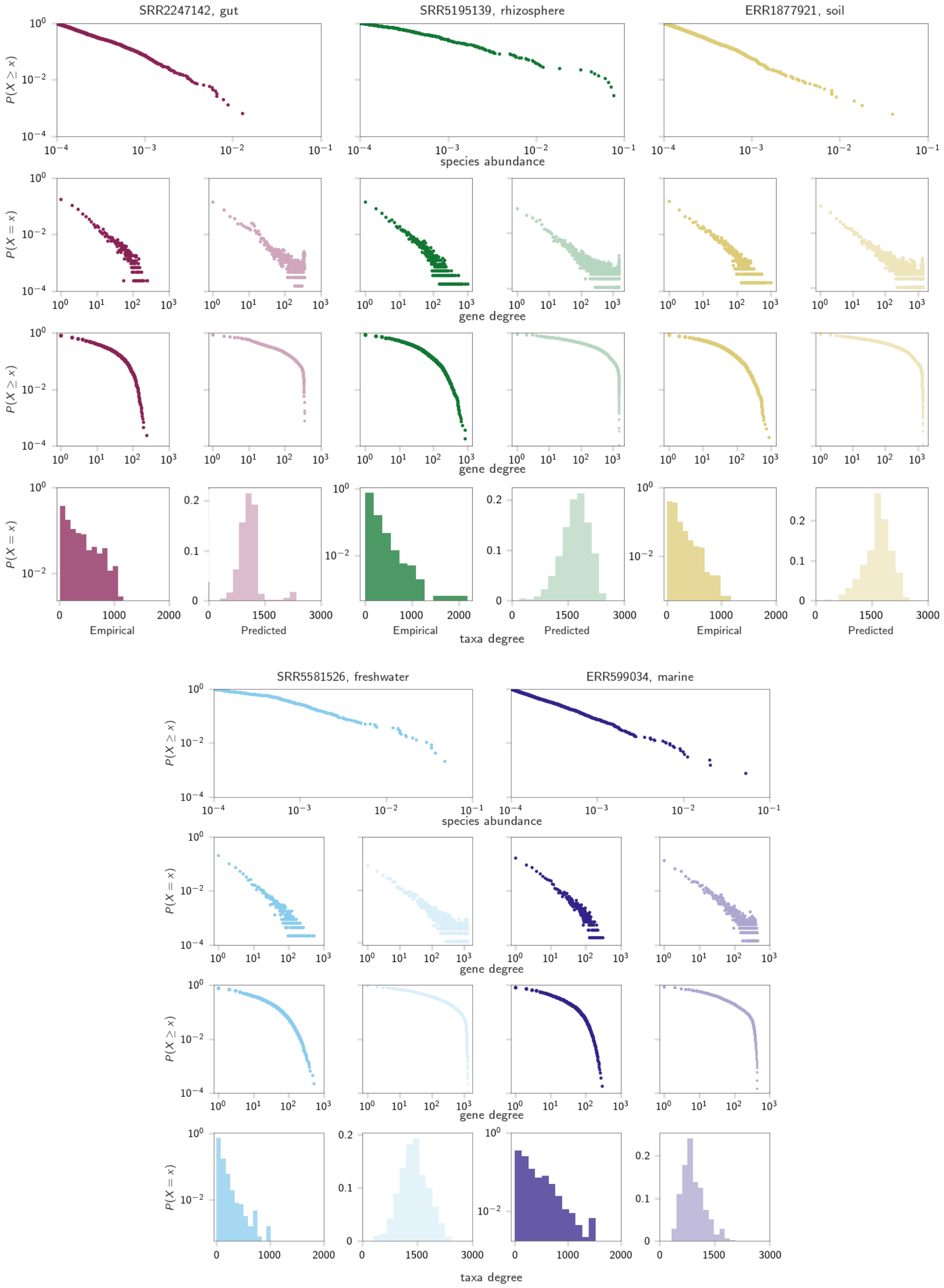


Figure F.2: Degree distributions in empirical networks annotated with Subsystems, showing representative samples from different biomes.

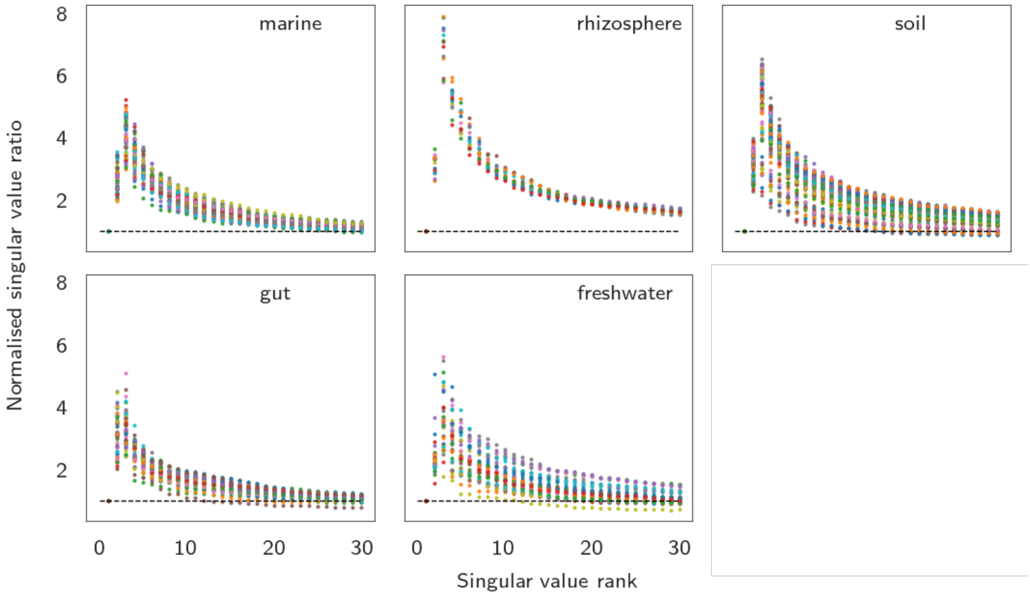


Figure F.3: Relative distribution ratios of the top 30 singular values across biomes. For each predicted network, the top 30 singular values were normalised to a maximum value of 1. For each sample, pairwise ratios of real-world and randomised singular values are plotted against their rank. The dotted line indicates a value of 1; if the distributions are the same, they should fall along or near the line.

Appendix G

Functional redundancy supplement

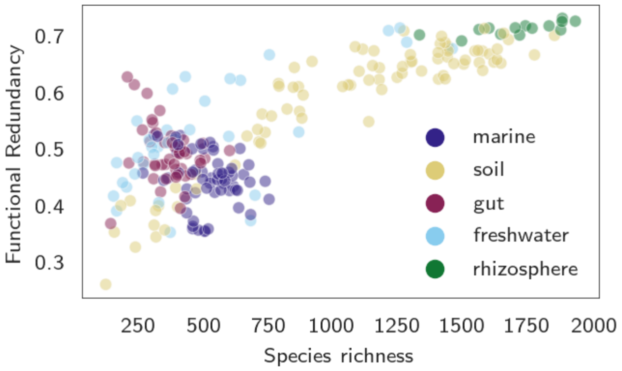


Figure G.1: Functional redundancy score vs. taxonomic diversity: communities sit within an envelope of values. Calculations undertaken with KEGG-annotated networks.

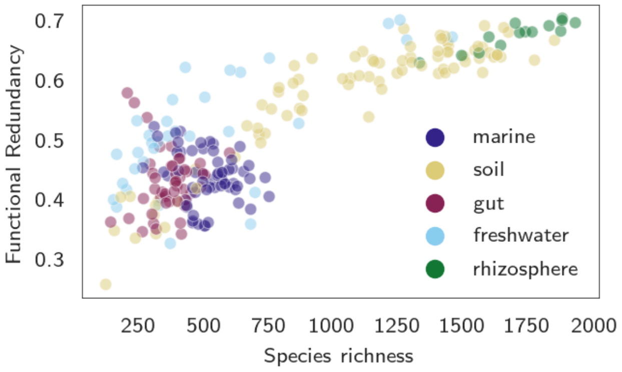


Figure G.2: Functional redundancy score vs. taxonomic diversity: communities sit within an envelope of values. Calculations undertaken with Subsystems-annotated networks.

Appendix H

NODF statistics supplement

Table H.1: Statistical results: identification of differences between KEGG NODF scores in the WGS and randomised networks (Mann-Whitney U test, two-sided). Here (E) denotes empirical networks and (P) denotes predicted networks. Note that most of the significant $p-$values place a higher NODF value on the randomised networks. [†]The first value in these columns is for the real-world networks, the second for the randomised.

Biome	n	Median[†] (E)	U (E)	p-value (E)	Median[†] (P)	U (P)	p-value (P)
Gut	46	2.48E-01/2.77E-01	574	1.60E-04	3.90E-01/3.83E-01	1116	6.53E-01
Marine	60	1.78E-01/1.96E-01	1070	1.29E-04	2.99E-01/2.86E-01	2337	4.86E-03
Freshwater	35	2.00E-01/2.05E-01	522	4.96E-01	3.48E-01/3.68E-01	444	1.02E-01
Soil	84	6.69E-02/6.83E-02	3304	4.78E-01	3.82E-01/4.10E-01	628	3.65E-20
Rhizosphere	23	1.08E-01/1.08E-01	264	1.00E+00	3.77E-01/4.04E-01	31	3.07E-07

Table H.2: Statistical results: identification of differences between Subsystems NODF scores in the WGS and randomised networks (Mann-Whitney U test, two-sided). Here (E) denotes empirical networks and (P) denotes predicted networks. Note that most of the significant $p-$values place a higher NODF value on the randomised networks. [†]The first value in these columns is for the real-world networks, the second for the randomised.

Biome	n	Median[†] (E)	U (E)	p-value (E)	Median[†] (P)	U (P)	p-value (P)
Gut	46	2.34E-01/2.55E-01	601	3.64E-04	3.37E-01/3.34E-01	1174	3.67E-01
Marine	60	1.76E-01/1.88E-01	1292	7.73E-03	2.81E-01/2.61E-01	2693	2.81E-06
Freshwater	35	1.96E-01/2.02E-01	534	5.94E-01	2.98E-01/3.07E-01	502	3.54E-01
Soil	84	7.04E-02/7.16E-02	3357	5.89E-01	3.22E-01/3.47E-01	625	3.34E-20
Rhizosphere	23	1.01E-01/1.00E-01	264	1.00E+00	3.21E-01/3.40E-01	36	5.47E-07

Appendix I

Clustering statistics supplement

Table I.1: Statistical results: identification of differences between KEGG average clustering scores in the WGS and randomised networks (Mann-Whitney U test, two-sided). Here (E) denotes the empirical networks, (P) the predicted networks. †The first value in these columns is for the real-world networks, the second for the randomised.

Biome	n	Median† (E)	U (E)	p-value (E)	Median† (P)	U (P)	p-value (P)
Gut	46	1.00E-01/7.52E-02	1847	7.40E-10	1.28E-01/9.66E-02	2007	1.29E-13
Marine	60	7.05E-02/5.72E-02	3161	9.28E-13	8.64E-02/7.08E-02	3471	1.82E-18
Freshwater	35	9.56E-02/8.20E-02	750	8.56E-03	1.08E-01/9.15E-02	900	5.30E-06
Soil	84	5.78E-02/5.13E-02	4346	3.62E-03	1.27E-01/1.19E-01	6268	7.70E-20
Rhizosphere	23	4.09E-02/3.50E-02	166	1.93E-03	1.35E-01/1.27E-01	171	8.65E-04

Table I.2: Statistical results: identification of differences between Subsystems average clustering scores in the WGS and randomised networks (Mann-Whitney U test, two-sided). Here (E) denotes the empirical networks, (P) the predicted networks. †The first value in these columns is for the real-world networks, the second for the randomised.

Biome	n	Median† (E)	U (E)	p-value (E)	Median† (P)	U (P)	p-value (P)
Gut	46	9.03E-02/7.21E-02	1786	1.34E-08	1.13E-01/8.73E-02	2043	1.50E-14
Marine	60	6.69E-02/5.75E-02	3017	1.71E-10	8.47E-02/7.02E-02	3570	1.58E-20
Freshwater	35	9.33E-02/8.07E-02	725	2.10E-02	9.25E-02/7.91E-02	875	2.32E-05
Soil	84	5.73E-02/5.32E-02	3973	8.81E-02	1.08E-01/9.92E-02	6336	9.92E-21
Rhizosphere	23	4.17E-02/3.65E-02	161	4.08E-03	1.15E-01/1.10E-01	174	5.22E-04

Appendix J

Works arising

The following list details journal articles arising from the work in this thesis that is currently submitted for review or in preparation, under my academic publishing name JC McKerral. Where applicable, a DOI for an article preprint is provided.

- Chapter 2: *Universal allometry from empirical parameters*, with coauthors Jerzy A. Filar, James G. Mitchell, Maria Kleshnina, and Louise Bartle. Contributions: JCM conceived the work and wrote the paper; JCM and JAF developed the model; JCM and MK analysed the model; JCM and LB sourced and analysed data; JGM contributed to writing, interpretation and insight. All authors edited the manuscript. DOI 10.1101/2021.05.20.444891.

- Chapter 3: *Synergy of turbulence and fishing reduce aquatic biomass,* with coauthors Justin R. Seymour, Trish J. Lavery, Paul J. Rogers, Thomas C. Jeffries, James S. Paterson, Ben Roudnew, Charlie Huveneers, Kelly Newton, Virginie van Dongen-Vogels, Nardi P. Cribb, Karina M. Winn, Renee J. Smith, Crystal L. Beckmann, Eloise Prime, Claire M. Charlton, Maria Kleshnina, Susanna R. Grigson, Marika Takeuchi, Laurent Seuront, James G. Mitchell. Contributions: JGM, JCM, JRS and LS conceived the work. JCM, JGM JRS wrote the paper. SRG, TJL, PJR, TCJ, JSP, BR, CH, KN, VvDV, NPC, KMW, RJS, CLB, EP, JRS and CMC gathered the data and helped with the analysis. JCM developed the model, gathered data, and did the analysis. MK contributed to model development. MT helped with analysis and contributed to writing, interpretation and insight. LS helped with analysis, and contributed to writing and interpretation and insight. DOI 10.1101/2021.10.04.459351.

- Chapters 4-6 (in preparation): *A universal genetic topology in microbial systems*, with coauthors Nima Dehmamy, Robert A Edwards, and James G. Mitchell. JCM conceived the work, did the analysis, and wrote the paper. RE assisted with bioinformatics pipelines and methods; ND assisted with analytic method design. JGM contributed to writing, interpretation and insight.