Exploring model-based approaches for simulating and analyzing cophylogenetic data

by

Wade Thomas Dismukes

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Ecology and Evolutionary Biology

Program of Study Committee:
Tracy A. Heath, Major Professor
Dean Adams
Jarad Niemi
Matt Hufford
John Nason

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2022

DEDICATION

I would like to dedicate this to my partner Angela Bunning who has been a constant source of support and love throughout graduate school.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, Dr. Tracy A. Heath for her guidance, immense patience and support throughout my research, personal development throughout graduate school, and the writing of this thesis. I could always rely on our meetings to make me feel encouraged and supported and ready to take on the next piece of my dissertation. I also express a huge thanks to my wonderful collaborators, Dr. Michael J. Landis, Dr. David H. Hembry and Dr. Mariana P. Braga, who contributed both directly in working with me on Chapter 2 and indirectly by being steadfast sources of support and inspiration. I would also like to thank my committee members for their efforts and contributions to this work: Dr. Dean Adams, Dr. John Nason, Dr. Matt Hufford and Dr. Jarad Niemi.

ABSTRACT

When two lineages with an intimate ecological association have a repeated history of cospeciation (*i.e.,* both speciate simultaneously), their evolutionary histories can become correlated through time. This correlated diversification can result in perfect concordance between pairs of phylogenies, a pattern often evoked in hosts and obligate symbionts. Yet, many of these presumed cases of cophylogeny do not show a perfect correspondence between host and symbiont phylogenies despite observations of obligate relationships between hosts and symbiont in nature. Previous researchers have primarily used model-free approaches to investigate cophylogenetics. My thesis focuses on moving cophylogenetics closer towards a model-based framework which will allow for more statistically rigorous quantification of cophylogenetic events and further understanding of how these assemblages of hosts and symbionts evolve.

Review contrasting the assumptions of and outcomes of cophylogenetic methods

Since previous reviews on this topic, new methods have been introduced to test hypotheses of codiversification, and existing methods have been applied in novel ways to understand the parallel histories of associated lineages. Moreover, researchers are now addressing these problems using larger phylogenetic datasets while also applying phylogenetic comparative methods to integrate ecological interactions into macroevolution. I outline different types of ecological interactions that can potentially leave signatures in the branching histories of host-symbiont systems. I contrast how different datasets may

violate the assumptions of cophylogenetic methods and how this may limit what conclusions can be drawn about the processes driving species interactions.

Developing a generative model for cophylogenetic data

Simulation is a necessary step when testing and examining phylogenetic methods. Despite this, there was no available simulation software for cophylogenetic data. I developed the cophylogenetic birth-death model and software to simulate under this model to address this gap. The cophylogenetic birth-death model simulates a pair of phylogenies, the host and the symbiont phylogenies, and their ecological interactions. This model uses independent speciation and extinction for the host and the symbiont and parameters that affect both phylogenies, including cospeciation and host-switching.

Using biogeographic models for examining cophylogenetic data

Recently probabilistic models such as the dispersal, extinction, and cladogenesis (DEC) have begun to be used in a phylogenetic context to estimate biogeographic parameters such as dispersal and extirpation. These models have a natural analogy to cophylogenetic data where the host phylogeny becomes the area cladogram, the symbiont phylogeny becomes the species tree, and the host-symbiont interaction matrix becomes the matrix representing the host ranges. As there is currently no method for estimating under the cophylogenetic birth-death model, I tested a Bayesian implementation of the DEC model to estimate symbiont dispersal to new hosts. To do this, I simulated data using treeducken to examine the DEC model's ability to estimate host dispersal.

Using machine learning to estimate cophylogenetic parameters

Without a straightforward means to calculate a likelihood for the cophylogenetic birth-death model, I pursued alternative means to estimate the parameters of this model. In particular, I used deep learning to estimate cophylogenetic events. I trained this deep neural network using data simulated under my cophylogenetic birth-death model and then tested the performance of this method. I provide a case study of using this deep neural network on empirical data of figs and fig wasps.

CHAPTER 1. GENERAL INTRODUCTION

Understanding the impact of ecological interactions on macroevolution is a central question in evolutionary biology. Nowhere is this dependence between ecological interaction and evolution more apparent than in host-symbiont systems. These systems have fascinated evolutionary biologists since Darwin (Darwin, 1862) and Wallace (Wallace, 1867). Fahrenholz (1913) proposed that more closely related symbionts would be found on closely related host species. In other words, over time, the host phylogeny and symbiont phylogeny should match. However, host and symbiont phylogenies seldom match in reality as has been observed as more phylogenies have become available. These systems have been examined through the use of cophylogenetic methods—a computational approaches that measure the support for one or more cophylogenetic hypotheses through the unified analysis of two phylogenies and the interactions among their taxa.

Cophylogenetic methods typically take two approaches for analyzing host-symbiont data: event-scoring methods and pattern-based statistical methods (Page, 2003, Dismukes et al., 2022). Event-scoring methods define a number of different events, assign a cost to each event type, and then map the symbiont phylogeny onto the host phylogeny using the events to explain incongruencies (Page, 1994, Charleston, 1998, Conow et al., 2010, Santichaivekin et al., 2020). The total cost of the events is calculated and this is repeated until the lowest cost mapping is found. Pattern-based statistical methods convert the host and symbiont phylogeny into phylogenetic distance matrices and use these to measure the cophylogenetic signal of the host and symbiont phylogenies (Legendre et al., 2002, Hommola et al., 2009, Balbuena et al., 2013, Hutchinson et al., 2017, Balbuena et al., 2020). This is often done in a hypothesis-testing framework to identify whether the cophylogenetic signal is greater than we would expect due to chance alone. More recently, a new type of analysis has emerged that uses stochastic processes to define cophylogenetic events and perform inference (Satler et al., 2019, Braga et al., 2020). In this dissertation, I set out to develop new ways of exploring cophylogenetic data.

Chapter 2 provides a comprehensive review of the current state of cophylogenetic methods and explores different aspects of cophylogenetic systems (currently in press in the journal *Annual Reviews in Ecology, Evolution and Systematics*; Dismukes et al., 2022). This review was motivated by the lack of clear guidelines available to help researchers choose which cophylogenetic methods are most appropriate for their own cophylogenetic systems and questions. To achieve this, my coauthors and I surveyed the development of methods used for cophylogenetics including pattern-based (Page, 2003, Stevens, 2004, De Vienne et al., 2013, Blasco-Costa et al., 2021), event-scoring (Charleston, 1998, Conow et al., 2010, Santichaivekin et al., 2020), and generative model-based methods (Satler et al., 2019, Braga et al., 2020). We provide additional guidance to reasarchers by categorizing cophylogenetic datasets based on shared characteristics that make certain methods more or less appropriate given the biological system under investigation. Chapter 2 concludes by identifying the limitations of currently available cophylogenetic methods to stimulate further theoretical development in this field.

Chapter 3 introduces the cophylogenetic birth-death process and the software package treeducken that simulates cophylogenetic data (published in the journal *Methods in Ecology and Evolution*; Dismukes and Heath, 2021). The cophylogenetic birth-death process is a generative model for datasets consisting of a host phylogeny, a symbiont phylogeny, and an association matrix describing the ecological interactions between the two. This chapter also introduces the R package, treeducken, that provides a straightforward method of simulating datasets under this model. We also provide a number of example vignettes that demonstrate how to use treeducken, and validation tests verifying that the simulations match model expectations.

Chapter 4 uses the cophylogenetic birth-death process (Chapter 3; Dismukes and Heath, 2021) as the basis for a deep learning method for analyzing cophylogenetic datasets. This deep learning method estimates the number of cophylogenetic events from datasets produced via the cophylogenetic birth-death process. We perform a simulation experiment to test the performance of this method. We also provide an empircal case study by applying this method to a dataset of Panamanian figs and fig wasps.

Chapter 5 explores the use of a Bayesian historical biogeography model to analyze cophylogenetic datasets. We use a straightforward analogy between historical biogeography and cophylogenetics as justification for using a historical biogeography model to estimate parameters of the cophylogenetic birth-death process. In this analogy, our host phylogeny becomes an area cladogram and our symbiont phylogeny becomes the species tree that inhabits that area cladogram.

1.1 References

Balbuena, J. A., Míguez-Lozano, R., and Blasco-Costa, I. (2013). PACo: A Novel Procrustes Application to Cophylogenetic Analysis. *PLOS ONE*, 8(4):e61048.

Balbuena, J. A., Pérez-Escobar, Ó. A., Llopis-Belenguer, C., and Llopis-Belenguer, I. (2020). Random tanglegram partitions (Random TaPas): An Alexandrian approach to the cophylogenetic Gordian knot. *Systematic Biology*. syaa033.

Blasco-Costa, I., Hayward, A., Poulin, R., and Balbuena, J. A. (2021). Next-generation cophylogeny: unravelling eco-evolutionary processes. *Trends in Ecology & Evolution*, 36(10):907–918.

Braga, M. P., Landis, M. J., Nylin, S., Janz, N., and Ronquist, F. (2020). Bayesian inference of ancestral host-parasite interactions under a phylogenetic model of host repertoire evolution. *Systematic Biology*, 69:1149–1162.

Charleston, M. (1998). Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences*, 149(2):191–223.

Conow, C., Fielder, D., Ovadia, Y., and Libeskind-Hadas, R. (2010). Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology*, 5(1):16.

Darwin, C. (1862). Letter to J.D. Hooker. *More letters of Charles Darwin Volume*, 2.

De Vienne, D., Refrégier, G., López-Villavicencio, M., Tellier, A., Hood, M., and Giraud, T. (2013). Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. *New Phytologist*, 198(2):347–385.

Dismukes, W., Braga, M. P., Hembry, D. H., Heath, T. A., and Landis, M. J. (2022). Cophylogenetic methods to untangle the evolutionary history of ecological interactions. *Annual Review of Ecology, Evolution, and Systematics*, in press.

Dismukes, W. and Heath, T. A. (2021). treeducken: An R package for simulating cophylogenetic systems. *Methods in Ecology and Evolution*, 12(8):1358–1364.

Fahrenholz, H. (1913). Ectoparasiten und abstammungslehre. *Zoologischer Anzeiger*, 41:371–374.

Hommola, K., Smith, J. E., Qiu, Y., and Gilks, W. R. (2009). A permutation test of host–parasite cospeciation. *Molecular Biology and Evolution*, 26(7):1457–1468.

Hutchinson, M. C., Cagua, E. F., and Stouffer, D. B. (2017). Cophylogenetic signal is detectable in pollination interactions across ecological scales. *Ecology*, 98(10):2640–2652.

Janzen, D. H. (1980). When is it coevolution? *Evolution*, 34(3):611–612.

Legendre, P., Desdevises, Y., and Bazin, E. (2002). A statistical test for host–parasite coevolution. *Systematic Biology*, 51(2):217–234.

Page, R. D. M. (1994). Parallel phylogenies: reconstructing the history of host-parasite assemblages. *Cladistics*, 10:155–175.

Page, R. D. M. (2003). *Tangled Trees: Phylogeny, Cospeciation, and Coevolution.* University of Chicago Press.

Price, P. W. (1980). *Evolutionary Biology of Parasites.* Princeton University Press.

Santichaivekin, S., Yang, Q., Liu, J., Mawhorter, R., Jiang, J., Wesley, T., Wu, Y.-C., and Libeskind-Hadas, R. (2020). eMPRess: a systematic cophylogeny reconciliation tool. *Bioinformatics*, in press.

Satler, J. D., Herre, E. A., Jandér, K. C., Eaton, D. A., Machado, C. A., Heath, T. A., and Nason, J. D. (2019). Inferring processes of coevolutionary diversification in a community of Panamanian strangler figs and associated pollinating wasps. *Evolution*, 73(11):2295–2311.

Stevens, J. (2004). Computational aspects of host–parasite phylogenies. *Briefings in Bioinformatics*, 5(4):339–349.

Thompson, J. N. (1994). *The Coevolutionary Process.* University of Chicago Press, Chicago, IL.

Thompson, J. N. (2005). *The Geographic Mosaic of Coevolution.* University of Chicago Press, Chicago, IL.

Wallace, A. R. (1867). Creation by law. *QJ Sci*, 4(16):470–488.

1.2 Tables and Figures

Table 1.1 Glossary of commonly used terms in cophylogenetics

Term	Definition
Coevolution	reciprocal natural selection in two or more interacting species (Janzen, 1980, Thompson, 1994, 2005)
Symbiosis	interaction between two or more organisms in close physical or physiological association. Symbioses may be mutualistic, antagonistic (in which case they are called "parasitic"), or commensalistic (Price, 1980, Thompson, 1994)
Host	in parasitism and commensalism, the organism which the parasite or commensal exploits for resources; in mutualistic symbioses often used to refer to the larger of two interacting mutualists
Symbiont	any organism engaged in symbiosis with another taxon; when used in opposition to "host", refers to the smaller of two organisms engaged in symbiosis
Cophylogenetic system	a host phylogeny, a symbiont phylogeny, and the interactions between extant tips connecting the two trees
Phylogenetic congruence	topological matching of two phylogenies of interacting clades above some statistical threshold
Cophylogenetic signal	a measure of the congruence of evolutionary history between two interacting clades
Host repertoire	the set of host taxa with which a symbiont associates/interacts

CHAPTER 2. COPHYLOGENETIC METHODS TO UNTANGLE THE EVOLUTIONARY HISTORY OF ECOLOGICAL INTERACTIONS

Wade Dismukes[1], Mariana P. Braga[2,3,4], David H. Hembry[5], Tracy A. Heath[1] and Michael J. Landis[2]

[1]*Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Bessey Hall, Ames, Iowa, 50010, USA*

[2]*Department of Biology, Washington University in St. Louis, St. Louis, MO, 63130, USA*

[3]*Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden*

[4]*Helsinki Life Science Institute, University of Helsinki, Helsinki, Finland*

[5]*Department of Biology, University of Texas Permian Basin, Odessa, TX, USA, 79762*

Modified from a manuscript published in *Annual Reviews in Ecology, Evolutuon, and Systematics*

Abstract

Myriad branches in the Tree of Life are intertwined through ecological relationships. Biologists have long hypothesized that intimate symbioses between lineages can influence diversification patterns to the extent that it leaves a topological imprint on the phylogenetic trees of interacting clades. Over the past few decades, cophylogenetic methods development has provided a toolkit for identifying such histories of codiversification, yet it is often difficult to determine which tools best suit the task at hand. In this review, we organize currently available cophylogenetic methods into three categories—pattern-based statistics, event-scoring methods, and more recently developed generative model-based methods—and discuss their assumptions and appropriateness for different types of cophylogenetic questions. We classify cophylogenetic systems based on their biological properties to provide a framework for empiricists investigating the macroevolution of symbioses. Finally, we provide recommendations for the next generation of cophylogenetic models that we hope will facilitate further methods development.

2.1 Introduction

All organisms are members of ecological communities, and interact with individuals of other species throughout their lives. These interactions vary in terms of kind, strength, and intimacy, *i.e.*, the degree of biological integration among associated organisms (Ollerton, 2006, Guimarães et al., 2007). Modern microbiology, ecology, and molecular biology, as well as the advent of new sequencing technologies, have revealed the immense, often cryptic, diversity of symbioses (high-intimacy interactions). It is estimated that >40% of the lineages in the Tree of Life are symbionts in the broad sense—parasites, mutualists, and commensals intimately dependent on the lives of others (Dobson et al., 2008, Wang and Qiu, 2006, National Research Council, 2007).

This diversity of coevolving symbiotic lineages is staggering in its extent and pervasiveness among Earth's ecosystems, ranging from gut bacteria in vertebrate megafauna to parasitic mistletoe plants on conifers and angiosperms; from chemosynthetic bacterial symbionts of deep-sea bivalves to galling, chewing, and leaf-mining herbivorous insects on angiosperms; from the mitochondria and chloroplasts of eukaryotes to green algae inhabiting the tissues of corals, sea slugs, lichens, and salamander eggs; and from specialized brood-pollination mutualisms like fig wasps and yucca moths to a yet mostly unmeasured diversity of mites and nematodes associated with most plants and animals

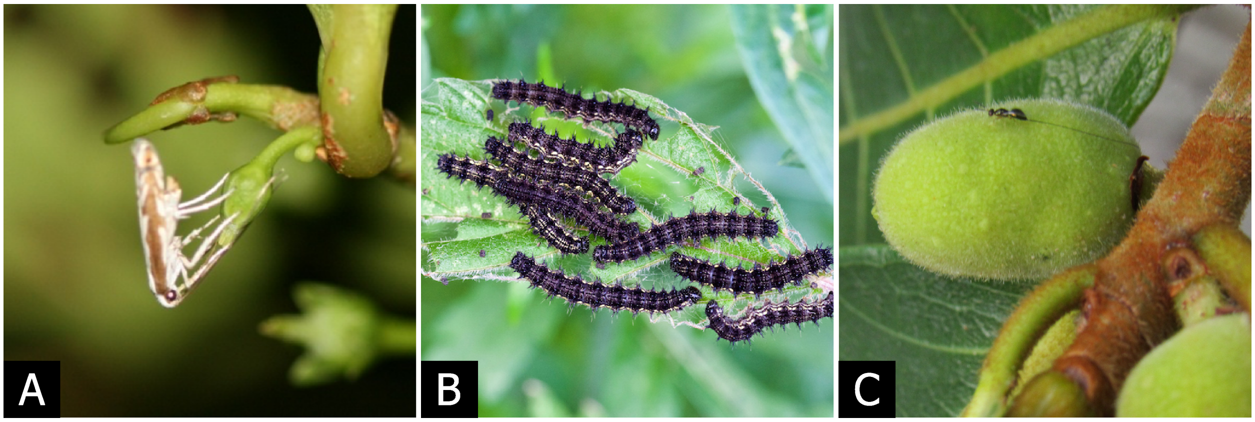


Figure 2.1 Examples of ecological interactions that provide biological motivation for this review. **A.** Leafflower moth (*Epicephala* sp.) pollinating its leafflower host (*Glochidion grayanum*, syn. *Phyllanthus grayanus*), an example of a brood pollination mutualism (Tahiti, French Polynesia, photo by David Hembry from Hembry et al. 2012). **B.** Parasitic caterpillar larvae of *Aglais urticae* on their host plant *Urtica dioca* (photo by Niklas Janz, used with permission). **C.** A fig syconium (*Ficus popenoei*) and a non-pollinating parasitic female wasp (*Idarnes* sp.) that is an antagonist of the fig/fig-wasp mutualism (photo by Kevin Quinteros, used with permission).

(three motivating examples are shown in Figure 2.1). Fossil data and phylogenetic inferences suggest that interspecific interactions, especially symbioses, can persist for millions, and sometimes hundreds of millions of years (*e.g.,* McKinney, 1995, Compton et al., 2010, Zeng and Wiens, 2021, Labandeira et al., 1994), effectively coupling the evolutionary histories of even long-diverged clades. Furthermore, organisms engaged in non-symbiotic and more diffuse interactions—*e.g.,* via pollination, seed dispersal, and predation—are often phylogenetically conservative in the taxa with which they interact (Thompson, 1994, Rezende et al., 2007). Such ecological interactions may also leave signatures in the macroevolutionary patterns of coupled lineages (Jablonski, 2008, Weber and Agrawal, 2014, Hembry and Weber, 2020).

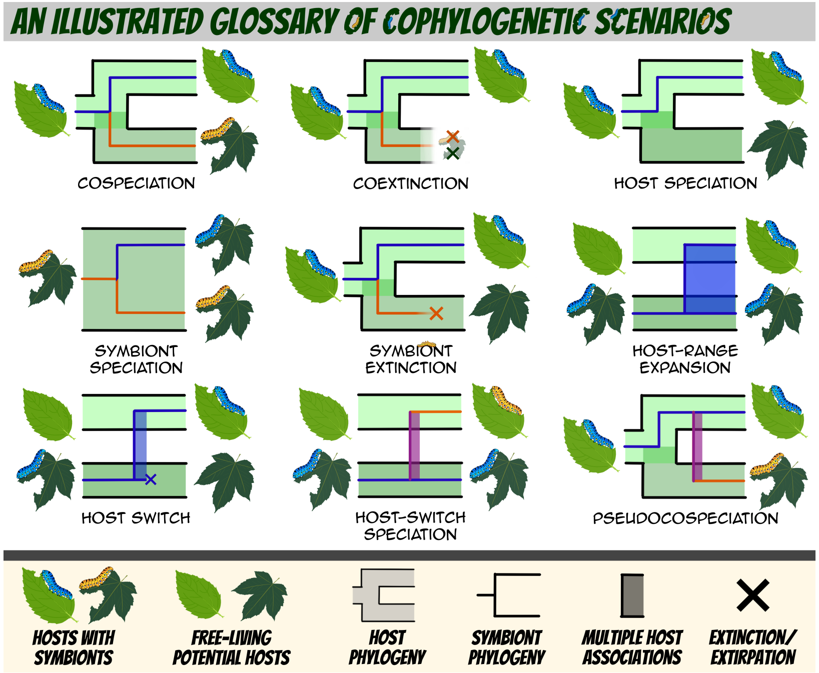


Figure 2.2 Cophylogenetic systems can result in a number of different event types or scenarios that can manifest in different patterns of the host and symbiont phylogenies. These are illustrated using a hypothetical host-parasite example based on figure 1.1 from Page (2003). We provide detailed definitions of each scenario in the **Glossary of Cophylogenetic Scenarios**.

Over a century ago Heinrich Fahrenholz (1913), a parasitologist and lice specialist, first proposed that the evolutionary history and taxonomy of parasites should closely reflect that of their hosts. The rise of molecular phylogenetics in the late twentieth century made it possible to test Fahrenholz's hypothesis within a statistical framework. Researchers found that while the topologies for many host and parasite trees may be similar, they rarely matched perfectly (*e.g.*, Hafner and Nadler, 1990, Cruaud et al., 2012) because events such as host-switching and symbiont speciation can result in incongruencies between the evolutionary history of host and symbionts. These findings spurred biologists to invent cophylogenetic methods, a new class of computational approaches that untangle *why* ecologically linked lineages exhibit similar or different diversification patterns.

2.2 Glossary of Cophylogenetic Terms

Cospeciation – coordinated speciation of both the host and the symbiont.

Coextinction – simultaneous extinction of both host and symbiont lineages.

Host speciation – speciation of the host lineage while the symbiont remains a single species associated with only one of the descendant host lineages. Also called 'missing-the-boat'.

Symbiont speciation – speciation of the symbiont lineage resulting in a single host species associated with two symbiont species. Also referred to as a duplication.

Symbiont extinction – independent extinction of the symbiont lineage, while the host lineage survives without an associated symbiont. Also referred to as a loss.

Host-range expansion – symbiont dispersal to a new host species while also maintaining its association with the ancestral host.

Host switch – transfer of the symbiont species from one host lineage to another, resulting in extirpation on the ancestral host. Also referred to as a 'host shift'.

Host switch speciation – speciation of a symbiont lineage where one descendant lineage transfers to a new host species, while the other remains associated with the ancestral host.

Pseudocospeciation – a match between host and symbiont phylogenies that is not due to host speciation (*i.e.,* not true coordinated cospeciation). Sometimes called phylogenetic tracking.

Cophylogenetic methods have historically been separated into two categories: global fit methods that assess the overall congruence between two phylogenetic trees, and event-based methods that map the symbiont phylogeny onto the host phylogeny using discrete events (Page, 2003). Explicitly, global-fit methods use summary statistics to compare two complete phylogenetic patterns and measure cophylogenetic congruence

(Hafner et al., 1994, Legendre et al., 2002, Balbuena et al., 2013), whereas event-based methods use predefined scoring systems (*e.g.,* parsimony) to search for optimal arrangements of historical events that can reconcile the topologies of two interacting clades (Brooks, 1985, Page, 1994, Ronquist, 2003, Conow et al., 2010). However, since the terms "event-based" and "global-fit" were coined, newer approaches have proliferated that use generative models and statistical inference to shed light on the underlying evolutionary processes responsible for producing cophylogenetic data (*e.g.,* Huelsenbeck et al., 2000, Braga et al., 2020, Dismukes and Heath, 2021). While model-based methods can result in estimates of historical events, the use of explicit stochastic models accounting for historical associations sets them apart from other "event-based" methods. In this review, we adopt an alternative set of categories for cophylogenetic methods: (1) *pattern-based methods* include approaches historically categorized as "global-fit" methods in addition to other previously uncategorized methods (Section 2.3.1), (2) *event-scoring methods* are event-based but model-free optimization methods (Section 2.3.2), and (3) *generative model–based methods* use probabilistic models to describe the generating process underlying cophylogenetic patterns (Section 2.3.3). While all cophylogenetic methods are typically informed by the same data sources—two phylogenies and a matrix that denotes which tips interact—the methods differ widely in their assumptions and the degree to which they are suitable for different types of cophylogenetic systems or taxonomic levels. As a consequence, choosing which methods are appropriate for a given system and set of biological hypotheses is a difficult task.

This review has two primary objectives: (1) to help researchers understand which cophylogenetic methods are best suited to their cophylogenetic systems and questions, and (2) to identify gaps in understanding or capability under our current means for cophylogenetic analysis to stimulate future development. We begin by summarizing recent developments among pattern-based and event-scoring methods, which have been reviewed in various contexts previously (Page, 2003, Stevens, 2004, De Vienne et al., 2013, Blasco-Costa et al., 2021), in addition to a newer emerging class of generative model-based cophylogenetic methods. We first define the major features of the cophylogeny problem, and then survey assorted available pattern based methods, event-scoring methods, and generative model-based methods, while assessing the assumptions, strengths, and weaknesses for each method family. We continue by discussing the types of cophylogenetic systems and data, dividing these into broad categories that can be analyzed using similar methods. Lastly, we survey aspects of cophylogenetic methodology and analysis that are important, yet underexplored areas of research in need of further study.

2.3 Current methods for cophlogenetic data

Biologists expect that many intimate ecological interactions generate predictable cophylogenetic patterns of topologies, divergence times, and associations between two clades. In an extreme example, if symbiont lineages always and only cospeciated in response to the divergences of their host lineages, and never switched hosts afterwards, one would expect to see identical (*i.e.,* completely congruent) topologies and branch-length distributions for both hosts and symbionts (Fahrenholz, 1913, Brooks, 1985, Paterson and Gray, 1997). Such a cophylogenetic history would generate a strong pattern of phylogenetic congruence, as pairs of random trees will seldom perfectly match in topology (Felsenstein, 1978). However, since even strongly associated clades do not exhibit total phylogenetic congruence, this raises the question: how can we measure the strength of cophylogenetic signal using these imperfect patterns?

Phylogenetic comparative methods explain patterns of trait variation among taxa as an outcome of descent with modification. Cophylogenetic methods are similar, except they investigate how trait variation patterns may be correlated both within and between clades.

Here, we focus on cophylogenetic methods that explain the distributions of ecological interactions between species-pairs from two separate clades. These cophylogenetic methods look at the codistribution of taxon interactions, clade memberships, divergence times, and/or associated traits to ask whether the topological similarities between two phylogenies are due to a shared evolutionary history of ecological interaction. Exactly what questions a researcher asks depends on the biology of the cophylogenetic system, the question at hand, available data, and what available methods are suitable for producing meaningful insights (see the sidebar: **Properties of Different Cophylogenetic Systems**).

Most cophylogenetic analyses require three pieces of information as input: two phylogenies and an interaction matrix. Time-calibrated phylogenies are generally preferred so that rates of evolutionary change or event times can be placed within a broader temporal and geological context, although trees with branches measured in relative units of time or in molecular substitution events can be used when data for calibrating trees to absolute time are unavailable. The interaction matrix defines which extant taxa between the two clades interact as informed by field observations, experimental evidence, or previously published reports. Most methods assume that the recorded interactions within a clade are fundamentally analogous in kind. For example, many butterfly species parasitize one set of plants as larvae, and pollinate a different set of plants as adults, but these antagonistic and mutualistic interactions evolve by different underlying mechanisms and should not be treated as directly comparable. Interaction matrix cells are usually recorded as discrete values (*e.g.,* present/absent host use), but some datasets record experimentally verified potential interactions or continuous values (*e.g.,* feeding intensity). Some methods accommodate other forms of auxiliary information, such as chemical volatiles, geographical range data, taxon morphology, etc. Our overview of methods, however, will focus primarily on those analyzing two-clade presence/absence interaction data.

In this section, we describe three classes of methods: pattern-based methods, event-scoring methods, and generative model–based methods. Pattern-based methods decompose the host and symbiont phylogeny into phylogenetic distance matrices that can then be used to test the extent to which the interactions could have been produced due to chance alone. Event-scoring methods map a symbiont phylogeny onto a host phylogeny using the classic cophylogenetic events (*e.g.,* host-switching, symbiont speciation, cospeciation). Each of these events is assigned a cost to determine the lowest cost mapping. The newer generative models encompass and use probabilistic models to describe the processes that produce observed cophylogentic patterns.

Table 2.1 Examples of cophylogenetic scenarios (Figure 2.3)

Scenario	Typical # of interactions per host taxon	Typical # of interactions per symbiont taxon	Cophylogenetic signal	Host switching	Putative Example
A	1	1	High	Low	Pocket gophers & lice[A]
B	1+	1	High	High	Coronaviridae & mammals[B]
C	1	1	Low	Medium	Anolis & plasmodium[C]
D	1+	1	Medium	High	Figs & fig wasps[D]
E	1	1	Low	High	Columbidae & feather lice[E]
F	1+	1+	High	Low	Angiosperms & Nymphalini[F]

[A] Hafner et al. (1994); [B] Anthony et al. (2017); [C] Charleston and Perkins (2003); [D] Cruaud et al. (2012), Satler et al. (2019); [E] Doña et al. (2017); [F] Janz et al. (2001)

2.3.1 Pattern-based methods

Pattern-based methods test whether two interacting phylogenies are more similar than would be expected by chance. Methods in this family differ primarily in how they operationalize the terms ‘similarity’ and ‘chance’. Similarity is generally defined through a test statistic that measures distance between two phylogenetic distance matrices, whose rows and columns are ordered to match the associations between the host and symbiont clades. Chance is frequently defined in terms of a null distribution of cophylogenetic

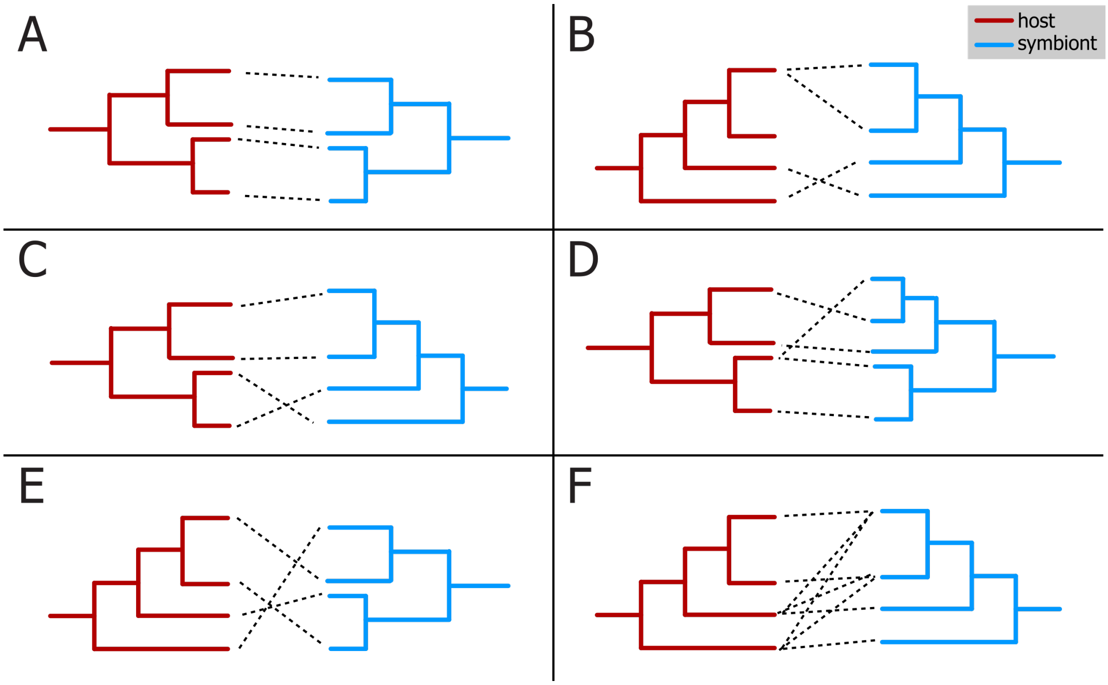


Figure 2.3 Six possible cophylogenetic scenarios with host phylogenies in dark red (left tree) and symbiont phylogenies in light blue (right tree). Interactions are shown by dashed lines. Putative examples are provided in the table below. We provide these simplified examples to recognize that the biological processes generating macroevolutionary patterns vary substantially depending on the system and on the taxonomic scale (see also Figure 2.4). Thus, it makes little sense to treat all these data the same analytically. See Table 2.1 for more information.

similarity scores for pairs of trees with randomly distributed interactions. Typically, cophylogenetic data are transformed into a simplified structure to measure similarity using classical statistical transformations and tests.

2.3.1.1 Distance matrix tests

Hafner and Nadler (1990) used a Mantel test to show that the phylogenetic distance matrix elements for a pair of host and symbiont clades (pocket gophers and pocket gopher lice, in their study) are more correlated than expected by chance. To measure the significance of the correlation, the test also simulates a null distribution of permuted matrices without cophylogenetic correlation structure. If the correlation between the two original distance matrices exceeds that for most permuted matrices, the null hypothesis of no cophylogenetic correlation may be rejected. Mantel tests extremely simple to apply, but they may suffer from low statistical power (*i.e.*, large trees needed to reject the null hypothesis) and inflated Type-I error in phylogenetic settings (Harmon and Glor, 2010).

In a subsequent study on gopher-lice interactions, Hafner et al. (1994) applied the Wilcoxon signed-rank test (Wilcoxon, 1992) to test whether the lengths of congruent branches between two host and symbiont phylogenies were consistently shorter or longer for the symbiont clade. The sum of signed-rank scores is first computed for the original phylogenies, and then a null distribution for the statistic is simulated via permutation, as above. In the case of Hafner et al. (1994), the Wilcoxon signed-rank test showed that, the molecular branch lengths (measured in expected number of nucleotide substitutions per site) in the parasite tree significantly outranked (were consistently longer than) host molecular branch lengths, implying that the lice lineages experienced greater amounts of genetic change when compared to their corresponding host lineages.

Both the Mantel and Wilcoxon signed-rank tests can typically be applied only to datasets with extremely specialized (one-to-one) relationships and that contain an equal number of taxa in both phylogenies. That said, Hommola et al. (2009) introduced a variant of the Mantel test that accommodates multiple interactions per parasite and/or host taxon.

Nonetheless, distance matrix methods often require that some taxa and/or interactions must be artificially removed from the analysis to conform with the test requirements.

2.3.1.2 Global-fit tests

Many global-fit tests rely upon principal coordinate analysis (PCoA) to transform the host and symbiont phylogenetic distance matrices into new coordinate systems, so that the overall similarity between the phylogenies may be compared (though see). As with the distance matrix tests, these tests also require a host phylogeny, a symbiont phylogeny, and an interaction matrix as input; unlike distance matrix tests, however, the two phylogenies may differ in size, and symbionts and hosts may have multiple interactions.

The first published global-fit method, Parafit (Legendre et al., 2002), computes an eponymous 'global fit' statistic by transforming the phylogenetic distances for hosts and symbionts into separate principal coordinate systems that are then aligned through an interaction matrix. Parafit next measures the global fit of the cophylogenetic system using a linear algebra framework (Legendre et al., 1997). The global fit is maximized when the host and symbiont matrices are perfectly identical in branch lengths and topologies among interacting species, and decreases as the two phylogenies grow incongruent. Parafit has also been extended to test for cophylogenetic patterns between three phylogenies (tritropic systems; Mramba et al., 2013). Nooney et al. (2017) refined the tritrophic hypothesis testing framework of Mramba et al. (2013), which can potentially generalize to larger numbers of interacting clades and richer cophylogenetic topologies, including networks.

Whereas Parafit was designed to test for correlations between the host upon symbiont phyogenies, the Procrustes Approach to Cophylogeny (PACo; Balbuena et al., 2013) adapts this method to test for the dependence of parasite phylogeny on a host phylogeny. To do so, PACo uses Procrustean superimposition to align the phylogenetic distances among parasites to those for the corresponding host distances. Hutchinson et al. (2017)

generalized PACo to allow for symmetrical dependencies between clades to model mutualistic host-symbiont relationships. Once both phylogenies share the same coordinate system, global fit of the cophylogenetic pattern is measured as the residual sum of squares, m^2, with small values indicating greater congruence. Signal for global and interaction-specific methods are tested using permutation and jackknifing approaches.

Because global-fit scores can be difficult to interpret biologically, Balbuena et al. (2020) introduced the Random TaPas method to measure cophylogenetic congruence in terms of patterns consistent with cospeciation and other coevolutionary events that are implied by a tanglegram. Random TaPas subsamples tanglegrams from the full distribution of cophylogenetic interactions, computes a distribution of significance scores against those subsamples using other global-fit methods (Legendre et al., 2002, Schardl et al., 2008, Balbuena et al., 2013), and then measures if frequencies of cospeciation exceed null expectations.

In addition to quantifying the global fit of cophylogenetic patterns, most global-fit methods (Legendre et al., 2002, Balbuena et al., 2013, 2020) can also measure the cophylogenetic significance of individual taxa and/or interactions.

2.3.1.3 Tree shape tests

Tree distance metrics, which are often used to measure difference among trees that share the same set of taxa—*e.g.,* different Bayesian posterior tree samples – can also be applied to study symbiont tree congruence. The program COMPONENT (Page, 1994), for example, measures how many taxa must be removed from the host and the symbiont phylogenies to produce a perfect cophylogenetic pattern. Because no single tree shape metric perfectly measures all differences between trees, Avino et al. (2019) tested for cophylogenetic signal using a panel of 18 tree shape metrics against data simulated under different cophylogenetic hypotheses. They found that data generated under different migration rates, speciation rates, and cophylogenetic event probabilities produced predictable tree shape metrics; for example, cophylogenetic Robinson-Foulds distances (Robinson and Foulds, 1979) predictably increased as host-switching rates in symbionts increased.

2.3.1.4 Take home message

Pattern-based methods benefit from fact that they are not explicit event-based models. These methods remain computationally efficient for large datasets with many interactions, while still being able to identify patterns that are consistent with a wide range of hypotheses of codiversification. Because these methods are fast, it is common practice to apply multiple pattern–based methods to establish a consensus for cophylogenetic signal. In doing so, the consensus is stronger if signal is shared across the three types of methods outlined above. Importantly, identifying significant signals for most pattern–based methods relies on null hypothesis testing. Null model definition, data set size, and the interpretation of a rejected null hypothesis should all be approached with extreme care. Because pattern-based methods do not provide an explicit generative model to produce cophylogenetic datasets, is is often difficult to interpret what the cophylogenetic test statistics mean biologically. Whether or not this meaning is relevant depends on the question at hand, which we discuss more in the final section.

2.3.2 Event-scoring methods

Event-scoring methods attempt to reconcile discordance between host and symbiont trees. These have previously been called event-based methods; however, we have separated them here to draw a distinction between more pattern-based and parsimony methods (this section), and event-based methods that rely on probabilistic generative models (Section 2.3.3). Event-scoring methods seek to connect cophylogenetic data to the events that could have produced them, but differ in their methodology. All event-scoring methods typically use a defined set of events to describe the possible ways the host and symbiont lineages evolved together. Reconstructed events—such as cospeciation, host-shifts, symbiont speciation (also called duplication), and symbiont extinction (also called loss)—are used to explain how the host and symbiont lineages evolved with respect to their shared cophylogenetic history.

2.3.2.1 Brooks Parsimony Analysis

Brooks Parsimony Analysis (BPA) was one of the first quantitative methods examining cophylogenetic systems in an algorithmic way. To do this, BPA treats each symbiont's relationship with specific hosts as a binary character of the host and then finds the most parsimonious mapping of that character onto the host tree (Brooks, 1981). In this method, the symbiont phylogeny is re-coded into a set of binary characters using additive binary coding to convert the cophylogenetic dataset into a phylogeny with character states. To do this, half of these characters are used to record the presence or absence of the symbiont taxa (*i.e.,* the tips of the symbiont phylogeny), and the remaining characters each represent interior nodes of the symbiont phylogeny. This mapping is then used to determine which events occur in which parts of the host and symbiont phylogenies. BPA is able to accommodate complexities such as hosts with multiple symbionts, and makes few assumptions about the underlying process. For example, there is no *a priori* assumption that cospeciation is the most likely event to occur using BPA (Brooks et al., 2015). Patterns of homoplasy and homology under the original BPA method can be difficult to interpret, however, as it lacks important cophylogenetic event types (*e.g.,* symbiont speciation; Siddall and Perkins, 2003). BPA has also widely been criticized for inaccurately counting events relating to lineage sorting and host switches (Siddall and Perkins, 2003, Brooks et al., 2004).

2.3.2.2 Generalized Parsimony Reconciliation

Other event-scoring methods take a different approach to reconciling a symbiont tree with its host phylogeny using events, either attempting to maximize the number of cospeciation events (Page, 1994) or assigning relative costs to each event and minimizing the total cost of all events (Conow et al., 2010, Charleston and Page, 2002, Merkle and Middendorf, 2005). Rather than determining events after the analysis, these methods define events prior to the analysis and assign each of them a cost. Specialized algorithms are used to map the symbiont phylogeny onto the host phylogeny until a lowest-cost mapping is found. Assigning sensible costs for events is crucial; for example, with a no-cost cospeciation event one assumes that cospeciation is quite likely to occur in their particular system. Indeed, the majority of these methods assign cospeciation a low cost by default, which implies that, if users do not change the default costs, cospeciation (*i.e.,* concordant host and symbiont phylogenies) is the most likely event to occur. As others have noted (De Vienne et al., 2013, Brooks et al., 2015), this can be a flawed assumption as concordance can be produced in a number of different ways—some cophylogenetically and some not.

One notable challenge researchers face when applying generalized parsimony reconciliation methods is choosing and specifying values for event costs for their study system. Methods have recently been introduced to assist in determining event costs (Baudet et al., 2015, Santichaivekin et al., 2021). The approach by Baudet et al. (2015) uses a simple birth-death model to generate simulated data and approximate Bayesian computation to determine appropriate costs for events. This, however, comes with the caveat that the costs will depend on the assumptions of the simulation model (for example, by default there is no host extinction in the model used in Baudet et al. 2015). Additionally, these methods can produce more than one optimal solution to mapping the symbiont tree onto the host tree, and in some cases this number of solutions can be exceedingly high (Hypša, 2006), making it difficult to interpret the results. Despite their challenges, these methods are very computationally efficient and can be used in combination with multiple data sources such as biogeographic data (Merkle and Middendorf, 2005) and time-calibrated phylogenies (Conow et al., 2010).

2.3.2.3 Finding events without cophylogenetic methods

Evolutionary biologists in some cases have sought evidence for the events inferred by event-scoring methods (cospeciation, duplication, host-shifts, and extinction) using cophylogenetic methods in tandem with non-cophylogenetic approaches, such as biogeographic or phylogenetic comparative methods (Althoff et al., 2012, Hembry et al., 2013), or without cophylogenetic methods altogether (Smith et al., 2008, Luo et al., 2017). The choice to use multiple lines of evidence or non-cophylogenetic methods may be motivated by concern over how to rigorously assign costs to unique events of interest (Luo et al., 2017), concern over how to distinguish phylogenetic tracking from true cospeciation (Althoff et al., 2012), or by a focus on a phylogenetic scale that is not amenable to cophylogenetic analyses (*e.g.,* a single putative cospeciation event; Smith et al. 2008). The use of multiple lines of evidence to identify events can be a powerful approach for testing hypotheses about cophylogenetic history, but it also testifies to the limitations inherent in using some currently available event-scoring methods on their own.

2.3.2.4 Take-home message

Event-scoring methods reconstruct the historical sequence of events that produced a cophylogenetic pattern, but they are heavily reliant on user-specified costs or post-hoc hypotheses. Recent work has introduced useful and highly efficient methods to determine meaningful event costs. Reconciliation analysis methods have seen widespread use and theoretical development in recent years (Drinkwater and Charleston, 2014, Drinkwater et al., 2016, Althoff et al., 2012, Flynn and Moreau, 2019)

2.3.3 Generative model-based methods

Understanding the evolutionary forces responsible for producing observed cophylogenetic patterns is the ultimate goal of researchers investigating the codiversification of interacting clades. To this end, statistical models that describe how host-symbiont associations change over time to generate present-day interactions are extremely useful. Building off of previous work that introduced statistical tests for cophylogenetic congruence (Huelsenbeck et al., 1997), the study by Huelsenbeck et al. (2000) was among the first to describe a probabilistic model capable of generating cophylogenetic data. Their model, in which a Poisson process generates host-switch events, assumes a strict one-to-one matching of hosts and symbionts, and is appropriate for interactions that exhibit such highly specialized partnerships. Though strong assumptions limit the range of cophylogenetic questions that this model can address, Huelsenbeck et al. (2000) nevertheless laid the foundation for future development of generative models for inferring cophylogenetic processes.

Newer developments in model-based approaches for cophylogenetic analysis take inspiration from two other types of models that describe correlated evolution: (1) the evolution of gene trees within species trees and (2) phylogenetic diversification driven by biogeographical processes. Such models can provide reasonable analogs to cophylogenetic patterns of species interactions, as demonstrated by recent studies applied to host-symbiont systems to explicitly estimate cophylogenetic parameters (Groussin et al., 2017, Satler et al., 2019, Braga et al., 2020, 2021).

2.3.3.1 Modeling species interactions

Two recent studies exemplify different strategies for modeling the evolution of host-symbiont interactions. In the first (Satler et al., 2019), the process producing cophylogenetic patterns between sympatric Panamanian strangler figs and their pollinating fig wasps was modeled as the combination of four evolutionary processes: host-switching, cospeciation, symbiont speciation, and symbiont extinction. This approach characterizes the evolutionary process similarly to the event-scoring methods (Section 2.3.2), but in a probabilistic framework. The second example study (Braga et al., 2020) models gains and losses of specific host taxa along the symbiont phylogeny to produce the observed present-day associations among parasitic Nymphalini butterfly species and angiosperm families. One major difference in the methods used by these studies is the treatment of events. Specifically, in Satler et al. (2019) the process happens along the host tree and the events are necessary to map the symbiont tree onto the host tree, whereas in Braga et al. (2020) the process happens along the symbiont tree and hosts are modeled as characters that evolve interdependently because phylogenetic distances between hosts affect the probability of host gains.

It may be reasonable to assume that tightly linked, obligate mutualisms—like fig trees and their wasp pollinators—are governed by a process that resembles the evolution of gene families within species phylogenies. Satler et al. (2019) used this reasoning to apply an existing gene-tree/species-tree method to elucidate the shared histories of host and mutualist lineages. Association patterns between genes and species are generated by a coevolutionary process that includes codivergence, gene duplication, gene loss, and gene transfer (Szöllősi et al., 2012, Szöllősi et al., 2013). For a system like the fig and fig-wasp mutualism, these events are analogous to cospeciation, symbiont speciation, symbiont extinction, and host switching. Satler et al. (2019) applied a gene-family evolution model (using the program ALE; Szöllősi et al. 2012 and Szöllősi et al. 2015) to reveal that a history of frequent host switching was responsible for the current associations observed in Panamanian strangler figs and their pollinators. The application of these models to the coevolution of mutualistic systems was initially introduced by Groussin et al. (2017) to estimate cospeciation and host switching in the evolutionary associations of mammals and their gut microbes. These studies (Groussin et al., 2017, Satler et al., 2019) demonstrate the potential for model-based approaches to yield deeper insights into the macroevolution of ecological interactions.

Braga et al. (2020) developed an approach for understanding host-repertoire evolution in clades of parasitic lineages that uses models adapted from the field of historical biogeography to describe species-area distributions (Ree et al., 2005, Landis et al., 2013). In this method, each parasitic lineage has a host repertoire, which is the set of possible hosts a parasitic species can exploit. When assessing ecological interactions using a biogeography-based model, the host repertoire corresponds to the set of possible areas that make up the geographic range of a species (Braga et al., 2020). Under this model, a parasitic lineage has a 'realized' and a 'fundamental' host repertoire, such that a range expansion to a new host species begins with a gain in the parasite's ability to exploit the new host (the new host becomes part of the fundamental host repertoire), followed by the parasite's switch to using the new host (the new host species is part of the realized host repertoire); these non-host, potential, and realized associations evolve in through an ordered two-step process, imbuing the host repertoire with macroevolutionary memory (Goldberg and Foo, 2020). In addition, phylogenetic distance-dependent host-range expansion rates are modeled to account for how closely related any new host is to all hosts currently used by a parasite. By modeling changes in the assemblage of host species exploited along the symbiont lineages, this method enables estimation of ancestral host repertoire, providing insights into the evolution of ecological interactions over time (Braga et al., 2021).

While these few applications of generative models highlight the potential of probabilistic methods for yielding significant insights into cophylogenetic patterns, it is clear that models should be selected carefully, with full awareness of their assumptions, limitations, and advantages. Certain models are clearly more suitable for analyzing certain types of cophylogenetic systems than others. For example, it may not be appropriate to adapt a gene-family evolution model to study the codiversification of host-parasite systems since these models do not allow hosts to escape an association with a parasite (*i.e.,* a species must be associated with at least one gene). Moreover, gene-tree/species-tree models restrict symbiont species only one host-species associate at a time. While there are workarounds for this limitation (*e.g.,* a symbiont with two hosts can be represented as two sister lineages in the symbiont tree, as in Satler et al. 2019), besides being biologically unrealistic in many systems, it is unclear how or if such transformations bias parameter estimates. The approach of Braga et al. (2020) was designed to allow symbionts to have multiple hosts in their repertoires at any given time, making this model suitable for studying the associations of generalist symbionts. However, their implementation does not allow host diversification and is currently appropriate only for systems where all host taxa are older than the symbiont clade. This restriction also means that analyses under this model do not estimate parameters associated with host evolution, though it may be possible to flip the focus to evaluate the symbiont repertoire of a clade of hosts.

Generative models and statistical inference are powerful tools for understanding how evolutionary processes influence ecological interactions. Very few studies, however, have investigated the accuracy and consistency of model-based cophylogenetic analysis. Host-symbiont association data generated under complex macroevolutonary models are essential if we want to understand the performance and limitations of all cophylogenetic methods. Recently introduced simulation tools (*e.g.,* Dismukes and Heath, 2021, Braga et al., 2020, Maliet et al., 2020) for generating cophylogenetic data under explicit coevolutionary models will yield deeper knowledge about the ways in which generative models can be applied to coevolutionary questions. For instance, Braga et al. (2020) used simulated data to show that their host-repertoire model could reliably estimate the true simulating rates of host repertoire evolution, and the effect of phylogenetic distances on host gain rates. Expanding the range of methods and statistical models evaluated using simulated data will be an important advancement for the field of cophylogenetics.

2.3.3.2 Take-home message

The types of inter-species interactions existing in nature are extraordinarily diverse and a "one-size-fits-all" model is neither possible nor desirable. Nevertheless, explicit models of codiversification and species interactions are necessary to gain greater understanding of the evolutionary forces responsible for generating present-day host-symbiont associations. Critically, future generative models must be motivated by the biological systems they seek to describe, with clearly defined assumptions and outcomes. Furthermore, methods integrating additional ecological (Clayton et al., 2015), biogeographical (Althoff et al., 2012, Hembry et al., 2013), genomic (Cai et al., 2021), and paleontological (Baets et al., 2021) data have the potential to enhance cophylogenetic analyses. Such data can provide more context for past and present host-symbiont associations by placing them in an environment, a space, and a time.

Designing biologically motivated models conditioned on a wide range of data sources, however, inherently leads to complex, parameter-rich processes. Statistical inference under complex macroevolutionary models is challenging to execute (*e.g.,* requiring detailed specification files and long analysis times). When complete, such analyses result in numerous parameter estimates that may be highly uncertain and/or poorly identifiable, making the output difficult to summarize and interpret. For this reason, it is essential that researchers applying these methods carefully consider the assumptions of the model and methods, clearly define their *a priori* hypotheses, and present results that adequately communicate parameter uncertainty. With great models comes great responsibility.

2.3.4 Methods summary

The three categories of cophylogenetic methods discussed above consider different types of cophylogenetic systems and interactions. Table 2.1 lists software tools and approaches that have implemented methods we have reviewed. When choosing a method to analyze cophylogenetic data, it is important to consider the nature of the ecological interactions, the phylogenetic data available, and with how many taxa symbionts and hosts associate.

Generative model–based methods and pattern–based methods both make use of probabilities in evaluating data. Event-scoring methods make use of parsimony scores. A major advantage of using probabilities rather than parsimony scores is the ability to evaluate uncertainty. Parsimony scores are limiting in the sense that they allow for the possibility of multiple lowest cost mappings with no way to determine which of these is the most plausible.

All of the cophylogenetic methods we described earlier are inference methods, in the sense they allow biologists to reconstruct how relationships between hosts and symbionts evolved, whether certain cophylogenetic relationships are phylogenetically conserved or clustered, and at what rates various cophylogenetic events occurred. However, only generative model–based methods are capable of simulating synthetic datasets. Making use of simulated data will enable detailed analysis of the performance of cophylogenetic inference under realistic evolutionary scenarios. Such studies will determine how accurately each method can estimate certain values from cophylogenetic data, how robust they are to different types of error, and the computational resources and time they require. When designing cophylogenetic studies, researchers can combine this information with detailed knowledge of the biological properties of their system and explicit *a priori* hypotheses to choose the most appropriate methods that will lead to reliable conclusions.

Table 2.1 Summary of reviewed cophylogenetic methods.

Method	System	Phylogeny	Interactions
PATTERN-BASED STATISTICS			
Mantel test[1]	S	B,D	1
Wilcoxon test[2]	S	B,D	1
Parafit[3]	S	B,D	M
MRCAlink[4]	S	B,D	M
PACo[5]	D,S	B,D	M
Random TaPas[6]	D,S	B	M
EVENT-SCORING METHODS			
BPA[7]	D	T	1,M
TreeMap[8]	D	T	1
Jane[9]	D	B,D	M
Tarzan[10]	D	D	M
COALA[11]	D	B,D	M
Jungles[12]	D	B,D	1
eMPRess[13]	D	B,D	M
DIVA[14]	D	T	M
CoRe-PA[15]	D	D	M
GENERATIVE MODEL–BASED METHODS			
Bayesian host switching[16]	D	D	1
DEC[17]	D	D	M
ALE[18]	D	D	1
Host repertoire evolution[19]	D	D	M

Each method differs in how it treats interactions (D: directional [*e.g.,* parasite on host]; S: symmetric [*e.g.,* plant and pollinator]), the types of phylogenetic data it can use (T: topology, B: topology and branch lengths, D: topology and divergence times), and in how many interactions it permits per taxon (1: 1-to-1 only, M: multiple).

References: [1]Hafner and Nadler (1990); [2]Hafner et al. (1994); [3]Legendre et al. (2002); [4]Schardl et al. (2008); [5]Balbuena et al. (2013), Hutchinson et al. (2017); [6]Balbuena et al. (2020); [7]Brooks (1981, 1990); [8]Page (1994); [9]Conow et al. (2010); [10]Merkle and Middendorf (2005); [11]Baudet et al. (2015); [12]Charleston (1998); [13]Santichaivekin et al. (2021); [14]Ronquist (1995); [15]Merkle et al. (2010); [16]Huelsenbeck et al. (2000); [17]Ree et al. (2005); [18]Szöllősi et al. (2012, 2015); [19]Braga et al. (2020)

2.4 Classifying cophylogenetic systems

The cophylogenetic methods toolbox contains a wide variety of approaches. Because these methods differ in their assumptions, it should come as no surprise that some methods are more applicable to certain classes of cophylogenetic problems over others. In this section, we characterize how distinct types of cophylogenetic systems interface with currently available cophylogenetic methods to guide biologists towards selecting the right 'tools for the job' in their own research. Table 2.1 summarizes how we view which methods are appropriate for which systems and datasets. In this section, we first describe the different properties of cophylogenetic systems which may impact their patterns and dynamics of codiversification. Second, we describe qualities of datasets. Third, we review what system and dataset properties are most congruent with which cophylogenetic methods. Finally, we discuss situations for which cophylogenetic methods are inappropriate and call for better awareness of what cophylogenetic biology and the rest of evolutionary ecology can learn from and inform each other.

2.4.1 Properties of cophylogenetic systems

Evolutionary ecologists classify species interactions (of which cophylogenetic systems are a subset) along a variety of axes (Thompson, 1994, Ollerton, 2006). Four of these axes represent properties that have important consequences for patterns and processes of codiversification, as well as for choice of cophylogenetic methods: types of interaction (*e.g.*, antagonism [including parasitism], mutualism, commensalism), whether one or both taxa depend on the interaction, degree of specialization (number of partner taxa), and transmission ecology (vertical or horizontal). Figure 2.4 uses a hypothetical system to illustrate various ways in which ecological associations may be distributed among host and symbiont taxa.

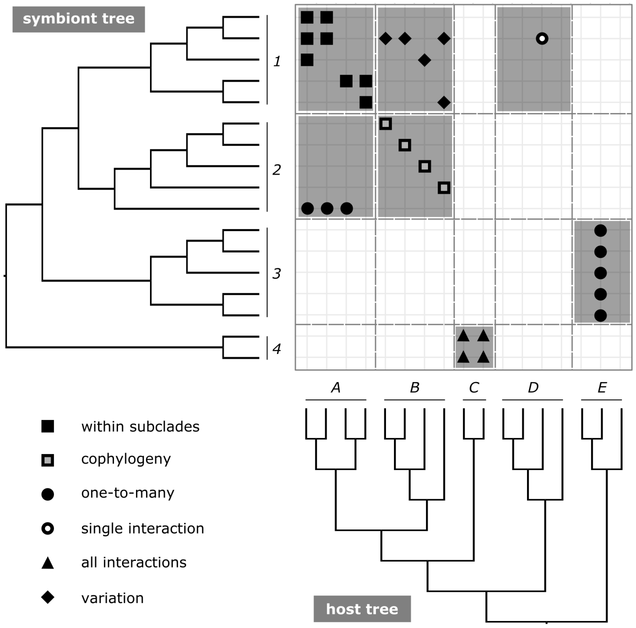


Figure 2.4 Hypothetical cophylogenetic system showing the different ways species-species interactions might be distributed within clade-clade interactions. The symbiont tree is split in four clades (1-4), and the host tree comprises five clades (A-E). In the interaction matrix, species-species interactions are represented by different point shapes, while clade-clade interactions are represented by grey-shaded rectangles. Specialization can be seen in several levels: 1) species-species – within interaction between clades 2 and B, each symbiont species interacts with a unique host species; 2) clade-species – all species in clade 3 are specialized to the same species in clade E; 3) species-clade – a single species in clade 2 interacts with several species within clade A; 4) subclade-subclade – each subclade within clade 1 only interacts with one subclade of A; 5) clade-clade – all species within clades 4 and C interact; all but one of the interactions of the sister clades 1 and 2 are with hosts from the sister clades A and B, and the exception is a taxonomically rare interaction, where a single species-species interaction occurs between clades 1 and D. Only the interactions between clades 2 and B show a perfect cophylogenetic pattern (considering only tree topology). When data are summarized at the clade level, different amounts of data (species-species interactions) are reduced to one clade-clade interaction. This might be desirable or necessary in some situations (*e.g.,* to compensate for uneven sampling effort or to reduce dataset size for tractability), but it adds an assumption about data distribution.

2.4.1.1 Types of interactions

Symbiotic interactions between pairs of clades—or other interactions with high biological intimacy—in which individuals of one taxon spend much or all of their life in close physical or physiological proximity to individuals of the other taxon (Ollerton, 2006) are the most tractable for cophylogenetic analyses. Depending on the benefits to individuals of each interacting clade, these interactions can be classified as antagonism (usually parasitism in cophylogenetic systems), commensalism, or mutualism. Although these three classes are each fascinating in their own right, here we primarily consider whether a given biological interaction's underlying processes of diversification are or are not compatible with the assumptions of different cophylogenetic estimation methods.

How two groups of symbionts interact changes which cophylogenetic methods can be applied to their study. Parasitic and commensal interactions are unidirectional, where the survival of one symbiont depends upon that of its host, but not vice versa. Mutualistic interactions are bidirectional in the sense that survival of both symbionts is positively affected by the interaction. (Note, however, that while both symbionts benefit, dependence in mutualisms may be asymmetric if one symbiont is more dependent on the interaction than is the other.) A symmetric method applied to a unidirectional host-parasite system would unintentionally search for the possibility that parasite phylogeny influenced host phylogeny (but see cases where parasites do influence host speciation: Shoemaker et al. 1999). A unidirectional method applied to a symmetric plant-pollinator system, on the other hand, would require that either the plant or the pollinator diversification was not influenced by the diversification of its partner clade. Applying a unidirectional method to a symmetric system twice, with each symbiont playing the role of 'host', can help identify mutually compatible bidirectional inferences, however. In a similar vein, symbiont loss (extinction) may be more biologically realistic in a parasitic or commensalistic system but less so in a given mutualistic system.

Not all systems can be categorized so easily and so applying multiple methods could be both appropriate and instructive. Systems are often defined with respect to a specific trait and/or life history stage, even though the same species may interact in a variety of different ways depending on which trait or life history stage is considered. Interactions between butterflies and angiosperms could potentially be defined as parasitic or mutualistic, depending on whether the question concerns larval herbivory of caterpillars on host plant tissues versus adult pollination of flowers. Similarly, the traits, behaviors, and genes facilitating such ecological interactions might involve either chemical defenses and counterdefenses in the first case, or cues and anatomical features in the second case. One goal of our guide is to aid the practicing biologist in defining the appropriate scope for the ecological interactions, taxa, and traits to be studied productively using cophylogenetic methods.

2.4.1.2 Specialization and generalization

Interacting species in a cophylogenetic system often vary in terms of the number of species with which they interact (Figure 2.4). At one extreme, all species might have one-to-one interactions (specialists – *e.g.,* interactions between clades 2 and B in Figure 2.4) and at another extreme, all species may interact with many species (generalists – *e.g.,* clade 1 in Figure 2.4). Systems in which the average level of specialization is high (in the language of species interaction network ecology, degree k is low for most taxa; Bascompte and Jordano 2013) are most amenable to the assumptions of cophylogenetic analyses. By contrast, there is a threshold above which too many species are too generalized to avoid violating the assumptions of cophylogenetic analyses, and alternative methods should be considered (Hembry and Weber, 2020).

2.4.1.3 Vertical and horizontal transmission

Life history traits may also skew how cophylogenetic variation is generated through either codiversification or host switching events. The most consequential of these for cophylogenetic analyses is whether transmission of symbionts to hosts across generations is vertical (parent-to-offspring, such as the mitochondria of eukaryotes or *Buchnera* symbionts of aphids) or horizontal (symbionts can switch hosts in each generation, as is the case for leafflower moths and leafflowers, or *Symbiodinium* algae symbiotic with corals). In systems where vertical transmission is the rule, cospeciation (contemporaneous codivergence) may be assumed to be common; indeed, some of the clearest phylogenetically congruent systems fall in this category (Hayward et al., 2021). In systems with horizontal transmission, the opposite is the case. Methods assuming that the horizontal transmission of symbionts via host switching is rare may then be prone to over-penalizing incongruencies in cophylogenetic patterns.

2.4.1.4 Other properties of cophylogenetic systems

The interaction data that are used to power cophylogenetic methods are generally based on presence or absence of *actual* interactions that are observed in the field. However, it is possible that many species have the *potential* to interact, but are unable to do so due to mitigating ecological or geographical circumstances. For example, if two similar host species A and B are geographically allopatric, and a parasite species is only sympatric with and uses host A (but not host B), the parasite may realize its potential to use host B if the two entered geographical sympatry. Although binary representations of the presence/absence of interactions are often used in cophylogenetics to simplify data representation and methodological complexity, most ecological interactions are, in truth, more accurately represented by complex sets of phenotypic interactions that are governed by large numbers of quantitative traits.

2.4.2 Dataset properties

2.4.2.1 Taxonomic rank and missing taxa

Cophylogenetic datasets must be carefully defined in an appropriate manner for the question of study. The scale and scope of taxon sets for each symbiont clade is often determined by several practical factors. Ideally, most cophylogenetic studies would use species-level (or even individual-level) datasets, but (1) data with such finescale taxonomic resolution are often not available and/or are difficult to obtain and (2) computational methods generally become less efficient as the number of taxa increases. If data are missing, it is important to consider *what* data points are missing. For example, if the dataset for a clade of highly vagile parasites contains only North American, but not South American, taxa, cophylogenetic methods may underestimate the degree of host switching and/or generalization in the system.

It is also necessarily easier to assemble cophylogenetic datasets that encode symbiotic associations between higher taxonomic ranks (between genera or families) rather than between lower-rank taxa (between species or populations). That said, data associated with higher ranks can variably misrepresent the degree of generalization or specialization in different lineages, which can in turn bias how methods perform. Event-scoring and generative model–based methods that are explicitly designed to model species-level interactions should be used with species-level taxa whenever possible; it is probably better to use only one or few species-level taxa to represent family-level variation, rather that lump all interactions together for a single family-level taxon representative.
Pattern-statistic methods will also be influenced by taxon sampling or the choice of

taxonomic rank, but it is harder to characterize exactly how so based on first principles, since they lack a mechanism for generating datasets.

2.4.2.2 Phylogenetic estimates

All cophylogenetic methods require topologies for both clades. Regarding branch length estimates, many pattern-based methods are solely informed by cophylogenetic distance matrices, and may be applied to unrooted phylogenies for clocklike genes. Event-scoring methods that rely on parsimony, in contrast, make no use of branch length information, and only require topology. Time-calibrated phylogenies are preferred, if not required, for several event-scoring and most generative model-based methods that reconstruct the chronology of cophylogenetic events. Phylogenetic error, both in terms of topology and branch lengths, can easily lead cophylogenetic estimates astray, *e.g.,* by inflating the required number of host-switching events to explain incongruence. Topologies with polytomies that represent uncertainty may be appropriate for pattern statistic methods, but less so for event-scoring and generative model-based methods. Applying the method of choice to a Bayesian posterior of trees can help assess the extent to which the results are sensitive to phylogenetic uncertainty, as well as uncertainty in divergence-time estimates if the phylogenies are time-calibrated. Balbuena et al. (2020) and Pérez-Escobar et al. (2015) have developed specific approaches for dealing with phylogenetic uncertainty using both simulated and empirical data.

2.4.3 When cophylogenetics are not appropriate

Many systems in which two clades interact are not suited for analysis with cophylogenetic methods. The interactions may be too generalized at the species level or too clearly phylogenetically incongruent to meet the assumptions of cophylogenetic methods. Accordingly, it has long been recognized in evolutionary ecology that many situations in which species interactions influence diversification or in which trait coevolution occurs cannot be detected using cophylogenetic methods (Thompson, 1994, 2005) and indeed, that in some if not many cases, coevolution between two clades should not result in phylogenetic congruence (Thompson, 2005, Poisot, 2015). There is a rich and burgeoning literature on ways to use phylogenetic comparative methods to detect the signature of species interactions in macroevolutionary data in situations where cophylogenetic approaches are not appropriate (see reviews by Weber et al., 2017, Harmon et al., 2019, Hembry and Weber, 2020). Cophylogenetic research and research applying phylogenetic comparative methods to the role of species interactions in macroevolution have largely proceeded independently of one another. However, these subfields are very closely related, and recent model-based methods development—*e.g.,* Braga et al. (2020)—begins to blur the distinction between them. We suggest that recent developments in each field may usefully inform the other in light of recent attention toward elucidating the role of ecological interactions among taxa in macroevolution.

2.5 Conclusion

Cophylogenetic methodology has advanced tremendously over the past several decades, and yet there is still far to go (Brooks, 1985, Page, 1994, Hafner et al., 1994). Today's cophylogenetic methods are varied in their assumptions, in what types of data they analyze, and in what types of estimates they produce. Yet it is as critical as ever for practicing biologists to carefully weigh which features of a method are most appropriate to study the question at hand. It is apparent to us that cophylogenetic studies do not yet benefit from the conveniences of many other macroevolutionary analysis frameworks—such as those relying on standard molecular models or phylogenetic comparative methods—because cophylogenetics currently lacks 'one-size-fits-all' methods. Why adequate methods are lacking largely boils down to our imperfect understanding of the evolutionary processes that generate cophylogenetic patterns, taxon sampling limitations, the inherent statistical and combinatorial complexity of cophylogenetic metrics and models, and the poor computational scalability of our inference methods for those approaches. As a result, every method typically must compromise something biological (realism), statistical (complexity or generality), or computational (speed or scalability) for it to be useful. In this review we have attempted to equip readers with a framework to locate where their datasets and hypotheses reside in the tangled frontier of cophylogenetics, navigate it safely, and use these methods productively and rigorously.

When do we care about cophylogenetic patterns vs. processes? It depends. Although there has been a strong tradition in phylogenetic comparative methods and macroevolution research to only make claims about pattern (Revell et al., 2008, Losos, 2011, Hembry and Weber, 2020), many of the goals in cophylogenetic research explicitly aim to draw conclusions about process. For example, it remains unclear to what extent true, contemporaneous cospeciation—a focal event that many cophylogenetic methods seek to identify—occurs. This ambiguity is not due to the state of cophylogenetic methods alone, as contemporaneous cospeciation can be challenging to demonstrate using additional lines of evidence or entirely non-cophylogenetic approaches (*e.g.,* population genomics). Certainly, there is value in knowing that two clades show a pattern of coarse phylogenetic congruence and that their evolutionary history has probably shaped their contemporary interaction patterns in some way. But we would argue that much of the motivation for cophylogenetic investigation is the explicit testing of hypotheses about the processes that generate these compelling patterns.

In this light, the recent development of model-based approaches in the field are especially important. Only a few model-based cophylogenetic approaches for simulating data (*e.g.,* Dismukes and Heath, 2021) and inferring historical processes (*e.g.,* Satler et al., 2019, Braga et al., 2020, Blasco-Costa et al., 2021) are currently available; a richer set of simulation tools as well as studies using simulations to understand the performance of methods are needed for the field to blossom. Cophylogenetic methods development is still lagging in terms of model design and performance assessment when compared with macroevolutionary methods from similar fields.

Given the extremely high-dimensional nature of cophylogenetic systems, and the inherent difficulty of defining closed-form likelihood equations for the appropriate generative models, we anticipate that likelihood-free deep learning methods will soon increase in popularity. Flexible, efficient, and mechanistic generative models for simulating large numbers of training datasets will be essential to train neural networks. That said, many challenges remain in terms of how to most efficiently structure cophylogenetic data for method input, how to define and/or select the best cophylogenetic summary statistics, and how to identify the limits of reasonable inference. Testing hypotheses of global cophylogenetic congruence might be feasible, but can we expect deep learning, or any new-fangled method, to correctly reconstruct all historical cospeciation and host-switching events with high confidence?

Our understanding for how intertwined symbiotic lineages diversify remains limited by the hypotheses we have considered, by the data we have gathered, and by the methods that we can employ. We are still far from knowing the limits of what cophylogenetic inferences can and cannot do. To stimulate future research towards this knowledge, we have listed a number of challenges (**Future Issues**) to consider, that have not been fully solved by current cophylogenetic approaches.

2.5.1 Future issues

Conceptual

- How do we distinguish true cospeciation from pseudocospeciation, and how common is each?
- How do we properly measure and model realized vs. fundamental partner repertoires that facilitate and prevent host-switching and host-range expansion in symbioses?
- What roles do pre-adaptation (Donoghue and Sanderson, 2015) and macroevolutionary memory (Goldberg and Foo, 2020) play in the gain and loss of species interactions?
- How important is functional trait evolution to the gain/loss of interactions? For example, using trait-matching (Nuismer and Harmon, 2015) for plant-pollinator systems (Muchhala and Thomson, 2009).
- How do we model interactions among more than two clades, *e.g.,* among figs, pollinating fig wasps, and antagonistic galling wasps? (Wang et al., 2019).

Methodological

- How can we better combine cophylogenetic analyses with other lines of evidence such as morphological, chemical, ecological, genomic, biogeographic, or paleontological data? (Baets et al., 2021)
- How do we perform inference under complex models, with large numbers of states and traits and interspecific biotic interactions (Quintero and Landis, 2020)?
- How well do these methods perform in terms of accuracy, robustness, etc. for different kinds of datasets and under different assumptions?

- What role does extinction play in inferring ancestral reconstructions accurately, and in inducing co-extinction events? (Rezende et al., 2007)
- How can we more rigorously quantify the phylogenetic conservatism of symbioses in cases where cophylogenetic methods themselves are not appropriate?

2.6 Disclosure Statement

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

2.7 Acknowledgements

We thank H.B. Shaffer, J. Balbuena, D. Althoff, M. Weber, and K. Segraves for helpful comments during the development of this review. We are very grateful to N. Janz and K. Quinteros for contributing photographs of ecological interactions used in Figure 1. This work was supported by a National Science Foundation (NSF USA) Graduate Research Fellowship (awarded to WD), NSF grants DBI-1759909 and DEB-1556853 (awarded to TAH), NSF grant DEB-2040347 (awarded to MJL), an International Postdoc Grant from the Swedish Research Council (2020-06422, awarded to MPB), and the University of Texas Permian Basin (to DHH).

2.8 References

Althoff, D. M., Segraves, K. A., Smith, C. I., Leebens-Mack, J., and Pellmyr, O. (2012). Geographic isolation trumps coevolution as a driver of yucca and yucca moth diversification. *Molecular Phylogenetics and Evolution*, 62(3):898–906.

Anthony, S. J., Johnson, C. K., Greig, D. J., Kramer, S., Che, X., Wells, H., Hicks, A. L., Joly, D. O., Wolfe, N. D., Daszak, P., et al. (2017). Global patterns in coronavirus diversity. *Virus evolution*, 3(1):vex012.

Avino, M., Ng, G. T., He, Y., Renaud, M. S., Jones, B. R., and Poon, A. F. (2019). Tree shape-based approaches for the comparative study of cophylogeny. *Ecology and Evolution*, 9(12):6756–6771.

Baets, K. D., Huntley, J. W., Klompmaker, A. A., Schiffbauer, J. D., and Muscente, A. (2021). The fossil record of parasitism: Its extent and taphonomic constraints. In *The Evolution and Fossil Record of Parasitism*, pages 1–50. Springer.

Balbuena, J. A., Miguez-Lozano, R., and Blasco-Costa, I. (2013). PACo: A Novel Procrustes Application to Cophylogenetic Analysis. *PLoS ONE*, 8(4):1–15.

Balbuena, J. A., Pérez-Escobar, Ó. A., Llopis-Belenguer, C., and Blasco-Costa, I. (2020). Random tanglegram partitions (Random TaPas): An Alexandrian approach to the cophylogenetic Gordian knot. *Systematic Biology*, 69(6):1212–1230.

Bascompte, J. and Jordano, P. (2013). *Mutualistic networks.* Princeton University Press, Princeton, NJ.

Baudet, C., Donati, B., Sinaimeri, B., Crescenzi, P., Gautier, C., Matias, C., and Sagot, M.-F. (2015). Cophylogeny reconstruction via an approximate Bayesian computation. *Systematic Biology*, 64(3):416–431.

Blasco-Costa, I., Hayward, A., Poulin, R., and Balbuena, J. A. (2021). Next-generation cophylogeny: unravelling eco-evolutionary processes. *Trends in Ecology & Evolution*, 36(10):907–918.

Braga, M. P., Janz, N., Nylin, S., Ronquist, F., and Landis, M. J. (2021). Phylogenetic reconstruction of ancestral ecological networks through time for pierid butterflies and their host plants. *Ecology Letters*, 24(10):2134–2145.

Braga, M. P., Landis, M. J., Nylin, S., Janz, N., and Ronquist, F. (2020). Bayesian inference of ancestral host–parasite interactions under a phylogenetic model of host repertoire evolution. *Systematic Biology*, 69(6):1149–1162.

Brooks, D. R. (1981). Hennig's parasitological method: A proposed solution. *Systematic Biology*, 30(3):229–249.

Brooks, D. R. (1985). Historical ecology: A new approach to studying the evolution of ecological associations. *Annals of the Missouri Botanical Garden*, pages 660–680.

Brooks, D. R. (1990). Parsimony analysis in historical biogeography and coevolution: methodological and theoretical update. *Systematic Zoology*, 39(1):14–30.

Brooks, D. R., Dowling, A. P., Van Veller, M. G., and Hoberg, E. P. (2004). Ending a decade of deception: a valiant failure, a not-so-valiant failure, and a success story. *Cladistics*, 20(1):32–46.

Brooks, D. R., Hoberg, E. P., and Boeger, W. A. (2015). In the eye of the cyclops: the classic case of cospeciation and why paradigms are important. *Comparative Parasitology*, 82(1):1–8.

Cai, L., Arnold, B. J., Xi, Z., Khost, D. E., Patel, N., Hartmann, C. B., Manickam, S., Sasirat, S., Nikolov, L. A., Mathews, S., et al. (2021). Deeply altered genome architecture in the endoparasitic flowering plant sapria himalayana griff.(rafflesiaceae). *Current Biology*, 31(5):1002–1011.

Charleston, M. (1998). Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences*, 149(2):191–223.

Charleston, M. and Page, R. (2002). TreeMap v. 2.0. 2. *Software distributed by authors.*

Charleston, M. A. and Perkins, L. (2003). Lizards, malaria, and jungles. *Tangled Trees: Phylogeny, Cospeciation, and Coevolution*, page 65.

Clayton, D. H., Bush, S. E., and Johnson, K. P. (2015). *Coevolution of life on hosts.* University of Chicago Press.

Compton, S. G., Ball, A. D., Collinson, M. E., Hayes, P., Rasnitsyn, A. P., and Ross, A. J. (2010). Ancient fig wasps indicate at least 34 Myr of stasis in their mutualism with fig trees. *Biology Letters*, 6(6):838–842.

Conow, C., Fielder, D., Ovadia, Y., and Libeskind-Hadas, R. (2010). Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology*, 5(1):1–10.

Cruaud, A., Rønsted, N., Chantarasuwan, B., Chou, L. S., Clement, W. L., Couloux, A., Cousins, B., Genson, G., Harrison, R. D., Hanson, P. E., et al. (2012). An extreme case of plant–insect codiversification: figs and fig-pollinating wasps. *Systematic Biology*, 61(6):1029–1047.

De Vienne, D., Refrégier, G., López-Villavicencio, M., Tellier, A., Hood, M., and Giraud, T. (2013). Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. *New Phytologist*, 198(2):347–385.

Dismukes, W. and Heath, T. A. (2021). treeducken: An R package for simulating cophylogenetic systems. *Methods in Ecology and Evolution*, 12(8):1358–1364.

Dobson, A., Lafferty, K. D., Kuris, A. M., Hechinger, R. F., and Jetz, W. (2008). Homage to Linnaeus: how many parasites? how many hosts? *Proceedings of the National Academy of Sciences*, 105(Supplement 1):11482–11489.

Doña, J., Sweet, A. D., Johnson, K. P., Serrano, D., Mironov, S., and Jovani, R. (2017). Cophylogenetic analyses reveal extensive host-shift speciation in a highly specialized and host-specific symbiont system. *Molecular Phylogenetics and Evolution*, 115:190–196.

Donoghue, M. J. and Sanderson, M. J. (2015). Confluence, synnovation, and depauperons in plant diversification. *New Phytologist*, 207(2):260–274.

Drinkwater, B. and Charleston, M. A. (2014). Introducing TreeCollapse: a novel greedy algorithm to solve the cophylogeny reconstruction problem. *BMC Bioinformatics*, 15(16):1–15.

Drinkwater, B., Qiao, A., and Charleston, M. A. (2016). Wispa: A new approach for dealing with widespread parasitism. *arXiv preprint arXiv:1603.09415*.

Fahrenholz, H. (1913). Ectoparasiten und abstammungslehre. *Zoologischer Anzeiger*, 41:371–374.

Felsenstein, J. (1978). Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. *Systematic Zoology*, 27(4):401–410.

Flynn, P. J. and Moreau, C. S. (2019). Assessing the diversity of endogenous viruses throughout ant genomes. *Frontiers in Microbiology*, 10:1139.

Goldberg, E. E. and Foo, J. (2020). Memory in trait macroevolution. *The American Naturalist*, 195(2):300–314.

Groussin, M., Mazel, F., Sanders, J. G., Smillie, C. S., Lavergne, S., Thuiller, W., and Alm, E. J. (2017). Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. *Nature Communications*, 8(1):1–12.

Guimarães, P. R., Rico-Gray, V., Oliveira, P., Izzo, T. J., dos Reis, S. F., and Thompson, J. N. (2007). Interaction intimacy affects structure and coevolutionary dynamics in mutualistic networks. *Current Biology*, 17(20):1797–1803.

Hafner, M. S. and Nadler, S. A. (1990). Cospeciation in host-parasite assemblages: comparative analysis of rates of evolution and timing of cospeciation events. *Systematic Zoology*, 39(3):192–204.

Hafner, M. S., Sudman, P. D., Villablanca, F. X., Spradling, T. A., Demastes, J. W., and Nadler, S. A. (1994). Disparate rates of molecular evolution in cospeciating hosts and parasites. *Science*, 265(5175):1087–1090.

Harmon, L. J., Andreazzi, C. S., Débarre, F., Drury, J., Goldberg, E. E., Martins, A. B., Melián, C. J., Narwani, A., Nuismer, S. L., Pennell, M. W., et al. (2019). Detecting the macroevolutionary signal of species interactions. *Journal of Evolutionary Biology*, 32(8):769–782.

Harmon, L. J. and Glor, R. E. (2010). Poor statistical performance of the Mantel test in phylogenetic comparative analyses. *Evolution*, 64(7):2173–2178.

Hayward, A., Poulin, R., and Nakagawa, S. (2021). A broadscale analysis of host-symbiont cophylogeny reveals the drivers of phylogenetic congruence. *Ecology Letters*.

Hembry, D. H., Kawakita, A., Gurr, N. E., Schmaedick, M. A., Baldwin, B. G., and Gillespie, R. G. (2013). Non-congruent colonizations and diversification in a coevolving pollination mutualism on oceanic islands. *Proceedings of the Royal Society B: Biological Sciences*, 280(1761):20130361.

Hembry, D. H., Okamoto, T., and Gillespie, R. G. (2012). Repeated colonization of remote islands by specialized mutualists. *Biology Letters*, 8(2):258–261.

Hembry, D. H. and Weber, M. G. (2020). Ecological interactions and macroevolution: a new field with old roots. *Annual Review of Ecology, Evolution, and Systematics*, 51:215–243.

Hommola, K., Smith, J. E., Qiu, Y., and Gilks, W. R. (2009). A permutation test of host–parasite cospeciation. *Molecular Biology and Evolution*, 26(7):1457–1468.

Huelsenbeck, J. P., Larget, B., and Rannala, B. (2000). A bayesian framework for the analysis of cospeciation. *Evolution*, 54:352–364.

Huelsenbeck, J. P., Rannala, B., and Yang, Z. (1997). Statistical tests of host-parasite cospeciation. *Evolution*, 51(2):410–419.

Hutchinson, M. C., Cagua, E. F., and Stouffer, D. B. (2017). Cophylogenetic signal is detectable in pollination interactions across ecological scales. *Ecology*, 98(10):2640–2652.

Hypša, V. (2006). Parasite histories and novel phylogenetic tools: alternative approaches to inferring parasite evolution from molecular markers. *International Journal for Parasitology*, 36(2):141–155.

Jablonski, D. (2008). Biotic interactions and macroevolution: extensions and mismatches across scales and levels. *Evolution*, 62(4):715–739.

Janz, N., Nyblom, K., and Nylin, S. (2001). Evolutionary dynamics of host-plant specialization: a case study of the tribe Nymphalini. *Evolution*, 55(4):783–796.

Labandeira, C. C., Dilcher, D. L., Davis, D. R., and Wagner, D. L. (1994). Ninety-seven million years of angiosperm-insect association: paleobiological insights into the meaning of coevolution. *Proceedings of the National Academy of Sciences*, 91(25):12278–12282.

Landis, M. J., Matzke, N. J., Moore, B. R., and Huelsenbeck, J. P. (2013). Bayesian analysis of biogeography when the number of areas is large. *Systematic Biology*, 62(6):789–804.

Legendre, P., Desdevises, Y., and Bazin, E. (2002). A statistical test for host–parasite coevolution. *Systematic Biology*, 51(2):217–234.

Legendre, P., Galzin, R., and Harmelin-Vivien, M. L. (1997). Relating behavior to habitat: solutions to thefourth-corner problem. *Ecology*, 78(2):547–562.

Losos, J. B. (2011). Seeing the forest for the trees: The limitations of phylogenies in comparative biology (American Society of Naturalists Address). *The American Naturalist*, 177(6):709–727.

Luo, S.-X., Yao, G., Wang, Z., Zhang, D., and Hembry, D. H. (2017). A novel, enigmatic basal leafflower moth lineage pollinating a derived leafflower host illustrates the dynamics of host shifts, partner replacement, and apparent coadaptation in intimate mutualisms. *The American Naturalist*, 189(4):422–435.

Maliet, O., Loeuille, N., and Morlon, H. (2020). An individual-based model for the eco-evolutionary emergence of bipartite interaction networks. *Ecology Letters*, 23(11):1623–1634.

McKinney, F. K. (1995). One hundred million years of competitive interactions between bryozoan clades: asymmetrical but not escalating. *Biological Journal of the Linnean Society*, 56(3):465–481.

Merkle, D. and Middendorf, M. (2005). Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information. *Theory in Biosciences*, 123(4):277–299.

Merkle, D., Middendorf, M., and Wieseke, N. (2010). A parameter-adaptive dynamic programming approach for inferring cophylogenies. *BMC Bioinformatics*, 11(1):1–10.

Mramba, L. K., Barber, S., Hommola, K., Dyer, L. A., Wilson, J. S., Forister, M. L., and Gilks, W. R. (2013). Permutation tests for analyzing cospeciation in multiple phylogenies: applications in tri-trophic ecology. *Statistical Applications in Genetics and Molecular Biology*, 12(6):679–701.

Muchhala, N. and Thomson, J. D. (2009). Going to great lengths: selection for long corolla tubes in an extremely specialized bat–flower mutualism. *Proceedings of the Royal Society B: Biological Sciences*, 276(1665):2147–2152.

National Research Council (2007). *Status of pollinators in North America*. The National Academies Press, Washington, DC.

Nooney, C., Barber, S., Gusnanto, A., and Gilks, W. R. (2017). A statistical method for analysing cospeciation in tritrophic ecology using electrical circuit theory. *Statistical Applications in Genetics and Molecular Biology*, 16(5-6):349–365.

Nuismer, S. L. and Harmon, L. J. (2015). Predicting rates of interspecific interaction from phylogenetic trees. *Ecology letters*, 18(1):17–27.

Ollerton, J. (2006). "Biological barter": patterns of specialization compared across different mutualisms. In *Plant-pollinator Interactions: From Specialization to Generalization*, pages 411–435. University of Chicago Press, Chicago, IL.

Page, R. D. M. (1994). Parallel phylogenies: reconstructing the history of host-parasite assemblages. *Cladistics*, 10:155–175.

Page, R. D. M. (2003). *Tangled Trees: Phylogeny, Cospeciation, and Coevolution*. University of Chicago Press.

Paterson, A. M. and Gray, R. D. (1997). Host-parasite co-speciation, host switching, and missing the boat. In *Host-parasite Evolution: General Principles and Avian Models*, pages 236–250. Oxford University Press, Oxford, UK.

Pérez-Escobar, O. A., Balbuena, J. A., and Gottschling, M. (2015). Rumbling orchids: How to assess divergent evolution between chloroplast endosymbionts and the nuclear host. *Systematic Biology*, 65(1):51–65.

Poisot, T. (2015). 23 when is co-phylogeny evidence of coevolution? *Parasite Diversity and Diversification: Evolutionary Ecology Meets Phylogenetics*, page 420.

Quintero, I. and Landis, M. J. (2020). Interdependent phenotypic and biogeographic evolution driven by biotic interactions. *Systematic Biology*, 69(4):739–755.

Ree, R. H., Moore, B. R., Webb, C. O., and Donoghue, M. J. (2005). A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution*, 59(11):2299–2311.

Revell, L. J., Harmon, L. J., and Collar, D. C. (2008). Phylogenetic signal, evolutionary process, and rate. *Systematic Biology*, 57(4):591–601.

Rezende, E. L., Lavabre, J. E., Guimarães, P. R., Jordano, P., and Bascompte, J. (2007). Non-random coextinctions in phylogenetically structured mutualistic networks. *Nature*, 448(7156):925–928.

Robinson, D. and Foulds, L. (1979). Comparison of weighted labelled trees. *Combinatorial mathematics, VI (Proc. Sixth Austral. Conf., Univ. New England, Armidale, 1978), Lecture Notes in Mathematics*, 748:119–126.

Ronquist, F. (1995). Reconstructing the history of host-parasite associations using generalised parsimony. *Cladistics*, 11(1):73–89.

Ronquist, F. (2003). Parsimony analysis of coevolving species associations. In Page, R., editor, *Tangled Trees: Phylogeny, Cospeciation and Coevolution*, pages 22–64. The University of Chicago Press Chicago.

Santichaivekin, S., Yang, Q., Liu, J., Mawhorter, R., Jiang, J., Wesley, T., Wu, Y.-C., and Libeskind-Hadas, R. (2021). empress: a systematic cophylogeny reconciliation tool. *Bioinformatics*, 37(16):2481–2482.

Satler, J. D., Herre, E. A., Jandér, K. C., Eaton, D. A., Machado, C. A., Heath, T. A., and Nason, J. D. (2019). Inferring processes of coevolutionary diversification in a community of Panamanian strangler figs and associated pollinating wasps. *Evolution*, 73(11):2295–2311.

Schardl, C. L., Craven, K. D., Speakman, S., Stromberg, A., Lindstrom, A., and Yoshida, R. (2008). A Novel Test for Host-Symbiont Codivergence Indicates Ancient Origin of Fungal Endophytes in Grasses. *Systematic Biology*, 57(3):483–498.

Shoemaker, D. D., Katju, V., and Jaenike, J. (1999). Wolbachia and the evolution of reproductive isolation between *Drosophila recens* and *Drosophila subquinaria. Evolution*, 53(4):1157–1164.

Siddall, M. E. and Perkins, S. L. (2003). Brooks parsimony analysis: a valiant failure. *Cladistics*, 19(6):554–564.

Smith, C. I., Godsoe, W. K., Tank, S., Yoder, J. B., and Pellmyr, O. (2008). Distinguishing coevolution from covicariance in an obligate pollination mutualism: asynchronous divergence in Joshua tree and its pollinators. *Evolution*, 62(10):2676–2687.

Stevens, J. (2004). Computational aspects of host–parasite phylogenies. *Briefings in Bioinformatics*, 5(4):339–349.

Szöllősi, G. J., Boussau, B., Abby, S. S., Tannier, E., and Daubin, V. (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences*, 109(43):17513–17518.

Szöllősi, G. J., Davín, A. A., Tannier, E., Daubin, V., and Boussau, B. (2015). Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678):20140335.

Szöllősi, G. J., Tannier, E., Lartillot, N., and Daubin, V. (2013). Lateral gene transfer from the dead. *Systematic Biology*, 62(3):386–397.

Thompson, J. N. (1994). *The Coevolutionary Process.* University of Chicago Press, Chicago, IL.

Thompson, J. N. (2005). *The Geographic Mosaic of Coevolution.* University of Chicago Press, Chicago, IL.

Wang, A.-Y., Peng, Y.-Q., Harder, L. D., Huang, J.-F., Yang, D.-R., Zhang, D.-Y., and Liao, W.-J. (2019). The nature of interspecific interactions and co-diversification patterns, as illustrated by the fig microcosm. *New Phytologist*, 224(3):1304–1315.

Wang, B. and Qiu, Y.-L. (2006). Phylogenetic distribution and evolution of mycorrhizas in land plants. *Mycorrhiza*, 16(5):299–363.

Weber, M. G. and Agrawal, A. A. (2014). Defense mutualisms enhance plant diversification. *Proceedings of the National Academy of Sciences*, 111(46):16442–16447.

Weber, M. G., Wagner, C. E., Best, R. J., Harmon, L. J., and Matthews, B. (2017). Evolution in a community context: on integrating ecological interactions and macroevolution. *Trends in Ecology & Evolution*, 32(4):291–304.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in Statistics*, pages 196–202. Springer.

Zeng, Y. and Wiens, J. J. (2021). Do mutualistic interactions last longer than antagonistic interactions? *Proceedings of the Royal Society B*, 288(1958):20211457.

CHAPTER 3. TREEDUCKEN: AN R PACKAGE FOR SIMULATING COPHYLOGENETIC SYSTEMS

Wade Dismukes[1] and Tracy A. Heath[1]

[1] *Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Bessey Hall, Ames, Iowa, 50010, USA*

Modified from a manuscript published in *Methods in Ecology and Evolution*

Abstract:

1. Cophylogenetic methods describe discordance between non-independent phylogenies.

2. Simulation is necessary for testing cophylogenetic methods, but few simulators exist that are capable of generating data under explicit and biologically meaningful models.

3. We present an R package, treeducken, for simulating host-symbiont evolution and gene tree species tree evolution.

4. treeducken provides a straightforward and reproducible interface for simulating cophylogenetic data. This allows easier performance testing of methods and has potential applications in machine learning (ML) and approximate Bayesian computation (ABC) approaches.

Corresponding author: Wade Dismukes, Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Bessey Hall, Ames, Iowa, 50010, USA; E-mail: waded@iastate.edu.

3.1 Introduction

Comparing phylogenetic trees provides insights into biological processes at various biological scales. These comparisons are common between the phylogenies of host and symbiont lineages or gene trees and species trees. Because of their shared histories, trees derived from associated evolutionary processes are expected to show matching patterns of diversification, yet this is often not the case. Thus, understanding the underlying causes of incongruence in phylogenies that have ecological and evolutionary links is a rich area of research (*e.g.,* Page and Charleston, 1998, de Vienne et al., 2013, Szöllosi et al., 2013, Balbuena et al., 2020). While there are several methods for untangling cophylogenetic patterns, tools for generating simulated datasets to test the performance of tree-comparison methods are not readily available. To address this gap, we have developed treeducken, an R package for simulating datasets under host-symbiont or gene-tree/species-tree scenarios. This simulation package can be useful for quantifying the patterns generated under different simulation conditions and for evaluating the accuracy of methods used to understand codiversifying lineages.

Cophylogenetic methods are used to explain the incongruence between host and symbiont phylogenies. These methods are often divided into two categories: global-fit and event-based. Global-fit methods attempt to measure the degree of dependency between host and symbiont trees (Balbuena et al., 2020). Event-based methods use events such as host-switching or cospeciation (*i.e.,* a host lineage and symbiont lineage speciating simultaneously) to explain the degree of congruence between host and symbiont trees

(Charleston, 1998, Conow et al., 2010, Santichaivekin et al., 2020). More recently, statistical models have been used to examine host and symbiont data (Baudet et al., 2015, Alcala et al., 2017, Satler et al., 2019, Braga et al., 2020). Models for understanding gene-tree and species-tree discordance have received much attention as genomic data have become widely available. The processes of gene duplication, gene loss, lateral gene transfer (LGT), and incomplete lineage sorting (ILS) contribute significantly to the discordance between the gene and species tree (Maddison, 1997). Models describing these discordance-causing processes (Rannala and Yang, 2003, Heled and Drummond, 2009, Rasmussen and Kellis, 2012) can provide more robust species tree estimates (Kubatko et al., 2009), and a deeper understanding of the nature of genomic evolution (Szöllosi et al., 2013).

Simulation tools are essential for validating the performance of methods for comparing phylogenetic trees. There are currently few existing cophylogenetic simulators available for generating host-symbiont datasets. The simulation program, CoRE-PA (Keller-Schmidt et al., 2011) allows two types of events in host trees (cospeciation and host speciation with sorting of symbionts), and two types of events in symbiont trees (host switching and symbiont speciation). Baudet et al. (2015) and Alcala et al. (2017) each introduced simulation methods that generate symbiont phylogenies along user-defined host trees under forward-time birth-death processes, though, neither approach simultaneously simulates the evolution of both the host and the symbiont. There are a number of tools for generating discordant species trees and gene trees, including DLCoalSim (Rasmussen and Kellis, 2012) and SimPhy (Mallo et al., 2016). However, no existing tool can simulate both gene trees, species trees, and cophylogenetic data while accounting for extinction.

Here we present an R package, treeducken, for simulating both cophylogenetic and gene-tree/species-tree data. treeducken currently generates data under two scenarios: the cophylogenetic birth-death model for simulating host-symbiont coevolution and a hierarchical model (Rasmussen and Kellis, 2012, Mallo et al., 2016) for simulating gene-tree/species-tree coevolution. The cophylogenetic birth-death model presented here extends the work of Keller-Schmidt et al. (2011) by allowing for extinction and host-shift or host-switch speciation and the tandem simulation of host and symbiont phylogenies. The hierarchical gene-family model used by treeducken simulates trees under processes implicated in gene-tree/species-tree discordance. The modular simulation framework provides a user-friendly and reproducible workflow allowing interoperability with existing R packages.

3.2 Model

treeducken simulates under two distinct types of processes: host-symbiont evolution and gene-species evolution. Here, we present the *cophylogenetic birth-death model* for simulating the tandem evolution of interacting hosts and symbionts (Figure 3.1). To simulate gene-species coevolution, we use a hierarchical *three-tree model* consisting of three levels: species, locus, and gene (Figure 3.2). These three levels model different causes of gene tree and species tree discordance including ILS, LGT, and gene duplication and loss (Rasmussen and Kellis, 2012).

3.2.1 Cophylogenetic birth-death model

The cophylogenetic birth-death model simulates a pair of phylogenies, the host phylogeny and the symbiont phylogeny, and their ecological interactions or associations. The resulting two trees and extant associations are intended to mimic the data used in cophylogenetic analyses. Ecological interactions are represented by a presence-absence matrix, hereafter the association matrix, with rows representing hosts and columns representing symbionts. The association matrix determines which hosts and symbionts can evolve together; for example, the model does not allow cospeciation in unassociated hosts and symbionts. The symbiont and host lineages can undergo speciation and extinction independently, and they can speciate together (cospeciation) or go extinct together (coextinction) (Figure 3.1). The user can also set symbiont lineages' maximum number of host associations at any given time. For example, with a host limit of three, a symbiont is only able to be associated with, at the most, three hosts.

The symbiont speciation and extinction rates determine the independent evolutionary history of the symbiont taxa (Figure 3.1a,b). Following a symbiont speciation event, both descendant lineages inherit their ancestral associations. The symbiont speciation event described here corresponds to the cophylogenetic event of duplication. The model also includes host expansion speciation, a special case of symbiont speciation where one of the descendant lineages, chosen at random, gains a novel association in addition to the symbiont's ancestral host repertoire (Figure 3.1c). The host-expansion event described here is similar to the spreading event described in Brooks et al. (1991). Users can use the `hs_mode = TRUE` argument to change the host-expansion rate to the host-switching rate. Here we define host-switching as a symbiont speciation where one descendant gains a randomly chosen novel association and the other descendant inherits the ancestral host repertoire. The symbiont tree has three parameters: symbiont speciation rate λ_S, symbiont extinction rate μ_S, and host expansion rate χ.

Cospeciation occurs when one host lineage bifurcates and a simultaneous speciation event occurs on one of the host's (randomly selected) symbiont lineages (Figure 3.1d). Following cospeciation, each descendant host lineage is then associated with one of the new descendant symbiont lineages. The remainder of the associations of the ancestral lineages of host or symbiont are sorted at random among the host's or symbiont's descendants. The cospeciation rate λ_C, is a shared parameter between the host and the symbiont tree.

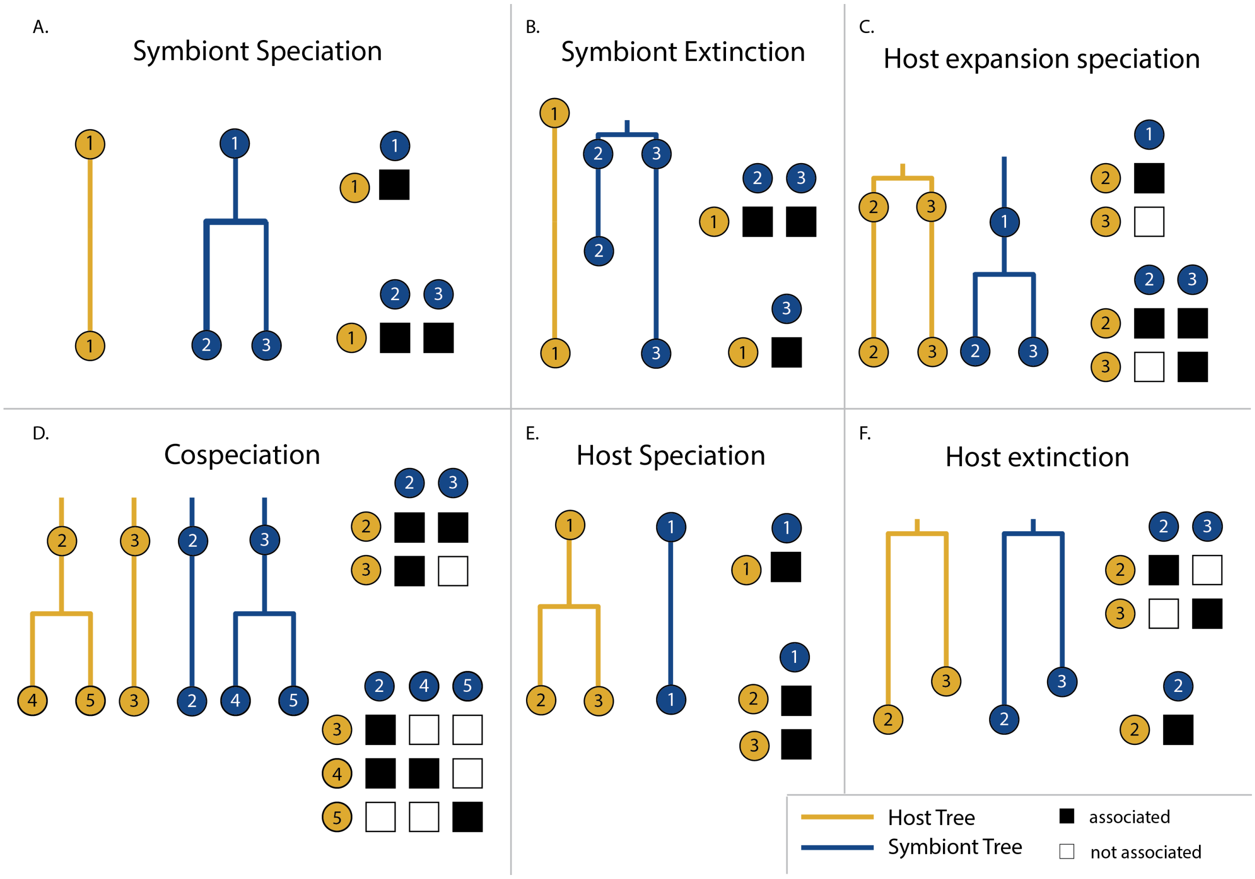


Figure 3.1 Cophylogenetic birth-death process simulation model including examples of each event. The host tree is shown in goldenrod and the symbiont tree in navy. A. Symbiont speciation event of symbiont 1 into descendant symbionts 2 and 3. The association matrix gains a column corresponding to the new symbionts, and both symbionts inherit the host association. B. Symbiont extinction of symbiont 2. The association matrix loses the corresponding column. C. Host expansion speciation. This is the same as the event in A except that one of the symbionts gains a random new host. D. Cospeciation of one host lineage and one symbiont lineage. A host lineage speciates and a randomly chosen associated symbiont speciates at the same time. Each descendant host lineage is associated with one of the descendant symbiont lineages. The ancestor host and symbiont lineage's remaining associations are sorted randomly among the descendants. E. Host speciation. A host lineage speciates, and the associations are randomly sorted on one or, in this case, both of the descendant lineages. F. Host extinction occurs. This results in a symbiont with no associations forcing symbiont extinction.

Host speciation—controlled by rate λ_H—describes events where the host speciates independently from its symbiont, *i.e.,* no cospeciation occurs (Figure 3.1e). This is equivalent to the failure-to-diverge event used in many event-based methods (Conow et al., 2010). Following a host speciation event, ancestral associations are sorted randomly on either or both descendant lineages. Host extinction refers to events where the host goes extinct (Figure 3.1f). If there are symbiont lineages left without a host following host extinction, then coextinction occurs. Host extinction—occurring at rate μ_H—is not independent of the symbiont phylogeny because host extinction can cause symbiont extinction. Coextinction does not occur if a host is left with no associations following a symbiont extinction event. For host-parasite systems this is adequate, but for obligate mutualisms, this may not be biologically realistic if hosts are unable to persist without their symbionts. Extending this model to accommodate these cases would be straightforward and is an area for further development.

The generating process is a birth-death model with parameters $\lambda_S, \mu_S, \chi, \lambda_C, \lambda_H,$ and μ_H conditioned on time. This forward-time simulation terminates after a pre-specified amount of time, resulting in two phylogenies and an association matrix. The cophylogenetic birth-death process simulation is performed using a single R function `sim_cophyloBD` that takes as input all six parameters of the model, the pre-specified amount of time, and the number of cophylogenetic datasets to simulate. The function outputs a list containing a host tree, a symbiont tree, the extant association matrix with hosts in rows and symbionts in columns, and a data frame containing all events that have occurred during the simulation.

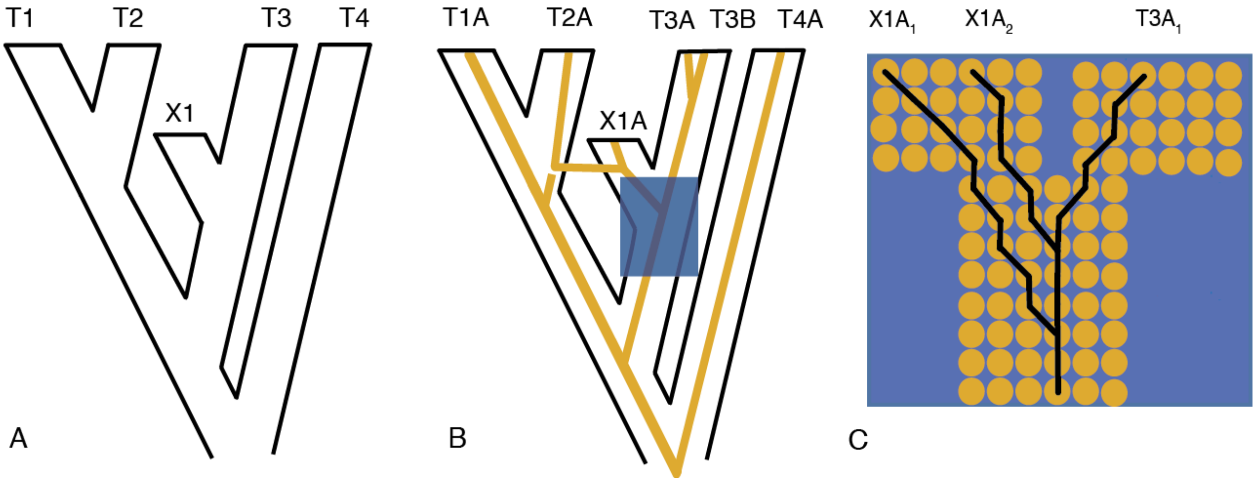


Figure 3.2 Three-tree simulation model of treeducken(A) The species tree (black) produced via a forward-time birth-death process. (B) A locus tree (goldenrod) simulated within the species tree using a second forward-time birth-death process coupled with LGTs. Capital letters following the tip name denote the locus identity. (C) A portion of the gene tree is backward simulated using a coalescent process within part of the locus tree (indicated by the blue box). Note that multiple individuals, denoted by subscripts, have been sampled.

3.2.2 Three-tree model

We implemented a hierarchical model, the three-tree model, to allow simulation from the species level down to coalescent sites (Rasmussen and Kellis, 2012, Mallo et al., 2016). The three-tree model has three levels: the species tree, the locus tree, and the gene tree (Rasmussen and Kellis, 2012). At the highest level is the species tree describing the history of speciation and extinction (Figure 3.2a). The next level down is the locus tree, evolving within the species tree, which models gene duplications and losses at a locus within a genome (Figure 3.2b). Within the locus tree, the gene tree models the dynamics of the multi-locus coalescent process allowing for ILS (Figure 3.2c). treeducken extends the three-tree model of Rasmussen and Kellis (2012) to simulate gene duplication, gene loss, ILS, and LGT along all lineages including those bound for extinction.

The three-tree model in treeducken first simulates the species tree (Figure 3.3a) under the birth-death process (Kendall, 1948, Gernhard, 2008) using the general sampling algorithm (GSA) given in Hartmann et al. (2010) for forward simulating a tree to a set number of tips with set birth and death rates. treeducken is also able to simulate species trees to a pre-specified time using the simple sampling algorithm (SSA) (Stadler, 2011). For the species tree, users can use host and symbiont trees simulated from the cophylogenetic simulator, or use one of the species tree simulation functions `sim_stBD` for the GSA simulation or `sim_stBD_t` for the SSA simulation.

Next, a locus tree is simulated within the full species tree—*i.e.*, the species tree containing records of extinct lineages (Figure 3.3b). This locus is simulated over the time spanned by the complete species tree under a birth-death process coupled with LGT. Within treeducken the species tree is used as input along with a gene birth rate, gene death rate, LGT rate, and a number of loci to the `sim_ltBD` function to simulate a set of locus trees under a birth death process.

Finally, gene trees are simulated backwards in time along the locus tree using a multilocus coalescent process (Figure 3.2c) (Rasmussen and Kellis, 2012). Gene trees can also be simulated along species trees using the multispecies coalescent process (Rannala and Yang, 2003). In treeducken locus trees or species trees are used as input for the multilocus `sim_mlc` for the multispecies coalescent `sim_msc` functions. Users are able to set the generation time, mutation rate, and effective population size. Each of these gene trees corresponds to a single coalescent independent site within its containing locus.

3.3 Usage

3.3.1 Description of the R package

treeducken is implemented as an R package (R Core Team, 2020) and is available on CRAN (https://cran.r-project.org/web/packages/treeducken/index.html). Validation testing was conducted in R and is available on Github (https://github.com/wadedismukes/treeduckenValidation). This package makes extensive use of the Rcpp and RcppArmadillo packages to wrap C++ code and improve performance (Eddelbuettel and Sanderson, 2014). In addition to the simulation functions described above, the package builds on previous R phylogenetics packages to provide functions for assistance in determining various simulation parameters, calculating summary statistics, and plotting host-symbiont tree sets (Harmon et al., 2007, Revell, 2012). All phylogenies are output in the format of the APE package as full phylogenies containing extinct tips (Paradis and Schliep, 2018). An example cophylogenetic plot is shown in (Figure 3.3). Integration with R allows for straightforward simulation of parameters from statistical distributions, and intuitive integration with other macroevolutionary and cophylogenetic tools (*e.g.,* phytools, PACo) (Revell, 2012, Hutchinson et al., 2017).

3.3.2 Simulating cophylogenetic datasets

To generate cophylogenetic data, we first set the cophylogenetic birth-death process parameters: host speciation and extinction rate, symbiont speciation and extinction rate, host expansion rate, and cospeciation rate and the prespecified simulation time. These rates are set relative to time; for example, a speciation rate of 0.1 corresponds to on average 1 speciation every 10 time units.

```
library(treeducken)
set.seed(54)
```

```
lambda_H <- rexp(n=1)
mu_H <- 0.0
lambda_C <- rexp(n=1)
time <- 1.0

lambda_S <- rexp(n=1)
mu_S <- 0.0
lambda_total_H <- lambda_H + lambda_C
lambda_total_S <- lambda_S + lambda_C
H_tips <- ave_tips_st(lambda = lambda_total_H, mu = mu_H, t =
    time)

S_tips <- ave_tips_st(lambda = lambda_total_S, mu = mu_S, t =
    time)
cophy_obj <- sim_cophyloBD(hbr = lambda_H, hdr = mu_H, sbr =
    lambda_S, sdr = mu_S, cosp_rate = lambda_C, host_exp_rate =
    0.0, time_to_sim = time, numbsim = 1)
plot(cophy_obj[[1]], col = "orange", lty = "dotted")
```

Cophylogenetic objects are output as a `cophy` object with many generic functions implemented including `summary`, `plot`, and `print`. These objects can be used with functions within `treeducken` to perform the ParaFit global fit test, or with existing packages for cophylogenetics (Legendre et al., 2002). We provide an example of using treeducken with the `paco` package (Hutchinson et al., 2017).

```
library(paco)
host_dist <- cophenetic(host_tree(cophy_obj[[1]]))
```

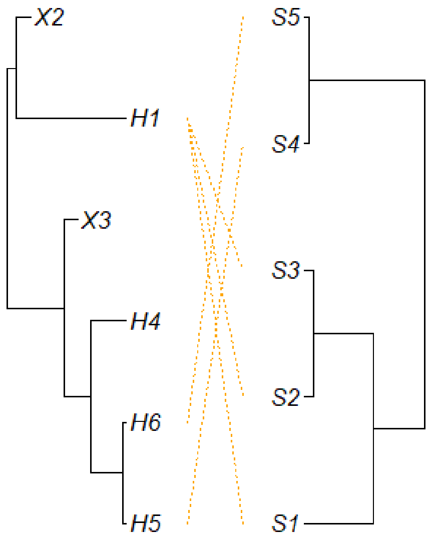


Figure 3.3 An example of the final output of the cophylogenetic simulation.

```
symb_dist <- cophenetic(symb_tree(cophy_obj[[1]]))
links <- association_mat(cophy_obj[[1])

rownames(links) <- cophy_obj[[1]]$host_tree$tip.label
colnames(links) <- cophy_obj[[1]]$symb_tree$tip.label
D <- paco::prepare_paco_data(H = host_dist, P = symb_dist, HP =
    links)
D <- paco::add_pcoord(D)
D <- paco::PACo(D, nperm=100, seed = 11, method="r0")
```

3.3.3 Simulating under the three tree model

In addition to the cophylogenetic simulations, treeducken can simulate using the three tree model to simulate gene tree and species tree discordance. For instance, we can use the host tree from the cophylogenetic example above to simulate a locus tree with gene birth and death rates. Then we use the locus tree and set mutation rate, generation time and population size to simulate under the multilocus coalescent. We can perform a similar simulation with the symbiont tree.

```r
host_tree <- host_tree(cophy_obj[[1]])
host_locus_trees <- sim_ltBD(host_tree, gbr = 0.2, gdr = 0.1, lgtr = 0.0, num_loci = 10)
host_gene_trees <- sim_mlc(host_locus_trees[[1]], effective_pop_size = 10000, generation_time = 1e-6, num_reps = 100)

symb_tree <- symb_tree(cophy_obj[[1]])
symb_locus_trees <- sim_ltBD(symb_tree, gbr = 0.3, gdr = 0.1, lgtr = 0.0, num_loci = 10)
symb_gene_trees <- sim_mlc(symb_locus_trees[[1]], effective_pop_size = 50000, generation_time = 1e-5, num_reps = 100)
```

3.4 Conclusion

treeducken adds to the phylogenetic analysis and simulation toolbox available in the R programming language (a list of those existing on the CRAN R project can be found here: https://cran.rproject.org/web/views/Phylogenetics.html). As a tree simulator, treeducken outputs phylogenetic trees in the format of the APE package, allowing it to be used in many other packages in the R ecosystem. This package fills a needed gap in the available phylogenetic tools in R by providing a straightforward means of simulating cophylogenetic data under a variety of various models. treeducken allows for a high degree of flexibility and complexity in simulations that make it straightforward to use for testing the performance of phylogenetic methods. We intend to improve this tool and increase its usability and functionality for empiricists and theoreticians using cophylogenetic methods.

3.5 Acknowledgments

Thanks to J. Justison for helpful comments on simulation of the cophylogenetic model, and J. Barido-Sottani, J. Satler, K. Quinteros, and B. Petrucci for code review. Also thank you to D. Mallo for helpful comments on an earlier version of this software and manuscript. This work was supported by an NSF GRFP award to WTD and NSF grant DEB-1556853 (Awarded to J.D. Nason, T.A. Heath, and E.A. Herre).

3.6 Data Accessibility

The R package, user manual, and example workflows are available on the Github page and CRAN. The validation tests are also available on Github.

3.7 Author Contributions

WTD and TAH designed the simulator and models. WTD implemented the methods. WTD and TAH wrote the manuscript.

3.8 References

Alcala, N., Jenkins, T., Christe, P., and Vuilleumier, S. (2017). Host shift and cospeciation rate estimation from co-phylogenies. *Ecology Letters*, 20(8):1014–1024.

Balbuena, J. A., Pérez-Escobar, Ó. A., Llopis-Belenguer, C., and Llopis-Belenguer, I. (2020). Random tanglegram partitions (Random TaPas): An Alexandrian approach to the cophylogenetic Gordian knot. *Systematic Biology*. syaa033.

Baudet, C., Donati, B., Sinaimeri, B., Crescenzi, P., Gautier, C., Matias, C., and Sagot, M.-F. (2015). Cophylogeny reconstruction via an approximate Bayesian computation. *Systematic Biology*, 64(3):416–431.

Braga, M. P., Landis, M. J., Nylin, S., Janz, N., and Ronquist, F. (2020). Bayesian inference of ancestral host–parasite interactions under a phylogenetic model of host repertoire evolution. *Systematic Biology*.

Brooks, D. R., McLennan, D. A., and McLennan, D. A. (1991). *Phylogeny, Ecology, and Behavior: A Research Program in Comparative Biology*. University of Chicago press.

Charleston, M. (1998). Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences*, 149(2):191–223.

Conow, C., Fielder, D., Ovadia, Y., and Libeskind-Hadas, R. (2010). Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology*, 5(1):16.

de Vienne, D. M., Refrégier, G., López-Villavicencio, M., Tellier, A., Hood, M. E., and Giraud, T. (2013). Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. *New Phytologist*, 198(2):347–385.

Eddelbuettel, D. and Sanderson, C. (2014). RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics & Data Analysis*, 71:1054–1063.

Gernhard, T. (2008). The conditioned reconstructed process. *Journal of Theoretical Biology*, 253(4):769–778.

Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E., and Challenger, W. (2007). GEIGER: investigating evolutionary radiations. *Bioinformatics*, 24(1):129–131.

Hartmann, K., Wong, D., and Stadler, T. (2010). Sampling trees from evolutionary models. *Systematic Biology*, 59(4):465–476.

Heled, J. and Drummond, A. J. (2009). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3):570–580.

Hutchinson, M. C., Cagua, E. F., Balbuena, J. A., Stouffer, D. B., and Poisot, T. (2017). paco: implementing Procrustean approach to cophylogeny in R. *Methods in Ecology and Evolution*, 8(8):932–940.

Keller-Schmidt, S., Wieseke, N., Klemm, K., and Middendorf, M. (2011). Evaluation of host parasite reconciliation methods using a new approach for cophylogeny generation. *University of Leipzig, Technical Report*, pages 11–013.

Kendall, D. G. (1948). On the generalized "birth-and-death" process. *Annals of Mathematical Statistics*, 19:1–15.

Kubatko, L. S., Carstens, B. C., and Knowles, L. L. (2009). STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, 25(7):971–973.

Legendre, P., Desdevises, Y., and Bazin, E. (2002). A statistical test for host–parasite coevolution. *Systematic Biology*, 51(2):217–234.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46:523–536.

Mallo, D., De Oliveira Martins, L., and Posada, D. (2016). SimPhy: Phylogenomic simulation of gene, locus, and species trees. *Systematic Biology*, 65(2):334–344.

Mooers, A., Gascuel, O., Stadler, T., Li, H., and Steel, M. (2012). Branch lengths on birth–death trees and the expected loss of phylogenetic diversity. *Systematic biology*, 61(2):195–203.

Page, R. D. M. and Charleston, M. A. (1998). Trees within trees: phylogeny and historical associations. *Trends in Ecology & Evolution*, 13:356–359.

Paradis, E. and Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3):526–528.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, 164(4):1645–1656.

Rasmussen, M. D. and Kellis, M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research*, 22(4):755–765.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223.

Santichaivekin, S., Yang, Q., Liu, J., Mawhorter, R., Jiang, J., Wesley, T., Wu, Y.-C., and Libeskind-Hadas, R. (2020). eMPRess: a systematic cophylogeny reconciliation tool. *Bioinformatics*, in press.

Satler, J. D., Herre, E. A., Jandér, K. C., Eaton, D. A., Machado, C. A., Heath, T. A., and Nason, J. D. (2019). Inferring processes of coevolutionary diversification in a community of Panamanian strangler figs and associated pollinating wasps. *Evolution*, 73(11):2295–2311.

Stadler, T. (2010). Sampling through time in birth–death trees. *Journal of Theoretical Biology*, 267(3):396–404.

Stadler, T. (2011). Simulating trees with a fixed number of extant species. *Systematic Biology*, 60(5):676–684.

Szöllosi, G. J., Tannier, E., Lartillot, N., and Daubin, V. (2013). Lateral gene transfer from the dead. *Systematic Biology*, 62(3):386–397.

3.9 Appendix: Validation of software and models

3.9.1 Validation tests

We performed validation tests at all the simulation levels available in treeducken to ensure that the simulations we performed fit theoretical or numerical expectations. The entire validation was conducted in R and is available on Github (https://github.com/wadedismukes/treeduckenValidation). In addition to validation testing, we performed unit testing to ensure the code ran as expected in a variety of scenarios.

3.9.1.1 Cophylogenetic birth-death model

We validated the host tree for the cophylogenetic birth-death model simulation. The symbiont tree and associations have no direct effect on the outcome of the host tree simulation directly. Thus, we compared the host tree output with results from the constant rate birth-death process simulated using the SSA. Specifically, we compared the expected number of tips, $E(N(t))$, of the host tree over a time, t, with the theoretical expectation given by Mooers et al. (2012): $E(N(t)) = 2e^{(\lambda-\mu)t}$ where λ and μ are the speciation and extinction rates of the constant rate birth-death process. Here the speciation rate is the sum of the host speciation rate (λ_H) and cospeciation rate (λ_C), and the extinction rate is the host extinction rate (μ_H). We simulated 10000 host-symbiont tree sets for $t = 2.0$ time units under 10 different host speciation rates, $\lambda_H = (0.1, 0.2, \ldots, 1.0)$, and a cospeciation rate of $\lambda_C = 1.0$. We set extinction rates set such that the turnover (i.e. μ/λ) was kept constant between simulation scenarios.

We validated the symbiont tree by comparing results with those found under the gene tree-species tree model. We used nine combinations of three different symbiont speciation rates and three different symbiont extinction rates with no host speciation or extinction, a cospeciation rate of $\lambda_C = 0.5$ and $t = 2.0$. We compared the average number of tips for each set of symbiont speciation and extinction rates to the theoretical expectation for the expected number of leaves of a locus tree: $E(N(t)) = \frac{ne^{(\lambda_S - \mu_S)t}}{1-m^{n-2}}$, with $m = \frac{\mu_S(1-e^{-(\lambda_S - l)t})}{\lambda_S - le^{-(\lambda_S - l)t}}$ with n being the number of species on the host tree, and λ_S and μ_S are the symbiont speciation and extinction rate respectively (Mallo et al., 2016).

3.9.1.2 Three-tree model

We validated each level of the three-tree model: species, locus, and gene. Species trees are able to be simulated using two different algorithms: the generalized sampling algorithm (GSA) Hartmann et al. (2010) and the simple sampling algorithm (SSA) Stadler (2010). To test the former, we simulated under six different birth rates (λ) with death rates (μ) set such that $\mu/\lambda = 0.25$ using 10000 replicate trees for each rate and 50 tips for each tree. The mean heights of the six different sets of birth and death rates were then compared to the mean heights of trees generated using `TreeSim` Stadler (2010). To test the latter, we simulated under the same six birth and death rates from above with 10000 replicates for each rate, and a fixed tree height, $t = 2.0$. The average number of tips for each set of birth and death rates were then compared to the theoretical expectation $E(N(t)) = 2e^{(\lambda - \mu)t}$ under this model (Mooers et al., 2012).

To determine if the locus trees were simulated correctly we tested that our trees had the expected number of tips. This was done using six different gene birth rates (δ) and gene loss rates (l) with a fixed species tree depth of 2.0, and $\lambda = 1.0$ and $\mu = 0.5$. The average number of tips for each set of gene birth and loss rates were then compared to the theoretical expectation: $E(N(t)) = \frac{ne^{(\delta - l)t}}{1-m^{n-2}}$ with $m = \frac{l(1-e^{-(\delta - l)t})}{\delta - le^{-(\delta - l)t}}$ with n being the number of species on the species tree.

Gene trees can be simulated using two different processes: the multispecies coalescent and the multilocus coalescent. The multispecies coalescent was tested by simulating 10 species trees each with $\lambda = 1.0$ and 25 tips. For each tree we then set population genetic parameters and simulated 100,000 gene trees. We set the generation time to 1×10^-6 time units per generation, and 1×10^-9 mutations per generation. The generation time value assumes our species tree is in time units of millions of years with 1 generation per year. The mutation rate was chosen to be a biologically realistic value based on the mutation rates of primates given in Rannala and Yang (2003). The effective population size was varied along orders of magnitudes from 10 to one million individuals. We then compared the time to most recent common ancestor (TMRCA) was to the expected TMRCA (Rannala and Yang, 2003) The multilocus coalescent was validated in a similar fashion, using the same parameters, but simulating under the multilocus coalescent. The simulated TMRCAs were then compared with the expectated TMRCA given in Rasmussen and Kellis (2012).

The validation tests were successful, showing little deviation from theoretical or numerical expectations.

CHAPTER 4. A DEEP LEARNING METHOD FOR COPHYLOGENETIC ANALYSIS

Wade Dismukes[1] and Tracy A. Heath[1]

[1]*Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Bessey Hall, Ames, Iowa, 50010, USA*

Modified from a manuscript to be submitted to *Systematic Biology*

4.1 Abstract

Cophylogenetic methods assess the discordance between host phylogeny and a symbiont phylogeny; however, few of these methods make use of generative models. Most of these methods rely on pattern-based statistics (also called global fit methods) or fitting of the symbiont tree onto the host tree using parsimony. Recent work has introduced a simple generative model, the cophylogenetic birth-death model, for simulating interacting host and symbiont phylogenies in tandem. Here I use the cophylogenetic birth-death model to train a deep learning method using summary statistics for the host and symbiont phylogeny to estimate parameters of the cophylogenetic birth-death model. These preliminary results explore the potential of machine learning in cophylogenetics and in related problems such as biogeography.

4.2 Introduction

More than 40% of the lineages in the Tree of Life are symbionts—*i.e.,* parasites, mutualists, or commensals—relying on other (often distantly related) lineages to survive (Dobson et al., 2008, Wang and Qiu, 2006, National Research Council, 2007). From brood-pollinating fig wasps and their host fig trees (Cruaud et al., 2012, Satler et al., 2019) to pocket gophers and their chewing lice (Hafner et al., 1994), many of these symbioses display a high degree of specificity and complexity and have fascinated evolutionary biologists since Darwin and Wallace (Darwin, 1862, Wallace, 1867). Heinrich Fahrenholz hypothesized that given a high degree of specificity, the topology of the symbiont phylogeny should match the phylogeny of its host (Fahrenholz, 1913). As more phylogenies of these systems became available, however, biologists studying coevolution found growing evidence of incongruence between host and symbiont evolutionary histories. Cophylogenetic methods were developed to better understand this discordance (Page, 2003).

Cophylogenetic methods are traditionally divided into two categories: pattern-based statistics or event-scoring methods (Dismukes et al., 2022). These methods use cophylogenetic datasets, which consist of a host tree, a symbiont tree, and a list of which extant hosts-symbiont associations (often summarized using an interaction matrix). Pattern-based statistic methods perform hypothesis testing on cophylogenetic datasets to determine the degree to which the host and symbiont trees and their ecological interactions are more similar than expected by chance alone (Legendre et al., 2002, Balbuena et al., 2020). Event-scoring methods, on the other hand, map the symbiont tree to the host phylogeny by reconstructing the possible history of events that produced the patterns observed in the cophylogenetic dataset (Conow et al., 2010, Santichaivekin et al., 2021). These events usually include cospeciation (host and symbiont speciate simultaneously), host-switching (symbiont jumps from one host to another), and independent symbiont speciation (also called duplication) (Page, 2003, Dismukes et al., 2022). These events are assigned costs, which are used to determine the lowest-cost mapping of the symbiont phylogeny onto the host phylogeny.

More recently, another group of methods has emerged that uses generative models to define these cophylogenetic events using stochastic processes. Several methods use generative models to simulate cophylogenetic datasets, including pure-birth models (Baudet et al., 2015, Alcala et al., 2017), those simulating host-repertoire evolution (Braga et al., 2020) and birth-death models (Dismukes and Heath, 2021). These cophylogenetic processes are complex and we do not yet have tractable solutions for calculating the probability densities of realizations of these models (*i.e.,* likelihood functions), thus making traditional statistical-inference methods unavailable.

The methods introduced by Baudet et al. (2015) and Alcala et al. (2017) both use approximate Bayesian computation (ABC) to estimate event rates under their respective generative models. Baudet et al. (2015) accomplish this by generating symbiont phylogenies on host phylogenies, and then calculating different distance metrics between the host and symbiont trees. The approach described by Alcala et al. (2017) converts the cophylogenetic dataset into a graph with symbiont edges, interaction edges, and host edges and then calculates graph summary statistics that can be used as input to the ABC method. Both Alcala et al. (2017) and Baudet et al. (2015) use a pure-birth simulation model (*i.e.,* no extinction) to generate symbiont phylogenies conditioned on their host phylogenies using their ABC methods. When using these ABC methods choosing the correct combination of summary statistics can be difficult to apply in practice (Beaumont, 2010). Indeed, the accuracy of ABC decreases rapidly with increasing numbers of summary statistics (Beaumont, 2010).

Machine learning offers a straightforward solution to this drawback by using algorithms to decide which of the summary statistics is more or less important (Kingma and Ba, 2014). Machine learning has seen prominent use in many computational biology fields including phylogeography (Smith et al., 2017), biogeography (Sukumaran et al., 2016) and population genetics (Schrider and Kern, 2017). Deep neural networks have also been used to accurately perform phylogenetic inference on multiple sequence alignments (Zou et al., 2020, Suvorov et al., 2020). The cophylogenetic birth-death model (Dismukes and Heath, 2021), offers a complex generative model that can produce the quantity of data required to train a deep neural network. As there is currently no traditional inference-based approach to analyzing cophylogenetic data under the cophylogenetic birth-death model, deep neural networks offer a suitable alternative that is scalable to datasets with a large number of hosts and symbionts. In addition, deep neural networks avoid the potential pitfall of choosing the wrong, not enough, or too many summary statistics.

We present a deep-learning method for estimating the number of cophylogenetic events. We build directly on the ABC method of Alcala et al. (2017) by using their general framework to generate summary statistics from cophylogenetic datasets. These summary statistics are then used to train a deep learning model that estimates the numbers of cophylogenetic events including symbiont speciation, host speciation, cospeciation, and host-switching. This method converts a cophylogenetic dataset into a graph incorporating the entire cophylogenetic dataset, calculates summary statistics, and uses these statistics as input for the deep neural network. We provide simulation results showing how well this model estimates cospeciation and host-switching on cophylogenetic datasets with and without extinction. Lastly, we analyze the cophylogenetic patterns of extant Panamanian strangler figs (*Ficus*) and their associated pollinating fig wasps (*Pegoscapus*) from Satler et al. (2019) to demonstrate an empirical use-case of this method.

4.3 Materials and Methods

We developed a deep learning method building on the ABC method of Alcala et al. (2017). We first simulate datasets using the cophylogenetic birth-death model and convert these cophylogenetic datasets into cophylogenetic graphs. These cophylogenetic graphs consist of host nodes and edges that formerly made up the host phylogeny, symbiont nodes and edges that formerly made up the symbiont phylogeny, and interaction edges that describing their associations. We then calculate summary statistics on this graph and use this as input for our deep neural network. The output of our neural network is the number of the different cophylogenetic events. The source code and scripts for conducting the analyses we describe are available on Github (`https://github.com/wadedismukes/cophyloML`).

4.3.1 Cophylogenetic birth-death model

We used the cophylogenetic birth-death model to simulate datasets under a range of conditions using the R package treeducken (Dismukes and Heath, 2021). The cophylogenetic birth-death model simulates a host phylogeny and a symbiont phylogeny, and their ecological interactions. The resulting trees and extant associations mimic the data used in cophylogenetic analyses. Ecological interactions are represented using a presence-absence matrix—called the association matrix—with rows representing hosts and columns representing symbionts. The association matrix determines which hosts and symbionts have intimate interactions and the potential to evolve in concert with one another (for detailed definitions of cophylogenetic scenarios see Dismukes et al., 2022). For example, this model does not allow cospeciation or coextinction to occur on noninteracting lineages. The host and symbiont lineages can undergo speciation and extinction independently, and interacting taxa can speciate or go extinct simultaneously (cospeciation and coextinction, respectively). This model additionally allows for host switching or host spreading. Host switching is defined here as an event in which a symbiont lineage speciates as one descendant lineage becomes associated with a new host, while the other retains the host repertoire of the ancestral symbiont lineage. Host spreading is defined as a symbiont speciation in which one descendant retains the ancestral host repertoire, and the other inherits the ancestral host repertoire and an additional novel host.

4.3.2 Summary statistics

We converted cophylogenetic datasets to networks using the R package igraph v1.3.0 (Csardi and Nepusz, 2006). These graphs have three types of edges: host edges, corresponding to host-tree branches; symbiont edges, corresponding to symbiont-tree edge; and ecological interaction edges that connect the tip nodes of the symbiont and the host phylogenies. The graphs also have two types of nodes: symbiont nodes and host nodes. This graph is used to calculate eleven summary statistics on the entire graph, the host graph, and the symbiont graph. The summary statistics (defined in Table 4.1) used were graph diameter (Wasserman et al., 1994), centralized Eigencentrality (Wasserman et al., 1994), graph density (Wasserman et al., 1994), mean and standard deviation of the coreness of the hosts Seidman (1983), mean and standard deviation of the coreness of the symbionts Seidman (1983), mean and standard deviation of the Jaccard similarity of the hosts (Adamic and Adar, 2003), and mean and standard deviation of the Jaccard similarity of the symbionts (Adamic and Adar, 2003). These summary statistics are a subset of those used in Alcala et al. (2017) and were chosen to provide an overall picture of the structure of the cophylogenetic network.

Table 4.1 Glossary of summary statistics used in the deep neural network. The coreness and Jaccard similarity are calculated for each node of either the host or the symbiont subgraph. We used the mean and standard deviation of the set of these values.

Summary statistic	**Graph definition**
diameter	The longest distance between two nodes.
density	The ratio of the number of edges and the number of possible edges.
centralized Eigencentrality	A measure of the influence of a node in a network. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high Eigenvector score means that a node is connected to many nodes who themselves have high score.
coreness	The k-core of graph is a maximal subgraph in which each vertex has at least degree k. The coreness of a vertex is k if it belongs to the k-core but not to the (k+1)-core.
Jaccard similarity	The ratio of the size of the intersection of set H and set S and the size of the union of set H and set S.

4.3.3 Parameter and model inference using neural networks

We used Python v3.7 to implement our deep feedforward neural network using TensorFlow 3.7.0 (Abadi et al., 2015), Keras 2.6.0 (Chollet et al., 2015), and sci-kit-learn 1.0.2 (Pedregosa et al., 2011). We tested several combinations of number of layers and neurons per layer, activation functions, regularization, loss functions, and optimizer. The one used in subsequent analyses is presented below. We provide a glossary of machine learning terms in Table 4.3 to familiarize readers with the basic terminology.

Our neural network architecture consisted of an input layer containing all summary statistics calculated on each cophylogenetic dataset. We used four sequential layers with the ReLU activation function (Agarap, 2018) with 64 neurons each with the final output layer containing four neurons for our estimates of host speciation, symbiont speciation, cospeciation and host-switching.

We optimized our neural network using the Adam algorithm with a learning rate of 0.001. Mean squared error (MSE) was the loss function, and the mean absolute error (MAE) as a metric. We randomly split each simulated dataset into a training set composed of 70% of our data and testing sets composed of the remaining 30% of the simulated data—20% and 10% respectively. The training set is used to train the model. The testing set is used to evaluate the final performance of the training of our model. We chose this 70%/30% split as that is a common choice used in the machine learning literature (Xu and Goodacre, 2018). The batch size was set to 256. We used the early stopping algorithm to prevent overfitting in our data (Yao et al., 2007).

Table 4.3 Glossary of machine learning terminology.

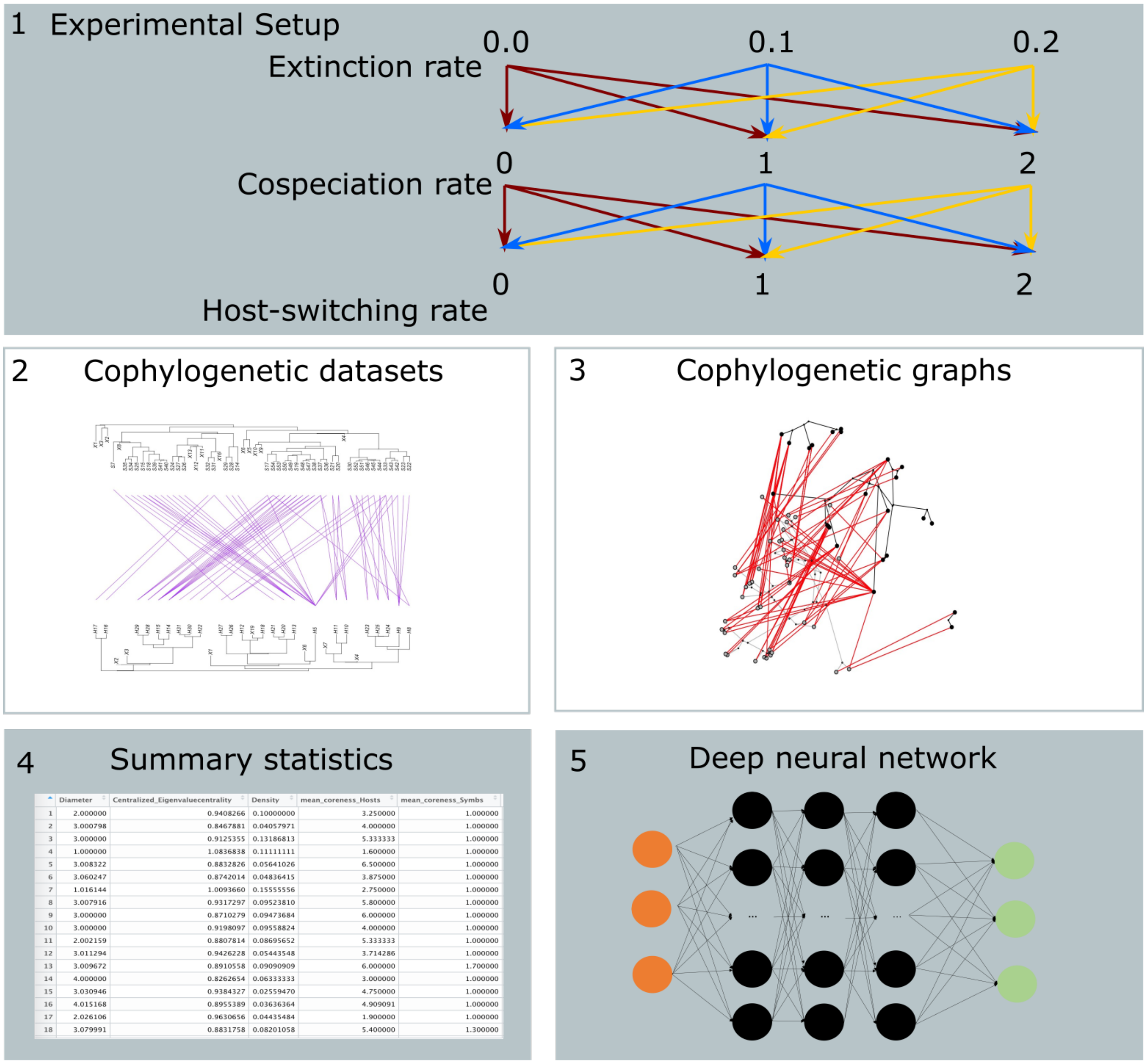
Term	**Definition**
layer	Neuron set in a neural network processing a set of input features, or the output of those neurons.
neuron	A node in a neural network which takes in multiple input values and generates one output value. The output value is calculated by applying an activation function (nonlinear transformation) to a weighted sum of input values.
activation function	A function (for example, ReLU) taking the weighted sum of all of the inputs from the previous layer to generate and pass an output value to the next layer.
learning rate	A scalar value used to train a model using gradient descent. In each iteration, the gradient descent algorithm multiplies the learning rate by the gradient. The resulting product is called the gradient step.
batch	The set of examples used in one iteration (that is, one gradient update) of model training.
loss function	A function which measures of how far a model's predictions are from its label.
ReLU	Short for Rectified Linear Unit. Defined by two rules: If input is negative or 0, output is 0. If input is positive, output is equal to input.

4.3.4 Evaluating performance

To test the performance of our machine learning method under the cophylogenetic birth-death model, we performed a simulation experiment where we varied host switching rates, cospeciation rates, and total extinction rates. The host-switching rates and cospeciation rates were set to 0, 1, or 2. Total extinction rates, where independent host and symbiont per-lineage extinction rates were set to be equal, were set at values of 0, 0.1, or 0.2. All permutations of these parameters resulted in a total of 27 simulation regimes (see Figure 4.1). These regimes attempted to capture a wide range of variation from independent host and symbiont phylogenies to highly dependent host and symbiont phylogenies. The remaining parameters of our cophylogenetic birth-death process were the same for all regimes. This included the host speciation rate of 1.5 and a symbiont speciation rate of 1.5. For each regime, we simulated 10,000 cophylogenetic datasets. Host-switching was set to switch mode for these analyses, and symbionts were allowed to be associated with at most four hosts. Host-switching is more commonly discussed in cophylogenetics literature so we use this here as it is more relevant to those researchers. Each of these was converted into a graph, summary statistics were calculated, and these statistics were fed into the deep learning model described above.

4.3.5 Empirical data

To demonstrate the efficacy of this method with empirical data we used a dataset of Panamanian strangler figs and their pollinating fig wasps (Satler et al., 2019). We simulated several test datasets that we suspected matched the biology of the system based on previous studies (*i.e.,* high host-switching rates, low cospeciation rates; Cruaud et al., 2012, Satler et al., 2019) using treeducken (Dismukes and Heath, 2021). Several parameters were set based on the biology of the system including host-switching which was set to the



	Diameter	Centralized_Eigenvaluecentrality	Density	mean_coreness_Hosts	mean_coreness_Symbs
1	2.000000	0.9408266	0.10000000	3.250000	1.000000
2	3.000798	0.8467881	0.04057971	4.000000	1.000000
3	3.000000	0.9125355	0.13186813	5.333333	1.000000
4	1.000000	1.0836838	0.11111111	1.600000	1.000000
5	3.008322	0.8832826	0.05641026	6.500000	1.000000
6	3.060247	0.8742014	0.04836415	3.875000	1.000000
7	1.016144	1.0093660	0.15555556	2.750000	1.000000
8	3.007916	0.9317297	0.09523810	5.800000	1.000000
9	3.000000	0.8710279	0.09473684	6.000000	1.000000
10	3.000000	0.9198097	0.09558824	4.000000	1.000000
11	2.002159	0.8807814	0.08695652	5.333333	1.000000
12	3.011294	0.9426228	0.05443548	3.714286	1.000000
13	3.009672	0.8910558	0.09090909	6.000000	1.700000
14	4.000000	0.8262654	0.06333333	3.000000	1.000000
15	3.030946	0.9384327	0.02559470	4.750000	1.000000
16	4.015168	0.8955389	0.03636364	4.909091	1.000000
17	2.026106	0.9630656	0.04435484	1.900000	1.000000
18	3.079991	0.8831758	0.08201058	5.400000	1.300000

Figure 4.1 Flowchart of the deep learning pipeline and simulation experiment. (1) First we varied in our simulations, each regime uses one extinction rate, one cospeciation rate, and one host-switching rate, totaling to 27 simulation regimes. (2) These simulation settings are used to produce cophylogenetic datasets such as the one shown, with purple lines showing the interactions between host and sybiont (3) Next, each of these is converted into a cophylogenetic graph where symbiont edges are shown in grey, host edges are shown in black, and the interactions edges are shown in red. (4) We then calculate summary statistics on these graphs. (5) These summary statistics are used as the features for a deep learning method.

"switch" setting and the maximum number of hosts per symbiont was set to two hosts. We used the phylogenies from Satler et al. (2019) and extracted the tree length from these to get the simulation time. Next, we calculated the expected birth and death rates of both phylogenies to obtain empirical estimates for some of the parameters for our cophylogenetic birth-death simulation model. We used the formulas given in Nee et al. (1994) to calculate the birth and death rates of both the host phylogeny and the symbiont phylogeny. The host-switching rates were a fraction of the symbiont speciation rate (0.5, 0.8). The cospeciation rates were set to a fraction of both the host and symbiont speciation rates (0.1, 0.2). The host and symbiont speciation rates were then adjusted to account for these new rates. For example, suppose we use a cospeciation value of 0.1 and a host-switching value of 0.5, and we have an estimated host speciation rate of 10 and a symbiont speciation rate of 8. Our cospeciation rate is then 1 with an adjusted host speciation rate of 9 and a symbiont speciation rate of 7.875. The host-switching rate, a type of symbiont speciation in our model, is then set to 4, and our symbiont speciation rate adjusted to 3.875. We simulated 10000 cophylogenetic datasets for each parameter setting. These simulated datasets were used to train a deep learning neural network, splitting this dataset into training and testing datasets. We then used the neural network to obtain predictions about our cophylogenetic events from our empirical dataset. The deep neural network used was the same as the one used in the simulation experiment.

4.4 Results and Discussion

4.4.1 Deep learning model accurately predicts cophylogenetic events

Our simulation experiment examined all combinations of three cospeciation rates (0, 1, 2), three host-switching rates (0, 1, 2), and three extinction rates (0, 0.1, 0.2). We evaluated the deep neural network's performance using MSE, R^2 score, and the accuracy of the estimated parameters. In a typical case, our loss function—*i.e.,* MSE dropped rapidly and stabilized after 200 epochs (see Appendix). The plots of the values of our loss functions as a function of the epoch for each simulation are also available in the GitHub repository. The loss values increased with both the simulated host-switching and cospeciation rates, but seemed to decrease in a number of cases with increasing extinction rate. The increased loss with host-switching and cospeciation rate was expected as more events of means more noise and thus higher error in our simulated data.

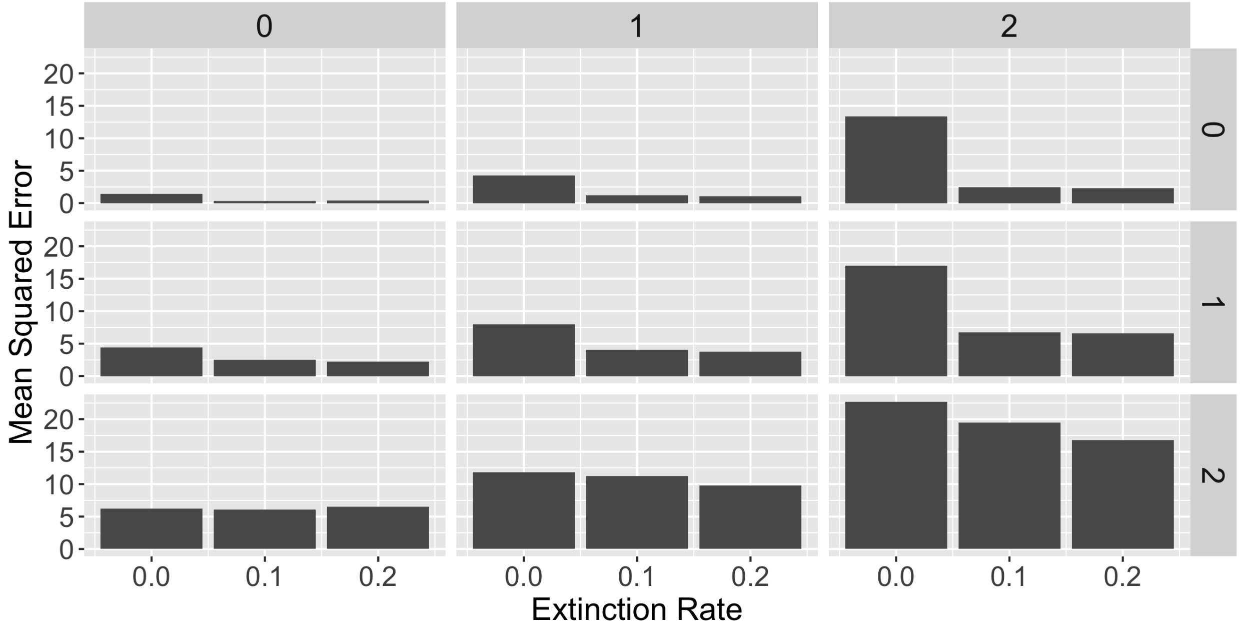


Figure 4.2 Mean squared error—the loss function for the deep neural network—-plotted for each of the 27 simulation regimes. The 0, 1, and 2 panels across the top of the figure are the host-switching rate. The 0, 1 and 2 panels across the right side of the figure are the cospeciation rates.

We also examined the R^2 value of our deep learning model for each of our simulation regimes (Figure 4.3). Figure 4.3 shows that the simplest case—*i.e.,* no cospeciation, no host switching—has the lowest R^2 values; whereas the highest cospeciation and host-switching rates produced the highest R^2 values. Interestingly, extinction seemed to have little effect on the value of R^2 in this simulation experiment. Figure 2 shows that a large portion of the variation in our datasets is explained by our model; although, this is not the case in the dataset generated without cospeciation or host switching. As few cophylogenetic datasets will have no cospeciation or host-switching this is unlikely to be of practical importance.

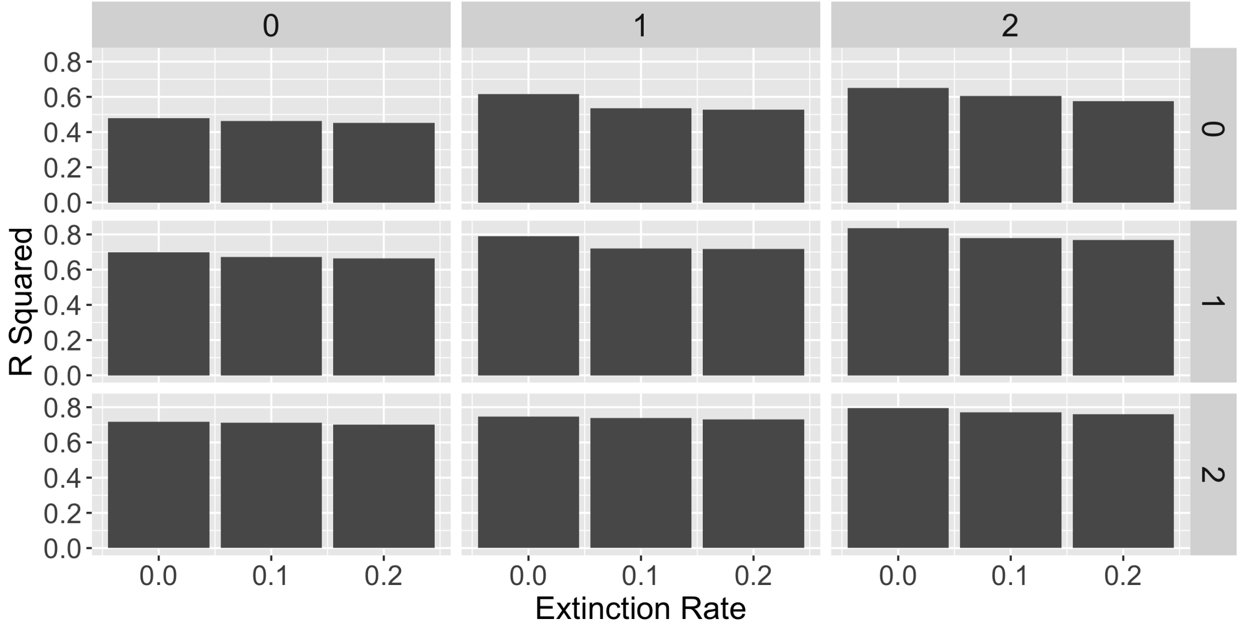


Figure 4.3 R^2 plotted for each of the 27 simulation regimes. The 0, 1, and 2 panels across the top of the figure are the host-switching rate. The 0, 1 and 2 panels across the right side of the figure are the cospeciation rates.

We also showed the accuracy of our model in predicting cospeciation (Figure 4.4), host switching (Figure 4.5), symbiont speciation (Figure 4.6), and host speciation (Figure 4.7). The model performed well at estimating the correct numbers of cospeciations. Extinction appeared to have little to no effect on the estimation of cospeciation events (Figure 4.4). Symbiont and host speciation were both accurately estimated with little impact from extinction (Figure 4.6). Host switching was consistently underestimated and seemed to be most impacted by the values of other parameters (Figure 4.5). In particular, the degree of cospeciation seemed to impact the accuracy of host-switch event estimates with this underestimation becoming more pronounced as the cospeciation rate increased. These results suggest that some of our summary statistics may not provide adequate signals for accurately estimating host switching. Including more summary statistics that capture different facets of our cophylogenetic datasets may increase this accuracy. One benefit of this method is that including more summary statistics is very straightforward. Considering this result, host switching versus host spreading could show a difference since host spreading is more likely to leave a lasting signal in the cophylogenetic dataset, particularly if host extinction is considered. For example, when a symbiont lineage undergoes a host-switching event one lineage obtains a single host species if that host subsequently goes extinct then the record of that host switch is essentially lost.

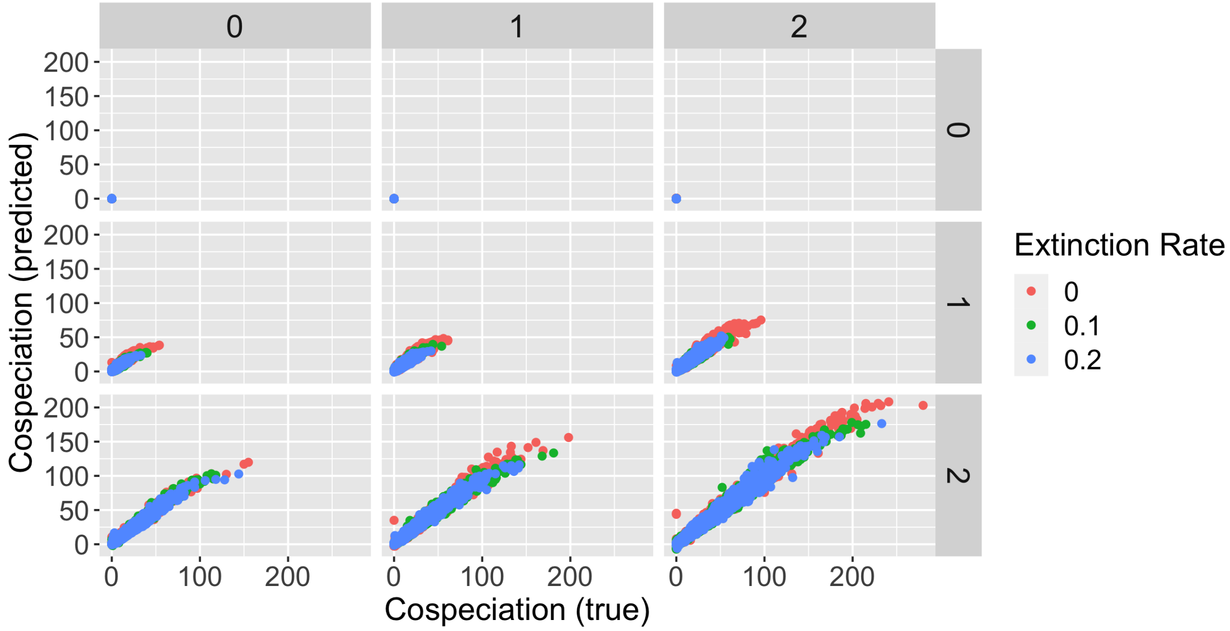


Figure 4.4 True cospeciation events plotted against cospeciation events predicted by the deep learning neural networks. The 0, 1, and 2 panels across the top of the figure are the host-switching rate. The 0, 1 and 2 panels across the right side of the figure are the cospeciation rates.

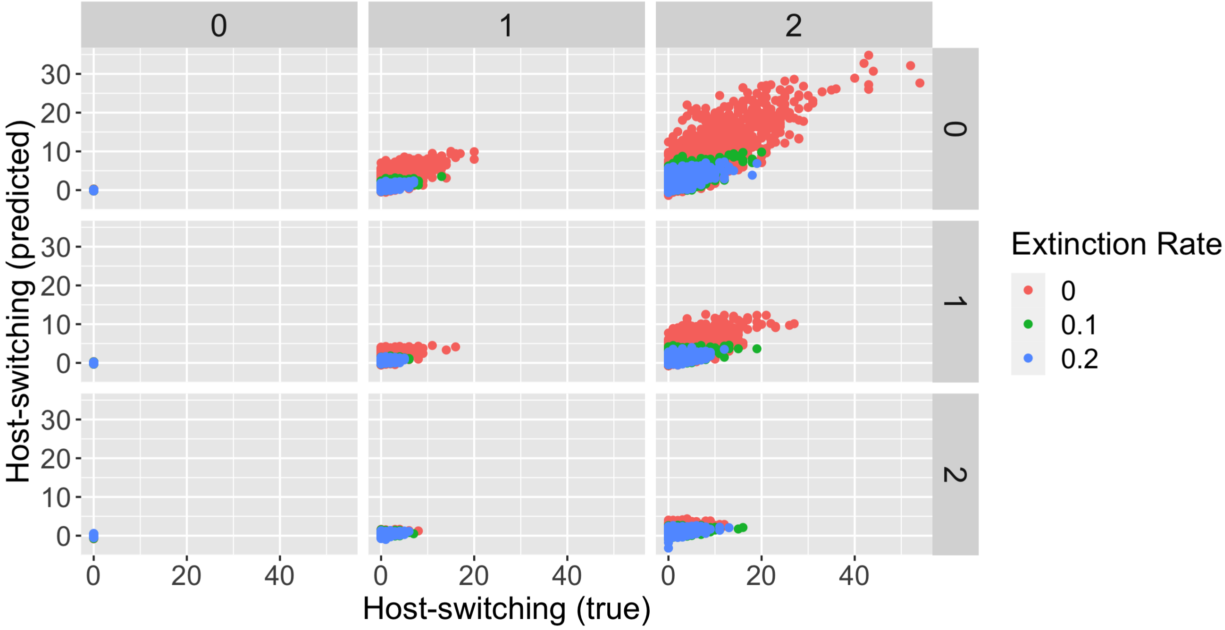


Figure 4.5 True host-switching events plotted against host-switching events predicted by the deep learning neural networks. The 0, 1, and 2 panels across the top of the figure are the host-switching rate. The 0, 1 and 2 panels across the right side of the figure are the cospeciation rates.

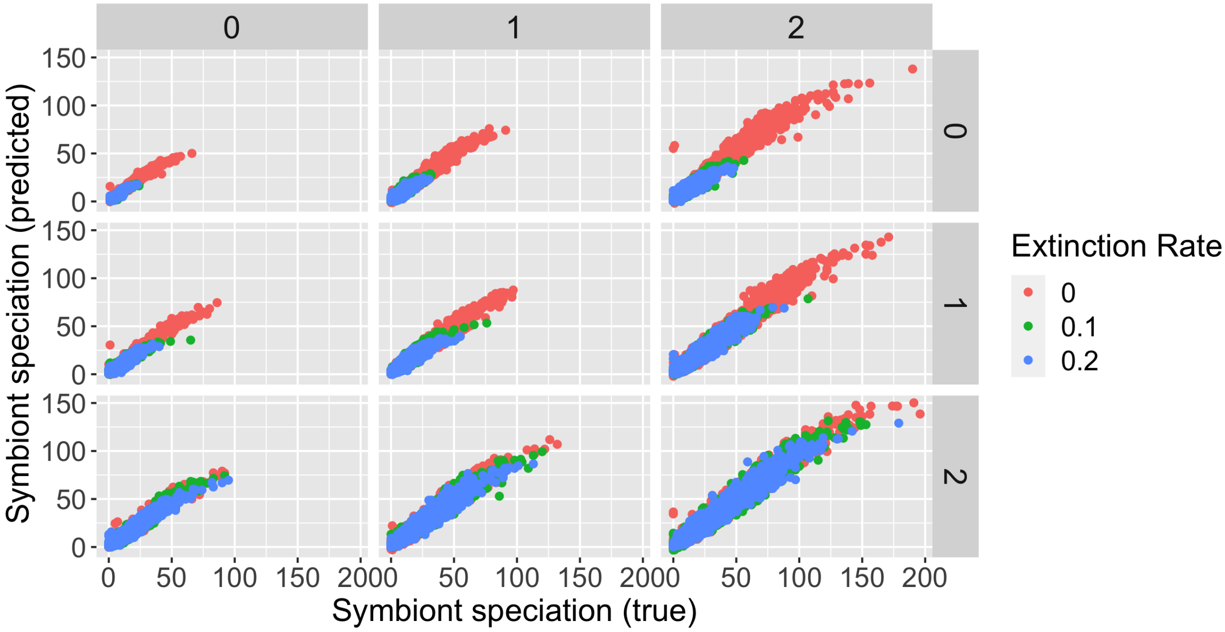


Figure 4.6 True symbiont speciation events plotted against symbiont speciation events predicted by the deep learning neural networks. The 0, 1, and 2 panels across the top of the figure are the host-switching rate. The 0, 1 and 2 panels across the right side of the figure are the cospeciation rates.

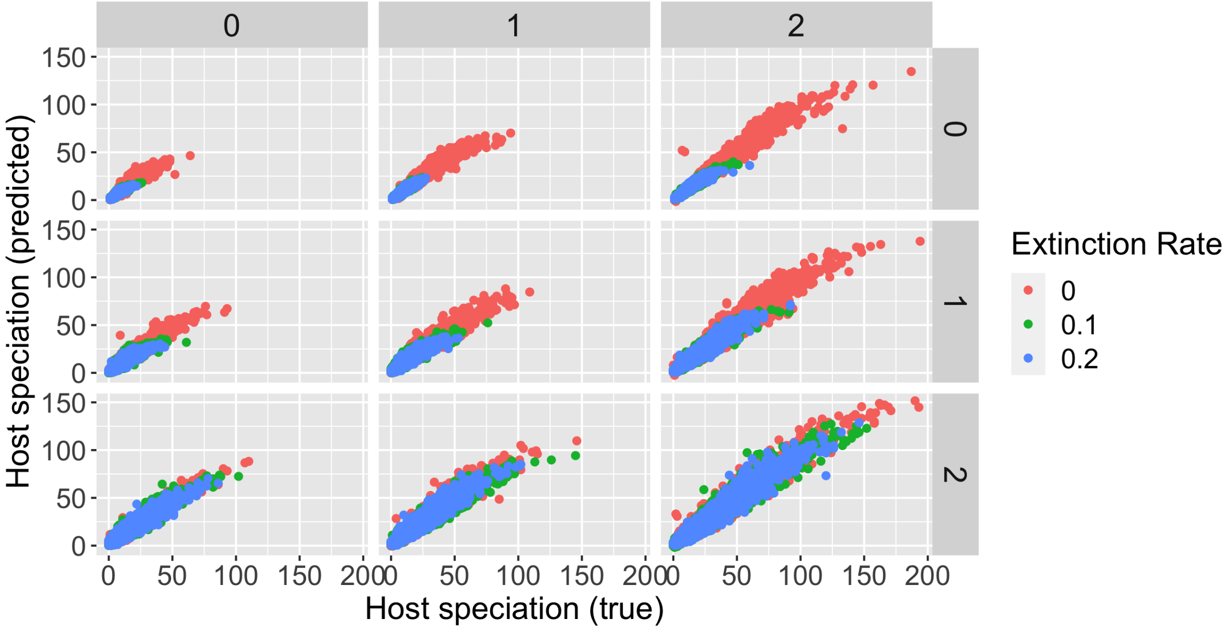


Figure 4.7 True host speciation events plotted against host speciation events predicted by the deep learning neural networks. The 0, 1, and 2 panels across the top of the figure are the host-switching rate. The 0, 1 and 2 panels across the right side of the figure are the cospeciation rates.

We examined the relationship between the size of our host trees and the relative error of parameters. This allowed us to investigate the possibility of the tree size increasing noise in our datasets. This is a potential problem because we currently cannot simulate datasets with set numbers of tips for the host or symbiont phylogenies. There was not a linear relationship between absolute error of host-switching and the size of the host tree (Figure 4.8). For cospeciation, there also no clear linear relationship between absolute error and the size of the host tree (Figure 4.9). Overall, these results suggest that our deep neural network can effectively estimate the number of cophylogenetic events.

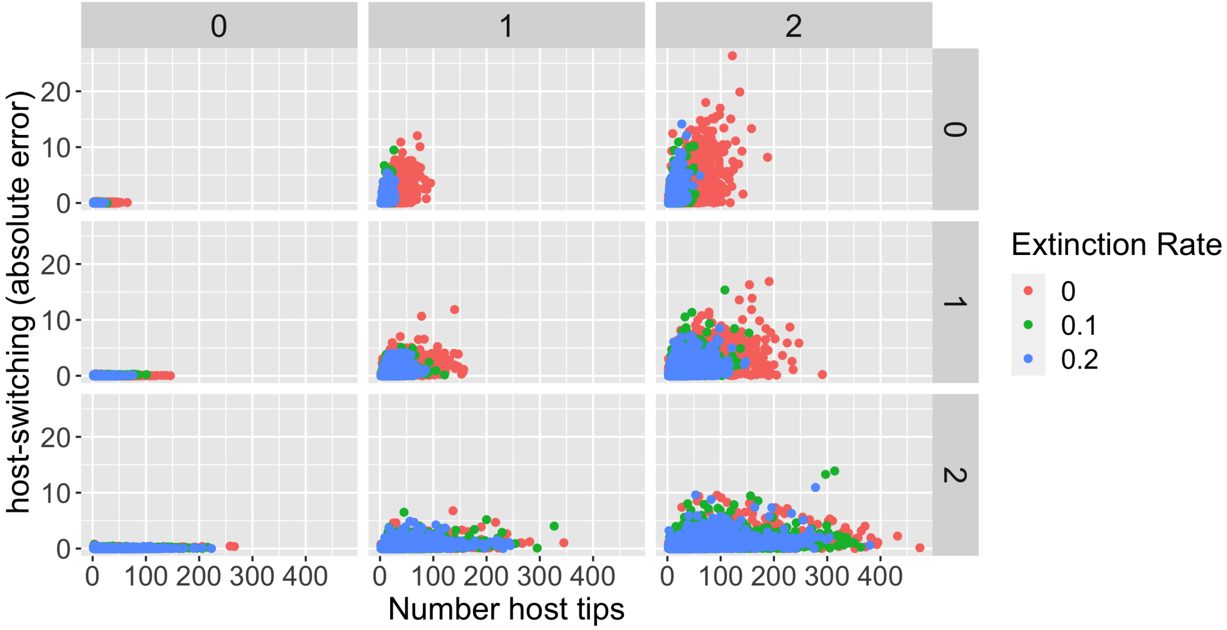


Figure 4.8 Absolute error of host-switching plotted against host tree size. The 0, 1, and 2 panels across the top of the figure are the host-switching rate. The 0, 1 and 2 panels across the right side of the figure are the cospeciation rates. Different colors show different extinction rates.

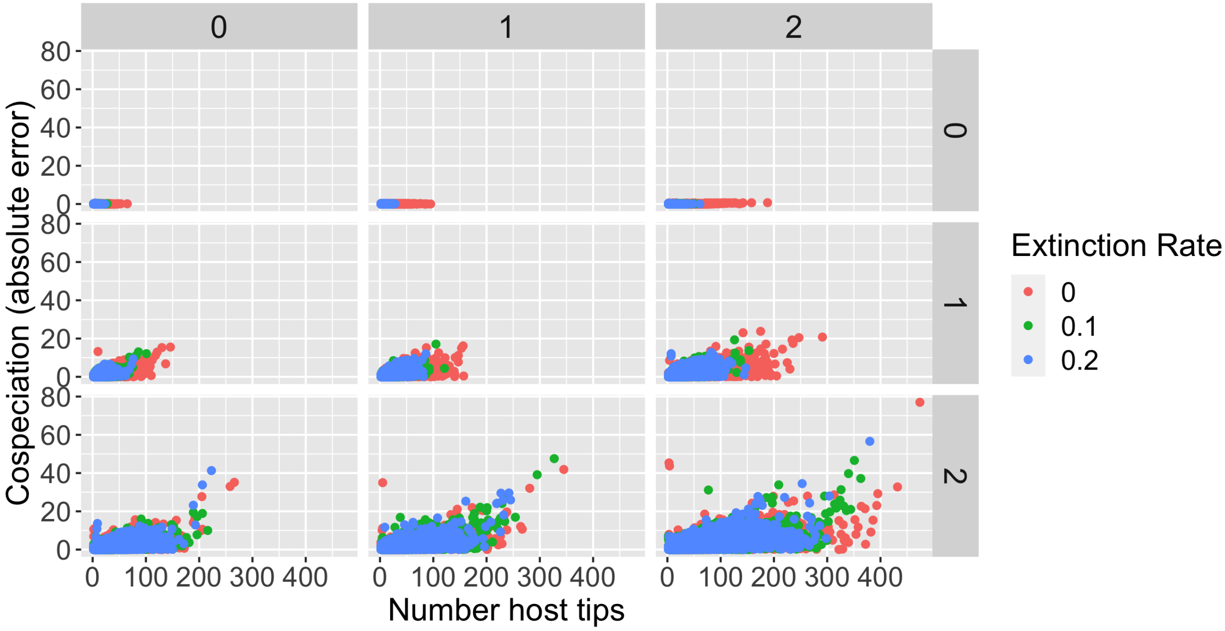


Figure 4.9 Absolute error of cospeciation plotted against host tree size. The 0, 1, and 2 panels across the top of the figure are the host-switching rate. The 0, 1 and 2 panels across the right side of the figure are the cospeciation rates. Different colors show different extinction rates.

4.4.2 Cophylogenetic analysis of figs and fig wasps

To demonstrate the utility of this method on an empirical dataset, we conducted a case study using a cophylogenetic dataset of fig trees and their associated fig-wasp pollinators (Satler et al., 2019). The results for these empirical analyses are shown in Table 4.5. Interestingly, these results contrast with those of Satler et al. (2019), where they used the amalgamated likelihood estimation method (ALE; Szöllősi et al., 2013) to produce average estimates for host-switching, cospeciations, and symbiont speciations. The ALE method is a gene-tree/species-tree approach that uses an ultrametric species tree along with a posterior distribution of gene family trees. Satler et al. (2019) treated the fig species as the host species tree, and a posterior distribution of fig-wasp species trees as gene family trees. Their average estimates for the numbers of different events were: 11.01 host switches, 5.92 cospeciations, and 0.32 symbiont speciations. Our results differed showing much lower values of host-switching events, less than two in all cases, and lower numbers of cospeciation events. We also showed much higher values of symbiont speciation.

There are a number of reasons that our results may not have matched those of Satler et al. (2019). It is possible that our simulated datasets did not adequately represent the empirical dataset and larger variety of simulation settings need to be performed. We intend to complete more simulations and further explore this empirical dataset using the deep neural network. In addition, the ALE method likens host-symbiont evolution to gene-tree/species-tree evolution, this comes with several possible issues that could be the cause of the difference between our datasets. This method does not allow symbionts to be shared by multiple hosts, because of this Satler et al. (2019) used a workaround wherein these lineages were duplicated. As these are being duplicated this could have inflated their symbiont speciation estimates. Our deep learning method does not require, and this may

Table 4.5 The results for the empirical case study.

	Host speciation	Symbiont speciation	Cospeciation	Host switching
Cospeciation 0.1, host-switching 0.5	8.912077	5.906229	1.349562	1.3684165
Cospeciation 0.1, host-switching 0.8	8.536900	7.472594	1.492852	0.8579537
Cospeciation 0.2, host-switching 0.5	6.875938	5.425841	2.970502	0.8682939
Cospeciation 0.2, host-switching 0.8	7.226309	5.982451	3.293928	0.5078039

produce different results. In addition, the ALE method only deals with evolution along the symbiont tree; whereas, our simulations account for both host and symbiont evolution.

4.5 Conclusion

Here, we used a deep neural network to count the number of cophylogenetic events in host-symbiont data. The results of our simulation experiments show that this method holds promise as a new cophylogenetic method with a number of advantages over previous methods. In particular, this method can simultaneously estimate cospeciation, symbiont speciation, host speciation, and host switching accurately without the need to assign costs. We also provided a case study on how to use this deep neural network in an empirical dataset. This method is also capable of dealing with symbionts be using multiple hosts as well as hosts with multiple symbionts. In addition, it is straightforward to estimate more cophylogenetic events such as symbiont dispersal and extirpation as this requires. We also provided a case study of how empiricists might use this method to analyze their own data.

Although our results did not coincide with previous analyses, we believe that this provides another tool that those studying host-symbiont evolution can use to corroborate inferences about their systems.

Our work suggests that deep learning has potential in the field of cophylogenetics as it can leverage both event-scoring methods, in the sense that these methods provide conceptual models and pattern-based statistic methods, which could provide an important source of summary statistics that could be used as features for neural networks. Although our findings are encouraging, additional work is needed on using deep neural networks for cophylogenetics. At present, there is no clear connection between cophylogenetic events themselves and the values of these summary statistics. Further work on the feature importance would be instructive in establishing these connections. For example, this can be done for deep neural networks by randomizing the entries for a single feature then retraining our model and then repeating this randomization for each of our features (Fisher et al., 2019). Our method also does not place events on a tree and does not produce a map of a symbiont tree on a host tree, as is done in several event-scoring methods (Conow et al., 2010, Santichaivekin et al., 2021, Charleston, 1998). Despite the present limitations, we believe that the use of deep learning and other machine learning approaches show promise for analyzing host-symbiont phylogenies.

4.6 References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the web. *Social networks*, 25(3):211–230.

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Alcala, N., Jenkins, T., Christe, P., and Vuilleumier, S. (2017). Host shift and cospeciation rate estimation from co-phylogenies. *Ecology Letters*, 20(8):1014–1024.

Balbuena, J. A., Pérez-Escobar, Ó. A., Llopis-Belenguer, C., and Blasco-Costa, I. (2020). Random tanglegram partitions (Random TaPas): An Alexandrian approach to the cophylogenetic Gordian knot. *Systematic Biology*, 69(6):1212–1230.

Baudet, C., Donati, B., Sinaimeri, B., Crescenzi, P., Gautier, C., Matias, C., and Sagot, M.-F. (2015). Cophylogeny reconstruction via an approximate Bayesian computation. *Systematic Biology*, 64(3):416–431.

Beaumont, M. A. (2010). Approximate bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, pages 379–406.

Braga, M. P., Landis, M. J., Nylin, S., Janz, N., and Ronquist, F. (2020). Bayesian inference of ancestral host–parasite interactions under a phylogenetic model of host repertoire evolution. *Systematic Biology*, 69(6):1149–1162.

Charleston, M. (1998). Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences*, 149(2):191–223.

Chollet, F. et al. (2015). Keras. https://keras.io.

Conow, C., Fielder, D., Ovadia, Y., and Libeskind-Hadas, R. (2010). Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology*, 5(1):1–10.

Cruaud, A., Rønsted, N., Chantarasuwan, B., Chou, L. S., Clement, W. L., Couloux, A., Cousins, B., Genson, G., Harrison, R. D., Hanson, P. E., et al. (2012). An extreme case of plant–insect codiversification: figs and fig-pollinating wasps. *Systematic Biology*, 61(6):1029–1047.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

Darwin, C. (1862). Letter to J.D. Hooker. *More letters of Charles Darwin Volume*, 2.

Dismukes, W., Braga, M. P., Hembry, D. H., Heath, T. A., and Landis, M. J. (2022). Cophylogenetic methods to untangle the evolutionary history of ecological interactions. *Annual Review of Ecology, Evolution, and Systematics*, in press.

Dismukes, W. and Heath, T. A. (2021). treeducken: An R package for simulating cophylogenetic systems. *Methods in Ecology and Evolution*, 12(8):1358–1364.

Dobson, A., Lafferty, K. D., Kuris, A. M., Hechinger, R. F., and Jetz, W. (2008). Homage to Linnaeus: how many parasites? how many hosts? *Proceedings of the National Academy of Sciences*, 105(Supplement 1):11482–11489.

Fahrenholz, H. (1913). Ectoparasiten und abstammungslehre. *Zoologischer Anzeiger*, 41:371–374.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81.

Hafner, M. S., Sudman, P. D., Villablanca, F. X., Spradling, T. A., Demastes, J. W., and Nadler, S. A. (1994). Disparate rates of molecular evolution in cospeciating hosts and parasites. *Science*, 265(5175):1087–1090.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Legendre, P., Desdevises, Y., and Bazin, E. (2002). A statistical test for host–parasite coevolution. *Systematic Biology*, 51(2):217–234.

National Research Council (2007). *Status of pollinators in North America.* The National Academies Press, Washington, DC.

Nee, S., May, R. M., and Harvey, P. H. (1994). The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1309):305–311.

Page, R. D. M. (2003). *Tangled Trees: Phylogeny, Cospeciation, and Coevolution*. University of Chicago Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Santichaivekin, S., Yang, Q., Liu, J., Mawhorter, R., Jiang, J., Wesley, T., Wu, Y.-C., and Libeskind-Hadas, R. (2021). empress: a systematic cophylogeny reconciliation tool. *Bioinformatics*, 37(16):2481–2482.

Satler, J. D., Herre, E. A., Jandér, K. C., Eaton, D. A., Machado, C. A., Heath, T. A., and Nason, J. D. (2019). Inferring processes of coevolutionary diversification in a community of Panamanian strangler figs and associated pollinating wasps. *Evolution*, 73(11):2295–2311.

Schrider, D. R. and Kern, A. D. (2017). Machine learning for population genetics: A new paradigm. *bioRxiv*, page 206482.

Seidman, S. B. (1983). Network structure and minimum degree. *Social Networks*, 5(3):269–287.

Smith, M. L., Ruffley, M., Espíndola, A., Tank, D. C., Sullivan, J., and Carstens, B. C. (2017). Demographic model selection using random forests and the site frequency spectrum. *Molecular Ecology*, 26(17):4562–4573.

Sukumaran, J., Economo, E. P., and Lacey Knowles, L. (2016). Machine learning biogeographic processes from biotic patterns: a new trait-dependent dispersal and diversification model with model choice by simulation-trained discriminant analysis. *Systematic Biology*, 65(3):525–545.

Suvorov, A., Hochuli, J., and Schrider, D. R. (2020). Accurate inference of tree topologies from multiple sequence alignments using deep learning. *Systematic Biology*, 69(2):221–233.

Szöllősi, G. J., Tannier, E., Lartillot, N., and Daubin, V. (2013). Lateral gene transfer from the dead. *Systematic Biology*, 62(3):386–397.

Wallace, A. R. (1867). Creation by law. *QJ Sci*, 4(16):470–488.

Wang, B. and Qiu, Y.-L. (2006). Phylogenetic distribution and evolution of mycorrhizas in land plants. *Mycorrhiza*, 16(5):299–363.

Wasserman, S., Faust, K., et al. (1994). Social network analysis: Methods and applications.

Xu, Y. and Goodacre, R. (2018). On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, 2(3):249–262.

Yao, Y., Rosasco, L., and Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315.

Zou, Z., Zhang, H., Guan, Y., and Zhang, J. (2020). Deep residual neural networks resolve quartet molecular phylogenies. *Molecular Biology and Evolution*, 37(5):1495–1507.

4.7 Appendix: Plots of MAE and Loss by epoch

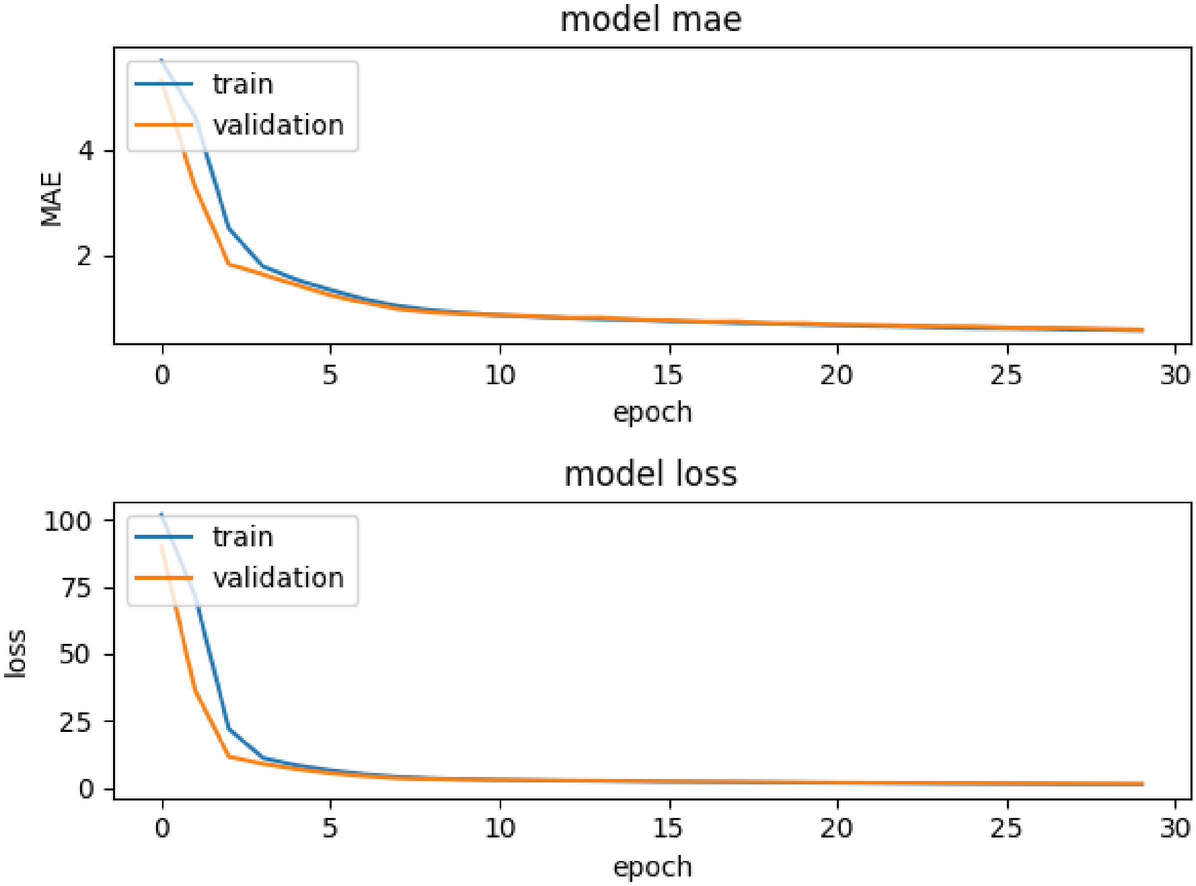


Figure 4.10 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0, cospeciation rate 0, and host-switching rate 0.

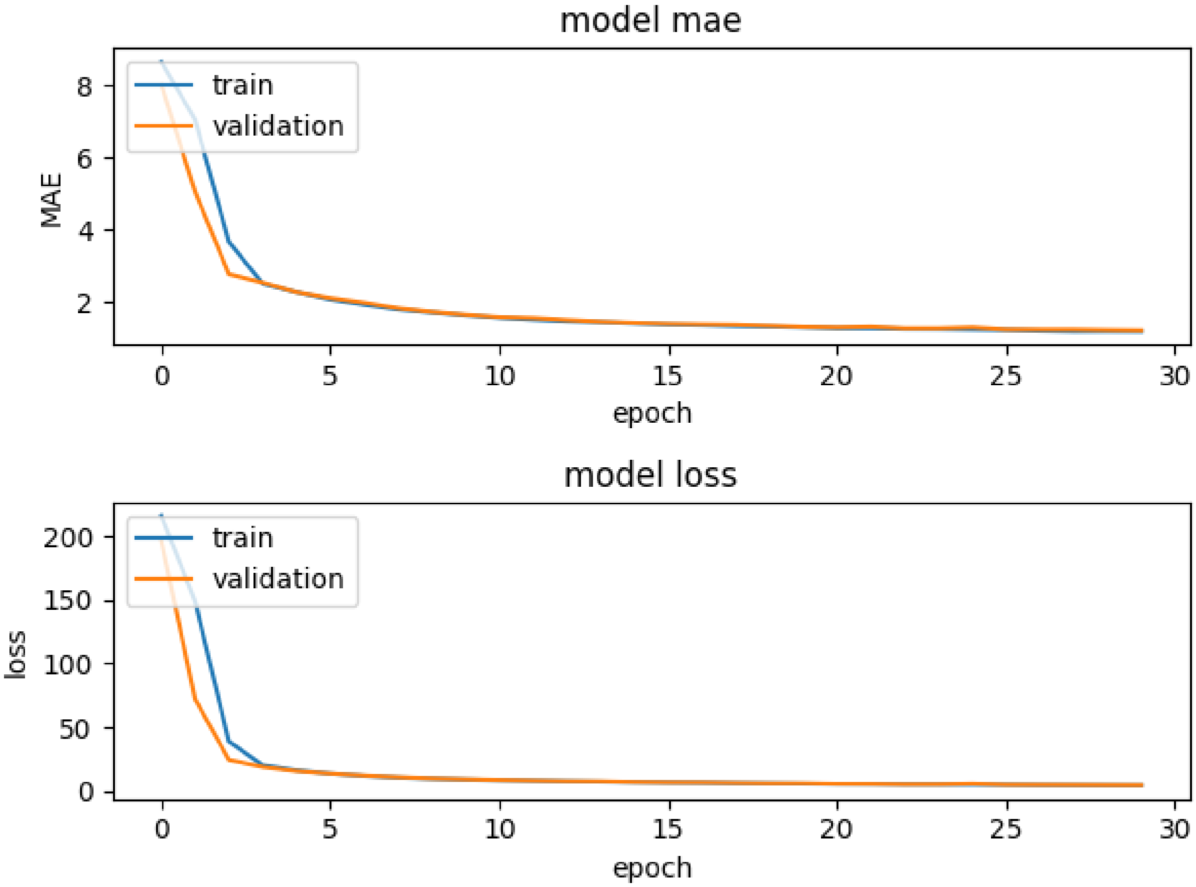


Figure 4.11 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0, cospeciation rate 0, and host-switching rate 1.

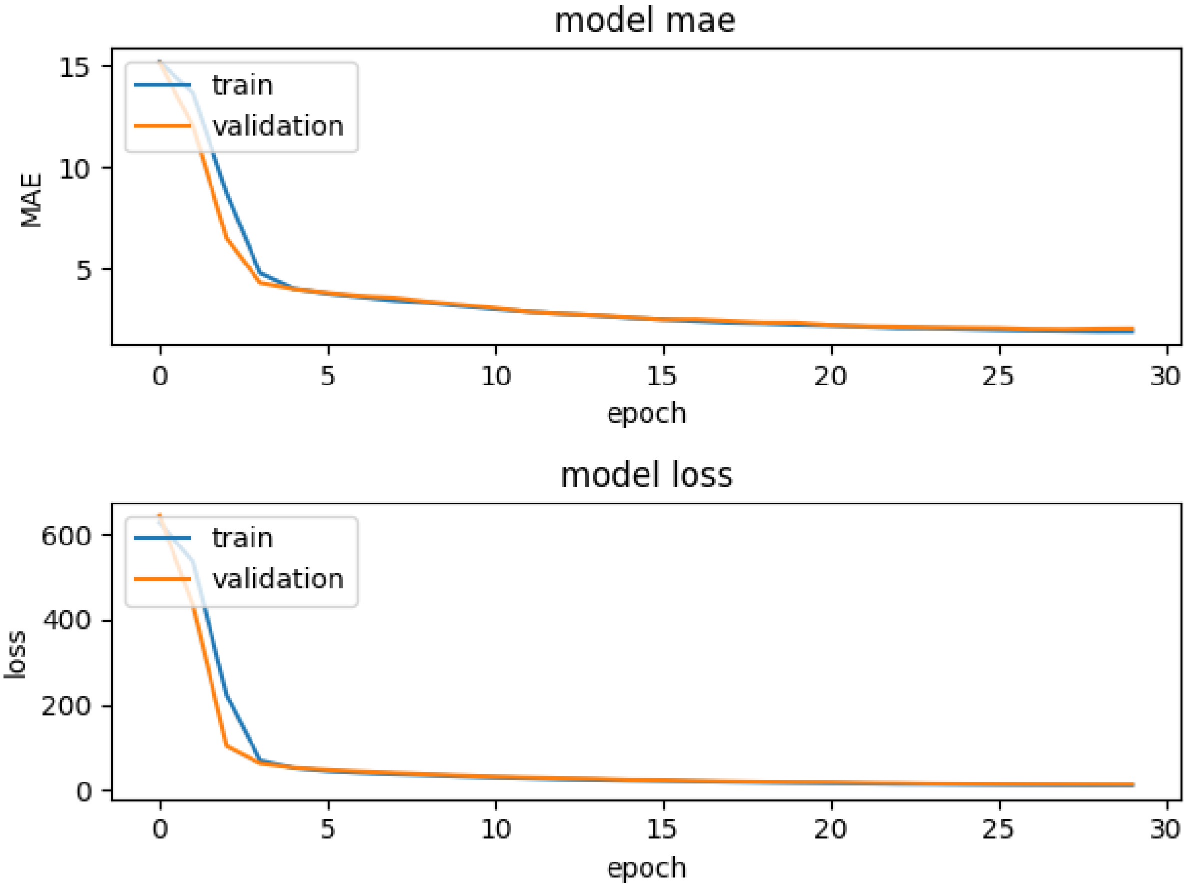


Figure 4.12 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0, cospeciation rate 0, and host-switching rate 2.

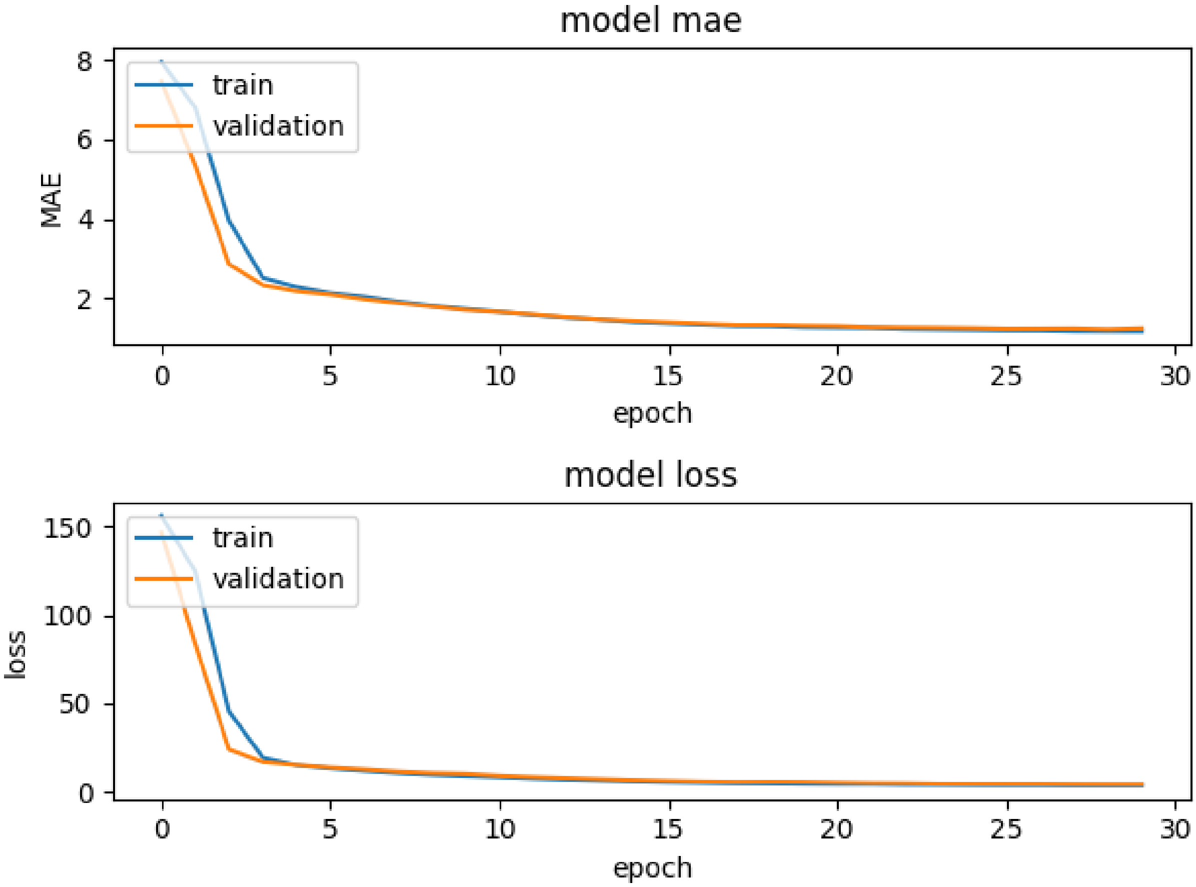


Figure 4.13 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0, cospeciation rate 1, and host-switching rate 0.

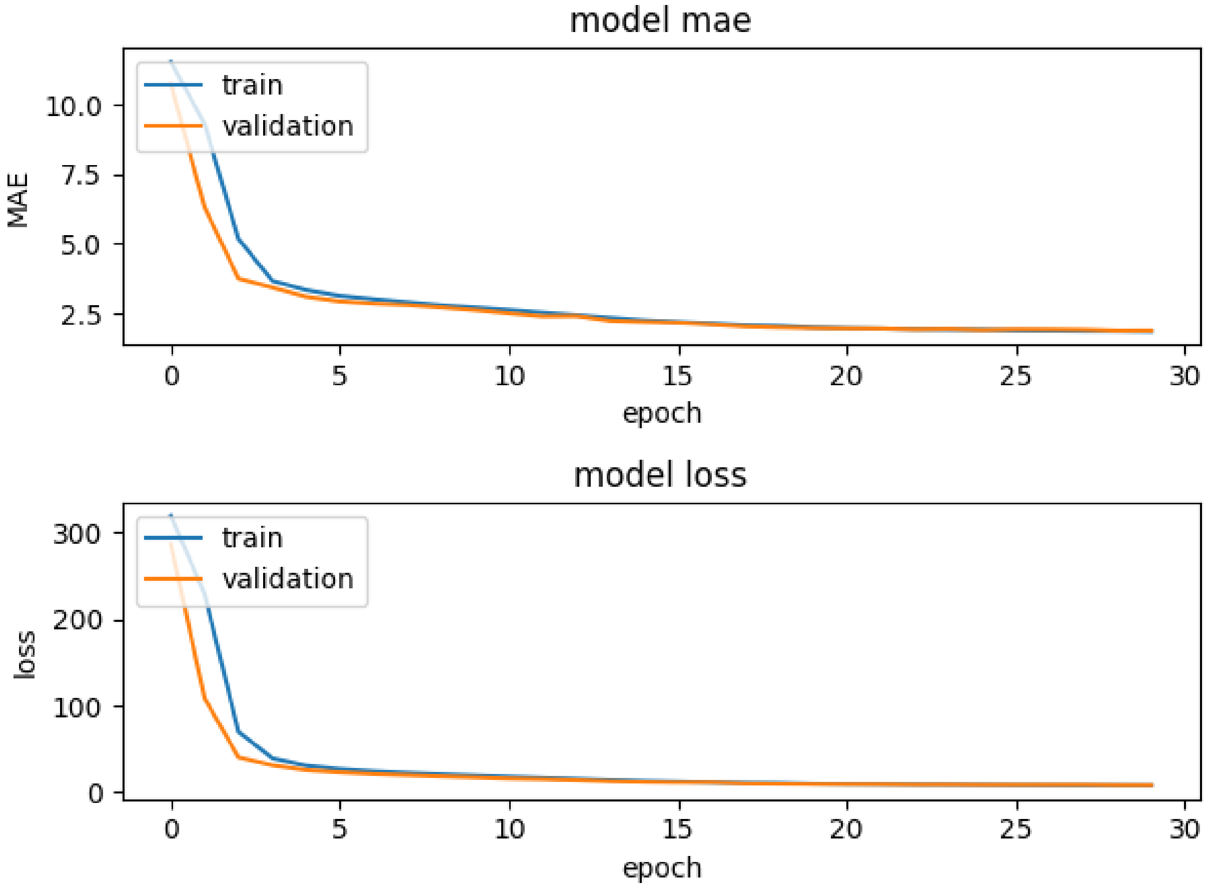


Figure 4.14 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0, cospeciation rate 1, and host-switching rate 1.

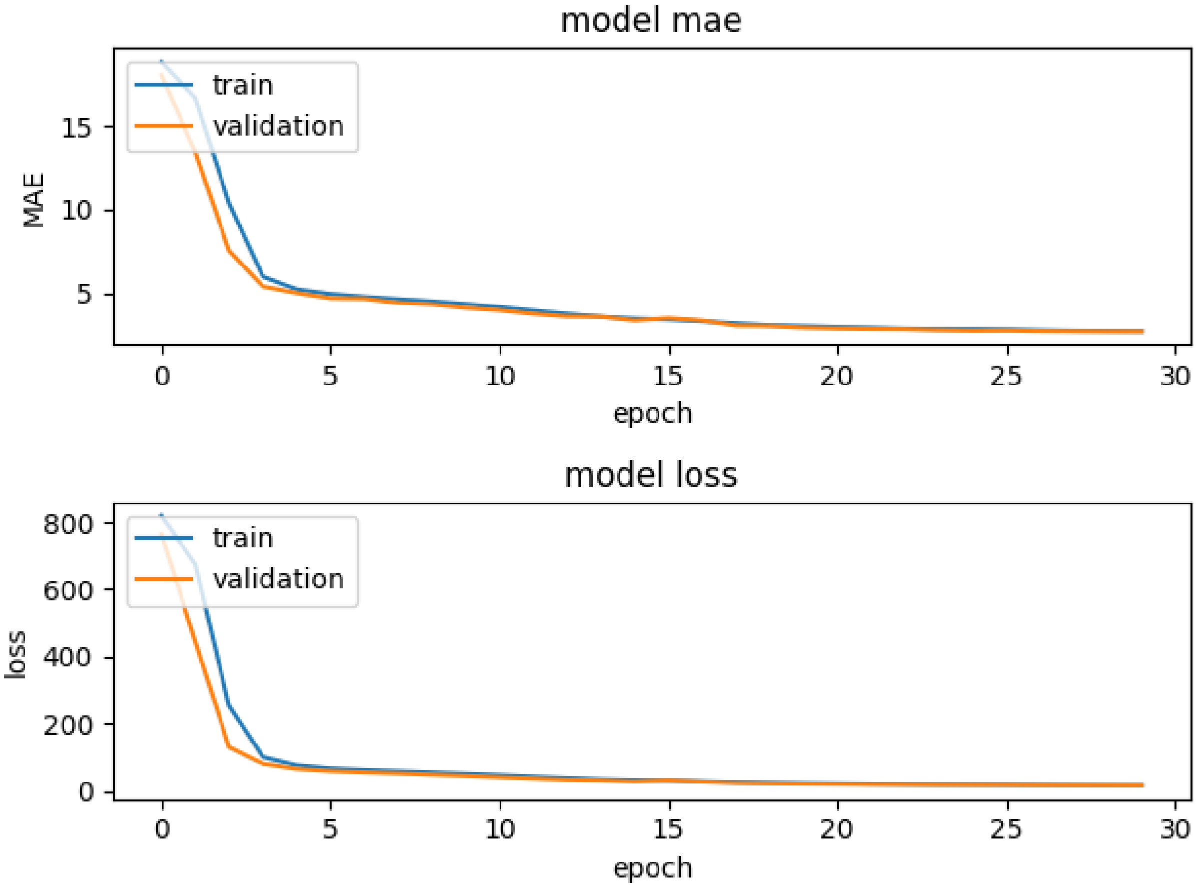


Figure 4.15 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0, cospeciation rate 1, and host-switching rate 2.

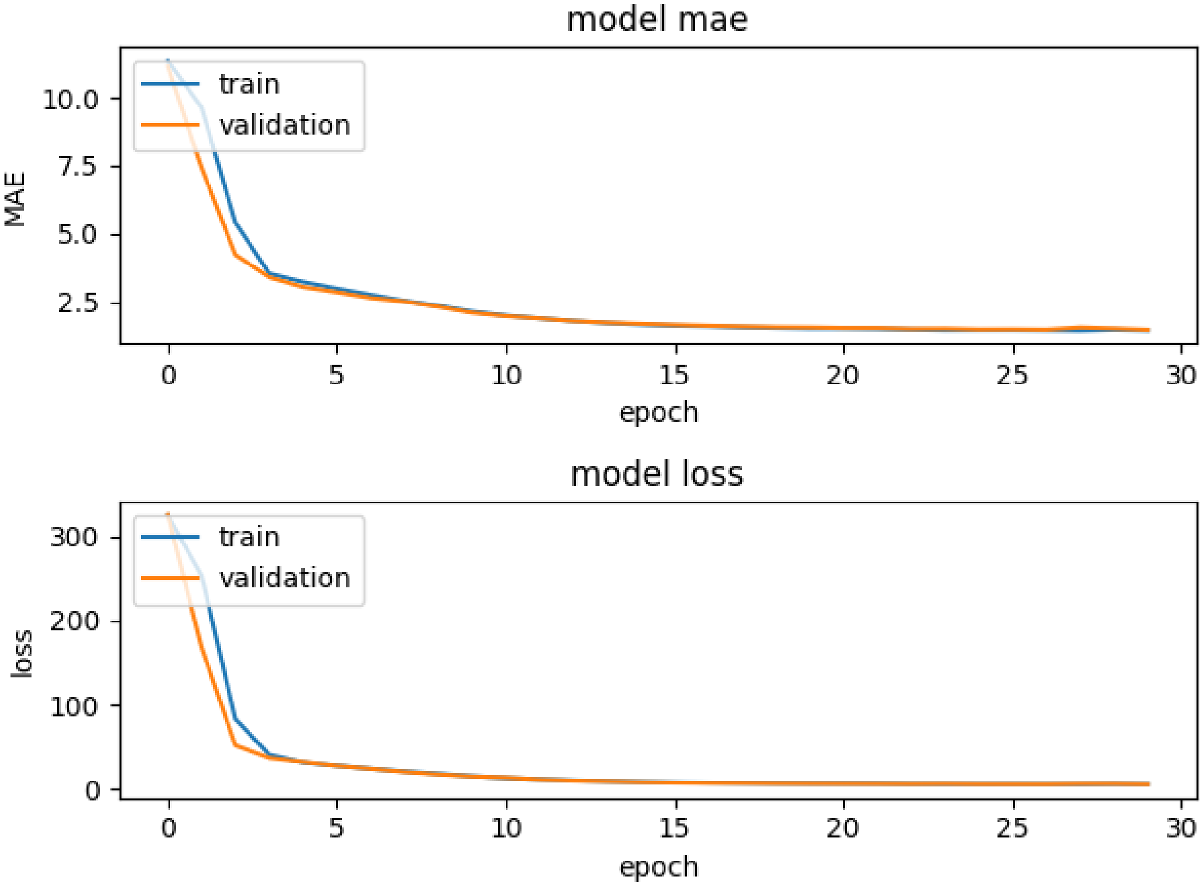


Figure 4.16 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0, cospeciation rate 2, and host-switching rate 0.

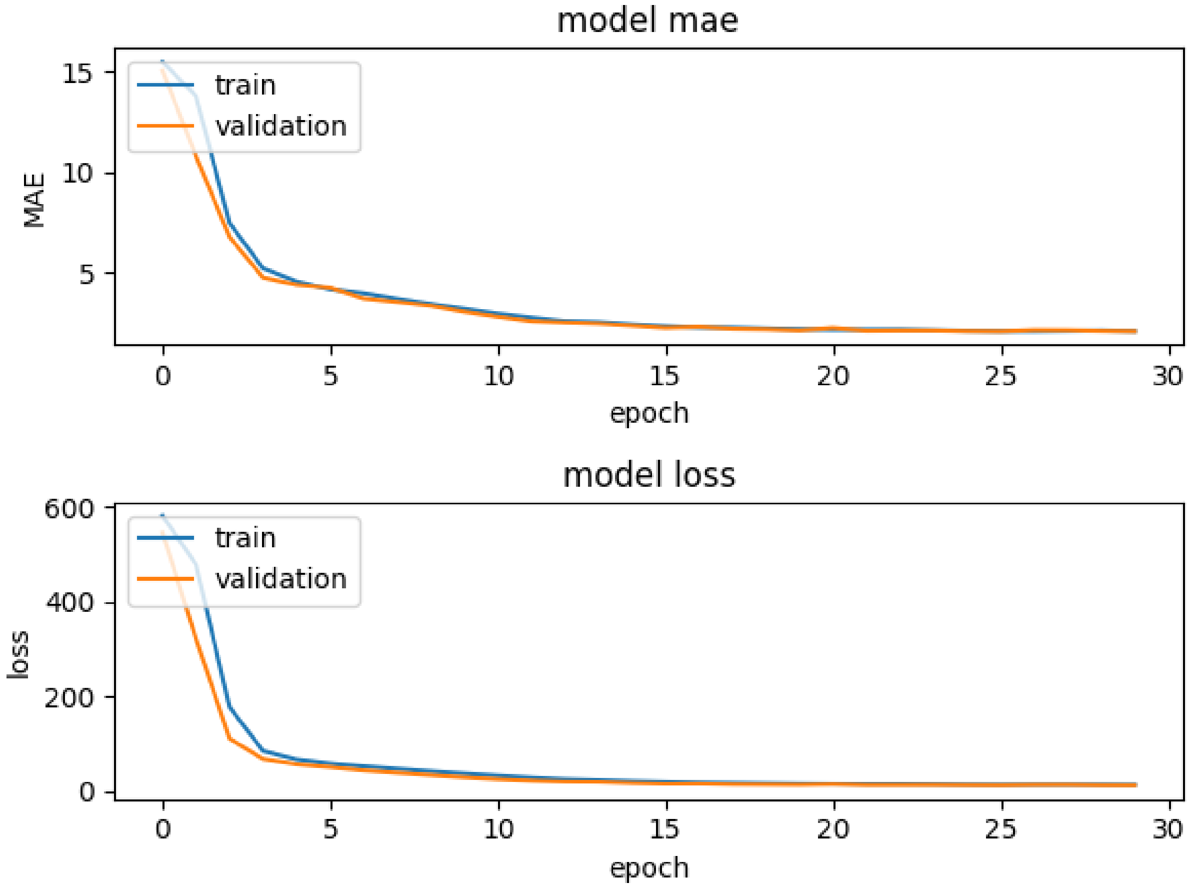


Figure 4.17 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0, cospeciation rate 2, and host-switching rate 1.

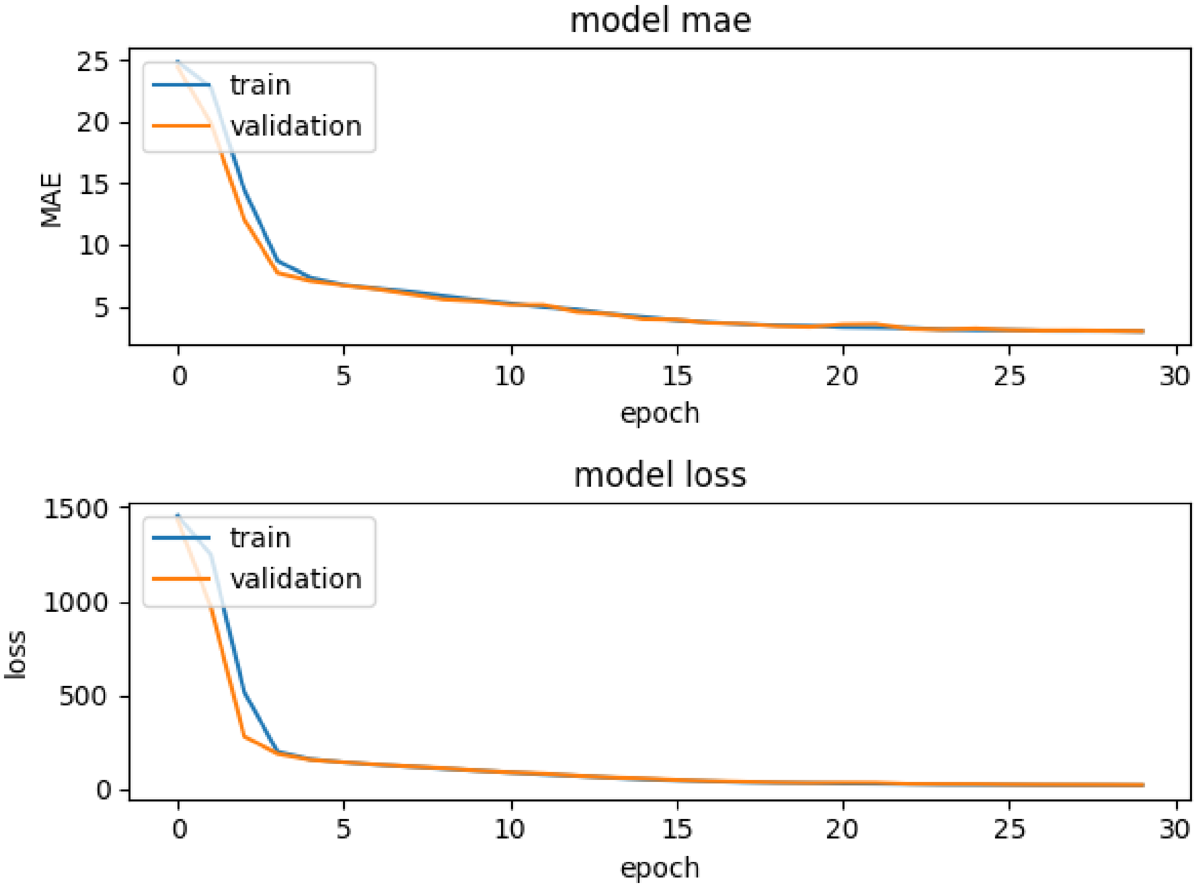


Figure 4.18 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0, cospeciation rate 2, and host-switching rate 2.

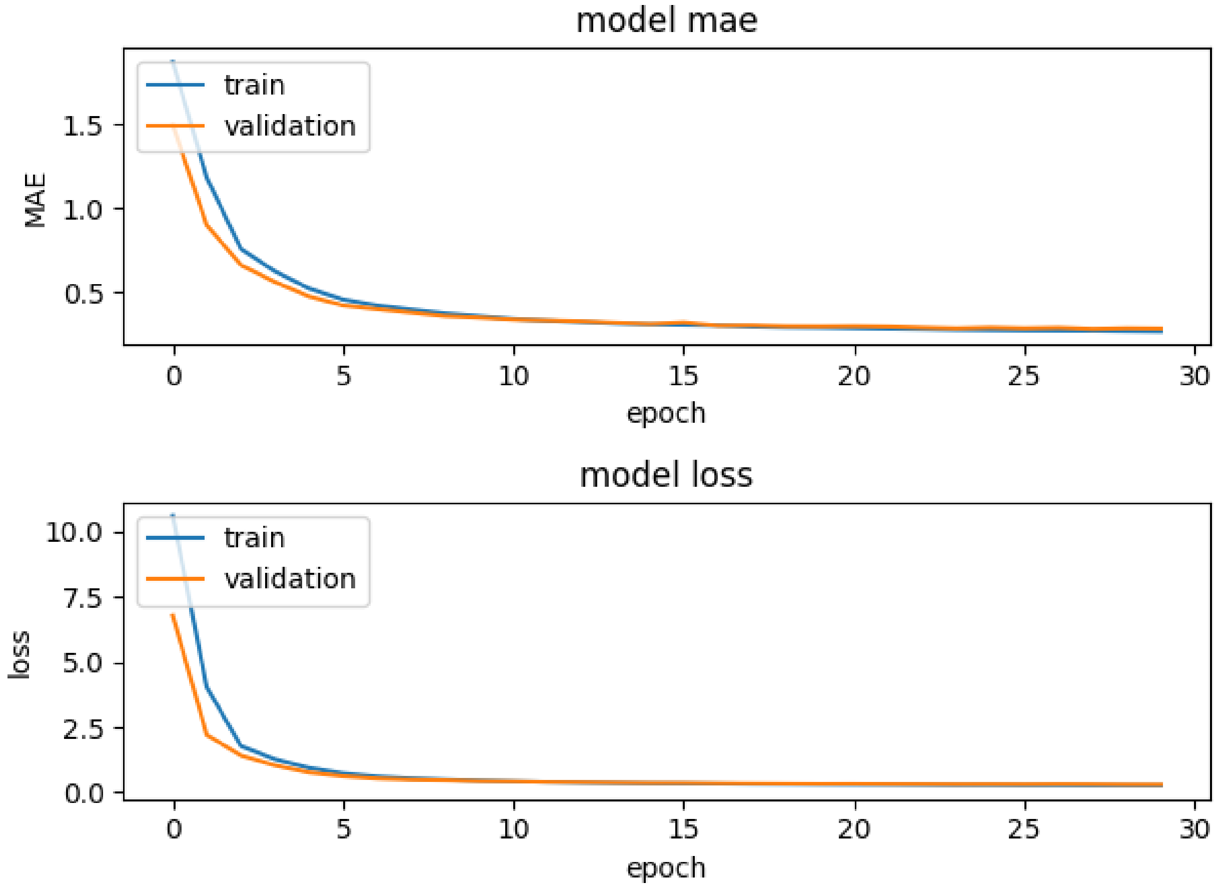


Figure 4.19 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.1, cospeciation rate 0, and host-switching rate 0.

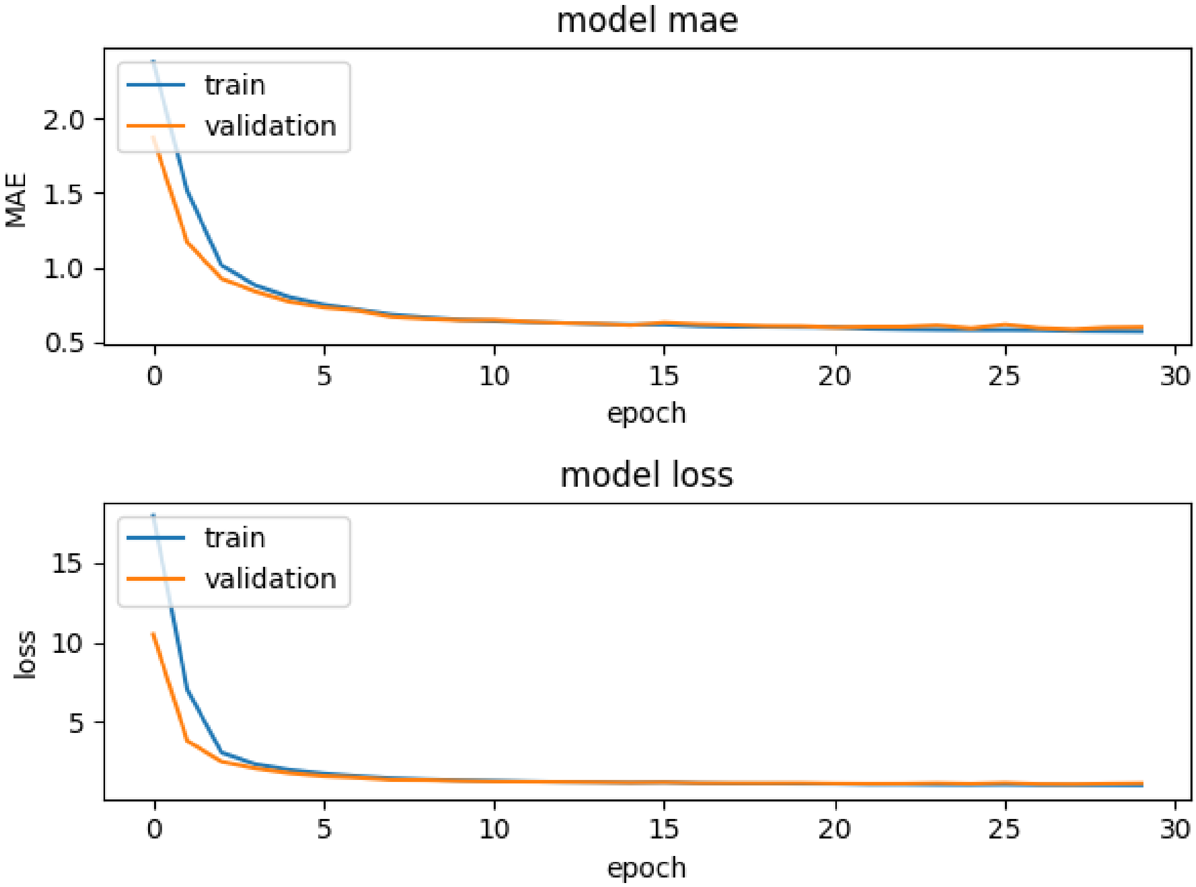


Figure 4.20 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.1, cospeciation rate 0, and host-switching rate 1.

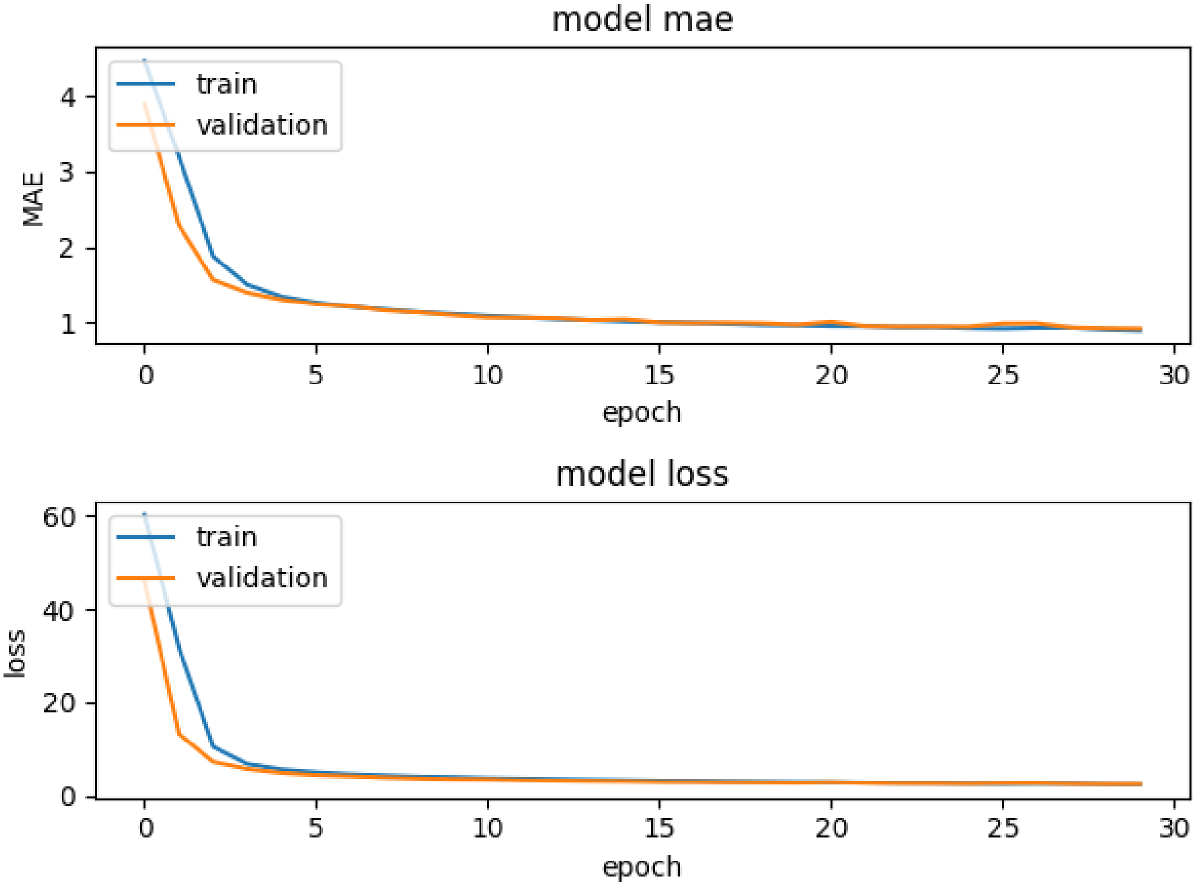


Figure 4.21 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.1, cospeciation rate 0, and host-switching rate 2.

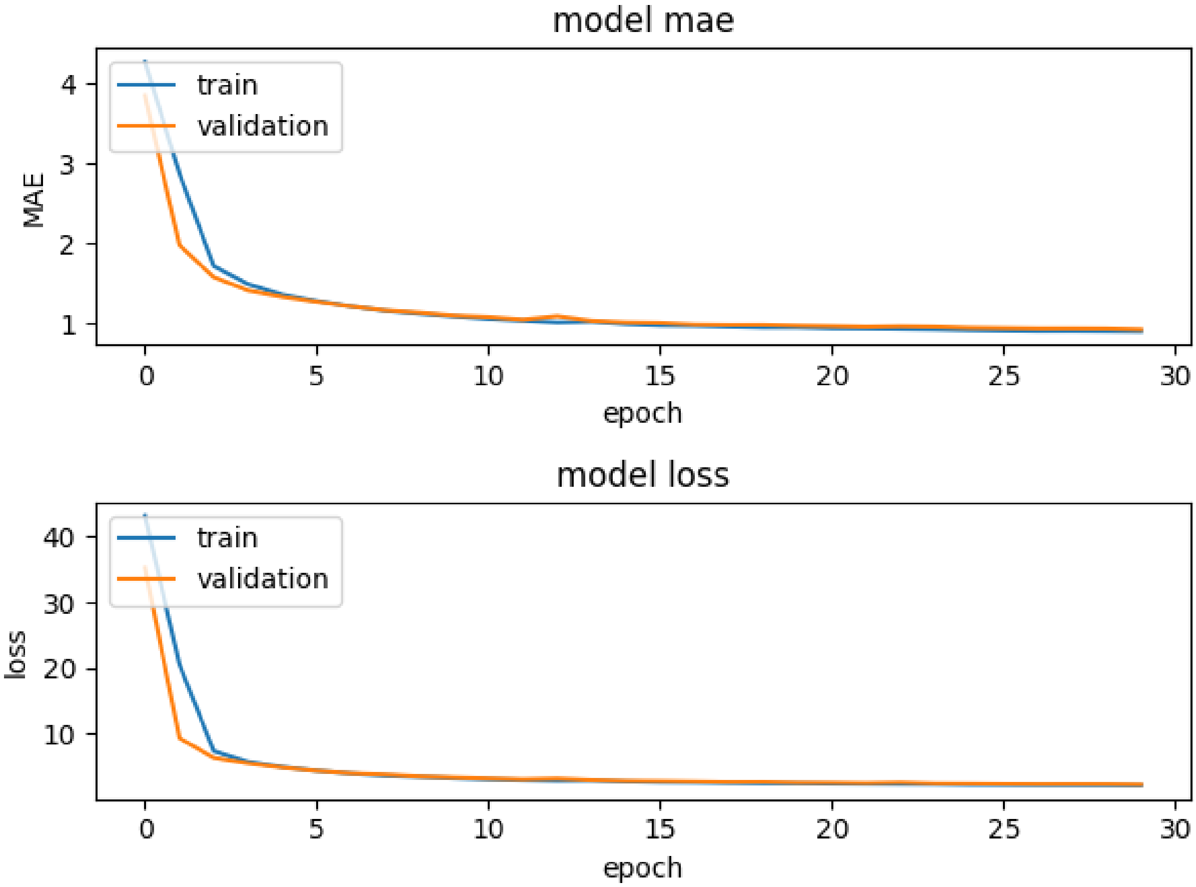


Figure 4.22 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.1, cospeciation rate 1, and host-switching rate 0.

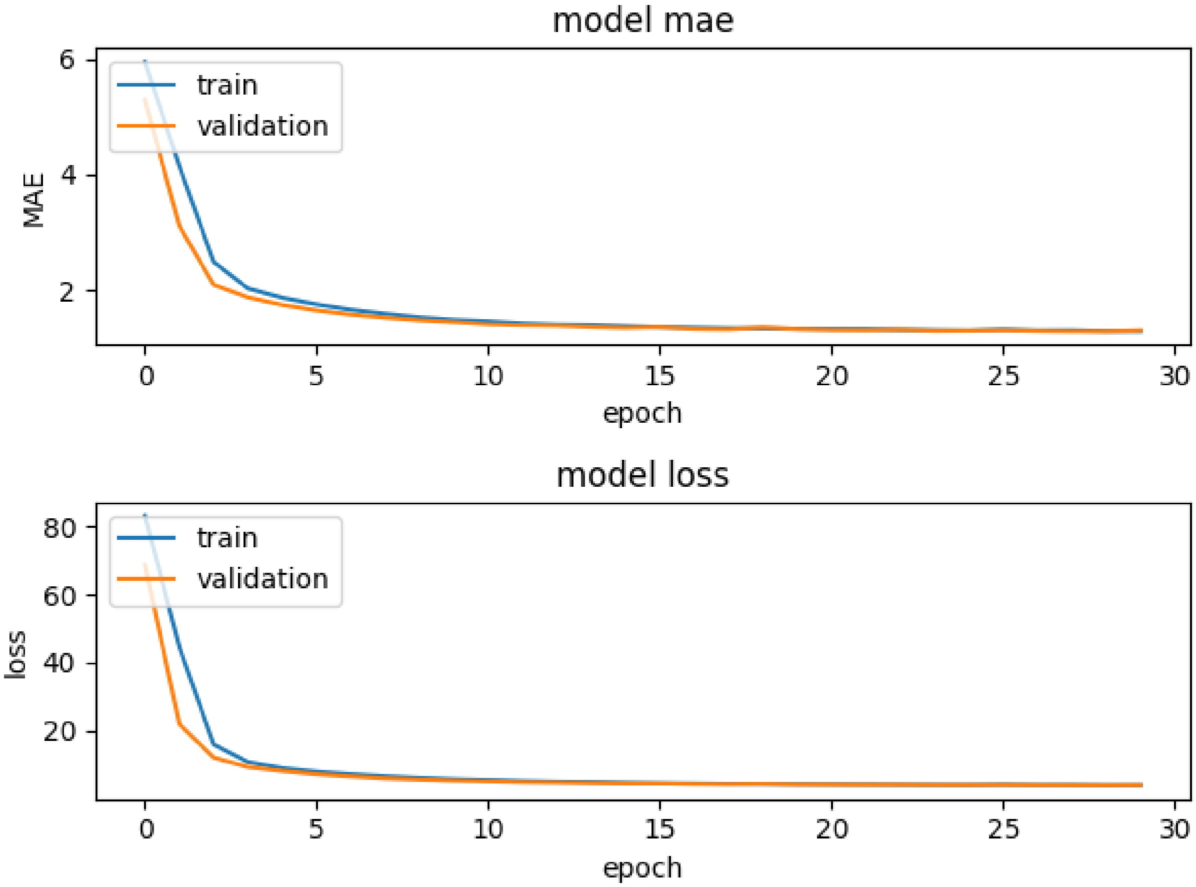


Figure 4.23 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.1, cospeciation rate 1, and host-switching rate 1.

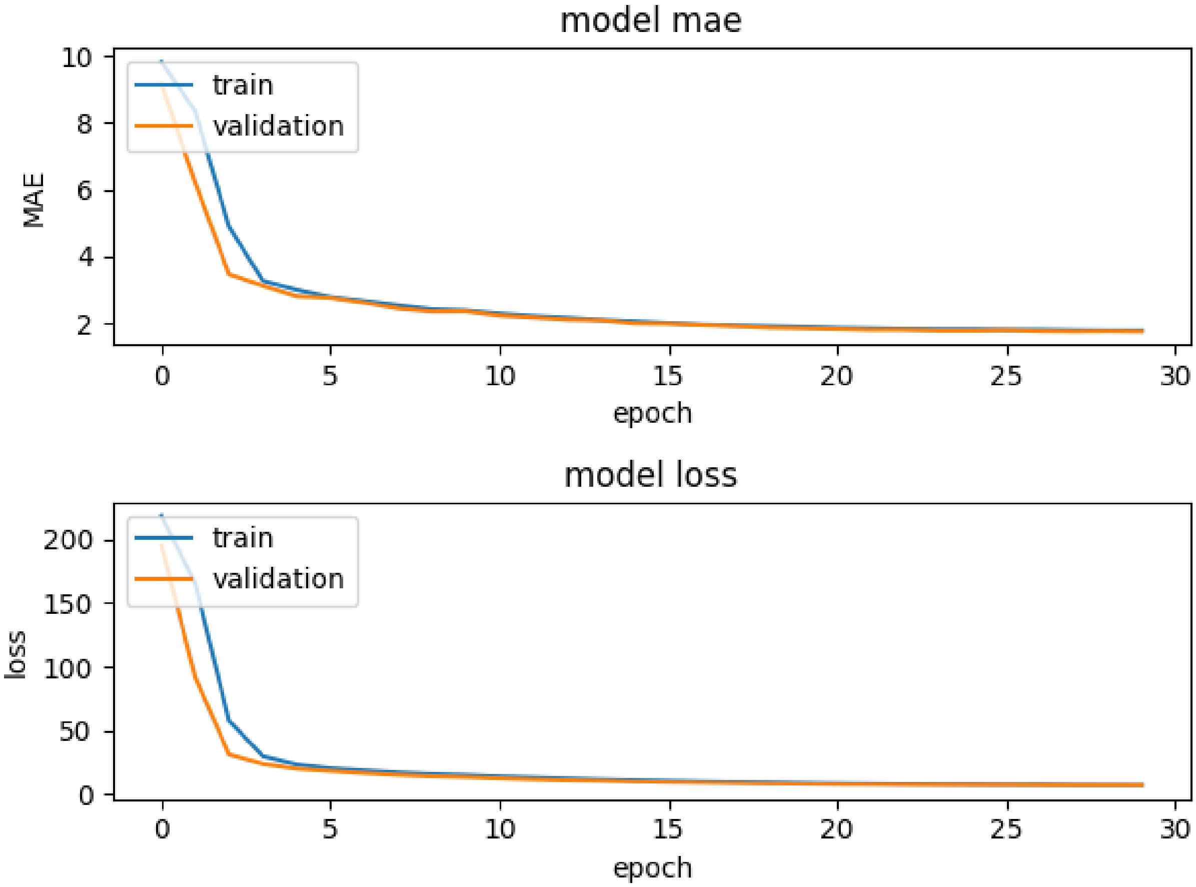


Figure 4.24 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.1, cospeciation rate 1, and host-switching rate 2.

2

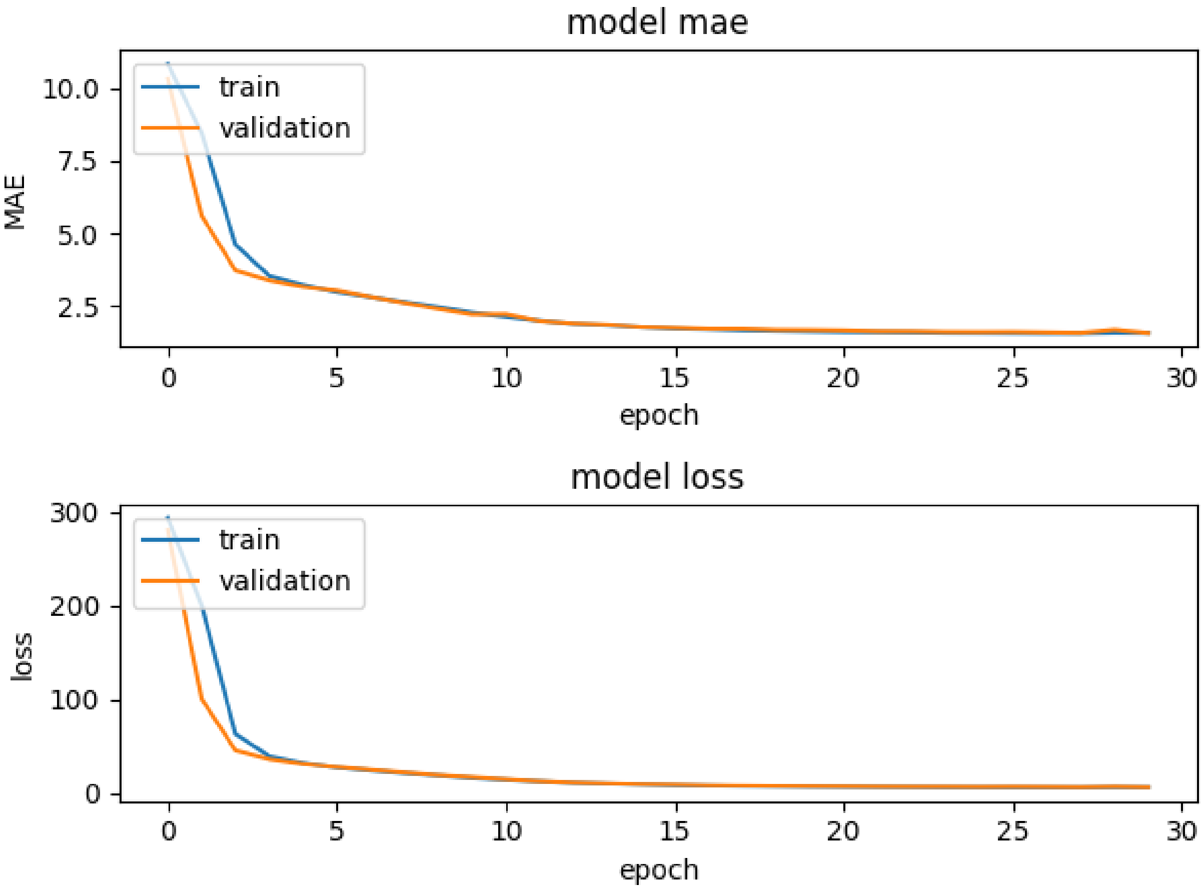


Figure 4.25 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.1, cospeciation rate 2, and host-switching rate 0.

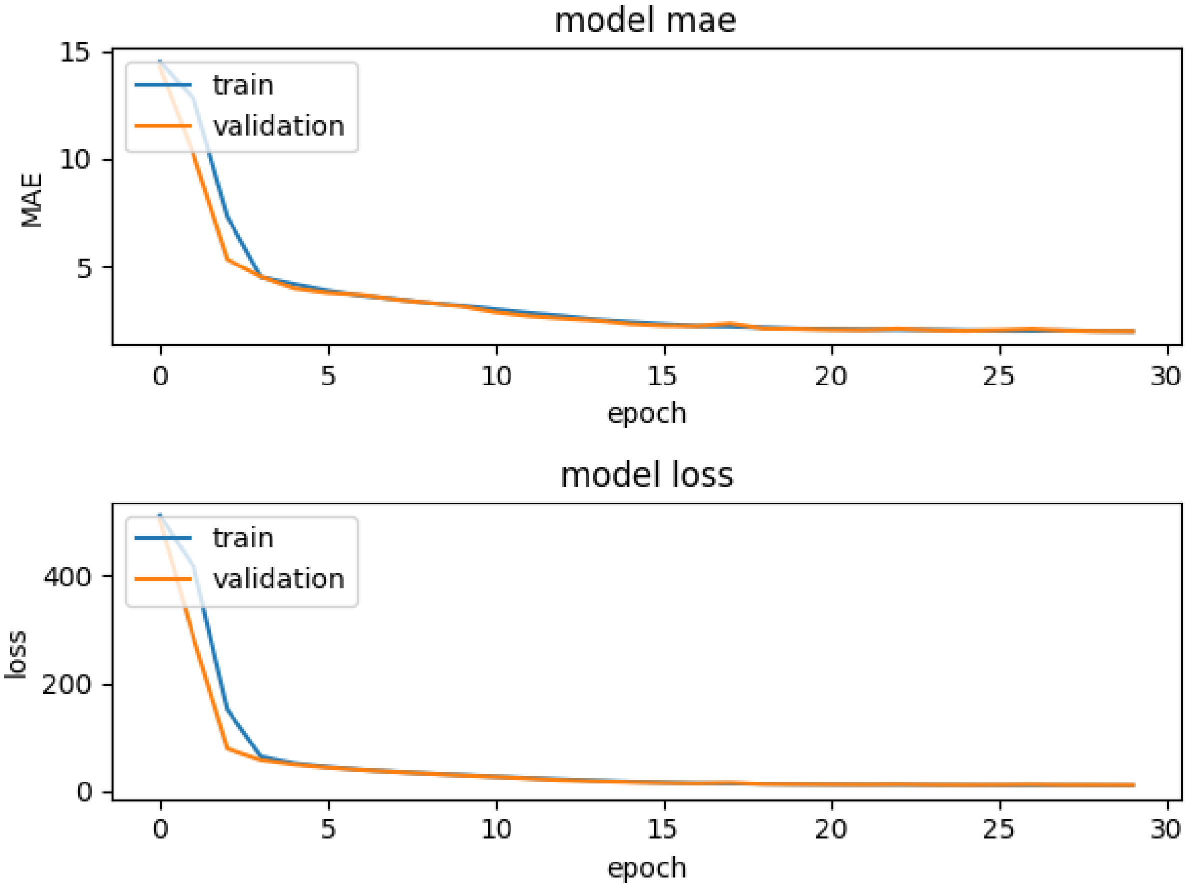


Figure 4.26 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.1, cospeciation rate 2, and host-switching rate 1.

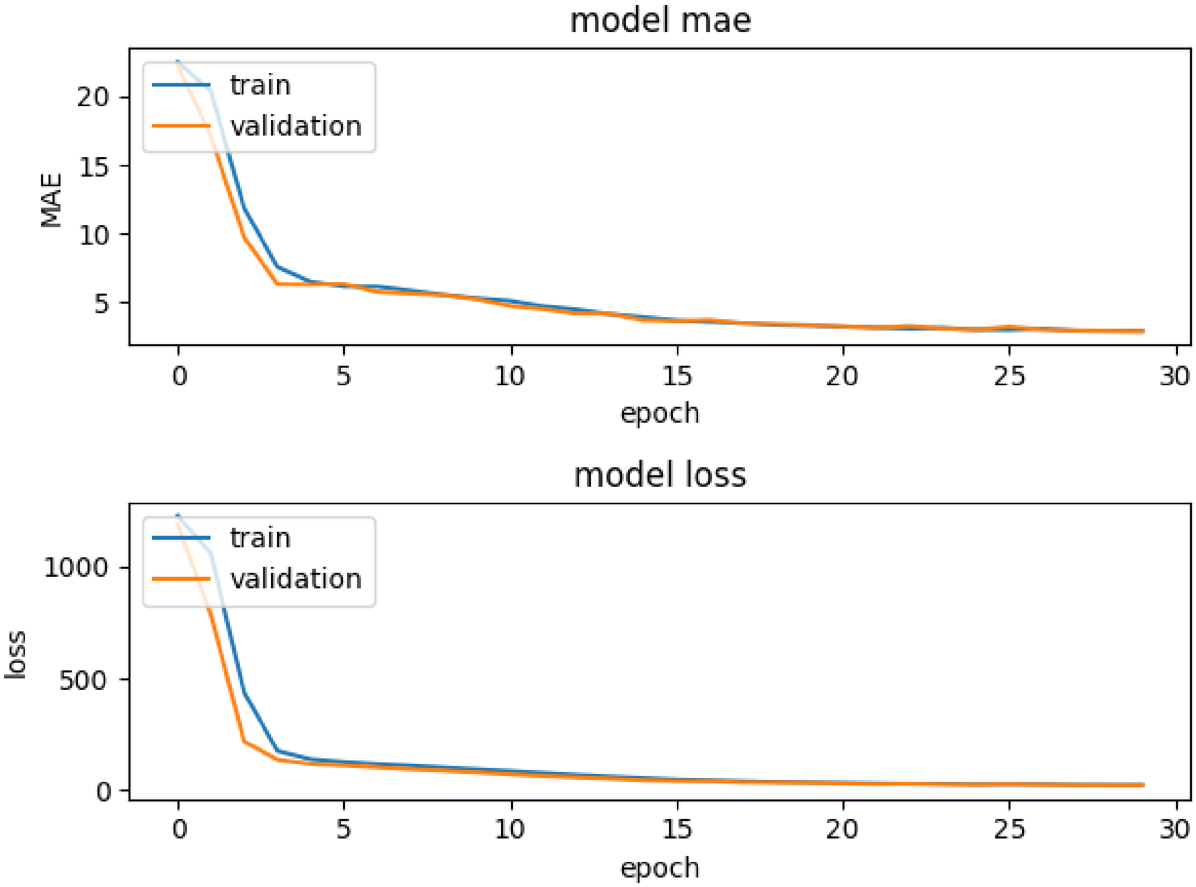


Figure 4.27 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.1, cospeciation rate 2, and host-switching rate 2.

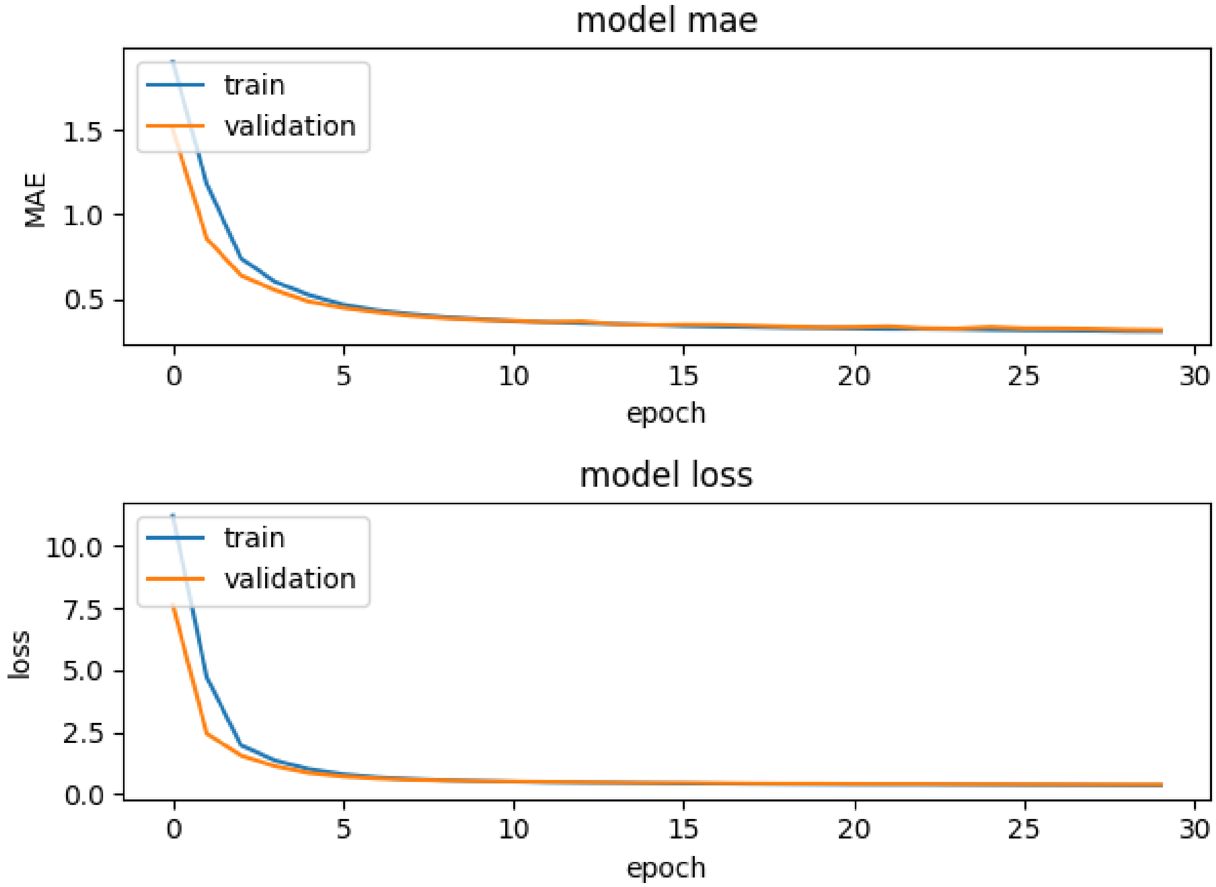


Figure 4.28 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.2, cospeciation rate 0, and host-switching rate 0.

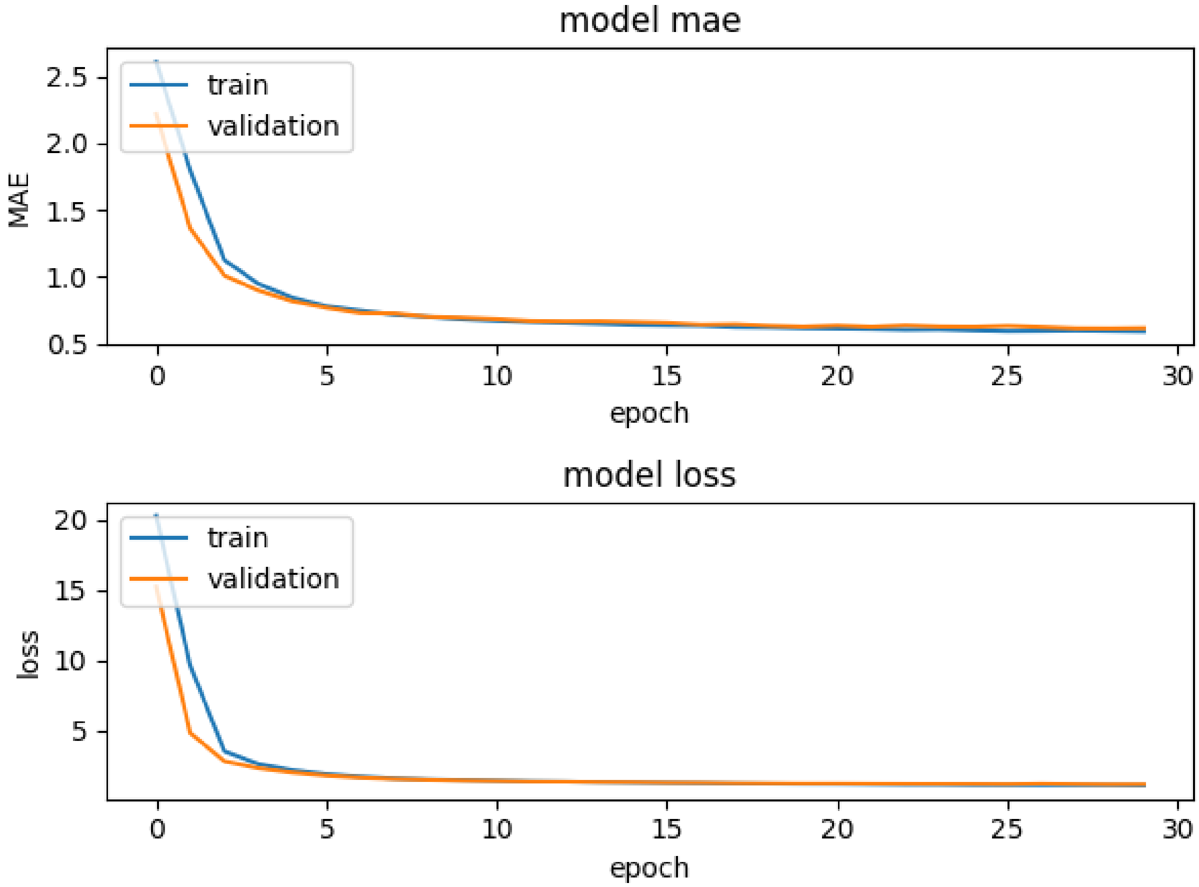


Figure 4.29 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.2, cospeciation rate 0, and host-switching rate 1.

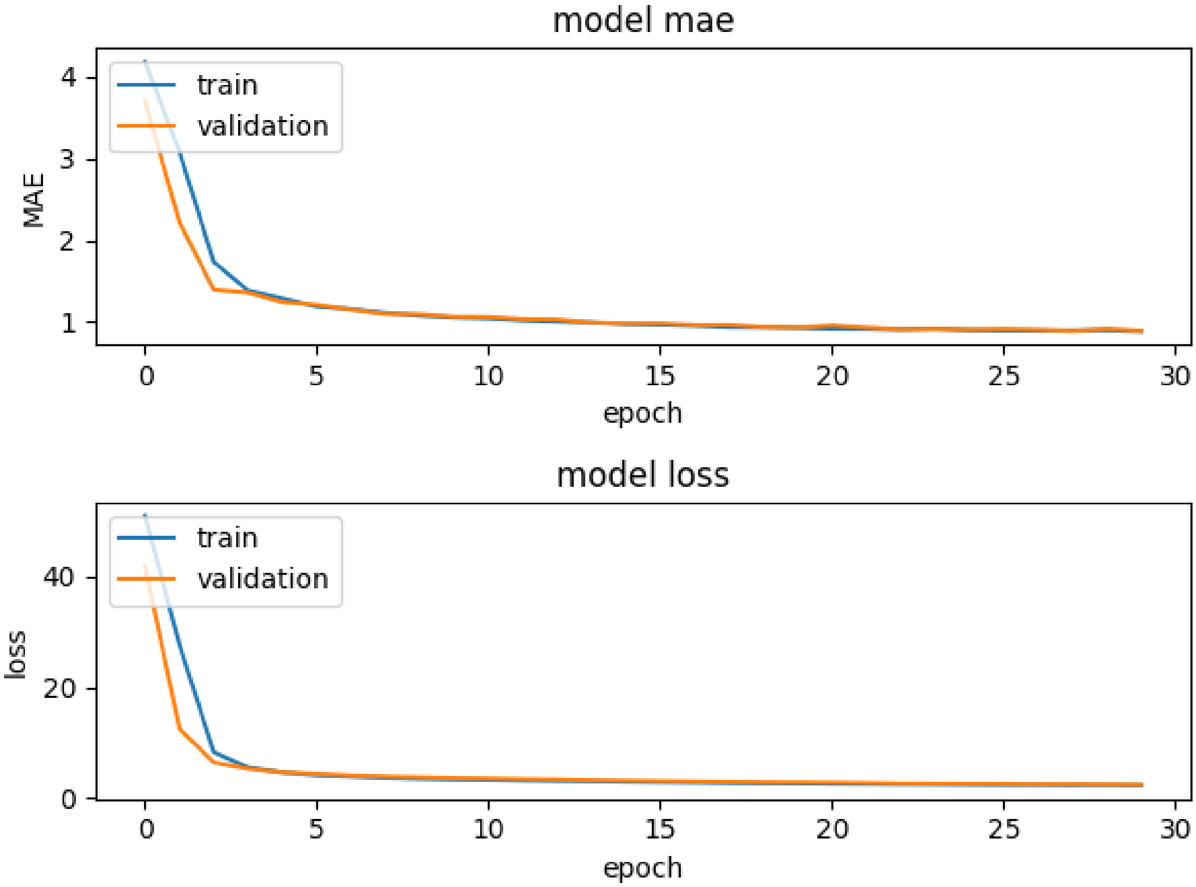


Figure 4.30 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.2, cospeciation rate 0, and host-switching rate 2.

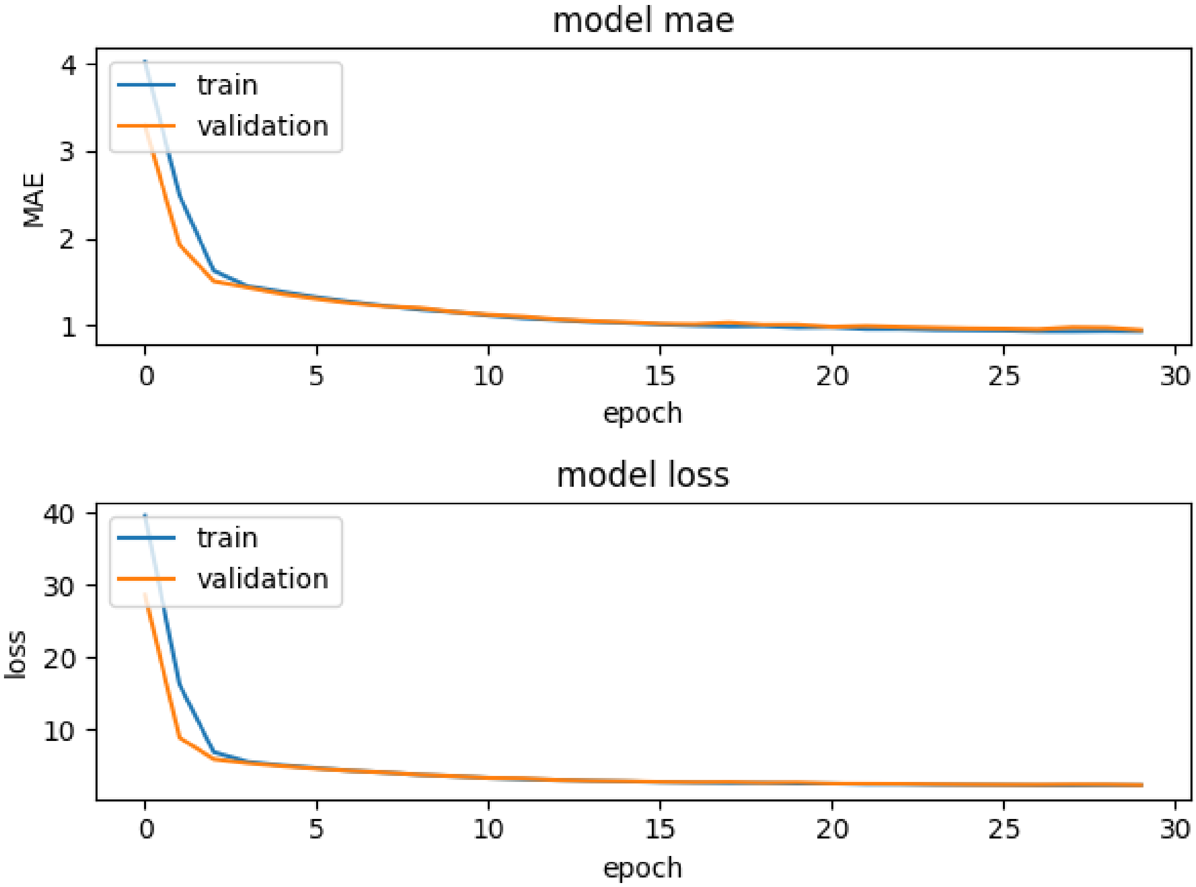


Figure 4.31 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.2, cospeciation rate 1, and host-switching rate 0.

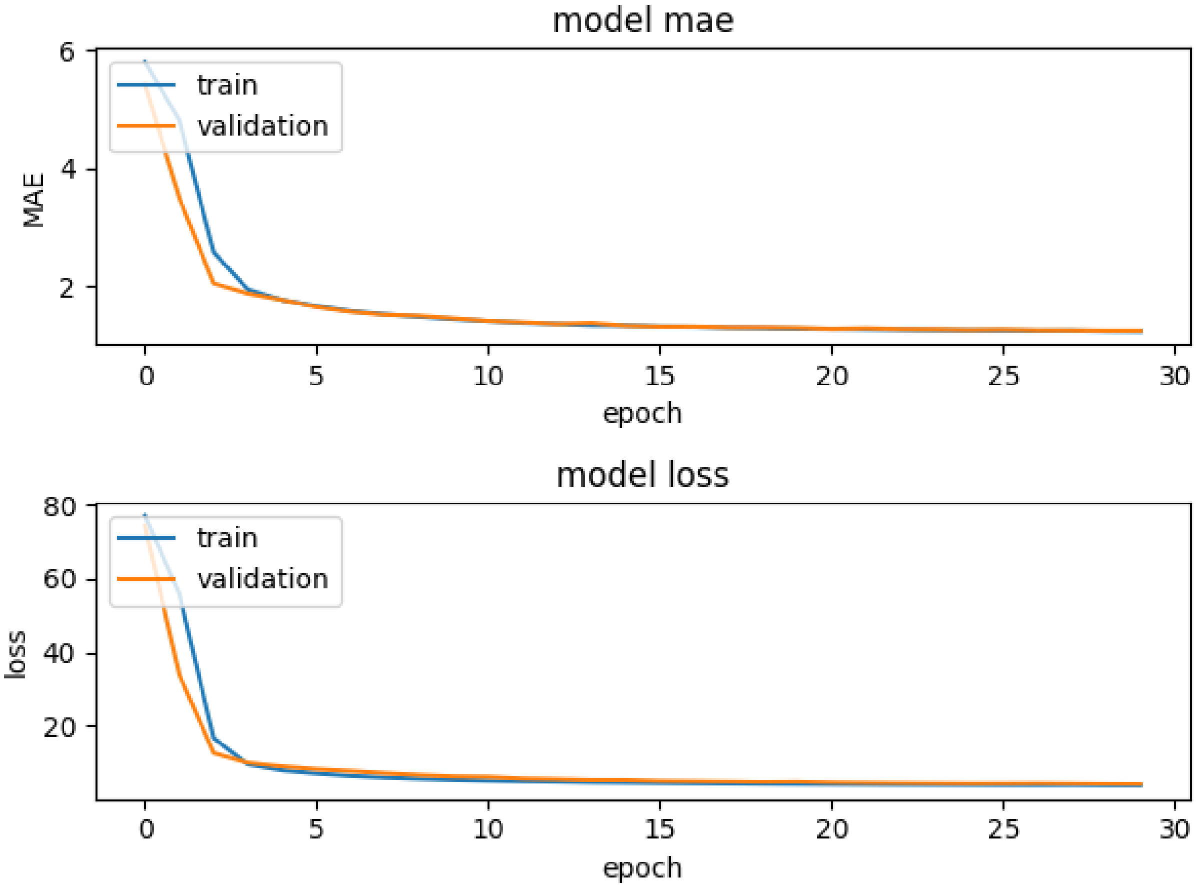


Figure 4.32 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.2, cospeciation rate 1, and host-switching rate 1.

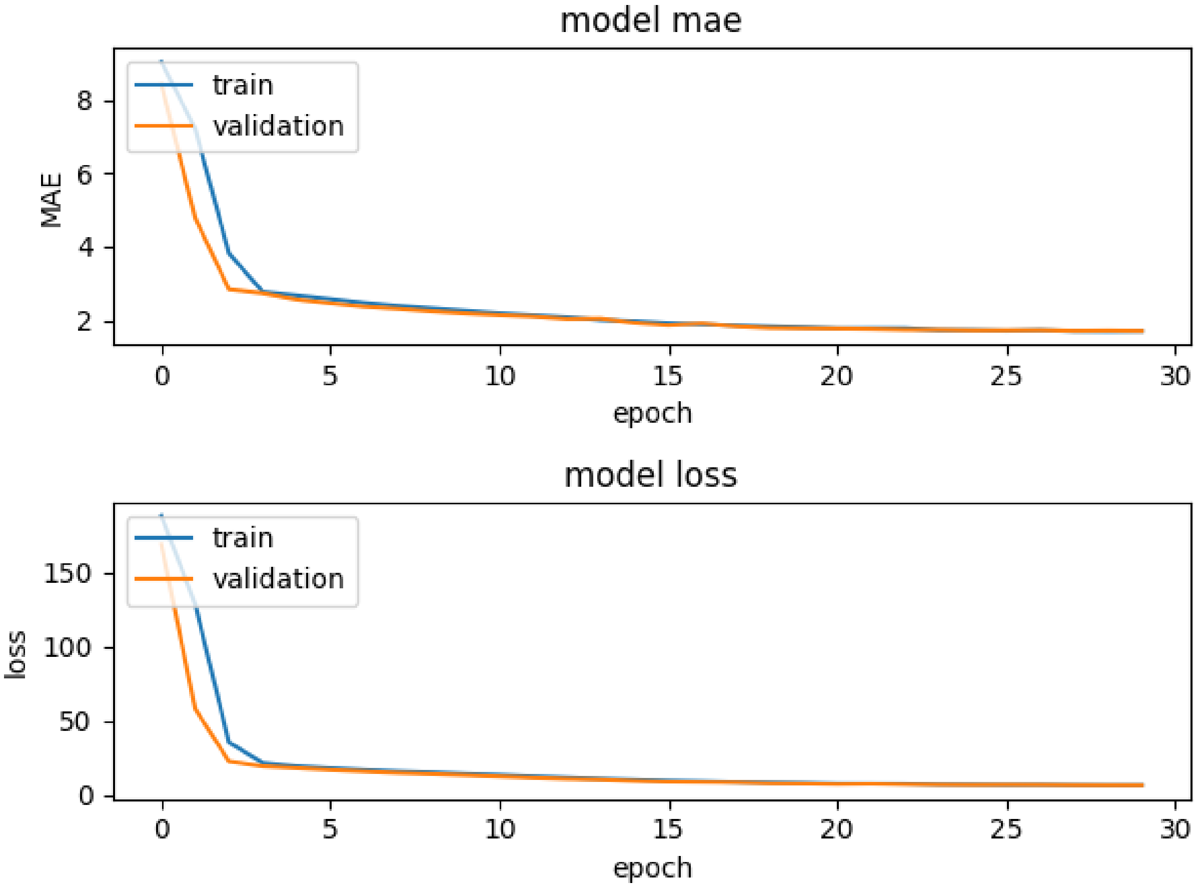


Figure 4.33 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.2, cospeciation rate 1, and host-switching rate 2.

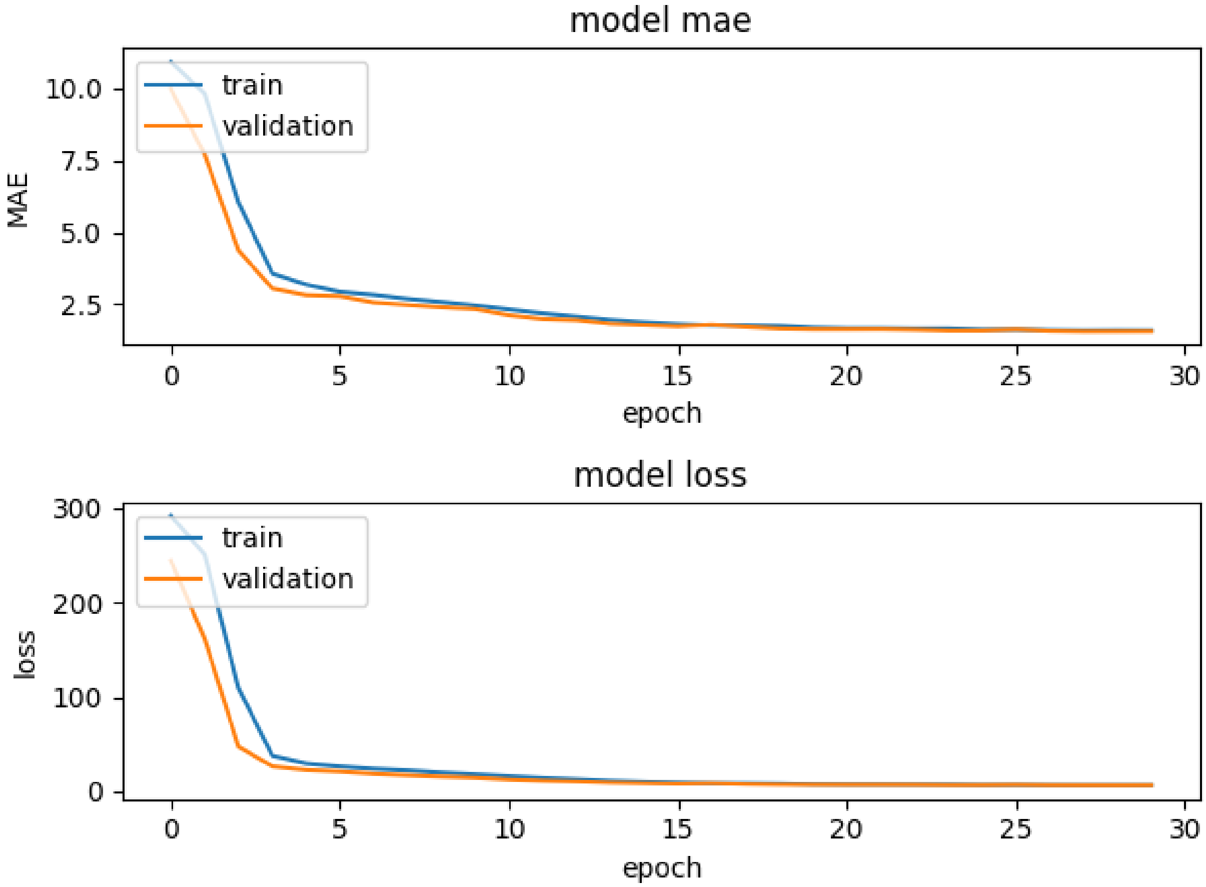


Figure 4.34 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.2, cospeciation rate 2, and host-switching rate 0.

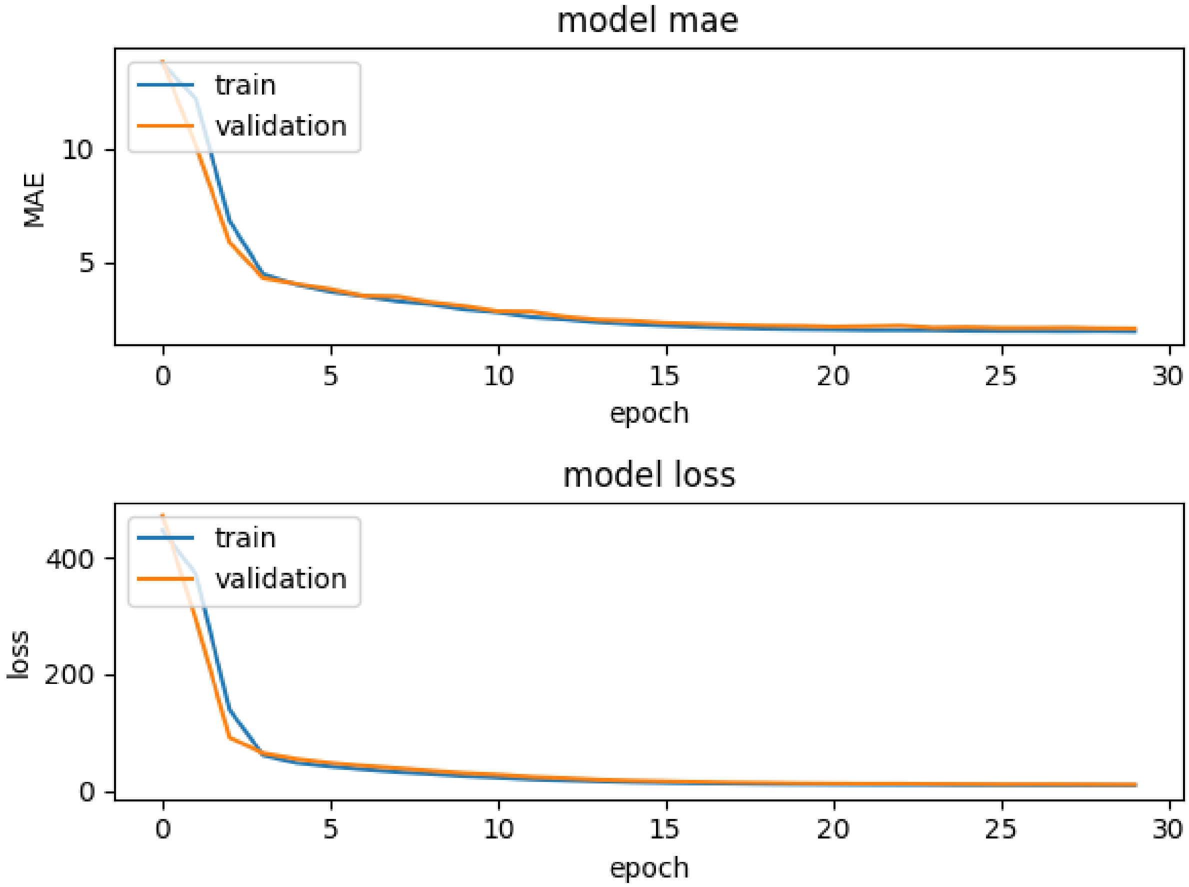


Figure 4.35 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.2, cospeciation rate 2, and host-switching rate 1.

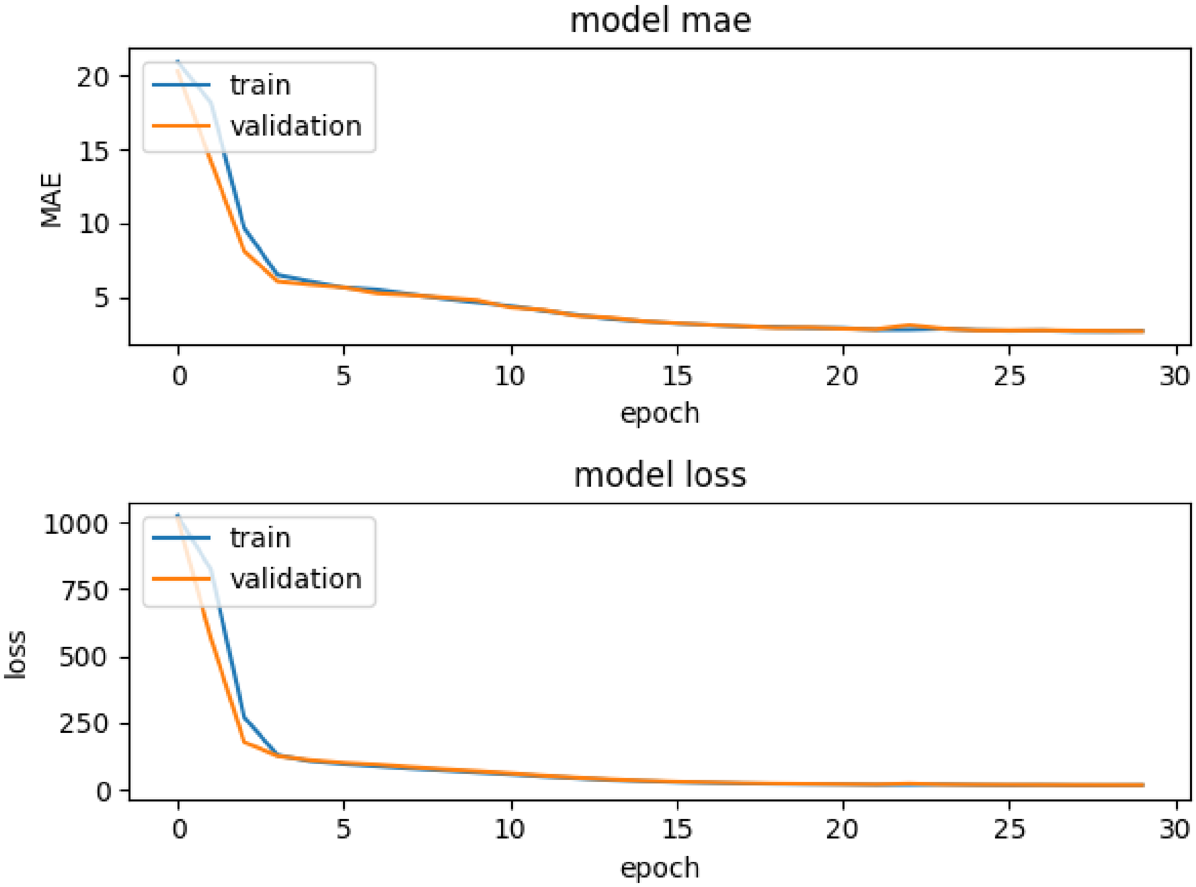


Figure 4.36 Mean absolute error and loss plotted by epoch for simulation settings: extinction rate 0.2, cospeciation rate 2, and host-switching rate 2.

CHAPTER 5. USING A HISTORICAL BIOGEOGRAPHY MODEL TO ESTIMATE PARAMETERS OF THE COPHYLOGENETIC BIRTHDEATH PROCESS

Wade Dismukes[1] and Tracy A. Heath[1]

[1] *Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Bessey Hall, Ames, Iowa, 50010, USA*

5.1 Abstract

Historical biogeography and host-symbiont evolution have a long history of sharing methodologies. While historical biogeography has begun using probabilistic generative models, host-symbiont methods rely on parsimony and pattern-based methods. This study uses an area cladogram and a species tree inhabiting those areas as an analog for a host tree with interacting symbionts. We use a Bayesian implementation of the dispersal-extinction-cladogenesis model to estimate parameters of a host-symbiont model. However, we found little information within host-symbiont data suggesting that this historical biogeography model does not align perfectly with the host-symbiont data.

5.2 Introduction

Symbioses are pervasive throughout the Tree of Life representing more than 40% of life on Earth (Dobson et al., 2008, Wang and Qiu, 2006, National Research Council, 2007). Despite this pervasiveness, our understanding of the processes that generate these highly dependent systems remains poorly understood (Weber et al., 2017). Cophylogenetic methods provide a computational framework that measures the support for one or more hypotheses through the unified analysis of cophylogenetic data consisting of host and symbiont phylogenies and the interactions among their taxa (Dismukes et al., 2022).

There is a long history tying together historical biogeography methods, gene-tree/species-tree analysis, and cophylogenetic methods (Page, 2003). Indeed, Brooks parsimony (Brooks, 1981) was originally intended for use on cophylogenetic data consisting of host and parasites despite being more well known for its use in historical biogeography analyses. For historical biogeography, an area cladogram is compared with a species tree that inhabits those areas, for gene tree/species tree comparisons, gene trees are compared with species trees, and for cophylogenetics, the host tree is compared with the symbiont tree. We might assume that these different tree structures would be identical in all of these cases and that discordance between them is caused by certain events. Discordance between area cladograms and species trees is caused, at least in part, by dispersal to new areas, extirpation within areas, speciation, and extinction within areas, and how the species "inherits" the areas when they split apart (Ree and Smith, 2008a). For gene and species tree, the significant causes of discordance are gene duplication, gene loss, lateral gene transfer (LGT), and incomplete lineage sorting (ILS) (Maddison, 1997). In cophylogenetics, discordance between host and symbiont phylogenies is attributed to symbiont speciation (also called duplication), symbiont extinction (also called loss), and host-switching (Dismukes et al., 2022). While these events differ fundamentally in how they happen, the patterns they leave among these tree structures can be similar (Page and Charleston, 1998).

Despite these similarities, the fields have since diverged in methodology. Gene-tree/species-tree methods have embraced the wide availability of genomic data and developed models that describe these discordance-causing processes (Rannala and Yang, 2003, Heled and Drummond, 2009, Rasmussen and Kellis, 2012, Szöllosi et al., 2013).

Similarly, in historical biogeography methods, probabilistic models have been developed such as the dispersal-extinction-cladogenesis (DEC) model (Ree and Smith, 2008a, Smith et al., 2008, Landis et al., 2013, Matzke, 2014). Cophylogenetics, on the other hand, continues to see development in newer parsimony methods (Santichaivekin et al., 2021) and pattern-based statistics that address the overall congruence of host and symbiont phylogenies (Balbuena et al., 2020).

More recently, Satler et al. (2019) used a gene tree/species tree model to perform an analysis on cophylogenetic data and Braga et al. (2020) adapted a biogeography model for analyzing host-use evolution among Nymphalid butterflies. Only one of these provides a generative model for host and symbiont systems (Braga et al., 2020), and a more general generative model is developed in (Dismukes and Heath, 2021). However, at present the cophylogenetic birth-death model lacks a method of inference.

Here we used a Bayesian implementation of the DEC model (Landis, 2017) to estimate the parameters analogous to those of the cophylogenetic birth-death model. We use an analogy to interpret these results, where our host tree is an area cladogram, the species tree of the biogeographic model is the symbiont tree, and the biogeographic events are interpreted as cophylogenetic events (see Figure 5.1). We used a simulation experiment to assess the results of using this DEC model.

5.3 Methods

5.3.1 Cophylogenetic birth-death model

We used the cophylogenetic birth-death model (Dismukes and Heath, 2021) to simulate data for our experiments. The cophylogenetic birth-death model allows simulation of host-symbiont evolution and produces cophylogenetic datasets consisting of a host phylogeny, a symbiont phylogeny, and an association matrix describing the interactions

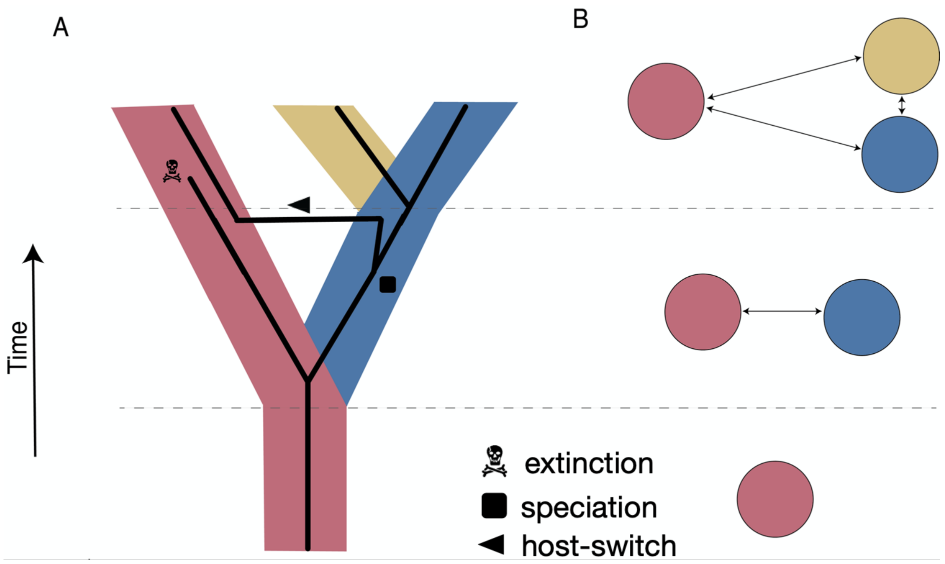


Figure 5.1 Analogy between biogeography and cophylogeny. A. Host tree represented as the thicker, colored tree with inset symbiont tree (thin, black). Cophylogenetic events are labeled. B. The host species as areas corresponding to those in A. Order of events: The symbiont species begins living within the red host. Host species branches into a blue host, symbiont species speciates with one on the red host and one on the blue host (cospeciation/allopatric speciation). The symbiont living on the blue host speciates (symbiont speciation). The new symbiont lineage disperses to the red host (host-switching). The original symbiont on the red host goes extinct (symbiont extinction). Another cospeciation event occurs producing yellow host and new symbiont lineages.

between extant hosts and symbionts. This model contains rates for the following cophylogenetic events: symbiont speciation, symbiont extinction, host speciation, host extinction, cospeciation, host-switching, symbiont dispersal, and symbiont extirpation. Cospeciation is defined as a host speciation that accompanies speciation of a symbiont that interacts with the speciating host (Page, 2003, Dismukes et al., 2022). Host-switching is defined here as a symbiont speciation in which one descendant lineage inherits the host repertoire of the ancestral symbiont and the other descendant lineage obtains a new host lineage (Dismukes et al., 2022). Symbiont dispersal and extirpation describe the symbiont obtain new hosts or losing hosts respectively without any branching of the symbiont lineages (Dismukes et al., 2022). This model is implemented in the R package treeducken (Dismukes and Heath, 2021).

5.3.2 Dispersal-extinction-cladogenesis model

The DEC model is used to estimate the ancestral host range of a phylogeny. The DEC model contains two pieces: an anagenetic part that models change in biogeographic range along lineages and a cladogenetic part that models how biogeographic ranges are inherited at splitting events (Ree et al., 2005, Ree and Smith, 2008b). Here biogeographic range is coded as presence-absence data. Anagenetic range evolution is modeled as a continuous time Markov chain (CTMC) with two types of events dispersal to new areas and extirpation within areas.

Cladogenetic range evolution is allowed to have a number of different events that describe how biogeographic ranges may be inherited at speciation nodes. These events include sympatric speciation—biogeographic range inherited by both descendants, jump dispersal—biogeographic range is inherited by one descendant lineage, the other descendant lineage disperses to a new host, allopatric speciation—the biogeographic range is split without overlap by the descendant lineages. Landis (2017) built on this model by allowing the biogeographic range to vary within different geologic epochs. We use a modified version of the DEC model used in Landis (2017).

We use this model as an analogy for our cophylogenetic birth-death model. Specifically, we treat the symbiont phylogeny as the species tree and the host phylogeny as a history of the areas—*i.e.,* an area cladogram. Each host speciation node defines an epoch where a new area (host) emerges that can be dispersed to. The anagenetic events of dispersal and extirpation correspond to symbiont dispersal and extirpation. The cladogenetic events also correspond in this analogy. Jump dispersal corresponds to host-switching, sympatric speciation corresponds to symbiont speciation, and allopatric speciation corresponds to cospeciation.

5.3.3 Analysis

We estimated posterior densities using Markov chain Monte Carlo (MCMC) in the software RevBayes (Höhna et al., 2014). Analysis scripts are available on GitHub https://github.com/wadedismukes/cbd_ana_simstudy. Analyses were performed on the Iowa State University high-performance computing cluster Pronto.

5.3.3.1 Simulation

We investigated the efficacy of using the DEC model for estimating the parameters of the cophylogenetic birth-death model. To do this, we performed a simulation experiment by varying a number of parameters that we thought would be detectable under the DEC model. Specifically, we varied the symbiont dispersal rate (0, 2, 4) in two situations: with cospeciation and host-switching and with only cospeciation giving us a total of six simulation regimes. The remaining parameters of our cophylogenetic birth-death model were equal among all of our simulation regimes. The simulation time was set to 1.0, and no extinction was allowed to occur. We simulated 1000 replicate cophylogenetic datasets for each parameter setting.

We also set the maximum number of hosts that a symbiont is able to be associated with to two hosts. The primary reason for this is that the DEC scales extremely poorly with the number of states. The probabilities of dispersal to new hosts and extirpation within hosts along symbiont lineages must account for all combinations of starting states and ending states. For example, for our dataset with two areas, there are four states which means $4 \times 4 = 16$ probability terms. This becomes even worse with the cladogenetic probabilities as we must account for before the host speciation event, after host speciation on the left descendant, and after host speciation on the right descendant. Extending our previous example gives us $4 \times 4 \times 4 = 64$ probability terms.

5.4 Results and Discussion

Our simulation experiment evaluated six parameter settings. We varied rates of symbiont dispersal from 0, 2, and 4 in simulations with cospeciation and with cospeciation and host-switching. We examined the accuracy of our estimate of dispersal rate as well as the probability of sympatric speciation, analogous to symbiont speciation in our model. All analyses were run until Gelman-Rubin convergence values for our parameters was <1.1 (Gelman and Rubin, 1992).

The results for these analyses are shown in Figure 5.2 and Figure 5.3. The dispersal rate was correctly estimated as 0 in the case where there was no dispersal rate. However, the other positive values of dispersal rate showed little difference from one another in our experiment. There also seemed to be a correlation between host-switching and dispersal rate parameter. This could be the result of the model not being able to identify the

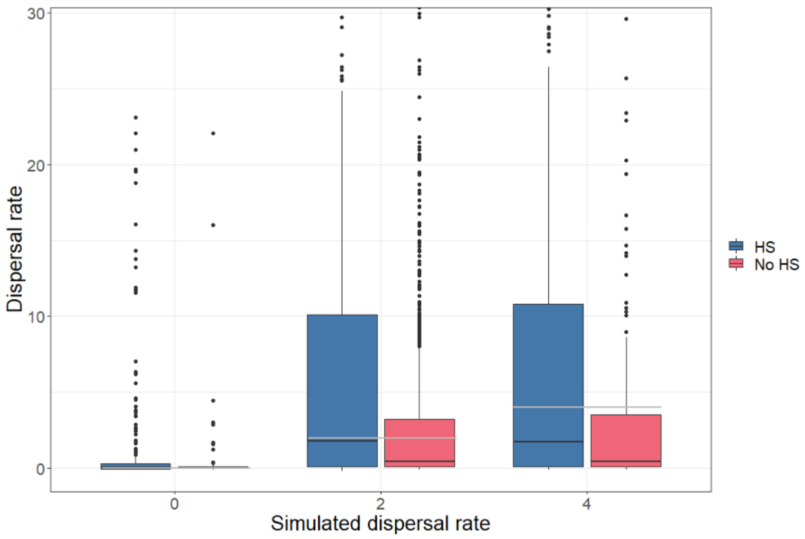


Figure 5.2 Boxplots showing the simulation results. Blue shows the host-switching (HS) and cospeciation simulations and red shows the only cospeciation simulations (no HS). The grey lines show the expectation for the dispersal rates. Each point is the maximum a posteriori (MAP) estimate for dispersal rate for one simulation.

difference between jump dispersal that occurs at host speciation nodes, and symbiont dispersal that occurs along lineages.

The results for the probability of sympatric speciation were essentially the prior we provided for our cladogenetic events parameter. This parameter was a vector of probabilities for two events: sympatric speciation and jump dispersal. In our cophylogenetic analogy, these correspond to symbiont speciation and host-switching respectively. We used a $Dirichlet(10, 10)$ prior on this parameter. It is possible that the signal in our simulated data did not provide ample evidence to update this prior, and that a $Dirichlet(1, 1)$ may have been more appropriate.

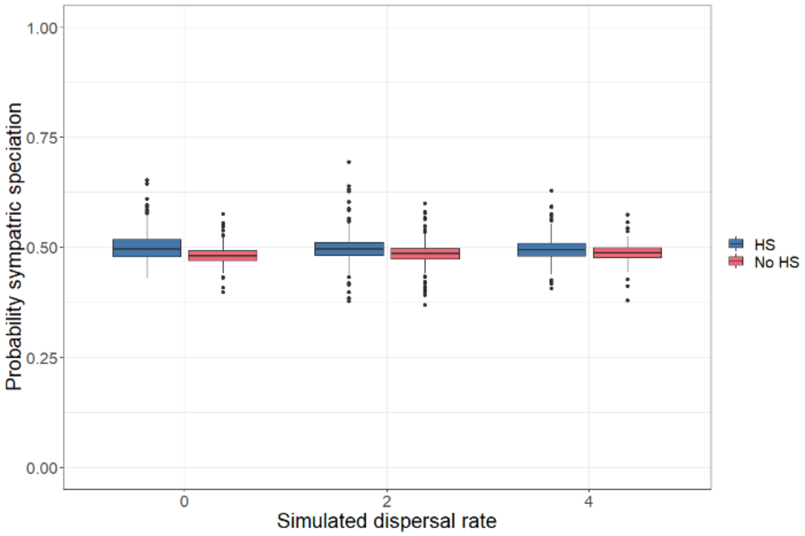


Figure 5.3 Box plots showing the probability of sympatric speciation. Blue shows the host-switching (HS) and cospeciation simulations and red shows the only cospeciation simulations (no HS). Each point is the maximum a posteriori (MAP) estimate for the probability of sympatric speciation for one simulation.

5.5 Conclusion

We performed a simulation experiment to assess the utility of a historical biogeography model for estimating the parameters of the cophylogenetic birth-death process. The results of this experiment were not promising. The analyses took between seven days and 34 days to run, and did not seem to work as intended. Biogeography models as a whole have shown promise for cophylogenetics (Braga et al., 2020). Here the issue could have been that the data did not provide adequate information for the complex model used. In addition, the DEC model is hampered by the poor scalability to many hosts which makes the model difficult to use for many cophylogenetic datasets.

5.6 References

Balbuena, J. A., Pérez-Escobar, Ó. A., Llopis-Belenguer, C., and Llopis-Belenguer, I. (2020). Random tanglegram partitions (Random TaPas): An Alexandrian approach to the cophylogenetic Gordian knot. *Systematic Biology*. syaa033.

Braga, M. P., Landis, M. J., Nylin, S., Janz, N., and Ronquist, F. (2020). Bayesian inference of ancestral host-parasite interactions under a phylogenetic model of host repertoire evolution. *Systematic Biology*, 69:1149–1162.

Brooks, D. R. (1981). Hennig's parasitological method: A proposed solution. *Systematic Biology*, 30(3):229–249.

Dismukes, W., Braga, M. P., Hembry, D. H., Heath, T. A., and Landis, M. J. (2022). Cophylogenetic methods to untangle the evolutionary history of ecological interactions. *Annual Review of Ecology, Evolution, and Systematics*, in press.

Dismukes, W. and Heath, T. A. (2021). treeducken: An R package for simulating cophylogenetic systems. *Methods in Ecology and Evolution*, 12(8):1358–1364.

Dobson, A., Lafferty, K. D., Kuris, A. M., Hechinger, R. F., and Jetz, W. (2008). Homage to Linnaeus: how many parasites? how many hosts? *Proceedings of the National Academy of Sciences*, 105(Supplement 1):11482–11489.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

Heled, J. and Drummond, A. J. (2009). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3):570–580.

Höhna, S., Heath, T. A., Boussau, B., Landis, M. J., Ronquist, F., and Huelsenbeck, J. P. (2014). Probabilistic graphical model representation in phylogenetics. *Systematic Biology*, 63(5):753–771.

Landis, M. J. (2017). Biogeographic dating of speciation times using paleogeographically informed processes. *Systematic Biology*, 66(2):128–144.

Landis, M. J., Matzke, N. J., Moore, B. R., and Huelsenbeck, J. P. (2013). Bayesian analysis of biogeography when the number of areas is large. *Systematic Biology*, 62(6):789–804.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46:523–536.

Matzke, N. J. (2014). Model Selection in Historical Biogeography Reveals that Founder-Event Speciation Is a Crucial Process in Island Clades. *Systematic Biology*, 63(6):951–970.

National Research Council (2007). *Status of pollinators in North America.* The National Academies Press, Washington, DC.

Page, R. D. M. (2003). *Tangled Trees: Phylogeny, Cospeciation, and Coevolution.* University of Chicago Press.

Page, R. D. M. and Charleston, M. A. (1998). Trees within trees: phylogeny and historical associations. *Trends in Ecology & Evolution*, 13:356–359.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, 164(4):1645–1656.

Rasmussen, M. D. and Kellis, M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research*, 22(4):755–765.

Ree, R. H., Moore, B. R., Webb, C. O., and Donoghue, M. J. (2005). A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. *Evolution*, 59(11):2299–2311.

Ree, R. H. and Smith, S. A. (2008a). Maximum Likelihood Inference of Geographic Range Evolution by Dispersal, Local Extinction, and Cladogenesis. *Systematic Biology*, 57(1):4–14.

Ree, R. H. and Smith, S. A. (2008b). Maximum Likelihood Inference of Geographic Range Evolution by Dispersal, Local Extinction, and Cladogenesis. *Systematic Biology*, 57(1):4–14.

Santichaivekin, S., Yang, Q., Liu, J., Mawhorter, R., Jiang, J., Wesley, T., Wu, Y.-C., and Libeskind-Hadas, R. (2021). empress: a systematic cophylogeny reconciliation tool. *Bioinformatics*, 37(16):2481–2482.

Satler, J. D., Herre, E. A., Jandér, K. C., Eaton, D. A., Machado, C. A., Heath, T. A., and Nason, J. D. (2019). Inferring processes of coevolutionary diversification in a community of Panamanian strangler figs and associated pollinating wasps. *Evolution*, 73(11):2295–2311.

Smith, C. I., Godsoe, W. K., Tank, S., Yoder, J. B., and Pellmyr, O. (2008). Distinguishing coevolution from covicariance in an obligate pollination mutualism: asynchronous divergence in Joshua tree and its pollinators. *Evolution*, 62(10):2676–2687.

Szöllosi, G. J., Tannier, E., Lartillot, N., and Daubin, V. (2013). Lateral gene transfer from the dead. *Systematic Biology*, 62(3):386–397.

Wang, B. and Qiu, Y.-L. (2006). Phylogenetic distribution and evolution of mycorrhizas in land plants. *Mycorrhiza*, 16(5):299–363.

Weber, M. G., Wagner, C. E., Best, R. J., Harmon, L. J., and Matthews, B. (2017). Evolution in a community context: on integrating ecological interactions and macroevolution. *Trends in Ecology & Evolution*, 32(4):291–304.

CHAPTER 6. GENERAL CONCLUSION

The vast diversity of symbioses in the natural world has captivated biologists since Darwin (Darwin, 1862) and Wallace (Wallace, 1867). Cophylogenetics emerged as a means to analyze host and symbiont phylogenies. My dissertation focused on new ways of exploring the macroevolution of hosts and their symbionts. My first goal was to survey the existing methodologies and how they are used by empiricists. Chapter 2 (Dismukes et al., 2022) offers a detailed review of cophylogenetic methods and provided guidance on how to group different types of host-symbiont systems based on their biological properties and resulting phylogenetic patterns. Next, I developed a generative model for cophylogenetic data that fills a gap in the current toolkit for cophylogenetic simulation (Keller-Schmidt et al., 2011, Baudet et al., 2015, Alcala et al., 2017). Indeed, none of the existing generative models allows speciation and extinction on both the host and symbiont trees. Chapter 3 (Dismukes and Heath, 2021) introduced the cophylogenetic birth-death model which does do this, and is implemented as an R package, treeducken. We then investigated ways of estimating the parameters of this generative model. Currently, there is no traditional inference approach for doing this so we proceeded in two different ways. One of these approaches was a deep learning approach that accurately estimated the cophylogenetic events. The other used an analogy between historical biogeography and cophylogenetics, using the dispersal-extinction-cladogenesis model to estimate parameters of the cophylogenetic birth-death model.

Chapter 3 introduced the first generative model for cophylogenetic data that allows for simulation of both the host and symbiont trees simultaneously with extinction for both trees. This model builds on past simulation models (Keller-Schmidt et al., 2011) to develop this approach. This generative model is available as an R package (Dismukes and Heath, 2021) making it straightforward to install and use. This simulator will assist theoreticians in assessing the performance of their models, and empiricists in exploring cophylogenetic datasets. More work could be done to make this simulator even more useful to cophylogenetic researchers. Specifically, I could introduce a means to simulate a symbiont tree on a host tree.

Chapters 4 and 5 explored new approaches for examining cophylogenetic data. Chapter 4 used this model to generate data that was used to train a deep neural network. This model is one of the first uses of machine learning to explore cophylogenetic data. While this deep learning approach is relatively simplistic, the method proved to be accurate at estimating the number of cophylogenetic events. In addition, I will use more summary statistics and investigate the importance of the different summary statistics more thoroughly. I also intend to make this deep learning approach more approachable for empiricists by creating an R package in which these analyses can be easily conducted.

In nearly all of these cases, ecological interactions are treated as binary characters; in reality, these interactions are mediated by phenotypic traits and thus are likely to contain more information about how these systems evolved than binary characters alone (Janzen, 1980). Recent work has discussed the importance of developing cophylogenetic methods that use traits as a key area of research in cophylogenetics (Blasco-Costa et al., 2021) At present, only a few cophylogenetic methods (Hutchinson et al., 2017, Adams and Nason, 2018) incorporate trait data, and these focus primarily on the traits themselves rather than parameters such as the rate of host-switching and cospeciation that may be of interest to those examining cophylogenetic systems. This, finding ways to incorporate trait data into my deep learning approach is a logical next step for these methods.

In nearly all of these cases, ecological interactions are treated as binary characters; in reality, these interactions are mediated by phenotypic traits and thus are likely to contain

more information about how these systems evolved than binary characters alone. Recent work has discussed the importance of developing cophylogenetic methods that use traits as a key area of research in cophylogenetics (Blasco-Costa et al. 2021). At present, only a few cophylogenetic methods (Adams and Nason 2018, Hutchinson 2017) incorporate trait data, and these focus primarily on the traits themselves rather than parameters such as the rate of host-switching and cospeciation that may be of interest to those examining cophylogenetic systems. In addition, none of these include a generative model for the diversification of a host and symbiont along with the respective coevolving traits of both host and symbiont. Adapting the cophylogenetic birth-death process to incorporate trait data would be a promising way of extending this model.

Chapter 5 used a historical biogeography model to estimate parameters of the cophylogenetic birth-death model developed in chapter three. The results of this study were not promising, with the posterior hardly being updated from the priors set. In addition, the scalability of this model is poor making it possibly a poor choice for large cophylogenetic datasets. However, this does not rule out historical biogeography models usefulness for investigating cophylogenetic data as Braga et al. (2020) showed with their method for estimating the ancestral host repertoire.

Overall we successfully developed new approaches for analyzing cophylogenetic data. The introduction of more model-based approaches, such as those used here, will spur development in other areas of cophylogenetics including event-scoring and pattern-based statistic methods. A more complete exploration of the parameter space of host-symbiont datasets is key to understanding the macroevolution of these complex systems. Ultimately, developing a likelihood function that provides actual estimates—as opposed to the surrogates used here—of the cophylogenetic birth-death process would be a particularly fruitful area of future research.

6.1 References

Adams, D. C. and Nason, J. D. (2018). A phylogenetic comparative method for evaluating trait coevolution across two phylogenies for sets of interacting species. *Evolution*, 72(2):234–243.

Alcala, N., Jenkins, T., Christe, P., and Vuilleumier, S. (2017). Host shift and cospeciation rate estimation from co-phylogenies. *Ecology Letters*, 20(8):1014–1024.

Baudet, C., Donati, B., Sinaimeri, B., Crescenzi, P., Gautier, C., Matias, C., and Sagot, M.-F. (2015). Cophylogeny reconstruction via an approximate Bayesian computation. *Systematic Biology*, 64(3):416–431.

Blasco-Costa, I., Hayward, A., Poulin, R., and Balbuena, J. A. (2021). Next-generation cophylogeny: unravelling eco-evolutionary processes. *Trends in Ecology & Evolution*, 36(10):907–918.

Braga, M. P., Landis, M. J., Nylin, S., Janz, N., and Ronquist, F. (2020). Bayesian inference of ancestral host-parasite interactions under a phylogenetic model of host repertoire evolution. *Systematic Biology*, 69:1149–1162.

Darwin, C. (1862). Letter to J.D. Hooker. *More letters of Charles Darwin Volume*, 2.

Dismukes, W., Braga, M. P., Hembry, D. H., Heath, T. A., and Landis, M. J. (2022). Cophylogenetic methods to untangle the evolutionary history of ecological interactions. *Annual Review of Ecology, Evolution, and Systematics*, in press.

Dismukes, W. and Heath, T. A. (2021). treeducken: An R package for simulating cophylogenetic systems. *Methods in Ecology and Evolution*, 12(8):1358–1364.

Hutchinson, M. C., Cagua, E. F., and Stouffer, D. B. (2017). Cophylogenetic signal is detectable in pollination interactions across ecological scales. *Ecology*, 98(10):2640–2652.

Janzen, D. H. (1980). When is it coevolution? *Evolution*, 34(3):611–612.

Keller-Schmidt, S., Wieseke, N., Klemm, K., and Middendorf, M. (2011). Evaluation of host parasite reconciliation methods using a new approach for cophylogeny generation. *University of Leipzig, Technical Report*, pages 11–013.

Wallace, A. R. (1867). Creation by law. *QJ Sci*, 4(16):470–488.