Recovering the Missing Links: From Genomic Signatures to Virus-Host Interaction Networks

by

G. Eric Bastien

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in the University of Michigan
2025

Doctoral Committee:

Associate Professor Melissa B. Duhaime, Co-Chair
Assistant Professor Luis Zaman, Co-Chair
Associate Professor Gideon Bradburd
Associate Professor Vincent Denef
Lecturer IV Cristina Mitrea
Associate Professor Robert Woods

G. Eric Bastien

gbastien@umich.edu

ORCID iD: 0000-0003-1644-600X

© G. Eric Bastien 2025

Dedication

À mon grand-père, dont le décès durant mon parcours doctoral m'a profondément touché, et à ma grand-mère, qui m'a appris à viser les étoiles—sachant que même lorsqu'on atterrit sur la lune, on peut toujours être heureux.

Merci pour votre curiosité sans faille et votre foi en moi.

Acknowledgements

First and foremost, I would like to thank my PhD advisor Melissa Duhaime for her support throughout this journey that started when I was a recent transfer undergraduate student arriving at the University of Michigan. Thank you for welcoming me to your lab and helping me feel like part of a greater scientific community. I will always remember the intellectual liberty you granted me when I was just an undergraduate and our incredibly satisfying discussions. I grew tremendously, both as a scientist and as a human being, and I owe so much of who I am today to your mentorship. It is incredible to think that we started working together 9 years ago.

I thank my co-PI Luis Zaman, who also accompanied me through this journey, mentoring me and encouraging me to pursue novel and complex ideas. I especially appreciate your encouragement and the time you made available when I needed feedback the most. I am grateful to my dissertation committee members for their valuable feedback, challenging questions, and guidance throughout this doctoral journey.

I gratefully acknowledge the National Science Foundation, which made this research possible. Thank you to the University of Michigan Rackham Graduate School and the Department of Ecology and Evolutionary Biology for providing additional funding. I would also like to thank the Department of Computational Medicine & Bioinformatics and LSA Technology Services for the technical expertise and resources that were essential to my research.

This dissertation would not have been possible without the unwavering support of the Duhaime lab. Thank you for years of brainstorming, discussions, and feedback. I thank the broader scientific community on the fourth floor of the Biological Science Building.

I would like to specifically thank Evelyn Faust, a master's student in Computational Medicine & Bioinformatics, whose discussions reinvigorated my research interests. This work would not have been possible without her tremendous help and curiosity. Similarly, I thank Emmalyn Campau, another master's student in Computational Medicine & Bioinformatics, who wrote extensive code that made much of this analysis possible.

I thank Cecelia Batterbee, who initially helped me with data collection from databases and literature. The dataset collected was instrumental to my dissertation and took months to collect.

The collaborative nature of this work reflects the cross-pollination of ideas, dreams, and goals that emerged from interactions within the Duhaime lab, the BSB neighborhood, the EEB community, the greater University of Michigan scientific community, and scientists I met through conferences and networking.

I thank Jessica Choi, Libby Davenport, Greg Krulin, Rachel Cable, and Marcin Dziuba for their friendship over the years. I'm grateful to the broader community of friends who embraced my hobby of board games and roleplaying—I especially thank the many groups I had the chance to be a game master, creating worlds limited only by our imaginations. The COVID years, though painful, brought unexpected gifts in the form of online friendships that sustained me through challenging times. I thank my parents for everything they given me.

Finally, reflecting on the last five years, it's remarkable to think of where this journey started. Perhaps it wasn't judicious to begin a doctoral program in the midst of a pandemic, but foresight on how things would evolve is rarely given. 2020 was also the year I married Ashley, who has been my companion, confidant, and anchor throughout this journey. Thank you for everything. Life would not be the same without you.

Table of Contents

Dedication ii

Acknowledgements iii

List of Tables x

List of Figures xi

Abstract xiv

Chapter 1 Introduction 1

1.1 The discovery of viruses 1

1.2 The virocell concept 2

1.3 Origins of viral ecology 3

1.4 Unveiling viral genomic diversity and ecological implications 4

1.5 The missing links 5

1.6 Network thinking in viral ecology 6

1.7 Thesis overview 7

1.8 References 10

Chapter 2 – Host Translational Machinery Reflects Viral Codon Usage Bias 20

2.1 Abstract 20

2.2 Introduction 21

2.3 Results 23

2.3.1 Virus and gene annotation 23

2.3.2 Viral codon usage patterns reflect GC content constraints and adaptation to host ... 23
2.3.3 Viral-encoded tRNAs compensate for virus-host codon usage mismatches ... 24
2.3.4 Viral genomes exhibit measurable adaptation to host cellular machinery ... 25
2.4 Discussion ... 26
2.5 Methods ... 28
2.5.1 Data acquisition of genomes and host range ... 28
2.5.2 Gene annotation of bacterial and viral genomes ... 28
2.5.3 Codon usage frequency calculations ... 29
2.5.4 Virus-encoded tRNAs and associated analysis ... 30
2.5.5 Codon bias feature extraction ... 30
2.5.6 tRNA-informed feature extraction ... 32
2.5.7 Statistical test to assess strength of genomic imprints ... 33
2.6 Figures ... 34
2.7 Supplementary figures ... 40
2.8 VHRnet: A Curated Resource of Laboratory-Verified Virus-Host Interactions ... 45
2.9 References ... 47
Chapter 3 – Virus-Host Interactions Predictor (VHIP): Machine Learning Approach to Resolve Microbial Virus-Host Interaction Networks ... 53
3.1 Abstract ... 53
3.2 Author summary ... 54
3.3 Introduction ... 54
3.4 Results ... 56
3.4.1 VHRnet, a manually curated host range dataset unparalleled in size and scope ... 56
3.4.2 VHRnet provides opportunity for holistic comparison of Host Prediction Tools (HPTs) ... 57

3.4.3 Genomic signals of virus adaptation to their host(s) are discernible at the species level 59

3.4.4 VHIP, a machine learning-based model, leverages the VHRnet dataset to predict virus-host interactions 61

3.5 Discussion 64

3.6 Methods 67

3.6.1 Collection of host range data and associated sequences 67

3.6.2 Comparison of existing virus-host prediction tools on complete host range experiments 68

3.6.3 Evaluation of commonly used features for virus-host predictions 68

3.6.4 Comparison of machine learning classifiers using the VHRnet dataset 69

3.6.5 Training, testing, and evaluation of VHIP 70

3.7 Acknowledgements 71

3.8 VHIP 2.0: Codon usage biases encode signals that improves virus-host interaction predictions from genomic sequences 72

3.9 Figures 74

3.10 Supplementary figures 81

3.11 References 90

Chapter 4 – A Generalizable Machine Learning Framework for Identifying Genomic Markers of Antibiotic Resistance Across Bacterial Species 97

4.1 Abstract 97

4.2 Introduction 97

4.3 Results 99

4.3.1 Data preparation and k-mer matrix construction 99

4.3.2 Determination of optimal of k-mer length for machine learning model to yield highest accuracy 99

4.3.3 Functional annotation of resistance-associated k-mer words 101

4.4 Discussion 101

4.5 Methods 103
4.5.1 Genome and antibiotic susceptibility 103
4.5.2 Generate of k-mer matrices for machine learning 103
4.5.3 Machine learning model training and assessment 104
4.5.4 Linking machine learning model features to genes 105
4.6 Figures 106
4.7 Supplementary Figures 109
4.8 Tables 110
4.9 References 114
Chapter 5 – Structural Properties of Virus-Host Interaction Networks Reveal Methodological Biases and Biological Patterns 119
5.1 Abstract 119
5.2 Introduction 120
5.3 Results 122
5.3.1 Data selection and virus-host interactions networks visualization from different methods 122
5.3.2 Virus-host interaction networks exhibit varied structures both through methodology and environment 123
5.3.3 Null model analysis to determine significance of VHIN's structure 124
5.4 Discussion 125
5.5 Methods 129
5.5.1 Data acquisition 129
5.5.2 Virus-host interaction predictions 130
5.5.3 Network analysis 130
5.5.4 Null model analysis 131
5.6 Figures 132

5.7 Supplementary figures 136
5.8 Tables 140
5.9 References 142
Chapter 6 Conclusion 152
6.1 Viral ecology in the contemporary age 152
6.2 Reconciling the definitions of host range 152
6.3 The continuum of viral outcomes 154
6.4 Applied viral ecology 155
6.5 References 156

List of Tables

Table 3.1. Comparison between input and output of existing host-prediction tools 88

Table 3.2. Pearson pairwise correlations between features of VHIP..... 89

Table 4.1. Bacteria genus and antibiotics considered for this study.....111

Table 4.2. Metrics of the best performing model for each bacteria-antibiotic combination.....113

Table 5.1. Date provenance, methodology used, number of viruses and hosts for each network considered in this study.....140

Table 5.2. Properties of the 11 virus-host interaction networks. 141

List of Figures

Figure 2.1. Viruses exhibit diverse codon usage patterns 34

Figure 2.2. Viruses have a general tendency to be enriched in certain codons and poor in others relative to their known hosts. 35

Figure 2.3. Viruses are enriched in A/U over G/C in their codons relative to their known hosts. 36

Figure 2.4. Viruses encode tRNAs that minimize mismatch with their known host(s). 37

Figure 2.5. Gene-informed signatures result in strongest signal to differentiate between compatible and incompatible virus-host pairs. 38

Figure 2.6. Viruses exhibit multiple dimensions of adaptation to their host. 39

Supplemental Figure 2.7. Distribution of predicted genes in virus genomes. 40

Supplemental Figure 2.8. Number of tRNAs predicted in virus genomes. 41

Supplemental Figure 2.9. Codon usage frequency of all viruses 42

Supplemental Figure 2.10. Null modeling of codon usage bias difference between viruses and hosts. 43

Supplemental Figure 2.11. Comparison of codon usage frequency relative to host 43

Figure 2.12. Comparing TAAI metrics between total pool of tRNAs and only host tRNA pool 44

Figure 3.1. Design of data collection, features computation, and evaluation of machine learning approach for virus-host interaction predictions. 74

Figure 3.2 VHRnet network visualization and content characteristics 75

Figure 3.3 Evaluation of the accuracy of existing virus-host prediction models using two complete host range datasets 76

Figure 3.4 Comparison of coevolution signals captures at the species taxonomic level 77

Figure 3.5 Model performance and feature evaluation 78

Figure 3.6 Model output and future directions 79

Figure 3.7. Model improvement by incorporating gene-derived metrics of virus-host adaptation. 80

Supplemental Figure 3.8. Distribution of all family classifications of viruses in VHRnet 81

Supplemental Figure 3.9. Distribution of all host genera represented in VHRnet 82

Supplemental Figure 3.10 Number of viruses tested against different host taxa. 83

Supplemental Figure 3.11 Kernel density of distance measurements for each virus-host pair and colored by interaction 84

Supplemental Figure 3.12. Feature distribution (diagonal plots) and features co-correlations (non-diagonal plots) 85

Supplemental Figure 3.13. Comparison of different machine learning models trained on VHRnet 86

Supplemental Figure 3.14. ROC curves from 100 bootstrapping iterations of the best model trained during the grid search 87

Supplemental Figure 3.15. F1 curves from 100 bootstrapping iterations of the best model during the grid search 87

Figure 4.1. Gradient Boosting Model classifier accuracy increases with k-length used. 106

Figure 4.2. Comparison of best performing models between Staphylococcus, Acinetobacter, Escherichia, and Pseudomonas. 107

Figure 4.3. Comparison between number of features generated through k-mer word decomposition and actual number of useful features for antibiotic phenotype prediction 108

Supplemental Figure 4.4. No correlation between the total number of points and Gradient Boosting Classifier's accuracy 109

Figure 5.1. Visualization of the virus-host interaction networks 132

Figure 5.2. Nestedness of the virus-host interaction networks 133

Figure 5.3. Centrality measurements of virus and host nodes 134

Figure 5.4. Conceptual diagram of the methods to reconstruct virus-host interaction network 135

Supplemental Figure 5.5. Distributions of VHIP-scores for predicted virus-host interaction networks 136

Supplemental Figure 5.6. Modularity of the virus-host interaction networks. 137

Supplemental Figure 5.7. Correlation between network architecture and network size ... 138

Supplemental Figure 5.8. Correlation between network architecture and virus-host ratios 139

Abstract

In a little more than a century, our understanding of viruses has shifted from being invisible to human sensing and knowledge to being recognized as ubiquitous biological forces shaping ecosystems and global biogeochemical cycles. Advancements in next-generation sequencing have accelerated virus discovery at an unprecedented rate, revealing staggering measures of viral diversity. However, this genomic revolution has exposed a critical knowledge gap: while sequencing enables rapid identification of viral genomes, determining the host organisms these viruses infect remains a grand fundamental challenge in microbial ecology.

This dissertation addresses this challenge by demonstrating that viral genomes encode detectable signatures of adaptation to their hosts. By leveraging computational and machine learning approaches, I show that these “genomic fingerprints" or “battle scars” of virus-host co-evolution can be systematically decided to accurately predict virus-host interactions. The central hypothesis guiding this work is that the evolutionary relationships between viruses and their hosts imprints molecular constraints—leaving signals that can be computationally extracted to identify virus-host relationships.

Chapter 2 establishes the theoretical groundwork for this concept by quantifying the extent to which virus-host adaptation is reflected at the genomic level. I adapted existing gene-cellular machinery compatibility frameworks to characterize molecular interactions governing virus-host relationships. The analysis demonstrates that viruses optimize their codon usage bias to match host cellular environments, while simultaneously adapting across multiple genomic dimensions. These findings reveal that viral dependence on host machinery creates detectable evolutionary constraints encoded in viral genomes.

Building on these insights, Chapter 3 presents the development of the Virus-Host Interaction Predictor (VHIP), a machine learning model specifically designed to predict virus-host associations from genome sequence data. VHIP achieves 92% accuracy in predicting virus-host relationships and is distributed as an open-source Python package, thus providing the scientific community with a scalable, accessible, and robust tool for predicting and investigating virus-host networks from sequence data.

Chapter 4 expands the genomic "battle scars" framework beyond virus-host adaptation by applying it to bacterial genomes. In this chapter, I developed a generalizable machine learning framework and with it show that genomic features can be harnessed to predict bacterial antibiotic resistance phenotypes.

In Chapter 5, I explored the architecture of virus-host interactions networks derived from different methodologies—including laboratory infection assays, Hi-C proximity ligation, and computational predictions from VHIP. Through comparative analyses of 11 networks, I characterize their structural properties, shedding light on how different approaches influence network topology and our understanding of virus-host interactions.

Chapter 6 synthesizes these findings in the context of viral ecological theory. It discusses conceptual advances in our understanding of host range and network ecology and proposes future research directions for integrating computational and experimental approaches in viral ecology.

Collectively, this dissertation establishes that signatures of genomic adaptation in viral and bacterial genomes are powerful indicators of ecological interactions and can be decided to reconstruct complex host-pathogen networks. By developing theoretical and practical frameworks—especially including the VHIP tool—this work advances computational viral ecology and opens new avenues for the application of network thinking in viral ecology.

Chapter 1 Introduction

1.1 The discovery of viruses

The discovery of viruses represents one of microbiology's most remarkable chapters. It began with a series of separate observations that challenged prevailing theories of infectious disease, particularly Koch's postulates requiring pathogens to be isolable and cultivable (1). Adolf Mayer provided the first reported evidence in 1886, demonstrating that tobacco mosaic disease could be transmitted between plants (2). Yet, no bacterial pathogen could be isolated using standard cultivation methods. Subsequent experiments revealed that this infectious agent could pass through porcelain filters designed to retain all known bacteria, suggesting the existence of submicroscopic infectious particles (2, 3). Within a decade, such discoveries extended across biology: foot-and-mouth disease in livestock (4), yellow fever in humans (5), and bacterial infections studied by Twort (6) and d'Herelle (7). Unbeknownst to these scientists, their observations were documenting similar biological entities—viruses. The modern conceptual framework to understand these as a unified category would not emerge until decades later (8–10).

What these early researchers could not see—but correctly inferred must be extraordinarily small (3, 6)—became visible with the development of electron microscopy in the 1940s (11–13). Electron microscopy revealed particles typically measuring 20-300 nanometers (14–16), consisting of nucleic acid cores surrounded by protein capsids (17, 18), and in some cases enclosed within lipid envelopes (14). This structural characterization, combined with decades of infectious disease studies (9, 10, 19), established the defining features of viruses: obligate intracellular parasitism (20), dependence on host cellular machinery for replication (21–23), and molecular mechanisms for host recognition (24–26), entry (20, 24), and exit (27, 28).

These converging insights transformed a collection of early independent observations into a unified scientific discipline: virology (9). The establishment of the International Committee on Taxonomy of Viruses in 1966 (29) marked a turning point, organizing international efforts to isolate, characterize, and classify viruses (30). As of August 2025, the National Center for Biotechnology Information (NCBI) records 18,735 viral genomes from isolated viruses infecting all branches of life. The IMG/VR database from the Joint Genome Institute (JGI) (31), funded by the Department of Energy, compiles viral genomes from experimental and metagenomic samples, and catalogs over 5 million viral genomes—the vast majority representing uncultured viruses discovered through environmental sequencing. Viruses went from undetectable agents of disease to being recognized as the most abundant biological entities on Earth (32–34), with important implications for ecosystem function (35–37), and evolutionary change across all domains of life (38, 39).

1.2 The virocell concept

While early virology research focused on viruses as isolated pathogens, recent decades have revealed that a complete understanding of viral biology requires studying them in the context of their hosts. Discoveries of bacteriophages conferring beneficial traits to hosts, from toxin production in *Vibrio cholerae* (40, 41) to increased fitness in marine cyanobacteria (42), challenged the pathogen-centric view.

The 'virocell' concept emerged from these observations, recognizing that viral infection creates distinct biological entities through sophisticated cellular reprogramming. Formally, the virocell represents an infected cell whose function shifts from producing two cells to producing virions, eliciting the metabolically active "living form" of the virus (43, 44). For example, viruses of cyanobacteria carry auxiliary metabolic genes (AMGs) encoding photosystem components that maintain energy production during infection (45–47). Giant viruses like mimiviruses possess their own transfer RNAs and translation factors, creating hybrid translation systems (48, 49). Other viruses provide tangible effects: *Salmonella* viruses confer resistance to competing viruses (50), while mycobacteriophages enhance

host stress vulnerability (51). These examples illustrate how infection generates novel cellular states with emergent phenotypic properties that interact with the environment in ways distinct from uninfected cellular counterparts (44, 52, 53).

A new arm of contemporary research treats the virocell as the fundamental unit of analysis rather than examining either component in isolation (54, 55). Such integrated approaches have revealed viral contributions to biogeochemical cycles (56, 57), host adaptation (58), and ecosystem stability (59, 60).

1.3 Origins of viral ecology

For much of the twentieth century, viruses remained invisible to science—not because they were absent, but because the tools to detect them did not exist. Established cultivation methods that had revealed pathogenic viruses failed to capture the majority of viruses in the environment. A breakthrough came in 1989 when Øivind Bergh and colleagues applied epifluorescence microscopy to Norwegian coastal waters (61). They found that seawater contained 10^7 viral particles per milliliter, concentrations 10 to 100 times higher than bacterial abundance (61). Within a year, Proctor and Fuhrman demonstrated that up to 40% of marine bacteria were visibly infected at any given time, suggesting that viral lysis, not grazing by protists, dominated microbial mortality (62).

With viruses now countable and trackable, existing ecological theories could be applied, marking the emergence of viral ecology as a scientific field. Those discoveries led to the recognition that viral activity could be on scales large enough to impact global biogeochemical cycles. The viral shunt hypothesis, for example, proposed that viruses redirect carbon flow in aquatic systems by lysing cells before grazers can consume them (37). Subsequent studies quantified this process and estimated that marine viruses return approximately 145 gigatons of carbon annually to the microbial loop (63), comparable to the ~150-175 gigatons of carbon fixed through global primary production (64).

The ecological implications extended beyond biogeochemical cycling. Building on Hutchinson's paradox of the plankton (65), which questioned how so many species could coexist in apparently homogeneous environments, Thingstad's "kill-the-winner" model explained how viruses maintain microbial diversity. As any bacterial population becomes dominant, corresponding virus populations expand to suppress its growth, preventing competitive exclusion between bacterial populations (59, 66). Similarly, the Red Queen hypothesis, developed by Van Valen in 1973, could now be applied to virus-host systems, describing how hosts and viruses engage in perpetual coevolutionary races (67).

1.4 Unveiling viral genomic diversity and ecological implications

While microscopy had revealed extensive viral abundance in the 1990s, spurring interest in puzzling out their ecological roles, the advent of genomic technologies in the late 1990s and early 2000s offered new opportunities to uncover viral diversity. Initial studies sequencing individual viral genomes revealed surprising genetic novelty. This pattern only intensified as sequencing technologies improved, and costs declined. Major ocean sampling expeditions, such as the Global Ocean Survey (2003-2010) (68) and Tara Oceans (2009-2013) (69), uncovered millions of previously unknown viral sequences (70). These surveys revealed that most viral genes have no known homologs, constituting a vast reservoir of genomic diversity aptly termed "viral dark matter", which is now recognized as the largest source of genetic innovation on Earth (71–73).

This genomic exploration also revealed mechanisms by which viral diversity persists across time and space through "viral seed banks." These reservoirs include both free viral particles persisting in the environment and integrated prophages within host genomes. Antarctic ice cores have yielded infectious viruses after 14,000 years of dormancy (74), while marine sediments harbor diverse viral communities dating back millennia (75). Free viruses can remain viable in seawater, soil, and other environments for extended periods, waiting to encounter suitable hosts (76, 77). These seed banks serve multiple ecological functions. They preserve genetic diversity across unfavorable conditions, enable rapid

response to environmental changes through both prophage induction and reactivation of environmental particles, and provide a temporal buffer against extinction.

The combination of quantitative ecological data from the 1990s and genomic insights from the 2000s had firmly established viruses as both numerically dominant and genetically innovative components of microbial ecosystems. Researchers began incorporating viral parameters into biogeochemical models that had previously considered only cellular organisms (78, 79), while the exponentially growing sequence databases—now containing over 5 million viral sequences in repositories like IMG/VR (31)—offered new insights into viral ecology.

1.5 The missing links

The ecological principles described in previous sections, from the virocell concept to viral impacts on biogeochemical cycles, depend on knowing which virus infects which host. Yet cultivation-based methods capture less than 1% of microbial diversity, and the remaining 99% of unculturable microorganisms and their viruses have remained inaccessible to traditional approaches (81–83). While metagenomics answered the question of 'who is there?', it created an equally challenging question: 'who do they infect?' These unknown virus-host associations represent the "missing links" in our understanding of viral ecology—connections that are essential for interpreting viral ecological roles but remain hidden in sequence data.

This knowledge gap motivated the development of computational host prediction methods. Early approaches focused on comparing k-mer frequency profiles between viruses and potential hosts, aiming to assign viruses to broad taxonomic groups like phyla or classes (84, 85). These methods assumed that viruses and their hosts share similar sequence composition patterns due to evolutionary constraints. As the field progressed, researchers refined k-mer comparison techniques and developed more sophisticated metrics that better describe the degree of similarity between virus and host sequences (86). The next iteration of tools moved beyond simple correlation of k-mer profiles. Machine learning

models began incorporating multiple genomic features, such as codon usage patterns, GC content, and oligonucleotide frequencies (87–89). More complex approaches employed Markov chain models to capture sequence dependencies and multi-layered neural networks to identify patterns in genomic data (87, 90–93). Some methods also integrated direct biological evidence, such as CRISPR spacer matches that provide molecular records of past infections (94, 95).

Despite these advances, fundamental limitations persist. Most critically, training data remains severely biased. While positive virus-host associations can be confirmed through cultivation, true negative data (confirmed non-interactions) are rarely available. Instead, assumptions are typically made, particularly that viruses have narrow host ranges and therefore any distant host represents a negative interaction. Growing evidence suggests many viruses infect broader taxonomic groups than previously thought, potentially invalidating these assumptions (96–101). The gap between viral discovery and host assignment continues to widen with each metagenomic study. Current computational methods achieve moderate accuracy at broad taxonomic levels but struggle with species-level predictions where ecological relevance lies. Without reliable virus-host predictions, millions of viral sequences remain unconnected to their ecological context, limiting our ability to incorporate viral processes into ecosystem models or predict how viral communities will respond to environmental change.

1.6 Network thinking in viral ecology

Network approaches have transformed our understanding of complex biological systems—from protein interactions (101) to food webs (102, 103) to disease transmission (104). In ecology, network analysis revealed how pollinator-plant relationships maintain biodiversity (105), how food web structure determines ecosystem stability (106), and how species loss cascades through communities (107–109). Epidemiology uses network models to predict disease spread and identify super-spreaders (110, 111). These successes suggest that network thinking could similarly advance viral ecology by revealing how virus-host interaction patterns shape microbial communities.

However, network analysis in viral ecology remains in its infancy, constrained by the missing links problem. While other fields can observe or experimentally determine interactions, virus-host networks depend on cultivation or computational prediction. The few existing network studies highlight both the potential and the constraints. Most examine simplified systems: multiple viruses infecting different strains of a single bacterial species, revealing strain-specific infection patterns and coevolutionary dynamics (112, 113). One broader study found that virus-host interactions appear nested within local communities—generalist viruses infecting subsets of hosts that specialists also infect—but form distinct modules when examined across spatial scales, suggesting that network structure correlates with geography (114).

These limited examples hint at what comprehensive network analysis could reveal. If we could construct complete and accurate virus-host networks from environmental samples, we could start identifying potential keystone viruses that disproportionately impact community stability, predict how viral pressure maintains microbial diversity, and ultimately model how network structure influences nutrient cycling.

1.7 Thesis overview

The challenge identified throughout this introduction, namely the missing links, is essential for reconstructing the ecological networks that shape microbial communities. This dissertation demonstrates that viral genomes encode quantifiable signatures of host adaptation that can be decoded to predict virus-host interactions. The central thesis argues that viruses and their hosts leave genomic "fingerprints" or "battle scars" through their interactions, measurable signatures created by viral dependence on host cellular machinery and host defenses against infection. This work identifies these bidirectional signatures using a compiled dataset of confirmed virus-host interactions from databases and public literature. Machine learning models trained on these adaptation signatures can reconstruct virus-host interaction networks from sequence data alone, predicting the missing links needed to apply network approaches to environmental viral communities.

Chapter 2 lays the groundwork by developing metrics that capture how viruses adjust their genes with host cellular machinery. These measures quantify the constraints viruses experience during infection and moves beyond simple sequence composition to assess functional compatibility. The chapters show that signals of adaptation are especially pronounced when analyzing viral codon usage biases in relation to their hosts.

Chapter 3 develops the Virus-Host Interaction Predictor (VHIP), a machine learning framework that predicts virus-host relationships with 92% accuracy—a notable improvement to existing methods. The chapter evaluates different genomic features and shows that gene-level adaptation signatures described in Chapter 2 outperform traditional sequence-based approaches. VHIP is distributed as open-source software, providing a practical tool for building virus-host networks from environmental sequences.

Chapter 4 extends the adaptation framework beyond virology, showing similar principles apply to bacterial genomes under antibiotic pressure. Machine learning models detect genomic changes left by antibiotic exposure, revealing both known and novel loci associated with resistance. This work illustrates how environmental pressures shape microbial genomes in predictable ways, validating the broader applicability of adaptation-based prediction methods.

Chapter 5 compares virus-host interaction networks constructed through three approaches: laboratory cultivation, computational prediction from genomic data, and Hi-C proximity ligation. This work explores how each method captures different aspects of viral ecology—laboratory networks show confirmed interactions, predicted networks reveal evolutionary relationships, and Hi-C networks capture DNA proximity between viruses and microbes. By examining how network architectures vary across methods, we demonstrate that integrating multiple approaches is essential for understanding the full complexity of virus-host interactions.

Chapter 6 synthesizes findings within viral ecology theory, discussing fundamental questions about host range definition, the nature and diversity of virus-host interactions, and more broadly the application of network thinking in viral ecology.

These chapters collectively establish genomic adaptation signatures as a promising approach to the missing links problem. By enabling accurate prediction of virus-host interactions, this dissertation advances descriptive viral ecology toward predictive understanding of viral communities through the lens of virus-host interaction networks.

1.8 References

1. Loeffler,F. (1884) Untersuchungen über die Bedeutung der Mikroorganismen für die Entstehung der Diphtherie beim Menschen : bei der Taube und beim Kalbe Berlin : K. Ges.-Amt.
2. Creager,A.N.H., Scholthof,K.-B.G., Citovsky,V. and Scholthof,H.B. (1999) Tobacco Mosaic Virus: Pioneering Research for a Century. *Plant Cell*, **11**, 301–308.
3. Rutherford,E., Martin,C.J., Arkwright,J.A., Barnard,J.E., Smith,K.M., Gye,W.E., Ledingham,J.C.G., Salaman,R.N., Twort,F.W., Andrews,C.H., *et al.* (1997) Discussion on "Ultra-microscopic viruses infecting animals and plants". *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character*, **104**, 537–560.
4. Loeffler,F. and Frosch,P. (1897) Summarischer Bericht über die Ergebnisse der Untersuchungen der Commission zur Erforschung der Maul- und Klauenseuche. **23**.
5. Kelly,H.A. (1923) Walter Reed and Yellow Fever McClure, Phillips.
6. Twort,F. (1915) An investigation on the nature of ultra-microscopic viruses, volume 186, chapter Part II.
7. d'Herelle,F. (1917) Sur un microbe invisible antagoniste des bacilles.
8. Baltimore,D. (1971) Expression of animal virus genomes. *Bacteriological Reviews*, **35**, 235–241.
9. Oldstone,M.B.A. (2014) History of Virology. *Encyclopedia of Microbiology*, 10.1016/B978-0-12-801238-3.00078-7.
10. Zuo,K., Gao,W., Wu,Z., Zhang,L., Wang,J., Yuan,X., Li,C., Xiang,Q., Lu,L. and Liu,H. (2024) Evolution of Virology: Science History through Milestones and Technological Advancements. *Viruses*, **16**, 374.
11. Luria,S.E., Delbrück,M. and Anderson,T.F. (1943) Electron Microscope Studies of Bacterial Viruses. *J Bacteriol*, **46**, 57–77.
12. Luria,S.E. and Anderson,T.F. (1942) The Identification and Characterization of Bacteriophages with the Electron Microscope. *Proceedings of the National Academy of Sciences*, **28**, 127-130.1.

13. Stanley,W. and Anderson,T.F. (1941) A STUDY OF PURIFIED VIRUSES WITH THE ELECTRON MICROSCOPE. *Journal of Biological Chemistry*, **139**, 325–338.
14. Burnet,F.M. (1956) Structure of Influenza Virus. *Science*, **123**, 1101–1104.
15. McMahon,A., Andrews,R., Groves,D., Ghani,S.V., Cordes,T., Kapanidis,A.N. and Robb,N.C. (2023) High-throughput super-resolution analysis of influenza virus pleomorphism reveals insights into viral spatial organization. *PLOS Pathogens*, **19**, e1011484.
16. Li,K., Shah,A., Sharma,R.K., Adkins,R., Marjanovic,T., Doyle,P.S. and Garaj,S. (2023) Metrology of Individual Small Viruses. *Advanced Materials Interfaces*, **10**, 2300385.
17. Kay,L.E. (1986) W. M. Stanley's Crystallization of the Tobacco Mosaic Virus, 1930-1940. *Isis*, **77**, 450–472.
18. Rossmann,M.G. (2013) Structure of viruses: a short history. *Quarterly Reviews of Biophysics*, **46**, 133–180.
19. Bos,L. (1999) Beijerinck's work on tobacco mosaic virus: historical context and legacy. *Philos Trans R Soc Lond B Biol Sci*, **354**, 675–685.
20. Dimitrov,D.S. (2004) Virus entry: molecular mechanisms and biomedical applications. *Nat Rev Microbiol*, **2**, 109–122.
21. Gerovac,M., Chihara,K., Wicke,L., Böttcher,B., Lavigne,R. and Vogel,J. (2024) Phage proteins target and co-opt host ribosomes immediately upon infection. *Nat Microbiol*, **9**, 787–800.
22. Walsh,D. and Mohr,I. (2011) Viral subversion of the host protein synthesis machinery. *Nat Rev Microbiol*, **9**, 860–875.
23. Nagy,P.D. and Pogany,J. (2012) The dependence of viral RNA replication on co-opted host factors. *Nat Rev Microbiol*, **10**, 137–149.
24. van Raaij,M.J. (2024) Bacteriophage Receptor Recognition and Nucleic Acid Transfer. In Mateu,M.G. (ed), *Structure and Physics of Viruses: An Integrated Guide*. Springer Nature Switzerland, Cham, pp. 593–628.
25. Grove,J. and Marsh,M. (2011) The cell biology of receptor-mediated virus entry. *J Cell Biol*, **195**, 1071–1082.
26. Baranowski,E., Ruiz-Jarabo,C.M. and Domingo,E. (2001) Evolution of cell recognition by viruses. *Science*, **292**, 1102–1105.

27. Young,R. (1992) Bacteriophage lysis: mechanism and regulation. *Microbiological Reviews*, **56**, 430–481.
28. Garoff,H., Hewson,R. and Opstelten,D.-J.E. (1998) Virus Maturation by Budding. *Microbiol Mol Biol Rev*, **62**, 1171–1190.
29. Fenner,F. (1976) The Classification and Nomenclature of Viruses. *Journal of General Virology*, **31**, 463–470.
30. Adams,M.J., Lefkowitz,E.J., King,A.M.Q., Harrach,B., Harrison,R.L., Knowles,N.J., Kropinski,A.M., Krupovic,M., Kuhn,J.H., Mushegian,A.R., *et al.* (2017) 50 years of the International Committee on Taxonomy of Viruses: progress and prospects. *Arch Virol*, **162**, 1441–1446.
31. Camargo,A.P., Nayfach,S., Chen,I.-M.A., Palaniappan,K., Ratner,A., Chu,K., Ritter,S.J., Reddy,T.B.K., Mukherjee,S., Schulz,F., *et al.* (2023) IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res*, **51**, D733–D743.
32. Filippini,M. and Middelboe,M. (2007) Viral abundance and genome size distribution in the sediment and water column of marine and freshwater ecosystems. *FEMS Microbiol Ecol*, **60**, 397–410.
33. Vincent,F. and Vardi,A. (2023) Viral infection in the ocean—A journey across scales. *PLOS Biology*, **21**, e3001966.
34. Suttle,C.A. (2007) Marine viruses — major players in the global ecosystem. *Nat Rev Microbiol*, **5**, 801–812.
35. Gazitúa,M.C., Gazitúa,M.C., Vik,D.R., Roux,S., Gregory,A.C., Bolduc,B., Widner,B., Mulholland,M.R., Hallam,S.J., Ulloa,O., *et al.* (2020) Potential virus-mediated nitrogen cycling in oxygen-depleted oceanic waters. *The ISME Journal*, **15**, 981–998.
36. Gao,Y., Lu,Y., Dungait,J.A.J., Liu,J., Lin,S., Jia,J. and Yu,G. (2022) The "Regulator" Function of Viruses on Ecosystem Carbon Cycling in the Anthropocene. *Front Public Health*, **10**, 858615.
37. Wilhelm,S.W. and Suttle,C.A. (1999) Viruses and Nutrient Cycles in the Sea: Viruses play critical roles in the structure and function of aquatic food webs. *BioScience*, **49**, 781–788.

38. Forterre,P. and Prangishvili,D. (2009) The Great Billion-year War between Ribosome- and Capsid-encoding Organisms (Cells and Viruses) as the Major Source of Evolutionary Novelties. *Annals of the New York Academy of Sciences*, **1178**, 65–77.
39. Arnold,B.J., Huang,I.-T. and Hanage,W.P. (2022) Horizontal gene transfer and adaptive evolution in bacteria. *Nat Rev Microbiol*, **20**, 206–218.
40. Waldor,M.K. and Mekalanos,J.J. (1996) Lysogenic Conversion by a Filamentous Phage Encoding Cholera Toxin. *Science*, **272**, 1910–1914.
41. Parker,C., Richardson,S.H. and Romig,W.R. (1970) Production of Bacteriophage-Associated Materials by Vibrio cholerae: Possible Correlation with Pathogenicity. *Infect Immun*, **1**, 417–420.
42. Ginzberg,D., Padan,E. and Shilo,M. (1976) Metabolic aspects of LPP cyanophage replication in the cyanobacterium *Plectonema boryanum*. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, **423**, 440–449.
43. Forterre,P. (2013) The virocell concept and environmental microbiology. *ISME J*, **7**, 233–236.
44. Howard-Varona,C., Lindback,M.M., Bastien,G.E., Solonenko,N., Zayed,A.A., Jang,H., Andreopoulos,B., Brewer,H.M., Glavina del Rio,T., Adkins,J.N., *et al.* (2020) Phage-specific metabolic reprogramming of virocells. *ISME J*, **14**, 881–895.
45. Lindell,D., Sullivan,M.B., Johnson,Z.I., Tolonen,A.C., Rohwer,F. and Chisholm,S.W. (2004) Transfer of photosynthesis genes to and from Prochlorococcus viruses. *PNAS*, **101**, 11013–11018.
46. Mann,N.H., Cook,A., Millard,A., Bailey,S. and Clokie,M. (2003) Bacterial photosynthesis genes in a virus. *Nature*, **424**, 741–741.
47. Lindell,D., Jaffe,J.D., Johnson,Z.I., Church,G.M. and Chisholm,S.W. (2005) Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, **438**, 86–89.
48. Raoult,D., Audic,S., Robert,C., Abergel,C., Renesto,P., Ogata,H., La Scola,B., Suzan,M. and Claverie,J.-M. (2004) The 1.2-Megabase Genome Sequence of Mimivirus. *Science*, **306**, 1344–1350.
49. Abrahão,J., Silva,L., Silva,L.S., Khalil,J.Y.B., Rodrigues,R., Arantes,T., Assis,F., Boratto,P., Andrade,M., Kroon,E.G., *et al.* (2018) Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. *Nat Commun*, **9**, 749.

50. Leavitt,J.C., Woodbury,B.M., Gilcrease,E.B., Bridges,C.M., Teschke,C.M. and Casjens,S.R. (2024) Bacteriophage P22 SieA-mediated superinfection exclusion. *mBio*, **15**, e0216923.
51. Li,Q., Zhou,M., Fan,X., Yan,J., Li,W. and Xie,J. (2016) Mycobacteriophage SWU1 gp39 can potentiate multiple antibiotics against Mycobacterium via altering the cell wall permeability. *Sci Rep*, **6**, 28701.
52. Lindback,M.M., Howard-Varona,C., Fudyma,J., Tfaily,M., Sullivan,M.B. and Duhaime,M.B. (2025) Virocell resource manipulation under nutrient limitation. *mSystems*, **10**, e00521-25.
53. Howard-Varona,C., Lindback,M.M., Fudyma,J.D., Krongauz,A., Solonenko,N.E., Zayed,A.A., Andreopoulos,W.B., Olson,H.M., Kim,Y.-M., Kyle,J.E., *et al.* (2024) Environment-specific virocell metabolic reprogramming. *ISME J*, **18**, wrae055.
54. Dheilly,N.M. (2014) Holobiont–Holobiont Interactions: Redefining Host–Parasite Interactions. *PLOS Pathogens*, **10**, e1004093.
55. Knowles,B., Bonachela,J.A., Behrenfeld,M.J., Bondoc,K.G., Cael,B.B., Carlson,C.A., Cieslik,N., Diaz,B., Fuchs,H.L., Graff,J.R., *et al.* (2020) Temperate infection in a virus-host system previously known for virulent dynamics. *Nat Commun*, **11**, 4626.
56. Zimmerman,A.E., Howard-Varona,C., Needham,D.M., John,S.G., Worden,A.Z., Sullivan,M.B., Waldbauer,J.R. and Coleman,M.L. (2020) Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nat Rev Microbiol*, **18**, 21–34.
57. Laber,C.P., Hunter,J.E., Carvalho,F., Collins,J.R., Hunter,E.J., Schieler,B.M., Boss,E., More,K., Frada,M., Thamatrakoln,K., *et al.* (2018) Coccolithovirus facilitation of carbon export in the North Atlantic. *Nat Microbiol*, **3**, 537–547.
58. Irwin,N.A.T., Pittis,A.A., Richards,T.A. and Keeling,P.J. (2022) Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat Microbiol*, **7**, 327–336.
59. Weitz,J.S., Stock,C.A., Wilhelm,S.W., Bourouiba,L., Coleman,M.L., Buchan,A., Follows,M.J., Fuhrman,J.A., Jover,L.F., Lennon,J.T., *et al.* (2015) A multitrophic model to quantify the effects of marine viruses on microbial food webs and ecosystem processes. *The ISME Journal*, **9**, 1352–1364.

60. Danovaro,R., Corinaldesi,C., Dell'Anno,A., Fuhrman,J.A., Middelburg,J.J., Noble,R.T. and Suttle,C.A. (2011) Marine viruses and global climate change. *FEMS Microbiol Rev*, **35**, 993–1034.

61. Bergh,Ø., BØrsheim,K.Y., Bratbak,G. and Heldal,M. (1989) High abundance of viruses found in aquatic environments. *Nature*, **340**, 467–468.

62. Proctor,L.M. and Fuhrman,J.A. (1990) Viral mortality of marine bacteria and cyanobacteria. *Nature*, **343**, 60–62.

63. Fuhrman,J.A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature*, **399**, 541–548.

64. Welp,L.R., Keeling,R.F., Meijer,H.A.J., Bollenbacher,A.F., Piper,S.C., Yoshimura,K., Francey,R.J., Allison,C.E. and Wahlen,M. (2011) Interannual variability in the oxygen isotopes of atmospheric CO2 driven by El Niño. *Nature*, **477**, 579–582.

65. Hutchinson,G.E. The Paradox of the Plankton. *THE AMERICAN NATURALIST*.

66. Thingstad,T.F. (2000) Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnology and Oceanography*, **45**, 1320–1328.

67. Ignacio-Espinoza,J.C., Ahlgren,N.A. and Fuhrman,J.A. (2020) Long-term stability and Red Queen-like strain dynamics in marine viruses. *Nat Microbiol*, **5**, 265–271.

68. Williamson,S.J., Rusch,D.B., Yooseph,S., Halpern,A.L., Heidelberg,K.B., Glass,J.I., Andrews-Pfannkoch,C., Fadrosh,D., Miller,C.S., Sutton,G., *et al.* (2008) The Sorcerer II Global Ocean Sampling Expedition: Metagenomic Characterization of Viruses within Aquatic Microbial Samples. *PLOS ONE*, **3**, e1456.

69. Brum,J.R., Ignacio-Espinoza,J.C., Roux,S., Doulcier,G., Acinas,S.G., Alberti,A., Chaffron,S., Cruaud,C., Vargas,C. de, Gasol,J.M., *et al.* (2015) Patterns and ecological drivers of ocean viral communities. *Science*, **348**.

70. Gregory,A.C., Zayed,A.A., Conceição-Neto,N., Temperton,B., Bolduc,B., Alberti,A., Ardyna,M., Arkhipova,K., Carmichael,M., Cruaud,C., *et al.* (2019) Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*, **177**, 1109-1123.e14.

71. Santiago-Rodriguez,T.M. and Hollister,E.B. (2022) Unraveling the viral dark matter through viral metagenomics. *Front Immunol*, **13**, 1005107.

72. Breitbart,M., Salamon,P., Andresen,B., Mahaffy,J.M., Segall,A.M., Mead,D., Azam,F. and Rohwer,F. (2002) Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences*, **99**, 14250–14255.

73. Roux,S., Hallam,S.J., Woyke,T. and Sullivan,M.B. (2015) Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife*, **4**, e08490.

74. Zhong,Z.-P., Tian,F., Roux,S., Gazitúa,M.C., Solonenko,N.E., Li,Y.-F., Davis,M.E., Van Etten,J.L., Mosley-Thompson,E., Rich,V.I., *et al.* (2021) Glacier ice archives nearly 15,000-year-old microbes and phages. *Microbiome*, **9**, 160.

75. Danovaro,R., Dell'Anno,A., Corinaldesi,C., Magagnini,M., Noble,R., Tamburini,C. and Weinbauer,M. (2008) Major viral impact on the functioning of benthic deep-sea ecosystems. *Nature*, **454**, 1084–1087.

76. Nicolas,A.M., Sieradzki,E.T., Pett-Ridge,J., Banfield,J.F., Taga,M.E., Firestone,M.K. and Blazewicz,S.J. (2023) A subset of viruses thrives following microbial resuscitation during rewetting of a seasonally dry California grassland soil. *Nat Commun*, **14**, 5835.

77. Tsai,Y.L., Tran,B. and Palmer,C.J. (1995) Analysis of viral RNA persistence in seawater by reverse transcriptase-PCR. *Appl Environ Microbiol*, **61**, 363–366.

78. Dalcin Martins,P., Danczak,R.E., Roux,S., Frank,J., Borton,M.A., Wolfe,R.A., Burris,M.N. and Wilkins,M.J. (2018) Viral and metabolic controls on high rates of microbial sulfur and carbon cycling in wetland ecosystems. *Microbiome*, **6**, 138.

79. Bratbak,G., Heldal,M., Thingstad,T.F., Riemann,B. and Haslund,O.H. (1992) Incorporation of viruses into the budget of microbial C-transfer. A first approach. *Marine Ecology Progress Series*, **83**, 273–280.

80. Pace,N.R. (1997) A Molecular View of Microbial Diversity and the Biosphere. *Science*, **276**, 734–740.

81. Ward,D.M., Weller,R. and Bateson,M.M. (1990) 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature*, **345**, 63–65.

82. Staley,J.T. and Konopka,A. (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol*, **39**, 321–346.

83. Robins,H., Krasnitz,M., Barak,H. and Levine,A.J. (2005) A Relative-Entropy Algorithm for Genomic Fingerprinting Captures Host-Phage Similarities. *J Bacteriol*, **187**, 8370–8374.

84. Ahmed,S., Saito,A., Suzuki,M., Nemoto,N. and Nishigaki,K. (2009) Host–parasite relations of bacteria and phages can be unveiled by Oligostickiness, a measure of relaxed sequence similarity. *Bioinformatics*, **25**, 563–570.

85. Ahlgren,N.A., Ren,J., Lu,Y.Y., Fuhrman,J.A. and Sun,F. (2017) Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res*, **45**, 39–53.

86. Galiez,C., Siebert,M., Enault,F., Vincent,J. and Söding,J. (2017) WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics*, **33**, 3113–3114.

87. Young,F., Rogers,S. and Robertson,D.L. (2020) Predicting host taxonomic information from viral genomes: A comparison of feature representations. *PLOS Computational Biology*, **16**, e1007894.

88. Zhang,M., Yang,L., Ren,J., Ahlgren,N.A., Fuhrman,J.A. and Sun,F. (2017) Prediction of virus-host infectious association by supervised learning methods. *BMC Bioinformatics*, **18**, 60.

89. Wang,W., Ren,J., Tang,K., Dart,E., Ignacio-Espinoza,J.C., Fuhrman,J.A., Braun,J., Sun,F. and Ahlgren,N.A. (2020) A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genom Bioinform*, **2**.

90. Amgarten,D., Iha,B.K.V., Piroupo,C.M., da Silva,A.M. and Setubal,J.C. (2022) vHULK, a New Tool for Bacteriophage Host Prediction Based on Annotated Genomic Features and Neural Networks. *PHAGE*, 10.1089/phage.2021.0016.

91. Roux,S., Camargo,A.P., Coutinho,F.H., Dabdoub,S.M., Dutilh,B.E., Nayfach,S. and Tritt,A. (2022) iPHoP: an integrated machine-learning framework to maximize host prediction for metagenome-assembled virus genomes. 10.1101/2022.07.28.501908.

92. Shang,J. and Sun,Y. (2022) CHERRY: a Computational metHod for accuratE pRediction of virus-pRokarYotic interactions using a graph encoder-decoder model. *Briefings in Bioinformatics*, 10.1093/bib/bbac182.

93. Dion,M.B., Plante,P.-L., Zufferey,E., Shah,S.A., Corbeil,J. and Moineau,S. (2021) Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. *Nucleic Acids Research*, **49**, 3127–3138.

94. Zhang,R., Mirdita,M., Karin,E.L., Norroy,C., Galiez,C. and Soeding,J. (2020) SpacePHARER: Sensitive identification of phages from CRISPR spacers in prokaryotic hosts. *bioRxiv*, 10.1101/2020.05.15.090266.

95. Beumer,A. and Robinson,J.B. (2005) A Broad-Host-Range, Generalized Transducing Phage (SN-T) Acquires 16S rRNA Genes from Different Genera of Bacteria. *Applied and Environmental Microbiology*, 10.1128/AEM.71.12.8301-8304.2005.

96. Feng,X., Yan,W., Wang,A., Ma,R., Chen,X., Lin,T.-H., Chen,Y.-L., Wei,S., Jin,T., Jiao,N., *et al.* (2021) A Novel Broad Host Range Phage Infecting Alteromonas. *Viruses*, **13**, 987.

97. Jensen,E.C., Schrader,H., Rieland,B., Thompson,T.L., Lee,K.W., Nickerson,K.W. and Kokjohn,T.A. (1998) Prevalence of broad-host-range lytic bacteriophages of Sphaerotilus natans, Escherichia coli, and Pseudomonas aeruginosa. *Applied and Environmental Microbiology*, **64**, 575–580.

98. Malki,K., Kula,A., Bruder,K., Sible,E., Hatzopoulos,T., Steidel,S., Watkins,S.C. and Putonti,C. (2015) Bacteriophages isolated from Lake Michigan demonstrate broad host-range across several bacterial phyla. *Virology Journal*, **12**, 164.

99. Keya,S., Kumar,G.S., Subrata,S., Chandra,M.S. and Chandra,G.T. (2007) Studies on Synonymous Codon and Amino Acid Usage Biases in the Broad-Host Range Bacteriophage KVP40. *Journal of Microbiology*, **45**, 58–63.

100. Bandín,I. and Dopazo,C.P. (2011) Host range, host specificity and hypothesized host shift events among viruses of lower vertebrates. *Vet Res*, **42**, 67.

101. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P., *et al.* (2000) A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.

102. May,R.M. (1972) Will a Large Complex System be Stable? *Nature*, **238**, 413–414.

103. Dunne,J.A., Williams,R.J. and Martinez,N.D. (2002) Food-web structure and network theory: The role of connectance and size. *Proceedings of the National Academy of Sciences*, **99**, 12917–12922.

104. Keeling,M.J. and Eames,K.T.D. (2005) Networks and epidemic models. *Journal of The Royal Society Interface*, **2**, 295–307.
105. Bascompte,J. and Jordano,P. (2007) Plant-Animal Mutualistic Networks: The Architecture of Biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 567–593.
106. Thébault,E. and Fontaine,C. (2010) Stability of Ecological Communities and the Architecture of Mutualistic and Trophic Networks. *Science*, **329**, 853–856.
107. Strona,G. and Lafferty,K.D. (2016) Environmental change makes robust ecological networks fragile. *Nat Commun*, **7**, 12462.
108. Solé,R.V. and Montoya,M. (2001) Complexity and fragility in ecological networks. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **268**, 2039–2045.
109. Dunne,J.A., Williams,R.J. and Martinez,N.D. (2002) Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecology Letters*, **5**, 558–567.
110. Newman,M.E.J. (2002) Spread of epidemic disease on networks. *Phys. Rev. E*, **66**, 016128.
111. Lloyd-Smith,J.O., Schreiber,S.J., Kopp,P.E. and Getz,W.M. (2005) Superspreading and the effect of individual variation on disease emergence. *Nature*, **438**, 355–359.
112. Weitz,J.S., Poisot,T., Meyer,J.R., Flores,C.O., Valverde,S., Sullivan,M.B. and Hochberg,M.E. (2013) Phage–bacteria infection networks. *Trends in Microbiology*, **21**, 82–91.
113. Flores,C.O., Meyer,J.R., Valverde,S., Farr,L. and Weitz,J.S. (2011) Statistical structure of host–phage interactions. *PNAS*, **108**, E288–E297.
114. Flores,C.O., Valverde,S. and Weitz,J.S. (2013) Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. *ISME J*, **7**, 520–532.

Chapter 2 – Host Translational Machinery Reflects Viral Codon Usage Bias[1]

2.1 Abstract

Viruses that infect bacteria universally rely on host cellular machinery to successfully replicate. Recent multi-omics studies have shed light on this fundamental dependency, revealing how viruses manipulate and utilize host resources during infection. However, the consistency of these patterns across viruses and their influence on viruses' gene adaptation remains unclear. To address this knowledge gap, we leveraged a curated dataset of lab-verified viruses-host interactions sourced from literature and public databases. This comprehensive resource includes genomic sequences for viruses and hosts, along with infection status. We hypothesized that host resource constraints, particularly those related to translational machinery, have shaped and continue to influence viruses' gene adaptation, resulting in capturable genomic signals. To test this hypothesis, we adapted existing metrics to describe adaptation of viruses to their hosts, which include codon and amino acid usage comparisons, Relative Synonymous Codon Usage (RSCU), and tRNA Amino Acid Accordance Indices (TAAI). Comparing all the viruses and host genes, we found a non-significant correlation between their amino acid usages and a weak correlation between their codon usages, indicating that viruses do not closely mimic host gene codon or amino acid patterns, as previously reported. However, metrics incorporating host tRNA genes (rather than total gene content) encoded a stronger signal, underscoring the critical role of host translational machinery in viruses' gene adaptation. Those results have broad implications. First, this pressure can be leveraged to improve predictive models of virus-

[1] This chapter is based on a manuscript in preparation co-authored with Evelyn Faust (co-first author), Morgan Lindback, Emmalyn Campau, Luis Zaman, and Melissa Duhaime

host interactions from genomic data. Second, it provides deeper insights into the trajectories of viruses-host coevolution. Finally, it offers exciting prospects for synthetic biology, potentially enabling the design of better-optimized viruses' genomes with enhanced expression efficiency and host specificity, paving the way for more effective viruses-based therapeutics and biotechnological applications. By linking genomic features to translational constraints, this study deepens our understanding of viruses' biology and its inherent dependence on host resources.

2.2 Introduction

Viruses of microbes obligately depend on their host cellular machinery and resources to produce progeny, often diverting cellular energy and resources towards viral replication at the expense of host functions (1). To do so, viruses remodel host cell conditions, optimizing them for viral replication and assembly rather than cellular replication. Specific mechanisms include metabolic reprogramming through precise timing of viral gene expression (2–6), recycling of host DNA and RNA nucleotides (2, 3, 7–9), and deployment of anti-sigma factors to redirect transcription from the host to viral genes (10–12). Over evolutionary timescales, these varied mechanisms leave behind genomic imprints—distinctive patterns in nucleotide composition, codon usage, and sequence motifs—that reflect the selective pressures viruses experience while adapting to specific host cellular contexts (13–18). These genomic signatures provide novel insights into viral adaptation strategies, yet distinguishing adaptation signals from background genomic patterns requires comparison between compatible and incompatible virus-host pairs (19).

Previous research has primarily focused on capturing whole-genome features that describe virus adaptation to known hosts. For example, previous studies have identified a persistent dinucleotide compositional bias in viral genomes characterized by a skew toward adenine/thymine (A/T) nucleotides over guanine/cytosine (G/C) (14, 19, 20). Viruses also exhibit specific codon usage biases (21), reflecting the biological constraints imposed by their dependency on the host translational machinery (1, 6, 22, 23). Oligonucleotide-based analyses, which examine frequency patterns of k-length nucleotide sequences, further

demonstrate that viruses tend to evolve genomic profiles that are similar to those of their hosts (24–26). These genomic sequence signatures reveal the evolutionary constraints experienced by viruses due to their dependency on host machinery, thereby offering genomic clues that can be leveraged by models to predict virus-host interactions (19, 24, 27–30).

K-mer approaches have identified broad, genome-wide patterns of viral adaptation (19, 24, 31, 32); however, because they aggregate nucleotide information from across the entire genome, they fail to capture the mechanistic processes by which viruses adapt to their hosts. This interpretive challenge compounds when analyzing viruses that encode machinery-related genes, as they may exhibit different genomic signatures when host machinery constraints are alleviated (33–35). Whole-genome metrics also overlook gene-level information and functional contexts that may be informative for a comprehensive understanding of genome sequence-derived signals. Furthermore, the absence of incompatible virus-host pair comparisons in previous work makes it challenging to accurately determine the strength and biological relevance of these genomic signals (19).

Metrics designed to capture genomic signals by quantifying bacterial self-adaptation—specifically, how well a genome is adapted to its own translational machinery—have not yet been applied to study viral adaptation. Historically, efforts to quantify the relationship between genomic sequences and their corresponding translational machinery have led to the development of metrics such as the Codon Adaptation Index (CAI) (21), tRNA Amino Acid Adaptation Index (TAAI) (36), and Frequency of Optimal Codons (FOP) (37, 38). These metrics provide valuable frameworks for the evaluation of genomic optimization. Notably, despite their potential applicability, these established metrics are rarely used to describe and quantify the adaptation of a specific virus to a specific host. To address the knowledge gap regarding viral genomic signatures, we leveraged VHRnet, a comprehensive and recently updated collection of prokaryotic virus-host interactions. Unlike other resources, VHRnet includes both compatible and incompatible virus-host pairs, providing an opportunity to distinguish true adaptation signals from background genomic patterns. By adapting metrics originally designed for bacterial self-adaptation to the virus-host context,

we sought to evaluate the manner and extent to which viruses optimize their genomes to those of their hosts. This has broad implications for improving virus-host interaction prediction modeling and phage genome optimization.

2.3 Results

2.3.1 Virus and gene annotation

To ensure consistent gene annotation across all viral and host sequences, we used the annotation tool BAKTA rather than relying on existing annotations from the NCBI RefSeq database. BAKTA enables rapid and comprehensive annotation of bacterial genomes and plasmids using a taxonomically untargeted database, allowing annotation even for rare species lacking reference genomes (39). Eight viruses yielded no predicted genes and were excluded from subsequent analyses. Across the remaining 4,393 viral genomes, we identified 383,483 genes, averaging 86.4 genes per genome (Fig S7). The vast majority (>99%) of predicted genetic elements were protein-coding sequences, with approximately half (49.56%) classified as hypothetical proteins of unknown function.

Viral-encoded tRNAs represented the second most abundant class of genetic elements found in viral genomes. We found that 32.71% of viruses encoded at least one tRNA, with an average of 6.5 tRNA genes and a maximum of 42 tRNA genes per viral genome (Fig S8). However, this distribution was heavily skewed toward viruses encoding a single tRNA.

2.3.2 Viral codon usage patterns reflect GC content constraints and adaptation to host

Using predicted genes from BAKTA, we calculated codon usage frequencies for each virus (Eq. 1). Viral genomes exhibited diverse codon usage patterns (Fig 1 and Fig S9). When viruses were organized by GC content, viruses with extreme GC content (either high or low) showed strong preferences for specific codons within each amino acid family, while viruses with moderate GC content displayed more balanced codon usage across synonymous codons.

Given that viruses depend largely on host cellular machinery for replication, we next examined codon usage differences between viruses and their known hosts. Host codon usage frequencies were calculated using the same methodology (Eq. 1), and we computed usage differences for all documented virus-host pairs in the VHRnet dataset as described in Methods. This analysis shows that viruses tend to be more enriched in certain codons and depleted in others relative to their hosts (Fig 2). To determine whether this codon usage bias represents statistically significant non-random behavior, we performed null modeling analysis. For each virus and host in a pair, we shuffled codon usage frequency values and computed the average frequencies for each codon. We then calculated the sum of squared differences between averaged shuffled virus and host codon frequencies to generate a null distribution of expected values. The observed differences differed significantly from the null distribution ($p < 0.0001$, $Z = 3795.16$), providing strong evidence that viral codon preferences are non-random (Fig S10).

Notably, viruses were predominantly A/U-rich, particularly at third codon positions (Fig 3). This pattern suggests that beyond GC content constraints, viruses exhibit a consistent bias toward A/U nucleotides in wobble positions, potentially reflecting adaptive strategies for utilizing host translational machinery.

2.3.3 Viral-encoded tRNAs compensate for virus-host codon usage mismatches

Previous studies have reported that some viruses encode their own tRNAs (33, 40). Subsequent research hypothesized and demonstrated on a small dataset that viral-encoded tRNAs compensate for codon usage mismatches between viruses and their hosts (33, 41). To determine if this trend holds in the expanded VHRnet dataset, we compared codon usage frequency differences (between viruses and hosts) across two categories: viruses encoding codon-complementary tRNA gene(s), and viruses not encoding codon-complementary tRNA genes. Specifically, for each codon, we compared the virus-host codon frequency difference between two groups: virus-host pairs where the virus encodes a tRNA complementary to that codon, versus pairs where the virus lacks such tRNA. We found that viruses encoding a tRNA for a given codon showed significantly higher usage of

that codon relative to their hosts compared to viruses lacking the corresponding tRNA (Fig 4 and Fig S11). This pattern indicates that viruses selectively encode tRNAs for codons they use more frequently than their hosts do, supporting previous findings that viral tRNAs help alleviate translational bottlenecks arising from virus-host codon usage disparities (33).

2.3.4 Viral genomes exhibit measurable adaptation to host cellular machinery

To quantify viral genome optimization for host cellular machinery, we modified an established bacterial self-adaptation metrics: tRNA Amino Acid Accordance Index (TAAI) (36). This metric was originally designed to quantify how well bacterial genomes are optimized for their own translational machinery, providing a framework with the potential to evaluate virus-host compatibility. We adapted TAAI to assess the accordance between viral amino acid requirements and host tRNA pool composition, capturing the broader translational demands that viruses place on their hosts. This metric directly addresses the fundamental biological constraints of viral replication.

To evaluate whether TAAO could distinguish biologically meaningful virus-host relationships, we compared the distribution between experimentally validated compatible and incompatible virus-host pairs from the VHRnet dataset. In addition, we also included other metrics, namely amino acid frequency, codon usage, and relative synonymous codon usage bias (RSCU). Although the difference between the compatible and incompatible pairs were significant, the degree by which they differentiated was small, especially when compared to the TAAI metric (Fig 5). Statistical analysis using the Wilcoxon rank-sum test revealed highly significant differences between compatible and incompatible pairs ($p < 0.001$). Notably, TAAI demonstrated the strongest discriminatory power with a large effect size (Cliff's delta = 0.68).

As demonstrated earlier, one third of the viruses encode their own tRNAs. We compared the complement of the virus to the host tRNA pool, and to the total pool of tRNAs which includes both host and virus-encoded tRNAs. Because viruses typically encode 1 or just a

few tRNAs, we found that the difference between those two metrics to be minimal (Fig S12).

2.4 Discussion

Our gene-level analysis revealed a striking preference for adenine (A) and uracil (U) codons in viral genomes compared to their hosts, particularly at the wobble position. This third codon position, which mediates 30-40% of all codon recognition events (42), offers crucial translational flexibility that viruses have evolved to exploit. While wobble base pairing itself is well-established, the remarkable degree of virus-host complementarity at this position represents a novel finding. This A/U bias confers multiple evolutionary advantages essential for compact viral genomes: it optimizes alignment with host tRNA pools (43), enables overlapping gene architecture through simultaneous multi-frame optimization (44, 45), and supports dual-coding sequences that maintain both protein-coding capacity and functional RNA secondary structures (46). This discovery reveals wobble position bias as a fundamental mechanism of viral translational resource optimization, suggesting that viruses fine-tune their codon usage patterns to hijack host translation machinery more efficiently than previously recognized. Interestingly, this translational optimization manifests differently across analytical metrics. While amino acid frequency profiles showed limited predictive power for identifying interaction (Cliff's delta = 0.21), codon-based metrics revealed a more nuanced picture. TAAI—which directly measures tRNA pool compatibility—emerged as specifically robust in distinguishing viral from host sequences (Cliff's delta = 0.68), underscoring the critical correlation of tRNA availability in viral codon usage bias.

The prevalence of viral-encoded tRNAs in approximately one-third of viruses (n=1437 out of 4393 viruses) further supports the observed pattern of virus-host translational resource optimization. Such viral-encoded tRNAs may act as a mechanism to circumvent host translational defense that could otherwise limit viral protein synthesis (47, 48). Excessive similarity in codon usage between virus and host can also be detrimental to the virus, as high viral expression can deplete host tRNA pools for heavily used codons and impair host

cellular functions that the virus also needs intact to successfully replicate (23, 49, 50), creating selective pressure on shared translational resources. We hypothesize that this intracellular resource competition has given rise to the pattern whereby virally encoded tRNAs are biased toward codons heavily used by the virus, relieving the virus of host tRNA limitations. Supporting this intracellular resource competition hypothesis, a previous study using a limited virus-host dataset suggested that viruses encode tRNAs to supplement host tRNA pools for codons that are more frequently used by the virus than the host (33, 51). Our analysis supported this trend across a more comprehensive virus-host dataset (n=10,228), demonstrating that viral tRNA selection is a targeted adaptation strategy rather than random tRNA retention.

Given these observed trends, we sought to develop a metric that explicitly quantifies the degree to which viral codon preferences align with host translational capacity. To this end, we adapted a metric originally designed to measure self-adaptation at the amino acid level – namely, the tRNA Amino Acid Accordance Index (TAAI) – to describe virus adaptation to host translational machinery at the amino acid level (TAAI). TAAI showed strong discriminatory power in differentiating compatible and incompatible virus-host pairs. This finding demonstrates that metrics based on the biological mechanisms of viral adaptation yield more robust signals than genome-wide approaches such as k-mer frequency correlations, which capture aggregate patterns that are more prone to be obscured by genomic noise.

Our analysis revealed that viral adaptation to hosts operates through multiple, interconnected mechanisms rather than a single evolutionary pathway (Fig 6). While previous studies have shown that viruses ameliorate k-mer composition and are typically AT-rich relative to their hosts (19, 20, 52), here we demonstrated that viruses also adapt their amino acid usage to better match host tRNA pools. This multifaceted strategy elicits the need to consider multiple evolutionary constraints and adaptation pathways. These findings have direct implications for phage engineering, presumably enabling enhanced viral fitness and adaptation to specific target hosts via rational modification of viral genomes to achieve more comprehensive alignment with host cellular resources. To

provide further insight into the processes underlying virus-host coevolution and the timing of viral adaptation events, directed evolution experiments could be used to mechanistically validate the patterns revealed here through comparative genomics approaches. For instance, such experimental approaches could demonstrate whether viruses evolve improved amino acid-tRNA matching when subjected to specific hosts, thereby confirming the adaptive significance of the patterns observed across diverse virus-host pairs. The genomic signatures discussed above reveal novel adaptive strategies that offer a foundation framework for future integrative work in biotechnology and microbial ecology, with applications ranging from improved phage engineering to enhanced ecological community modeling.

2.5 Methods

2.5.1 Data acquisition of genomes and host range

The VHRnet dataset was used for this comparative genomic study, which contains both infection (n=4031) and non-infection (n=6197) virus-host pairs from 4402 unique hosts and 835 unique viruses. The VHRnet dataset is a compiled collection of host range data from public databases and literature and its strength lies in its inclusion of both infection and non-infection pairs. Genomes for the viruses and hosts were retrieved using the NCBI accession number from VHRnet.

2.5.2 Gene annotation of bacterial and viral genomes

To maintain uniformity in the source of genes used for feature extraction and analysis, we opted to annotate genes of all viruses and hosts in VHRNet using Bakta (version 1.10.4). This computational tool was designed for rapid and comprehensive annotation of bacterial genomes and plasmids, leveraging a taxonomically untargeted database to provide annotations even for rare species without reference genomes (39). This database also includes phage-related unique protein sequences. We used the lightweight version of the database (db-light version 5.1) for optimal run times and storage requirements with minimal sacrifice to annotation completeness. Bakta also detects and filters short open-

reading-frames (sORFs < 30 amino acids in length), as well as tRNA genes via the integrated tRNAScan-SE tool (53).

Using Bakta, we annotated a mean of 4959.07 (SD = 10156.28) coding DNA sequence (CDS) genes for each host and 85.67 (SD = 78.42) CDS genes for each virus in the VHRnet dataset. Of these, an average of 0.02% (SD = 0.15%) of host CDS genes and 0.02% (SD = 4.08%) of virus CDS genes were excluded from feature calculations because they contained unknown translated amino acids residues and imprecise codons with degenerate nucleotides. Eight total viruses were discarded due to a lack of predicted CDS gene annotations. Bakta also predicted a mean of 80.47 (SD = 59.43) tRNA genes for each host and 2.13 (SD = 5.98) tRNA genes for each virus (with a notable right-skewed distribution visible in Fig. [#] 'tRNA distribution in viral genomes'). Of these, a mean of 1.4% (SD = 2.7%) of host tRNA genes and 18.3% (SD = 32.4%) of virus tRNA genes were excluded from feature calculations due to uncertainty in annotations designated as 'pseudogene' or lacking an associated amino acid.

2.5.3 Codon usage frequency calculations

For each host and virus genome, we computed the total number of each codon from all predicted genes. We calculated relative codon usage frequency as:

$$C_i = \frac{n_i}{N} \qquad (Equation\ 1)$$

- Where n_i is the number of occurrences of codon i and N is the total number of codons in the analyzed gene sequences.

We then computed the difference in relative codon usage frequency between known virus-host pairs. To visualize these differences, we developed custom code to generate a codon wheel, where the inner circle represents 1st nucleotide position, middle circle the 2nd nucleotide position, and the outer circle the 3rd nucleotide position. The purple color represents codon enrichment in the virus and orange represents enrichment in the host. The visualization displays the standardized difference, which is defined as such:

$$standardized\ difference = \frac{(C_{virus} - C_{host})}{\sigma_{host}} \qquad (Equation\ 2)$$

- Where C_{virus} and C_{host} are the codon frequencies of virus and host, respectively. And σ_{host} is the standard deviation of host codon usage frequency.

To determine if there was structure in the virus-host codon usage bias difference, we ran a null model analysis, where the difference within pairs were randomly shuffled but keeping the amino acid encoding the same (so only synonymous codons could have their values swapped). To summarize the structure of the heatmap we computed the sum of squared differences across codons.

2.5.4 Virus-encoded tRNAs and associated analysis

We determined the prevalence of tRNAs in viral genomes using the output from Bakta, which employs tRNAscan-SE, a tool specifically designed to predict tRNA genes (53). Pseudo-tRNAs were subsequently removed. To investigate whether viral-encoded tRNAs mitigate codon usage mismatches between viruses and their hosts, we compared codon usage patterns between tRNA-encoding and non-tRNA-encoding viruses in our dataset. For each virus-host pair in VHRnet, we calculated host-relative viral codon usage as the difference between viral and host codon frequencies for all 61 sense codons. We then categorized virus-host pairs based on whether the virus encoded tRNAs complementary to each specific codon. For each codon, we calculated the mean host-relative viral codon usage for two groups: (1) virus-host pairs where the virus encodes tRNAs complementary to that codon, and (2) pairs where the virus does not encode complementary tRNAs. The difference between these group means represents the effect of viral tRNA encoding on codon usage bias for each codon (Fig. 4). Statistical significance of codon usage differences between tRNA-encoding and non-tRNA-encoding virus groups was assessed using the Mann-Whitney U test.

2.5.5 Codon bias feature extraction

To describe virus adaptation to potential hosts, we developed metrics to analyze signals from Bakta-annotated genes. We first generated common metrics like codon frequency, amino acid frequency, and relative synonymous codon usage (RSCU).

We calculated codon frequency profiles for each virus and host in VHRnet using Equation 1, where n_i is the number of occurrences of codon i, and N is the total number of relevant codons present. Codon frequency was calculated for all codons in three different subsets: (1) frequencies of all 65 codons (64 unique codons and the special case of intragenic selenocysteine-encoding TGA), (2) frequencies of amino acid encoding codons (i.e. total codons excluding stop codons), and (3) frequencies of degenerate codons (i.e. codons belonging to synonymous codon families with >1 members encoding the same amino acid or stop signal). In the descriptions of each downstream derived feature, we specify which codon frequency variation is used.

We also calculated amino acid frequency profiles for all hosts and viruses in VHRnet using Equation 3 below, where n_i is the number of occurrences of amino acid i, and N is the total number of amino acids encoded by an individual's analyzed genes. Amino acid frequencies were generated for all 21 amino acids including selenocysteine.

$$A_i = \frac{n_i}{N} \qquad (Equation\ 3)$$

To account for synonymous codon families that each encode the same amino acid (also called degenerate codons), we calculated Relative Synonymous Codon Usage (RSCU) profiles for all viruses and hosts using Equation 4 (21). In the RSCU equation below, $n_{i,a}$ is the number of occurrences of codon i encoding amino acid a, N_a is the number of occurrences of all codons encoding amino acid a, and k_a is the number of synonymous codon family members that encode amino acid a. The denominator $N_a\ /\ k_a$ is the expected frequency of codon i assuming equally abundant synonymous a-encoding codons. We calculated RSCU for all codons.

$$RSCU_i = \frac{n_{i,a}}{N_a/k_a} \qquad (Equation\ 4)$$

We then developed metrics to compare codon frequency, amino acid frequency, and RSCU profiles between each virus and host paired in VHRnet. First we performed linear regressions between virus and host profiles of the same type (i.e. codon frequencies, amino acid frequencies, or RSCU values) using the 'polyfit' function available through the numpy module (1.26.4) in Python (version 3.12.4). We then extracted the slope and R2 values to function as scores of virus-host matching at the codon and amino acid levels and input features for machine learning using VHIP (54).

2.5.6 tRNA-informed feature extraction

The aforementioned codon bias features describe similarity between virus and host gene content, representing general resource use alignment for gene maintenance and expression. However, we also developed features to represent virus-host alignment within specific resource constraints – namely, tRNA availability. To estimate tRNA availability in a given virocell, we first calculated tRNA gene copy frequencies at both the amino acid and codon/anticodon levels. We calculated T_a (or the frequency of tRNA isoacceptor genes associated with carrying amino acid a) according to Equation 5.1, where n_a is the copy number of tRNA genes carrying amino acid a, and N is the the total number of tRNA genes present in the genome. tRNA frequencies were computed for all 21 associated amino acids. We also calculated T_i (or the frequency of tRNA genes with a precise anticodon complementary to codon i) according to Equation 5.2, where n_i is the copy number of tRNA genes complementary to codon i, and N is the total number of tRNA genes present in the genome. tRNA frequencies were computed for all anticodon-complementary codons that encode amino acids. We calculated tRNA frequency profiles for all viruses and hosts in VHRnet both at the amino acid (T_a) and anticodon levels (T_i) in two different contexts: host tRNA genes alone, or both host and virus tRNA genes together.

$$T_a = \frac{n_a}{N} \qquad (Equation\ 5.1)$$

$$T_i = \frac{n_i}{N} \qquad (Equation\ 5.2)$$

To measure the degree of accordance between viral translational demand (i.e. amino acid frequencies) and the available tRNA pool, we modified an existing metric called the tRNA Amino Acid Accordance Index or TAAI, originally developed by Du. et. al. (36) to measure translational self-optimization within singular uninfected organisms. To do this, we performed a Spearman Rank Correlation between all amino acid frequencies in the virus (vector of A_i values from Equation 3) and their cognate tRNA frequencies at the amino acid level (vector of T_a values from Equation 5.1) in the host or virocell (denoted 'host*') using the 'spearmanr' function available in the scipy.stats Python module (1.13.1). The correlation coefficient (TAAI, defined in Equation 6) was calculated for every virus-host pairing in VHRnet and functions as a score of resource matching at the tRNA level. For TAAI, values closer to 1 indicate more alignment between viral amino acid or codon frequencies and tRNA availability in the host or virocell environment.

$$TAAI = Spearman(virus\{A_i\}, host\{T_a\}) \qquad (Equation\ 6)$$

2.5.7 Statistical test to assess strength of genomic imprints

To determine whether observed genomic signatures reflect genuine viral adaptation rather than random variation, we used Wilcoxon rank-sum tests to compare codon usage, amino acid composition, relative synonymous codon usage (RSCU), and tRNA adaptation index (TAAI) between infection and non-infection events in VHRnet. Effect sizes were quantified using Cliff's delta as a post-hoc measure to assess the magnitude of differences between infection and non-infection conditions.

2.6 Figures

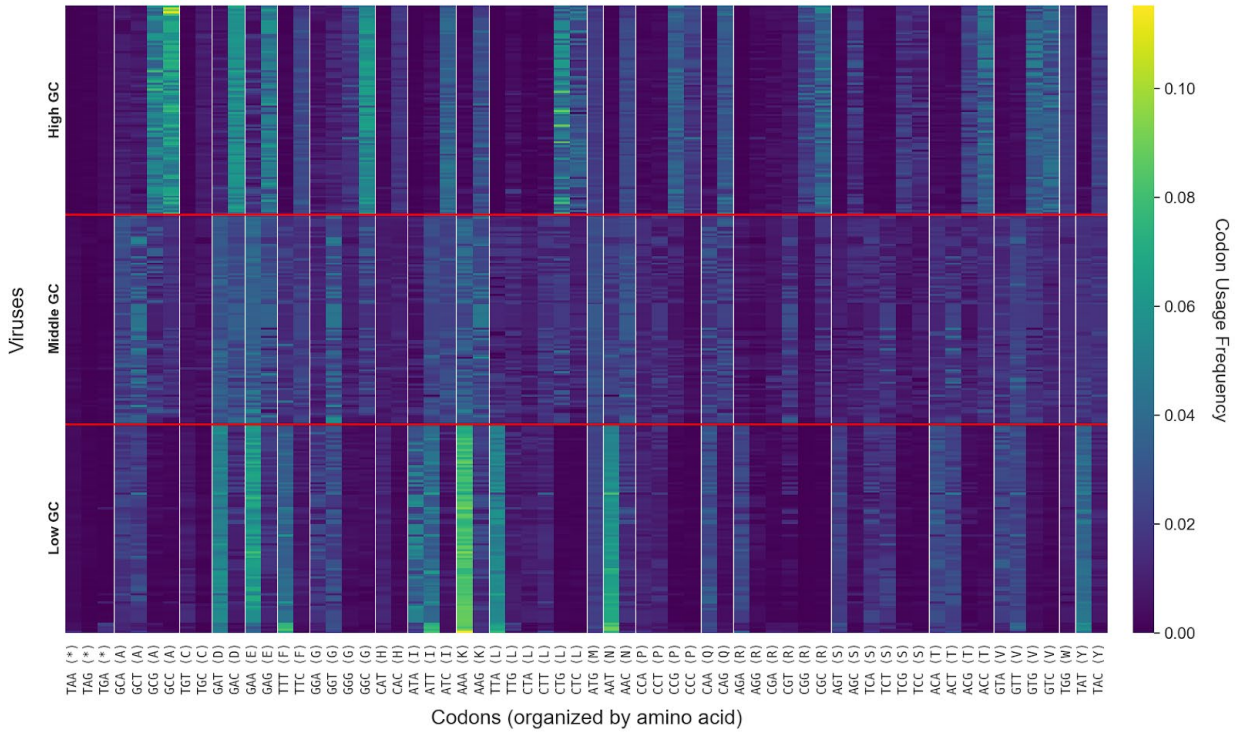


Figure 2.1. Viruses exhibit diverse codon usage patterns

Each column represents one of the 64 possible codons. Each row represents a unique virus. The color (from blue to yellow) shows the relative frequency of that codon across all its codons across each row. This figure is showing a subset of all the viruses in VHRnet, with XXX number of viruses with high GC content, middle GC content, and low GC content.

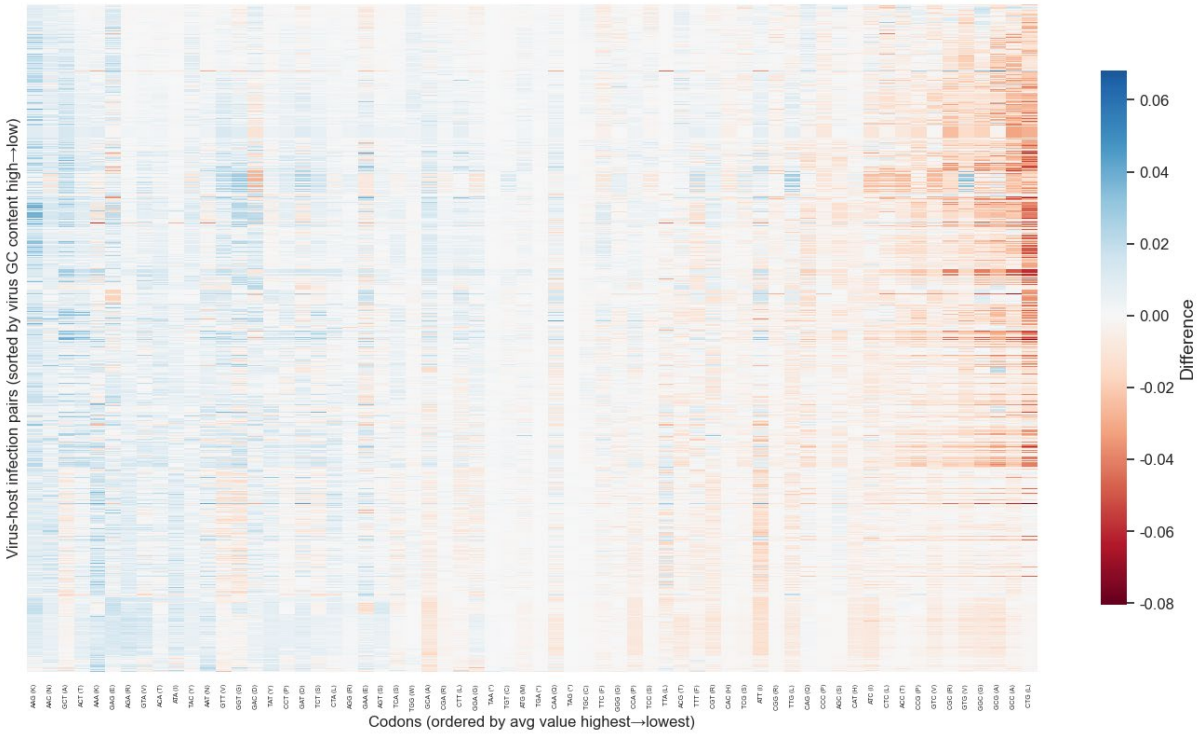


Figure 2.2. Viruses have a general tendency to be enriched in certain codons and poor in others relative to their known hosts.

Each column represents one of the 64 possible codons, which are organized from most enriched in the virus to least enriched (left to right). Each row is a virus-host pair from VHRnet. The color represents the difference in the codon usage frequency between the virus and its known hosts.

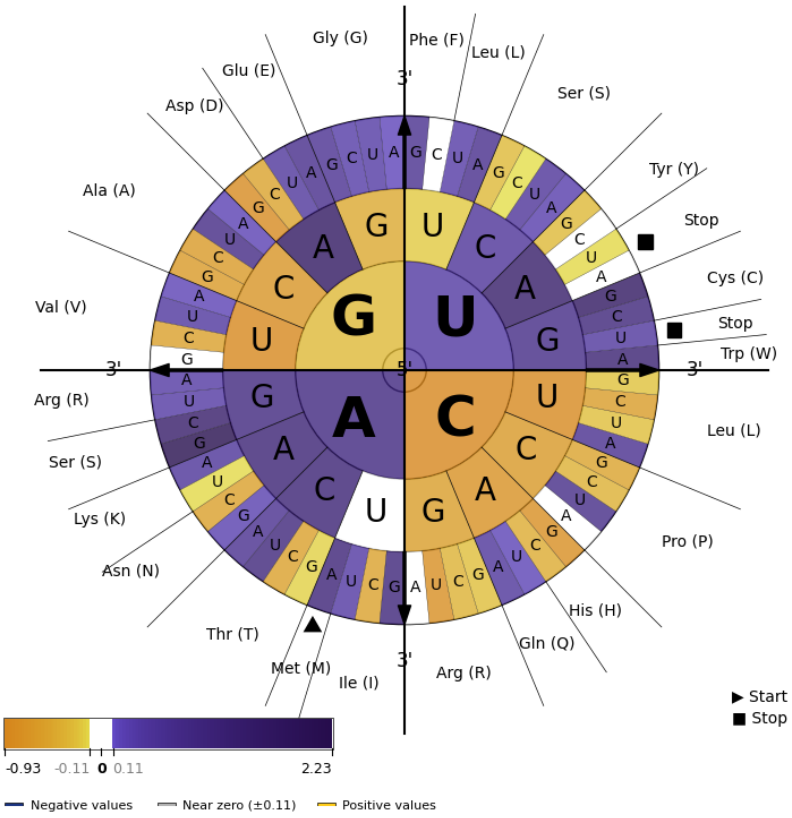


Figure 2.3. Viruses are enriched in A/U over G/C in their codons relative to their known hosts.

The wheel displays the standard genetic code organized by codon position, reading from 5' (center) to 3' (periphery). The three concentric rings represent the first (innermost), second (middle), and third (outermost) nucleotide positions of each codon. Colors shows nucleotide position or codon preference of viruses related to their known host (using equation 2), where purple represents preference in virus genome and yellow in host genome

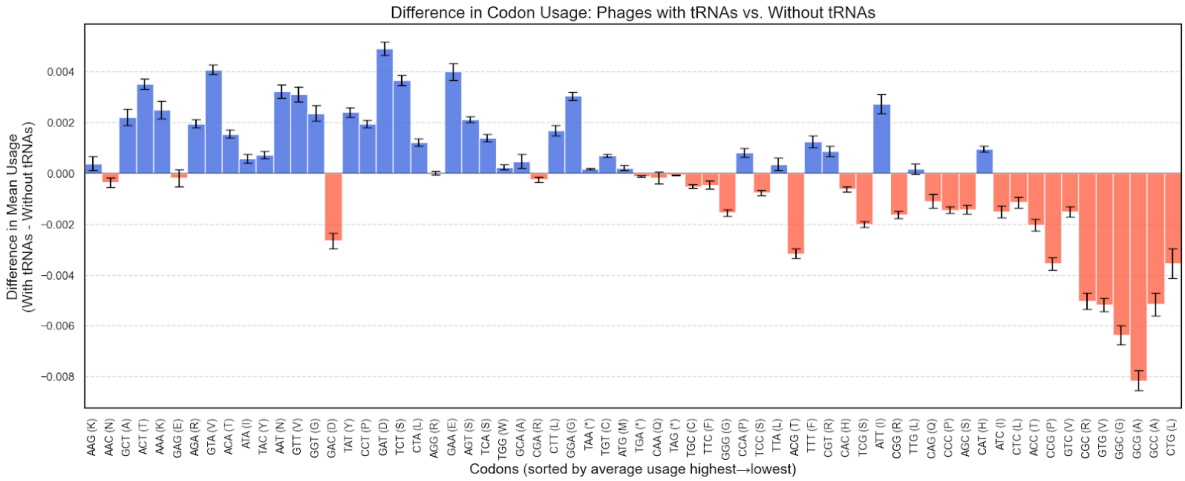


Figure 2.4. Viruses encode tRNAs that minimize mismatch with their known host(s).

Each column represents one the 64 possible codons, which are organized from most enriched in the virus to least enriched (left to right). The height of the bars represents the difference between mean host-relative codon usage in viruses encoding complementary-codon tRNA gene(s), and mean host-relative codon usage in viruses that do not encode such tRNA genes. Error bars indicate the range of possible differences defined by the Mann-Whitney U test. The blue color indicates when a codon is, on average, host-relatively more frequent in viruses that encode tRNAs complementary to that codon compared to viruses that do not. Blue bars are shown more localized on the left side of the plot, which also shows codons that are more enriched in the virus.

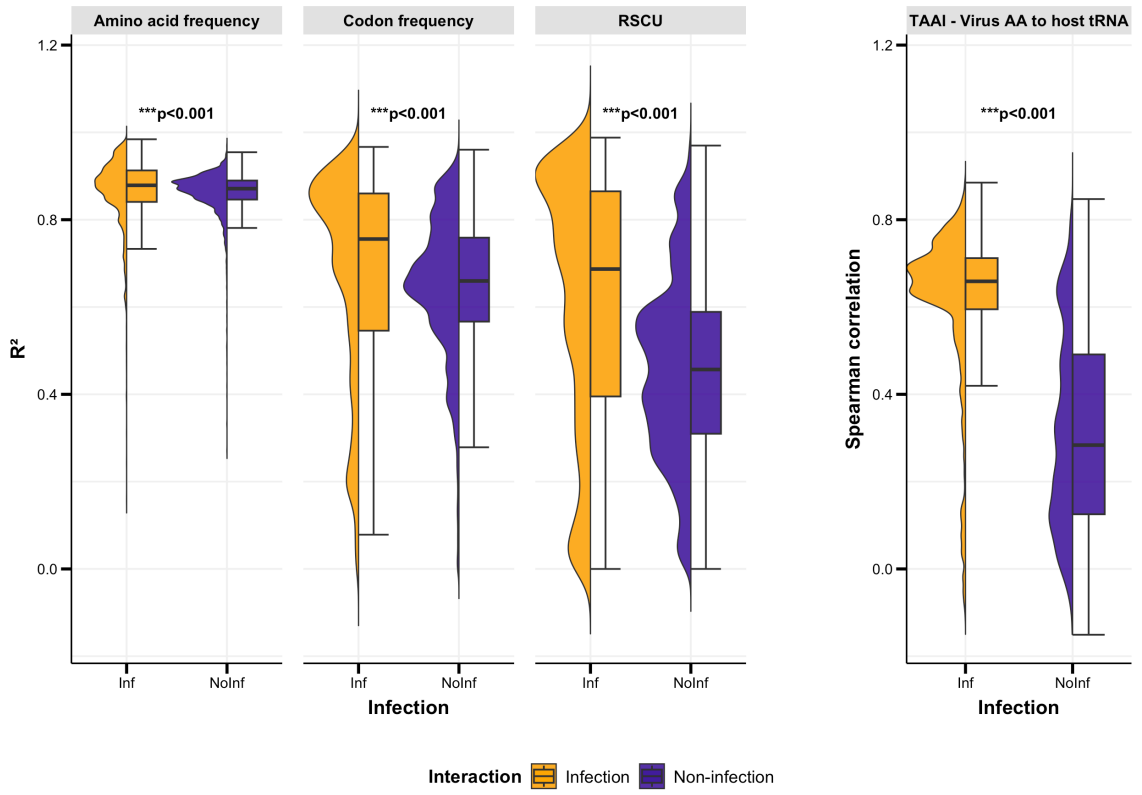


Figure 2.5. Gene-informed signatures result in strongest signal to differentiate between compatible and incompatible virus-host pairs.

Notably, the TAAI metric (right plot), which captures the correlation between amino acids needs of the virus to the tRNA pool of the host shows the strongest discriminatory power to differentiate between infection and non-infection events.

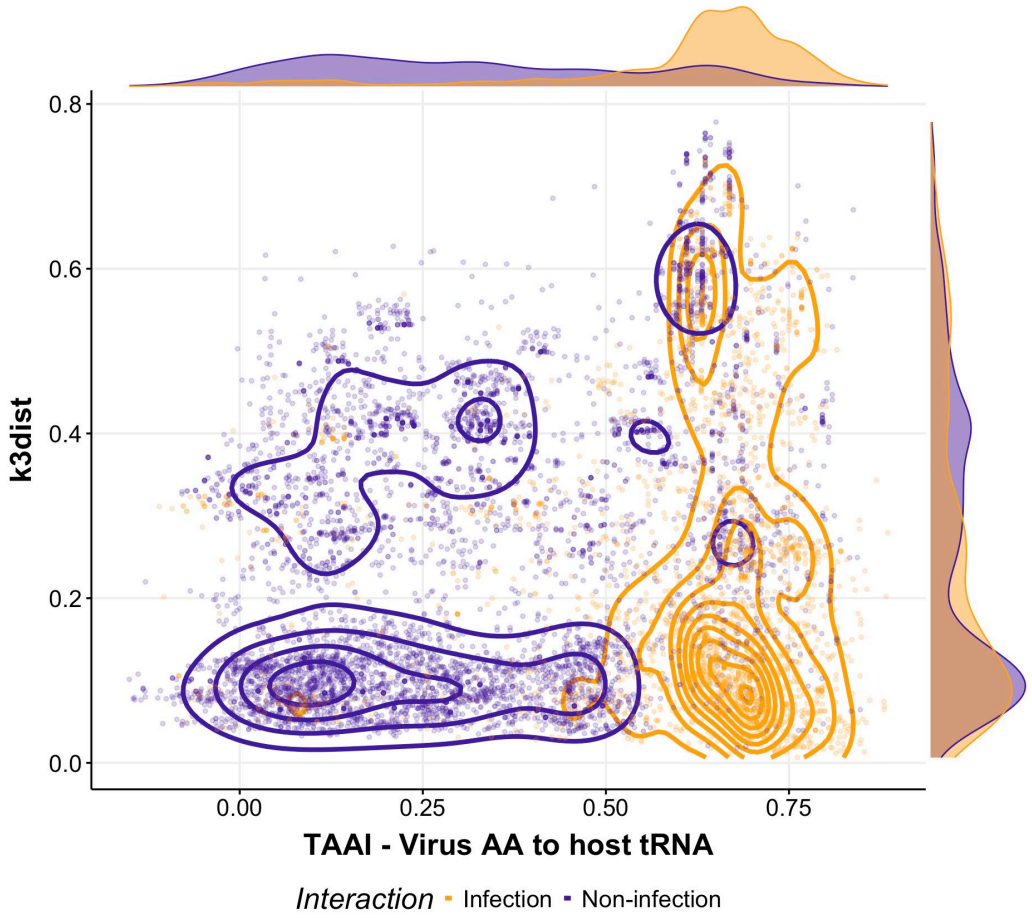
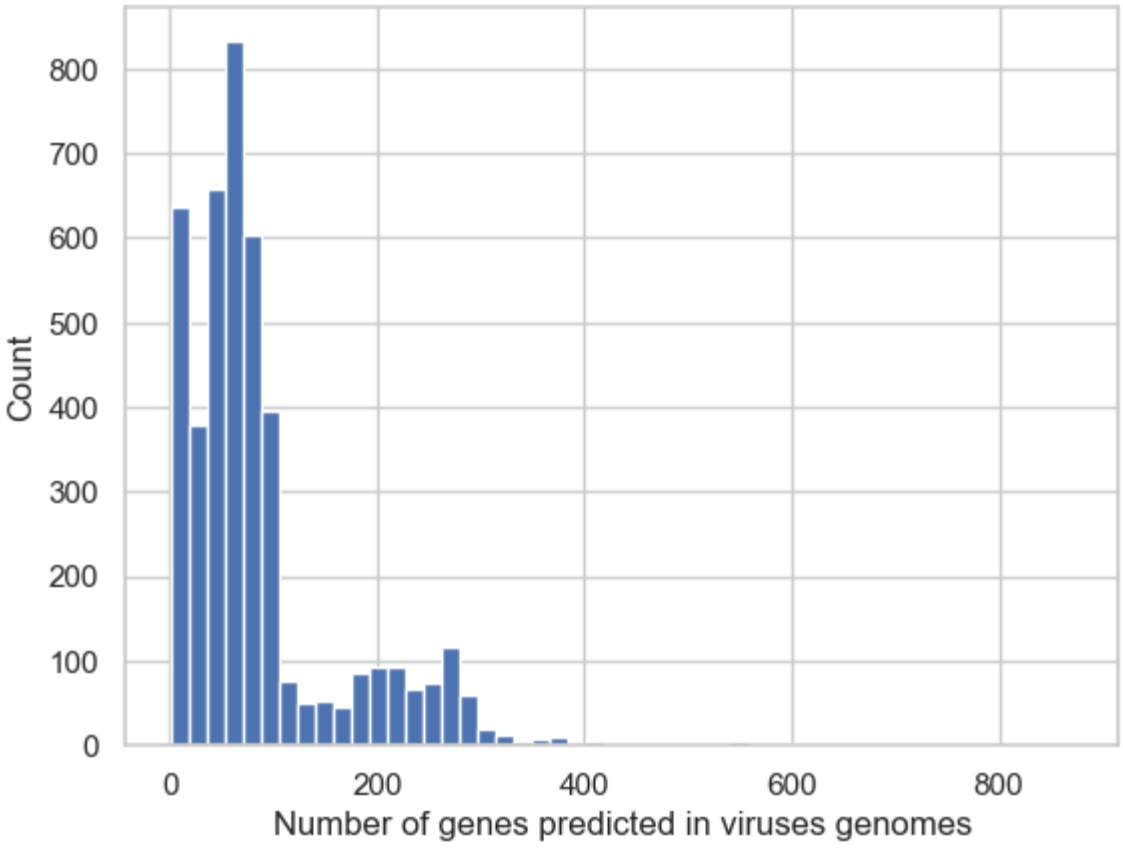


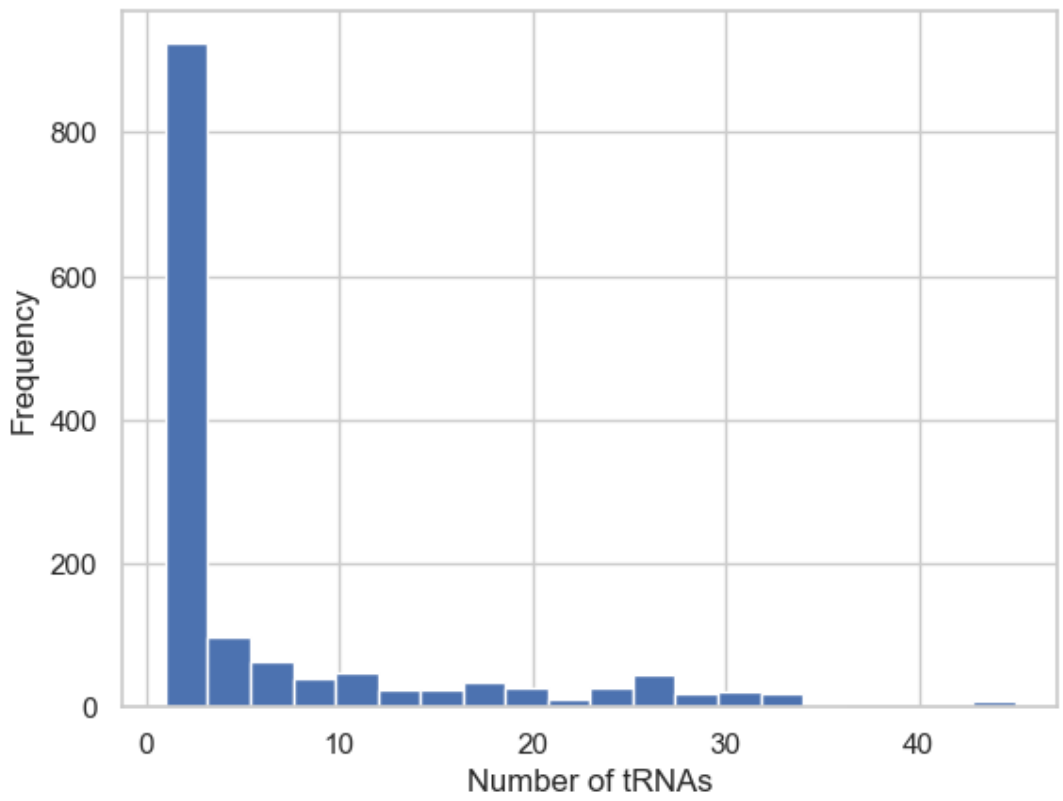
Figure 2.6. Viruses exhibit multiple dimensions of adaptation to their host.

Viruses show adaptation in multiple dimensions. When only considering k-mer distances of virus-host pairs, infection and non-infection pairs overlap (right side distribution). Similarly, some overlap exists between those infections and non-infections for looking at TAAI. However, when those metrics are combined, a clearer signal of adaptation emerges.

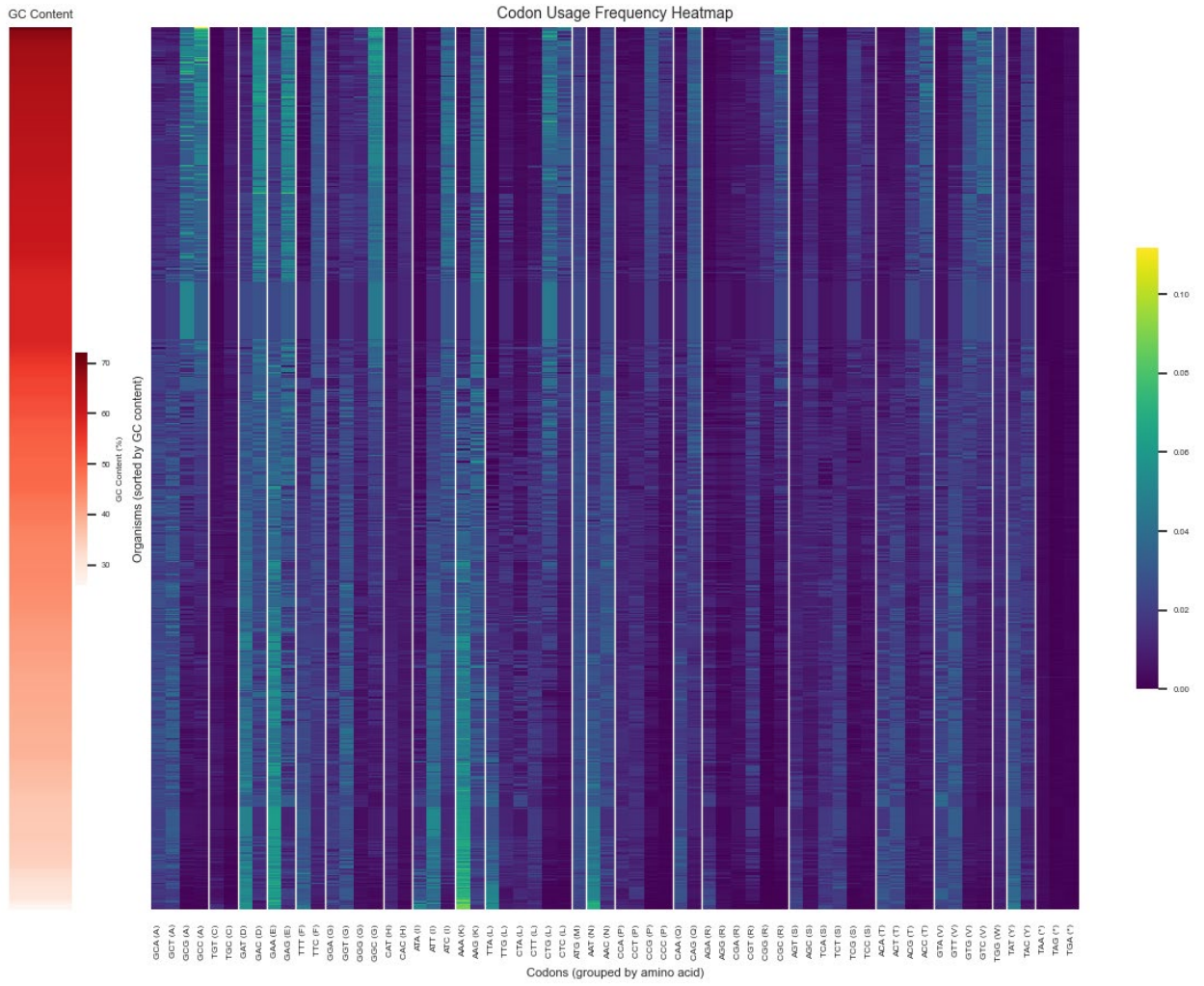
2.7 Supplementary figures



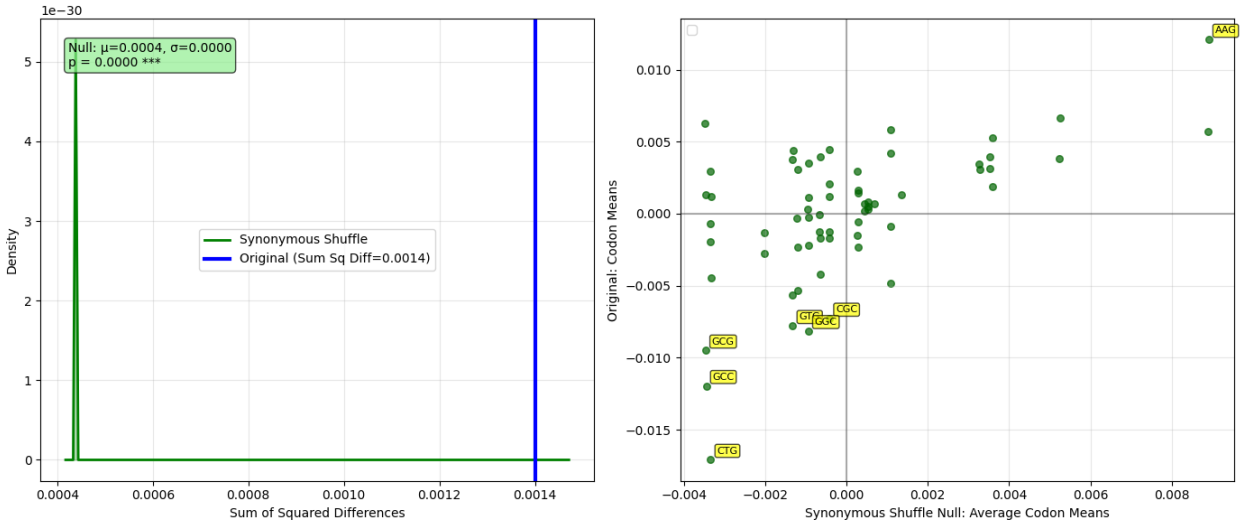
Supplemental Figure 2.7. Distribution of predicted genes in virus genomes.



Supplemental Figure 2.8. Number of tRNAs predicted in virus genomes.

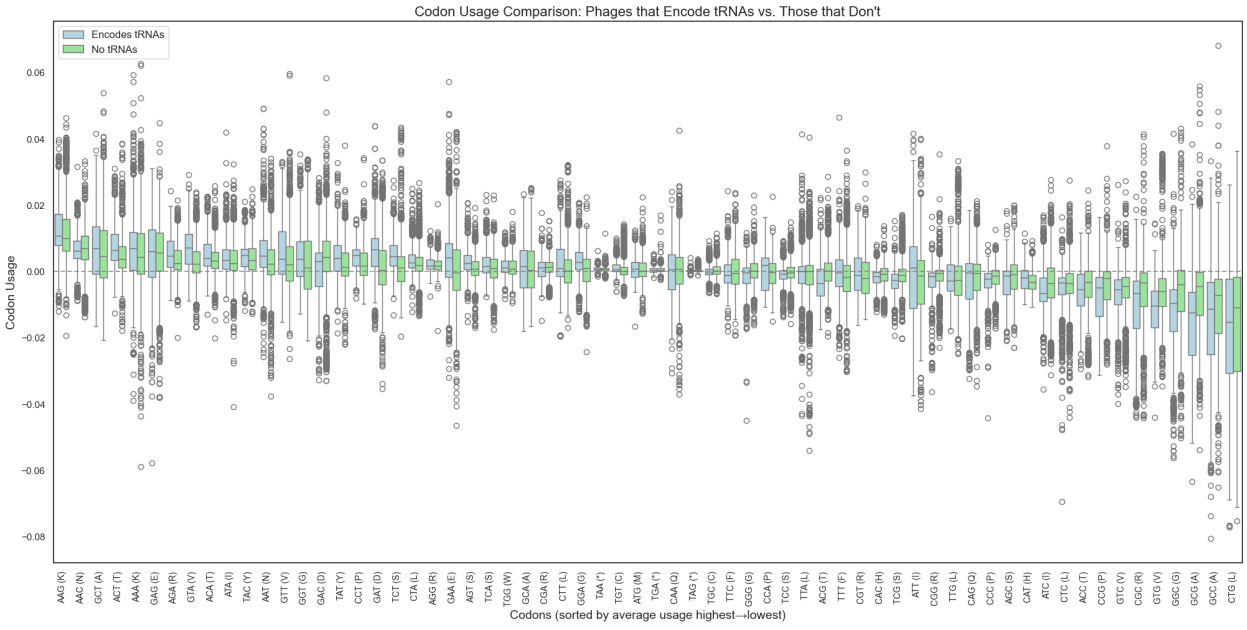


Supplemental Figure 2.9. Codon usage frequency of all viruses



Supplemental Figure 2.10. Null modeling of codon usage bias difference between viruses and hosts.

Left plot shows distribution of null models (green) compared to the empirical value (blue vertical line). Right plot shows which codons were the most different from the null modeling.



Supplemental Figure 2.11. Comparison of codon usage frequency relative to host

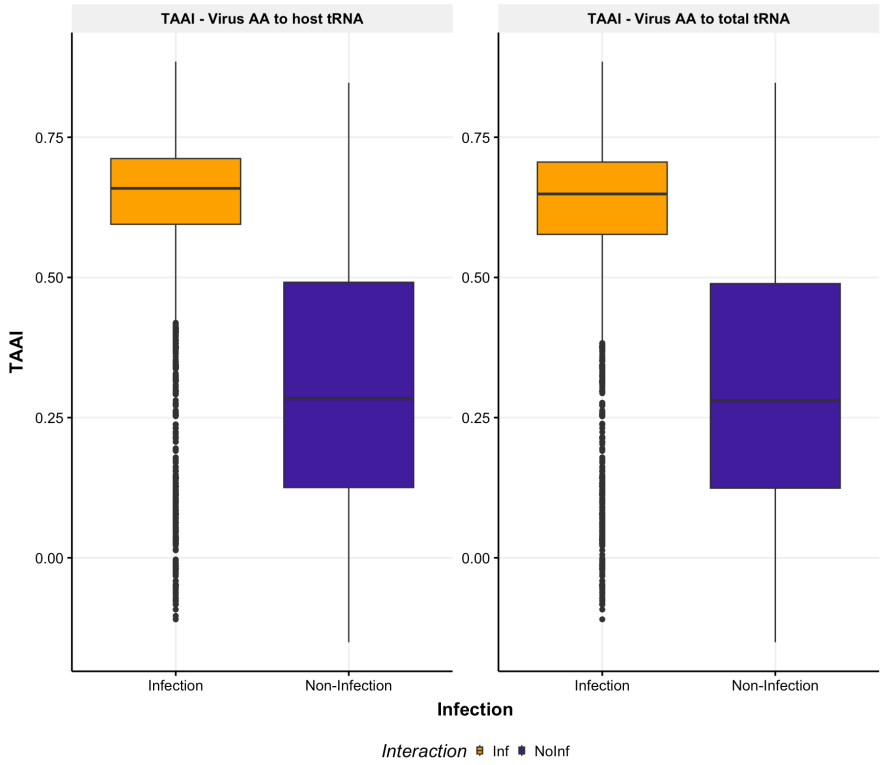


Figure 2.12. Comparing TAAI metrics between total pool of tRNAs and only host tRNA pool

Since viruses encode few tRNAs, including those tRNAs in the TAAI calculation did not make a meaningful difference in the values computed (comparing infection and non-infection).

2.8 VHRnet: A Curated Resource of Laboratory-Verified Virus-Host Interactions[2]

Comprehensive datasets linking genomic sequences to virus-host infection outcomes (herein 'interactions') are essential to advance viral and microbial ecology (55–59) and (micro)biotechnological applications (57, 60–62), however, they are lacking. Existing resources, such as NCBI RefSeq, provide viral genomic data, but lack host range information beyond the report of a single host. While numerous databases catalog virus-host associations (59, 63–65), most contain successful infections only and omit reports of non-infection, creating systematic biases in our understanding of host-range determinants (54). If negative interactions are reported, they are published as virus host range studies but are rarely compiled into open-source accessible datasets. Specialized virus-host datasets that do report infection outcomes of many-hosts-to-one-virus often omit genomic sequences or focus on narrow taxonomic groups (66). This separation of positive and negative interaction data prevents linkage of viral genomic features to host range patterns at the scales at which genomic data are available. The lack of integrated genomic and host range data hinders the discovery of infection determinants and the prediction of virus-host interactions. Here, we present VHRnet (Virus-Host Range network), which addresses these limitations by combining genomic sequences with curated lab-verified documentation of both infection and non-infection data.

VHRnet was first compiled in August 2018 from two primary sources (54). First, we programmatically extracted metadata from viral genomes in NCBI RefSeq, utilizing the annotations from the "/lab_host" structured comment in GenBank files to identify the reported one-to-one host interactions (i.e., one host for the one virus). Second, we conducted literature searches of host-range studies for viruses with available genomic sequences that reported both positive and negative infection outcomes. This expanded our

[2]This section is based on a manuscript co-authored with Cecelia Batterbee and Melissa Duhaime, currently under review at ASM Microbiology Resource Announcement.

data structures beyond NCBI's one-to-one reporting to include many-to-one. When genome sequences of a host strain were not available, a genome of that species was randomly selected from among those sequenced to be a species representative genome. To keep the dataset current with the rapidly expanding literature on viruses of microbes, we updated VHRnet to include new genomes and infection reports available as of 31 September 2024, adding 1,379 new interactions.

The current VHRnet dataset contains 10,228 laboratory-verified interactions among 2,292 unique viruses and 375 bacterial and archaeal host species. With 4,149 infections (40.6%) and 6,079 non-infections (59.4%), the dataset is more balanced, which is critical for unbiased algorithm development that uses these classes.

VHRnet is structured as a tabular dataset, where each row represents a virus-host pair. Each interaction record includes NCBI genome accession numbers of the virus, host taxonomic information and the infection outcomes (infection/non-infection).
By compiling lab-verified infection and non-infection outcomes from virus-host interactions reported across numerous studies, VHRnet provides a robust foundation for advancing our quantitative understanding of virus-host coevolution and for developing computational models and tools that support fundamental viral ecology research and biotechnological applications.

VHRnet is publicly available at the following link:
https://github.com/DuhaimeLab/VHRNet.

2.9 References

1. Walsh,D. and Mohr,I. (2011) Viral subversion of the host protein synthesis machinery. *Nat Rev Microbiol*, **9**, 860–875.
2. Enav,H., Mandel-Gutfreund,Y. and Béjà,O. (2014) Comparative metagenomic analyses reveal viral-induced shifts of host metabolism towards nucleotide biosynthesis. *Microbiome*, **2**, 9.
3. Thompson,L.R., Zeng,Q., Kelly,L., Huang,K.H., Singer,A.U., Stubbe,J. and Chisholm,S.W. (2011) Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *PNAS*, **108**, E757–E764.
4. Chevallereau,A., Pons,B.J., van Houte,S. and Westra,E.R. (2021) Interactions between bacterial and phage communities in natural environments. *Nat Rev Microbiol*, 10.1038/s41579-021-00602-y.
5. Blasdel,B.G., Chevallereau,A., Monot,M., Lavigne,R. and Debarbieux,L. (2017) Comparative transcriptomics analyses reveal the conservation of an ancestral infectious strategy in two bacteriophage genera. *ISME J*, **11**, 1988–1996.
6. Howard-Varona,C., Lindback,M.M., Bastien,G.E., Solonenko,N., Zayed,A.A., Jang,H., Andreopoulos,B., Brewer,H.M., Glavina del Rio,T., Adkins,J.N., *et al.* (2020) Phage-specific metabolic reprogramming of virocells. *ISME J*, **14**, 881–895.
7. Berglund,O., Karlström,O. and Reichard,P. (1969) A new ribonucleotide reductase system after infection with phage t4*. *Proceedings of the National Academy of Sciences*, **62**, 829–835.
8. Wikner,J., Vallino,J.J., Steward,G.F., Smith,D.C. and Azam,F. (1993) Nucleic acids from the host bacterium as a major source of nucleotides for three marine bacteriophages. *FEMS Microbiol Ecol*, **12**, 237–248.
9. Lee,J.Y., Li,Z. and Miller,E.S. (2017) Vibrio Phage KVP40 Encodes a Functional NAD+ Salvage Pathway. *J Bacteriol*, **199**, e00855-16.
10. Lambert,L.J., Wei,Y., Schirf,V., Demeler,B. and Werner,M.H. (2004) T4 AsiA blocks DNA recognition by remodeling σ70 region 4. *The EMBO Journal*, **23**, 2952–2962.
11. Hinton,D.M. (2010) Transcriptional control in the prereplicative phase of T4 development. *Virology Journal*, **7**, 289.

12. Severinova,E., Severinov,K. and Darst,S.A. (1998) Inhibition of Escherichia coli RNA polymerase by bacteriophage T4 AsiA. *J Mol Biol*, **279**, 9–18.
13. Holmudden,M., Gustafsson,J., Bertrand,Y.J.K., Schliep,A. and Norberg,P. (2024) Evolution shapes and conserves genomic signatures in viruses. *Commun Biol*, **7**, 1–12.
14. Belalov,I.S. and Lukashev,A.N. (2013) Causes and Implications of Codon Usage Bias in RNA Viruses. *PLOS ONE*, **8**, e56642.
15. Carbone,A. (2008) Codon Bias is a Major Factor Explaining Phage Evolution in Translationally Biased Hosts. *J Mol Evol*, **66**, 210–223.
16. Lázaro,E., Arribas,M., Cabanillas,L., Román,I. and Acosta,E. (2018) Evolutionary adaptation of an RNA bacteriophage to the simultaneous increase in the within-host and extracellular temperatures. *Sci Rep*, **8**, 8080.
17. Simmonds,P., Aiewsakun,P. and Katzourakis,A. (2019) Prisoners of war — host adaptation and its constraints on virus evolution. *Nat Rev Microbiol*, **17**, 321–328.
18. Bahir,I., Fromer,M., Prat,Y. and Linial,M. (2009) Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Molecular Systems Biology*, **5**, 311.
19. Bastien,E.G., Cable,R.N., Zaman,L., Batterbee,C., Wing,A.J. and Duhaime,M.B. (2023) Virus-Host Interactions Predictor (VHIP): machine learning approach to resolve microbial virus-host interaction networks. 10.1101/2023.11.03.565433.
20. Simón,D., Cristina,J. and Musto,H. (2021) Nucleotide Composition and Codon Usage Across Viruses and Their Respective Hosts. *Front. Microbiol.*, **12**.
21. Sharp,P.M. and Li,W.-H. (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, **15**, 1281–1295.
22. Jenkins,G.M. and Holmes,E.C. (2003) The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res*, **92**, 1–7.
23. Chen,F., Wu,P., Deng,S., Zhang,H., Hou,Y., Hu,Z., Zhang,J., Chen,X. and Yang,J.-R. (2020) Dissimilation of synonymous codon usage bias in virus–host coevolution due to translational selection. *Nat Ecol Evol*, **4**, 589–600.

24. Ahlgren,N.A., Ren,J., Lu,Y.Y., Fuhrman,J.A. and Sun,F. (2017) Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res*, **45**, 39–53.
25. Iwasaki,Y., Abe,T. and Ikemura,T. (2023) Oligonucleotide usage in coronavirus genomes mimics that in exon regions in host genomes. *Virology Journal*, **20**, 39.
26. Greenbaum,B.D., Levine,A.J., Bhanot,G. and Rabadan,R. (2008) Patterns of Evolution and Host Gene Mimicry in Influenza and Other RNA Viruses. *PLOS Pathogens*, **4**, e1000079.
27. Amgarten,D., Iha,B.K.V., Piroupo,C.M., da Silva,A.M. and Setubal,J.C. (2022) vHULK, a New Tool for Bacteriophage Host Prediction Based on Annotated Genomic Features and Neural Networks. *PHAGE*, 10.1089/phage.2021.0016.
28. Coutinho,F.H., Zaragoza-Solas,A., López-Pérez,M., Barylski,J., Zielezinski,A., Dutilh,B.E., Edwards,R. and Rodriguez-Valera,F. (2021) RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns*, **2**, 100274.
29. Roux,S., Camargo,A.P., Coutinho,F.H., Dabdoub,S.M., Dutilh,B.E., Nayfach,S. and Tritt,A. (2022) iPHoP: an integrated machine-learning framework to maximize host prediction for metagenome-assembled virus genomes. 10.1101/2022.07.28.501908.
30. Shang,J. and Sun,Y. (2022) CHERRY: a Computational metHod for accuratE pRediction of virus-pRokarYotic interactions using a graph encoder-decoder model. *Briefings in Bioinformatics*, 10.1093/bib/bbac182.
31. Ren,J., Ahlgren,N.A., Lu,Y.Y., Fuhrman,J.A. and Sun,F. (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, **5**, 69.
32. Edwards,R.A., McNair,K., Faust,K., Raes,J. and Dutilh,B.E. (2016) Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol Rev*, **40**, 258–272.
33. Bailly-Bechet,M., Vergassola,M. and Rocha,E. (2007) Causes for the intriguing presence of tRNAs in phages. *Genome Res.*, **17**, 1486–1495.
34. Abergel,C., Legendre,M. and Claverie,J.-M. (2015) The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol Rev*, **39**, 779–796.

35. Moniruzzaman,M., Martinez-Gutierrez,C.A., Weinheimer,A.R. and Aylward,F.O. (2020) Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat Commun*, **11**, 1710.
36. Du,M.-Z., Wei,W., Qin,L., Liu,S., Zhang,A.-Y., Zhang,Y., Zhou,H. and Guo,F.-B. (2017) Co-adaption of tRNA gene copy number and amino acid usage influences translation rates in three life domains. *DNA Res*, **24**, 623–633.
37. Ikemura,T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *Journal of Molecular Biology*, **146**, 1–21.
38. Ikemura,T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol*, **2**, 13–34.
39. Schwengers,O., Jelonek,L., Dieckmann,M.A., Beyvers,S., Blom,J. and Goesmann,A. (2021) Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial Genomics*, **7**, 000685.
40. Weiss,S.B., Hsu,W.T., Foft,J.W. and Scherberg,N.H. (1968) Transfer RNA coded by the T4 bacteriophage genome. *Proceedings of the National Academy of Sciences*, **61**, 114–121.
41. Lucks,J.B., Nelson,D.R., Kudla,G.R. and Plotkin,J.B. (2008) Genome Landscapes and Bacteriophage Codon Usage. *PLOS Computational Biology*, **4**, e1000001.
42. Varani,G. and McClain,W.H. (2000) The G·U wobble base pair. *EMBO reports*, **1**, 18–23.
43. Lomeli-Ortega,C.O. and Balcázar,J.L. (2024) Why tRNA acquisition could be relevant to bacteriophages? *Microbial Biotechnology*, **17**, e14464.
44. Schlub,T.E. and Holmes,E.C. (2020) Properties and abundance of overlapping genes in viruses. *Virus Evol*, **6**, veaa009.
45. Brandes,N. and Linial,M. (2016) Gene overlapping and size constraints in the viral world. *Biol Direct*, **11**, 26.
46. Warnecke,T., Weber,C.C. and Hurst,L.D. (2009) Why there is more to protein evolution than protein function: splicing, nucleosomes and dual-coding sequence. *Biochem Soc Trans*, **37**, 756–761.

47. van den Berg,D.F., van der Steen,B.A., Costa,A.R. and Brouns,S.J.J. (2023) Phage tRNAs evade tRNA-targeting host defenses through anticodon loop mutations. *Elife*, **12**, e85183.
48. Burman,N., Belukhina,S., Depardieu,F., Wilkinson,R.A., Skutel,M., Santiago-Frangos,A., Graham,A.B., Livenskyi,A., Chechenina,A., Morozova,N., *et al.* (2024) A virally encoded tRNA neutralizes the PARIS antiviral defence system. *Nature*, **634**, 424–431.
49. Alonso,A.M. and Diambra,L. (2020) SARS-CoV-2 Codon Usage Bias Downregulates Host Expressed Genes With Similar Codon Usage. *Front. Cell Dev. Biol.*, **8**.
50. Frumkin,I., Lajoie,M.J., Gregg,C.J., Hornung,G., Church,G.M. and Pilpel,Y. (2018) Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proceedings of the National Academy of Sciences*, **115**, E4940–E4949.
51. Prabhakaran,R., Chithambaram,S. and Xia,X. (2014) Aeromonas phages encode tRNAs for their overused codons. *Int J Comput Biol Drug Des*, **7**, 168–182.
52. Pride,D.T., Wassenaar,T.M., Ghose,C. and Blaser,M.J. (2006) Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics*, **7**, 8.
53. Chan,P.P., Lin,B.Y., Mak,A.J. and Lowe,T.M. (2021) tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research*, **49**, 9077–9096.
54. Bastien,G.E., Cable,R.N., Batterbee,C., Wing,A.J., Zaman,L. and Duhaime,M.B. (2024) Virus-host interactions predictor (VHIP): Machine learning approach to resolve microbial virus-host interaction networks. *PLOS Computational Biology*, **20**, e1011649.
55. Breitbart,M. (2011) Marine Viruses: Truth or Dare. *Annu. Rev. Mar. Sci.*, **4**, 425–448.
56. de Jonge,P.A., Nobrega,F.L., Brouns,S.J.J. and Dutilh,B.E. (2019) Molecular and Evolutionary Determinants of Bacteriophage Host Range. *Trends in Microbiology*, **27**, 51–63.
57. Bisen,M., Kharga,K., Mehta,S., Jabi,N. and Kumar,L. (2024) Bacteriophages in nature: recent advances in research tools and diverse environmental and biotechnological applications. *Environ Sci Pollut Res*, **31**, 22199–22242.

58. Sudweeks,J. and Hauert,C. (2025) The impact of simultaneous infections on phage-host ecology. *Theoretical Population Biology*, **161**, 42–49.
59. Albrycht,K., Rynkiewicz,A.A., Harasymczuk,M., Barylski,J. and Zielezinski,A. (2022) Daily Reports on Phage-Host Interactions. *Front. Microbiol.*, **13**.
60. Stone,E., Campbell,K., Grant,I. and McAuliffe,O. (2019) Understanding and Exploiting Phage–Host Interactions. *Viruses*, **11**.
61. Federici,S., Nobs,S.P. and Elinav,E. (2021) Phages and their potential to modulate the microbiome and immunity. *Cell Mol Immunol*, **18**, 889–904.
62. Liu,R., Li,Z., Han,G., Cun,S., Yang,M. and Liu,X. (2021) Bacteriophage ecology in biological wastewater treatment systems. *Appl Microbiol Biotechnol*, **105**, 5299–5307.
63. Gao,N.L., Zhang,C., Zhang,Z., Hu,S., Lercher,M.J., Zhao,X.-M., Bork,P., Liu,Z. and Chen,W.-H. (2018) MVP: a microbe–phage interaction database. *Nucleic Acids Res*, **46**, D700–D707.
64. Lamy-Besnier,Q., Brancotte,B., Brancotte, Ménager,H., Ménager and Debarbieux,L. (2021) Viral Host Range database, an online tool for recording, analyzing and disseminating virus–host interactions. *Bioinformatics*, **37**, 2798–2801.
65. Mihara,T., Nishimura,Y., Shimizu,Y., Nishiyama,H., Yoshikawa,G., Uehara,H., Hingamp,P., Goto,S. and Ogata,H. (2016) Linking Virus Genomes with Host Taxonomy. *Viruses*, **8**, 66.
66. Maffei,E., Shaidullina,A., Burkolter,M., Heyer,Y., Estermann,F., Druelle,V., Sauer,P., Willi,L., Michaelis,S., Hilbi,H., *et al.* (2021) Systematic exploration of Escherichia coli phage–host interactions with the BASEL phage collection. *PLOS Biology*, **19**, e3001424

Chapter 3 – Virus-Host Interactions Predictor (VHIP): Machine Learning Approach to Resolve Microbial Virus-Host Interaction Networks[3]

3.1 Abstract

Viruses of microbes are ubiquitous biological entities that reprogram their hosts' metabolisms during infection in order to produce viral progeny, impacting the ecology and evolution of microbiomes with broad implications for human and environmental health. Advances in genome sequencing have led to the discovery of millions of novel viruses and an appreciation for the great diversity of viruses on Earth. Yet, with knowledge of only *"who is there?"* we fall short in our ability to infer the impacts of viruses on microbes at population, community, and ecosystem-scales. To do this, we need a more explicit understanding *"who do they infect?"* Here, we developed a novel machine learning model (ML), Virus-Host Interaction Predictor (VHIP), to predict virus-host interactions (infection/non-infection) from input virus and host genomes. This ML model was trained and tested on a high-value manually curated set of 8849 virus-host pairs and their corresponding sequence data. The resulting dataset, 'Virus Host Range network' (VHRnet), is core to VHIP functionality. Each data point that underlies the VHIP training and testing represents a lab-tested virus-host pair in VHRnet, from which meaningful signals of viral adaptation to host were computed from genomic sequences. VHIP departs from existing virus-host prediction models in its ability to predict multiple interactions rather than

[3] This chapter is a reprint of a publication with the following citation: Bastien,G.E., Cable,R.N., Batterbee,C., Wing,A.J., Zaman,L. and Duhaime,M.B. (2024) Virus-host interactions predictor (VHIP): Machine learning approach to resolve microbial virus-host interaction networks. *PLOS Computational Biology*, **20**, e1011649.

predicting a single most likely host or host clade. As a result, VHIP is able to infer the complexity of virus-host networks in natural systems. VHIP has an 87.8% accuracy rate at predicting interactions between virus-host pairs at the species level and can be applied to novel viral and host population genomes reconstructed from metagenomic datasets.

3.2 Author summary

The ecology and evolution of microbial communities are deeply influenced by viruses. Metagenomics analysis, the non-targeted sequencing of community genomes, has led to the discovery of millions of novel viruses. Yet, through the sequencing process, only DNA sequences are recovered, begging the question: which microbial hosts do those novel viruses infect? To address this question, we developed a computational tool to allow researchers to predict virus-host interactions from such sequence data. The power of this tool is its use of a high-value, manually curated set of 8849 lab-verified virus-host pairs and their corresponding sequence data. For each pair, we computed signals of coevolution to use as the predictive features in a machine learning model designed to predict interactions between viruses and hosts. The resulting model, Virus-Host Interaction Predictor (VHIP), has an accuracy of 87.8% and can be applied to novel viral and host genomes reconstructed from metagenomic datasets. Because the model considers all possible virus-host pairs, it can resolve complete virus-host interaction networks and supports a new avenue to apply network thinking to viral ecology.

3.3 Introduction

The development of metagenomic sequence analyses has led to unprecedented discoveries in microbial science (1, 2), owing to the ability to study viruses and cellular microbes in their quintessential contexts. In particular, metagenomics has shed light on the genomic diversity and ubiquity of viruses (3–7). Those viral populations are reconstructed from metagenomic data and identified as novel based on their sequence similarity to known viruses, expanding the total number of distinct uncultured viruses to millions the last decade alone (3, 8–11). With these discoveries, there is mounting interest in characterizing

how viruses impact microbial communities (12). During infection, viruses influence ecological processes at multiple scales by shaping host population dynamics (13, 14), modulating horizontal gene transfer (15, 16), and reprogramming host metabolic pathways that can modulate the flux of environmental nutrients (17–20). Given the central role viruses play in the ecology and evolution of their microbial hosts, there is great interest in including them in biogeochemical modeling by leveraging global-scale metagenomic data. While the population genomes of viruses and their microbial hosts can be reconstructed from metagenomic data with increasing accuracy, the inclusion of viruses in community metabolic models that can one day be scaled to ecosystems is impeded by the absence of arguably the most important question about any novel virus: who does it infect (11, 21, 22)?

To bridge this knowledge gap, various approaches have been explored. Phylogeography-based approaches have been successful in predicting virus-host associations in eukaryotic systems (23–25). However, this approach is not applicable for viruses infecting microbes, as microbes show weak patterns of biogeography (26). Another method leverages patterns of coevolution that can be extracted from the genomic sequences of the viruses and their host to predict the most likely taxa that can be infected by a given virus (27–31). This approach of leveraging genomic signals to predict virus-host association is possible because viruses rely on their host machinery to complete their life cycle and evolve to better utilize those resources by matching the codon biases of their host (20, 32, 33). This results in a meaningful and capturable signal that is embedded in the sequences of the viruses and their known host (34, 35). Those host prediction tools (HPTs) are, however, limited in scope. They typically only allow viral sequences as input, which restricts testing to host sequences that already exist in pre-defined reference databases, or they require sufficient expertise and resources from users to re-train the models to include new hosts sequences. These limitations restrict the applicability of such tools, as they are difficult to use to study viral host range in natural community contexts with newly discovered host populations. In addition, they typically focus on predicting the most likely taxa a virus infects, which does not reflect the breadth of natural virus-host range profiles (36, 37) that can span different taxonomic levels (37–40). Further, as they do not predict non-infection

interactions, that is, a virus's inability to infect a host, existing HPTs cannot be used to resolve virus-host interaction networks.

To address these limitations, we collected lab-verified viral-host interactions (infection and non-infection data) from public databases and literature and compiled them into a single dataset named 'Virus Host Range network' (VHRnet). This data is essential for exploring the strengths of the virus-host coevolution signals, assessing existing HPTs, and for developing a novel model that can predict both infection and non-infection relationships for all pairwise sets of viruses and putative hosts the prediction model may someday encounter. In this study the VHRnet data were used to (1) quantify and evaluate genome-derived signals of coevolution captured in lab-validated virus-host pairs, (2) develop a machine learning model that leverages these virus-microbe coevolutionary signals, and (3) assess the accuracy of the model in predicting interactions (i.e., infection or non-infection) at the species level. The resulting model developed and described here is named VHIP for Virus-Host Interaction Predictor.

3.4 Results

3.4.1 VHRnet, a manually curated host range dataset unparalleled in size and scope

To train and test machine learning model approaches to predict virus-host relationships, we aimed to compile the most comprehensive host range dataset available, wherein all viral and host genome sequences are also publicly available (Fig 1A). For this, we relied on the fact that every virus in the NCBI RefSeq database has a host associated with it via the `/lab_host`tag in the GenBank file controlled syntax. In addition, for each virus in the RefSeq database, we searched published studies beyond the genome reports for documented host range trials. A total of 8849 lab-tested interactions were collected and compiled (Fig 2A). The majority of interactions in this dataset were non-infection (n = 6079), owing to a small number of large-scale host range studies that diligently tested and reported all virus-host pairs in their study (Fig 2B), rather than only reporting cases of infection, as is done in the vast majority of virus-host studies. This resulting dataset was

named VHRnet, for Virus-Host Range network. It comprises 375 unique host species and 2292 unique viruses. The majority (94.7%) of the viruses belong to the Caudovirales order (Figs 2C and S8), which may be driven by culture techniques biases (36, 41). There are biases in the host taxa represented as well, with the majority of the tested hosts human pathogens (Figs 2D and S9), partially driven in the recent surge of phage therapy research (42).

The majority of the viruses (n = 1962, 84.2%) were reportedly tested against a single host; these pairings came from the `/lab_host`tag of their NCBI GenBank files. Of the remaining viruses (n = 369, 18.8%), most (81.8%) were tested against different host species with no cross-genus tests. The percentages of viruses tested across clades gets increasingly smaller at higher taxonomic levels: 78.5% were tested at most across families, 77.5% were tested across order, 9.2% were tested across class, and 0% were tested across phyla or domain (S10 Fig). However it is important to note that the numbers of cross-clade host testing were heavily influenced by two large host-range scale studies (herein the 'Staphylococcus study' (40) and 'Nahant study' (36)), in which every potential virus and host pair was experimentally tested, resulting in a "complete" virus-host network (Fig 2A and 2B) (36, 40). 70.8% of the VHRnet pairs came from those two studies alone. When excluding those two large host range studies, only 82 viruses in VHRnet (4.01%) were tested against multiple species. Out of those viruses, 65.6% were tested against hosts belonging to different genera, 52.4% were tested against different families, 46.3% against different order, 41.2% against different class, and 0 viruses were tested against different hosts belonging to different phyla.

3.4.2 VHRnet provides opportunity for holistic comparison of Host Prediction Tools (HPTs)

The diligent lab-testing of all virus-host pairs from Staphylococcus and Nahant studies (Fig 2A and 2B) presented an opportunity to directly compare existing HPTs (Table 1) on the same lab-validated infection/non-infection data. While HPT benchmarking is common, such a direct comparison of HPTs has been challenging given that each tool uses similar but

not identical datasets for model testing and training. For the most objective assessment, testing datasets must exclude training data, which is not possible when part of the data used to compare outputs of one model was included in the training sets of other benchmarked models. In assessing the accuracy of HPTs against the Nahant Collection and Staphylococcus datasets in this study, there were three possible outcomes: (1) correct host was predicted, (2) incorrect host was predicted, (3) a host was predicted that was not experimentally tested ("untested", Fig 3). Note that if there are multiple known hosts for a virus, these models only need to predict one correct host to obtain 100% accuracy in this evaluation.

We found that the accuracy of each HPT decreased as we considered more resolved host taxonomic levels (i.e., from domain to species level). Furthermore, the number of predicted virus-host pairs that were not experimentally tested also increased with increasing host taxonomic resolution (Fig 3). Existing HPTs performed better in predicting hosts for viruses in the Staphylococcus study than for viruses in the Nahant study, with overall fewer wrong predictions. Regardless, existing HPTs do not have the resolution to reliably predict species level virus-host interactions. Moreover, due to the limitation that stems from the reference data used by HPTs (that is, only one known host per virus), these tools typically deliver only their highest scoring host prediction as output, which can lead to uncertainty in output data interpretation and downstream use. For instance, if there are five total predictions, all of low quality scores, the best of the poor predictions will be reported as the "most likely host", rather than returning a prediction of "no infection". Similarly, if five hosts are predicted with similarly high scores, only a single prediction is chosen, rather than five "infection" predictions. This forced one-to-one prediction output model design limits the use of current HPTs and does not reflect the dynamics of virus-host relationships. Motivated to move beyond the one-to-one output design, we next leveraged the VHRnet data to identify and quantify genome-derived signals of coevolution between known infection pairs (relative to non-infection pairs) at the virus and host species level. We hypothesized that these data could be used to develop a many-to-many HPT design.

3.4.3 Genomic signals of virus adaptation to their host(s) are discernible at the species level

Signals of coevolution were computed for each virus-host pair in VHRnet (Fig 1A and 1B), which can be broadly divided into two categories: sequence homology (Fig 1C) and sequence composition (Fig 1D) (41). For sequence homology, virus and host genomes were scanned for stretches of DNA sequence homology, which may indicate past virus-host interactions, such as CRISPR activity or horizontal gene transfer (HGT). We first identified bacteria and archaea in the VHRnet dataset with predicted CRISPR-Cas systems, then retrieved their associated CRISPR spacer motifs, which are short DNA sequences acquired from previous exposure to a foreign genetic element as part of the CRISPR-Cas adaptive immune response. Viral genomes were then scanned for the presence of these identical motifs. Stretches of non-CRISPR sequence homology, which we attributed to HGT events, between viruses and their putative host sequences were similarly identified between VHRnet viruses and hosts, with the difference being their search was not restricted to predicted CRISPR spacer motifs (minimum identity percentage 80 with minimum hit length 500 bp). Across the VHRnet data, instances of shared sequence homology between viral and host sequences were rare: sequence matches to CRISPR spacers were identified in 184 viruses out of 2292 (8.03%) and sequence matches attributed to HGT were identified in 340 viruses (14.83%). That the frequency of CRISPR matches was low was not surprising given that an estimated 50% of bacteria and 10% of archaea lack CRISPR-Cas viral defense systems (43).

In addition to sequence homology, we quantitatively evaluated signals of virus-host coevolution based on sequence composition. Because viruses rely on their host machinery to complete their life cycle (44), their genomes have a tendency towards matching the nucleotide usage biases of their hosts over coevolutionary time (the process of 'genome amelioration') (45–47). We computed k-mer profiles at k = 3, 6, and 9 for all viral and host sequences in VHRnet. We used both the Euclidean and the d2* distance metrics (35) to compute k-mer profile similarities between viruses and hosts and then to evaluate which measure encoded the strongest signals for virus-host predictions given our study design.

As previously reported (35), the d2* metric outperformed the Euclidean distance metric in its ability to resolve sequence composition-based signals of virus-host coevolution (S11 Fig). In other applications, d2* has been shown to predict viral hosts at the genus level with twice the accuracy of the Euclidean metric (35). The d2* metric differs from other distance metrics as it takes into consideration the background oligonucleotide patterns of the sequences being compared (48). While the Euclidean metric remains the most popular for binning metagenomic contigs in the reconstruction of microbial populations from metagenomes, our results suggest that the d2* may be a better metric for binning as well, especially with continued optimization to reduce its computational burden. The d2* algorithm is rewritten here as a Python package to ease accessibility.

As longer k-mers ("words") are considered, the number of possible words increases exponentially, e.g., 4 possible letters raised to the power of k means that for a 7 character sequence stretch (47), 16,384 words are possible. The length of the sequence must be sufficiently long, such that the frequency of each word is likely to be detected within the virus and host genomes being considered. If the k-mer is too long relative to the genome length, zeroes accrue in the k-mer frequencies table, which leads to spurious distance values, a behavior that has been previously posited by Edwards et al, 2016 (34). Here we assessed the impact of this behavior on prediction performance, given that we could now explicitly compare k-mer profile distances of known infection and known non-infection cases. Although the signal for certain virus-host pairs may get stronger at higher k-lengths, this is not a universal trend (35) (Fig 4A). At k = 9, some known non-infection pairs have smaller distances to known infection pairs (Fig 4A, orange arrow), regardless of the distance metric used. This does not happen when using k-lengths of 3 or 6. Thus we do not recommend using k-mers larger than 6 for purposes of virus-host predictions.

While it is commonly recognized that the %G+C contents of viruses are typically very similar to those of their hosts (49), the %G+C differences between viruses and hosts (%G+C of virus—%G+C of host) have also been previously recognized as a feature that could provide valuable information for virus-host predictions (34). As %G+C difference is not currently considered by existing HPTs, we evaluated the potential of this signal to capture

coevolutionary relationships. A decades-old study of 59 virus-host associations reported that viruses were on average 4% richer in AT (49). This trend was attributed to the fact that G and C are energetically more expensive nucleotides to synthesize than A and T, that ATP is an abundant cellular molecule and more readily available for genome incorporation, and that there are more diverse pathways (and fewer metabolic bottlenecks) to synthesize A and T, as compared to, e.g., C (49). In our set of 8849 virus-host interactions in VHRnet, we found a remarkably consistent result: viruses are on average 3.5% AT-richer relative to their host (Fig 4B), horizontal boxplot along top). We found that while the %G+C difference of a majority (68%) of the virus-host pairs fall within a narrow range of -4% to 4% (Fig 4C), the overall range of %G+C differences was quite broad, ranging from -40% to 32% (Fig 4C). These trends support that both the magnitude and direction of %G+C difference may be distinguishing features of virus-host infection pairs.

3.4.4 VHIP, a machine learning-based model, leverages the VHRnet dataset to predict virus-host interactions

Machine-learning model approaches are well-suited for capturing relationships between data features and have been applied to leverage genome-derived signals of virus-host coevolution using limited host range data in previous studies (50, 51). We sought to leverage VHRnet virus-host sequence pairs and the features we identified to assess, develop, and evaluate the performance of different machine learning models for virus-host predictions. Inclusion of infection and non-infection data points in VHRnet allows us to train a model predicting either infection or non-infection without any assumptions about virus host range. Before training machine learning-based models, we evaluated the Pearson pairwise correlations between the virus-host genomic signals (features used for machine-learning approach) (S12 Fig, and Table 2) to determine whether any of the features are strongly correlated. Strongly correlated features typically do not bring additional information for prediction and would increase the complexity of the ML model, which is generally avoided when designing sound ML models (52, 53). Of all our evaluated features, only the k-3 and k-6 features were strongly correlated (Pearson = 0.94, S4 Table). However, when comparing the k-3 distance against %G+C difference and k-6 distance against %G+C

difference, different patterns emerged (Fig 4D), suggesting that both the k-3 distance and k-6 distance encode different signals that can be leveraged by machine learning model approaches.

For a single feature to be strong enough to serve as a singular predictor of infection, there would need to be no overlap between infection and non-infection data points. This was not observed. For every feature, there was an overlap in the infection and non-infection density plots (Figs 4A and S12 diagonal plots). This strategy identifies regions of feature overlap and regions of distinction (i.e., valuable non-redundant information encoded in the features; green arrows Fig 4A) was also used to evaluate the power of combining features, such as the decision above to keep both k-3 and k-6 distances in the model (Fig 4D). Further, while virus-host genome amelioration was observed in the k-mer frequencies (resulting in low virus-host k-mer distances), we observed that viruses also have a strong tendency to remain AT-rich relative to their hosts (Fig 4B), consistent with prior reports (49).

We next evaluated the performance of different machine learning algorithms given our selected features (%G+C difference, k-3 distance, k-6 distance, homology hits). The Gradient Boosting Classifier and Random Forest performed best out of the classifiers we tested, with an average accuracy of 87.5% and 88.3% respectively (S13 Fig). However, the Gradient Boosting Classifier was used rather than the Random Forest, as it achieved comparable results with shallower trees, meaning fewer decision nodes were needed to reach comparable prediction performance. This is considered best practice for yielding the highest accuracy, while not overfitting the model (52). The best performing model was then assessed on the hold-out dataset. We bootstrapped this approach 100 times and tracked which set of hyperparameters yielded the highest accuracy on the hold-out set for each iteration (Fig 1E). Across the 100 best models from each bootstrapping iteration, the average area under the receiver operator characteristic (AUROC) value was 0.93 ± 0.004 (S14 Fig). AUROC is a metric used to evaluate the accuracy of classification models, whereby a value of 1 represents a model with perfect predictions, a 0 represents only incorrect predictions, and 0.5 represents a model that makes random predictions.

We trained the machine learning model a final time using the entire dataset and the best hyperparameters determined from the performance analysis. The model can predict virus-host species interactions with 87.8% accuracy, defined as the number of correct predictions divided by total number of predictions on the test set. To assess VHIP performance, we considered all possible prediction outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). False positives are commonly known as Type I error, and it represents cases when the model predicted infection but the virus-host pair is a case of non-infection. False negatives are commonly known as Type II errors, where the model predicted non-infection when it should have predicted infection. The AUROC score (the area under the receiver operating characteristics) for VHIP is 0.94 (Fig 5A); this is a measurement between the false positive rate (Type I error) against the true positive rate (model accurately predicting infection). The F1 score for VHIP is 0.93 (Fig 5B), which is the harmonic mean of the precision (number of true positive results divided by the number of all positive results, TP / (TP + FP)) and recall (number of true positive results divided by the number of all samples that should have been identified as positive (TP / (TP + FN)). The F1 score ignores the true negative and may be misleading for unbalanced classes (54). Finally, we computed Matthew's Coefficient Correlation (MCC), which considers all four possible outcomes (TP, TN, FP, and FN) (Fig 5C), and VHIP's scored 0.75 (where 1 represents a model that is perfectly accurate and 0 a model that is making random predictions).

Interestingly, all the features used by the model contain information that can be leveraged for predictions, but based on feature importance determined by the model during the training phase, sequence composition features are the most useful for virus-host predictions (Fig 5D), with %G+C difference encoding the strongest signal the model leverages. Note that because the presence of sequence homology matches between viruses and hosts are rare, we combined instances of homology between viral genomes and CRISPR spacers and between viral genomes and putative host genomes into a single feature termed `homology`. The model is available through conda-forge and PyPI. The source code is available on Github.

To assess the effect of data provenance on model performance, we trained a machine learning model on a subsampled dataset containing 3159 data points but containing an equal amount from each source (i.e., NCBI, Nahant Collection, and the Staphylococcus study). This resulted model has worse score metrics than VHIP (accuracy: 0.829, F1 score: 0.77, ROC: 0.89, and Matthew's correlation: 0.639), which is expected as less data was used during the training phase of the model. This inferior model was then applied on the data unused during training and it predicted interactions with a 91% accuracy rate. This may suggest that data provenance is not a significant driver of the accuracy of VHIP when trained on the entire dataset.

Comparing VHIP to existing host prediction tools is not straightforward. Existing HPTs aim to identify the host for a given virus (i.e., answering the question "what taxa does this virus infect?"). Whereas VHIP was designed to answer a fundamentally distinct question, where given a virus and a list of hosts, the response is "whether infection is predicted to occur or not occur for each host". To enable comparison, we challenged HPTs against known virus-host pairs and recorded their top predictions. Since HPTs were trained on virus-host pairs extracted from the virus NCBI database, we used virus-host pair associations from the Nahant Collection and Staphylococcus study that were not used during training of VHIP, ensuring novel data points across all tools. For each virus-host pair evaluation, we queried VHIP's infection prediction and checked if the correct host was included in each HPT's output. Out of 214 data points, VHIP correctly predicted infections 63.5% of the time. The accuracy of iPHoP, VHMnet, vHULK, and CHERRY at predicting correct hosts was 0.93%, 4.2%, 4.2%, and 2.3% respectively.

3.5 Discussion

In this study, we make available the most comprehensive dataset of experimentally verified virus-microbe interactions from publicly available databases and literature. This holistic dataset allowed a reevaluation of assumptions about the prevalence of narrow versus broad viral host ranges. We assessed taxonomic biases in virus-microbe culturing and

found that the lack of consistency in testing and reporting viral host range against a taxonomically diverse panel of hosts may perpetuate the notion that viruses are specialists. Further, recent technological advances and lab experiments are also challenging this conventional view of viral specialists, as well as revealing the diversity and complexity of virus-host interactions (17, 18, 55). For instance, the Hi-C metagenomic pipeline, which links DNA based on physical proximity before sequencing, frequently results in highly nested networks where viruses' host ranges can span broad taxa (56, 57). These metagenome-based findings are now routinely supported by host range experiments when a broader range of hosts is challenged (36–39, 58). Additional lab experiments testing viruses against diverse host panels and consistent reporting are necessary to ascertain the relative specificity versus breadth of viral host ranges.

Importantly, the dataset of experimentally verified virus-host microbe interactions compiled in this study allowed for the development of a new resource, VHIP, that deviates from existing virus-microbe prediction tools and opens a new avenue to the study of virus-host interaction networks. VHIP is distinct by design in that it predicts infection/non-infection for any given virus-host pair. This approach has multiple benefits. First, VHIP takes both viral and putative host sequences as input, allowing a user to consider all viral and cellular populations recovered from a sampled community (Fig 6A). Second, VHIP may predict a virus to infect multiple different hosts, more accurately reflecting the nature of viral host ranges. Finally, VHIP can resolve complete virus-host interaction networks, which is only possible if a model can explicitly predict both positive and negative relationships between viruses and their potential hosts. Owing to these central design differences, it is impossible to fully compare the accuracy of VHIP with that of existing HPTs. This is because HPTs are designed to return a single host or a list of predicted hosts, and their accuracy calculation considers only the highest score. This is not a problem when the tools are trained on a dataset where there is only one known host, but this limitation becomes an issue when predicting hosts for a novel virus since the predictions are limited to the pool of taxa on which those models were trained. For VHIP, every virus-microbe pair combination is considered, such that the accuracy is defined by the ability of VHIP to accurately infer both infection and non-infection interactions.

The underlying assumption of predicting virus-microbe associations by leveraging genomic signals is that those signals are similar regardless of virus or host taxonomic assignment and/or environmental conditions (e.g., viral adaptation at the genome scale is similar the human gut microbiome to viruses in the oceans). If this assumption is violated, then one must be careful with interpreting predictions from sequence-based tools. To assess this issue, we used a subset of VHRnet containing a smaller number of virus-microbe pairs that contained an equal amount of data points from each data source (NCBI, Nahant Collection, and Staphylococcus). This model was then applied on the data not used for the training/testing phase of the model and obtained an 89% accuracy rate at predicting infection and non-infection events. Furthermore, when using the full dataset, VHRnet was divided into two sets: a set for the training/testing phase of the machine learning model, and a hold-out set to assess the model performance. Because variation can arise from how the data is divided between the two sets, this pipeline was bootstrapped 100 times. For each iteration, there was very little difference in the performance of VHIP (S14 and S15 Figs).

The perspective shift from predicting the most likely taxa a virus can infect to considering all possible virus-host pairs is necessary to resolve virus-host interaction networks (Fig 6A). Interaction networks are mathematical objects that capture and quantify the multitude of potential interactions between species, which provide a common framework for investigations across scales (Fig 6). In such networks, the nodes represent viral and host populations, and an edge connects a virus to a host if it is predicted to infect it. Additional data, sequence or otherwise, can be depicted with networks. Edges can be colored by properties of the interaction that depend on the unique combination of a given virus and host (e.g., viral fitness on a given host, whether infection is lysogenic or lytic) or by the phenotypic properties of the infected cell (virocell) during infection (Fig 6B). Population sizes could be represented by scaling the sizes of nodes, and frequency of infection could be encoded in the width of the edges (Fig 6C). By considering all possible virus-host pairs, such networks can be used to better understand microbial predator-prey interaction patterns and tease apart the underlying processes occurring across scales (50,

59, 60) (Fig 6D), inspiring hypothesis linking such structures and ecosystem properties. Further, virus-host networks explicitly model possible routes of infection-mediated horizontal gene transfer, facilitating the study of gene flow at both population and community scales (Fig 6E). At ecosystem scales, the predicted virus-host infection networks allow us to move beyond ecogenomics as the study of diversity across gradients and instead study how ecological interactions are structured across physicochemical, temporal, and spatial gradients. This shift allows for the integration of multilayer network theory into microbial ecology and opens new opportunities to study ecological complexity (59) (Fig 6F).

The application of network analyses permeates studies in ecology and evolution (61, 62) and contributes to the understanding of community assembly (60, 63, 64), robustness and resilience (64, 65), and species coexistence (66) across biological disciplines. Yet, the application of network thinking to the study of virus-microbe interaction is still in its infancy. The tool developed and presented here represents an important step with the power to leverage metagenomic data to answer the question "who infects whom?" from uncultured sequence data and supports new avenues to apply network thinking to viral ecology.

3.6 Methods

3.6.1 Collection of host range data and associated sequences

GenBank formatted viral genome files (n = 2621) were downloaded from the NCBI RefSeq database (Aug. 2018). Virus-host pairs were retrieved from two sources: (1) metadata fields in the viral genome GenBank file under the host or lab_host description, and (2) literature search reporting host range data. For the first source of host data, the 'lab_host' tag of each of the 2621 viral GenBank files was used to associate the virus with the host used in the sequencing project. If available, the host strain genome was downloaded from RefSeq. If a genome sequence for the host strain was not available, but a genome of the host species was sequenced, a representative genome of the host species was randomly chosen

and downloaded from RefSeq. For the second source of host range data, the 'Title', 'Journal', and 'Author' tags of each viral GenBank file were used to identify primary journal articles. Lab-verified infection and non-infection data for the sequenced viruses were recorded from the identified reference articles. Additional studies that reported host range data for the sequenced viruses were identified via manual literature searches of the virus name. The data was compiled into a single file, named VHRnet for Virus-Host Range network.

3.6.2 Comparison of existing virus-host prediction tools on complete host range experiments

The viral sequences belonging to the *Staphylococcus* study and the Nahant collection study were given as input for the following predictions tools using their default settings: VirHostMatcher-Net (July 2021 version) (30), vHULK (v1.0.0) (31), CHERRY (v1.0.0) (27), iPHoP (v1.2.0) (67), and RaFAH (v0.3) (28). To evaluate accuracy of those tools, the highest score prediction from each tool was considered. There are three possible outcomes: HPT correctly predicted a species that the virus can infect, HPT predicted a species that the virus cannot infect, and HPT predicted a host that was not tested experimentally. In cases where a virus could infect multiple different hosts, a tool only had to predict at least one host among the known hosts to be considered to have 100% accuracy. iPHoP differs from existing HPTs since it can return 0, 1, or multiple predicted hosts for a given virus. We considered separately the best prediction by iPHoP versus the set of hosts predicted by iPHoP when evaluating its performance on the complete host range studies. For RaFAH, it can only return predictions at the genus level so no assessment of its accuracy at the species level was not possible. To calculate accuracy of those tools at higher taxonomic levels, we considered the taxonomic level of the highest score of the predicted species. For example, if a tool predicted E. coli as the most likely host, to determine the accuracy at the phylum level, the phylum of the known hosts were compared to the phylum of E. coli.

3.6.3 Evaluation of commonly used features for virus-host predictions

The most commonly used features are sequence composition (i.e., how similar the pattern of k-mer frequencies between the viruses and their hosts are) and sequence homology (i.e., the presence of a DNA match between a virus and a host) (Fig 1C and 1D). The %G+C content was calculated for all the viral and host genomes using a custom Python script. The difference in %G+C content between viral and host genomes was defined as: viral%G+C—host%G+C. A custom Python script was used to generate k-mer profiles for the viruses and hosts using k-length of 3, 6, and 9, and to calculate similarities for each virus-host pair using the d2* distance metric and the Euclidean distance metric.

Sequence homology, a stretch of DNA that matches between a virus and a host, was used to identify evidence of prior infection (e.g., in the form of remnant integrated prophages, horizontal gene transfer events, or CRISPR spacers). A BLASTn was run between all viral genomes belonging to VHRnet to the NCBI RefSeq (Aug. 2018) sequences database for bacteria and archaea (minimum identity percentage 80 and minimum length 500bp). CRISPR spacers were identified using the CRISPRCasFinder tool (v4.2.20) on all sequences of the NCBI RefSeq sequence database with the following settings: -keepAll -log -cas -ccvRep -rcfowce -getSummaryCasfinder -prokka -meta. Spacer sequences were extracted from the CRISPRCasFinder output. Since spacers are typically 30 to 35 nucleotides long, a BLASTn with the short setting flag was performed between viruses against spacers. Only virus-spacers hits with no mismatches were kept for the CRISPR feature. A Pearson pairwise correlation was performed to assess correlations between features (Table 2).

3.6.4 Comparison of machine learning classifiers using the VHRnet dataset

The signals of coevolution in combination with the knowledge of infection/non-infection for each pair constitute the input needed to explore machine learning model approaches. The input data was first randomly shuffled to ensure that any intrinsic non-random ordering of the data was removed and thus would not influence machine learning behavior. In addition, because the ratio of non-infection to infection in VHRnet is imbalanced (68.7 to 31.3), the host range data was first downsampled to reach a ratio of 60/40 of non-infection to infection. Different machine learning classifiers were tested using the scikit-learn

package (v1.3) in Python (v3.10), namely AdaBoost, GradientBoostingClassifier, KNeighborsClassifier, RandomForest, StochasticDescentGradient, and SupportVectorClassifier (SVC). Each machine learning model has different values settings, herein referred as hyperparameters, that control the learning process during the training phase. The combination of hyperparameters that results in a robust model is both dependent on the training dataset and the features being considered. To determine the best performing machine learning model given the study design for virus-host predictions, we performed a broad grid search (using the GridSearchCV module from scikit-learn) to explore different combinations of hyperparameters.

During the grid search, 70% of the input data was used as the training/testing set and the remaining 30% was kept as a hold-out set. A shuffle split (n = 10) was used to divide the training/testing set into 10 splits, where 9 splits are used to train the model and the remaining one is used to assess performance of the model. This is repeated until each split has been used as the test set. Once the best performing set of hyperparameters was determined using the training/testing set, it was then evaluated on the hold-out set (Fig 1E). This entire process, including the downsampling of non-infection data to obtain a 60/40 ratio of infection to non-infection events, was bootstrapped 50 times for each type of machine learning classifier, except for SVC for which a single grid search was performed due to the runtime required. Code and analysis of the grid search is available here: https://github.com/DuhaimeLab/VHIP_analyses_Bastien_et_al_2023

3.6.5 Training, testing, and evaluation of VHIP

From the previous analysis, we determined that the Gradient Boosting Classifier performed best and therefore is the most appropriate model for virus-host predictions given our study design. We ran a more exhaustive grid search. During the grid search, the data was again downsample to reach a 60/40 ratio of non-infection to infection. For each iteration (n = 100), 70% of the host range data was used for training/testing of the mode, and the remaining 30% kept as hold-out to evaluate the best set of hyperparameters for that iteration (Fig 1E). In addition, when assessing the best combination of hyperparameters,

the AUROC (S14 Fig) and F1 score (S15 Fig) were also computed to assess the model performance and consistency across each iteration. From this pipeline, we determined that the best combination of hyperparameters are: max_depth = 15, learning_rate = 0.75, and loss = exponential. Finally, the model was trained one more time using a shuffle split (n = 10), where 70% of the data was used for training and the remaining 30% for testing the model. The ROC, F1, and MCC scores were calculated using functions provided by the scikit-learn metrics module. Finally, the accuracy of the model was calculated as (TP + TN) / (TP + TN + FP + FN). VHIP is available through conda-forge and PyPI. The source code is available at: https://github.com/DuhaimeLab/VirusHostInteractionPredictor.

To assess the effect of data provenance on our machine learning model, the data was subsampled such that there were equal amounts of data points from each source (n = 3159). Because the subsampling can introduce randomness, this pipeline was bootstrapped 100 times. The accuracy, AUROC, F1 score, and Matthew's correlation were computed for each iteration of the model.

To compare VHIP's prediction ability to existing tools, we evaluated each tool in their ability to recover known infection virus-host pairs. For this assessment, only pairs from the Nahant Collection and Staphylococcus study in the test set were considered to ensure novel data points across all tools. For each virus-host pair evaluation, we queried VHIP's and checked if the correct host was included in each HPT's output.

3.7 Acknowledgements

We thank K. Shedden, B. Hegarty, and M. Moreno for helpful discussions and suggestions and Geoffrey Hannigan for early discussions and collaboration that inspired the hunt for more data.

3.8 VHIP 2.0: Codon usage biases encode signals that improves virus-host interaction predictions from genomic sequences[4]

Metagenomics approaches, enabled by increasingly accessible sequencing technologies, have uncovered a remarkable degree of novel viral diversity across myriad ecosystems (3–7). This perspective has offered insights into the potential roles' viruses play within their microbial communities. However, since community sequencing efforts predominantly yield genomic data rather than organisms at hand for lab testing, this leaves a critical question unanswered: which host(s) do these newly discovered viruses infect?

To address this challenge, we previously developed a machine learning model, Virus-Host Interaction Predictor (VHIP), which is designed to predict virus-host associations from genomic signatures—imprints resulting from past virus-host interactions that are encoded in the virus and host genomes (68). The original model employed a training dataset of 8849 lab-validated virus-host infection pairs compiled in 2018 (68). From this set, it leveraged k-mer similarities, GC content differences, and sequence homology to predict virus-host interactions using publicly available host range data with 87.8% accuracy.

To improve the predictive power of VHIP, we have since expanded the training dataset to include new viruses whose genomes have been deposited in NCBI repositories since the initial compilation in 2018. This expansion was supplemented by a comprehensive literature review of scientific articles reporting host ranges of these newly included viruses. This effort resulted in the addition of 1,379 virus-host pairs, bringing the total training set for VHIP 2.0 to 10,228 pairs.

In this model update, we expanded the feature set to include new metrics of virus-host adaptation from genomic signatures, including codon usage bias, amino acid frequency distributions, and relative synonymous codon usage (RSCU). For each metric, we fitted linear models to quantify the degree of alignment (R2 values) between viral and their

[4] This section is based on a manuscript in preparation, co-authored with Evelyn Faust and Melissa Duhaime

hosts. Additionally, we adapted the Translational Adaptation Analysis Index (TAAI) metric, originally designed to describe how well an organism's genes are adapted to its own machinery (69), to instead quantify viral gene adaptation to host machinery.

The integration of expanded host range data and novel gene-derived features significantly improved model performance. Accuracy increased from 87.8% to 92.0%, with corresponding improvements in F1 score (0.93 to 0.96), AUROC (0.94 to 0.96), and Matthews Correlation Coefficient (0.75 to 0.78) (Fig 7). This enhanced model, VHIP 2.0, is made available as a Python package through both pip and conda channels. The source code is publicly accessible on GitHub, accompanied by comprehensive documentation and usage examples (https://github.com/DuhaimeLab/VirusHostInteractionPredictor). VHIP 2.0 inputs are gene and genome sequences of viruses and hosts of interest. The genomic signals are then calculated from the sequences, which are subsequently used by the model to determine infection status for each possible virus-host pair. The model reports include both the confidence score associated with the prediction and the features used to predict the interaction between each virus-host pair.

3.9 Figures

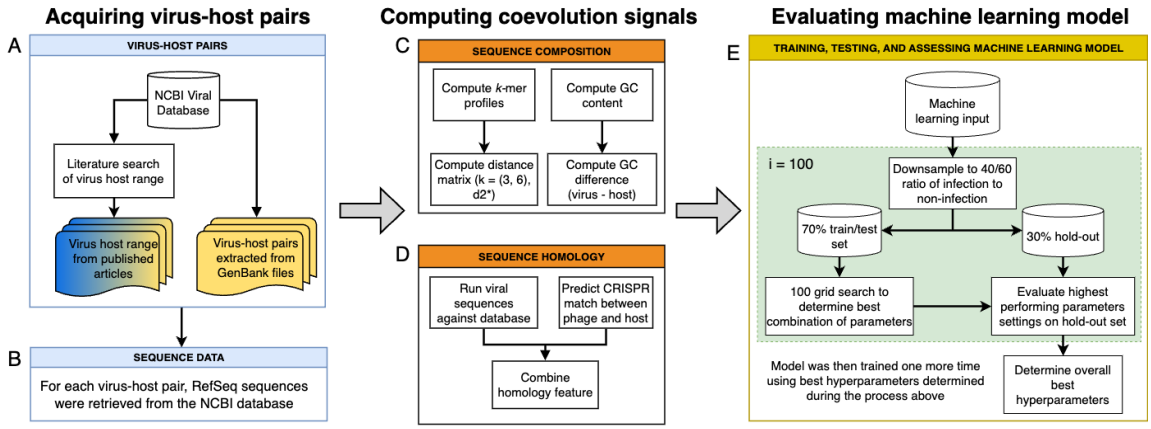


Figure 3.1. Design of data collection, features computation, and evaluation of machine learning approach for virus-host interaction predictions.

A. Metadata describing virus-host pairs was retrieved and compiled from NCBI and literature. Blue and yellow color indicates provenance of infection versus non-infection information, i.e., non-infection data came only from literature studies. **B.** Sequences of each virus and host were retrieved. Signals of coevolution were calculated for each virus-host pair, including **C.** sequence composition and **D.** sequence homology. **E.** The virus-host pairs and their computed signals of coevolution were used to train VHIP. To determine the best parameters for the machine learning model, a grid search was performed and bootstrapped 100 times on a training/testing set and evaluated on a hold-out set. The model was then retrained for a final time using the best hyperparameters.

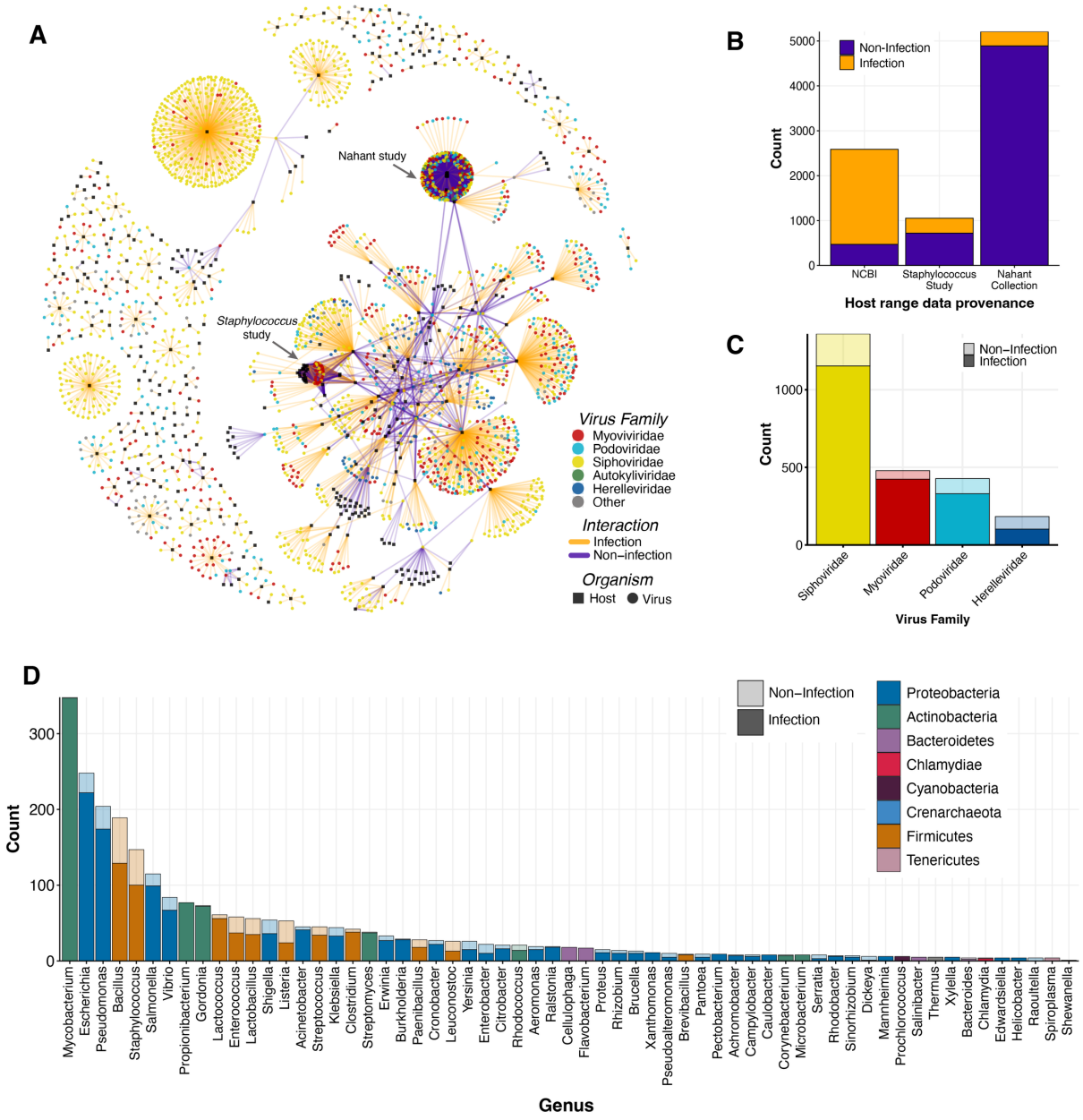


Figure 3.2 VHRnet network visualization and content characteristics

A. Network visualization of VHRnet, where an edge connects a viral node to a host node if that pair has been experimentally tested. The edge is colored by the interaction class (infection/non-infection) and the nodes by virus taxonomy or whether it is a host node (black squares are host species). **B.** Origin of known lab-tested infection and non-infection data across dataset sources. **C.** Distribution of family classifications for subset of viruses in VHRnet collected from NCBI. Lighter transparency represents the proportion of non-infection reports by viral family, relative to the solid portion, which represents known infection reports by viral family. **D.** Distribution of host genera for subset of hosts in VHRnet collected from NCBI. Lighter color transparency represents the proportion of non-infection relative to infection (solid color).

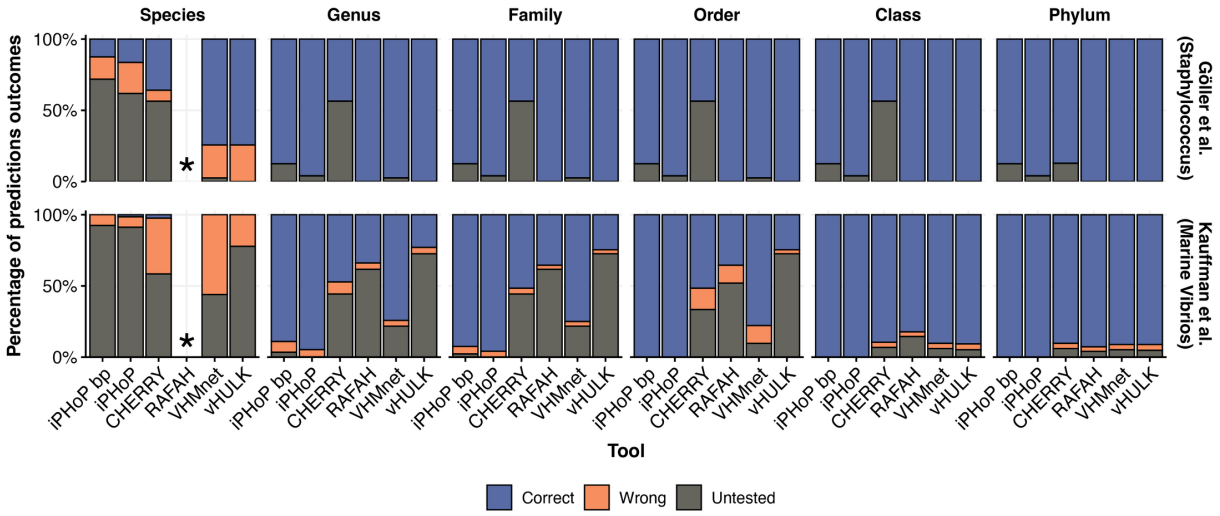


Figure 3.3 Evaluation of the accuracy of existing virus-host prediction models using two complete host range datasets

Accuracy is evaluated at different host taxa thresholds, from species to phylum. The relative proportion of correct, incorrect, and untested predictions is represented by bar color. Untested predictions are hosts predicted by an HPT but that are not experimentally validated as correct or incorrect. The prediction that received the highest score by each HPT was evaluated. Because iPHoP outputs multiple predictions, both the highest scoring prediction ("iPHoP bp") and the full set of predictions ("iPHoP") were reported. *RAFAH does not return predictions at the species level.

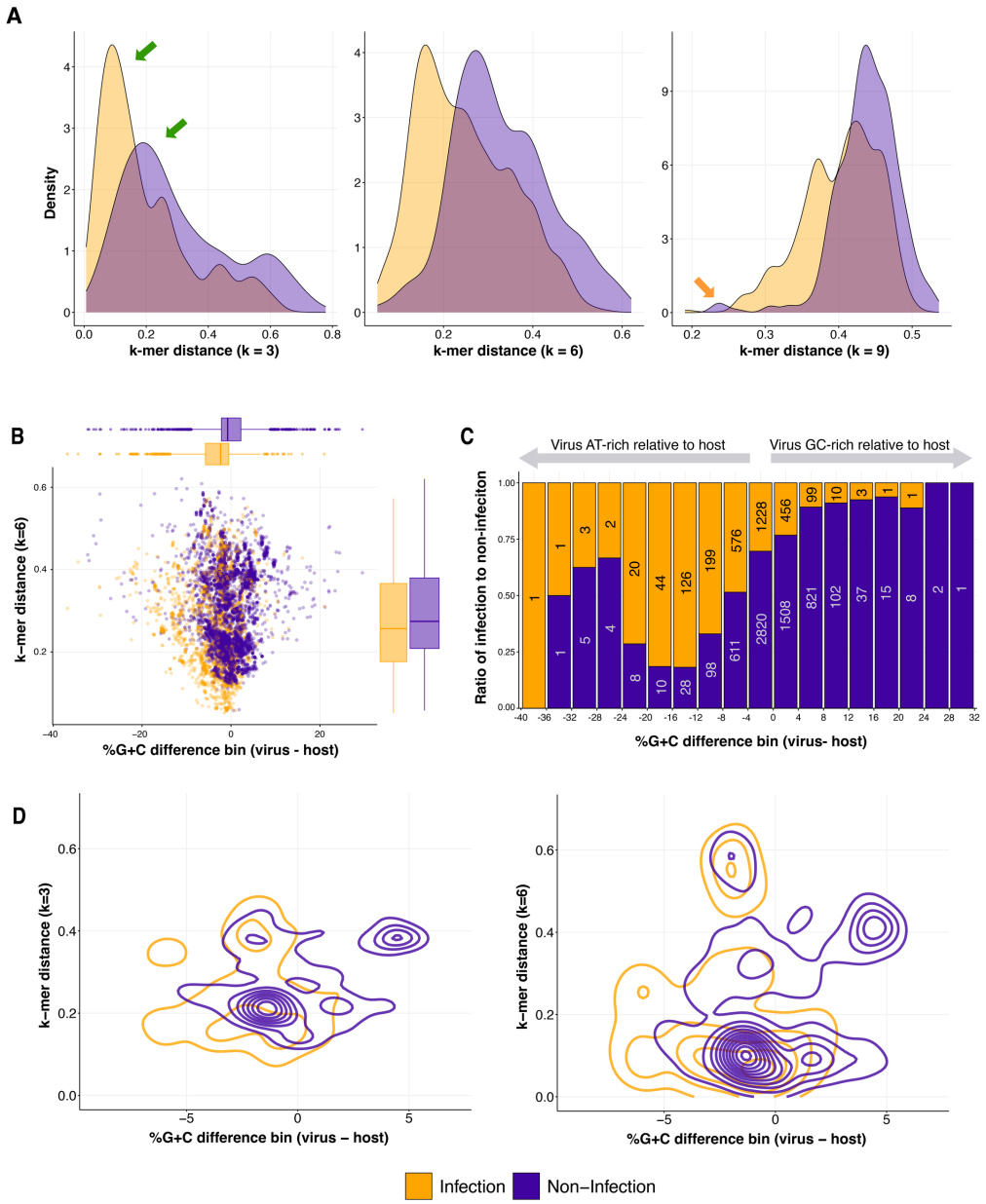


Figure 3.4 Comparison of coevolution signals captures at the species taxonomic level

A. Density plots of k-mer distances using k-length of 3, 6, and 9 between viruses and their hosts, colored by known interaction. **B.** Dot plot of %G+C difference bin against *k*-mer distance using d2*, colored by known interaction. Box and whisker plots represent the distribution values for the x-axis and y-axis, colored by known interaction (middle bar is median, quartile shows 25th and 75th percentile, and whiskers shows 1.5 times interquartile range). **C.** Stacked bar plot showing the ratio of lab-verified infection to non-infection for %G+C difference bins. Numbers inside the bars note the number of events observed for each %G+C difference bin. Note that this number decreases sharply towards the extremes. **D.** Density plot of %G+C difference and *k*-mer distance between viruses and lab-tested hosts, colored by known interaction. Left plot is with k set to 3 and the right plot is with k set to 6.

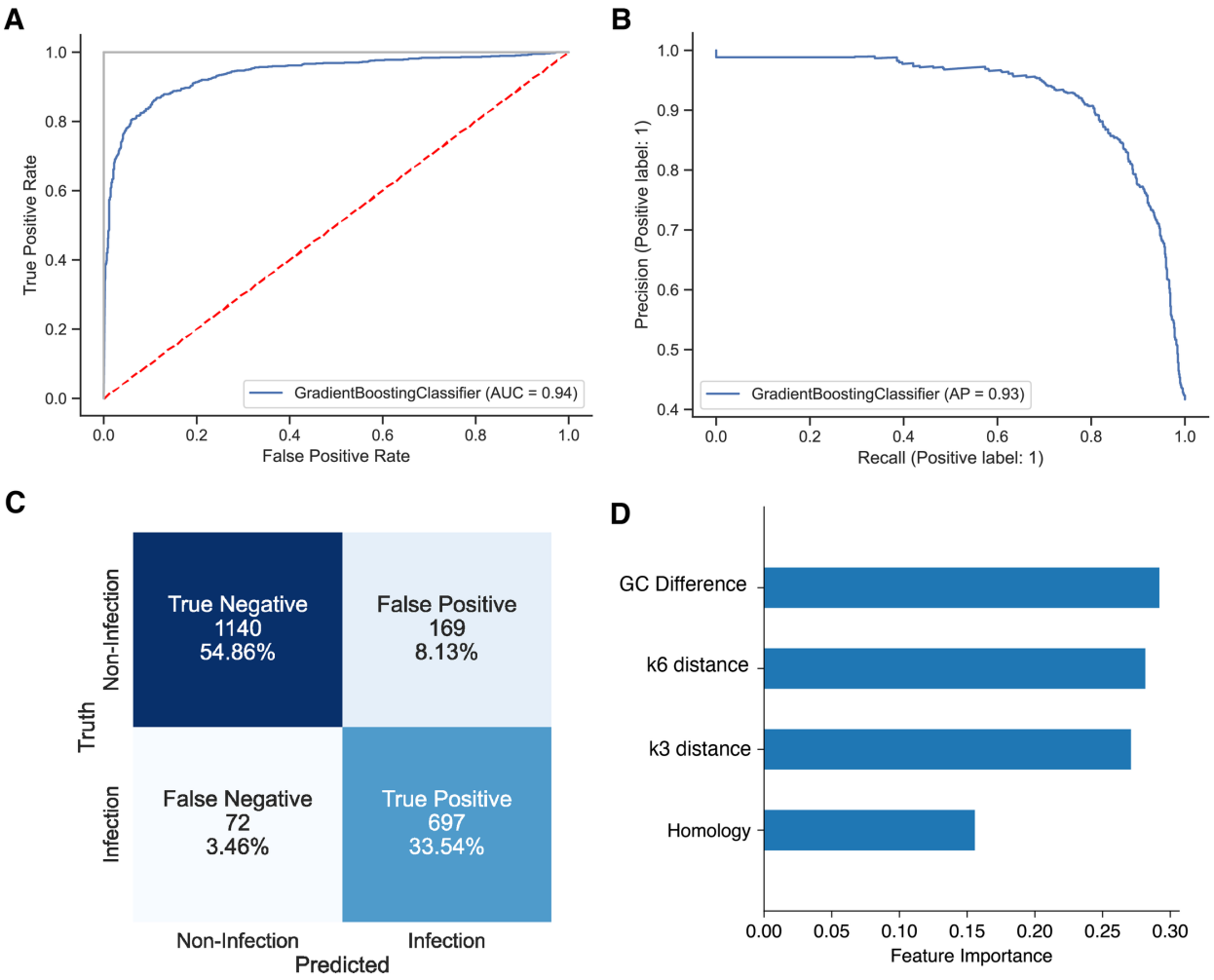


Figure 3.5 Model performance and feature evaluation

A. ROC curve of final trained Gradient Boosting Classifier. Red dashed line represents a model that makes random predictions, whereas the gray line represents a model that has 100% accuracy. The blue line is the behavior for VHIP. **B.** F1 curve of the final trained Gradient Boosting Classifier. **C.** Confusion matrix showing all possible outcomes, counts, and percentages. **D.** Relative importance of each feature that VHIP used to predict virus-host interactions.

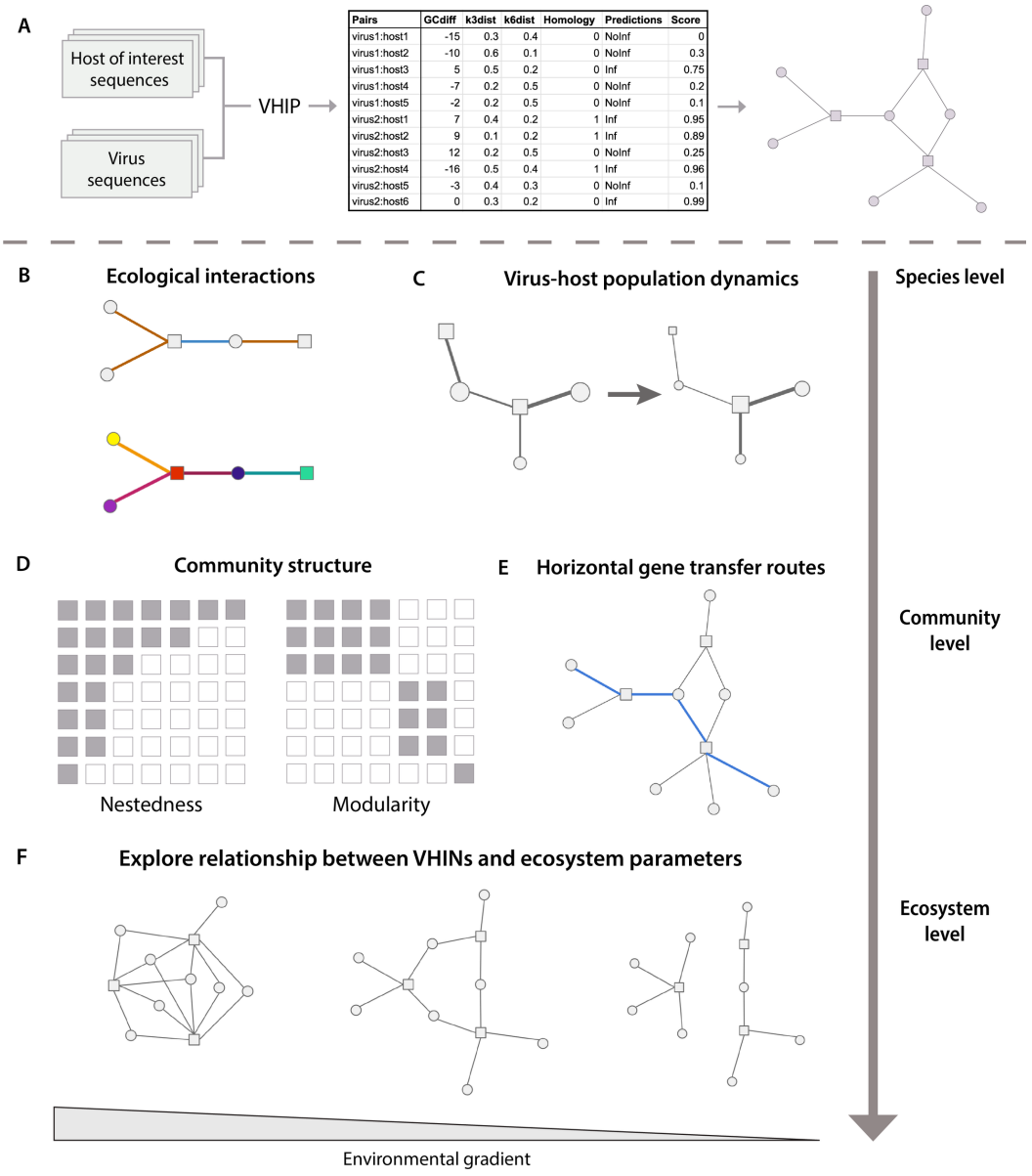


Figure 3.6 Model output and future directions

A. VHIP predicts interactions between all possible virus-host combinations from the input sequences of viruses and potential hosts. Because all possible pairs are considered, the VHIP predictions can be visualized as a bipartite network of virus-host interactions. **B.** Additional information describing the nodes (biological entities) and edges (interactions) can be mapped onto the network. **C.** Virus-host population dynamics can be coded and visualized in the nodes and edges. **D.** Community structure can be studied by considering the entire set of predicted interactions. **E.** Edges in the network can be thought of as different possible routes for horizontal gene transfer to occur. **F.** By comparing networks across space, time, and/or environmental gradients can help understand how virus-host interactions are structured and their underlying processes manifest at ecosystem scales.

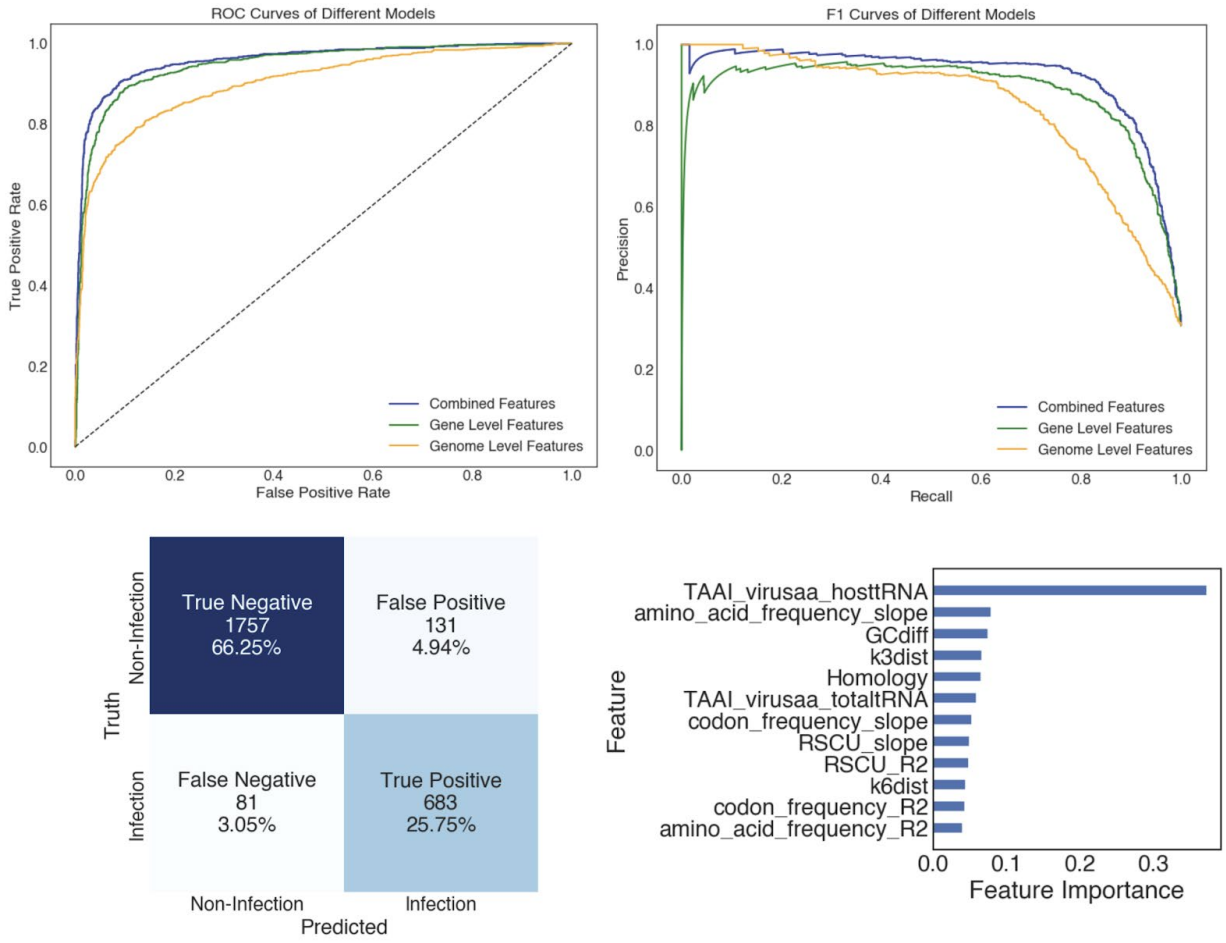
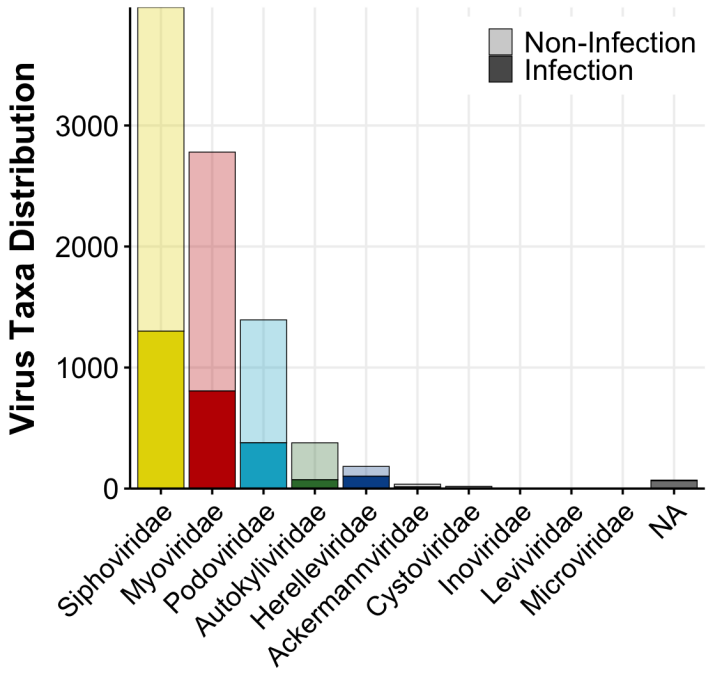


Figure 3.7. Model improvement by incorporating gene-derived metrics of virus-host adaptation.

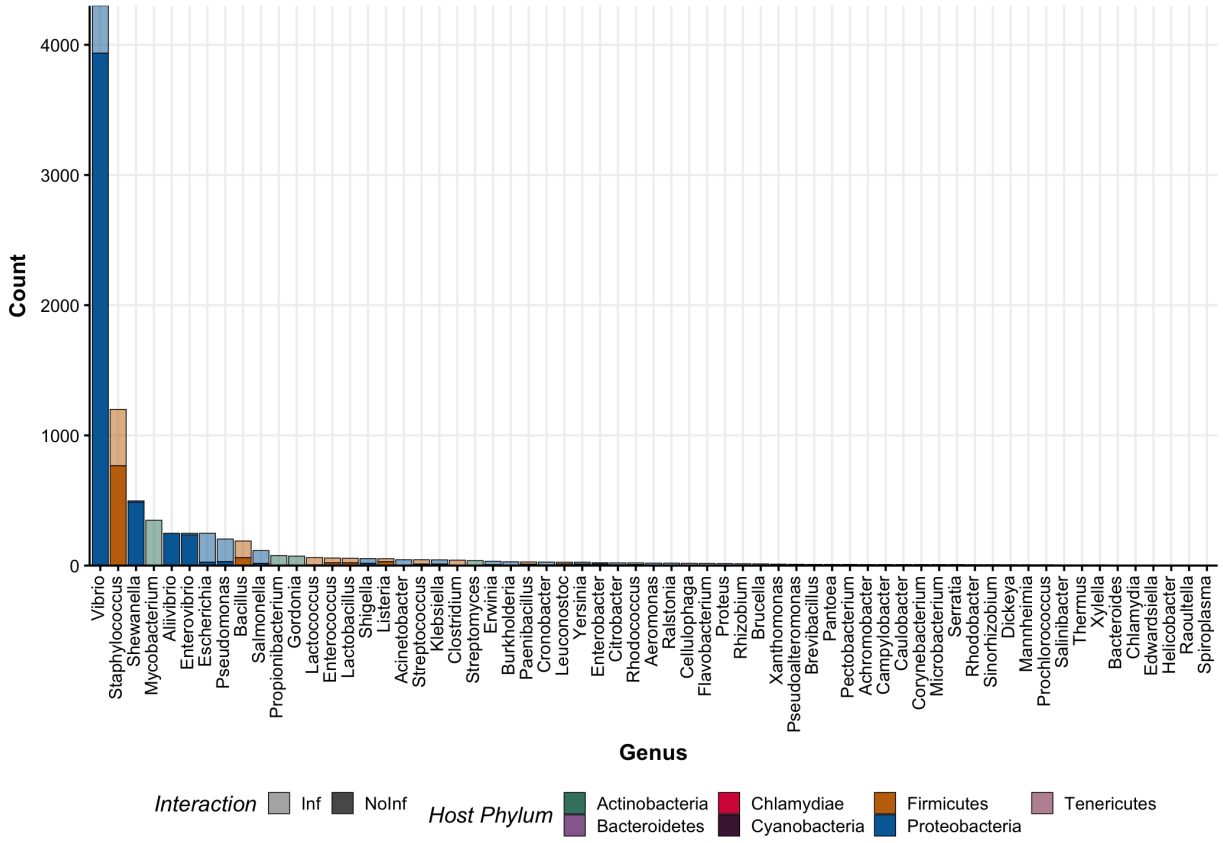
A. ROC curves machine learning models, comparing the performance of models between models that uses only gene-level features, only genome-level features, and combined. The dashed line represents a model that makes random predictions. **B.** F1 curves. **C.** Confusion matrix of the combined model showing all possible outcomes. **D.** Relative importance of each feature when considering all features. The adapted TAAI metric is by the most informative for predicting virus-host interactions.

3.10 Supplementary figures



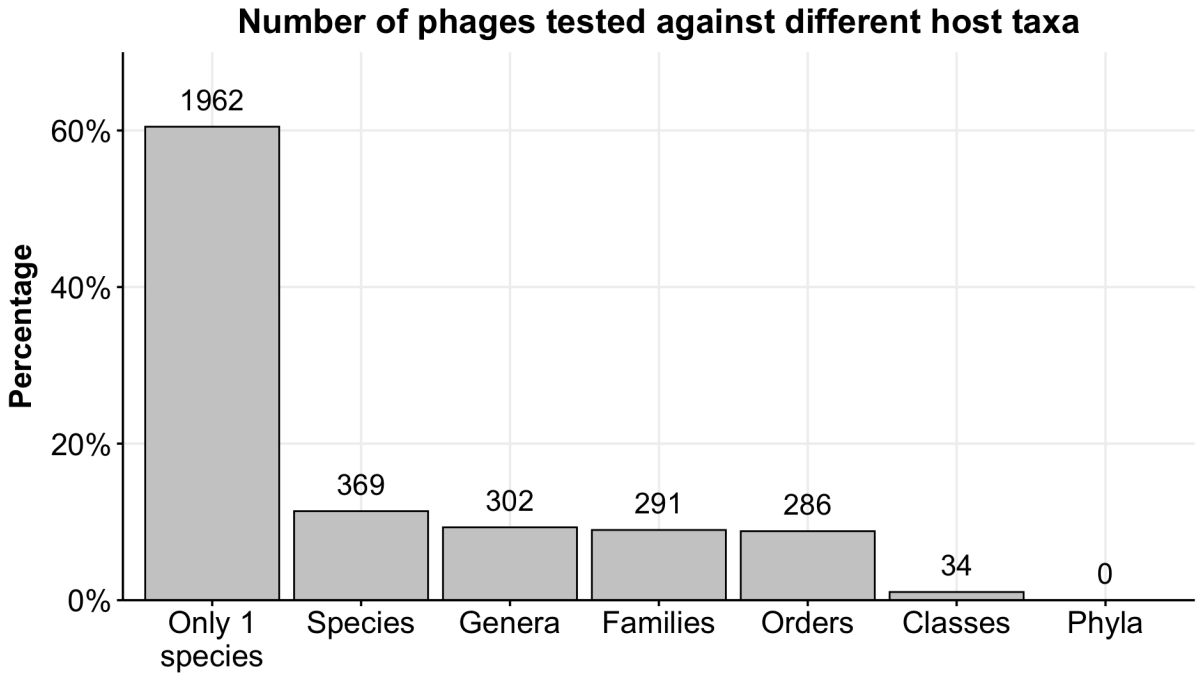
Supplemental Figure 3.8. Distribution of all family classifications of viruses in VHRnet

Lighter transparency represents the proportion of non-infection reports by viral family, relative to the solid portion, which represents known infection reports by viral family.



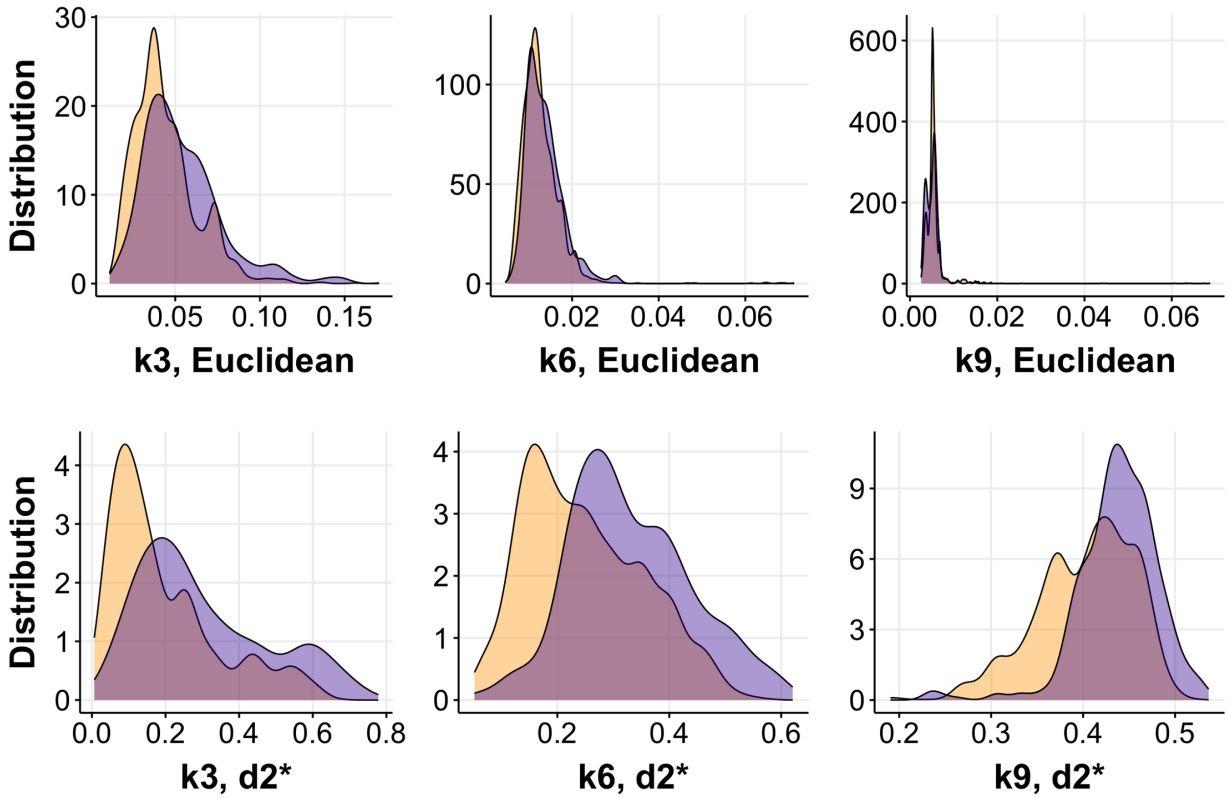
Supplemental Figure 3.9. Distribution of all host genera represented in VHRnet

Lighter color transparency represents the proportion of non-infection relative to infection (solid color).



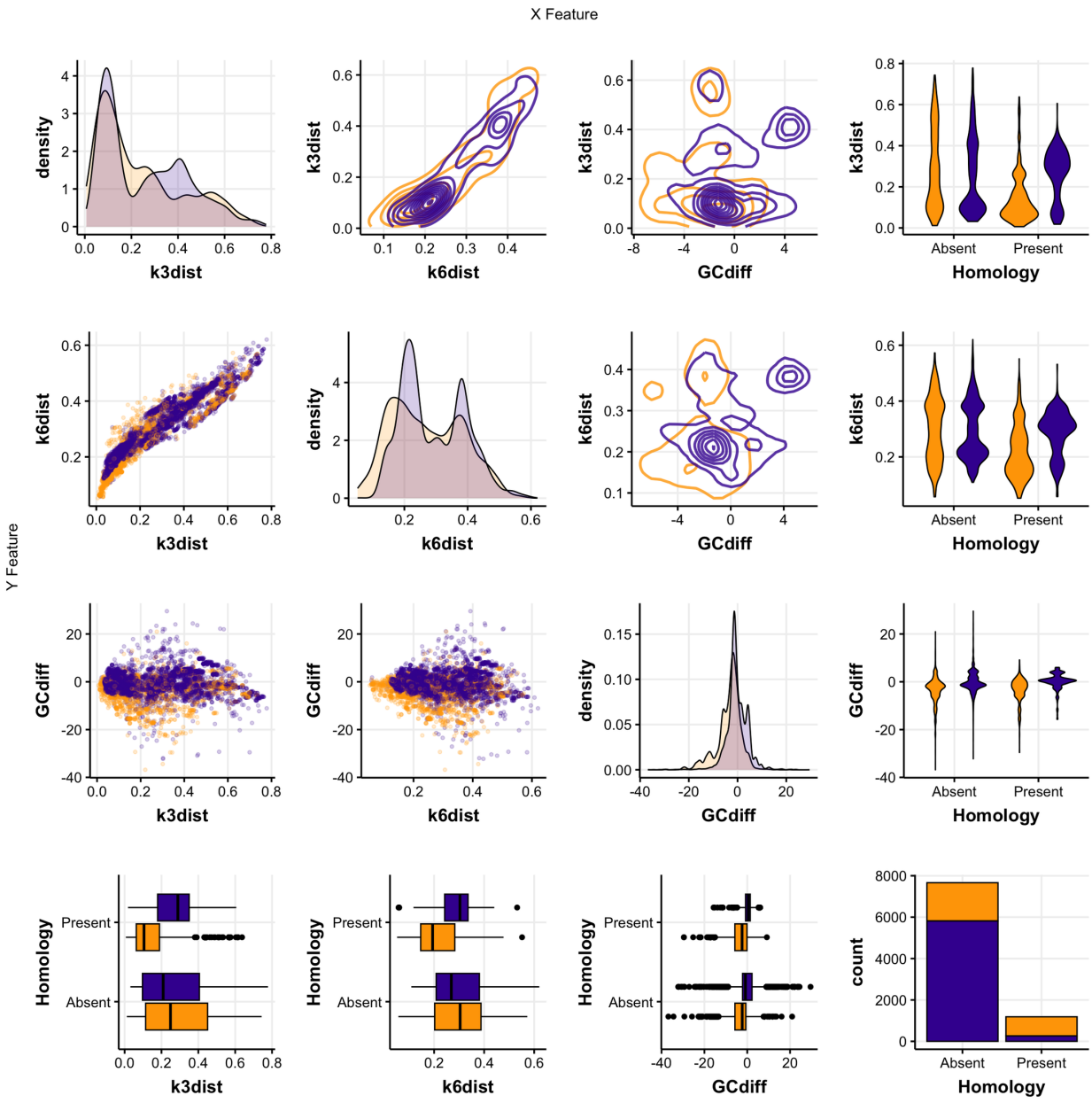
Supplemental Figure 3.10 Number of viruses tested against different host taxa.

X-axis represents the number of viruses that have been tested.

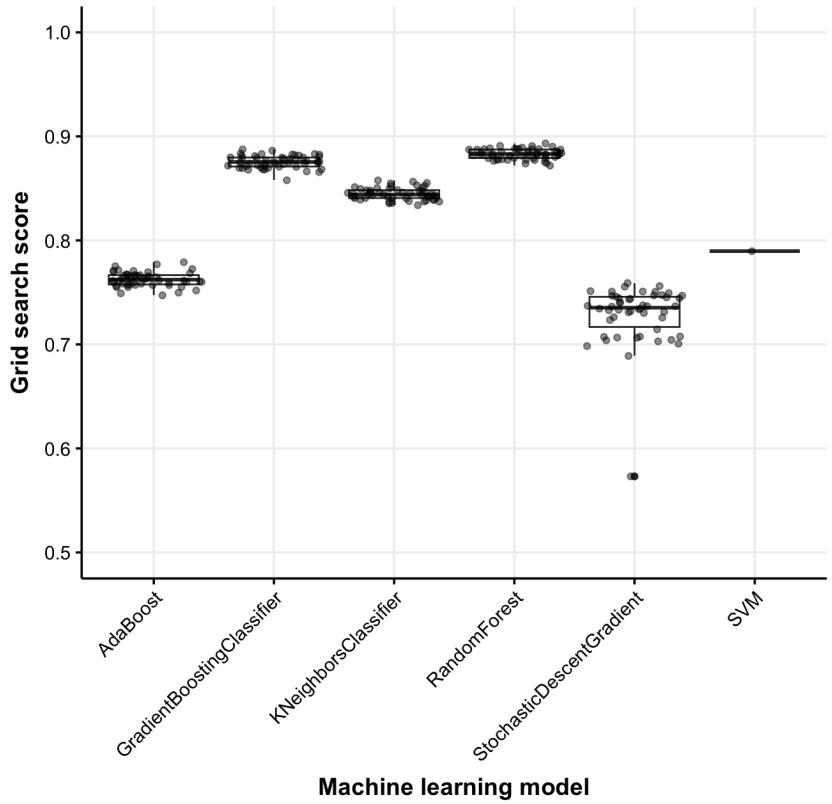


Supplemental Figure 3.11 Kernel density of distance measurements for each virus-host pair and colored by interaction

Yellow is for infection, blue for non-infection. The top row used the Euclidean distance to compute the similarity between the k-mer profiles of the virus and its host, while the second row uses the d2* distance metric. Each column represents a different length of k-mer used to create the k-mer profiles (k-length of 3 versus 6 versus 9). The d2* distance metric is a more appropriate metric than the Euclidean distance metric for the purpose of virus-host prediction since it encodes some evolutionary signals (the peaks for the no-infection and infection are separated).

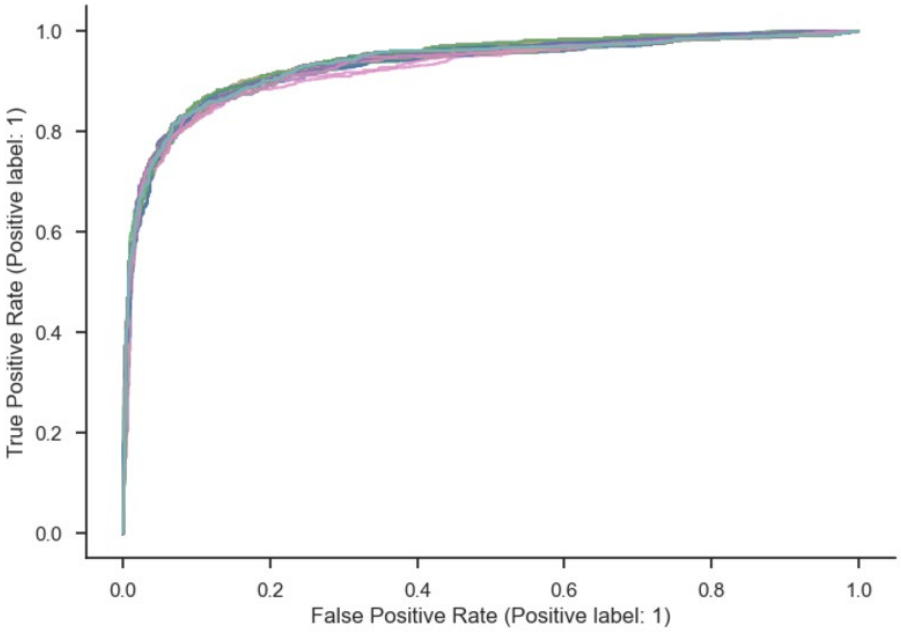


Supplemental Figure 3.12. Feature distribution (diagonal plots) and features co-correlations (non-diagonal plots)

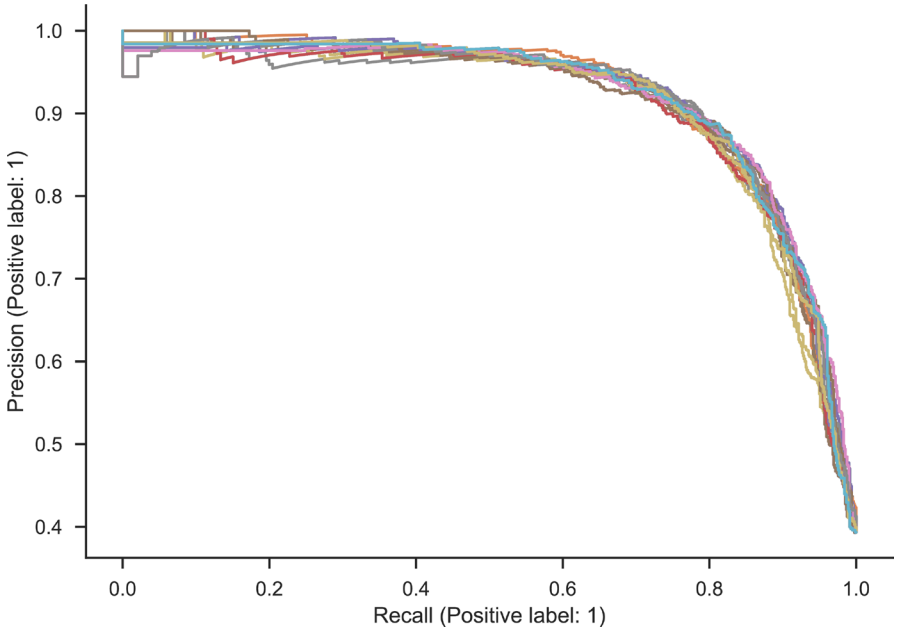


Supplemental Figure 3.13. Comparison of different machine learning models trained on VHRnet

For each type of machine learning model, a grid search was performed to determine the best combinations of parameters. This plot shows the accuracy of the best performing model. This was bootstrapped 50 times (except for SVM since the fit algorithm is O(n^2)).



Supplemental Figure 3.14. ROC curves from 100 bootstrapping iterations of the best model trained during the grid search



Supplemental Figure 3.15. F1 curves from 100 bootstrapping iterations of the best model during the grid search

Tables

Tool name	*Type*	*Tool output*	*Considerations*
IPHoP	Modular machine-learning framework	Most likely taxa a virus can infect, a score associated with the prediction, and the main method that was used to determine the prediction.	Uses the output of other existing tools. It may return more than 1 lijkely host or 0. By default, only return results for which the score is above 0.9. It uses the following tools: CRISPRCasFinder, VirHostMatcher, WIsH, PHP, and RAFAH.
CHERRY	Graph encoder-decoder model.	Most likely taxa a virus can infect and a score associated with that prediction.	Model can be retrained with additional virus-host pairs knowledge (only infection events are considered). The tool can be used to predict viruses infecting prokaryotes of interest, where in this case, prokaryote sequences are given as input instead of virus sequences.
RAFAH	Random forest machine learning	For each virus, it returns a score value against each genus included in the RAFAH database. It emphasizes a single genus (column Predicted_Host) based on the score value (Winner_Score).	Cannot make predictions at the species level.
VirHostMatcher-Net	Network inferred	Can only work if viruses of interest share at least 1 marker gene with a known virus reference.	One of the options allows the user to return multiple predicted hosts for viruses by defining topN parameter.
vHULK	Deep neural network	Predicts most likely genus and species a virus can infect with their respective score values.	Only work if viruses of interest share at least 1 marker gene with a known virus reference. Relies on gene annotation for accurate predictions.

Table 3.1. Comparison between input and output of existing host-prediction tools

	%G+C difference	k-3 distance	k-6 distance
%G+C difference	0.0316		
k-3 distance	0.0296	0.1286	
k-6 distance	0.1286	0.201	0.201

Table 3.2. Pearson pairwise correlations between features of VHIP

3.11 References

1. Gilbert,J.A. and Dupont,C.L. (2011) Microbial Metagenomics: Beyond the Genome. *Annual Review of Marine Science*, **3**, 347–371.
2. Garza,D.R. and Dutilh,B.E. (2015) From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. *Cell. Mol. Life Sci.*, **72**, 4287–4308.
3. Gregory,A.C., Zayed,A.A., Conceição-Neto,N., Temperton,B., Bolduc,B., Alberti,A., Ardyna,M., Arkhipova,K., Carmichael,M., Cruaud,C., *et al.* (2019) Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*, **177**, 1109-1123.e14.
4. Coutinho,F.H., Silveira,C.B., Gregoracci,G.B., Thompson,C.C., Edwards,R.A., Brussaard,C.P.D., Dutilh,B.E. and Thompson,F.L. (2017) Marine viruses discovered via metagenomics shed light on viral strategies throughout the oceans. *Nat Commun*, **8**, 1–12.
5. Benler,S., Yutin,N., Antipov,D., Rayko,M., Shmakov,S., Gussow,A.B., Pevzner,P. and Koonin,E.V. (2021) Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome*, **9**, 78.
6. Mokili,J.L., Rohwer,F. and Dutilh,B.E. (2012) Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology*, **2**, 63–77.
7. Rosario,K. and Breitbart,M. (2011) Exploring the viral world through metagenomics. *Current Opinion in Virology*, **1**, 289–297.
8. Brum,J.R., Ignacio-Espinoza,J.C., Roux,S., Doulcier,G., Acinas,S.G., Alberti,A., Chaffron,S., Cruaud,C., Vargas,C. de, Gasol,J.M., *et al.* (2015) Patterns and ecological drivers of ocean viral communities. *Science*, **348**.
9. Roux,S., Emerson,J.B., Eloe-Fadrosh,E.A. and Sullivan,M.B. (2017) Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ*, **5**, e3817.
10. Camargo,A.P., Nayfach,S., Chen,I.-M.A., Palaniappan,K., Ratner,A., Chu,K., Ritter,S.J., Reddy,T.B.K., Mukherjee,S., Schulz,F., *et al.* (2023) IMG/VR v4: an expanded database

of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res*, **51**, D733–D743.

11. Roux,S., Adriaenssens,E.M., Dutilh,B.E., Koonin,E.V., Kropinski,A.M., Krupovic,M., Kuhn,J.H., Lavigne,R., Brister,J.R., Varsani,A., *et al.* (2019) Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat Biotechnol*, **37**, 29–37.
12. Breitbart,M. (2011) Marine Viruses: Truth or Dare. *Annu. Rev. Mar. Sci.*, **4**, 425–448.
13. Gayder,S., Parcey,M., Nesbitt,D., Castle,A.J. and Svircev,A.M. (2020) Population Dynamics between Erwinia amylovora, Pantoea agglomerans and Bacteriophages: Exploiting Synergy and Competition to Improve Phage Cocktail Efficacy. *Microorganisms*, **8**, 1449.
14. Maslov,S. and Sneppen,K. (2017) Population cycles and species diversity in dynamic Kill-the-Winner model of microbial ecosystems. *Scientific Reports*, **7**, 39642.
15. McDaniel,L.D., Young,E., Delaney,J., Ruhnau,F., Ritchie,K.B. and Paul,J.H. (2010) High Frequency of Horizontal Gene Transfer in the Oceans. *Science*, **330**, 50–50.
16. Soucy,S.M., Huang,J. and Gogarten,J.P. (2015) Horizontal gene transfer: building the web of life. *Nat Rev Genet*, **16**, 472–482.
17. Forterre,P. (2013) The virocell concept and environmental microbiology. *ISME J*, **7**, 233–236.
18. Howard-Varona,C., Lindback,M.M., Bastien,G.E., Solonenko,N., Zayed,A.A., Jang,H., Andreopoulos,B., Brewer,H.M., Glavina del Rio,T., Adkins,J.N., *et al.* (2020) Phage-specific metabolic reprogramming of virocells. *ISME J*, **14**, 881–895.
19. Zimmerman,A.E., Howard-Varona,C., Needham,D.M., John,S.G., Worden,A.Z., Sullivan,M.B., Waldbauer,J.R. and Coleman,M.L. (2020) Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nat Rev Microbiol*, **18**, 21–34.
20. Enav,H., Mandel-Gutfreund,Y. and Béjà,O. (2014) Comparative metagenomic analyses reveal viral-induced shifts of host metabolism towards nucleotide biosynthesis. *Microbiome*, **2**, 9.
21. Duhaime,M.B., Kottmann,R., Field,D. and Glöckner,F.O. (2011) Enriching public descriptions of marine phages using the Genomic Standards Consortium MIGS standard. *Stand in Genomic Sci*, **4**, 271–285.

22. Coclet,C. and Roux,S. (2021) Global overview and major challenges of host prediction methods for uncultivated phages. *Current Opinion in Virology*, **49**, 117–126.
23. Albery,G.F., Eskew,E.A., Ross,N. and Olival,K.J. (2020) Predicting the global mammalian viral sharing network using phylogeography. *Nat Commun*, **11**, 2260.
24. Albery,G.F., Becker,D.J., Brierley,L., Brook,C.E., Christofferson,R.C., Cohen,L.E., Dallas,T.A., Eskew,E.A., Fagre,A., Farrell,M.J., *et al.* (2021) The science of the host–virus network. *Nat Microbiol*, **6**, 1483–1492.
25. Poisot,T., Ouellet,M.-A., Mollentze,N., Farrell,M.J., Becker,D.J., Brierley,L., Albery,G.F., Gibb,R.J., Seifert,S.N. and Carlson,C.J. (2023) Network embedding unveils the hidden interactions in the mammalian virome. *Patterns*, **4**, 100738.
26. Meyer,K.M., Memiaghe,H., Korte,L., Kenfack,D., Alonso,A. and Bohannan,B.J.M. (2018) Why do microbes exhibit weak biogeographic patterns? *ISME J*, **12**, 1404–1413.
27. Shang,J. and Sun,Y. (2022) CHERRY: a Computational metHod for accuratE pRediction of virus-pRokarYotic interactions using a graph encoder-decoder model. *Briefings in Bioinformatics*, 10.1093/bib/bbac182.
28. Coutinho,F.H., Zaragoza-Solas,A., López-Pérez,M., Barylski,J., Zielezinski,A., Dutilh,B.E., Edwards,R. and Rodriguez-Valera,F. (2021) RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns*, **2**, 100274.
29. Roux,S., Camargo,A.P., Coutinho,F.H., Dabdoub,S.M., Dutilh,B.E., Nayfach,S. and Tritt,A. (2022) iPHoP: an integrated machine-learning framework to maximize host prediction for metagenome-assembled virus genomes. 10.1101/2022.07.28.501908.
30. Wang,W., Ren,J., Tang,K., Dart,E., Ignacio-Espinoza,J.C., Fuhrman,J.A., Braun,J., Sun,F. and Ahlgren,N.A. (2020) A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genom Bioinform*, **2**.
31. Amgarten,D., Iha,B.K.V., Piroupo,C.M., da Silva,A.M. and Setubal,J.C. (2022) vHULK, a New Tool for Bacteriophage Host Prediction Based on Annotated Genomic Features and Neural Networks. *PHAGE*, 10.1089/phage.2021.0016.
32. Lucks,J.B., Nelson,D.R., Kudla,G.R. and Plotkin,J.B. (2008) Genome Landscapes and Bacteriophage Codon Usage. *PLOS Computational Biology*, **4**, e1000001.

33. Esposito,L.A., Gupta,S., Streiter,F., Prasad,A. and Dennehy,J.J. (2016) Evolutionary interpretations of mycobacteriophage biodiversity and host-range through the analysis of codon usage bias. *Microbial Genomics*, **2**, e000079.
34. Edwards,R.A., McNair,K., Faust,K., Raes,J. and Dutilh,B.E. (2016) Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol Rev*, **40**, 258–272.
35. Ahlgren,N.A., Ren,J., Lu,Y.Y., Fuhrman,J.A. and Sun,F. (2017) Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res*, **45**, 39–53.
36. Kauffman,K.M., Hussain,F.A., Yang,J., Arevalo,P., Brown,J.M., Chang,W.K., VanInsberghe,D., Elsherbini,J., Sharma,R.S., Cutler,M.B., *et al.* (2018) A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature*, **554**, 118–122.
37. Malki,K., Kula,A., Bruder,K., Sible,E., Hatzopoulos,T., Steidel,S., Watkins,S.C. and Putonti,C. (2015) Bacteriophages isolated from Lake Michigan demonstrate broad host-range across several bacterial phyla. *Virology Journal*, **12**, 164.
38. Beumer,A. and Robinson,J.B. (2005) A Broad-Host-Range, Generalized Transducing Phage (SN-T) Acquires 16S rRNA Genes from Different Genera of Bacteria. *Applied and Environmental Microbiology*, 10.1128/AEM.71.12.8301-8304.2005.
39. Feng,X., Yan,W., Wang,A., Ma,R., Chen,X., Lin,T.-H., Chen,Y.-L., Wei,S., Jin,T., Jiao,N., *et al.* (2021) A Novel Broad Host Range Phage Infecting Alteromonas. *Viruses*, **13**, 987.
40. Göller,P.C., Elsener,T., Lorgé,D., Radulovic,N., Bernardi,V., Naumann,A., Amri,N., Khatchatourova,E., Coutinho,F.H., Loessner,M.J., *et al.* (2021) Multi-species host range of staphylococcal phages isolated from wastewater. *Nat Commun*, **12**, 6965.
41. Krishnamurthy,S.R. and Wang,D. (2017) Origins and challenges of viral dark matter. *Virus Research*, **239**, 136–142.
42. Pires,D.P., Costa,A.R., Pinto,G., Meneses,L. and Azeredo,J. (2020) Current challenges and future opportunities of phage therapy. *FEMS Microbiology Reviews*, **44**, 684–700.
43. Burstein,D., Sun,C.L., Brown,C.T., Sharon,I., Anantharaman,K., Probst,A.J., Thomas,B.C. and Banfield,J.F. (2016) Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat Commun*, **7**, 10613.

44. Breitbart,M., Bonnain,C., Malki,K. and Sawaya,N.A. (2018) Phage puppet masters of the marine microbial realm. *Nature Microbiology*, **3**, 754–766.
45. Carbone,A. (2008) Codon Bias is a Major Factor Explaining Phage Evolution in Translationally Biased Hosts. *J Mol Evol*, **66**, 210–223.
46. Pride,D.T., Wassenaar,T.M., Ghose,C. and Blaser,M.J. (2006) Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. *BMC Genomics*, **7**, 8.
47. Lawrence,J.G. and Ochman,H. (1997) Amelioration of Bacterial Genomes: Rates of Change and Exchange. *J Mol Evol*, **44**, 383–397.
48. Song,K., Ren,J., Zhai,Z., Liu,X., Deng,M. and Sun,F. (2013) Alignment-Free Sequence Comparison Based on Next-Generation Sequencing Reads. *Journal of Computational Biology*, **20**, 64–79.
49. Rocha,E.P.C. and Danchin,A. (2002) Base composition bias might result from competition for metabolic resources. *Trends in Genetics*, **18**, 291–294.
50. Hannigan,G.D., Duhaime,M.B., Koutra,D. and Schloss,P.D. (2018) Biogeography and environmental conditions shape bacteriophage-bacteria networks across the human microbiome. *PLOS Computational Biology*, **14**, e1006099.
51. Hannigan,G.D., Duhaime,M.B., Ruffin,M.T., Koumpouras,C.C. and Schloss,P.D. (2018) Diagnostic Potential and Interactive Dynamics of the Colorectal Cancer Virome. *mBio*, **9**.
52. Topçuoğlu,B.D., Lesniak,N.A., Ruffin,M.T., Wiens,J. and Schloss,P.D. (2020) A Framework for Effective Application of Machine Learning to Microbiome-Based Classification Problems. *mBio*, **11**.
53. Cai,J., Luo,J., Wang,S. and Yang,S. (2018) Feature selection in machine learning: A new perspective. *Neurocomputing*, **300**, 70–79.
54. Chicco,D. and Jurman,G. (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, **21**, 6.
55. Hobbs,Z. and Abedon,S.T. (2016) Diversity of phage infection types and associated terminology: the problem with 'Lytic or lysogenic'. *FEMS Microbiology Letters*, **363**, fnw047.

56. Hwang,Y., Rahlff,J., Schulze-Makuch,D., Schloter,M. and Probst,A.J. Diverse Viruses Carrying Genes for Microbial Extremotolerance in the Atacama Desert Hyperarid Soil. *mSystems*, **6**, e00385-21.
57. Cheng,Z., Li,X., Palomo,A., Yang,Q., Han,L., Wu,Z., Li,Z., Zhang,M., Chen,L., Zhao,B., *et al.* (2023) Virus impacted community adaptation in oligotrophic groundwater environment revealed by Hi-C coupled metagenomic and viromic study. *Journal of Hazardous Materials*, **458**, 131944.
58. Sakowski,E.G., Arora-Williams,K., Tian,F., Zayed,A.A., Zablocki,O., Sullivan,M.B. and Preheim,S.P. (2021) Interaction dynamics and virus–host range for estuarine actinophages captured by epicPCR. *Nat Microbiol*, **6**, 630–642.
59. Pilosof,S., Porter,M.A., Pascual,M. and Kéfi,S. (2017) The multilayer nature of ecological networks. *Nature Ecology & Evolution*, **1**, 1–9.
60. Barberán,A., Bates,S.T., Casamayor,E.O. and Fierer,N. (2012) Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J*, **6**, 343–351.
61. Proulx,S.R., Promislow,D.E.L. and Phillips,P.C. (2005) Network thinking in ecology and evolution. *Trends in Ecology & Evolution*, **20**, 345–353.
62. Segar,S.T., Fayle,T.M., Srivastava,D.S., Lewinsohn,T.M., Lewis,O.T., Novotny,V., Kitching,R.L. and Maunsell,S.C. (2020) The Role of Evolution in Shaping Ecological Networks. *Trends in Ecology & Evolution*, **35**, 454–466.
63. Montoya,J.M. and Solé,R.V. (2003) Topological properties of food webs: from real data to community assembly models. *Oikos*, **102**, 614–622.
64. Weitz,J.S., Poisot,T., Meyer,J.R., Flores,C.O., Valverde,S., Sullivan,M.B. and Hochberg,M.E. (2013) Phage–bacteria infection networks. *Trends in Microbiology*, **21**, 82–91.
65. Kéfi,S., Miele,V., Wieters,E.A., Navarrete,S.A. and Berlow,E.L. (2016) How Structured Is the Entangled Bank? The Surprisingly Simple Organization of Multiplex Ecological Networks Leads to Increased Persistence and Resilience. *PLOS Biology*, **14**, e1002527.
66. Allesina,S. and Levine,J.M. (2011) A competitive network theory of species diversity. *PNAS*, **108**, 5638–5642.
67. Roux,S., Camargo,A.P., Coutinho,F.H., Dabdoub,S.M., Dutilh,B.E., Nayfach,S. and Tritt,A. (2023) iPHoP: An integrated machine learning framework to maximize host

prediction for metagenome-derived viruses of archaea and bacteria. *PLOS Biology*, **21**, e3002083.

68. Bastien,E.G., Cable,R.N., Zaman,L., Batterbee,C., Wing,A.J. and Duhaime,M.B. (2023) Virus-Host Interactions Predictor (VHIP): machine learning approach to resolve microbial virus-host interaction networks. 10.1101/2023.11.03.565433.
69. Du,M.-Z., Wei,W., Qin,L., Liu,S., Zhang,A.-Y., Zhang,Y., Zhou,H. and Guo,F.-B. (2017) Co-adaption of tRNA gene copy number and amino acid usage influences translation rates in three life domains. *DNA Res*, **24**, 623–633

Chapter 4 – A Generalizable Machine Learning Framework for Identifying Genomic Markers of Antibiotic Resistance Across Bacterial Species

4.1 Abstract

As antibiotic resistance poses an escalating global health threat, the rapidly expanding repository of bacterial genomic data offers a valuable opportunity for both predicting resistance phenotypes from genomes and characterizing resistance-associated genes. To leverage this genomic information, we developed a generalizable framework that identifies genomic markers of antibiotic resistance through k-mer-based decomposition of genomes. This is made feasible by leveraging Gradient Boosting Classifiers machine learning models. Our framework successfully extracted predictive genomic signatures from genome sequence data across multiple bacterial genera and antibiotic classes, with accuracy varying among bacterial genera and correlating with underlying genomic characteristics. Functional annotation of resistance-associated k-mers demonstrated the framework's dual capability to both predict antibioitc resistance phenotype and discover genes of interest, successfully identifying well-characterized resistance determinants while revealing novel genomic sites potentially associated with resistance.

4.2 Introduction

The emergence of antibiotic resistance in bacterial populations represents one of the most pressing global health concerns of our time (1, 2). Bacteria employ diverse strategies to counteract antibiotics (3–5), which includes target site modifications through single nucleotide polymorphisms (6–8), enzymatic degradation of antibiotics (9, 10), reduced membrane permeability (11), and active efflux systems (12). These resistance mechanisms

result in genomic signatures that can be detected through sequence analysis (13). In addition, these mechanisms may also impose metabolic costs (13, 14), leading to additional detectable genomic adaptations (14–16). While specific resistance mechanisms have been elucidated for specific bacteria-antibiotic pairs (17), there remains room for improvement in developing generalizable, interpretable computational approaches that identify and leverage genomic signatures to predict the antibiotic resistance phenotype for a given genomic sequence. Such methods could enable rapid determination of antibiotic resistance phenotypes, complementing laboratory approaches, and guide experimental research by narrowing down candidate resistance mechanisms for targeted investigation (18).

Current computational tools for predicting antibiotic resistance from genomic data face several limitations that restrict their broader application. Many developed tools are constrained to specific bacterial genera or limited sets of antibiotics (19–22), reducing their generalizability across the diverse landscape of bacterial pathogens and therapeutic agents, or focuses on SNPs which can only capture a narrow range of mechanisms that lead to resistance to antibiotic (6, 23). Additionally, some tools rely heavily on curated databases of known resistance genes, which provide high accuracy for well-characterized resistance mechanisms but may struggle to identify novel resistance pathways and require computationally costly updates as new resistance mechanisms are being discovered and reported (24). Furthermore, many current machine learning approaches prioritize predictive performance over biological interpretability, effectively functioning as "black boxes" that provide resistance predictions without revealing the underlying genomic features driving those predictions (25). This lack of interpretability severely limits their utility for guiding experimental validation and mechanistic understanding of resistance development (23, 26, 27).

To address these limitations, we developed a broadly generalizable computational framework that combines machine learning techniques with interpretable feature selection to identify genomic signatures associated with antibiotic resistance. We applied this framework to multiple bacterial species and antibiotic classes. Furthermore, we successfully rediscovered well-characterized resistance mutations while also identifying

novel genomic sites potentially associated with resistance for *Escherichia* and ciprofloxacin. By maintaining biological interpretability throughout the analysis pipeline, our method provides mechanistic insights that can guide targeted experimental validation and contribute to a deeper understanding of antibiotic resistance evolution. The database-independent nature of our framework enables rapid model updates as new genomic data becomes available.

4.3 Results

4.3.1 Data preparation and k-mer matrix construction

AMR phenotypes and corresponding sequences were downloaded from the BV-BRC database. Bacteria-antibiotic pairs were selected based on data availability, with a minimum threshold of 150 genomes per antibiotic resistance phenotype to ensure there was enough data to train machine learning models. Table 1 presents the bacterial general considered in this study, along with the numbers of susceptible and resistant strains for each antibiotic. Following preprocessing steps detailed in Methods, the final dataset consisted of matrices where rows represented individual bacterial strains with known AMR phenotypes and columns represented unique k-mer words of specified length. Matrix elements were populated with binary values (0 or 1) indicating the absence or presence of each k-mer word in the corresponding genome sequence. K-mer matrices were constructed for each bacteria-antibiotic combination in Table 1.

4.3.2 Determination of optimal of k-mer length for machine learning model to yield highest accuracy

A critical parameter in this approach was the selection of optimal k-mer length, as this choice represents a tradeoff between predictive performance and computational feasibility. As k-length increases, the number of possible k-mer words that are generated grows exponentially ($4^k$where k is length used), providing a deeper genomic decomposition of sequences that could correlate with antibiotic resistance phenotype. However, this expansion creates a high-dimensional feature space that presents multiple challenges: (1)

an unfavorable feature-to-sample ratio that increases overfitting risk, (2) substantially increased computational requirements for model training, and (3) inclusion of numerous non-informative k-mers that contribute noise rather than signal for resistance prediction.

To address the high-dimensionality posed by k-mer data, we implemented two complementary strategies to minimize overfitting risks. First, we constrained model complexity by limiting the decision tree depth to a maximum of three levels, thereby restricting classification rules to combinations of at most three k-mer features. Second, we used Gradient Boosting Classifier as our primary machine learning framework, leveraging its inherent robustness with high-dimensional, sparse datasets (28), though such machine learning models take considerably more time to train than other classifier models such as Random Forest. Model training was performed using 70% of the data for training and 30% for testing. No cross-validation was performed due to the training and computational time required to train a single model. Accuracy of models are reported from its accuracy on the testing set. In addition, we computed the F1 score, AUROC score, and Matthew's Correlation score (MCC) (Fig 1 and Table 2).

To determine optimal k-mer lengths, we trained Gradient Boosting Classifiers across k-values ranging from 6 to 12 for each bacterial genus-antibiotic combination. Classification accuracy generally improved with increasing k-length, with diminishing returns between accuracy and training time (Fig 1). This suggests that longer k-mer sequences typically capture more informative genomic patterns relevant to antibiotic resistance. Interestingly, in some cases we saw decreased accuracy at high k-length. These findings demonstrate that k-mer length optimization requires empirical evaluation for each specific prediction task, balancing the trade-off between feature richness and model complexity. Interestingly, we found no significant correlation between total number of data points and accuracy achieved. However, certain genera of bacteria were easier to predict antibiotic resistance, and this correlated with genome length. For example, models accuracy for predicting antibiotic resistance for *Staphylococcus* were between 95.5% and 99.6%, whereas for Pseudomonas the accuracy was between 63.2% and 90.7% (Fig 2).

4.3.3 Functional annotation of resistance-associated k-mer words

For each machine learning model trained on a bacteria-antibiotic pair, we extracted the number of predictive k-mers identified during training. This number increased with k-length before plateauing, indicating that longer k-mers do not necessarily capture additional meaningful genomic signatures for resistance prediction (Fig 3).

To explore the biological basis of model predictions, we mapped predictive k-mers back onto the genomes. Due to time constraints, we performed this analysis only for *Escherichia coli* and ciprofloxacin as proof of concept. Bacterial genomes were annotated using BAKTA to generate comprehensive feature tables encompassing coding sequences. We aligned k-mer sequences back onto the reference genomes. We then determined whether k-mer positions overlapped with annotated features and extracted genes containing at least one predictive k-mer.

The DNA gyrase subunit A gene (gyrA) accounted for 75% of total feature importance. This is consistent with its role as the primary ciprofloxacin target and is a known site of resistance mutations (29). Additionally, yeeY, contributed approximately 20% of feature importance despite lacking known associations with antibiotic resistance, suggesting potential population structure effects or uncharacterized resistance mechanisms. The remaining genes—rsmF, parC, hybO, ebgA, yegH, and clpB—formed a long tail with minimal individual contributions. Notably, parC (topoisomerase IV subunit C), a well-established secondary fluoroquinolone target (30), appeared in this list.

4.4 Discussion

In this study, we showed that k-mer-based machine learning approaches can predict antibiotic-resistance phenotypes across bacterial genera. We implemented two strategies to mitigate the risk associated with overfitting owing to the high-dimensional nature of the k-mer approach. First, we selected the Gradient Boosting Classifier as it is more robust in handling sparse, high-dimensional datasets (28, 31, 32). Second, we limited the decision tree depth to a maximum of three levels, restricting the classification rules to combinations

of at most three k-mer words. These design choices proved useful for achieving predictive performance despite the exponential growth in the feature space with increasing k-mer length, where the number of possible k-mers scales exponentially.

In addition, our results establish a correlation between specific k-mer patterns and antibiotic resistance phenotypes across the tested bacterial–antibiotic combinations (17, 33, 34). Accuracy generally improved with increasing k-mer length, albeit with diminishing returns. This represents a tradeoff between feature richness (i.e., the number of features available during training) and model complexity (i.e., the number of features indicative of antibiotic resistance). However, the optimal k-mer length varied between 9 and 12 depending on the bacterial genus and antibiotic, suggesting that no universal k-value exists, and should be empirically determined.

Although our method demonstrated broad generalizability across multiple bacterial genera, we reported notable differences in prediction accuracy that correlated with bacterial genome characteristics. *Staphylococcus* species consistently yielded the highest prediction accuracies, while *Pseudomonas* proved most challenging to predict, and Acinetobacter and Escherichia failing in between. This pattern may reflect the differences in genomic architecture (35, 36), regulatory pathways (37, 38), and phylogenetic diversity (39). For example, *Pseudomonas* species have evolved many different antibiotic resistance mechanisms, such as complex transcriptional regulation systems (40), efflux pump systems (41, 42), and membrane composition changes (43, 44), which may be particularly difficult to capture entirely with only k-mer-based approaches. In contrast, *Staphylococcus* aureus employs resistance strategies centered on discrete, well-characterized genetic determinants (45). These distinct genetic markers create clear genomic signatures (46, 47) that can be reliably detected using a k-mer-based approach.

To validate the biological relevance and interpretability of our approach, we mapped the predicted k-mers back to their genomic contexts for *Escherichia* and ciprofloxacin. We show that our approach recovered well-established resistance determinants, while also identifying novel genomic sites potentially associated with resistance. This framework

could guide experimental efforts by prioritizing genes and/or genomic regions for targeted investigations, potentially accelerating the discovery of novel resistance mechanisms. Finally, the framework could theoretically be extended to assess the antibiotic resistance potential at the community level. By correlating k-mer words with antibiotic resistance, it may be possible to rapidly evaluate antibiotic resistance potential at the community level without requiring culturing-based methods or species-level resolution. Such an application would be particularly valuable for the surveillance of antibiotic-resistant samples.

4.5 Methods

4.5.1 Genome and antibiotic susceptibility

AMR phenotypes and corresponding sequences were downloaded from the BV-BRC dataset, a repository focusing on bacteria that pose threats to human health. The selection of bacteria and antibiotics for analysis was primarily determined by data availability, with priority given to pairs with the largest number of observations. This prioritization of larger datasets was essential for training robust machine learning models, as sufficient sample sizes are required to achieve reliable predictive performance and minimize overfitting risks.

4.5.2 Generate of k-mer matrices for machine learning

K-mer profiles were generated for each sequence within every bacterium-antibiotic combination across a range of k-lengths from 6 to 12 nucleotides. For each k-length, all possible overlapping k-mer words were extracted from the genomic sequences and compiled into binary presence/absence matrices, where rows represented individual bacterial strains and columns represented unique k-mer words. Each matrix element was populated with a binary value (1 for presence, 0 for absence) indicating whether the corresponding k-mer was found in that particular genome sequence. To reduce the dimensionality of the feature space and focus on discriminatory genomic signatures, k-mer words that were invariant across all strains within a bacteria-antibiotic pair were systematically removed. Specifically, k-mers that were either universally present in all

strains or completely absent from all strains were excluded from the analysis, as these invariant features would provide no discriminatory power for distinguishing between resistant and susceptible phenotypes during machine learning classification. This preprocessing step substantially reduced the number of features while retaining only those k-mers that varied between resistant and susceptible strains, thereby focusing the analysis on potentially informative genomic signatures associated with antibiotic resistance.

4.5.3 Machine learning model training and assessment

To address the high-dimensional nature of the data, we implemented two mitigating strategies to overfitting for our binary classification tasks. First, we restricted model complexity by only allowing shallow decision trees (tree_depth = 3), preventing spurious associations between k-mer features and antibiotic resistance phenotypes. Second, we selected gradient boosting classifiers (with the following default parameters: depth = 4, learning rate = 0.25. loss= exponential) due to their sequential error-correction approach, which builds models iteratively to minimize prediction errors and demonstrates greater resilience to overfitting compared to ensemble methods like Random Forests. Random Forests, while effective in many contexts, can struggle in high-dimensional settings where most features are uninformative. Their random feature selection approach may result in many poorly performing trees that degrade overall prediction accuracy when averaged. All machine learning models were implemented using the scikit-learn library in Python.

Due to the substantial computational requirements imposed by the large, high-dimensional k-mer datasets, we did not perform hyperparameter tuning for the gradient boosting classifiers. Instead, we used a 70/30 split for training and testing data, with model accuracy reported from the testing set. Since no hyperparameter optimization was conducted on the testing data, the reported accuracy represents true performance on unseen data, providing an unbiased estimate of model generalizability. In addition to accuracy, we computed comprehensive performance metrics including the F1 score (harmonic mean of precision and recall), Matthews Correlation Coefficient (MCC) to account for potential class imbalances, and Area Under the Receiver Operating Characteristic curve (AUROC) to

evaluate the models' ability to discriminate between resistant and susceptible phenotypes across different classification thresholds.

4.5.4 Linking machine learning model features to genes

To verify the model's true ability to detect antibiotic resistance phenotype and understand the biological basis of predictions, we sought to identify which k-mer words were the most informative and determine their genomic locations, specifically whether they mapped to known genes or regulatory regions. Due to computational constraints and time limitations, we focused this analysis on the *Escherichia* and ciprofloxacin combination, which achieved 93.4% accuracy using k-length 10, representing one of our best-performing models. Following training of the Gradient Boosting Classifier, we extracted the most informative k-mer features by ranking them according to their feature importance scores calculated by the scikit-learn implementation. Feature importance in gradient boosting reflects how frequently and effectively each k-mer was used across all decision trees in the ensemble for making classification decisions. We selected the top-ranked k-mers that cumulatively contributed to the majority of the model's predictive power for subsequent genomic mapping analysis.

To map these informative k-mer sequences back to their genomic contexts, we performed exact sequence matching against a representative set of *Escherichia* coli genomes from our dataset. Prior to mapping, all *Escherichia* coli genomes were comprehensively annotated using BAKTA, which identified coding sequences.

We then systematically assessed whether each informative k-mer's genomic positions overlapped with annotated genetic elements, categorizing associations by feature type (e.g., coding sequences, promoter regions, intergenic regions) and compiling ranked lists of genes containing resistance-predictive k-mers based on their cumulative feature importance scores.

4.6 Figures

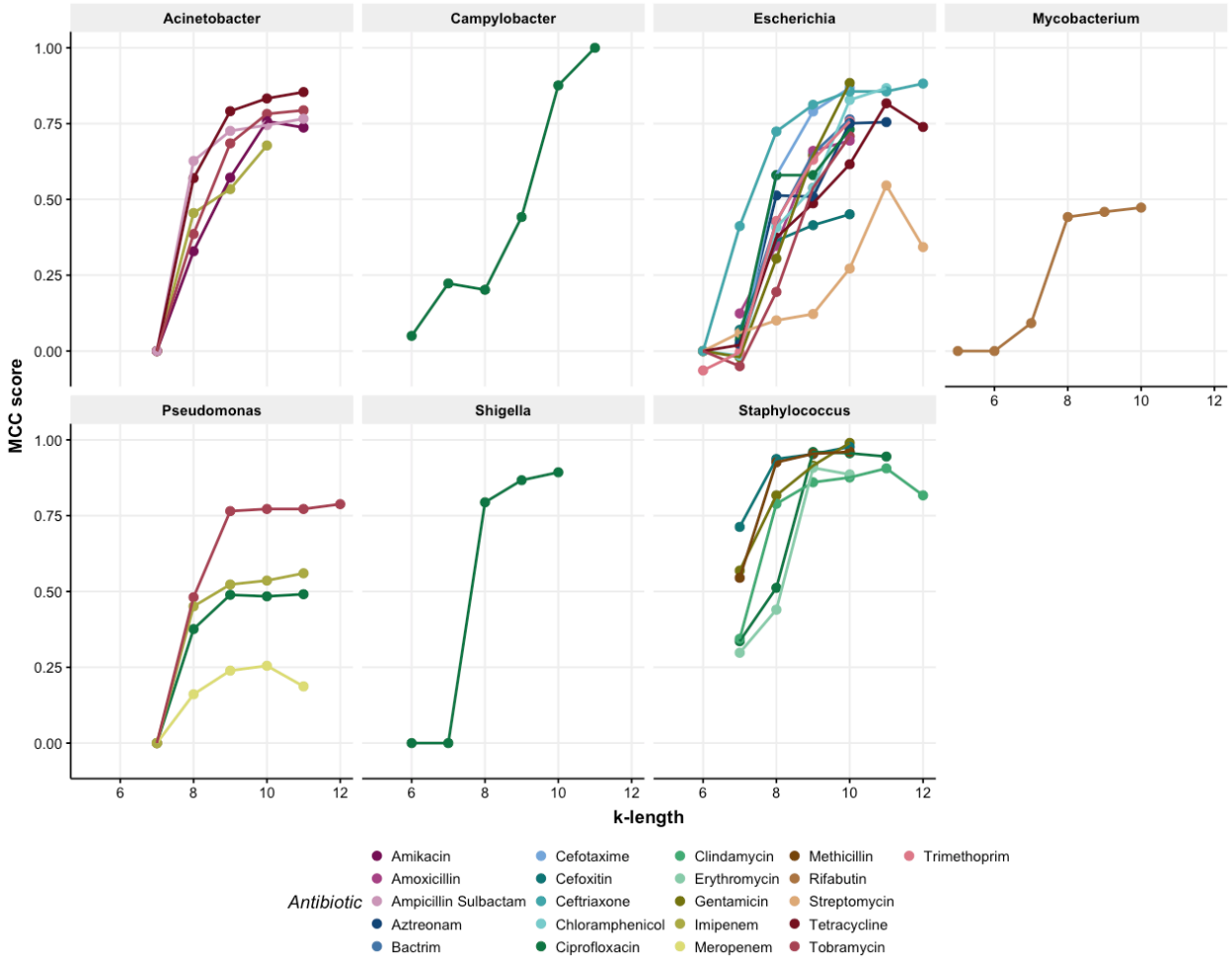


Figure 4.1. Gradient Boosting Model classifier accuracy increases with k-length used.

Note that MCC score is used rather than accuracy, as it considers all possible model predicting outcomes (i.e., true positive, true negative, false positive, and false negative) equally, which is a better representation of the model's performance for imbalanced datasets.

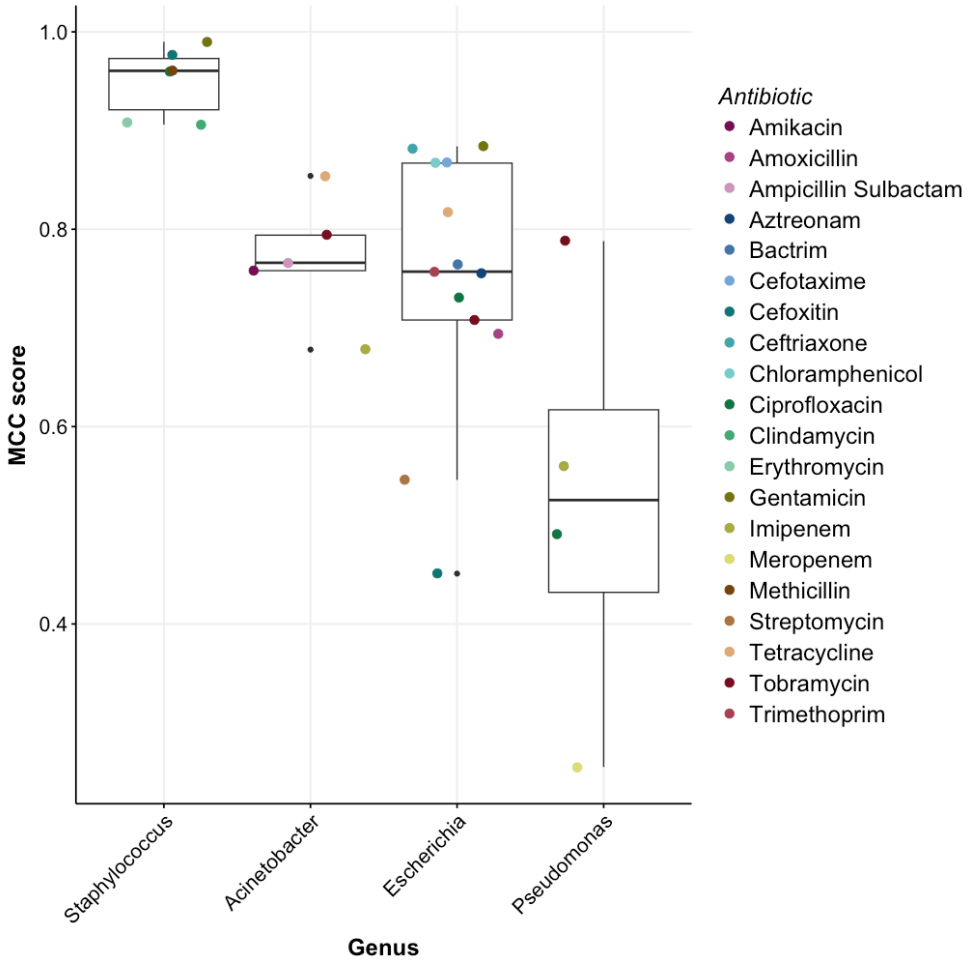


Figure 4.2. Comparison of best performing models between Staphylococcus, Acinetobacter, Escherichia, and Pseudomonas.

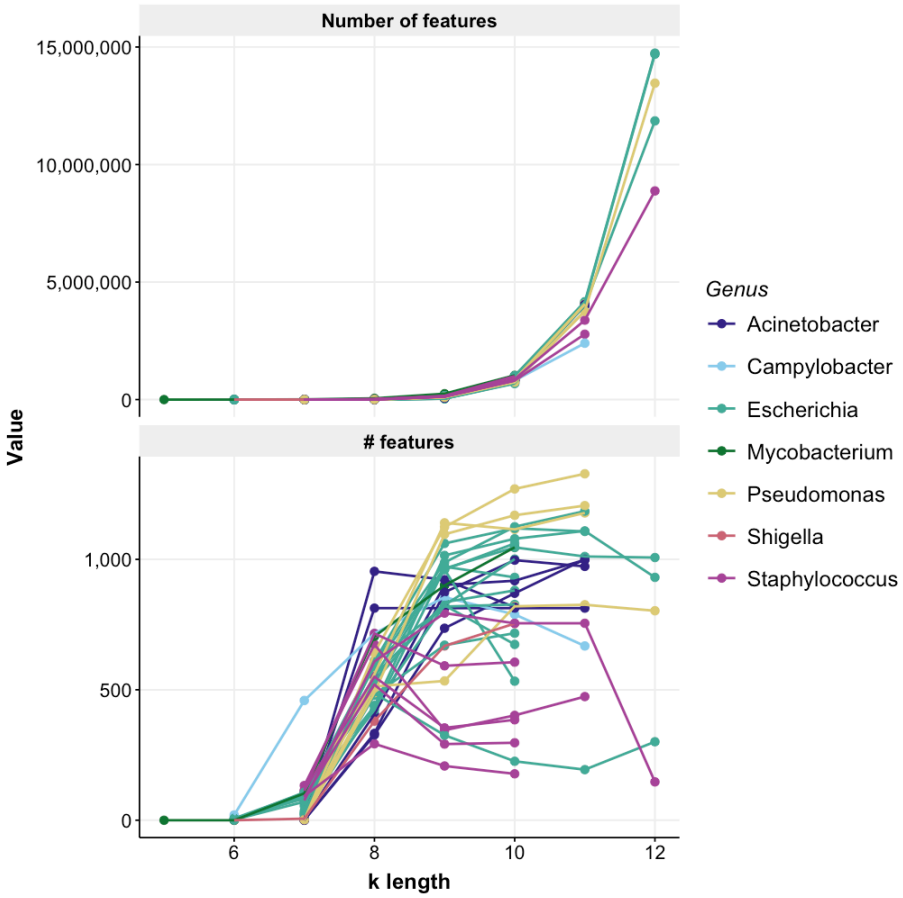
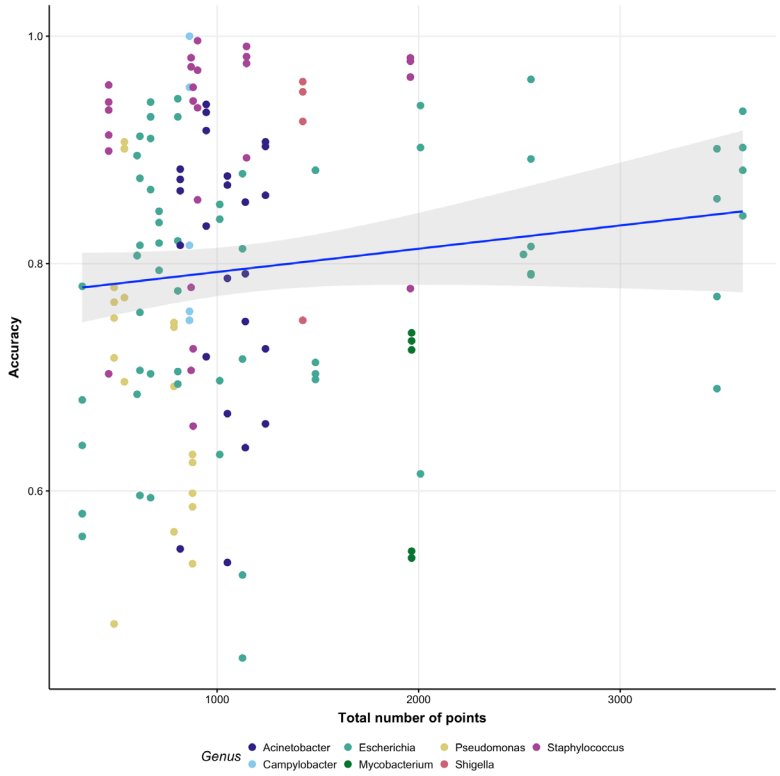


Figure 4.3. Comparison between number of features generated through k-mer word decomposition and actual number of useful features for antibiotic phenotype prediction

Top figure shows the number of features that went into the Gradient Boost Classifiers as a function of k-length. Colors represent different genus. Note that the number of features going into the model varies between genera. This is because words that were present or absent in all strains of interest were removed. Bottom figure shows the actual number of words used by the model during training. Notably, the number of features used plateaus, suggesting that k-mer words of length between 9 and 11 might encode the strongest signal regarding antibiotic resistance.

4.7 Supplementary Figures



Supplemental Figure 4.4. No correlation between the total number of points and Gradient Boosting Classifier's accuracy

4.8 Tables

Genus	Antibiotic	Laboratory method	Number of susceptible strains	Number of resistant strains
Acinetobacter	Tobramycin	Broth dilution	520	720
Acinetobacter	Amikacin	Broth dilution	589	462
Acinetobacter	Imipenem	Broth dilution	709	431
Acinetobacter	Tetracycline	Broth dilution	290	656
Acinetobacter	Ampicillin Sulbactam	Broth dilution	449	368
Campylobacter	Ciprofloxacin	Broth dilution	595	268
Escherichia	Cefotaxime	Broth dilution	1260	749
Escherichia	Tetracycline	Broth dilution	250	367
Escherichia	Bactrim	Disk diffusion	2435	1045
Escherichia	Ciprofloxacin	Disk diffusion	2990	618
Escherichia	Tobramycin	Broth dilution	1138	350
Escherichia	Cefoxitin	Broth dilution	547	165
Escherichia	Gentamicin	Broth dilution	2093	464
Escherichia	Chloramphenicol	Mixed methods	574	231
Escherichia	Amoxicillin	Agar dilution	612	401
Escherichia	Streptomycin	Broth dilution	186	145
Escherichia	Ceftriaxone	Broth dilution	349	321
Escherichia	Trimethoprim	Broth dilution	581	545
Escherichia	Aztreonam	Broth dilution	409	194
Mycobacterium	Rifabutin	MIC	1033	932

Pseudomonas	Ciprofloxacin	Mixed methods	432	354
Pseudomonas	Tobramycin	Mixed methods	349	191
Pseudomonas	Meropenem	Mixed methods	384	494
Pseudomonas	Imipenem	Mixed methods	236	253
Shigella	Ciprofloxacin	Agar dilution	1048	377
Staphylococcus	Ciprofloxacin	Disk diffusion	522	349
Staphylococcus	Gentamicin	Disk diffusion	678	225
Staphylococcus	Cefoxitin	Mixed methods	841	305
Staphylococcus	Clindamycin	Mixed methods	271	191
Staphylococcus	Methicillin	Mixed methods	1129	830
Staphylococcus	Erythromycin	Disk diffusion	516	365

Table 4.1. Bacteria genus and antibiotics considered for this study.

Genus	Antibiotic	Optimal k-length	Number of useful features	Accuracy	AUROC	F1 score	MCC score
Acinetobacter	Tobramycin	11	998	90.7%	0.956	0.907	0.794
Acinetobacter	Amikacin	10	813	87.7%	0.946	0.877	0.758
Acinetobacter	Imipenem	10	817	85.4%	0.931	0.85	0.678
Acinetobacter	Tetracycline	11	973	94.0%	0.958	0.941	0.854
Acinetobacter	Ampicillin Sulbactam	11	1000	88.3%	0.937	0.883	0.766
Campylobacter	Ciprofloxacin	11	668	100%	1	1	1
Escherichia	Cefotaxime	10	674	93.9%	0.96	0.938	0.868
Escherichia	Tetracycline	11	1109	91.2%	0.954	0.912	0.817
Escherichia	Bactrim	10	827	90.1%	0.942	0.901	0.764
Escherichia	Ciprofloxacin	10	881	93.4%	0.953	0.93	0.731
Escherichia	Tobramycin	10	931	88.2%	0.891	0.876	0.708
Escherichia	Cefoxitin	10	1001	84.6%	0.911	0.82	0.451
Escherichia	Gentamicin	10	533	96.2%	0.967	0.962	0.884
Escherichia	Chloramphenicol	11	1107	94.5%	0.979	0.945	0.867
Escherichia	Amoxicillin	10	717	85.2%	0.92	0.853	0.694
Escherichia	Streptomycin	11	194	78.0%	0.894	0.779	0.546
Escherichia	Ceftriaxone	12	1007	94.2%	0.982	0.941	0.882
Escherichia	Trimethoprim	10	1061	87.9%	0.83	0.879	0.757

Escherichia	Aztreonam	11	1185	89.5%	0.96	0.896	0.755
Mycobacterium	Rifabutin	10	1046	73.9%	0.777	0.736	0.473
Pseudomonas	Ciprofloxacin	11	1178	74.8%	0.85	0.748	0.491
Pseudomonas	Tobramycin	12	803	90.7%	0.901	0.908	0.788
Pseudomonas	Meropenem	10	1270	63.2%	0.713	0.627	0.255
Pseudomonas	Imipenem	11	1206	77.9%	0.882	0.779	0.56
Shigella	Ciprofloxacin	10	754	96.0%	0.972	0.96	0.893
Staphylococcus	Ciprofloxacin	9	345	98.1%	0.984	0.981	0.96
Staphylococcus	Gentamicin	10	178	99.6%	1	0.996	0.99
Staphylococcus	Cefoxitin	10	297	99.1%	1	0.991	0.977
Staphylococcus	Clindamycin	11	755	95.7%	0.975	0.956	0.906
Staphylococcus	Methicillin	10	385	98.1%	0.992	0.981	0.961
Staphylococcus	Erythromycin	9	592	95.5%	0.98	0.955	0.908

Table 4.2. Metrics of the best performing model for each bacteria-antibiotic combination

Note that the accuracy is determined from the test set. However, since no tuning of hyperparameters was performed, the accuracy reported is from data points that the machine learning model did not see during training.
Number of useful features is defined as k-mer words that was at least used once by the machine learning model.

4.9 References

1. Murray,C.J.L., Ikuta,K.S., Sharara,F., Swetschinski,L., Aguilar,G.R., Gray,A., Han,C., Bisignano,C., Rao,P., Wool,E., *et al.* (2022) Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*, **399**, 629–655.
2. Aljeldah,M.M. (2022) Antimicrobial Resistance and Its Spread Is a Global Threat. *Antibiotics*, **11**, 1082.
3. Darby,E.M., Trampari,E., Siasat,P., Gaya,M.S., Alav,I., Webber,M.A. and Blair,J.M.A. (2023) Molecular mechanisms of antibiotic resistance revisited. *Nat Rev Microbiol*, **21**, 280–295.
4. Blair,J.M.A., Webber,M.A., Baylay,A.J., Ogbolu,D.O. and Piddock,L.J.V. (2015) Molecular mechanisms of antibiotic resistance. *Nat Rev Microbiol*, **13**, 42–51.
5. Peterson,E. and Kaur,P. (2018) Antibiotic Resistance Mechanisms in Bacteria: Relationships Between Resistance Determinants of Antibiotic Producers, Environmental Bacteria, and Clinical Pathogens. *Front. Microbiol.*, **9**.
6. Zankari,E., Allesøe,R., Joensen,K.G., Cavaco,L.M., Lund,O. and Aarestrup,F.M. (2017) PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *J Antimicrob Chemother*, **72**, 2764–2768.
7. Knopp,M. and Andersson,D.I. (2018) Predictable Phenotypes of Antibiotic Resistance Mutations. *mBio*, **9**, 10.1128/mbio.00770-18.
8. Mosquera-Rendón,J., Moreno-Herrera,C.X., Robledo,J. and Hurtado-Páez,U. (2023) Genome-Wide Association Studies (GWAS) Approaches for the Detection of Genetic Variants Associated with Antibiotic Resistance: A Systematic Review. *Microorganisms*, **11**, 2866.
9. Bush,K. and Bradford,P.A. (2019) Interplay between β-lactamases and new β-lactamase inhibitors. *Nat Rev Microbiol*, **17**, 295–306.
10. Drawz,S.M. and Bonomo,R.A. (2010) Three Decades of β-Lactamase Inhibitors. *Clin Microbiol Rev*, **23**, 160–201.

11. Pagès,J.-M., James,C.E. and Winterhalter,M. (2008) The porin and the permeating antibiotic: a selective diffusion barrier in Gram-negative bacteria. *Nat Rev Microbiol*, **6**, 893–903.
12. Li,X.-Z., Plésiat,P. and Nikaido,H. (2015) The Challenge of Efflux-Mediated Antibiotic Resistance in Gram-Negative Bacteria. *Clin Microbiol Rev*, **28**, 337–418.
13. Melnyk,A.H., Wong,A. and Kassen,R. (2015) The fitness costs of antibiotic resistance mutations. *Evol Appl*, **8**, 273–283.
14. Rajer,F. and Sandegren,L. (2022) The Role of Antibiotic Resistance Genes in the Fitness Cost of Multiresistance Plasmids. *mBio*, **13**, e03552-21.
15. Levin,B.R., Perrot,V. and Walker,N. (2000) Compensatory Mutations, Antibiotic Resistance and the Population Genetics of Adaptive Evolution in Bacteria. *Genetics*, **154**, 985–997.
16. Pal,A. and Andersson,D.I. (2024) Bacteria can compensate the fitness costs of amplified resistance genes via a bypass mechanism. *Nat Commun*, **15**, 2333.
17. Jaillard,M., Palmieri,M., van Belkum,A. and Mahé,P. (2020) Interpreting k-mer–based signatures for antibiotic resistance prediction. *Gigascience*, **9**, giaa110.
18. Jang,H.Y., Song,J., Kim,J.H., Lee,H., Kim,I.-W., Moon,B. and Oh,J.M. (2022) Machine learning-based quantitative prediction of drug exposure in drug-drug interactions using drug label information. *NPJ Digit Med*, **5**, 88.
19. Ha,A.D., Moniruzzaman,M. and Aylward,F.O. (2023) Assessing the biogeography of marine giant viruses in four oceanic transects. *ISME COMMUN.*, **3**, 43.
20. Ren,Y., Chakraborty,T., Doijad,S., Falgenhauer,L., Falgenhauer,J., Goesmann,A., Hauschild,A.-C., Schwengers,O. and Heider,D. (2022) Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics*, **38**, 325–334.
21. Kavvas,E.S., Catoiu,E., Mih,N., Yurkovich,J.T., Seif,Y., Dillon,N., Heckmann,D., Anand,A., Yang,L., Nizet,V., *et al.* (2018) Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. *Nat Commun*, **9**, 4306.
22. Avershina,E., Sharma,P., Taxt,A.M., Singh,H., Frye,S.A., Paul,K., Kapil,A., Naseer,U., Kaur,P. and Ahmad,R. (2021) AMR-Diag: Neural network based genotype-to-phenotype

prediction of resistance towards β-lactams in *Escherichia coli* and *Klebsiella pneumoniae*. *Computational and Structural Biotechnology Journal*, **19**, 1896–1906.

23. Sakagianni,A., Koufopoulou,C., Feretzakis,G., Kalles,D., Verykios,V.S., Myrianthefs,P. and Fildisis,G. (2023) Using Machine Learning to Predict Antimicrobial Resistance—A Literature Review. *Antibiotics*, **12**, 452.
24. Arango-Argoty,G., Garner,E., Pruden,A., Heath,L.S., Vikesland,P. and Zhang,L. (2018) DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, **6**, 23.
25. Macesic,N., Polubriaginof,F. and Tatonetti,N.P. (2017) Machine learning: novel bioinformatics approaches for combating antimicrobial resistance. *Current Opinion in Infectious Diseases*, **30**, 511.
26. Giacobbe,D.R., Marelli,C., Guastavino,S., Mora,S., Rosso,N., Signori,A., Campi,C., Giacomini,M. and Bassetti,M. (2024) Explainable and Interpretable Machine Learning for Antimicrobial Stewardship: Opportunities and Challenges. *Clinical Therapeutics*, **46**, 474–480.
27. Kim,J.I., Maguire,F., Tsang,K.K., Gouliouris,T., Peacock,S.J., McAllister,T.A., McArthur,A.G. and Beiko,R.G. Machine Learning for Antimicrobial Resistance Prediction: Current Practice, Limitations, and Clinical Perspective. *Clin Microbiol Rev*, **35**, e00179-21.
28. Boldini,D., Grisoni,F., Kuhn,D., Friedrich,L. and Sieber,S.A. (2023) Practical guidelines for the use of gradient boosting for molecular property prediction. *Journal of Cheminformatics*, **15**, 73.
29. Weigel,L.M., Steward,C.D. and Tenover,F.C. (1998) gyrA mutations associated with fluoroquinolone resistance in eight species of Enterobacteriaceae. *Antimicrob Agents Chemother*, **42**, 2661–2667.
30. Vila,J., Ruiz,J., Goñi,P. and De Anta,M.T. (1996) Detection of mutations in parC in quinolone-resistant clinical isolates of Escherichia coli. *Antimicrob Agents Chemother*, **40**, 491–493.
31. Blagus,R. and Lusa,L. (2015) Boosting for high-dimensional two-class prediction. *BMC Bioinformatics*, **16**, 300.

32. Moradigaravand,D., Palm,M., Farewell,A., Mustonen,V., Warringer,J. and Parts,L. (2018) Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. *PLOS Computational Biology*, **14**, e1006258.
33. Davis,J.J., Boisvert,S., Brettin,T., Kenyon,R.W., Mao,C., Olson,R., Overbeek,R., Santerre,J., Shukla,M., Wattam,A.R., *et al.* (2016) Antimicrobial Resistance Prediction in PATRIC and RAST. *Sci Rep*, **6**, 27930.
34. Břinda,K., Callendrello,A., Ma,K.C., MacFadden,D.R., Charalampous,T., Lee,R.S., Cowley,L., Wadsworth,C.B., Grad,Y.H., Kucherov,G., *et al.* (2020) Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. *Nat Microbiol*, **5**, 455–464.
35. Yakkala,H., Samantarrai,D., Gribskov,M. and Siddavattam,D. (2019) Comparative genome analysis reveals niche-specific genome expansion in Acinetobacter baumannii strains. *PLOS ONE*, **14**, e0218204.
36. Saati-Santamaría,Z., Baroncelli,R., Rivas,R. and García-Fraile,P. (2022) Comparative Genomics of the Genus Pseudomonas Reveals Host- and Environment-Specific Evolution. *Microbiology Spectrum*, **10**, e02370-22.
37. Santajit,S. and Indrawattana,N. (2016) Mechanisms of Antimicrobial Resistance in ESKAPE Pathogens. *Biomed Res Int*, **2016**, 2475067.
38. Tierney,A.R. and Rather,P.N. Roles of two-component regulatory systems in antibiotic resistance. *Future Microbiol*, **14**, 533–552.
39. Yurtseven,A., Buyanova,S., Agrawal,A.A., Bochkareva,O.O. and Kalinina,O.V. (2023) Machine learning and phylogenetic analysis allow for predicting antibiotic resistance in M. tuberculosis. *BMC Microbiology*, **23**, 404.
40. Huang,H., Shao,X., Xie,Y., Wang,T., Zhang,Y., Wang,X. and Deng,X. (2019) An integrated genomic regulatory network of virulence-related transcriptional factors in Pseudomonas aeruginosa. *Nat Commun*, **10**, 2931.
41. Lorusso,A.B., Carrara,J.A., Barroso,C.D.N., Tuon,F.F. and Faoro,H. (2022) Role of Efflux Pumps on Antimicrobial Resistance in Pseudomonas aeruginosa. *Int J Mol Sci*, **23**, 15779.

42. Masuda,N., Sakagawa,E., Ohya,S., Gotoh,N., Tsujimoto,H. and Nishino,T. (2000) Substrate Specificities of MexAB-OprM, MexCD-OprJ, and MexXY-OprM Efflux Pumps in Pseudomonas aeruginosa. *Antimicrob Agents Chemother*, **44**, 3322–3327.
43. Huszczynski,S.M., Lam,J.S. and Khursigara,C.M. (2019) The Role of Pseudomonas aeruginosa Lipopolysaccharide in Bacterial Pathogenesis and Physiology. *Pathogens*, **9**, 6.
44. Guest,R.L., Lee,M.J., Wang,W. and Silhavy,T.J. (2023) A periplasmic phospholipase that maintains outer membrane lipid asymmetry in Pseudomonas aeruginosa. *Proceedings of the National Academy of Sciences*, **120**, e2302546120.
45. Miragaia,M. (2018) Factors Contributing to the Evolution of mecA-Mediated β-lactam Resistance in Staphylococci: Update and New Insights From Whole Genome Sequencing (WGS). *Front Microbiol*, **9**, 2723.
46. Gordon,N.C., Price,J.R., Cole,K., Everitt,R., Morgan,M., Finney,J., Kearns,A.M., Pichon,B., Young,B., Wilson,D.J., *et al.* (2014) Prediction of Staphylococcus aureus Antimicrobial Resistance by Whole-Genome Sequencing. *J Clin Microbiol*, **52**, 1182–1191.
47. Bradley,P., Gordon,N.C., Walker,T.M., Dunn,L., Heys,S., Huang,B., Earle,S., Pankhurst,L.J., Anson,L., de Cesare,M., *et al.* (2015) Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. *Nat Commun*, **6**, 10063.

Chapter 5 – Structural Properties of Virus-Host Interaction Networks Reveal Methodological Biases and Biological Patterns[5]

5.1 Abstract

Viruses are central to microbial ecology and evolution, influencing global biogeochemical cycles through lysis of marine bacteria and driving microbial evolution through horizontal gene transfer. To better understand the impact viruses exert on microbial communities, we can consider the bipartite networks that represent virus-host interactions. However, generating virus-host interaction networks presents significant methodological challenges, as different approaches introduce different biases that have the potential to affect the network's architecture. While laboratory-resolved infection assays provide definitive proof of infection, they are inherently limited to cultivable viruses and hosts, capturing only a fraction of microbial diversity. Hi-C proximity ligation offers cultivation-independent detection of physical associations between viral and bacterial DNA, potentially capturing current infections but is unable to distinguish true infections from other DNA co-associations. Finally, computational predictions based on genomic signatures leverage signals of viral adaptation to hosts. This allows the analysis of complete communities but with varying accuracy that may reflect evolutionary history rather than current infection potential. In this study, we characterized the network architecture of 11 virus-host interaction networks generated using these three methodological approaches, examining both whole-network properties (nestedness, modularity, diameter, and density) and node-level centrality measurements. Despite methodological differences, we found consistent

[5] This chapter contains work conducted in collaboration with Emmalyn Campau, Luis Zaman, and Melissa Duhaime. A revised version will be prepared for publication.

architectural features across networks, including small diameters (4-6 steps), short average path lengths (1.8-3.4), and significant nestedness that exceeded null model expectations. The consistently low betweenness centrality of viruses across all methods suggests extensive host range overlap and functional redundancy in viral communities, while the divergent connectivity patterns highlight how each method captures different aspects of virus-host interaction biology. These findings demonstrate that while each methodological approach introduces specific biases that influence network topology, the structural patterns across methods reveal fundamental organizational principles of virus-host communities, suggesting that integrating multiple approaches will be essential for advancing our understanding of viral ecology and the role of viruses in ecosystem function.

5.2 Introduction

Contemporary metagenomics studies identify more novel viral sequences in single environmental samples than were characterized throughout the 20th century (1–3), revealing the diversity and ubiquity of viruses (4–6). Viruses influence global biogeochemical cycles through lysis of 20-40% of marine bacteria daily (7, 8), releasing cellular contents into microbial food webs (9, 10). Through horizontal gene transfer, viruses contribute to the mosaic nature of bacterial genomes and drive microbial evolution (7, 11). During infection, virus-host encounters produce virocells—entities in which viruses reprogram bacterial metabolism for viral replication (12–14). These cellular-level processes scale to influence ecosystem function, yet the mechanisms connecting individual infections to community patterns remain poorly defined due to our limited ability to identify virus-host interactions. This knowledge gap has historically relegated virus-host interactions to a black box in ecosystem models, where community-scale impacts could be measured but the underlying interaction mechanisms remained opaque (15, 16). Recent cultivation-independent methods, including Hi-C proximity ligation (17–20) and computational predictions (21–28), have enabled identification of virus-host pairs beyond cultivation constraints. However, frameworks for understanding how these paired interactions scales to community structure and function remain underdeveloped--a gap

that is notable given the successful application of network approaches in other areas of ecology.

In food webs, network topology influences stability through weak interactions and modular organization that limits perturbation spread (29–31). Plant-pollinator networks display nested structures where generalist species support specialist interactions, contributing to community persistence (32, 33). Mycorrhizal networks exhibit hub-based architectures that facilitate resource distribution among plants (34, 35). These examples show that the pattern of species interactions can predict how ecosystems respond to change. Such frameworks for virus-host interaction networks could reveal how perturbations propagate through interactions, map the possible routes of horizontal gene transfer between bacterial populations, and potentially determine whether network structure predicts ecosystem function better than viral or host identity alone, paving the way for more accurate incorporation of viruses into ecosystem models (36, 37). However, before such applications can be realized, we must first characterize the structural properties of virus-host interaction networks.

However, current understanding of virus-host network topology is limited, and different methods for resolving virus-host interactions networks may yield different network structures. Previous work has examined strain-level patterns using infection assays, but these methods can only capture a small fraction of environmental diversity due to the "uncultured majority" (38, 39). Moreover, it is unclear whether strain-level patterns extend to community-level networks. Another approach uses Hi-C proximity ligation to detect physical associations between viral and bacterial DNA without cultivation (40, 41), though it cannot readily distinguish infection from other DNA co-associations (42, 43). Computational methods use diverse genomic signatures to predict interactions across uncultivated diversity with varying accuracy (23, 24, 44, 45). Each method introduces specific biases (46) that may shape the resulting network topology (47–49).

To address this knowledge gap, this study characterizes virus-host network architecture obtained through three different approaches for resolving these interactions: laboratory

infection assays, Hi-C proximity ligation, and computational predictions. We examined whole-network properties (nestedness, modularity, diameter, density) and measurements of node centralities of viruses and hosts. By characterizing the structural properties of networks generated through each method, we aim to establish expectations for virus-host network analysis and explore how methodological inherent biases may influence these patterns.

5.3 Results

5.3.1 Data selection and virus-host interactions networks visualization from different methods

We generated 11 virus-host interaction networks from three distinct methodological approaches (Table 1). Laboratory-resolved networks are constructed from infection assays where viruses and hosts are mixed together and plaque formation indicates successful infection. While this approach provides definitive proof of infection, it is inherently limited to cultivable viruses and hosts (50–52). We analyzed two laboratory-resolved datasets that are publicly available: the Nahant collection (53, 54), which focused primarily on Vibrio species and their viruses, and a Staphylococcus study that examined Staphylococcus and their viruses isolated from wastewater samples (55).

Hi-C proximity ligation provides an alternative cultivation-independent approach by capturing physical associations between viral and bacterial DNA before sequencing (17, 18, 56). While Hi-C has been used to capture virus-host interaction in various contexts (19, 56, 57), we focused our analysis on one Hi-C dataset from hydrothermal vent samples (17) because it was the only study that provided publicly available processed virus-host interaction data, rather than raw sequencing reads alone.

The third approach utilizes computational predictions from metagenomic data of viruses and hosts using Virus-Host Interaction Predictor 1.0 (VHIP) (58). VHIP is a machine learning model that achieves 87.8% accuracy using genome-wide features (i.e., genomic

patterns similarities between viruses and hosts) and considers all possible virus-host pairs. We analyzed three VHIP-predicted studies: samples from the Atacama Desert (59), wetland ecosystems (60), and a harmful algal bloom (HAB) time series dataset spanning July to October 2024 (61, 62). For each prediction made, VHIP also reports confidence scores. The confidence score distributions for these predictions showed bimodal patterns, with most values concentrated at the extremities (0 and 1), indicating high confidence in predictions of either non-infection or infection respectively (Fig S5). Figure 1 visualizes all networks as adjacency matrices, revealing visually distinct structural patterns across methodologies and environments.

5.3.2 Virus-host interaction networks exhibit varied structures both through methodology and environment

5.3.2.1 Network-level architecture

The diameter of the largest connected component, which measures the maximum shortest path between two nodes in the network, remained consistently small across all networks (Table 2), ranging from 4 to 6 steps. Similarly, the average shortest path length showed limited variation, spanning 1.8 to 3.4, with VHIP-predicted networks showing values at the lower end of this range. Network density, calculated as the proportion of realized or predicted edges relative to all possible edges, varied across methodologies. Values ranged from as low as 6.11% Nahant Collection to 61.8% for the VHIP-predicted Atacama network. Nestedness quantifies the degree to which specialist viruses interact with subsets of the hosts that generalists interact with. To calculate nestedness, we applied the NODF algorithm (63), where the algorithm's output values range from 0 to 100. 0 represents a completely non-nested structure and 100 indicates perfect nestedness. Most networks had moderate to high nestedness, ranging from 55 to 85.5, except for the Nahant collection which had remarkably low nestedness at 9.2. To compute modularity, we used the Infomap algorithm (64, 65), which identified communities by optimizing the map equation that measures information flow through the network. Only two networks contained more than one module: the Nahant collection and the Hi-C dataset (Table 2 and Fig S6). Notably, we

found no significant correlations between these network-level metrics and network size or virus-to-host ratio (Fig S7 and Fig S8).

5.3.2.2 Node-level centralities

To describe the centralities of individual nodes within the bipartite nature of virus-host networks, we computed three different centrality metrics separately for viral and host nodes. Betweenness centrality quantifies how often a node lies on the shortest path between each other nodes. For the virus-host interaction networks, high betweenness centrality may highlight a node's importance connecting otherwise separated parts of the interaction network (66). Closeness centrality measures the average distance from a node to all other nodes, with higher values indicating more central positions that could facilitate rapid spread of perturbations or genetic material (67). Finally, eigenvector centrality assigns importance based on connections to other important nodes (i.e., other nodes that are well connected too).

Analysis of these centrality metrics (Fig 3) revealed consistent patterns across studies for betweenness centrality, with notably low values for viruses. This suggests that multiple paths connect hosts through shared viruses, indicating viral host range overlap that reduces the betweenness centrality of individual viruses. In contrast, some hosts values are higher, potentially identifying them as infection hotspots susceptible to diverse viral populations. Closeness centrality showed similar distributions for viruses and hosts in VHIP-predicted networks, while the Nahant collection displayed unusually low host closeness values, possibly reflecting the specialized nature of laboratory-cultivable *Vibrio*-phage interactions. Finally, eigenvector centrality varied considerably across networks.

5.3.3 Null model analysis to determine significance of VHIN's structure

To determine whether observed network properties arose from structural organization rather than the result of random chance, we generated random networks for each dataset.

We generated 1,000 null networks for each network while maintaining the same number of infections but randomly reassigning them among all possible virus-host pairs.

Comparison of the empirical networks to their null distributions revealed that observed nestedness consistently exceeded null expectations across all networks ($p < 0.001$). This finding indicates that the nested structure of virus-host interactions cannot be explained solely by the number of infection events, suggesting that ecological, evolutionary, or other processes shape host range distributions.

In addition, we compared centrality distributions between empirical networks and null models using Kolmogorov-Smirnov tests. All centrality measurements showed statistically significant differences from null expectations for both viral and host nodes ($p <0.0001$ for all comparisons). These results strongly suggest that the position and importance of individual nodes within virus-host networks reflect non-random processes, possibly arising from variation in host susceptibility, viral infectivity, and coevolutionary dynamics (68, 69).

5.4 Discussion

The structure of virus-host interaction networks provides insights that extend beyond individual virus-host pairs, revealing patterns of co-evolution and ecological dynamics at the community level. Our analysis of 11 networks across the three methodological approaches revealed consistent architectural features: small network diameters (i.e., 4-6), short average path lengths (i.e., 1.8-3.4), and significant nestedness. This suggests that virus-host communities share fundamental organizational principles despite the vast differences in sampling methods and environments. These network architectures capture the emergent properties that remain hidden when examining isolated interactions (70, 71). The consistently low betweenness centrality observed for viruses across all networks indicates extensive host range overlap, where multiple viruses can infect the same hosts, which could facilitate horizontal gene transfer (67, 72–74). Network topology reflects evolutionary pressure, ecological processes, and the flow of energy and genetic material

through microbial communities (75–77). The observed patterns likely depend on nutrient availability and other environmental factors that shape both viral infection strategies and host susceptibility (78–81), but this remains beyond the scope of this study. However, interpreting these structures requires careful consideration of the methodological biases inherent in each approach used to construct virus-host networks (Fig 4).

Laboratory-based infection assays have long served as the gold standard for confirming virus-host interactions, yet they capture only a small fraction of environmental viral diversity owing to the great uncultured majority (82, 83). The rate of novel virus discovery through metagenomics far outperforms our ability to test interactions experimentally, leaving the vast majority of virus-host pairs unknown (84, 85). This raises a fundamental question: do the network properties observed in laboratory-resolved subsets—particularly the exceptionally low nestedness and high modularity of the Nahant collection—reflect specialized coevolutionary dynamics or are they artifacts of cultivation bias? Laboratory methods may fail to detect infections that require specific environmental conditions that are absent in laboratory settings (86–90). Complex infection mechanisms, such as those requiring intermediate hosts or specific community contexts (87, 91, 92), cannot be easily tested using plaque assays. Furthermore, viral isolation procedures may eliminate mutation variants crucial for infection (93–95), particularly given the lock-and-key co-evolutionary dynamics between viruses and their hosts (96, 97). The structural differences we observed between laboratory-resolved networks and those from cultivation-independent methods, particularly the 10-fold difference in network density, suggest that laboratory-resolved networks may underestimate both the connectivity of virus-host systems.

Hi-C proximity ligation offers a cultivation-independent alternative that captures the physical associations between viral and bacterial DNA (98), potentially recovering a broader range of interactions. The Hi-C network from hydrothermal vents exhibited intermediate properties between laboratory and computational methods, with moderate density and the presence of multiple modules, suggesting ecological compartmentalization. Notably, Hi-C networks exhibit high connectivity with viral host ranges that are

substantially larger than expected from laboratory studies. This expanded host range could reflect genuine biological phenomena or methodological artifacts. This is because Hi-C cannot distinguish true infections from other mechanisms that generate DNA proximity, such as viral DNA within bacterial prey without infection, environmental DNA associations, or viruses attached to cell surfaces without successful infection (43). The strength of the method in capturing cultivation-resistant interactions must be weighed against the uncertainty in the biological meaning of the detected associations.

Computational predictions based on genomic signatures provide the broadest coverage of viral diversity, but introduce different biases and an error rate to consider. Our VHIP-predicted networks consistently showed the highest density (up to 61.8% for Atacama) and strong nestedness, patterns that persisted across diverse environments, from deserts to lakes and wetlands. These methods rely on signals of virus-host coevolution, which may create a temporal disconnect between evolutionary history and the current infection capability. The bimodal distribution of VHIP confidence scores we observed, concentrated at 0 and 1, suggests that the machine learning model identifies clear genomic signatures of either interaction or non-interaction, yet these signatures may reflect evolutionary history rather than the current infection potential. A virus may retain genomic signatures of past host adaptation after losing its ability to infect the host (99). Conversely, recently established virus-host pairs may lack detectable co-evolutionary signals despite active infection (68). The consistently shorter path lengths in VHIP networks compared with other methods could result from over connection due to persistent evolutionary signals or incorrect predictions, creating networks that appear more connected than current infection patterns.

The significant nestedness observed across nearly all networks, ranging from 55 to 85.5 except for the Nahant collection, deserves particular attention. Our null model analysis confirmed that this pattern cannot arise from random edge placement alone, suggesting that fundamental ecological or evolutionary processes shape virus-host communities toward nested architectures, in agreement with previous studies (100, 101). In nested networks, specialist viruses infect subsets of hosts infected by generalists, which can

enhance community stability and facilitate viral coexistence. The anomalously low nestedness of the Nahant collection may reflect the specialized nature of cultivable *Vibrio* viruses or indicate that laboratory cultivation selects against the generalist strategies that drive nestedness in natural communities.

Future work should consider how each step in the bioinformatics pipeline, from identifying, defining, and assembling viral and host species to delineating microbial populations, affects the genomic signatures used for computational predictions. Small decisions in sequence clustering, genome completeness thresholds, or population boundaries can cascade into significantly different genome assemblies (102), which subsequently may affect predictions based on genome features. Similarly, the choice of null models profoundly influenced our interpretation of the network structure (103). While this study employed a rather simple null model maintaining constant edge density, alternative models preserving degree distribution could reveal different aspects of network organization, such as whether the observed patterns stem from variation in host susceptibility or viral host range.

Finally, the divergent host range estimates across the methods warrant particular attention. Our results showed that laboratory methods suggest relatively specialized viruses with low connectivity, Hi-C reveals broader association patterns with intermediate density, and computational predictions often indicate extensive potential host ranges with network densities exceeding 60%. Rather than viewing these as incompatible and contradictory and keeping in mind that these networks cannot be easily compared because of the scope and intent of each study, each method may capture different aspects of virus-host interaction biology. Laboratory methods measure successful productive infection under artificial conditions; Hi-C detects physical proximity that may include true infections, abortive or failed infections, or other associations; and computational methods reveal evolutionary potential (past, current, and possibly future) for interaction. The consistent finding of low viral betweenness centrality across all methods suggests that regardless of the detection approach, viral communities exhibit substantial functional redundancy, with multiple viruses capable of infecting or predicted to infect the same host populations.

Despite their individual limitations, combining multiple approaches to construct virus-host networks offers the potential to get closer to biological truth. The network structure itself may capture community-level properties (i.e., stability, resilience, and evolutionary potential) that does not rely on perfect accuracy of the individual edges. The patterns revealed across methods, even with their biases, provide a new perspective for studying viral ecology. Moving forward, explicitly acknowledging and accounting for methodological biases will be crucial for interpreting network patterns into an ecological understanding of virus-host interaction networks.

5.5 Methods

5.5.1 Data acquisition

5.5.1.1 Laboratory-resolved interactions

Experimentally validated virus-host interactions from the Nahant Collection and Staphylococcus datasets were obtained from the previously compiled VHRnet dataset. These represent direct laboratory confirmations of viral infectivity against specific bacterial hosts.

5.5.1.2 Hi-C linkage

Virus-host linkages determined through Hi-C proximity ligation were obtained from a hydrothermal vent study (17). While Hi-C generates quantitative linkage values representing the frequency of DNA-DNA interactions between viruses and hosts, we converted these to binary infection matrices (1 = interaction present, 0 = interaction absent) to be consistent with the other datasets.

5.5.1.3 Computational predictions datasets

Three datasets were used for predictive methods:

- The Atacama Desert dataset (59), which included 84 viruses' genomes and 73 hosts genomes
- The Wetland dataset (60), which included 2036 viruses' genomes and 26 hosts genomes
- The Harmful Algal Bloom (HAB) dataset (61), for which the virus-host associations predictions were made part of a previous study.

5.5.2 Virus-host interaction predictions

We applied VHIP version 1.0 (58) to predict virus-host interactions for each dataset independently. VHIP is a machine learning model that evaluates all possible virus-host pairs within a dataset by calculating features representing viral adaptations to specific hosts. All predictions were performed using default VHIP parameters. A total of 6,132 predictions were made for the Atacama Desert dataset, and 52,936 predictions were made for the Wetland dataset. VHIP also returns score values for each prediction.

5.5.3 Network analysis

We characterized each virus-host network using four key metrics. Density was calculated as the proportion of realized interactions out of all possible virus-host pairs. Diameter measured the maximum shortest path length between any two connected nodes, computed using NetworkX (v. 3.5) for the largest connected component. Modularity was assessed using the Infomap (v. 2.8.0) algorithm (64), which detects community structure by optimizing the map equation to identify groups of densely interconnected nodes. Finally, nestedness was quantified using the NODF (Nestedness based on Overlap and Decreasing Fill) metric (63), which measures the degree to which specialist interactions form subsets of generalist interactions.

For centrality measurements, given the bipartite nature of our networks, the measurements were calculated separately for viral and host nodes. Betweenness centrality quantifies how often a node lies on shortest paths between other nodes. Closeness

centrality measures the inverse of average shortest path length to all other nodes. Eigenvector centrality reflects both the number and importance of neighboring nodes. All centrality calculations were performed using NetworkX (v. 3.5) functions in Python 3.13 (betweenness_centrality, closeness_centrality, eigenvector_centrality).

All network analysis code is available at:
https://github.com/DuhaimeLab/VirusHostNetworkAnalysis

5.5.4 Null model analysis

To assess statistical significance of observed network properties, we generated null models by randomly redistributing interactions while preserving network density. For each empirical network, we randomly reassigned the positions of interactions (1s) within the bipartite matrix, maintaining the total number of interactions but randomizing their distribution among virus-host pairs. Network properties were recalculated for each randomized network. This process was repeated 1,000 times to generate null distributions for comparison with empirical values.

5.6 Figures

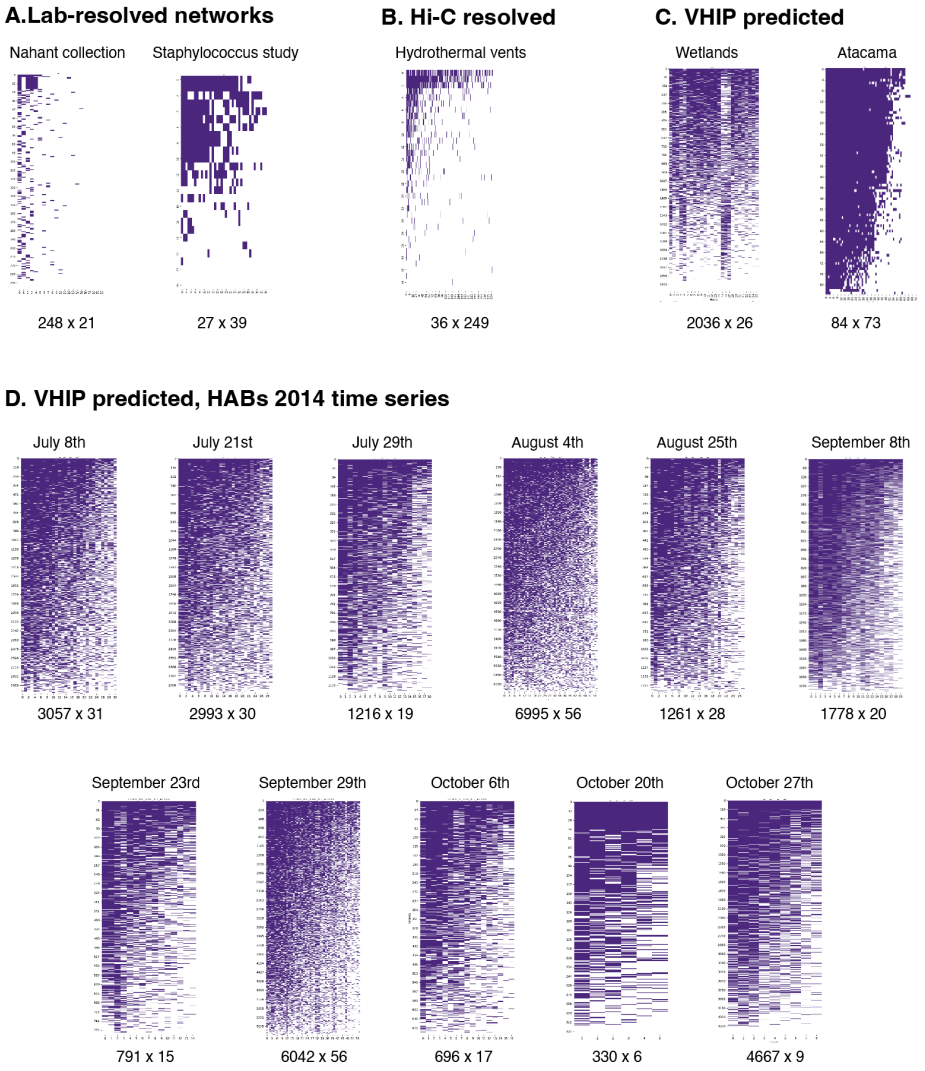


Figure 5.1. Visualization of the virus-host interaction networks

For each network, each row represents a virus, and each column represents a host. The numbers below each visualization indicate the total number of viruses and hosts in each network. Rows are arranged from generalist viruses (top) to specialist viruses (bottom), while columns are ordered from most susceptible hosts (left) to most resistant hosts (right).

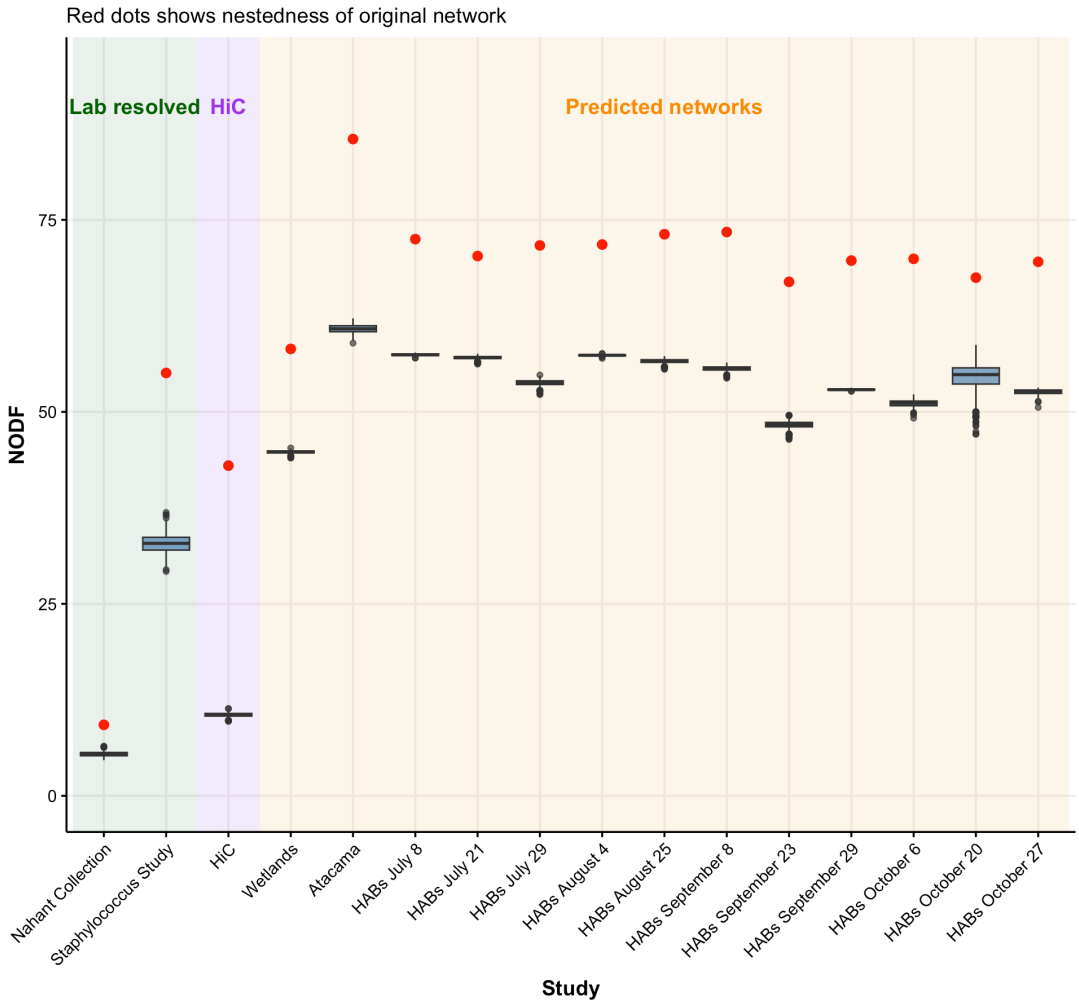


Figure 5.2. Nestedness of the virus-host interaction networks

Background color represents the different methods used to construct the virus-host interaction networks. The NODF algorithm was used to calculate nestedness. Red dots show nestedness of original network, whereas the boxplot shows the distribution of nestedness of the null models. Note that VHIP-predicted are considerably higher than lab-resolved or Hi-C.

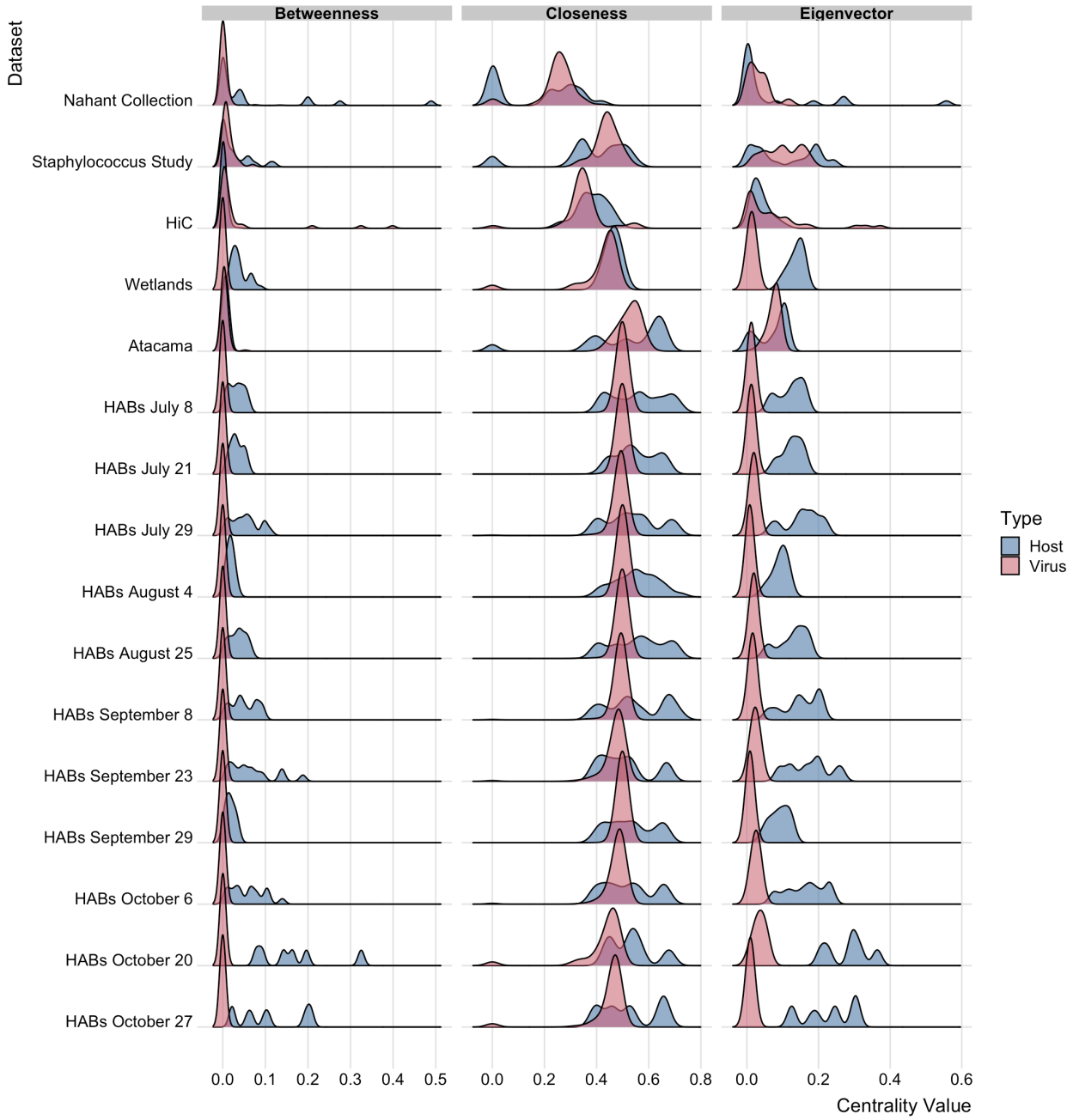


Figure 5.3. Centrality measurements of virus and host nodes

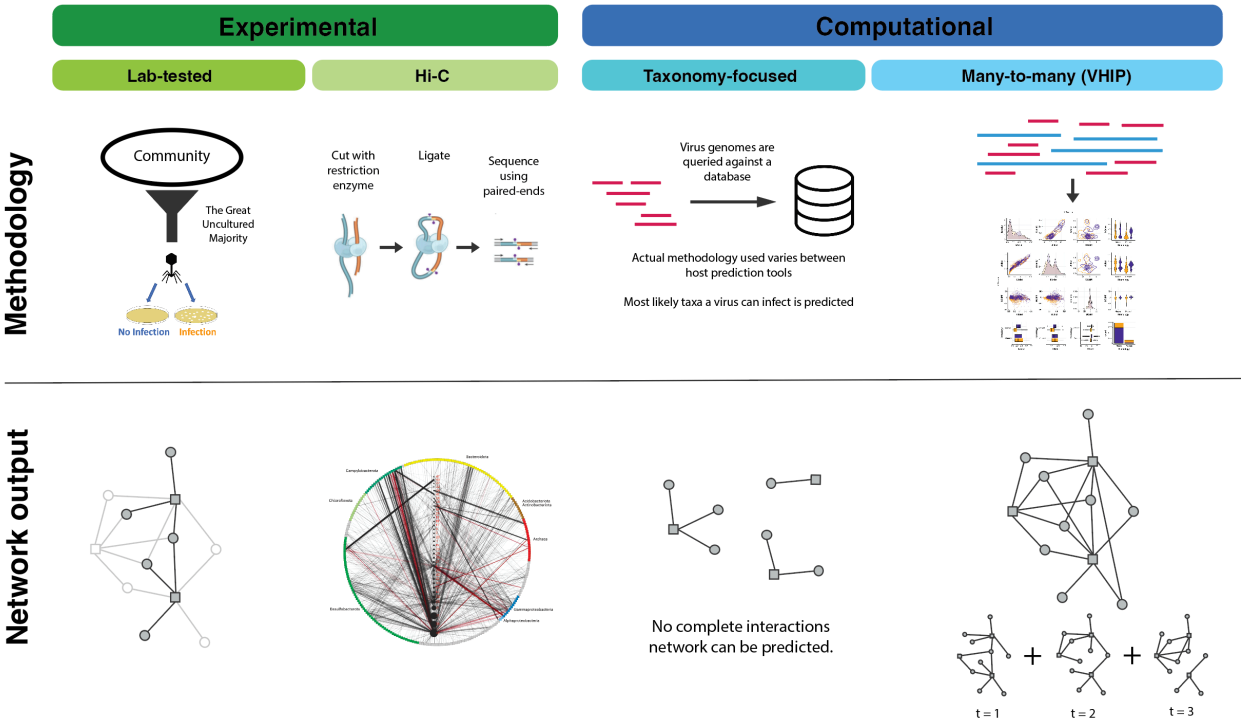
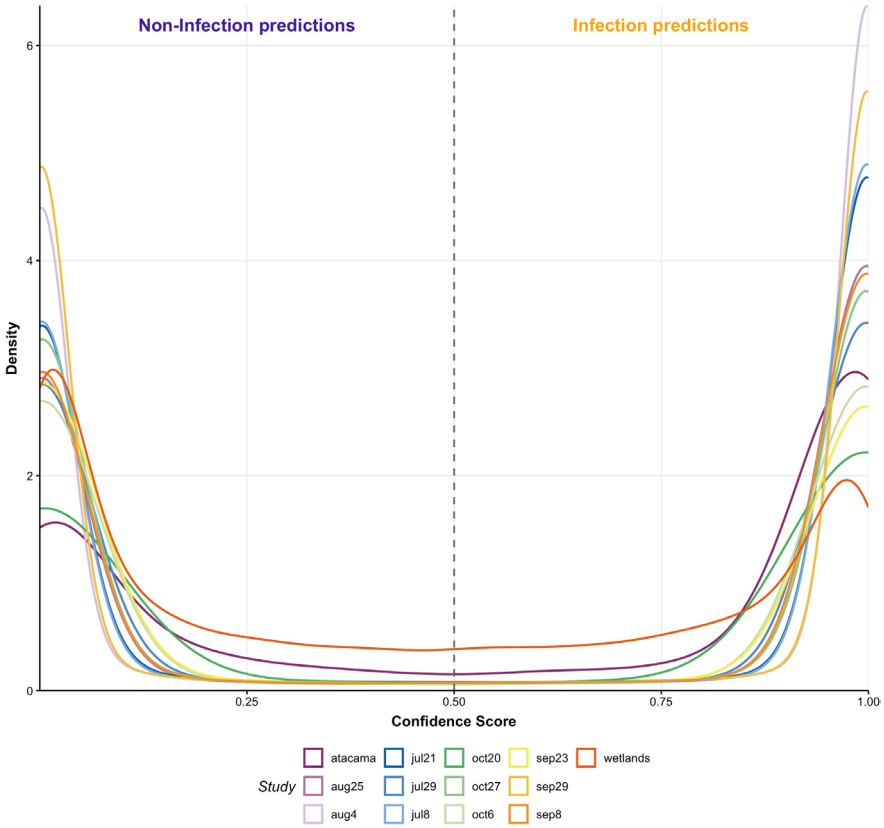


Figure 5.4. Conceptual diagram of the methods to reconstruct virus-host interaction network

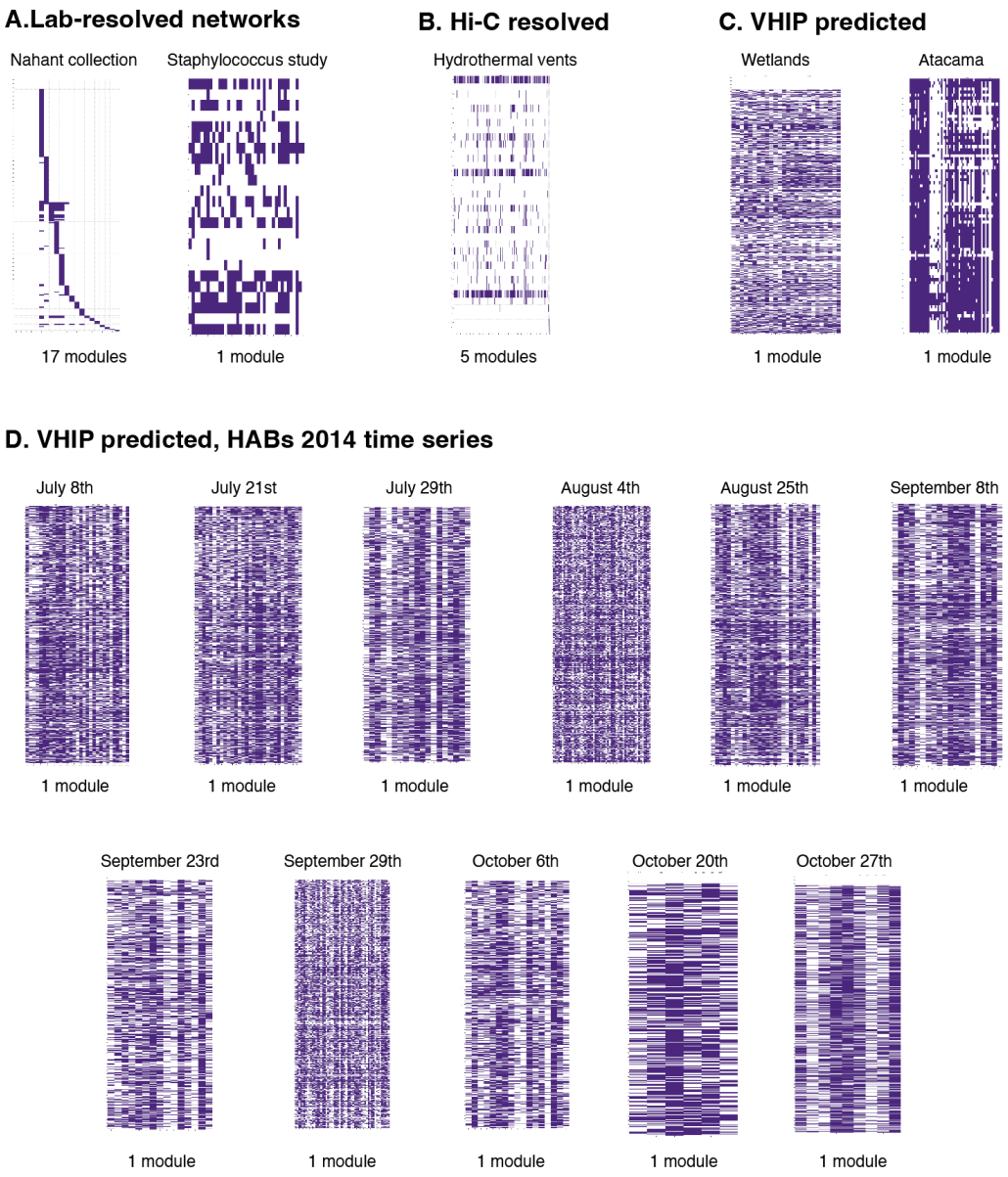
Each method has inherent limitations and scope, which affects our interpretation of the result networks. For lab-tested networks, only a fraction of the microbial community can be recovered and studied in the lab. Hi-C, while capturing the total community genomic sequences, cannot affirm infection. Taxonomy-focused methods predict the most likely taxa virus can infect but cannot resolved interaction networks. Many-to-many methods, such as VHIP, can predict total community interactions, but cannot definite proof that those predictions are real.

5.7 Supplementary figures



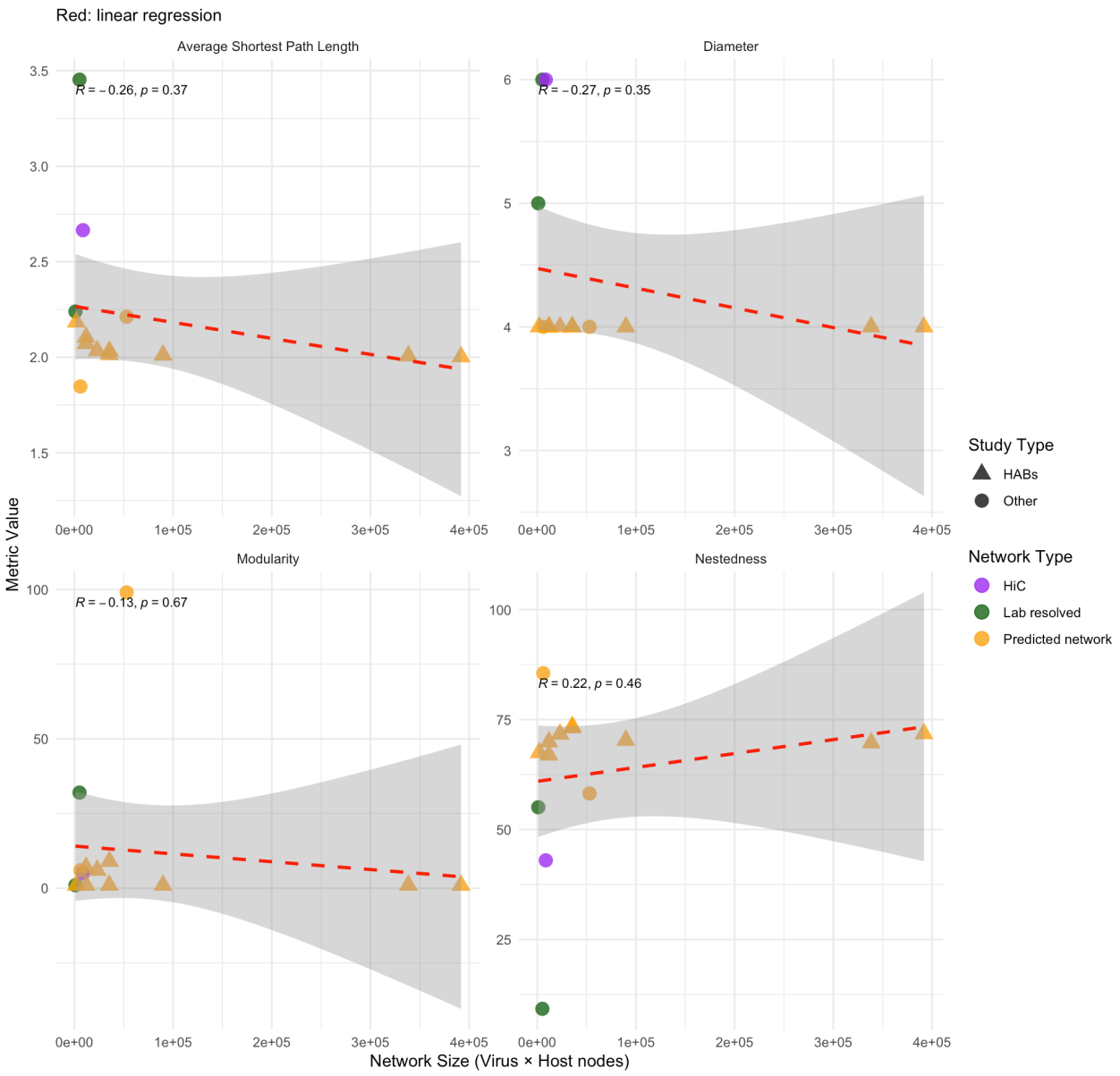
Supplemental Figure 5.5. Distributions of VHIP-scores for predicted virus-host interaction networks

Confidence score near 0 means VHIP strongly predicted non-infection, whereas scores near 1 represents strongly predicted infection. Values in between represent various confidence in the prediction, where 0.5 represents a random prediction.

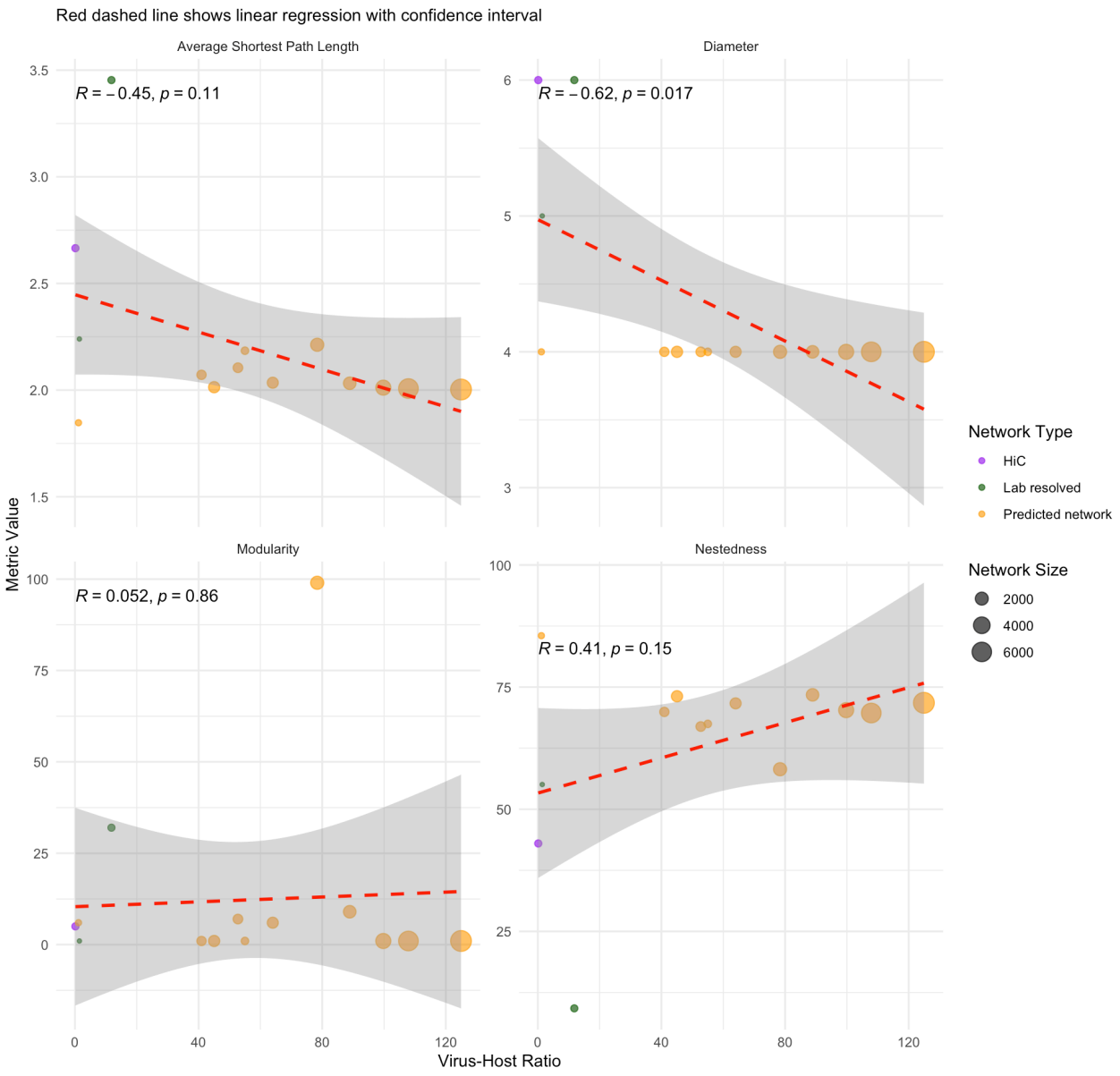


Supplemental Figure 5.6. Modularity of the virus-host interaction networks.

Only two studies resulted in modules (as determined by the Infomap algorithm), the lab-resolved Nahant collection and the Hi-C resolved hydrothermal vents.



Supplemental Figure 5.7. Correlation between network architecture and network size



Supplemental Figure 5.8. Correlation between network architecture and virus-host ratios

5.8 Tables

Data	Methods to determine interactions	Number of viruses	Number of hosts	Ratio of viruses to host	Number of possible pairs
Nahant Collection	Lab-tested (plaque assays)	248	21	11.81	5,208
Staphylococcus Study	Lab-tested (plaque assays)	27	39	0.69	1,053
Atacama	VHIP predicted from metagenomics	84	73	1.15	6,132
Wetland	VHIP predicted from metagenomics	2036	26	78.31	52,936
Hi-C Hydrothermal vents	Hi-C ligation	36	246	0.149	8856
HABs July 8th 2014	VHIP predicted from metagenomics sequences	3057	31	98.61	94,767
HABs July 21st 2014	VHIP predicted from metagenomics sequences	2993	30	99.77	89,790
HABs July 29th 2014	VHIP predicted from metagenomics sequences	1216	19	64	23,104
HABs August 4th 2014	VHIP predicted from metagenomics sequences	6995	56	124.91	391,720
HABs August 25th 2014	VHIP predicted from metagenomics sequences	1261	28	45.04	35,308
HABs September 8th 2014	VHIP predicted from metagenomics sequences	1778	20	88.9	35,560
HABs September 23rd 2014	VHIP predicted from metagenomics sequences	791	15	52.73	11,865
HABs September 29th 2024	VHIP predicted from metagenomics sequences	6042	56	107.89	338,352
HABs October 6th 2014	VHIP predicted from metagenomics sequences	696	17	40.94	11,832
HABs October 20th 2014	VHIP predicted from metagenomics sequences	330	6	55	1,980
HABs October 27th 2014	VHIP predicted from metagenomics sequences	4667	9	518.56	42,003

Table 5.1. Date provenance, methodology used, number of viruses and hosts for each network considered in this study.

Data	Diameter of largest component	Average of shortest path length	Network density (%)	Nestedness (NODF)	Number of modules (infomap)
Nahant collection	6	3.453	6.11	9.2	17
Staphylococcus study	5	2.24	31.6	55.1	1
Atacama	4	1.847	61.8	85.5	1
Wetland	4	2.213	44.5	58.2	1
Hi-C Hydrothermal vents	6	2.665	9.9	43.0	5
HABs July 8th 2014	4	2.01	57.9	72.5	1
HABs July 21st 2014	4	2.012	57.6	70.3	1
HABs July 29th 2014	4	2.035	54.1	71.7	1
HABs August 4th 2014	4	2.003	57.7	71.8	1
HABs August 25th 2014	4	2.014	57.1	73.1	1
HABs September 8th 2014	4	2.032	56.1	73.4	1
HABs September 23rd 2014	4	2.105	48.2	66.9	1
HABs September 29th 2024	4	2.007	53	69.7	1
HABs October 6th 2014	4	2.072	51.2	69.9	1
HABs October 20th 2014	4	2.185	56	67.5	1
HABs October 27th 2014	4	2.106	52.9	69.5	1

Table 5.2. Properties of the 11 virus-host interaction networks.

5.9 References

1. Brum,J.R., Ignacio-Espinoza,J.C., Roux,S., Doulcier,G., Acinas,S.G., Alberti,A., Chaffron,S., Cruaud,C., Vargas,C. de, Gasol,J.M., *et al.* (2015) Patterns and ecological drivers of ocean viral communities. *Science*, **348**.
2. Gregory,A.C., Zayed,A.A., Conceição-Neto,N., Temperton,B., Bolduc,B., Alberti,A., Ardyna,M., Arkhipova,K., Carmichael,M., Cruaud,C., *et al.* (2019) Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell*, **177**, 1109-1123.e14.
3. Paez-Espino,D., Eloe-Fadrosh,E.A., Pavlopoulos,G.A., Thomas,A.D., Huntemann,M., Mikhailova,N., Rubin,E., Ivanova,N.N. and Kyrpides,N.C. (2016) Uncovering Earth's virome. *Nature*, **536**, 425–430.
4. Edwards,R.A. and Rohwer,F. (2005) Viral metagenomics. *Nat Rev Microbiol*, **3**, 504–510.
5. Suttle,C.A. (2007) Marine viruses — major players in the global ecosystem. *Nat Rev Microbiol*, **5**, 801–812.
6. Roux,S., Krupovic,M., Daly,R.A., Borges,A.L., Nayfach,S., Schulz,F., Sharrar,A., Matheus Carnevali,P.B., Cheng,J.-F., Ivanova,N.N., *et al.* (2019) Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat Microbiol*, **4**, 1895–1906.
7. Tian,F., Wainaina,J.M., Howard-Varona,C., Domínguez-Huerta,G., Bolduc,B., Gazitúa,M.C., Smith,G., Gittrich,M.R., Zablocki,O., Cronin,D.R., *et al.* (2024) Prokaryotic-virus-encoded auxiliary metabolic genes throughout the global oceans. *Microbiome*, **12**, 159.
8. Suttle,C.A. (1994) The significance of viruses to mortality in aquatic microbial communities. *Microb Ecol*, **28**, 237–243.
9. Wilhelm,S.W. and Suttle,C.A. (1999) Viruses and Nutrient Cycles in the Sea: Viruses play critical roles in the structure and function of aquatic food webs. *BioScience*, **49**, 781–788.
10. Coutinho,F.H., Silveira,C.B., Gregoracci,G.B., Thompson,C.C., Edwards,R.A., Brussaard,C.P.D., Dutilh,B.E. and Thompson,F.L. (2017) Marine viruses discovered

via metagenomics shed light on viral strategies throughout the oceans. *Nat Commun*, **8**, 1–12.

11. Jiang,S.C. and Paul,J.H. (1998) Gene transfer by transduction in the marine environment. *Appl Environ Microbiol*, **64**, 2780–2787.
12. Howard-Varona,C., Lindback,M.M., Bastien,G.E., Solonenko,N., Zayed,A.A., Jang,H., Andreopoulos,B., Brewer,H.M., Glavina del Rio,T., Adkins,J.N., *et al.* (2020) Phage-specific metabolic reprogramming of virocells. *ISME J*, **14**, 881–895.
13. Forterre,P. (2013) The virocell concept and environmental microbiology. *ISME J*, **7**, 233–236.
14. Rosenwasser,S., Ziv,C., Creveld,S.G. van and Vardi,A. (2016) Virocell Metabolism: Metabolic Innovations During Host–Virus Interactions in the Ocean. *Trends in Microbiology*, **24**, 821–832.
15. Guidi,L., Chaffron,S., Bittner,L., Eveillard,D., Larhlimi,A., Roux,S., Darzi,Y., Audic,S., Berline,L., Brum,J.R., *et al.* (2016) Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, **532**, 465–470.
16. Wieczynski,D.J., Yoshimura,K.M., Denison,E.R., Geisen,S., DeBruyn,J.M., Shaw,A.J., Weston,D.J., Pelletier,D.A., Wilhelm,S.W. and Gibert,J.P. (2023) Viral infections likely mediate microbial controls on ecosystem responses to global warming. *FEMS Microbiol Ecol*, **99**, fiad016.
17. Hwang,Y., Roux,S., Coclet,C., Krause,S.J.E. and Girguis,P.R. (2023) Viruses interact with hosts that span distantly related microbial domains in dense hydrothermal mats. *Nat Microbiol*, **8**, 946–957.
18. Marbouty,M., Cournac,A., Flot,J.-F., Marie-Nelly,H., Mozziconacci,J. and Koszul,R. (2014) Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *eLife*, **3**, e03318.
19. Du,Y., Fuhrman,J.A. and Sun,F. (2023) ViralCC retrieves complete viral genomes and virus-host pairs from metagenomic Hi-C data. *Nat Commun*, **14**, 502.
20. Wu,R., Davison,M.R., Nelson,W.C., Smith,M.L., Lipton,M.S., Jansson,J.K., McClure,R.S., McDermott,J.E. and Hofmockel,K.S. (2023) Hi-C metagenome sequencing reveals soil phage–host interactions. *Nat Commun*, **14**, 7666.

21. Coutinho,F.H., Zaragoza-Solas,A., López-Pérez,M., Barylski,J., Zielezinski,A., Dutilh,B.E., Edwards,R. and Rodriguez-Valera,F. (2021) RaFAH: Host prediction for viruses of Bacteria and Archaea based on protein content. *Patterns*, **2**, 100274.
22. Shang,J. and Sun,Y. (2022) CHERRY: a Computational metHod for accuratE pRediction of virus-pRokarYotic interactions using a graph encoder-decoder model. *Briefings in Bioinformatics*, 10.1093/bib/bbac182.
23. Bastien,E.G., Cable,R.N., Zaman,L., Batterbee,C., Wing,A.J. and Duhaime,M.B. (2023) Virus-Host Interactions Predictor (VHIP): machine learning approach to resolve microbial virus-host interaction networks. 10.1101/2023.11.03.565433.
24. Roux,S., Camargo,A.P., Coutinho,F.H., Dabdoub,S.M., Dutilh,B.E., Nayfach,S. and Tritt,A. (2022) iPHoP: an integrated machine-learning framework to maximize host prediction for metagenome-assembled virus genomes. 10.1101/2022.07.28.501908.
25. Galiez,C., Siebert,M., Enault,F., Vincent,J. and Söding,J. (2017) WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics*, **33**, 3113–3114.
26. Wang,W., Ren,J., Tang,K., Dart,E., Ignacio-Espinoza,J.C., Fuhrman,J.A., Braun,J., Sun,F. and Ahlgren,N.A. (2020) A network-based integrated framework for predicting virus–prokaryote interactions. *NAR Genom Bioinform*, **2**.
27. Lu,C., Zhang,Z., Cai,Z., Zhu,Z., Qiu,Y., Wu,A., Jiang,T., Zheng,H. and Peng,Y. (2021) Prokaryotic virus host predictor: a Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biology*, **19**, 5.
28. Edwards,R.A., McNair,K., Faust,K., Raes,J. and Dutilh,B.E. (2016) Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol Rev*, **40**, 258–272.
29. Neutel,A.-M., Heesterbeek,J.A.P. and de Ruiter,P.C. (2002) Stability in Real Food Webs: Weak Links in Long Loops. *Science*, **296**, 1120–1123.
30. Grilli,J., Rogers,T. and Allesina,S. (2016) Modularity and stability in ecological communities. *Nat Commun*, **7**, 12031.
31. McCann,K., Hastings,A. and Huxel,G.R. (1998) Weak trophic interactions and the balance of nature. *Nature*, **395**, 794–798.

32. Bascompte,J., Jordano,P., Melian,C.J. and Olesen,J.M. (2003) The nested assembly of plant-animal mutualistic networks. *Proceedings of the National Academy of Sciences*, **100**, 9383–9387.
33. Bastolla,U., Fortuna,M.A., Pascual-García,A., Ferrera,A., Luque,B. and Bascompte,J. (2009) The architecture of mutualistic networks minimizes competition and increases biodiversity. *Nature*, **458**, 1018–1020.
34. Beiler,K.J., Simard,S.W. and Durall,D.M. (2015) Topology of tree–mycorrhizal fungus interaction networks in xeric and mesic Douglas-fir forests. *Journal of Ecology*, **103**, 616–628.
35. Simard,S.W., Beiler,K.J., Bingham,M.A., Deslippe,J.R., Philip,L.J. and Teste,F.P. (2012) Mycorrhizal networks: Mechanisms, ecology and modelling. *Fungal Biology Reviews*, **26**, 39–60.
36. Stouffer,D.B. and Bascompte,J. (2011) Compartmentalization increases food-web persistence. *Proceedings of the National Academy of Sciences*, **108**, 3648–3652.
37. Gravel,D., Massol,F. and Leibold,M.A. (2016) Stability and complexity in model meta-ecosystems. *Nat Commun*, **7**, 12457.
38. Whitman,W.B., Coleman,D.C. and Wiebe,W.J. (1998) Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences*, **95**, 6578–6583.
39. Stewart,E.J. (2012) Growing Unculturable Bacteria. *Journal of Bacteriology*, **194**, 4151–4160.
40. Kent,A.G., Vill,A.C., Shi,Q., Satlin,M.J. and Brito,I.L. (2020) Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C. *Nat Commun*, **11**, 4379.
41. Bickhart,D.M., Watson,M., Koren,S., Panke-Buisse,K., Cersosimo,L.M., Press,M.O., Van Tassell,C.P., Van Kessel,J.A.S., Haley,B.J., Kim,S.W., *et al.* (2019) Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. *Genome Biology*, **20**, 153.
42. Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, **43**, 1059–1065.

43. Shatadru,R.N., Solonenko,N.E., Sun,C.L. and Sullivan,M.B. (2025) Synthetic community Hi-C benchmarking provides a baseline for virus-host inferences. *bioRxiv*, 10.1101/2025.02.12.637985.

44. Young,F., Rogers,S. and Robertson,D.L. (2020) Predicting host taxonomic information from viral genomes: A comparison of feature representations. *PLOS Computational Biology*, **16**, e1007894.

45. Zhang,M., Yang,L., Ren,J., Ahlgren,N.A., Fuhrman,J.A. and Sun,F. (2017) Prediction of virus-host infectious association by supervised learning methods. *BMC Bioinformatics*, **18**, 60.

46. Coclet,C. and Roux,S. (2021) Global overview and major challenges of host prediction methods for uncultivated phages. *Current Opinion in Virology*, **49**, 117–126.

47. Fründ,J., McCann,K.S. and Williams,N.M. (2016) Sampling bias is a challenge for quantifying specialization and network structure: lessons from a quantitative niche model. *Oikos*, **125**, 502–513.

48. Aguiar,M.A.M. de, Newman,E.A., Pires,M.M., Yeakel,J.D., Boettiger,C., Burkle,L.A., Gravel,D., Jr,P.R.G., O'Donnell,J.L., Poisot,T., *et al.* (2019) Revealing biases in the sampling of ecological interaction networks. *PeerJ*, **7**, e7566.

49. Macgregor,C.J., Evans,D.M. and Pocock,M.J.O. (2017) Estimating sampling completeness of interactions in quantitative bipartite ecological networks: incorporating variation in species' specialisation. 10.1101/195917.

50. Connon,S.A. and Giovannoni,S.J. (2002) High-Throughput Methods for Culturing Microorganisms in Very-Low-Nutrient Media Yield Diverse New Marine Isolates. *Applied and Environmental Microbiology*, **68**, 3878–3885.

51. Kogure,K., Simidu,U. and Taga,N. (1979) A tentative direct microscopic method for counting living marine bacteria. *Can. J. Microbiol.*, **25**, 415–420.

52. Staley,J.T. and Konopka,A. (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol*, **39**, 321–346.

53. Kauffman,K.M., Brown,J.M., Sharma,R.S., VanInsberghe,D., Elsherbini,J., Polz,M. and Kelly,L. (2018) Viruses of the Nahant Collection, characterization of 251 marine Vibrionaceae viruses. *Scientific Data*, **5**, 1–11.

54. Kauffman,K.M., Hussain,F.A., Yang,J., Arevalo,P., Brown,J.M., Chang,W.K., VanInsberghe,D., Elsherbini,J., Sharma,R.S., Cutler,M.B., *et al.* (2018) A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature*, **554**, 118–122.

55. Göller,P.C., Elsener,T., Lorgé,D., Radulovic,N., Bernardi,V., Naumann,A., Amri,N., Khatchatourova,E., Coutinho,F.H., Loessner,M.J., *et al.* (2021) Multi-species host range of staphylococcal phages isolated from wastewater. *Nat Commun*, **12**, 6965.

56. Marbouty,M., Thierry,A., Millot,G.A. and Koszul,R. (2021) MetaHiC phage-bacteria infection network reveals active cycling phages of the healthy human gut. *eLife*, **10**, e60608.

57. Chen,Y., Wang,Y., Paez-Espino,D., Polz,M.F. and Zhang,T. (2021) Prokaryotic viruses impact functional microorganisms in nutrient removal and carbon cycle in wastewater treatment plants. *Nat Commun*, **12**, 5398.

58. Bastien,G.E., Cable,R.N., Batterbee,C., Wing,A.J., Zaman,L. and Duhaime,M.B. (2024) Virus-host interactions predictor (VHIP): Machine learning approach to resolve microbial virus-host interaction networks. *PLOS Computational Biology*, **20**, e1011649.

59. Hwang,Y., Rahlff,J., Schulze-Makuch,D., Schloter,M. and Probst,A.J. Diverse Viruses Carrying Genes for Microbial Extremotolerance in the Atacama Desert Hyperarid Soil. *mSystems*, **6**, e00385-21.

60. Dalcin Martins,P., Danczak,R.E., Roux,S., Frank,J., Borton,M.A., Wolfe,R.A., Burris,M.N. and Wilkins,M.J. (2018) Viral and metabolic controls on high rates of microbial sulfur and carbon cycling in wetland ecosystems. *Microbiome*, **6**, 138.

61. Wing,A.J., Hegarty,B., Bastien,E., Denef,V., Evans,J., Dick,G. and Duhaime,M. (2024) Tracking Microcystis viruses and infection dynamics across distinct phases of a Microcystis-dominated bloom. 10.1101/2024.05.24.595742.

62. Berry,M.A., Davis,T.W., Cory,R.M., Duhaime,M.B., Johengen,T.H., Kling,G.W., Marino,J.A., Uyl,P.A.D., Gossiaux,D., Dick,G.J., *et al.* (2017) Cyanobacterial harmful algal blooms are a biological disturbance to Western Lake Erie bacterial communities. *Environmental Microbiology*, **19**, 1149–1162.

63. Almeida-Neto,M., Guimarães,P., Guimarães Jr,P.R., Loyola,R.D. and Ulrich,W. (2008) A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. *Oikos*, **117**, 1227–1239.

64. Farage,C., Edler,D., Eklöf,A., Rosvall,M. and Pilosof,S. (2021) Identifying flow modules in ecological networks using Infomap. *Methods in Ecology and Evolution*, **12**, 778–786.

65. Rosvall,M. and Bergstrom,C.T. (2008) Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, **105**, 1118–1123.

66. Banerjee,S., Schlaeppi,K. and van der Heijden,M.G.A. (2018) Keystone taxa as drivers of microbiome structure and functioning. *Nat Rev Microbiol*, **16**, 567–576.

67. Irwin,N.A.T., Pittis,A.A., Richards,T.A. and Keeling,P.J. (2022) Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat Microbiol*, **7**, 327–336.

68. Simmonds,P., Aiewsakun,P. and Katzourakis,A. (2019) Prisoners of war — host adaptation and its constraints on virus evolution. *Nat Rev Microbiol*, **17**, 321–328.

69. Koskella,B. and Bergelson,J. (2020) The study of host–microbiome (co)evolution across levels of selection. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **375**, 20190604.

70. Albert,R. (2007) Network Inference, Analysis, and Modeling in Systems Biology. *Plant Cell*, **19**, 3327–3338.

71. Bhalla,U.S. and Iyengar,R. (1999) Emergent Properties of Networks of Biological Signaling Pathways. *Science*, **283**, 381–387.

72. Borodovich,T., Shkoporov,A.N., Ross,R.P. and Hill,C. (2022) Phage-mediated horizontal gene transfer and its implications for the human gut microbiome. *Gastroenterol Rep (Oxf)*, **10**, goac012.

73. Touchon,M., Moura de Sousa,J.A. and Rocha,E.P. (2017) Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Current Opinion in Microbiology*, **38**, 66–73.

74. Gregory,A.C., Solonenko,S.A., Ignacio-Espinoza,J.C., LaButti,K., Copeland,A., Sudek,S., Maitland,A., Chittick,L., dos Santos,F., Weitz,J.S., *et al.* (2016) Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genomics*, **17**, 930.

75. Coyte,K.Z., Schluter,J. and Foster,K.R. (2015) The ecology of the microbiome: Networks, competition, and stability. *Science*, **350**, 663–666.
76. Jordán,F. and Scheuring,I. (2004) Network ecology: topological constraints on ecosystem dynamics. *Physics of Life Reviews*, **1**, 139–172.
77. Ings,T.C., Montoya,J.M., Bascompte,J., Blüthgen,N., Brown,L., Dormann,C.F., Edwards,F., Figueroa,D., Jacob,U., Jones,J.I., *et al.* (2009) Review: Ecological networks – beyond food webs. *Journal of Animal Ecology*, **78**, 253–269.
78. Mandakovic,D., Rojas,C., Maldonado,J., Latorre,M., Travisany,D., Delage,E., Bihouée,A., Jean,G., Díaz,F.P., Fernández-Gómez,B., *et al.* (2018) Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. *Sci Rep*, **8**, 5875.
79. Muscatt,G., Cook,R., Millard,A., Bending,G.D. and Jameson,E. Viral metagenomics reveals diverse virus-host interactions throughout the soil depth profile. *mBio*, **14**, e02246-23.
80. Ji,M., Fan,X., Cornell,C.R., Zhang,Y., Yuan,M.M., Tian,Z., Sun,K., Gao,R., Liu,Y. and Zhou,J. Tundra Soil Viruses Mediate Responses of Microbial Communities to Climate Warming. *mBio*, **14**, e03009-22.
81. Roughgarden,J. (2024) Lytic/Lysogenic Transition as a Life-History Switch. *Virus Evol*, **10**, veae028.
82. Coutinho,F.H., Edwards,R.A. and Rodríguez-Valera,F. (2019) Charting the diversity of uncultured viruses of Archaea and Bacteria. *BMC Biology*, **17**, 109.
83. Rappé,M.S. and Giovannoni,S.J. (2003) The Uncultured Microbial Majority. *Annual Review of Microbiology*, **57**, 369–394.
84. Emerson,J.B., Roux,S., Brum,J.R., Bolduc,B., Woodcroft,B.J., Jang,H.B., Singleton,C.M., Solden,L.M., Naas,A.E., Boyd,J.A., *et al.* (2018) Host-linked soil viral ecology along a permafrost thaw gradient. *Nat Microbiol*, **3**, 870–880.
85. Breitbart,M. (2011) Marine Viruses: Truth or Dare. *Annu. Rev. Mar. Sci.*, **4**, 425–448.
86. Fister,S., Robben,C., Witte,A.K., Schoder,D., Wagner,M. and Rossmanith,P. (2016) Influence of Environmental Factors on Phage–Bacteria Interaction and on the Efficacy and Infectivity of Phage P100. *Front. Microbiol.*, **7**.

87. Chevallereau,A., Pons,B.J., van Houte,S. and Westra,E.R. (2021) Interactions between bacterial and phage communities in natural environments. *Nat Rev Microbiol*, 10.1038/s41579-021-00602-y.

88. Hernández,S. and Vives,M.J. (2020) Phages in Anaerobic Systems. *Viruses*, **12**, 1091.

89. Jończyk,E., Kłak,M., Międzybrodzki,R. and Górski,A. (2011) The influence of external factors on bacteriophages—review. *Folia Microbiol (Praha)*, **56**, 191–200.

90. Wells,L.E. and Deming,J.W. (2006) Effects of temperature, salinity and clay particles on inactivation and decay of cold-active marine Bacteriophage 9A. *Aquatic Microbial Ecology*, **45**, 31–39.

91. Silpe,J.E. and Bassler,B.L. (2019) A Host-Produced Quorum-Sensing Autoinducer Controls a Phage Lysis-Lysogeny Decision. *Cell*, **176**, 268-280.e13.

92. Visnapuu,A., Van der Gucht,M., Wagemans,J. and Lavigne,R. (2022) Deconstructing the Phage–Bacterial Biofilm Interaction as a Basis to Establish New Antibiofilm Strategies. *Viruses*, **14**, 1057.

93. Sime-Ngando,T. (2014) Environmental bacteriophages: viruses of microbes in aquatic ecosystems. *Front. Microbiol.*, **5**.

94. Clarke,D.K., Duarte,E.A., Moya,A., Elena,S.F., Domingo,E. and Holland,J. (1993) Genetic bottlenecks and population passages cause profound fitness differences in RNA viruses. *J Virol*, **67**, 222–228.

95. Dion,M.B., Oechslin,F. and Moineau,S. (2020) Phage diversity, genomics and phylogeny. *Nat Rev Microbiol*, **18**, 125–138.

96. Stern,A. and Sorek,R. (2011) The phage-host arms race: Shaping the evolution of microbes. *BioEssays*, **33**, 43–51.

97. Koskella,B. and Brockhurst,M.A. (2014) Bacteria–phage coevolution as a driver of ecological and evolutionary processes in microbial communities. *FEMS Microbiology Reviews*, **38**, 916–931.

98. Marbouty,M., Baudry,L., Cournac,A. and Koszul,R. (2017) Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. *Science Advances*, **3**, e1602105.

99. Holmudden,M., Gustafsson,J., Bertrand,Y.J.K., Schliep,A. and Norberg,P. (2024) Evolution shapes and conserves genomic signatures in viruses. *Commun Biol*, **7**, 1–12.

100. Flores,C.O., Meyer,J.R., Valverde,S., Farr,L. and Weitz,J.S. (2011) Statistical structure of host–phage interactions. *PNAS*, **108**, E288–E297.
101. Weitz,J.S., Poisot,T., Meyer,J.R., Flores,C.O., Valverde,S., Sullivan,M.B. and Hochberg,M.E. (2013) Phage–bacteria infection networks. *Trends in Microbiology*, **21**, 82–91.
102. Wu,L.-Y., Wijesekara,Y., Piedade,G.J., Pappas,N., Brussaard,C.P.D. and Dutilh,B.E. (2024) Benchmarking bioinformatic virus identification tools using real-world metagenomic data across biomes. *Genome Biology*, **25**, 97.
103. Connor,N., Barberán,A. and Clauset,A. (2017) Using null models to infer microbial co-occurrence networks. *PLOS ONE*, **12**, e0176751.

Chapter 6 Conclusion

6.1 Viral ecology in the contemporary age

Fourteen years ago, Mya Breitbart, in her "Marine Viruses: Truth or Dare" review, issued several dares, important knowledge gaps that defined the challenges faced in viral ecology (1). Among these dares was a need to link viruses to their hosts and characterize the virus-host range. This dissertation addressed this challenge. The development of VHIP, a computational tool that leverages genomic scars embedded in genomic sequences of viruses and hosts, offers a complementary strategy to existing methods. This expansion of methods, from Hi-C sequencing techniques that capture spatial proximity between viral and host genetic material (2, 3), higher throughput laboratory techniques (4, 5), and microscopy (6), to machine learning frameworks that detect genomic patterns (7), reflects the field's recognition that virus-host interactions operate as complex many-to-many networks rather than one-to-one relationships.

However, beyond technological and methodological advances, conceptual maturation is needed regarding virus-host interactions. The challenge of linking viruses to hosts is intrinsically tied to our understanding of the concept of host range. As we develop better methods to identify who infects whom, we must simultaneously refine our frameworks to understand the nature, breadth, and dynamics of these interactions.

6.2 Reconciling the definitions of host range

Host range describes two distinct properties that are often conflated. One virus might infect hundreds of strains within a single species. Another might infect just a few strains spanning multiple genera. Both cases have led researchers to call them 'broad host range' viruses,

highlighting the term's inherent ambiguity (8). This lack of clarity stems from viral ecology's interdisciplinary origins. Different research communities brought their own interpretations of host range, and these varied meanings persist throughout the literature (8–10). In phage therapy, researchers focus on strain-level specificity by identifying viruses that target pathogenic bacteria without affecting beneficial microbiota, and monitoring how quickly resistance evolves (11, 12). Here, host range operates at the resolution of individual mutations that alter receptors or defense systems (13–15). In viral ecology, attention shifts to viruses infecting diverse taxonomic groups, as these generalists may drive horizontal gene transfer between distant lineages or persist longer in changing environments (16, 17). These different applications require different definitions of host range.

These seemingly conflicting definitions of host range reflect different temporal scales of virus-host interactions. Strain-level host range captures rapid evolutionary dynamics—viruses gaining or losing the ability to infect specific strains through mutations in receptor-binding proteins or host acquisition of defense systems (18–20). Laboratory evolution experiments demonstrate these changes occurring over days to months (21, 22). By contrast, the ability to infect different bacterial species may reflect deeper evolutionary adaptations (23, 24). A virus capable of infecting both *Escherichia* and *Pseudomonas* must possess molecular machinery compatible with the cellular environments of each host. Such broad compatibility suggests a more nuanced coevolutionary history.

This perspective aligns with the niche-filling model of viral evolution, which proposes that viruses rapidly optimize fitness within specific host populations before becoming "imprisoned" in co-adapted relationships (25). Building upon that framework, we propose that *realized host range* describes immediate infection dynamics, which specific strains a virus can infect right now, and is measurable through laboratory assays. This reflects the virus exploring and optimizing its niche space through rapid adaptation. At evolutionary timescales, we propose that the *potential host range* encompasses deeper compatibility between viral and host molecular machinery, reflecting the boundaries of the niche. Computational approaches that detect these signatures of long-term adaptation

complement laboratory studies of immediate infection, together providing a more comprehensive picture of host range.

By recognizing that host range operates on different timescales, from the immediacy of receptor-ligand interactions to the compatibility of molecular systems, we can reconcile the field's varied definitions. Time becomes the unifying framework: strain-level specificity reflects recent adaptation within the constraints of the evolutionary niche, while species-level generalism reveals coevolutionary relationships that define the niche boundaries. This framework suggests that understanding host range requires not just asking "who infects whom?" but also "over what timescale?", and whether we are observing a virus exploring its niche (realized host range) or revealing the limits of its evolutionary boundaries (potential host range).

6.3 The continuum of viral outcomes

Traditional approaches to virus-host interactions have largely relied on binary classifications of infection versus non-infection, a framework that oversimplifies these relationships. This view ignores mutualistic relationships where viruses confer environment-dependent benefits to their hosts (26). Furthermore, the diversity of viral life cycles (i.e., acute infections, persistent infections, lysogenic cycles, and chronic infections) creates a continuum of interaction states that cannot be adequately represented by simple binary categories (27–29). Even 'failed' infections where viruses cannot replicate still alter host physiology (30, 31). These context-dependent outcomes reveal that virus-host interactions are far more nuanced than binary classifications would otherwise suggest.

While this dissertation did not incorporate virus fitness data, though such data would have been highly valuable, the absence reflects a broader challenge in the field. The availability of comprehensive virus fitness data remains limited, constraining researchers' ability to fully understand the mechanistic drivers of viral success and adaptation. However, the signals of viral adaptation that we have identified throughout this work are inherently and intrinsically linked to fitness landscapes, and future investigations should explore fitness

metrics to provide deeper insights into the ecological and evolutionary pressures shaping virus-host interactions.

6.4 Applied viral ecology

The insights from this dissertation extend beyond academic understanding into practical applications across multiple sectors. In medicine, the antibiotic resistance crisis has renewed interest in phage therapy (32–34). Network-based approaches to predicting virus-host interactions can guide the design of therapeutic phage cocktails that target pathogens while preserving beneficial microbiota. In agriculture, understanding host range enables the development of phages as biocontrol agents, sustainable alternatives to chemical pesticides that selectively target plant pathogens (35, 36). Similarly, biotechnology applications benefit from improved host prediction, allowing more precise engineering of viral vectors for gene delivery and protein production (37, 38). While much work remains to fully translate these insights into practice, understanding virus-host interactions at scale represents an important step toward more effective use of viruses in applied settings.

6.5 References

1. Breitbart,M. (2011) Marine Viruses: Truth or Dare. *Annu. Rev. Mar. Sci.*, **4**, 425–448.
2. Du,Y., Fuhrman,J.A. and Sun,F. (2023) ViralCC retrieves complete viral genomes and virus-host pairs from metagenomic Hi-C data. *Nat Commun*, **14**, 502.
3. Hwang,Y., Roux,S., Coclet,C., Krause,S.J.E. and Girguis,P.R. (2023) Viruses interact with hosts that span distantly related microbial domains in dense hydrothermal mats. *Nat Microbiol*, **8**, 946–957.
4. Labonté,J.M., Swan,B.K., Poulos,B., Luo,H., Koren,S., Hallam,S.J., Sullivan,M.B., Woyke,T., Eric Wommack,K. and Stepanauskas,R. (2015) Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J*, **9**, 2386–2399.
5. Jarett,J.K., Džunková,M., Schulz,F., Roux,S., Paez-Espino,D., Eloe-Fadrosh,E., Jungbluth,S.P., Ivanova,N., Spear,J.R., Carr,S.A., *et al.* (2020) Insights into the dynamics between viruses and their hosts in a hot spring microbial mat. *ISME J*, **14**, 2527–2541.
6. Baquero,D.P., Medvedeva,S., Martin-Gallausiaux,C., Pende,N., Sartori-Rupp,A., Tachon,S., Pedron,T., Debarbieux,L., Borrel,G., Gribaldo,S., *et al.* (2024) Stable coexistence between an archaeal virus and the dominant methanogen of the human gut. *Nat Commun*, **15**, 7702.
7. Roux,S., Camargo,A.P., Coutinho,F.H., Dabdoub,S.M., Dutilh,B.E., Nayfach,S. and Tritt,A. (2023) iPHoP: An integrated machine learning framework to maximize host prediction for metagenome-derived viruses of archaea and bacteria. *PLOS Biology*, **21**, e3002083.
8. Ross,A., Ward,S. and Hyman,P. (2016) More Is Better: Selecting for Broad Host Range Bacteriophages. *Front. Microbiol.*, **7**.
9. Hyman,P. and Abedon,S.T. (2010) Bacteriophage host range and bacterial resistance. *Adv Appl Microbiol*, **70**, 217–248.
10. Holtappels,D., Alfenas-Zerbini,P. and Koskella,B. (2023) Drivers and consequences of bacteriophage host range. *FEMS Microbiology Reviews*, **47**, fuad038.
11. Nilsson,A.S. (2014) Phage therapy—constraints and possibilities. *Upsala Journal of Medical Sciences*, **119**, 192–198.

12. Oromí-Bosch,A., Antani,J.D. and Turner,P.E. (2023) Developing Phage Therapy That Overcomes the Evolution of Bacterial Resistance. *Annual Review of Virology*, **10**, 503–524.
13. de Jonge,P.A., Nobrega,F.L., Brouns,S.J.J. and Dutilh,B.E. (2019) Molecular and Evolutionary Determinants of Bacteriophage Host Range. *Trends in Microbiology*, **27**, 51–63.
14. Labrie,S.J., Samson,J.E. and Moineau,S. (2010) Bacteriophage resistance mechanisms. *Nat Rev Microbiol*, **8**, 317–327.
15. Rostøl,J.T. and Marraffini,L. (2019) (Ph)ighting Phages: How Bacteria Resist Their Parasites. *Cell Host Microbe*, **25**, 184–194.
16. Beumer,A. and Robinson,J.B. (2005) A Broad-Host-Range, Generalized Transducing Phage (SN-T) Acquires 16S rRNA Genes from Different Genera of Bacteria. *Applied and Environmental Microbiology*, 10.1128/AEM.71.12.8301-8304.2005.
17. Hall,J.P.J., Brockhurst,M.A. and Harrison,E. (2017) Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. *Philos Trans R Soc Lond B Biol Sci*, **372**, 20160424.
18. Strobel,H.M., Labador,S.D., Basu,D., Sane,M., Corbett,K.D. and Meyer,J.R. (2024) Viral Receptor-Binding Protein Evolves New Function through Mutations That Cause Trimer Instability and Functional Heterogeneity. *Mol Biol Evol*, **41**, msae056.
19. Meyer,J.R., Dobias,D.T., Weitz,J.S., Barrick,J.E., Quick,R.T. and Lenski,R.E. (2012) Repeatability and Contingency in the Evolution of a Key Innovation in Phage Lambda. *Science*, **335**, 428–432.
20. Gerbino,K.R., Borin,J.M., Ardell,S.M., Lee,J.J., Corbett,K.D. and Meyer,J.R. (2024) Bacteriophage Φ21's receptor-binding protein evolves new functions through destabilizing mutations that generate non-genetic phenotypic heterogeneity. *Virus Evol*, **10**, veae049.
21. Gupta,A., Peng,S., Leung,C.Y., Borin,J.M., Medina,S.J., Weitz,J.S. and Meyer,J.R. (2022) Leapfrog dynamics in phage-bacteria coevolution revealed by joint analysis of cross-infection phenotypes and whole genome sequencing. *Ecology Letters*, **25**, 876–888.

22. Borin,J.M., Avrani,S., Barrick,J.E., Petrie,K.L. and Meyer,J.R. (2021) Coevolutionary phage training leads to greater bacterial suppression and delays the evolution of phage resistance. *Proceedings of the National Academy of Sciences*, **118**, e2104592118.
23. Nepal,R., Houtak,G., Karki,S., Dhungana,G., Vreugde,S. and Malla,R. (2022) Genomic characterization of three bacteriophages targeting multidrug resistant clinical isolates of Escherichia, Klebsiella and Salmonella. *Arch Microbiol*, **204**, 334.
24. Zhang,H.-Z., Shu,M., Yang,W.-Y., Pan,H., Tang,M.-X., Zhao,Y.-Y., Zhong,C. and Wu,G.-P. (2023) Isolation and characterization of a novel *Salmonella* bacteriophage JNwz02 capable of lysing *Escherichia coli* O157:H7 and its antibacterial application in foods. *LWT*, **173**, 114251.
25. Simmonds,P., Aiewsakun,P. and Katzourakis,A. (2019) Prisoners of war — host adaptation and its constraints on virus evolution. *Nat Rev Microbiol*, **17**, 321–328.
26. Roossinck,M.J. (2011) The good viruses: viral mutualistic symbioses. *Nat Rev Microbiol*, **9**, 99–108.
27. Berg,M., Goudeau,D., Olmsted,C., McMahon,K.D., Yitbarek,S., Thweatt,J.L., Bryant,D.A., Eloe-Fadrosh,E.A., Malmstrom,R.R. and Roux,S. (2021) Host population diversity as a driver of viral infection cycle in wild populations of green sulfur bacteria with long standing virus-host interactions. *ISME J*, **15**, 1569–1584.
28. Weitz,J.S., Li,G., Gulbudak,H., Cortez,M.H. and Whitaker,R.J. (2019) Viral invasion fitness across a continuum from lysis to latency†. *Virus Evolution*, **5**.
29. Howard-Varona,C., Hargreaves,K.R., Abedon,S.T. and Sullivan,M.B. (2017) Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *ISME J*, **11**, 1511–1520.
30. Guegler,C.K., Teodoro,G.I.C., Srikant,S., Chetlapalli,K., Doering,C.R., Ghose,D.A. and Laub,M.T. (2023) A phage-encoded RNA-binding protein inhibits the antiviral activity of a toxin–antitoxin system. *Nucleic Acids Res*, **52**, 1298–1312.
31. Vendrell-Fernández,S., Beamud,B., Haydar,Y.A., Sousa,J.A. de, Burlaud-Gaillard,J., Kornobis,E., Raynal,B., Vinh,J., Bikard,D. and Ghigo,J.-M. (2025) Incomplete lytic cycle of a widespread Bacteroides bacteriophage leads to the formation of defective viral particles. *PLOS Biology*, **23**, e3002787.

32. Hatfull,G.F., Dedrick,R.M. and Schooley,R.T. (2022) Phage Therapy for Antibiotic-Resistant Bacterial Infections. *Annual Review of Medicine*, **73**, 197–211.
33. Cui,L., Watanabe,S., Miyanaga,K., Kiga,K., Sasahara,T., Aiba,Y., Tan,X.-E., Veeranarayanan,S., Thitiananpakorn,K., Nguyen,H.M., *et al.* (2024) A Comprehensive Review on Phage Therapy and Phage-Based Drug Development. *Antibiotics*, **13**, 870.
34. Subramanian,A. (2024) Emerging roles of bacteriophage-based therapeutics in combating antibiotic resistance. *Front. Microbiol.*, **15**.
35. Czajkowski,R., Roca,A. and Matilla,M.A. (2025) Harnessing Bacteriophages for Sustainable Crop Protection in the Face of Climate Change. *Microb Biotechnol*, **18**, e70108.
36. Holtappels,D., Fortuna,K., Lavigne,R. and Wagemans,J. (2021) The future of phage biocontrol in integrated plant protection for sustainable crop production. *Current Opinion in Biotechnology*, **68**, 60–71.
37. Li,H.R., Harb,M., Heath,J.E., Trippett,J.S., Shapiro,M.G. and Szablowski,J.O. (2024) Engineering viral vectors for acoustically targeted gene delivery. *Nat Commun*, **15**, 4924.
38. Gencay,Y.E., Jasinskytė,D., Robert,C., Semsey,S., Martínez,V., Petersen,A.Ø., Brunner,K., de Santiago Torio,A., Salazar,A., Turcu,I.C., *et al.* (2024) Engineered phage with antibacterial CRISPR–Cas selectively reduce E. coli burden in mice. *Nat Biotechnol*, **42**, 265–274.