Diversification Models and Neural Inference

Tianjian Qin

university of groningen

Diversification Models and Neural Inference

PhD thesis

to obtain the degree of PhD of the
University of Groningen
on the authority of the
Rector Magnificus Prof. J.M.A. Scherpen
and in accordance with
the decision by the College of Deans.
This thesis will be defended in public on
Monday 4 May 2026 at 14.00 hours

by

Tianjian QIN

Master of Science in Wetland Ecology,
born on 5 January 1994.

This thesis has been approved by:

Supervisor: Prof. dr. R.S. Etienne
Co-supervisor: Dr. L.L. Valente
Co-supervisor: Dr. K.J. van Benthem

Assessment committee:

Rector Magnificus,	chair
Prof. dr. R.S. Etienne,	University of Groningen
Dr. L.L. Valente,	Naturalis Biodiversity Center
Dr. K.J. van Benthem,	University of Groningen
Independent members:	
Prof. dr. E.C. Wit,	Università della Svizzera italiana
Prof. dr. F. Hartig,	Universität Regensburg
Prof. dr. M. Beaumont,	University of Bristol

This thesis was supported by the University of Groningen–China Scholarship Council joint scholarship program and carried out under the auspices of Groningen Institute for Evolutionary Life Sciences.

Keywords:	macroevolution; diversity dependence; deep learning
Printed by:	HY Printing
Cover:	Tianjian Qin
Style:	Tianjian Qin, modified from Moritz Beller

The author set this thesis in LaTeX using the Libertinus, Inconsolata, and Noto CJK fonts.

An electronic version of this thesis is available at
https://github.com/EvoLandEco/Thesis/.

rijksuniversiteit
groningen

Diversificatiemodellen en neurale inferentie

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus prof. dr. ir. J.M.A. Scherpen
en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op
maandag 4 mei 2026 om 14.00 uur

door

Tianjian QIN

Master of Science in Wetland Ecology,
geboren op 5 Januari 1994.

Dit proefschrift is goedgekeurd door de

promotor: Prof. dr. R.S. Etienne
copromotor: Dr. L.L. Valente
copromotor: Dr. K.J. van Benthem

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. R.S. Etienne,	Rijksuniversiteit Groningen
Dr. L.L. Valente,	Naturalis Biodiversity Center
Dr. K.J. van Benthem,	Rijksuniversiteit Groningen
Onafhankelijke leden:	
Prof. dr. E.C. Wit,	Università della Svizzera italiana
Prof. dr. F. Hartig,	Universität Regensburg
Prof. dr. M. Beaumont,	University of Bristol

Dit proefschrift werd ondersteund door het gezamenlijke beursprogramma van de Rijksuniversiteit Groningen en de China Scholarship Council en uitgevoerd onder auspiciën van het Groningen Institute for Evolutionary Life Sciences.

Trefwoorden:	macroevolutie; diversiteitsafhankelijkheid; deep learning
Druk:	HY Printing
Omslag:	Tianjian Qin
Vormgeving:	Tianjian Qin, aangepast van Moritz Beller

De auteur zette dit proefschrift in LaTeX met behulp van de Libertinus-, Inconsolata- en Noto CJK-lettertypen.

Een elektronische versie van dit proefschrift is beschikbaar op
https://github.com/EvoLandEco/Thesis/.

This dedication is still under construction, much like my life choices.

Contents

Summary

This thesis investigates the drivers of species diversification, and to what extent deep learning methods can recover different drivers from extant phylogenies. In particular, I focus on the effects of ecological limits (constraints to the numbers of species that can coexist) and evolutionary relatedness on speciation and extinction rates. First, I introduce a birth–death model, eve, in which speciation rates depend on both species richness and evolutionary relatedness (ER) measured at different phylogenetic scales. I show that tree shape and the distribution of speciation rates across lineages depend strongly on the scale at which ER acts, and that negative species richness dependence can partly mask the influence of ER on standard tree statistics. The model generates a wide range of empirically realistic, often imbalanced, phylogenies.

Second, I develop an ensemble neural-network framework for parameter estimation in diversification models. Combining dense, graph and recurrent neural networks trained on tree topologies, branching times and summary statistics, the method yields estimates faster than maximum-likelihood approaches and is less sensitive to tree size for constant-rate and diversity-dependent models. However, both likelihood-based and neural network estimators struggle under protracted speciation, highlighting limits imposed by the information content of trees.

Third, I use the eve model—which couples ecological limits with ER effects on speciation and extinction—as a testbed to map when neural networks can and cannot infer diversification mechanisms from phylogenies. In many cases the neural networks struggle to tell the three scenarios apart, and when the trees carry little information the estimated parameters tend to drift back toward average values. Strong global richness dependence further erodes recoverability, whereas sufficiently strong ER effects can create narrow regions of practical identifiability. Together, these results delineate the prospects and limits of using flexible diversification models and deep learning to unravel evolutionary dynamics from extant phylogenies.

Samenvatting

In dit proefschrift onderzoek ik welke krachten soortendiversificatie sturen en in hoeverre deep-learningmethoden verschillende drijvers kunnen terugvinden uit fylogenieën van nog levende soorten. Daarbij richt ik mij vooral op het effect van ecologische grenzen (beperkingen aan het aantal soorten dat kan samenleven) en evolutionaire verwantschap (ER) op soortvormings- en uitstervingssnelheden. Allereerst introduceer ik een geboorte–sterftemodel, `eve`, waarin soortvormingssnelheden afhangen van zowel soortrijkdom als ER, gemeten op verschillende fylogenetische schalen. Ik laat zien dat de boomvorm en de verdeling van soortvormingssnelheden over lijnen sterk afhangen van de schaal waarop ER werkt, en dat negatieve afhankelijkheid van soortrijkdom de invloed van ER op standaard boomstatistieken deels kan maskeren. Het model genereert een breed scala aan empirisch realistische, vaak ongebalanceerde fylogenieën.

Vervolgens ontwikkel ik een ensemble-benadering met neurale netwerken voor parameterschatting in diversificatiemodellen. Door dichte, grafische en recurrente neurale netwerken te combineren—getraind op boomtopologieën, vertakkingstijden en samenvattende statistieken—verkrijgt deze methode schattingen sneller dan maximum-likelihoodmethoden en is zij minder gevoelig voor boomgrootte bij constant-rate en diversiteitsafhankelijke modellen. Zowel likelihood-gebaseerde als neurale schatters hebben echter moeite met geprotraheerde soortvorming, wat de grenzen laat zien die worden opgelegd door de informatie-inhoud van fylogenieën.

Ten slotte gebruik ik het `eve`-model—dat ecologische grenzen koppelt aan ER-effecten op soortvorming en uitsterven—als testomgeving om in kaart te brengen wanneer neurale netwerken diversificatiemechanismen wel of niet kunnen achterhalen. In veel gevallen hebben de neurale netwerken moeite om de drie scenario's uit elkaar te houden, en wanneer de bomen weinig informatie bevatten schuiven de geschatte parameters terug richting gemiddelde waarden. Sterke globale afhankelijkheid van soortrijkdom vermindert de herleidbaarheid verder, terwijl voldoende sterke ER-effecten smalle gebieden van praktische identificeerbaarheid kunnen creëren. Gezamenlijk schetsen deze resultaten de mogelijkheden en beperkingen van het gebruik van flexibele diversificatiemodellen en deep learning om evolutionaire dynamiek uit fylogenieën van nog levende soorten te ontrafelen.

摘要

本论文研究物种分化（由物种的形成与灭绝共同决定的过程）的驱动因素，并探讨深度学习方法是否能够仅凭现存物种的系统发育树识别这些驱动机制。论文重点关注两类影响：生态容量/生态限制（即能够共存的物种数量上限），以及物种之间的进化亲缘关系（evolutionary relatedness, ER）对物种形成与灭绝速率的效应。

首先，本文提出一个新的出生 – 死亡模型 eve。在该模型中，物种分化过程同时受到物种丰富度与 ER 的影响，并允许 ER 在不同系统发育尺度上起作用。研究表明，ER 的作用尺度会显著改变系统发育树的结构，并影响物种分化速率在谱系间的分布；同时，较强的负向丰富度依赖会在一定程度上掩盖 ER 在常用树统计量中留下的痕迹。该模型能够生成多样且与经验数据更贴近、往往高度不平衡的系统发育树。

其次，本文提出一个用于物种分化模型参数估计的集成学习框架。该框架联合使用全连接神经网络、图神经网络与循环神经网络，能够同时利用系统发育树的拓扑结构、分枝时间与汇总统计量。在应用于多种经典模型时，该方法相较最大似然估计更快，并且其估计精度对树规模的敏感性更低。然而，在“延迟物种形成”模型下，无论是最大似然方法还是神经网络方法都难以精确复原参数。这表明模型参数的可推断性有可能受限于系统发育树本身所携带的信息量。

最后，本文以 eve 模型为检验平台，系统性评估神经网络在何种条件下能够从系统发育树中准确推断物种分化机制。结果显示，在许多情况下，神经网络难以准确区分 eve 模型对应的三类机制情景；当系统发育树所包含的信息量较低时，参数估计往往会回到“平均值”附近。较强的全局丰富度依赖效应会进一步削弱参数与情景的可推断性，而足够强的 ER 效应则可能在参数空间中形成狭窄但可有效识别的区域。综上所述，本论文系统讨论了深度学习框架用于物种分化模型参数估计时的表现、限制与应用前景。

Acknowledgments

This thesis would not exist without the people who made these past years not only possible, but truly enjoyable.

First and foremost, I thank my supervisors—Rampal, Luis, and Koen. You gave me the freedom to explore, the structure to finish, and the calm reassurance to keep going when things became messy. I have learned a great deal from your clarity, standards, and generosity. Just as importantly, I am grateful that along the way you became not only my mentors, but also friends for life.

To my parents, thank you for your steady faith and encouragement, even from far away. To my wife, thank you for being my anchor—for the warmth, patience, and quiet strength that carried me through the long stretches. To my family and friends, thank you for the laughter, the check-ins, the distractions at the right moments, and for reminding me that there is a world beyond revision cycles.

I am also grateful to my colleagues at the University of Groningen and Wageningen University & Research for the welcoming atmosphere, thoughtful conversations, and the coffee breaks that somehow turned into ideas. Working with you made research feel far less like a solitary marathon.

Finally, a special thank-you to Mitou, our cat, and to all the other cats who contributed in their own way—by keeping me company, insisting on timely breaks, and maintaining the firm belief that keyboards are simply deluxe pillows.

秦天健

Groningen, the Netherlands
December 2025

1

Syntroduction

> *It is interesting to contemplate an entangled bank, clothed with many plants of many kinds, with birds singing on the bushes, with various insects flitting about, and with worms crawling through the damp earth, and to reflect that these elaborately constructed forms, so different from each other, and dependent on each other in so complex a manner, have all been produced by laws acting around us.*

— Charles Darwin, *On the Origin of Species* (1859)

Charles Darwin's seminal work, *On the Origin of Species*, published in 1859, introduced the scientific theory that populations evolve over generations through a process of natural selection. This theory was revolutionary at the time, providing a unifying explanation for the diversity of life on Earth.

Darwin proposed that individuals within a species exhibit variations in their traits, and those with characteristics better suited to their environment have higher chances of surviving and reproducing. Over successive generations, these advantageous traits become more prevalent, leading to evolutionary changes. This concept of natural selection challenged the once prevailing notion of fixed, unchanging species and suggested a common ancestry among diverse forms of life.

Darwin's work laid the foundation for modern evolutionary biology, influencing a wide range of scientific disciplines. By explaining how species adapt and change over time, Darwin's theory has become central to our understanding of life's *complexity* and *interconnectedness*. Today, these ideas are made quantitatively explicit in macroevolutionary models and phylogenetic analyses, which use branching patterns and branch lengths to study how speciation and extinction have shaped the tree of life–an endeavor at the core of this thesis.

This chapter draws in part on the introductions to the following chapters.

1

1.1 From Phylogenetic Trees to Networks

Tree-like diagrams for organizing knowledge predate evolutionary theory. Early examples such as Augustin Augier's 1801 "Arbre botanique" depicted nature's perfect, God-given order without temporal or evolutionary interpretation [1, 2]. In 1809, Jean-Baptiste Lamarck introduced a branching diagram in his *Philosophie zoologique* that portrayed parallel lineages but did not invoke common descent [3]. Edward Hitchcock later published one of the first tree-like paleontological charts in 1840 [4], and Robert Chambers used a branching diagram in his 1844 *Vestiges of the Natural History of Creation* to tentatively apply a tree metaphor to the history of life [5]. By 1858, Heinrich Georg Bronn had proposed a hypothetical "tree of life," yet still without a clear mechanism for evolutionary change [6].

Charles Darwin formalized the evolutionary meaning of trees in 1859 by presenting an abstract diagram with a generational scale in *On the Origin of Species* (Figure 1.1). This tree depicts hypothetical species evolving over a sequence of time intervals, with branching lines representing divergence, extinction, and the origin of new species. Darwin used this figure to illustrate how small, heritable differences can accumulate gradually and transform varieties into distinct species. The term "phylogeny" itself, denoting the evolutionary relationships of species through time, was coined by Ernst Haeckel, who expanded Darwin's concept into richly annotated trees of life [7]. Haeckel's 1866 phylogeny depicted three kingdoms—Plantae, Protista, and Animalia—and is often cited as one of the earliest comprehensive "trees of life" spanning multiple major lineages [8, 9]. His later "Pedigree of Man" traced all life back to Monera and placed humans at the top, reflecting a strongly hierarchical view of evolution.

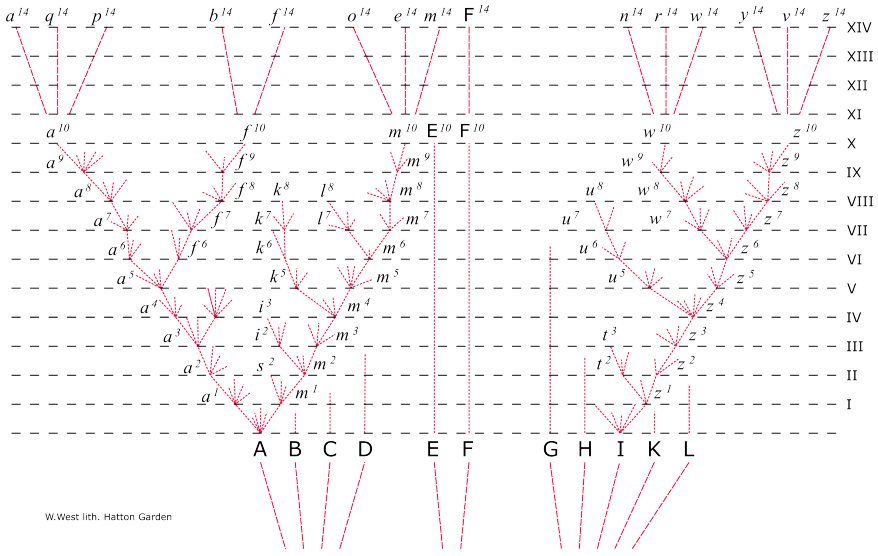


Figure 1.1: Darwin's abstract tree diagram illustrating evolutionary divergence over time. Modified from *Inductiveload, Wikimedia Commons.*

1.1.1 What Is a Phylogenetic Tree?

A phylogenetic tree, or phylogeny, is a branching diagram that represents hypothesized evolutionary relationships among biological entities (e.g. species, populations, or genes) based on similarities and differences in their traits or genetic compositions. A typical phylogeny contains the following elements:

1) **Root**: The most recent common ancestor of all entities represented in the tree.
2) **Branches**: Line segments representing evolutionary lineages.
3) **Nodes**: Points where branches join or split, typically interpreted as speciation or divergence events.
4) **Tips**: Terminal nodes representing the sampled taxa (extant or extinct) included in the analysis.

Trees can be *rooted*, indicating a direction of time from the root to the tips, or *unrooted*, which depict relationships among taxa without specifying an ancestral root. Mathematically, a phylogeny is a special case of a graph. Its topology can be rearranged visually—for example, by rotating subtrees around internal nodes or changing layout (rectangular, radial, unrooted)—without altering the underlying relationships, as long as connections among nodes are preserved (see Figure 1.2). For purely topological trees (those without meaningful branch lengths), even stretching or compressing branches does not affect the encoded relationships.

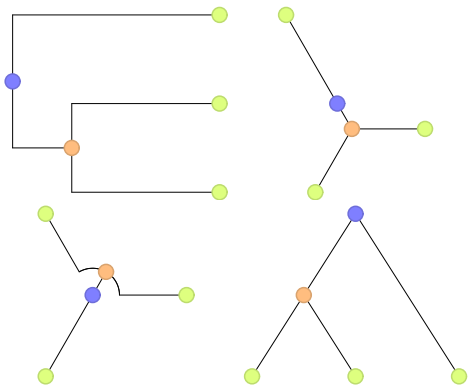


Figure 1.2: Different layouts of the same phylogenetic tree. The blue circle marks the root, orange circles indicate internal nodes, and green circles indicate tips. Rearranging branches without changing their connections preserves the underlying topology.

While phylogenies may include *polytomies* (nodes with more than two descendants), they are often depicted as fully bifurcating trees. In many cases, multifurcations are treated as temporary placeholders for uncertainty that could be resolved into a series of bifurcations with additional data. Importantly, standard phylogenies are spatially implicit: they do not encode explicit geographic locations or distances.

1

Historically, researchers assembled phylogenies manually from morphological and anatomical comparisons. Today, phylogenies are inferred from biological data using computational methods, and they serve as the backbone for a wide variety of downstream analyses. In many molecular phylogenetic applications, trees reconstructed from extant data are treated as *ultrametric*: all tips terminate at the same distance from the root and correspond to lineages observed at the present. This assumption underlies many phylogenetic algorithms and statistics.

1.1.2 From Morphology and Similarity to Molecular Phylogenies

Early phylogenetic inference relied heavily on morphology and anatomy. Researchers examined structural features of organisms to identify *homologous* traits—characters inherited from a common ancestor [9]. For example, the presence of four limbs in tetrapods is a classic homologous character. Shared homologous traits can be used to construct trees illustrating evolutionary relationships among groups (e.g. tetrapods in Figure 1.3). A major challenge in such analyses is distinguishing homologous traits from *analogous* traits that evolved independently (e.g. wings in birds and insects), a distinction critical for accurate phylogenetic reconstruction [10].

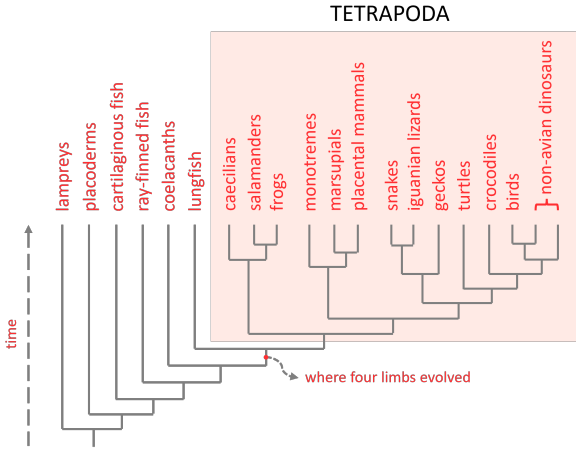


Figure 1.3: Example tree based on homologous characters (the presence of four limbs) in tetrapods. Recreated from Irisarri et al. [10].

Phenetics, or numerical taxonomy, later emerged as a method for classifying organisms by overall similarity, often using quantitative measurements across many traits [11]. The goal was to achieve objectivity by applying statistical methods to multivariate datasets, while avoiding explicit assumptions about evolutionary history. However, phenetic approaches were criticized for sometimes grouping taxa based on superficial similarity arising from convergent evolution, rather than true common ancestry [12]. Some taxonomists hoped that analyzing a sufficiently large number of characters would better cluster taxa descended from the nearest common ancestor, but this proved impractical: assembling and scoring many reliable characters was labor-intensive, and there is little guidance on which clustering or distance measures were appropriate in different situations [13, 14].

The advent of large-scale DNA sequencing transformed phylogenetics. Molecular data allow researchers to compare DNA or protein sequences across a wide variety of organisms and to infer trees that depict their evolutionary relationships (see Figure 1.4 for an example). These molecular phylogenies often corroborate traditional, morphology-based classifications but, in some cases, have prompted major revisions where genetic evidence reveals previously unrecognized relationships [15, 16]. Fossil records play a complementary role: they provide temporal context, document extinct lineages, and offer morphological information not accessible from living taxa alone. Integrating fossil data with molecular trees enables calibration of molecular clocks and estimation of divergence times [17–19]. Despite these advances, reconciling molecular and fossil evidence remains challenging due to incomplete fossil records [20], limitations on DNA preservation in ancient specimens [21], and complications from processes such as horizontal gene transfer [22, 23] and convergent evolution [24, 25].

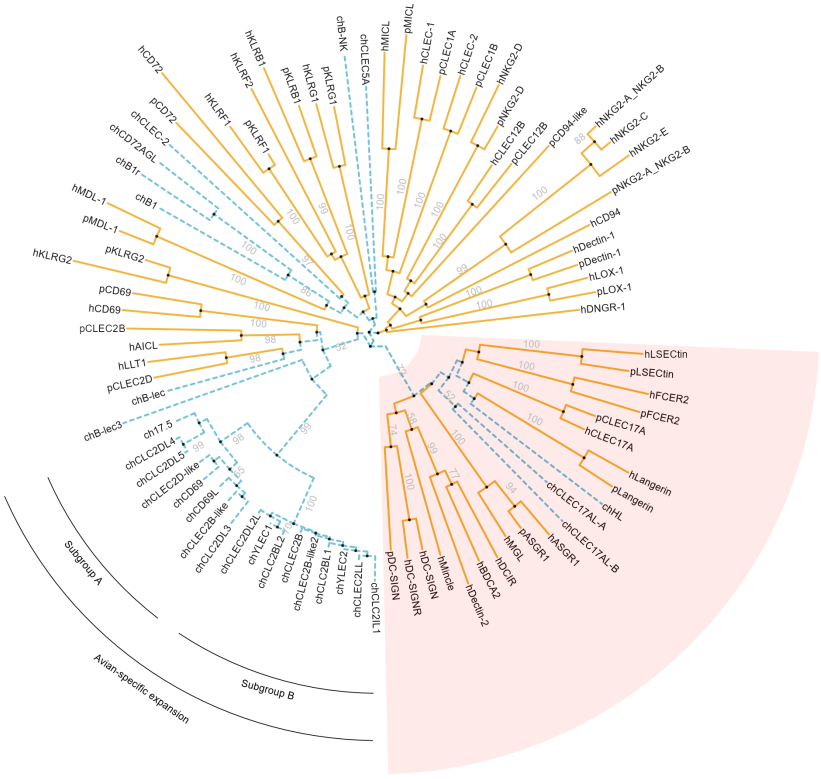


Figure 1.4: Example of a modern phylogeny based on DNA sequence data. Recreated from a `ggtree` illustration.

1

1.1.3 Alternative Tree Forms

The classical phylogenetic tree is often presented as a strictly bifurcating hierarchy where all sampled taxa are tips, ancestors are unsampled, and evolution proceeds exclusively via splitting. In practice, several extensions of this basic model are widely used to capture more nuanced evolutionary scenarios. These include, for example, sampled ancestors [18, 26–28], anagenesis [29, 30], and polytomies [31–34].

Phylogenies can explicitly incorporate fossil taxa, often scored for morphological characters, alongside extant species. Rather than using fossils only as node calibrations, *tip-dating* and *total-evidence* approaches treat fossil species as tips with known ages and integrate them into the branching process [17, 19]. In these frameworks, both morphological and molecular characters contribute to the inference of topology and divergence times, and fossils may act as sampled ancestors. Including fossil taxa can break up long branches, reveal stem lineages leading to modern clades, and substantially alter interpretations of trait evolution and biogeography [20, 26]. Trees that integrate fossils in this way provide a more explicit, time-aware view of macroevolutionary history.

1.1.4 Branch Lengths, Molecular Clocks, and Tree Types

Before interpreting phylogenetic tree properties or attempting to extract quantitative information from a tree, it is essential to understand what branch lengths represent in a given case. Three broad categories are commonly distinguished (see Figure 1.5):

1) **Cladogram**: branch lengths carry no quantitative meaning; only the branching order matters.

2) **Phylogram**: branch lengths are proportional to the amount of genetic change (e.g. expected substitutions per site).

3) **Chronogram**: branch lengths are proportional to time, so that the path length from root to a tip reflects elapsed evolutionary time.

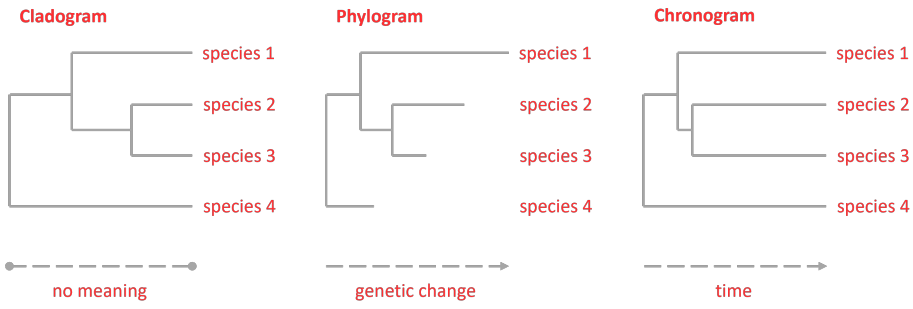


Figure 1.5: Three types of trees with different interpretations of branch length, adapted from Kellis *et al.* [35]. **Left:** a cladogram with no meaningful branch lengths. **Middle:** a phylogram in which branch lengths represent genetic change since divergence. **Right:** a chronogram (ultrametric tree) in which branch lengths represent time since divergence. All three trees share identical topology.

The classical molecular clock hypothesis assumes that genetic change accumulates at a

roughly constant rate, so that branch lengths in a phylogram are proportional to divergence times when appropriately scaled [36]. Empirical studies, however, show that rates vary widely across lineages. Factors such as metabolic rate [37], generation time [38], and environmental conditions [39, 40] all influence substitution rates. Applying a single, global clock across distantly related taxa can therefore lead to inaccurate divergence time estimates.

To accommodate rate variation, a range of *local*, *discrete*, and *relaxed* clock models have been developed. These allow rates to differ among branches or among clades while still linking molecular change to time [41–43]. More flexible "mixed" branch-length models also allow for site-specific rate variation across the genome, improving both phylogeny inference and divergence-time estimation [44]. When comparing phylogenetic metrics across trees, it is crucial to account for these differences: the same statistic may have different interpretations on a cladogram, a phylogram, or a chronogram.

1.1.5 What We Can Learn from Phylogenies

Many of the hardest questions in evolution are hard because the past is unobserved: we do not directly see historical speciation, extinction, or the tempo of trait change. For many clades (especially those lacking a rich fossil record), the most information-rich record we often have is a phylogeny of extant taxa, which encodes both branching order and the timing of divergences. For example, when a novel pathogen spreads, time-calibrated phylogenies built from genome sequences can be used to reconstruct introductions, identify transmission clusters, and estimate how rapidly an epidemic is growing [45–47]. In conservation, tree-based measures such as phylogenetic diversity and evolutionary distinctiveness help prioritize areas or lineages that represent disproportionate amounts of unique evolutionary history [48, 49]. In community ecology and invasion biology, the phylogenetic structure of assemblages (clustering vs. overdispersion) is often used as a proxy for shared traits and niche similarity, informing hypotheses about environmental filtering, competition, and biotic resistance [50–52].

These examples share a common logic: branching patterns summarize shared ancestry, while branch lengths carry information about the timing (or amount) of change. Phylogenies are therefore not merely static pictures of history; they are quantitative objects from which we can extract information about trait evolution, diversification dynamics, and community structure. Two broad classes of tools are especially important: phylogenetic comparative methods and tree summary statistics.

Phylogenetic Comparative Methods

Phylogenetic comparative methods use the shared evolutionary history encoded in a tree to test hypotheses about trait evolution while accounting for the non-independence of species. A foundational contribution is Felsenstein's method of *phylogenetically independent contrasts*, which transforms trait data on a tree into a set of contrasts that are statistically independent under a Brownian-motion model [53]. These contrasts can then be analyzed using standard regression or correlation techniques, while controlling for phylogenetic relatedness [54].

Generalized least squares and related approaches extend this framework by explicitly mod-

eling the covariance structure among species as a function of the tree and an evolutionary model [55, 56]. This allows hypothesis testing about processes such as adaptive evolution, constraint, or niche conservatism. Recent developments have pushed comparative methods into the high-dimensional regime. Penalized likelihood and Bayesian frameworks can accommodate many traits simultaneously and regularize complex models, enabling studies of multivariate trait evolution [57–59].

Tree Summary Statistics

A large family of summary statistics has been proposed to quantify different aspects of tree shape, size, and timing (see [60] for an overview). Although this thesis will discuss model-specific statistics later, it is useful here to highlight four broad classes:

1) **Tree balance metrics**: These quantify how evenly lineages branch across the tree. Classical examples include the Sackin index, which sums the depths of all tips [61], and Colless-type indices of imbalance [60]. Imbalanced trees (with many short branches on one side of a split and long branches on the other) may indicate heterogeneous diversification rates or adaptive radiations.

2) **Branching-time statistics**: These describe the temporal pattern of divergence events. The gamma statistic, for instance, summarizes whether branching events are concentrated early or late relative to a constant-rate birth–death model [62]. Deviations from the null expectation can suggest changes in diversification rates through time.

3) **Tree size and depth**: Simple metrics such as the number of tips (species richness) and tree height (the distance from root to most distant tip) provide basic information about the scale of the clade and the time available for diversification.

4) **Spectral and graph-theoretic metrics**: These approaches treat the phylogeny as a graph and summarize its structure via the eigenvalues of matrices such as the (distance) Laplacian or normalized Laplacian. The resulting spectral density profiles and derived statistics capture global features of connectivity and branching that are only partly reflected in classical imbalance or gamma-type metrics [60, 63, 64].

The interpretation of these statistics depends strongly on what branch lengths represent (cladogram vs. phylogram vs. chronogram) and on how the tree was reconstructed. For example, comparing a gamma statistic or a Laplacian spectrum across trees inferred under different clock models or with different fossil calibrations can confound biology with methodology: apparent differences may simply reflect rescaled or redistributed branch lengths.

1.1.6 Beyond Trees: Phylogenetic Networks

The tree model implicitly assumes that lineages only split and never rejoin. However, many evolutionary processes violate this assumption. *Reticulate* processes such as hybridization, introgression, horizontal gene transfer, and recombination create histories in which lineages merge as well as split (Figure 1.6). In such cases, representing evolution as a single bifurcating tree can be misleading: different genes may have different genealogies, and no single tree can capture all the relevant relationships [22, 23, 25, 65].

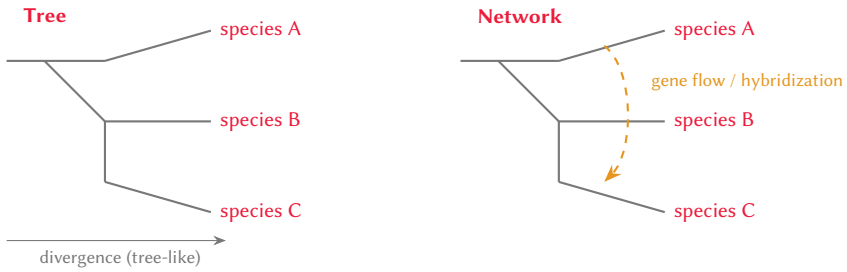


Figure 1.6: From a strictly bifurcating tree (**left**) to a simple phylogenetic network (**right**) with a reticulation event connecting two lineages.

Phylogenetic networks generalize trees by allowing edges that represent reticulation events [65]. Conceptually, a phylogenetic network is a graph in which some nodes may have more than one parent, corresponding, for example, to a hybrid species with two ancestral lineages or to a genome that has acquired genes by lateral transfer. Several types of networks are commonly distinguished:

1) **Implicit (split) networks**: These networks, such as neighbor-net or related methods, are largely exploratory. They visualize conflicting phylogenetic signals as cycles or boxes, without requiring that internal nodes correspond to actual ancestral taxa (Figure 1.7). Such networks are useful for highlighting data conflict or potential reticulation [65, 66].

2) **Explicit (hybridization or species) networks**: These represent hypothesized evolutionary histories where some internal nodes are hybrids or result from gene flow between lineages [67–69]. Here, nodes often have a direct biological interpretation, and reticulation edges correspond to specific evolutionary events (e.g. hybrid speciation in plants or introgression among closely related animals).

3) **Ancestral recombination graphs (ARGs)**: Widely used in population genetics, ARGs describe the genealogy of genetic segments in a recombining population. Recombination events introduce reticulation so that different parts of a genome trace back through different ancestral paths [70, 71].

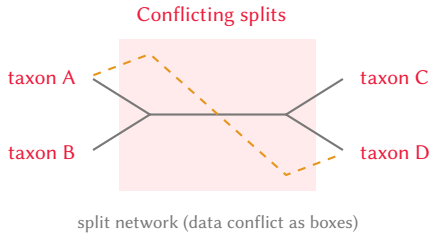


Figure 1.7: Toy split network for four taxa. Two incompatible splits form a “box”, illustrating conflicting phylogenetic signal that cannot be represented on a single bifurcating tree.

Reticulate evolution is especially prominent in microbes, where horizontal gene transfer can blur species boundaries and produce a "web of life" rather than a simple tree [22, 23, 72, 73]. In plants, hybridization and polyploidization are frequent and can drive rapid diversification [74, 75]. Even in animals, genome-scale data have revealed pervasive introgression events (e.g. among hominins), challenging purely tree-like views of evolution [76, 77]. In all of these cases, networks provide a potentially more faithful qualitative model of evolutionary history than trees alone.

At the same time, networks are more complex to infer and interpret than trees, and the data requirements for confidently identifying reticulation events are substantial [65]. For many questions, a tree remains an adequate and more parsimonious representation. A practical strategy is therefore to treat phylogenetic trees as a baseline model and to invoke network representations when there is clear evidence that simple branching is insufficient [67, 69]. In this thesis, the focus will be on *trees with calendar time units (chronograms)* and *tree-based statistics*, while keeping in mind that many real evolutionary histories are embedded in richer, network-like structures.

1.2 Phylogenetic Birth–Death Models

Phylogenetic birth–death models provide a simple but powerful probabilistic framework for describing how speciation ("birth") and extinction ("death") events generate branching trees through time. In the classical constant-rate model, each lineage gives rise to new species at rate λ and goes extinct at rate μ, independently of other lineages and of time [78–80]. When $\lambda > \mu$, the expected number of lineages grows approximately exponentially, and in the absence of extinction ($\mu = 0$) the process reduces to a pure-birth Yule model. Because empirical phylogenies often include only extant species, we typically observe a *reconstructed* tree, in which extinct lineages have been pruned away. For such reconstructed trees, constant-rate birth–death theory predicts that lineage-through-time (LTT) plots—tracking the number of lineages with descendants at the present—should exhibit a characteristic upturn near the present, often called the "pull of the present", as illustrated in the left panel of Figure 1.8 [54, 78, 81]. Intuitively, this arises because lineages that survive to the present have not had enough time to go extinct, so the observed extinction rate near the present is reduced and net diversification appears to accelerate [82].

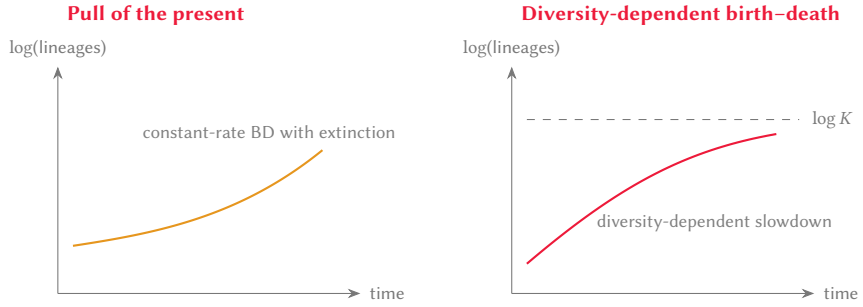


Figure 1.8: Lineage-through-time (LTT) plots. **Left**: approximate expected log LTT under a constant-rate birth–death model with extinction, where the slope increases from the net diversification rate $\lambda - \mu$ early on toward the speciation rate λ near the present, illustrating the pull of the present. **Right**: diversity-dependent model with carrying capacity K, where the expected log LTT (red) gradually approaches $\log K$ (dashed line) as diversification slows down.

Nonetheless, many empirical phylogenies do not meet this simple expectation. Across a wide range of clades, LTT curves instead show a *slowdown* in lineage accumulation toward the present: the log-lineage count bends downward rather than upward, suggesting that net diversification has decreased over time [83–86]. This pattern contrasts with what constant-rate models predict and also with some fossil-based reconstructions of diversity, which often show long-term saturation or fluctuating equilibria rather than recent slowdowns [82, 87]. Reconciling these discrepancies has motivated a variety of extensions to the basic birth–death framework.

Several classes of mechanisms can generate decelerating LTT curves [54, 85]:

1) **Time-dependent diversification rates**: Speciation and/or extinction rates may change through time due to extrinsic factors (e.g. climatic shifts, geological events) or intrinsic evolutionary innovations. For instance, models in which speciation rates decline monotonically through time—sometimes interpreted as decreasing ecological

1

opportunity—can produce early bursts of diversification followed by slowdowns [26, 85, 88].

2) **Protracted speciation**: Under protracted birth–death models, speciation is treated as a multi-stage process rather than an instantaneous event. Lineages first enter an incipient-species state and only later complete speciation, or may fail to do so and be reabsorbed [89, 90]. Because recently initiated speciation events are not yet counted as full species, protracted speciation reduces recent apparent diversification and counteracts the pull of the present. This process is illustrated in Figure 1.9.

3) **Negative diversity-dependent diversification**: In diversity-dependent diversification (DDD) models, speciation rates decrease (and/or extinction rates increase) as the number of species in a clade rises. This can reflect competition for finite resources or niche space, leading to an effective carrying capacity K for diversity [82, 91, 92]. Early in a clade's history, when diversity is low, speciation proceeds nearly unhindered; as diversity approaches K, the net diversification rate declines and LTT curves tend to level off, as in the right panel of Figure 1.8 [93].

4) **Incomplete and non-random sampling**: Empirical phylogenies often include only a fraction of the extant diversity [62, 94, 95]. Missing taxa—especially when sampling is biased toward deeper nodes and recent lineages are underrepresented—inflate terminal waiting times and can make lineage accumulation appear to slow toward the present even if diversification rates were constant [94]. Interestingly, protracted speciation can be viewed as a related form of incomplete sampling: the underlying process contains incipient species, but the reconstructed tree records only completed species. Not all incipient lineages are "observed" as distinct taxa.

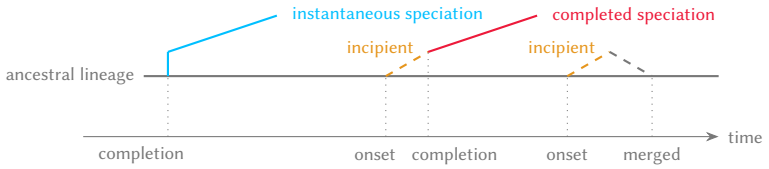


Figure 1.9: Illustration of protracted speciation. A new lineage enters an incipient-species stage (dashed orange) before becoming a fully recognized species (solid red), delaying the appearance of new species in reconstructed phylogenies relative to instantaneous speciation. A new lineage may fail and merge back.

In negative diversity-dependent models, species richness is usually treated as a single scalar state variable that summarizes competition, niche occupancy, and other ecological interactions. In practice, species act as proxies for mechanistic factors such as functional traits, ecological niches, and biotic interactions that ultimately shape ecological limits [96, 97]. This implicitly assumes that all species contribute equally to filling ecological space.

Yet species may not be equivalent entities, as illustrated in Figure 1.10. Closely related species tend to share similar traits and ecological requirements due to common ancestry; they may compete more strongly, respond similarly to environmental change, or share enemies [96, 98–100]. Phylogenetic or evolutionary relatedness thus provides a convenient

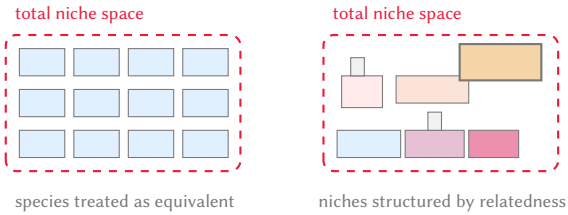


Figure 1.10: Conceptual illustration of species non-equivalence. **Left:** all species are treated as equivalent units occupying similar amounts of niche space. **Right:** niche widths and positions differ, reflecting evolutionary relatedness and ecological similarity.

proxy for ecological similarity and potential interaction strength. For example, in invasion biology, species that are evolutionarily distant from the resident community can be more successful colonists, presumably because they occupy distinct niches and face less competition from close relatives [101–103].

However, being *too* distantly related may also imply a poor match to the local abiotic environment, so that environmental filtering prevents establishment even if competition is weak. This tension between reduced biotic resistance and reduced abiotic fit is one facet of Darwin's naturalization "conundrum" and indicates that performance may peak at intermediate relatedness rather than changing monotonically [104–106]. Taken together, the impact of an additional species on diversification dynamics is very likely to hinge on *how closely related* it is to other members of the clade, a concept illustrated in Figure 1.11.

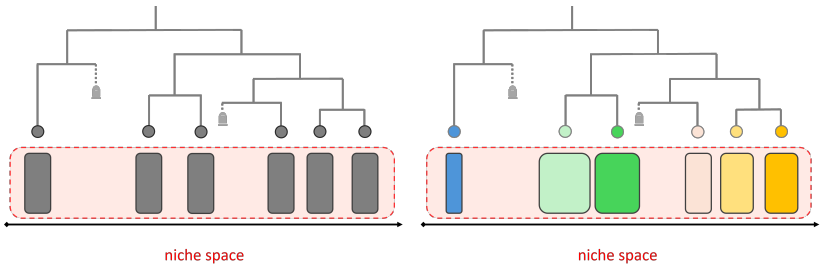


Figure 1.11: Species may not be equivalent entities; their traits and niche spaces can be influenced by their evolutionary relatedness. The left panel illustrates a scenario where all species are considered equivalent, with no regard for relatedness. In contrast, the right panel depicts a scenario where evolutionary relatedness affects traits and niche spaces. The phylogenetic trees in both panels are identical. Tombstone signs mark species that have gone extinct. The red dashed boxes represent the total hypothetical niche space, while the smaller colored boxes indicate the niche spaces occupied by individual species. The color gradient of these smaller boxes reflects the degree of similarity between species, and their width represents the amount of niche space each species occupies.

Most empirical and theoretical work on diversity-dependent diversification has focused on single clades and has assumed that all species within the focal clade contribute equally to limiting diversity [82, 107, 108]. Fewer studies have explored cross-clade or cross-lineage interactions, or explicitly distinguished between the effects of close vs. distant relatives [109]. A recent island-frog study [110] showed that diversity-dependence among closely

1

related species had much stronger effects on colonization and diversification than interactions with more distantly related species, underscoring the importance of phylogenetic scale.

To date, few birth–death models have made evolutionary relatedness an explicit determinant of diversification rates. A natural next step is to use *phylogenetic metrics* based on branch lengths—such as pairwise distances or measures of evolutionary distinctiveness—to modulate speciation and extinction rates in simulations. This is particularly appealing because most birth–death simulation algorithms produce chronograms (branch lengths proportional to time), so time since divergence is directly available and can be interpreted as a proxy for accumulated genetic and ecological differences.

At the same time, birth–death theory reveals that the distribution of branch lengths in simulated trees can be unintuitive. For example, under a pure-birth Yule process, the expected lengths of internal and pendant branches are surprisingly similar and on average shorter than might be naively expected [111]. Because many inference methods are sensitive to branch length distributions, it is crucial to be explicit about what branch lengths represent (time, amount of genetic change, or arbitrary units, as we discussed before) and how they were estimated [80, 82]. When we tie diversification rates to branch-length-based metrics, we implicitly assume that divergence time is a reasonable proxy for the accumulation of genetic and ecological differences. This assumption may hold in some systems but not in others.

Overall, birth–death models have evolved from simple constant-rate descriptions to increasingly rich frameworks that incorporate time-dependence, diversity-dependence, protracted speciation and many more [112–116]. In this thesis, we will introduce a new extension that incorporates dependence on *both diversity and evolutionary relatedness*. Together, these models provide a conceptual bridge between individual-level evolutionary or ecological processes and the large-scale patterns observed in phylogenies and the fossil record. To assess their adequacy and to infer their parameters from data, we rely on a suite of statistical tools, which we discuss next.

1.3 Inference of Diversification from Phylogenies

Inference tools play a central role in connecting birth–death models to empirical data. Given a reconstructed, time-calibrated phylogeny and, where available, fossil information, our goal is to infer diversification parameters (e.g. speciation and extinction rates) and to assess whether patterns such as time-dependence, diversity-dependence, or trait-dependence are supported. Time-calibrated phylogenies encode both branching times and topological relationships among species and thus provide a complementary source of information to the often incomplete fossil record [117]. When the assumed birth–death model approximates the true underlying process, it is possible to estimate speciation and extinction rates from such trees [79, 118, 119].

1.3.1 Likelihood-Based/Bayesian Methods

Classical approaches to diversification inference rely on *likelihood*-based frameworks. For a given birth–death model and parameter values, one can, in many cases, write down the probability of observing a particular reconstructed tree (or at least its ranked branching times) under that model [54, 78, 79, 120]. Maximizing this likelihood over the parameter space yields maximum likelihood estimates (MLEs) of speciation and extinction rates [119, 121]; alternatively, embedding the likelihood in a Bayesian framework allows one to obtain posterior distributions for parameters via Markov chain Monte Carlo (MCMC) sampling [122].

Over the past two decades, likelihood-based models have been extended in several directions. Time-varying models allow speciation and extinction rates to change as functions of time or environmental covariates [26, 88]. Diversity-dependent models explicitly link rates to the current number of species in the clade [82, 114, 123]. Trait-dependent or state-dependent models (e.g. BiSSE and its extensions) allow speciation and extinction to depend on discrete or continuous traits evolving along the tree [112, 113, 124]. Likelihood-based methods for these models underpin many macroevolutionary analyses in the literature, and they remain the standard against which new approaches are often compared.

However, likelihood-based inference also faces important challenges. Closed-form likelihood expressions are only available for relatively simple models; more complex models can require intricate derivations or numerical approximations that are computationally expensive [125, 126]. For many biologically realistic scenarios—for example, models with strong lineage heterogeneity or multiple interacting clades—deriving a tractable likelihood is prohibitively difficult. Even when a likelihood is available, identifiability issues can limit what can be learned from a single extant-only tree: different combinations of speciation and extinction rates may produce virtually indistinguishable reconstructed trees, especially when extinction is high [87, 127]. Small trees further exacerbate estimation bias; for example, MLEs of extinction often collapse to zero for modest-sized trees even when the true extinction rate is nonzero, whereas estimates improve as tree size (number of tips) increases [121]. Likelihood-based methods with many free parameters can also be prone to overfitting and may yield unstable estimates when data are limited [128].

1

1.3.2 Approximate Bayesian Computation

When likelihoods are unavailable or too costly to compute, *Approximate Bayesian Computation* (ABC) offers an alternative route to parameter inference. ABC approximates the posterior distribution of parameters by (i) simulating data under proposed parameter values, (ii) summarizing both simulated and observed data by a set of summary statistics, and (iii) retaining parameter values that produce simulated statistics close to the observed ones [129, 130]. In principle, ABC can handle any diversification model so long as it is easy to simulate phylogenies from it.

In phylogenetics, however, applications of ABC remain relatively scarce [131–134]. A central difficulty is choosing informative yet low-dimensional summary statistics that capture the relevant aspects of tree shape and branching times. Simple statistics, such as tree balance indices, the gamma statistic, or LTT-based features, may not be sufficient to distinguish between, say, time-dependent and diversity-dependent scenarios. In contrast, including many statistics can lead to a high-dimensional summary space in which ABC performs poorly. As a result, while ABC is conceptually attractive for complex birth–death models, careful design of summary statistics and distance metrics is critical and has so far limited its routine use in diversification studies [134].

1.3.3 Machine Learning and Deep Learning Approaches

Rapid advances in machine learning (ML)—particularly deep learning—offer new possibilities for inferring macroevolutionary parameters from phylogenetic data. Deep neural networks are function approximators that can learn complex, nonlinear relationships between inputs and outputs from large training datasets [135]. Figure 1.12 illustrates a simple example of a neural network architecture. In our context, the inputs are phylogenetic trees (or representations thereof), and the outputs are diversification parameters or model labels. Unlike classical ABC, deep learning can learn directly from high-dimensional or raw representations of trees without requiring hand-crafted summary statistics.

A range of network architectures and tree representations has been explored:

1) **Feature-based encodings and feed-forward networks**: One strategy is to compute vectors of tree-level features (e.g. branching-time summaries, balance indices, LTT-derived quantities) and feed them into standard multilayer perceptrons. This is straightforward to implement and has been used in several early applications of neural networks to phylogenetic data [126, 136]. Its performance, however, depends heavily on the choice of features.

2) **Sequence encodings and recurrent neural networks**: Another approach is to serialize information about the tree, for example by ordering internal node ages from root to tips and treating the resulting sequence as a time series. Recurrent neural networks (RNNs), including LSTM and GRU variants, are designed to handle such sequential data and can therefore capture temporal dependencies in branching patterns [137, 138]. This allows them to learn from the ordered history of diversification events rather than from aggregated summaries.

3) **Graph-based representations and graph neural networks**: Because phylogenies are graph-structured, graph neural networks (GNNs) offer an attractive way to

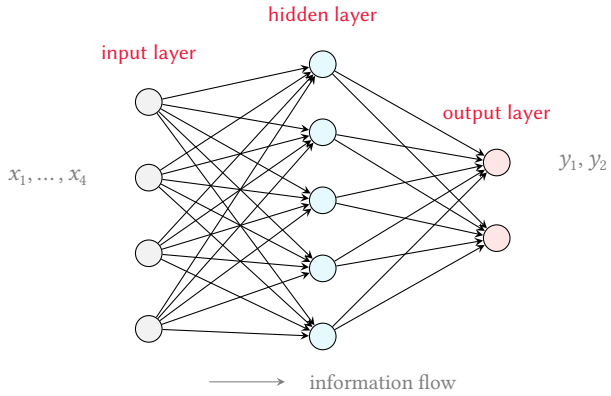


Figure 1.12: Illustrative multi-layer perceptron with one hidden layer. Each neuron in a layer is fully connected to all neurons in the next layer, and information flows from left to right. This feed-forward architecture represents a basic building block of neural networks; deeper, convolutional, recurrent, or more complex models can all be viewed as extensions or refinements of this simple layered structure.

process them directly. In GNNs, each node (e.g. a speciation event or species) aggregates information from its neighbors over multiple message-passing layers, yielding a learned representation of the entire tree that can be used for parameter prediction [139–143]. This avoids discarding structural information but requires careful design of node and edge features.

4) **Topology-aware vectorizations and convolutional networks**: Recently, compact bijective encodings or ladderized vectorizations of trees have been proposed, which map each tree to a fixed-length vector or matrix encoding both topology and branch lengths. Convolutional neural networks (CNNs), originally developed for image recognition, can then be applied to these encodings to detect local patterns associated with particular parameter values [126, 143–146]. These methods can automatically learn rich summary features without manual feature engineering.

Empirical evaluations of deep learning for diversification inference are encouraging. Lambert et al. [126] trained CNN-based models on simulated phylogenies under both homogeneous birth–death and state-dependent models and found that their networks could accurately recover speciation and extinction rates, often matching or exceeding the performance of MLE while being orders of magnitude faster at prediction. Similarly, Voznica et al. [144] and Reiman et al. [146] demonstrated that deep learning models can infer epidemiological parameters from pathogen phylogenies. Lajaaiti et al. [143] and Qin et al. [147] compared multiple architectures (e.g. feed-forward, CNN, GNN) and highlighted trade-offs between flexibility, data requirements, and robustness across models.

Despite their promise, neural network approaches also have limitations. We must explicitly define a loss function that quantifies the discrepancy between predictions and targets and then minimize this objective using backpropagation (see Figure 1.13 for an illustration) and gradient-based numerical optimization [148]. Training neural networks requires large

1

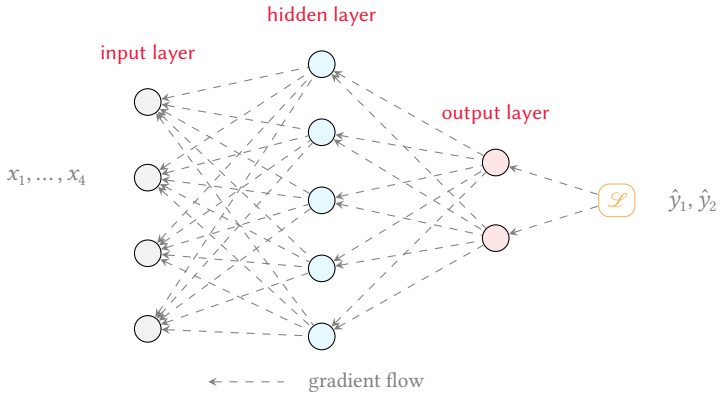


Figure 1.13: Training view of a simple multi-layer perceptron. Dashed gray arrows indicate the flow of gradients during backpropagation: the loss $\mathscr{L}$ is computed from the output predictions and then propagated backwards through the output, hidden, and input layers to update the network parameters.

numbers of simulated trees that adequately span the parameter space, and trained models may generalize poorly to empirical data that lie outside this space [147]. Uncertainty quantification is less straightforward than in Bayesian frameworks, although approaches such as Bayesian neural networks or ensemble methods can provide approximate credibility intervals [149, 150]. Neural-network-based estimators are not immune to the non-identifiability issues that affect classical likelihood approaches. If different parameter combinations generate phylogenies with indistinguishable or weakly distinct patterns, then no supervised learning method—no matter how flexible—can reliably disentangle those parameters from extant trees alone.

Neural models are also less interpretable than likelihood-based ones: they can predict parameter values accurately but generally do not yield simple analytical expressions or mechanistic insight. Consequently, deep learning should be viewed as complementary to, rather than a replacement for, traditional methods. For example, neural networks can quickly explore complex models or large datasets, generating hypotheses that can then be tested more formally with likelihood-based approaches where feasible.

After all, whether we maximize a likelihood in MLE or minimize a loss for a neural network, both procedures rely on numerical optimizers navigating a potentially rugged or even non-convex objective landscape and can therefore suffer from similar issues such as sensitivity to initial values and convergence to local optima (see Figure 1.14) [151, 152].

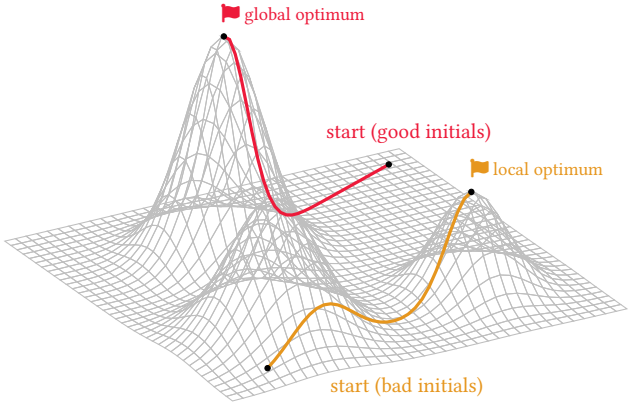


Figure 1.14: Illustrative log-likelihood surface with one global optimum and two local optima. The colored curves show two optimization trajectories: one (orange) that gets trapped in a local maximum and one (red) that reaches the global maximum, depending on the starting point. This highlights how, for complex and multi-modal likelihood functions, gradient-based MLE can be highly sensitive to initial values and may converge to suboptimal solutions, complicating reliable parameter estimation and uncertainty assessment. Training neural networks relies on similar gradient-based optimization on equally rugged loss landscapes, so deep-learning approaches to likelihood-free inference can suffer from the same issues, even though they operate on a different objective.

1

1.4 Emerging Directions and Future Prospects

Recent theoretical work has shown that some limitations of diversification inference are *structural* rather than merely computational. Under very general birth–death formulations, extant phylogenies alone cannot uniquely identify past speciation and extinction trajectories: an infinite number of distinct rate histories can be congruent with the same reconstructed tree [153]. This result highlights an inherent limitation of treating a single, time-calibrated tree as the sole data object: even with infinite sequence data and perfect topology, much information about diversification histories is simply not present in the tree. A growing body of work therefore focuses on model families that are both biologically interpretable and statistically identifiable. For example, Legried and Terhorst [154, 155] demonstrate that broad classes of piecewise-constant or piecewise-polynomial birth–death models are identifiable from sufficiently large chronograms, providing guidance for rate parameterizations. In parallel, methods such as `CRABS` explore the *congruence classes* of birth–death models that share the same likelihood, allowing users to summarize which qualitative features of inferred rate dynamics (e.g. the presence of slowdowns) are robust [156]. Together, these developments suggest that future diversification studies will increasingly rely on constrained, regularized rate models and on explicit exploration of congruent alternatives, rather than on unconstrained, highly flexible trajectories.

A second key direction is to move beyond extant timetrees as the only representation of macroevolutionary history by incorporating richer data sources. Fossil-inclusive models such as the fossilized birth–death (FBD) process jointly describe diversification, fossil sampling, and extant diversity in a single probabilistic framework [27]. By embedding fossils directly into the tree and explicitly modeling sampling, FBD-based analyses can tighten divergence-time estimates and constrain diversification rates beyond what is possible with extant-only trees. More generally, joint Bayesian frameworks that combine molecular sequences, fossils, traits, and environmental covariates offer a way to mitigate the information bottleneck inherent to trees alone [157–161].

Moreover, as we discussed earlier, processes such as hybridization, introgression, horizontal gene transfer and recombination produce reticulate patterns of ancestry that are poorly captured by a strictly bifurcating tree [65]. Phylogenetic networks generalize trees by allowing reticulation edges and can represent conflicting signals or non-treelike histories [65]. Recent work argues that such networks will be increasingly important in biodiversity research, as they can capture historical connectivity and gene flow among lineages that a single tree would obscure [162]. In the context of diversification, network-based models remain methodologically demanding and are only beginning to be explored.

A complementary set of advances concerns *inference technology* itself. On the one hand, approximate Bayesian methods based on optimization, such as variational Bayesian phylogenetic inference (VBPI), are beginning to provide scalable alternatives to Markov chain Monte Carlo for full Bayesian phylogenetics. VBPI approximates the posterior over trees and model parameters using a flexible graphical model, trained by stochastic gradient ascent [163]. Extensions that incorporate normalizing flows further enrich the branch-length component of the approximation and can closely match MCMC posteriors at a fraction of the computational cost [163]. On the other hand, *simulation-based inference* (SBI) uses

neural density estimators to learn likelihoods, posteriors or likelihood ratios directly from simulator output, without requiring closed-form likelihoods [164]. Benchmarks across scientific domains show that state-of-the-art SBI methods based on neural likelihood or posterior estimation can be both more sample-efficient and more accurate than classical ABC, particularly in high-dimensional settings [165]. For diversification models, SBI potentially offers a promising route. Once a neural estimator has been trained on simulated trees spanning a parameter space, it can provide approximate posteriors for many empirical trees at a much lower cost.

Looking ahead, we expect progress in diversification inference to come from a *joint evolution* of models, data and algorithms. Theoretical work on identifiability and congruence highlights which aspects of diversification histories can be recovered from extant trees and which require additional information. Richer data sources—including fossils, traits, spatial distributions and genomic signatures of gene flow—offer routes to overcome some of the inherent limitations of tree-based representations. Advances in machine learning, from VBPI to SBI inference, provide scalable tools for exploiting these data under increasingly complex and realistic models.

1

1.5 Scope and Structure of this Thesis

This thesis investigates how much information about macroevolutionary diversification can be extracted from extant phylogenies when we move beyond simple, constant-rate birth–death models and combine them with modern machine learning. On the modeling side, we extend classical diversification frameworks by allowing speciation and extinction to depend not only on species richness but also on evolutionary relatedness among lineages. On the inference side, we develop and evaluate neural-network-based methods that learn from simulated trees, probing when such approaches can recover diversification parameters and mechanisms, and when potential limits in the information content of trees make this impossible. Throughout, we focus on chronograms in which branch lengths represent time, and we treat neural networks not as black-box replacements for likelihood-based inference, but as flexible tools to map out the boundaries of what can and cannot be learned from phylogenetic trees under increasingly complex birth–death models.

The main body of the thesis consists of three research chapters, each addressing a different aspect of this program:

1) **Diversity, evolutionary relatedness, and tree shape.** We introduce a birth–death model (`eve`) in which speciation rates depend explicitly on measures of evolutionary relatedness between species, in addition to overall species richness. Using a suite of relatedness metrics that operate at different phylogenetic scales (from lineage-specific to clade-wide), we simulate trees and analyze how these scales influence standard tree statistics. We show that whole-tree relatedness effects generate smaller and more balanced trees, with speciation rates distributed evenly across tips, whereas lineage-specific effects yield different, often more imbalanced patterns. We also demonstrate that negative richness dependence can mask the signatures of relatedness in some statistics, and we identify combinations of richness and relatedness effects that reproduce the imbalanced trees commonly seen in empirical phylogenies.

2) **Neural-network estimation of diversification parameters.** We develop an ensemble deep-learning framework for likelihood-free inference of diversification parameters from time-calibrated trees. Our approach combines multiple architectures—a dense feed-forward network, a graph neural network, and a long short-term memory recurrent network—and allows the ensemble to learn simultaneously from graph-structured representations of phylogenies, their branching-time sequences, and vectors of tree-summary statistics. Using simulated trees from constant-rate, diversity-dependent, and protracted birth–death models, we compare our ensemble to classical maximum likelihood estimators and existing convolutional-network approaches. We find that the ensemble delivers estimates that are faster to compute and less sensitive to tree size than MLE in several scenarios, while matching or exceeding MLE accuracy when phylogenetic signals are strong, but that both MLE and neural methods struggle under protracted speciation. This chapter highlights both the potential and the limitations of neural networks as practical tools for parameter estimation in diversification models.

3) **Neural recoverability and complex diversification models.** We use the `eve` model as a testbed to examine when neural networks can recover diversification

mechanisms from extant trees. Training graph neural networks and long short-term memory classifiers on simulated trees, we ask how well they can distinguish among scenarios in which diversification depends on phylogenetic diversity, evolutionary distinctiveness, or nearest-neighbor distance, and we complement this with regression networks that aim to recover the underlying parameters. By analyzing classification accuracy, probability calibration, regression errors, and their dependence on tree size and the strength and sign of richness and relatedness effects, we map out regions of parameter space where scenarios and parameters are practically recoverable, and others where trees potentially carry too little information. This chapter provides an empirical perspective on recoverability in richly parameterized diversification models and clarifies how far neural inference can be pushed before additional data or constraints become non-negligible.

2

Diversity, Evolutionary Relatedness, and Tree Shape

Slowdowns in lineage accumulation are often observed in phylogenies of extant species. One explanation is the presence of ecological limits to diversity and hence to diversification. Previous research has examined whether and how species richness (SR) impacts diversification rates, but rarely considered the evolutionary relatedness (ER) between species, although ER can affect the degree of interaction between species, which likely sets these limits. To understand the influences of ER on species diversification and the interplay between SR and ER, we present a simple birth–death model in which the speciation rate depends on the ER. We use different metrics of ER that operate at different scales, ranging from branch/lineage-specific to clade-wide scales. We find that the scales at which an effect of ER operates yield distinct patterns in various tree statistics. When ER operates across the whole tree, we observe smaller and more balanced trees, with speciation rates distributed more evenly across the tips than in scenarios with lineage-specific ER effects. Importantly, we find that negative dependence of speciation masks the impact of ER on some of the tree statistics. Our model allows diverse evolutionary trajectories for producing imbalanced trees, which are commonly observed in empirical phylogenies but have been challenging to replicate with earlier models.

Qin, T., Valente, L.†, & Etienne, R.† (2025). Impact of evolutionary relatedness on species diversification and tree shape. *Journal of Theoretical Biology* [166]. † indicates joint senior authors.
Awarded the inaugural *Denise Kirschner Best Student Paper Prize* (2025).

2.1 Introduction

Phylogenetic trees are important tools to estimate past processes that may explain the current species richness of clades, such as diversification rates. Over the last two decades, the increasing availability of DNA sequence data and of tools to reconstruct phylogenetic trees from these data has led to the development of birth–death models that use molecular phylogenies as a source of information to study diversification dynamics [82, 83, 167, 168]. Lineage-through-time (LTT) plots, semi-logarithmic plots that track the number of lineages that have descendants at the present through time, are a powerful way of summarizing diversification dynamics. If per-lineage rates of speciation and extinction have been constant through time, the accumulation of lineages increases through time exponentially, with an even stronger increase close to the present, a phenomenon called the "pull-of-the-present" [78, 81]. However, a large number of empirical phylogenies display a different pattern: they often show recent deceleration [83–86], which contrasts with findings from the fossil record [82, 87].

Several hypotheses have been put forward to explain the observed slowdowns in phylogenies of extant species [85], such as time-dependent speciation rate [85], protracted speciation [89] and negative diversity-dependent diversification [82, 91, 92]. In negative diversity-dependent diversification models, the speciation rate declines with increasing species diversity (SR).

Current models of diversity-dependent diversification assume that speciation rate depends on global SR, regardless of evolutionary relationships of the species. The general underlying idea is that species diversification results in the occupation of available niche spaces, leaving limited opportunities for subsequent species to use those niches (Figure 2.1), because evolving clades compete for finite ecological resources [93]. In these models, SR acts as a proxy for more mechanistic factors such as functional traits, niches and ecological interactions, which ultimately may influence ecological limits [96, 97]. While past ecological interactions between species are difficult to infer, useful proxies are evolutionary relatedness (ER) metrics, which quantify the phylogenetic distance between taxa. ER can represent the ecological roles of the species well [98, 100], as closely related species are more likely to share trait states and ecological functions than distantly related species [96], and evolutionary distance may be related to fitness and niche differences [99], which in turn can affect speciation and extinction. Here, we may learn from insights from phylogenetic community ecology [51, 169, 170] and invasion biology. For example, alien species that are evolutionarily distant from those in the local community may have a greater chance of establishing [101], so using ER as a proxy to quantify open niche spaces may explain the success of plant invaders [102, 103].

The potentially powerful phylogenetic proxies for ecological interactions, however, have been the subject of intense debate because the underlying assumptions are often not robustly supported [100, 171]. Indeed, there is mixed evidence regarding the effectiveness of these proxies. On the one hand, Tucker et al. [172] concluded that phylogenetic diversity is a useful proxy for functional diversity, because phylogenetic diversity correlates more strongly with functional diversity when trait dimension increases (multidimensional trait space), although that correlation is weakened when trait evolution models become increas-

ingly complex. On the other hand, Mazel et al. [173] found that in many cases phylogenetic diversity only poorly captures functional diversity, even less so than random selection. Furthermore, Venail et al. [174] found that trait and functional variation among species is largely explained by SR but not phylogenetic relatedness (a concept similar to ER). Recently, Pie et al. [175] found that sympatry with closely-related species does not lead to decreasing speciation rates in a variety of vertebrate clades, implying that the underlying mechanisms of diversity-dependent diversification remain unconfirmed.

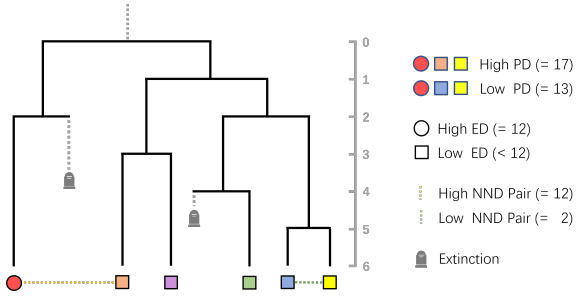


Figure 2.1: Illustration of how evolutionary relatedness (ER) is computed within clades and species in a phylogeny, as measured by three metrics: PD (phylogenetic diversity), ED (evolutionary distinctiveness), and NND (nearest neighbor distance). The tree on the left represents a phylogeny with extant and extinct (marked by tombstone icons) species. The numbered axis in the middle marks the branch lengths in evolutionary time (million-years). The colored circle and squares at the tips of the phylogeny represent corresponding species. On the right of the plot, the species denoted by a gray circle, an orange square and a yellow square form a combination of high PD (17), while the species denoted by the gray circle, the blue square and the yellow square form a combination of low PD (13). The species with the highest ED value (12) is represented by a circle, and those with lower ED values (less than 12) are represented by a square. The yellow and green dashed-lines between species illustrate two pairs of species, one with a low NND value (2) and one with a high NND value (12).

While these concerns indicate that support for a role of ER on diversification is not clear, this may be because we do not know what signal ER is expected to leave in phylogenetic trees and extant communities [93, 131, 176]. Most studies have only considered diversity-dependence within clades of phylogenetically closely related species [82, 107, 108], whereas only a few have considered diversity-dependence between clades of phylogenetically disparate species [109]. A recent study [110] using empirical data on island frogs found that a model with diversity-dependence between closely related species was preferred over one with diversity-dependence also occurring with more distantly related species, indicating that interactions of species with close relatives, but not with distant relatives, negatively affect colonization and diversification. This suggests an important role of ER in species diversification and the need for distinguishing between different levels of ER and scales at which it may have an effect on diversification. However, none of the above-mentioned studies has investigated whether ER directly impacts macro-evolutionary dynamics [177].

Assuming ER affects diversification rates, what signatures does this effect leave in phylogenetic trees of communities? What emergent phylogenetic patterns are expected if species facilitate or compete more strongly with their close relatives? Can phylogenetic limits imposed by ER be differentiated from those imposed by SR? And how do ER effects operat-

ing at the whole clade versus more lineage-specific scales affect phylogenies? To address these questions, here we present a new phylogenetic birth–death simulation model that incorporates both SR and ER. The model allows for positive, neutral, or negative effects of both SR and ER on diversification rates. We measure ER using three different mechanisms of how ER can affect diversification (see Methods) each considering a different scale of the effect of ER. We use the new model to simulate communities under a stochastic branching process where speciation rate can vary according to SR and ER, and analyze whether ER leaves a signature on the diversification dynamics by looking at various tree summary statistics. We aim to provide the first expectations for the effects of these processes on phylogenetic trees by developing a simple simulation model where the effects of SR and ER can be studied independently or in combination. While inferring parameters from empirical datasets is beyond the scope of this study, our model presents a tool that can be used in future simulation-based parameter estimation methods, for example for generating training datasets for neural network models. Our R package `evesim` on GitHub contains functions to generate simulated phylogenies using our model [178].

2.2 Methods

2.2.1 Model

We employed a phylogenetic stochastic model that simulates the processes of species birth and death over time [78]. The model is a mechanistic process-based model that is parameterized by speciation and extinction rates along with the effect sizes of SR and ER. It allows us to explore different evolutionary trajectories from phylogenetic trees given specific parameter settings at the start of the simulation. We assumed that all species on the same tree belong to the same clade, thus our study focuses on species diversification patterns within a clade, and the phylogenies simulated include all the extant species in the clade. We investigated different types of effects of ER on species diversification by considering three measures of ER: phylogenetic diversity (PD, community-level metric), evolutionary distinctiveness (ED, per-lineage metric), and nearest neighbor distance (NND, per-lineage metric).

PD measures the amount of evolutionary history represented by a group of species and is commonly used to determine how species occupy different niches [172]. We calculated PD using Faith's index [48], which represents the total branch length of a phylogenetic tree reconstructed from all species in a clade. ED quantifies the uniqueness of each of the species relative to other species and is a valuable tool in conservation efforts [179]. Per species, ED is defined as the sum of the pairwise distances between focal species and all other species, divided by the number of species minus one. NND also quantifies the uniqueness of a species, but only on a very local phylogenetic scale, as it is defined as the phylogenetic distance (branching length of the path) between a focal species and its nearest neighbor. NND measures the degree to which each of the species is locally clustered within specific clades, and is less sensitive to higher-level phylogenetic structures than ED [50]. Unlike PD, which is a clade-level metric and assigns a single value for all species in the clade, ED and NND are lineage-specific metrics, with each lineage assigned its own value.

There are many other choices of ER metrics, e.g. the "Fair Proportion" index [180] and the

"Evo-Heritage" metric [181]. However, we aimed to use metrics that are computationally efficient.

λ_i, the speciation rate of a specific lineage i, is given by

$$\lambda_i = \max\left(\lambda_0 + \beta_N N + \beta_\Phi \Phi_i,\ 0\right) \tag{2.1}$$

where λ_0 is the intrinsic speciation rate of all the lineages, N is the species richness (across all lineages), Φ_i represents ER (either PD, ED or NND) of the lineage i, β_N is a coefficient to adjust the effect size of species richness on the speciation rate and β_Φ is a coefficient to adjust the effect size of ER on the speciation rate. β_N and β_Φ can be positive, zero, or negative. We note that by negative ER we mean negative β_Φ and thus that as species become evolutionarily less related, speciation rate decreases. Furthermore, if we set both β_N and β_Φ to zero, the model reduces to a standard birth–death model.

When using PD as an ER metric, we assume that Φ_i is the same for every lineage i, so the speciation rates of all lineages are equal. Unlike PD, both ED and NND are calculated separately for every lineage.

To account for the strong correlation between phylogenetic diversity and time when using PD, we included an offset method in our model to compensate for the inflation of branch lengths in the phylogenetic tree. The method subtracts the tree age t from Φ_t, which is the phylogenetic diversity at time t. The adjusted Φ'_t is then given by

$$\Phi'_t = \Phi_t - t. \tag{2.2}$$

We assumed that the extinction rate μ is fixed to the intrinsic rate of extinction which is constant through time and across lineages:

$$\mu = \mu_0. \tag{2.3}$$

2.2.2 Simulations

We ran a series of simulations of the model under different scenarios in order to investigate the effect and signature of ER on various phylogenetic summary statistics. We started the simulations with two ancestral lineages, setting the values of ER of both lineages to zero. We used the Gillespie algorithm [182], in which the waiting time between two evolutionary events is sampled from an exponential distribution with a mean equal to the inverse of the sum of the rates of all possible events. The probability of each event is proportional to its own rate relative to the sum of the rates of all possible events.

Two types of events can occur: speciation and extinction. When a speciation event happens, a lineage at the tip of the tree bifurcates into two lineages. When an extinction event happens, a species is marked as extinct. The simulation lasts for a predetermined time, which equals the crown age of the final phylogeny. A successful simulation is conditional on survival of both crown lineages; the simulation will start over if one of the crown

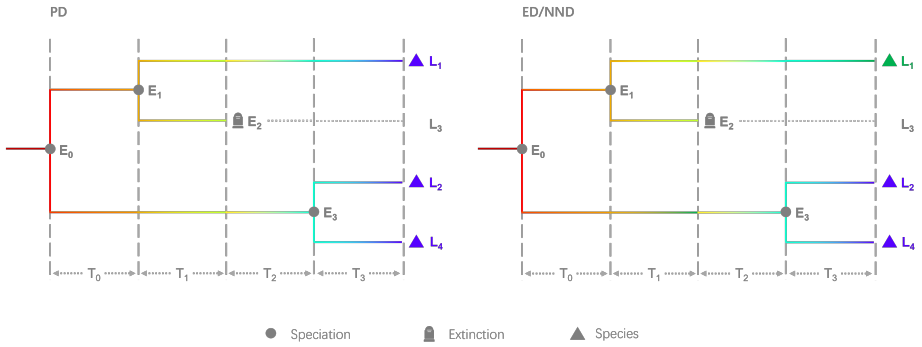


Figure 2.2: Illustration representing a stochastic simulation under the assumptions of positive speciation and extinction rates alongside a negative coefficient of evolutionary relatedness, exemplified through a scenario where phylogenetic diversity (PD, left panel) or evolutionary distinctiveness (ED, right panel) or nearest neighbor distance (NND, also right panel) acts as clade-wide (PD) or lineage-specific (ED and NND) constraint. Although ED and NND scenarios are illustrated in the same panel, the underlying processes are different. Two event types are depicted: speciation (solid gray circles) and extinction (tombstone symbols). During a speciation event, a lineage at a tree tip bifurcates into two lineages (e.g., E_0, E_1 and E_3). An extinction event marks a species as extinct (e.g., E_2). In the left panel, the branch color transition from red to blue signifies the variation in the speciation rate of lineages along PD. In the right panel, the branch color transition from red to blue or green signifies the variation of per-lineage speciation rates within the phylogeny due to lineage specific (ED or NND) phylogenetic diversity-dependence. The simulation unfolds as follows: it initiates from E_0 with two ancestral lineages L_1 and L_2, setting the initial PD value to 0. Speciation rates for all extant lineages (initially L_1 and L_2) are derived from PD. The first time interval, T_0, extends the branch lengths of L_1 and L_2. Prior to sampling event E_1, PD is recalculated based on extant lineages, and speciation rates are updated. Event E_1 illustrates $L1$ bifurcating into L_1 and L_3. The subsequent time interval T_1 further extends the branch lengths of L_1, L_2 and L_3. Before sampling each event, PD and speciation rates are updated. Event E_2 marks the extinction of L_3, halting its branch growth while L_1 and L_2 continue to extend through time interval T_2. Event E_3 illustrates the speciation of L_2 into L_2 and L_4. The final time interval, T_3, stops the simulation because the cumulative time ($T_0 + T_1 + T_2 + T_3$) surpasses a pre-determined time threshold, T. T_3 is then set to $T - (T_0 + T_1 + T_2)$. The branch lengths of L_1, L_2, and L_4 extend by T_3, marking the simulation endpoint.

lineages goes extinct entirely (see Figure 2.2 for the illustration of the simulation). Our GitHub repository eve [183] contains the codebase for the current study.

We simulated phylogenies using a variety of parameter combinations (see Table 2.1). We assumed a crown age of 6 time units, which can be interpreted as 6 million years. All combinations of parameters in Table 2.1 were used, to a total of 135 combinations, and each was repeated for the three different ER scenarios: PD, ED and NND. Thus, we had a total of 405 parameter sets. For each set we simulated 100 phylogenetic trees using the Peregrine high performance computing cluster of the University of Groningen. In our preliminary tests, we increased the number of replicates from 100 to 300 and then 1000 for the fastest parameter settings, but we did not observe noticeable trend changes with the increased number of replicates. Due to hardware limitations, we retained the number of 100 for consistency across all combinations. The parameter sets were chosen such that the simulation can be finished within the time and resource limits of the cluster. Moreover, the effects of β_Φ on the diversification process are inherently non-symmetrical because positive β_Φ results in a positive feedback, which has a much greater influence on the final

Table 2.1: Parameters used in the simulations.

Parameter	Value
Intrinsic speciation rate	0.4, 0.5, 0.6
Intrinsic extinction rate	0, 0.1, 0.2
Crown age	6
Coefficient species richness N (β_N)	-0.04, -0.02, 0
Coefficient evolutionary relatedness (β_Φ)	-0.04, -0.02, 0, 0.001, 0.002
Evolutionary relatedness metric	PD, ED, NND

size of the phylogenies. For this reason, we kept the positive β_Φ values relatively small. Typically, when setting $\lambda_0 = 0.6$, $\mu_0 = 0$, $\beta_N = 0$, and $\beta_\Phi < -0.1$, the output phylogeny will be very small (often less than five lineages) and hence not very meaningful. If $\beta_\Phi > 0.002$, then the resulting phylogenies will be very large (often more than 500 lineages), which creates computational problems (many simulation steps with compututionally demanding calculation of the phylogenetic metrics).

We deliberately chose a relatively simple model to assess the effect of ER on diversification, rather than including a variety of realistic ecological interactions linked with ER. It allows us to explore parameter space better, and it may facilitate future parameter estimation from phylogenies. Most importantly, it allows gaining broad generalizable insights based on the fundamental processes we are interested in (i.e., ER and SR effects on diversification).

2.2.3 Data Analysis

Raw output from the simulations was processed using the `eve` package into data compatible with statistics functions in other R packages. For each parameter set, the `treestats` [60] and `eve` packages were used to calculate summary statistics for all extant trees, excluding those with only two extant lineages. The statistics chosen were the following: the J One balance index [184], the Gamma statistic [62], mean branch length, mean pairwise distance (MPD) [51] and the Rogers J index of imbalance (Rogers hereafter) [185].

The effects of SR and ER were measured by the values of β_N and β_Φ, respectively. The effects of the scale at which the effect operates, from whole-clade, to species-specific were determined by the ER mechanism (one of the three ER measures: PD (whole-clade), ED (intermediate) and NND (species-specific).

2.2.4 Speciation Rate Evenness

In the ED and NND scenarios, speciation rates are expected to vary between tree tips. We measured how these rates are distributed in phylogenies by adopting a concept similar to measuring species evenness in a community, but instead quantifying the evenness of speciation rates across lineages weighted by their phylogenetic distances. Each phylogeny of n lineages was given by a correlation matrix C with each lineage i's speciation rate represented by λ_i ($i = 1, 2, 3, \ldots, n$). The phylogenetic evenness index E is then defined as

$$E = \frac{\lambda \operatorname{diag}(C)^\top m - m^\top C m}{\lambda^2 - \bar{\lambda}_i \lambda} \tag{2.4}$$

which was originally proposed by Helmus et al. [98].

In the equation, λ denotes $\operatorname{sum}(\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n)$ and $\bar{\lambda}_i$ denotes $\operatorname{mean}(\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n)$, $\operatorname{diag}(C)$ denotes a column vector comprising the diagonal elements of the correlation matrix C (see below), m denotes an $n \times 1$ column vector containing values of λ_i. The value of E ranges between 0 and 1. The maximum value (i.e., 1) of E only occurs when speciation rates among lineages are equal and the tree is completely balanced (star-like). Values of E less than 1 represent increasing unevenness of speciation rates among the lineages. E is sensitive to tree topology such that trees with uniform speciation rates can have different values of evenness.

The correlation matrix C is the standardized pairwise distance matrix. Because we only consider ultrametric (no extinct lineages) and binary (fully resolved) phylogenies, C can be computed as:

$$C = \frac{2t - R}{2t} \tag{2.5}$$

where R is the pairwise distance matrix of all the lineages of the phylogeny of n lineages, with r_{ij} $(i = 1, 2, 3, \ldots, n; j = 1, 2, 3, \ldots, n)$ represents the pairwise phylogenetic distance between lineage i and j:

$$R = \begin{bmatrix} 0 & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & 0 & r_{23} & \cdots & r_{2n} \\ r_{31} & r_{32} & 0 & \cdots & r_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \cdots & 0 \end{bmatrix}. \tag{2.6}$$

As shown in Equation 2.5 and Equation 2.6, all elements of $\operatorname{diag}(C)$ are equal to one, Equation 2.4 can thus be simplified as:

$$E = \frac{n}{n-1}(1 - m'^\top C m') \tag{2.7}$$

where $m' = m/\lambda$.

Note that this simplification is only possible for ultrametric trees with $n \geq 2$, a formal proof of Equation 2.5 and the derivation of Equation 2.7 are provided in Appendix G and Appendix H.

In the case of simulated data, the models and parameters are already known to differ. Testing for statistical significance between treatments is thus meaningless; the distributional changes sufficiently demonstrate the influence of the parameters. Therefore, we focused on how summary statistics vary with the strength of ER and SR to explain the model's power and effects.

2.2.5 Data Visualization

For visualization purposes, we selected one representative tree for each tree set of 100 trees representing a single parameter set. This tree was chosen among the complete trees based on its index vector, which has the smallest mean Mahalanobis distance [186] to the other trees in the same tree set (as defined in Equation 2.8 below). In order to calculate the Mahalanobis distance for each tree, we first defined the index vector for the i-th tree in a set of trees resulted from k-th parameter set, where k denotes one of the parameter sets in the 405 combinations in our simulation, as

$$v_{i,k} = \left(v_{\mathrm{J}}^{i,k}, \quad v_{\mathrm{G}}^{i,k}, \quad v_{\mathrm{P}}^{i,k}, \quad v_{\mathrm{M}}^{i,k}, \quad v_{\mathrm{R}}^{i,k}\right)^{\top} \tag{2.8}$$

where each element of $v_{i,k}$ represents a summary statistic of the i-th tree in the k-th set. $v_{\mathrm{J}}^{i,k}$ denotes the J One balance index, $v_{\mathrm{G}}^{i,k}$ denotes the Gamma statistic, $v_{\mathrm{P}}^{i,k}$ denotes PD, $v_{\mathrm{M}}^{i,k}$ denotes the mean pairwise distance, $v_{\mathrm{R}}^{i,k}$ denotes the Rogers balance index and $\top$ denotes the transpose of vector. These five statistics were selected based on a clustering dendrogram of various summary statistics, where they were found to be less correlated and more evenly spaced than other statistics (see Appendix C).

The Mahalanobis distance of $v_{i,k}$ is given as

$$d_{\mathrm{M}}(v_{i,k}) = \sqrt{(v_{i,k} - \bar{v}_k)^{\top} S^{-1} (v_{i,k} - \bar{v}_k)} \tag{2.9}$$

where $v_{i,k}$ is the index vector for the i-th tree of the k-th set. $\bar{v}_k$ comprises five mean values of each element respectively in the index vectors in the k-th set, $\top$ denotes the vector transpose, S is the covariance matrix of these vectors and S^{-1} denotes the inverse of S.

For each representative tree (i.e. the tree with the smallest Mahalanobis distance to the other trees), we mapped the historical speciation rates onto the tree using the `eve` package (see Appendix B for details). The variation of speciation rates along tree lineages is represented by a continuous color scale. The actual rate values corresponding to colors were calculated as mean speciation rates (mean value of speciation rates at the beginning and at the end) of the time frames between evolutionary events in a simulation.

LTT plots were generated using the `eve` package for each tree set by summarizing all the trees in the set and then displaying the mean LTT curve along with shading representing the confidence interval of the set. The LTT plots were based on extant species only.

2.3 Results

2.3.1 Simulation Performance

All simulation jobs were completed within seven hours on the computing cluster, with most of the parameter sets finishing within one hour.

When extinction rates are 0 and no species richness effect is present, a deceleration in lineage accumulation is observed in the PD scenario, but only when β_{Φ} is negative, that is,

higher ER in the communities results in lower speciation rate on all lineages (see Figure 2.3 and the animations in Appendix E).

2.3.2 Effects of Evolutionary Relatedness

We observed a hierarchical structuring in tree topology patterns, from PD to ED to NND. Under the PD scenario, trees exhibit a higher degree of balance (interpreted from the J One balance index) than in the ED scenario, which in turn exhibits more balanced trees than the NND scenario. This pattern is more prominent when β_Φ decreases from zero towards more negative values, that is, when ER reduces the speciation rates even more. As β_Φ increases from negative to positive values, that is, the ER effect on the speciation rates transitions from a reduction to augmentation, the differences in tree balance between PD, ED and NND scenarios decrease, and the trees exhibit generally lower degree of balance (see the animations in Appendix A ending with J_One).

As β_Φ increases from -0.04 to 0 (ER effect becomes neutral), a general decrease in the mean pairwise distances of the trees is observed. However, no substantial change occurs when β_Φ shifts from 0 to positive values (i.e., a beneficial effect of ER on speciation). When β_Φ is negative, the trees under the PD scenario exhibit larger mean pairwise distances than those in the ED scenario and the trees under the ED scenario exhibit larger mean pairwise distances than those in the NND scenario. As β_Φ increases from -0.04 to 0, the disparities in mean pairwise distances among PD, ED, and NND scenarios decrease (see the animations in Appendix A ending with MPD and MBL). The patterns of mean branch lengths across β_Φ settings are similar to those of mean pairwise distances.

The distribution of internal nodes, or the concentration of speciation events, were interpreted using the Gamma statistic. When β_Φ is negative, the speciation events in the PD scenario are located more closely to the root than in the ED scenario, and the events in the ED scenario are closer to the root than those in the NND scenario. These differences are reduced as β_Φ increases from negative values towards zero. This pattern changes as β_Φ shifts from zero towards positive values, particularly when β_N is 0. As β_Φ increases from -0.04 to 0.002, the distributions of speciation events in all three scenarios change from being closer to the root to being more evenly distributed throughout the temporal range of the phylogeny (see the animations in Appendix A ending with Gamma).

The trees become notably larger (more species) as β_Φ increases from -0.04 to 0. This increase is minimized when β_Φ increases from 0 to 0.002. When β_Φ is negative, the tree sizes are generally largest in the NND scenario, second largest in the ED scenario and smallest in the PD scenario. This pattern does not persist when β_Φ is zero or positive. The pattern is reversed when β_Φ is positive, in which case the PD scenario generally has the largest tree sizes and the NND scenario the smallest (see the animations in Appendix A ending with SR).

Of all model parameters, β_Φ has the strongest effect on the evenness of the speciation rates across lineages. In all parameter combinations, the disparities of speciation rate evenness between PD, ED and NND scenarios decline with increasing β_Φ (see Figure 2.4B and the animations in Appendix A ending with ERE).

2.3.3 Interaction between the Effects

The effects of β_Φ (for instance the degree of disparities of the statistics between PD, ED and NND scenarios, and the changes of the statistics along with varying β_Φ) appear more prominent when β_N imposes a less pronounced (i.e. less negative) effect on speciation, that is, the SR effect on the speciation rates is less pronounced, with the strongest effects observed at $\beta_N = 0$ which means the SR effect is neutral. From another point of view, the influence of β_N on the tree summary statistics is more prominent when β_Φ shifts towards more negative values (see Figure 2.4A).

When negative SR effect on speciation rate becomes weaker (β_N increases from -0.04 to 0, SR effect on the speciation rates shifts from reduction to neutral), the trees become larger and less balanced, mean branch lengths and mean pairwise distances of the trees become smaller, the distribution of speciation events changes from being closer to the root to being more evenly distributed, and the degrees of disparities of speciation rate evenness among PD, ED and NND increase (see Figure 2.4A for details and see all the animations in Appendix A starting with beta_n for animated comparisons). Generally, stronger negative SR effect reduces the differences of tree statistics among PD, ED and NND scenarios, it also reduces the differences of the statistics between different β_Φ settings. This is particularly true for the Gamma statistic and J One balance index.

The above described patterns, however, do not apply to speciation rate evenness. When the negative effect of SR on speciation rate becomes stronger (β_N decreases from 0 to -0.04), the disparities of speciation rate evenness among PD, ED and NND scenarios appear more prominent when β_Φ value becomes more negative.

2.3.4 Effects of Intrinsic Speciation and Extinction Rates

Higher intrinsic speciation rates (λ_0) lead to less balanced phylogenetic trees and more disparities among the scenarios. Higher intrinsic extinction rates (μ_0), however, lead to more balanced trees and less disparities. The effects of λ_0 and μ_0 are intertwined with both β_Φ and β_N. The effects are described in detail in Appendix F. Besides Figure 2.4, we also produced a set of animated plots (Appendix A), showing how tree summary statistics we measured change along with β_Φ, β_N, λ_0 and μ_0, to help interpret the patterns resulting from different levels of evolutionary relatedness effects (as determined by β_Φ) and between the three different versions of ER dependence.

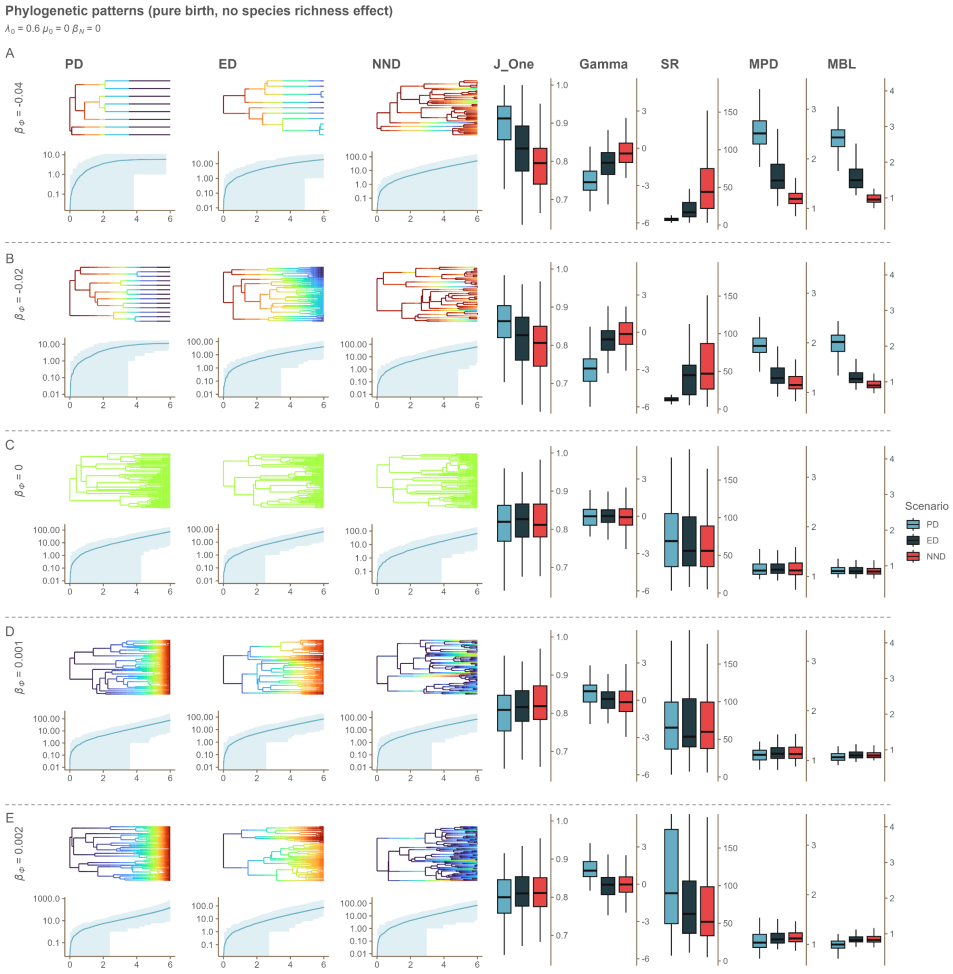


Figure 2.3: Overview of results of simulations of a pure birth process (no extinction, see Appendix E for other speciation and extinction rate combinations), with dependence of speciation on evolutionary relatedness. For all parameter sets shown in the figure, the intrinsic speciation rate at the start of the simulation (λ_0) is 0.6, the intrinsic extinction rate (μ_0) is 0, and the coefficient of the species richness effect (β_N) is 0. The results are grouped by rows according to varying levels of β_Φ (the coefficient of the evolutionary relatedness effect). On the left, representative simulated phylogenies and associated lineage-through time (LTT) plots are shown for dependence of speciation rates on three metrics of ER (left - phylogenetic diversity (PD), middle - evolutionary distinctiveness (ED), and right - nearest neighbor distance (NND) scenario). The colors mapped onto the trees represent the values of the speciation rate for each of the lineages during simulation time. The rates increase from blue to red. The trees colored in green are cases where the speciation rate remains unchanged throughout the simulation. Blue line in the LTT plots represents lineage accumulation through time and the shaded area is the 95% confidence interval. On the right of each row, boxplots for the corresponding summary statistics are shown: J_ONE - J One balance index; Gamma - Gamma statistic; SR - total number of extant lineages; MPD - mean pairwise distance; MBL - the mean branch length.

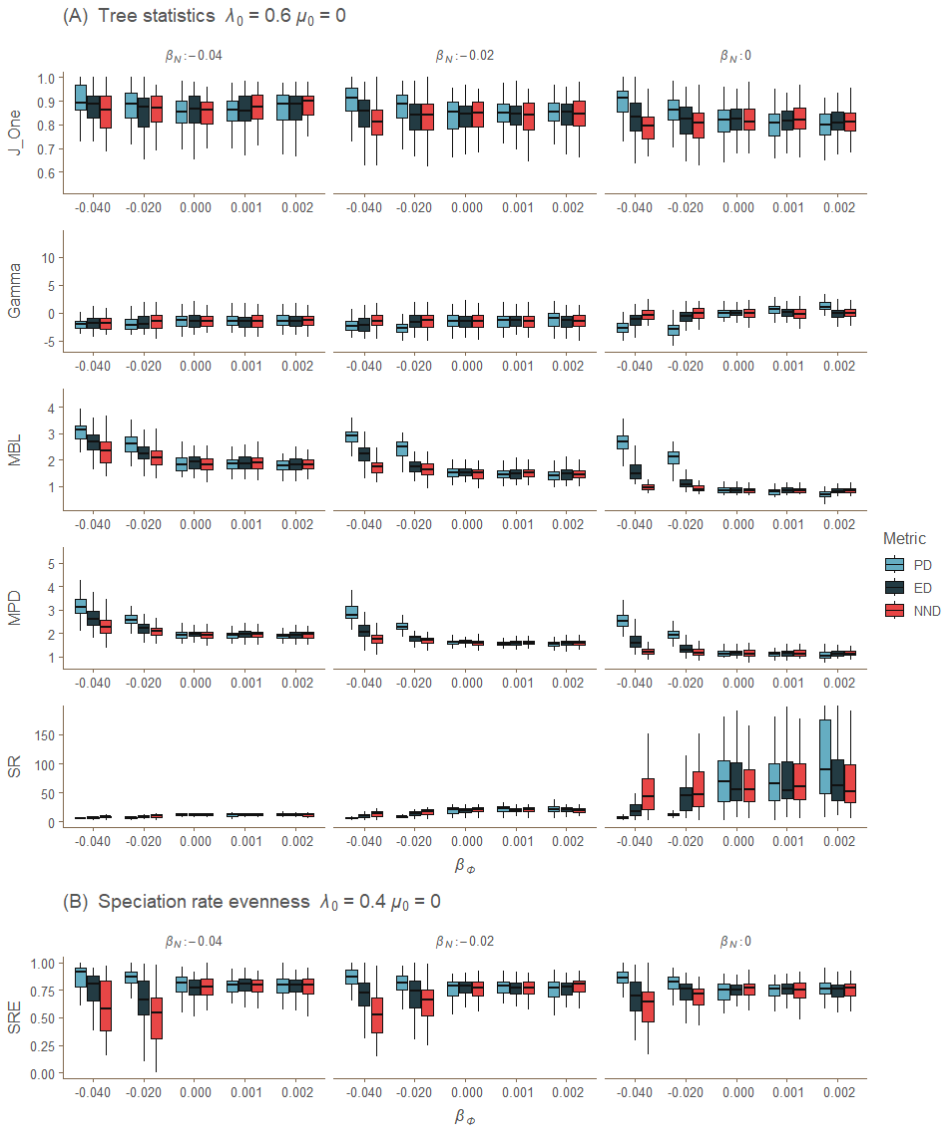


Figure 2.4: Tree summary statistics for the three scenarios of dependence of speciation rates (dependence on phylogenetic diversity (PD), evolutionary distinctiveness (ED) and nearest neighbor distance dependence (NND)), for various levels of the evolutionary relatedness effect (β_Φ) and the species richness effect (β_N). λ_0 - initial speciation rate; μ_0 - fixed extinction rate; x-axis: strength of ER effects; y-axis: value of the statistics. The following statistics are shown in panel A: J_ONE - J One balance index; Gamma - Gamma statistic; MBL - the mean branch length; MPD - mean pairwise distance; SR - total number of extant lineages. The folloing statistic is shown in panel B: SRE - speciation rate evenness. SRE is shown in a separated panel because it has different patterns and is in a different parameter setting. We pick SRE from this parameter setting to better present the disparities among different scenarios and across different levels of the ER effect.

2.4 Discussion

2.4.1 From Clade-Wide to Lineage-Specific

Evidence for ecological limits on diversity has been found or suggested in many studies [187, 188], and ecological limits may exert influence on diversification by directly impacting rates of speciation [93]. Our simulations further suggested that, from clade-wide to lineage-specific, different extents of ecological limits result in unique phylogenetic tree properties, when using ER as a proxy.

When phylogenetic diversity acts as a proxy regulating speciation rate of all lineages (PD scenario), a negative ER effect (or β_Φ, the corresponding parameter in our simulations) can be interpreted as a niche space limitation on clade expansion as phylogenetic diversity increases, which applies equally to all species in the clade. Under this interpretation, an increase in phylogenetic diversity within a clade leads to saturation of niches. Conversely, a positive ER effect provides the clade with the potential to exploit additional resources as phylogenetic diversity increases, potentially resulting in a rapid accumulation of species as the potential for biotic interactions increases.

By relating (clade-wide, PD scenario) speciation rate to ER, the PD scenario exhibits similarities to models with diversity carrying capacity (e.g. diversity-dependent diversification model, DDD), with negative ER effect (negative β_Φ) on the speciation rates imposing an upper limit of phylogenetic diversity carrying capacity. The role of negative SR effect (negative β_N) in our model is similar to the concept of carrying capacity in the DDD model, negative ER and SR effects in our simulations both markedly reduce average tree sizes, given a constant simulation time.

In the ED and NND scenarios, a negative ER effect indicates a scenario where more distantly related species are less likely to speciate, which may reflect speciation through an adaptive dynamics scenario where competition for similar resources leads to trait divergence to escape a fitness value [189, 190]. It may also reflect environmental filtering [191], where closely related species matching the environment are more likely to speciate. In contrast, a positive ER effect indicates that more distantly related species are more likely to produce descendants, for example, due to species adapting to different environmental conditions and thereby exploring new niche space [192].

By relating lineage-specific speciation rates to ER, the ED and NND scenarios can account for evolutionary trajectories in which niches are either conserved (more similar than expected), constrained (diverging within a restricted range of available niches), or divergent (less similar than expected), based on different settings of ER effects (see the illustration of niche conservatism by Pyron et al. [193]). Negative ER effect also trims the average tree sizes in the ED scenario, but this effect seems diminished or even absent in the NND scenario. Negative effects of ER decrease from PD scenario to ED scenario to NND scenario; along this "gradient" of metrics, ecological limits exert increasingly smaller influence on overall species richness of the communities, due to their impacts being more concentrated to close relatives (see also the speciation rate transitioning in Figure 2.3).

The unevenness of speciation rates among the tips increases from PD to ED and NND (note that although in PD scenario all tips of a tree have the same speciation rate, PD trees

from the same parameter setting could still have different unevenness due to different tree topologies). When the ER effect is negative, the more distinct lineages are less likely to spawn new lineages, leading to a "clustering" where closer lineages have closer speciation rates. As the effect of ER goes from PD to ED to NND, large "clusters" break apart into small and separate clades on the tree, thereby increasing unevenness. When we shift the ER effect to positive, more distinct lineages are more prone to speciate, which may cause an "overdispersion" of the rates rather than a "clustering" (Figure 2.4B). In the ED and NND scenarios, the speciation rates among the tips are directly determined by their distances to the clade (ED) or their immediate neighbors (NND), so the observed unevenness could also be interpreted as an unevenness of ER.

In general, we observed hierarchical patterns in various summary statistics (e.g. J One balance index, Gamma statistic and speciation rate evenness) from clade-wide to lineage-specific ecological limits (scenarios, i.e. PD, ED, NND) when ER effect is negative. These patterns fade as ER effect shifts from negative to neutral ($\beta_\Phi = 0$). We did not symmetrically explore the positive ER effects due to computational limitations as explained above.

We observed prominent difference of tree sizes between scenarios, with PD trees being the smallest, ED trees being noticeably larger and NND trees being the largest on average. Tree size often shows strong correlation with phylogenetic tree statistics, even after applying size correction methods, regardless of the underlying model [60]. This may explain why β_Φ impacts the balance and mean pairwise distances of PD trees, even though speciation rates across lineages are consistently uniform. To verify if tree size confounds our interpretation, we compared all the statistics from our results with the tree size correlation curves by Janzen and Etienne [60]. We found that the correlations between statistics and tree size in our results are opposite to patterns observed by Janzen and Etienne [60] as tree size increases from 10 to 100. This implies that our findings are unlikely to be due to an effect of tree size. Rather, if the effect of tree size would be corrected for, we would expect our patterns to be even more pronounced.

2.4.2 Species Richness Reduces Evolutionary Relatedness Signature

Debates in ecology persist regarding the efficacy of phylogenetic metrics serving as proxies for species richness or functional diversity [173, 194]. Our model introduces a more flexible scenario wherein species richness and phylogenetic (or evolutionary) relatedness operate concurrently. Their impacts on species diversification can be independently positive, neutral, or negative. We observed a stronger impact from ER as SR effect becomes neutral (β_N shifts from negative values to 0) on speciation rate in communities, with the most pronounced ER impact occurring when the SR effect is neutral ($\beta_N = 0$), where SR has no influence on species diversification. Furthermore, we noted that as SR imposes a more negative effect, the disparities in various tree statistics (e.g., J One balance index, mean branch length, mean pairwise distance, and Gamma statistic) among PD, ED, and NND scenarios become less evident or even disappear. Based on these patterns, we suggest that when speciation rate is limited by species richness, the signature of evolutionary relatedness is concealed. However, evolutionary relatedness can still play a complementary role in explaining macroevolutionary patterns if the impact of species richness is minor and the impact of evolutionary relatedness is substantial.

2.4.3 Diverse Evolutionary Trajectories Cause Tree Imbalance

Imbalanced phylogenies are frequently observed in empirical research, and their occurrence has been attributed to various factors, including errors in phylogenetic data, incomplete species sampling, and biases introduced by reconstruction methods; additionally, such imbalance may reflect variations in evolutionary rates within trees [195, 196]. Simulations based on earlier stochastic models often show discrepancies when compared to empirical evidence, highlighting a potential misalignment between model predictions and actual observations [196], although there is significant variation in the degree of imbalance between empirical clades [60].

Our model offers more diverse evolutionary trajectories including lineage-specific scenarios that allow each lineage within a tree to have distinct speciation rates. Our analysis reveals notable differences in speciation rate variations within trees across different scenarios (see Figure 2.4B, the differences in speciation rate evenness), from PD through ED to NND. This hierarchical pattern reflects a shift in the impact of evolutionary relatedness–becoming increasingly concentrated among closer relatives–and a corresponding increase in speciation rate variation among lineages. Phylogenies simulated under our model can be more balanced or less balanced compared to a simple birth–death scenario. For example, in the NND scenario where trees often exhibit large speciation rate variation when ER imposes negative effect, stronger negative ER effect results in higher speciation rate variation and the corresponding trees exhibit greater imbalance. See Appendix D for a comparison of phylogenetic imbalance between empirical trees and trees simulated under our model.

2.4.4 Extinction Process and Empirical Application

Although macroevolutionary dynamics are often studied through dependence of speciation rates on (phylogenetic) diversity metrics, the speed at which species go extinct may also be linked to ecological factors. Competition may play a non-negligible role in increasing the extinction risk of species, but its effects can vary depending on the context. For example, empirical evidence from Bengtsson [197] on rock pool zooplankton demonstrates that interspecific competition increases local extinction rates; Dangremond et al. [198] found that the proximity of an invasive grass increased seed predation on an endangered lupine species and accelerated its decline; Timmermann [199] reviewed the extinction of Neanderthals and highlighted how competitive pressures, coupled with resource exploitation efficiency, played a significant role in Neanderthals' demise. Overall, competition for finite ecological resources can accelerate species extinction, particularly for species that are more vulnerable to antagonistic ecological interactions.

If ER-dependent extinction were incorporated in our simulation, we expect that under the PD scenario, negative ER effects should lead to larger trees due to decreasing community-wide extinction rates, similar to the positive feedback loop of ER-dependent speciation. Positive ER effects on extinction would likely result in smaller trees. For the ED scenario, negative ER effects would mean distinct species are less likely to go extinct, making trees less balanced, while positive ER effects would promote the extinction of distinct species, leading to an auto-balancing process. For the NND scenario, ER effects are more local, which may result in different effects than for PD or ED.

Enabling both ER-dependent speciation and extinction, with all possible positive/negative

combinations, could lead to even more diverse evolutionary trajectories. However, without data, drawing concrete implications would be challenging. We stress that increased extinction can eliminate some of the information contained in phylogenies, which would further burden our interpretation from phylogenetic trees.

In some real-world scenarios, it is possible that ER effects are so strong that SR is not dominating, which could explain the significant imbalance observed in empirical phylogenies, even when SR effects might otherwise obscure this pattern. However, fitting complex models to empirical phylogenies, such as through maximum-likelihood estimation or approximate Bayesian computation, presents mathematical and technical challenges [123, 134]. Alternative methods such as neural networks can directly infer parameters from empirical phylogenies, but as models grow more complex, it becomes more difficult to recover parameters accurately, regardless of the method used [200]. Nonetheless, it should still be possible to identify which ER scenarios generate evolutionary patterns that are closest to empirical data through neural network classification tasks. Such analyses may enable us to re-evaluate how ecological factors shape biodiversity.

2.5 Appendix

A) Animation of Main Simulation

Animated plots were generated to aid the explanation of model behaviors and phylogenetic patterns of the main simulation. The plots were named in a format of "parameter_statistic.gif" where parameter refers to the changing parameter in the animation and statistic refers to the summary statistic the plot shows. For example, "beta_n_Gamma.gif" is an animation showing the Gamma statistics among different parameter settings and transitioning in β_N.

The full set of animation files is available at:

https://github.com/EvoLandEco/Thesis-Appendix/

B) Rate-Mapping Algorithm for Visualization

Algorithm: Map lineage histories to colored phylogeny segments

Require: Rooted phylogeny $\mathcal{T}$ with branch lengths; discrete-time history $H(t, \ell)$
Ensure: Segment table S with rows $(x, y) \longrightarrow (x', y')$ and value v

Compute phylogram coordinates (x_u, y_u) for all nodes/tips
$S \longleftarrow \emptyset$
procedure TRAVERSE(u)
  for all child edges $u \longrightarrow c$ with length L **do**
    Choose representative descendant tip $r \in$ subtree(c)
    $H_r \longleftarrow H(\cdot, r)$ restricted to $[x_u,\ x_u + L]$ (snap to grid if needed)
    for $i = 1$ to $|H_r| - 1$ **do**
      $t_0 \longleftarrow H_r.time[i]$, $t_1 \longleftarrow H_r.time[i+1]$
      $v \longleftarrow \frac{H_r.val[i] + H_r.val[i+1]}{2}$
      Append $(t_0, y_c) \longrightarrow (t_1, y_c)$ with value v to S
    end for
    Append $(x_u, y_u) \longrightarrow (x_u, y_c)$ with value $H_r.val[1]$ to S
    if c is internal **then**
      TRAVERSE(c)
    end if
  end for
end procedure
TRAVERSE(root($\mathcal{T}$))
return S

C) Heatmaps

Heatmaps were plotted to visualize the correlation matrix of a number of tree statistics under different ER scenarios. A representative tree was selected based on the heatmaps, using the statistics that are less correlated and more evenly spaced on the heatmap. There are three heatmaps, each presenting one scenario.

(Figures on next page.)

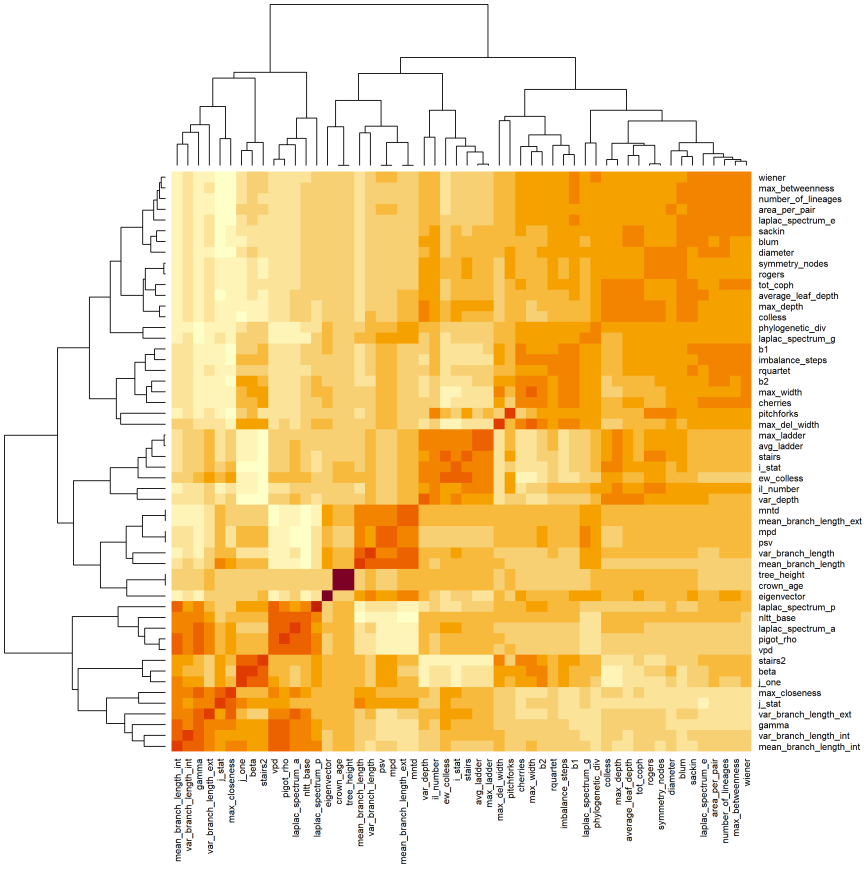


Figure 2.5: Correlation heatmap of tree statistics for phylogenies simulated under the PD scenario of the eve model. Each cell shows the pairwise Pearson correlation between two summary statistics; rows and columns are ordered by hierarchical clustering to highlight groups of strongly correlated metrics.

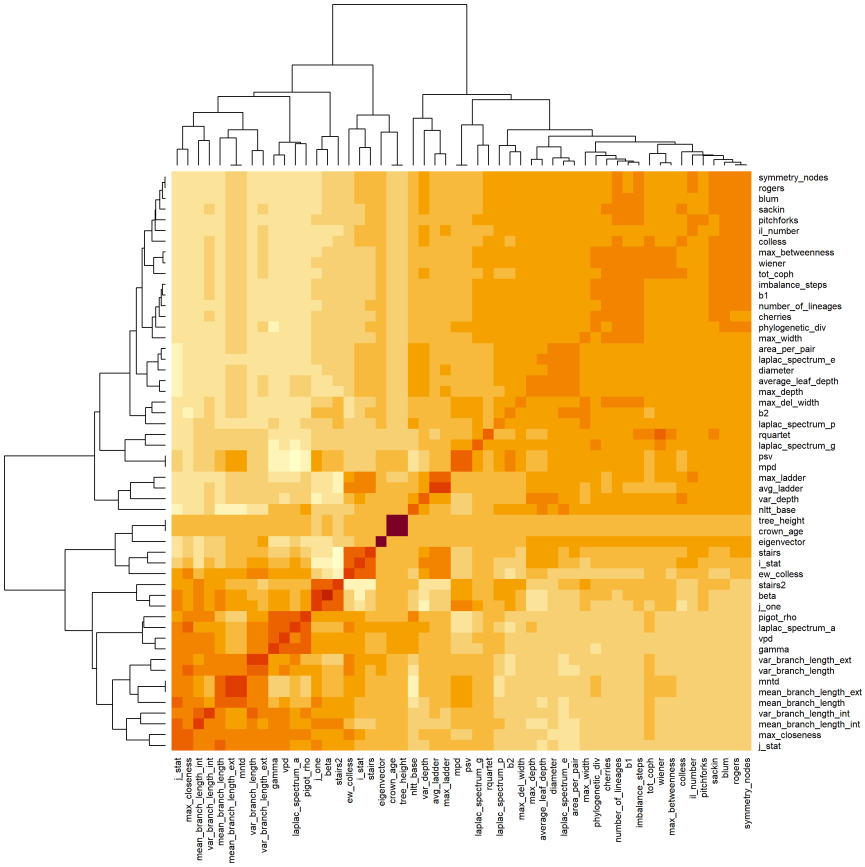


Figure 2.6: Correlation heatmap of tree statistics for phylogenies simulated under the ED scenario of the eve model. Each cell shows the pairwise Pearson correlation between two summary statistics; rows and columns are ordered by hierarchical clustering to highlight groups of strongly correlated metrics.

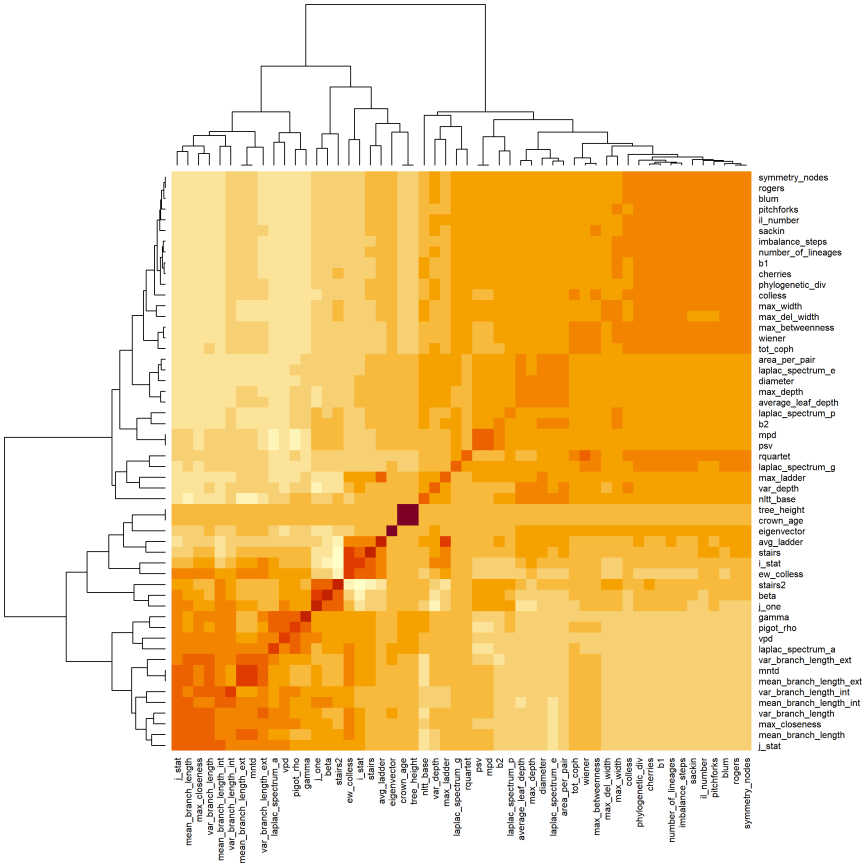


Figure 2.7: Correlation heatmap of tree statistics for phylogenies simulated under the NND scenario of the eve model. Each cell shows the pairwise Pearson correlation between two summary statistics; rows and columns are ordered by hierarchical clustering to highlight groups of strongly correlated metrics.

D) Tree Imbalance

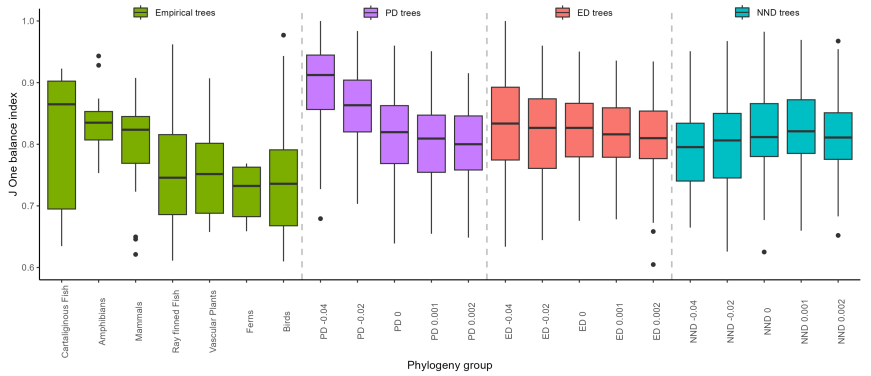


Figure 2.8: Comparison of phylogenetic tree imbalance between empirical clades and clades simulated using our model. The y-axis shows the J One balance index, with lower values indicating higher degree of imbalance. The leftmost group of boxes shows the tree imbalance computed from empirical trees obtained from Janzen and Etienne [60]. The x-axis indicates the taxonomic group of the empirical trees; there are seven empirical sub-clades. In the remaining groups of boxes, the x-axis indicates the ER scenario (PD, ED, or NND) and the effect size of ER (β_Φ) used for simulating the trees. The remaining parameters are fixed: speciation rate $\lambda_0 = 0.6$, extinction rate $\mu_0 = 0$ and effect size of SR $\beta_N = 0$. The plots show that under certain parameter values, we can generate trees with higher levels of imbalance, bringing the simulated phylogenies closer to empirical phylogenies. For example, with a net diversification rate ($\lambda_0 - \mu_0$) = 0.6, $\beta_\Phi > 0.001$, and $\beta_N = 0$, we can generate phylogenies that are more imbalanced than those of mammals and amphibians under the PD and ED scenarios. Further increasing the speciation rate and β_Φ could yield even more imbalanced trees. Under the NND scenario we estimate that $\beta_\Phi < -0.01$ can produce phylogenies close to empirical trees. We observed substantial variation in imbalance between empirical clades, with some clades, such as ferns and birds, much more imbalanced, while others, such as cartilaginous fish, more balanced. Given the smaller sample size and potential biases in empirical phylogenies, these findings should be interpreted with caution.

E) Tree Statistics and LTT Plots with Representative Trees

Tree statistics and LTT plots with representative trees were plotted for all the parameter settings of the main simulation. The plots were named in a format of "λ_0_μ_0_β_N.png". For example, "0.6_0_0.png" presents the results under the parameter setting $\lambda_0 = 0.6$, $\mu_0 = 0$ and $\beta_N = 0$.

The full set of figures is available at:

https://github.com/EvoLandEco/Thesis-Appendix/

F) Effects of Intrinsic Speciation and Extinction Rates

We found that the intrinsic speciation rate (λ_0) also showed a substantial impact on tree statistics. The effects of this parameter are intertwined with both β_Φ and β_N (see animations in Appendix A starting with lambda and mu).

As λ_0 increases, the differences in tree balance across scenarios increase, particularly when β_Φ has smaller positive effect or the effect is more negative. The trees become generally less balanced as λ_0 increases. The effects of λ_0 are stronger when β_Φ is more negative and β_N is closer to zero, with the most profound case being $\beta_\Phi = -0.04$ and $\beta_N = 0$ where increasing λ_0 results in most evident increase of disparities of balance between PD, ED and NND scenarios (see the animations in Appendix A starting with lambda and ending with J_One).

In contrast to λ_0, an increase in the intrinsic extinction rate (μ_0) results in more balanced trees and less disparities of balance among the PD, ED and NND scenarios. The reduction of disparities is more prominent when β_Φ becomes more negative. The most prominent case is when $\beta_\Phi = -0.04$ and $\beta_N = 0$ where increasing μ_0 results in most evident reduction of disparities of balance between PD, ED and NND scenarios (see the animations in Appendix A ending with J_One) (see the animations in Appendix A starting with mu and ending with J_One).

An increase in λ_0 from 0.4 to 0.6 generally leads to a reduction in mean branch length and mean pairwise distance, while an increase in μ_0 from 0 to 0.2 generally leads to an increase in these statistics (see the animations in Appendix A starting with lambda and mu, ending with MPD and MBL).

The disparities of speciation events distributed across the phylogeny among PD, ED and NND scenarios, as measured by the gamma statistic, become more prominent as λ_0 increases and less prominent as μ_0 increases. The distribution of speciation events changes from being closer to the root to being more evenly distributed as μ_0 increases (see the animations in Appendix A starting with lambda and mu, ending with Gamma).

When β_Φ effect is negative, increasing λ_0 has the largest effect of increasing the tree sizes in the NND scenario. This effect is smaller in the ED scenario and even smaller in the PD scenario. Increasing μ_0 has the largest effect of reducing the tree sizes in the NND scenario; this effect is smaller in the ED scenario and even smaller in the PD scenario. When β_Φ effect is zero or positive, these trends are diminished (see the animations in Appendix A starting with lambda and mu, ending with SR).

Increasing λ_0 only has the effect of reducing the disparities of speciation rate evenness among PD, ED and NND when β_Φ is negative and β_N is 0. μ_0 has no obvious effect on the evenness of the speciation rates (see Figure 2.4B and the animations in Appendix A starting with lambda and mu, ending with ERE).

G) Correlation Matrix in Ultrametric Phylogenetic Trees

This section presents a proof of the relationship between the correlation matrix and the phylogenetic distance matrix for any ultrametric phylogeny under the Brownian motion model of trait evolution [53, 201]. Specifically, we prove that the correlation matrix C is given by

$$C = \frac{2t - R}{2t} \tag{2.10}$$

where t is the crown age of the phylogeny, and R is the phylogenetic distance matrix.

Definition 1 (Phylogenetic Tree). *A phylogenetic tree is denoted as $\mathcal{T} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ is the set of nodes, including the root, internal nodes, and tips (leaves) representing taxa, and $\mathcal{E}$ is the set of edges (branches) connecting the nodes. Each edge $e \in \mathcal{E}$ has an associated branch length $l_e > 0$.*

Definition 2 (Ultrametric Tree). *An ultrametric tree is a rooted phylogenetic tree in which all tips (leaves) are equidistant from the root. This implies that the total path length from the root to any tip is the same for all tips.*

Definition 3 (Brownian Motion Model). *In the Brownian motion model of trait evolution, the variance of trait change along a branch of length l_e is $\sigma^2 l_e$, where σ^2 is the evolutionary rate. Trait changes along different branches are independent unless they share common ancestry.*

Definition 4 (Variance of a Trait). *For each tip $i \in \mathcal{L}$, let P_i denote the set of edges (branches) on the path from the root to tip i. The variance of a neutral trait X_i at tip i is defined as*

$$V_{ii} = \mathrm{Var}(X_i) = \sigma^2 \sum_{e \in P_i} l_e. \tag{2.11}$$

Definition 5 (Covariance Between Traits). *The covariance between traits at tips i and j is defined as*

$$V_{ij} = \mathrm{Cov}(X_i, X_j) = \sigma^2 \sum_{e \in P_i \cap P_j} l_e, \tag{2.12}$$

where $P_i \cap P_j$ is the set of shared edges on the paths from the root to tips i and j.

Definition 6 (Correlation Coefficient). *The correlation coefficient between tips i and j is defined as*

$$C_{ij} = \frac{V_{ij}}{\sqrt{V_{ii} V_{jj}}}. \tag{2.13}$$

Definition 7 (Phylogenetic Distance). *The phylogenetic distance R_{ij} between tips i and j is defined as*

$$R_{ij} = \sum_{e \in P_i} l_e + \sum_{e \in P_j} l_e - 2 \sum_{e \in P_i \cap P_j} l_e. \tag{2.14}$$

By virtue of Equation 2.11 this simplifies to:

$$R_{ij} = \frac{V_{ii} + V_{jj} - 2V_{ij}}{\sigma^2}. \tag{2.15}$$

Theorem. *For any ultrametric phylogeny under the Brownian motion model, the correlation matrix C is given by:*

$$C = \frac{2t - R}{2t}, \tag{2.16}$$

where t is the crown age of the phylogeny, and R is the phylogenetic distance matrix.

Proof. We begin by rearranging Equation 2.15 as

$$V_{ij} = \frac{V_{ii} + V_{jj} - \sigma^2 R_{ij}}{2}. \tag{2.17}$$

Substituting Equation 2.17 into the correlation coefficient formula in Equation 2.13 leads to:

$$C_{ij} = \frac{V_{ii} + V_{jj} - \sigma^2 R_{ij}}{2\sqrt{V_{ii}V_{jj}}}. \tag{2.18}$$

Because the tree is ultrametric, the total path length from the root to any tip i equals the crown age t. Hence, for any tip i and j, recall from Equation 2.11, we get:

$$V_{ii} = V_{jj} = \sigma^2 \sum_{e \in P_i} l_e = \sigma^2 t. \tag{2.19}$$

Substituting $V_{ii} = V_{jj} = \sigma^2 t$ into Equation 2.18 leads to

$$C_{ij} = \frac{2\sigma^2 t - \sigma^2 R_{ij}}{2\sigma^2 t}. \tag{2.20}$$

Then we factor out σ^2:

$$C_{ij} = \frac{2t - R_{ij}}{2t}. \tag{2.21}$$

Because C_{ij} denotes all the elements in the correlation matrix C, and R_{ij} denotes all the elements in the phylogenetic distance matrix R, Equation 2.21 implies:

$$\boxed{C = \frac{2t - R}{2t}}. \tag{2.22}$$

This completes the proof. □

H) Simplification of the Speciation Rate Evenness Computation

In this section we continue with the simplification of Equation 2.4. For binary trees, the number of tips $n \geq 2$. If $\mathrm{diag}(R)$ represents the vector comprising the diagonal elements of R, by definition of R, all elements in $\mathrm{diag}(R)$ are equal to 0, thus all elements in $\mathrm{diag}(C)$ are equal to 1. Then we have

$$\mathrm{diag}(C)^\top m = \lambda. \tag{2.23}$$

Recall that

$$\bar{\lambda}_i = \frac{\lambda}{n}. \tag{2.24}$$

So Equation 2.4 can be rewritten as

$$E = \frac{\lambda^2 - m^\top C m}{\lambda^2 - \frac{\lambda^2}{n}} \tag{2.25}$$

Factoring out λ^2 leads to

$$E = \frac{n}{n-1}\left(1 - \frac{m^\top C m}{\lambda^2}\right) \tag{2.26}$$

Defining $m' = m/\lambda$, we get

$$\boxed{E = \frac{n}{n-1}\left(1 - m'^\top C m'\right)} \tag{2.27}$$

3

Neural Network Estimation of Diversification Parameters

Species diversification is characterized by speciation and extinction, the rates of which can, under some assumptions, be estimated from time-calibrated phylogenies. However, maximum likelihood estimation methods (MLE) for inferring rates are limited to simpler models and can show bias, particularly in small phylogenies. Likelihood-free methods to estimate parameters of diversification models using deep learning have started to emerge, but how robust neural network methods are at handling the intricate nature of phylogenetic data remains an open question. Here we present a new ensemble neural network approach to estimate diversification parameters from phylogenetic trees that leverages different classes of neural networks (dense neural network, graph neural network, and long short-term memory recurrent network) and simultaneously learns from graph representations of phylogenies, their branching times and their summary statistics. Our best-performing ensemble neural network (which adjusts the graph neural network result using a recurrent neural network) delivers estimates faster than MLE and shows less sensitivity to tree size for constant-rate and diversity-dependent speciation scenarios. It performs well compared to an existing convolutional network approach. However, like MLE, our approach still fails to recover parameters precisely under a protracted birth–death process. Our analysis suggests that the primary limitation to accurate parameter estimation is the amount of information contained within a phylogeny, as indicated by its size and the strength of effects shaping it. In cases where MLE is unavailable, our neural network method provides a promising alternative for estimating phylogenetic tree parameters. If detectable phylogenetic signals are present, our approach delivers results that are comparable to MLE but without inherent biases.

Qin, T., van Benthem, K., Valente, L.†, & Etienne, R.† (2025). Parameter estimation from phylogenetic trees using neural networks and ensemble learning. *Systematic Biology* [147]. † indicates joint senior authors.

3.1 Introduction

Identifying the underlying mechanisms shaping biodiversity is an important goal in the fields of evolutionary biology and ecology. Species diversification processes can often be characterized by speciation and extinction rates, which can be estimated from time-calibrated phylogenies [79] as long as the assumed model structure of diversification resembles the true underlying data generation process [87]. Time-calibrated phylogenies contain branching times and evolutionary relationships (captured in the topology) between species and offer a complementary source of information to the often incomplete fossil record [117]. The increasing availability of reconstructed phylogenies has empowered many studies seeking explanations for the underlying diversity patterns using modeling approaches [119, 121, 202]. One type of models — birth–death models — are often used to estimate speciation, extinction and diversification rates from reconstructed phylogenetic trees [78, 79, 118, 203].

Likelihood-based approaches, such as maximum likelihood estimation (MLE) and Bayesian inference, can be used to infer not only speciation and extinction rates, but also possibly existing evolutionary and ecological signals, such as diversity-dependence or trait-dependence of rates, from branching times and other information sources [82, 112, 114, 123, 124]. However, Etienne et al. [121] showed that MLE estimates of the clade-level carrying capacity in diversity-dependent diversification models tend to be close to (but evidently higher than) the number of species observed in the phylogeny, which also biases the other parameters.

An alternative to these likelihood-based approaches for parameter estimation is Approximate Bayesian Computation (ABC), which approximates the posterior distribution of parameters without requiring explicit calculation of a likelihood function. ABC is often seen as a good substitute to MLE when a likelihood function of a model is not available, as long as simulations of the model are fast and tractable [125, 129, 130]. However, studies using ABC for parameter estimation in phylogenetics remain scarce [131–134]. This is partly due to the fact that it is often difficult to identify adequate summary statistics and potential distance metrics in ABC, which makes the application and development of this potentially powerful approach challenging.

A promising class of tools that may help overcome the limitations of likelihood-based methods and ABC are machine learning approaches, such as neural networks. Neural networks are comprised of layers of nodes, or ”neurons”, which process input data and learn to recognize patterns between input and output data from training data [135]. Classic feed-forward neural networks have achieved good results in tasks such as image recognition and natural language processing [136]. Another class of neural networks, graph neural networks, are designed specifically for graph-structured data, such as social networks, molecular structures, and ecological interaction networks. They can capture the dependencies and relationships inherent in data types that can be naturally represented as graphs [139] and have shown strong performance in various tasks involving graph representation learning [140–142]. Phylogenetic trees can also be viewed as graphs, suggesting that graph neural networks have potential applicability in phylogenetics. Recurrent neural networks, another type of neural network, are designed to handle sequential data, such as time series, by maintaining a memory of previous inputs [137, 138]. Recurrent neural networks can

process inputs of varying lengths and capture time-dependent features, making them particularly well-suited for tasks where the order of data points is crucial, such as learning parameters from branching times when viewed as time-series data.

Owing to the rapid development of both hardware capability and deep learning algorithms, applications of neural networks in phylogenetic analyses have started to emerge [126, 143–146, 204–208]. For instance, phylogenetic deep learning approaches have been shown to provide reliable estimates of parameters in epidemiological, birth–death, and trait-dependent speciation models [126, 143, 144, 209]. Despite their potential, employing neural networks for estimating parameters based on the whole phylogenetic tree, especially those associated with diversification, poses significant challenges and requires further systematic research regarding their performance, accuracy and robustness. Specifically, feed-forward linear neural networks usually require a large amount of data to be able to generalize well on the patterns within the data [136]; producing graph representations for graph-level learning can be challenging given the need to aggregate information across diverse graph sizes and topologies [140]; the ability of the recurrent neural networks to predict parameters from whole sequences is often challenging [137]. Hence, how robust neural network methods are at handling the intricate nature of phylogenetic data remains an open question.

In this study, we explore the potential of neural networks in research on species diversification using phylogenies. We first develop various neural network architectures and protocols for transforming phylogenetic trees and branching times into compatible formats. In addition to new neural networks presented here, for comparison with existing methods, we also investigate a one-dimensional convolutional neural network architecture described by Lajaaiti et al. [143], Voznica et al. [144]. We then investigate predictive performance of underlying parameters on simulated data for ensemble learning strategies. These strategies combine different neural network classes to maximize data utilization and enhance performance. We also assess the determinants of estimation accuracy and robustness for both neural network and MLE methods under various diversification scenarios. Finally, we implement our trained neural networks on empirical phylogenetic datasets and compare their estimations to those of MLE.

Our analyses encompass three different diversification scenarios for which likelihood-based inference approaches already exist: a constant-rate birth–death (BD) scenario, with constant speciation and extinction rates over time [120]; a diversity-dependent diversification (DDD) scenario, where the number of species in a clade negatively affects the speciation rate [82]; and a protracted birth–death (PBD) scenario, where speciation takes time and does not always proceed to completion [90]. Applying our new methodology to phylogenetic trees simulated under a broad range of the parameter space, our findings indicate that neural network approaches are as effective, if not more so, than MLE in recovering parameters from phylogenetic data simulated under these stochastic processes. Trained neural networks can be conveniently applied to empirical trees for parameter estimation. To facilitate this, we present a new R package, `EvoNN`, capable of performing such analyses based on phylogenetic trees (empirical or simulated) supplied by the user [210].

3.2 Methods

3.2.1 Software Environment and Computational Budget

We used a hybrid programming environment with `Python 3.7.1`, `CUDA 12.2.2` [211], `PyTorch 1.12.1` [212], `PyTorch Geometric 2.3.1` [213], and `R 4.2.1` [214]. Simulations, data transformation, and maximum likelihood estimation were handled through parallel CPU computations on the Hábrók high-performance computing cluster of the University of Groningen. The total computational budget for these processes was approximately 3000 hours (used CPU time). Our neural networks were trained, optimized and evaluated on the NVIDIA A100 and V100 tensor core GPUs of the Hábrók cluster. The estimated computational budget was 1500 hours (used GPU time, excluding CPU time for dataset loading and saving). We implemented a user-friendly illustrative tool to estimate parameters from phylogenetic trees using pretrained neural networks from this study in the new R package `EvoNN`.

3.2.2 Simulation Approaches

To train the neural networks, we simulated phylogenetic trees using different functions from different R packages. For each simulated dataset we kept trees with only extant lineages, mimicking reconstructed phylogenies. The settings for the parameters used to simulate the trees were selected to limit the maximum total number of nodes (including root, internal and tip nodes, here and after, we always refer to the total number of nodes) for the trees in each dataset. After simulation, we further filtered out all trees containing more than 3000 nodes to avoid the creation of excessively large matrices that could deplete the available memory space allocated to the GPUs during the GNN training process. Such trees are uncommon under the settings we used – typically fewer than 5 trees with more than 3000 nodes (1500 tips) are present within each set of phylogenies we acquired from simulation. We also filtered out all trees containing less than 5 nodes (3 tips) to ensure successful data transformation and summary statistic computation. Small trees inherently carry limited informational content. The exclusion of these trees is unlikely to impact performance of the neural networks on the remaining trees (typically fewer than 100 out of 100,000 trees with less than 5 nodes were removed for each parameter setting).

To consider different diversification processes, we simulated 100,000 random birth–death trees (BD phylogenies), 100,000 diversity-dependent trees (DDD phylogenies) and 100,000 protracted birth–death trees (PBD phylogenies). The amount of simulated data is bounded by the resource and time limits of the computing cluster. All trees have an identical crown age of 10 time units ($t = 10$) to reduce both the complexity of the data and the computational burden. For simulating BD trees, we used the `rlineage` function from R package `ape` [215] to generate complete trees and then pruned all the extinct lineages; for DDD trees, we used the `dd_sim` function from R package `DDD` [82]; for PBD trees we used the `pbd_sim` function from our R package `eveGNN` (a codebase of phylogeny simulation, data transformation, neural network training and MLE computation for our study), which is similar to the function with the same name in the original R package `PBD` [89], but only outputs necessary data for our study.

In our simulation approach we randomly sampled the (log) parameters required for each

scenario (BD, DDD and PBD) from uniform distributions. The upper bound for the extinction rates were proportionally dependent on the drawn speciation rate to avoid cases where extinction rates could be larger than speciation rates, because in such cases the whole tree likely goes extinct. Furthermore, to prevent a huge number of events, which would deplete available computational time and memory, we also imposed an overall cap of 1.5 on the extinction rates.

Our choice of speciation and extinction rate ranges in the DDD scenario was informed by both computational considerations and biological context. For instance, speciation rates in birds have been estimated to range from 0.1 to 1 events per lineage per million years, while extinction rates often fall between 0 and 0.5 events per lineage per million years [216, 217].

See Table 3.1 for the detailed parameter distribution settings used in the simulations. Note that the Gillespie algorithm is scale-invariant, and therefore the absolute rate magnitudes are interchangeable with the length of the simulated time interval.

Table 3.1: Parameter settings for the simulated tree datasets. The type column specifies which function is used to generate the trees. The columns specify the crown age (age), the number of trees in the data set (N), the lower (a) and the upper (b) bounds of the parameters for the tree simulations, all the parameters being sampled from $U(a, b)$, except for λ_1 of the protracted birth–death scenario. λ_1 is computed as $\lambda_1 = 10^i$ where i is sampled from $U(-3, 1)$. U denotes uniform distribution. Sub-table A shows the parameter distributions of the constant-rate birth–death model and the diversity-dependent-diversification model, λ: intrinsic speciation rate/birth rate; μ: intrinsic extinction rate/death rate; K: clade-level carrying capacity. Sub-table B shows the parameter distributions of the protracted birth–death model, λ_1: speciation-initiation rate of good species; λ_2: speciation-completion rate; λ_3: speciation-initiation rate of incipient species; μ_1: extinction rate of good species; μ_2: extinction rate of incipient species. *In diversity-dependent-diversification simulations, the maximum extinction rate is capped at 1.5 if $0.9\lambda > 1.5$.

A: Parameter settings for BD and DDD trees

Type	**Age**	**N**	λ_0		μ_0		K	
			a	b	a	b	a	b
BD	10	100,000	0.1	0.8	0.0	$0.9\lambda_0$	-	-
DDD	10	100,000	0.1	4.0	0.0	$0.9\lambda_0$*	10	1000

B: Parameter settings for PBD trees

Type	**Age**	**N**	λ_1		$\log_{10}(\lambda_2)$		λ_3		μ_1		μ_2	
			a	b	a	b	a	b	a	b	a	b
PBD	10	100,000	0.1	1.0	-3	1	0.1	1.0	0.0	$0.8\lambda_1$	0.0	$0.8\lambda_3$

3.2.3 Data Preparation

We employed three different basic neural network architectures: a dense neural network (DNN), a graph neural network (GNN), and a long short-term memory (LSTM) recurrent network, as illustrated in Figure 3.1 (see Appendix E for a detailed description). Each of these architectures was refined through validation and required different input data. For the DNN, the input data consisted of a total of 54 summary statistics (Appendix Q) for each simulated tree. In the GNN, the full phylogeny was interpreted as a graph and

could in that form be used as input data (as illustrated in Figure 3.2). In the LSTM, we treated branching times of the phylogenies as sequential or time-series data [137]. Given its recurrent architecture, LSTM is adept at sequence prediction tasks, making it particularly suitable for estimating tree parameters from entire sequences of branching times.

Therefore, our data comprises three major components: the phylogenetic trees, their corresponding summary statistics, and their branching times, to maximize the use of available data. The functions needed for the data transformations are either available in PyTorch or implemented in our package `eveGNN` and described in more detail in Appendix A, Appendix B, and Appendix C.

3.2.4 Ensemble Learning Strategies

To leverage all available data and improve prediction accuracy, we combined GNN, DNN, and LSTM using bagging, stacking, and boosting, which are typical ensemble learning strategies [218]. With bagging, we trained GNN, DNN and LSTM independently on the same dataset, translated their original outputs to parameter predictions (we will use "readout" hereafter to refer to this translation) and then aggregated the predictions. We used four aggregation methods: mean, median, max and min.

With stacking, we use GNN, DNN and LSTM in the same architecture but without their own readout layers. Instead, we combined the features learned from DNN, LSTM and GNN and fed them to a meta-learner comprising linear neural network layers that learns the best readout parameter predictions from these combined features. GNN, DNN, LSTM and the meta-learner were trained simultaneously.

With boosting, the neural networks were trained sequentially. Boosting strategies offer various pathways for enhancing model performance. We started with a GNN to make initial predictions and explored the effectiveness of both DNN and LSTM for correcting residuals, either individually or in sequence. We used "Boost SS" to refer to correcting GNN's residuals by DNN (from summary statistics, thus "SS"); "Boost BT" to refer to correcting GNN's residuals by LSTM (from branching times, thus "BT"); "Boost SS+BT" to refer to correcting GNN's residuals by DNN and then correcting DNN's residuals of residuals by LSTM; "Boost BT" to refer to correcting GNN's residuals by LSTM (from branching times); "Boost BT+SS" to refer to correcting GNN's residuals by LSTM and then correcting LSTM's residuals of residuals by DNN.

See Figure 3.3 for a simplified illustration of the ensemble learning strategies.

3.2.5 Training Neural Networks

Prior to training, each dataset of 100,000 trees was randomly shuffled and subsequently divided into two segments. The first segment, consisting of 90% of the dataset, was allocated for training purposes, while the remaining 10% was used as the validation dataset for monitoring and fine-tuning the neural network performance. The training session is carried out by epochs, each consisting of three major steps: first, performing forward pass on the training dataset; second, assessing the prediction accuracy; and lastly, performing back-propagation (adjusting the weights of the neuron connections to improve the neural network performance). Back-propagation, requires quantifying the error between the

neural networks' predictions and the actual ground truth values. We quantified the 'total loss' as the sum of the residual error and other terms for facilitating neural network training. We represented total loss using a loss function that sums up all the loss terms (see Appendix D for more detail).

We use the `AdamW` (Adaptive Moment Estimation with decoupled weight decay) optimizer [219] to iteratively update the neural networks' parameters to minimize the loss function. We used default `AdamW` argument settings. During training, we adopted mini-batches of size 64 (data of 64 simulated trees per mini-batch) to reduce GPU memory usage. The total number of epochs was manually optimized per architecture to avoid underfitting and overfitting. This was done by comparing the loss metrics for the training dataset to those of the validation datasets at every epoch. Overfitting is indicated by a training loss that continues to decrease while the validation loss starts to increase, whereas underfitting is suggested by both training and validation losses being high and decreasing at a similar rate. Analyzing these loss trends over time can help to optimize hyper-parameters ("settings" that might alter neural network behavior or impact performance). Early in the study, we used early stopping to identify the optimal training duration. Once we observed that our simulated datasets converged at highly consistent time points (because the simulated datasets are homogeneous between simulations and splits), we switched to manually setting training lengths–guided by diagnostic metrics–to achieve the best balance between training and testing performance.

Under the BD scenario, the neural networks were trained to predict two parameters: birth rate (λ) and death rate (μ). Under the DDD scenario, the neural networks were trained to predict three parameters: speciation rate (λ), extinction rate (μ) and carrying capacity (K). Under the PBD scenarios, the neural networks were trained to predict five parameters: speciation rate of the good species (λ_1), speciation completion rate (λ_2), speciation rate of the incipient species (λ_3), extinction rate of the good species (μ_1) and extinction rate of the incipient species (μ_2).

To enable a direct comparison between our networks and an existing method, we additionally implemented an one-dimensional convolutional neural network architecture (CNN1D here and after) described by Lajaaiti et al. [143], Voznica et al. [144] to estimate parameters from branching times.

3.2.6 MLE as Baseline Benchmark

Maximum likelihood estimation (MLE) approaches have been developed for the BD, DDD and PBD scenarios [82, 89, 123]. Per scenario, we simulated additional testing datasets each comprising 10,000 phylogenies using the same parameter spaces as the training datasets. For these testing datasets, we adopted the MLE approaches to estimate the parameters of each phylogeny from their branching times under different scenarios. For the BD trees, we estimated their birth and death rates. For the DDD trees, we estimated their speciation rate, extinction rate and clade-level carrying capacity. There is a limitation in the MLE approach for PBD trees because to allow for a computation of the likelihood, the speciation initiation rates of good species and incipient species need to be equal [123]. We therefore only estimated four initial parameters: speciation initiation rate (for both good and incipient species, assuming they are the same), speciation-completion rate, extinction rate of good

species and extinction rate of incipient species, although in our simulation we have five independently sampled parameters. The MLE prediction accuracies are used as a baseline benchmark to evaluate the prediction accuracies of the neural networks on tree parameter estimation.

We implemented two different MLEs: one with ground truth parameter values set as the starting point of the MLE searching process, the other with a starting point randomly sampled from the parameter space of the simulation. We consider the first type of benchmarks as a best-case MLE performance (as in real applications ground truth parameters are not known) and the other type as a naive-case MLE performance, which mimics the pragmatic approach if true parameter values are not known. Note that in practice it is possible to achieve much better performance than the naive-case, e.g. by optimizing from several random starting points to avoid being stuck in local optima.

We also explored the effectiveness of different optimization and integration approaches for MLE on the DDD phylogenies. We used the "Simplex" [220] optimizer. See Appendix G for reasons and for a detailed comparison between the optimizers.

3.2.7 Performance Analysis

We used the same testing dataset as was used for MLE parameter estimation, to obtain neural network parameter predictions. To evaluate the performance of each method (MLE benchmarks and different neural network architectures), we analysed the patterns of residuals (differences between ground truth and predicted values, which can be viewed as the goodness of fit) by examining their relation to true values and the total node counts of the phylogenetic trees, which include root, internal and tip nodes. Considering the complex nature of residual patterns, which may vary according to specific characteristics of the simulation processes (for instance, carrying capacity effects in DDD and protracted speciation in PBD), as well as the performance and robustness of the estimation methods, we calculated error metrics locally for three different phylogeny size ranges, as a global metric could be misleading.

Our observation is mainly based on simulated trees under the DDD scenario, because it involves more evolutionary mechanisms than the simple birth–death scenario while containing fewer parameters than the protracted birth–death scenario. This simplifies our analyses on the neural network performance while maintaining enough complexity to challenge the capability of our proposed methods. From this case study we identified and selected the most effective MLE optimization algorithm, neural network architecture and ensemble strategy, which we then applied to BD and PBD scenarios. We therefore only analysed the best-performing neural network methods against the naive and best MLE cases on BD and PBD. Additionally, for the BD scenario, we computed net diversification rate ($\lambda - \mu$) and extinction-to-speciation ratio (μ/λ); for the PBD scenario, we computed a composite parameter called the mean duration of speciation from the speciation completion rate, the speciation rate of incipient species and the extinction rate of incipient species, because MLE can arguably better estimate the mean duration of speciation than the original parameters [123].

3.2.8 Robustness Analysis

We assessed the robustness of the estimation results of both neural networks and MLE by measuring the consistency with which these approaches produce similar estimates for phylogenies generated under identical parameter settings. We also applied this bootstrapping method to assess the estimation robustness of empirical trees; the details are provided in the following section. In the previous simulations, each parameter combination was sampled and used only once, whereas in the robustness analysis we repeatedly use identical parameter settings to generate sets of phylogenies (bootstrapping) under the DDD scenario. Even when the same parameters are used, the resulting phylogenies can vary substantially in size, topology and structure due to stochasticity. Such an evaluation helps assess the neural networks' ability to abstract the underlying parameter influences from the phylogenetic data, regardless of heterogeneity. For each parameter combination, 1000 trees were generated randomly. We used a total of 80 sets of parameter combinations, thus 80,000 phylogenies in total. Specifically, we used all combinations of speciation rates $\lambda = 1.0, 1.5, 2.0, 2.5, 3.0$, extinction rates $\mu = 0.2, 0.4, 0.6, 0.8$ and carrying capacities $K = 200, 400, 600, 800$.

MLE is computationally more expensive than predicting from already trained neural networks, and computational time rapidly increases with the size of the phylogenies. We thus performed MLE on only 2000 simulated phylogenies. To ensure fair visual and numerical comparisons when plotting the results of these analyses, extreme MLE estimates were not shown in the figures (they exceeded the fixed range of the y-axis) and excluded from the computation of the mean absolute errors of the MLE estimates. Neural network results were randomly sub-sampled to match the MLE data count, maintaining equivalent visual density and facilitating a more accurate performance comparison between approaches. For the neural networks, the mean absolute errors were computed on the complete dataset without sub-sampling and exclusion. In Figure 3.4, Figure 3.5, Figure 3.6 and Figure 3.7, on average (we simulated the testing datasets many times throughout the study), out of 2000 samples, 5-150 samples per MLE figure panel, and 1-3 per neural network figure panel fell beyond the axis range. For the most underperforming method (Boost BT+SS) 600-1500 samples fell beyond the axis range.

We did not analyze the robustness of BD and PBD scenarios, because BD is a special case of DDD (if we set carrying capacity to an infinite value) and PBD related parameters can hardly be estimated accurately using MLE methods [123].

The complete code base for this study, including simulations, data processing, neural network training, evaluation, and both data analysis and visualization tools, is available in the GitHub repository `eveGNN` [221].

3.2.9 Misspecification Analysis

To further assess the performances of MLE and neural network approaches when confronted with misspecified assumptions, we treated simulated birth–death (BD) trees as if they were produced under a diversity-dependent diversification (DDD) model. Specifically, we applied MLE algorithm designed for the DDD scenario and used two neural networks–GNN and Boost BT–trained on DDD trees to estimate the speciation rate λ, extinction rate μ and clade-

level carrying capacity K. Under a true BD process viewed through the DDD perspective, $K \longrightarrow \infty$, so the value of K is infinite.

To compare naive and best-case performances of MLE under estimation method misspecification, we set the starting points of the optimizer by two strategies: one in which the initial K was set to 10,000 (naive-case) and another in which K was set to ∞ (best-case). In both cases, the starting points of λ and μ were set to the true parameters used to simulate the BD trees. For each case, we summarized the proportions of MLE estimates falling into four intervals relative to the total number of nodes N_{nodes}: (1) between 2 N_{nodes} and 5 N_{nodes}, (2) between 5 N_{nodes} and 20 N_{nodes}, (3) between 20 N_{nodes} and ∞, and (4) exactly ∞ (by exactly we mean the MLE algorithm gives infinite K estimate).

We computed the expected N_{nodes} under the BD process as a function of the net diversification rate and simulation time t,

$$\mathbb{E}[N_{\text{nodes}}(t)] \;=\; 2\,\mathbb{E}[N_{\text{tips}}(t)] - 1 \;=\; 2\,e^{(\lambda-\mu)t} \;-\; 1 \tag{3.1}$$

where $t = 10$ for all trees. This expectation guided the interpretation of estimated K magnitudes relative to the total number of nodes of the trees being estimated.

3.2.10 Empirical Tree Estimation

We deployed pre-trained neural networks to estimate phylogenetic parameters from a dataset of 199 empirical phylogenetic trees curated by Condamine et al. [222], with a tip count ranging from 20 to 1500. To align with the training conditions of our neural networks, which were trained on simulated phylogenies spanning exactly 10 time units (Myr), we rescaled the crown ages of all empirical trees to this duration. The parameter estimations we present are therefore rescaled. All the selected empirical trees are reconstructed phylogenies and fully bifurcated (each root or internal node has exactly two descendants). If an empirical tree fails an ultrametric (all tip-ends are aligned at the present) test due to branch length precision issue, we forced all its tips to end exactly at the present by extending the shorter tips to align with the longest one. See Appendix P for meta information of the empirical trees.

We used two distinct neural networks, each pre-trained on simulated trees from one of two evolutionary scenarios (BD or DDD) to estimate parameters from the empirical trees. For the BD scenario, we estimated the parameters λ (speciation rate) and μ (extinction rate); for the DDD scenario, we estimated λ, μ, and K (carrying capacity). We did not estimate parameters for the PBD scenarios because neither neural networks, nor MLE approaches could recover the individual parameters accurately from the simulated phylogenies. In addition to our neural network estimates, we used MLE methods for parameter estimation to provide a comparative assessment of the results. The MLE methods were set to use default starting points of likelihood optimization, as we do not know the true parameters of the empirical phylogenies.

We used the same bootstrapping method described before to quantify the uncertainties of both MLE and neural network estimates from empirical data. The process involves three main steps: first, estimating parameters from empirical phylogenies using MLE and pre-trained neural networks; second, simulating a set of phylogenies under a specified

diversification scenario (such as BD, DDD, or PBD) using the MLE and neural network estimates; and third, re-estimating parameters from the simulated phylogenies using MLE and neural networks. The estimates derived from the bootstrapped phylogenies form a distribution.

We applied this uncertainty computation to a selected set of empirical phylogenies from the Condamine dataset [222] under the DDD scenario (see Appendix H for details). The criteria for selection were phylogenies with more than 300 and less than 1000 nodes, and maximum likelihood estimates (MLE) of K (carrying capacity) being less than 1000. The distributions of MLE and neural network estimates from the bootstrapped phylogenies was compared to the original MLE and neural network estimates from empirical phylogenies. For each set of parameters estimated from empirical phylogenies, we bootstrapped 1000 simulated phylogenies.

Our R package `EvoNN` [210] provides functions to perform the uncertainty (bootstrap) analyses.

3.2.11 Supplementary Studies

To account for the potential effects of under- and over-representation of phylogenies of different sizes in our datasets, we conducted a supplementary study to explore whether the patterns we observed persist in a dataset with a re-balanced distribution of phylogeny sizes, see Appendix M, Figure 3.33) for details.

To explore the generalization ability of the neural networks when facing data with true parameters completely outside the training space, as well as to compare neural network performances between extant-only phylogenies and complete phylogenies with extinct species, we also conducted another supplementary study. MLE and neural network approaches were examined on in-distribution and out-of-distribution datasets crossed with extant and complete trees. See Appendix N for details.

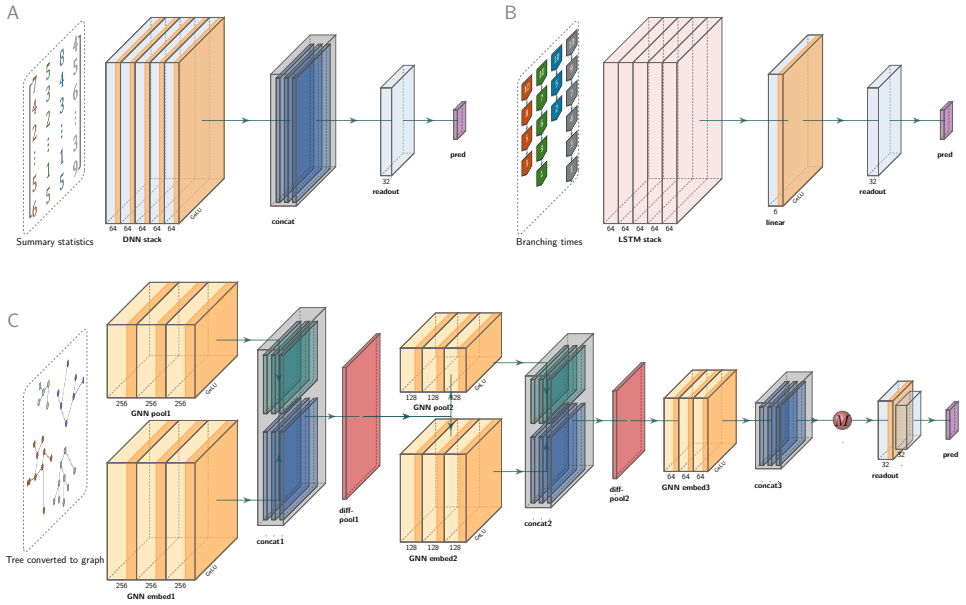


Figure 3.1: Illustration of the neural network architectures. From left to right, for each neural network, the inputs are filtered through the layers, and the network ultimately outputs the final predictions of the parameters through the readout layers. A: dense neural network (DNN), whose input data are summary statistics. The major component of the DNN is a stack comprising five linear layers (”DNN stack”), each followed by a Batch Normalization for 1D Inputs operator (BatchNorm1D, not shown in figure) and Gaussian Error Linear Units (GELU, the orange band within the boxes). Learned features from all the linear layers within the stacks are collected and concatenated (”concat”). A single linear readout layer (”readout”) outputs n predicted parameters (”pred”). B: long short-term memory recurrent neural network (LSTM), whose input data are the branching times. The major component of the LSTM is a stack of five LSTM recurrent neural network layers (”LSTM stack”). Learned features are processed by a linear layer accompanied by a GELU (”linear”), then passed to a single linear readout layer (”readout”) that outputs n predicted parameters (”pred”). C: graph neural network (GNN), whose input data is a graph representation of the phylogeny. GNN is assembled from five modules. Each module comprises the same number of GraphSAGE (sample-and-aggregate graph convolutional neural network) operators. Each operator is accompanied by a BatchNorm1d (not shown in the figure) operator and then a GELU activation function (illustrated by the orange bands within the yellow boxes). Learned features from all the GraphSAGE operators within a module are collected and concatenated. The differentiable pooling (DiffPool) technique is adopted to perform graph coarsening. In the first coarsening operation, the graph data inputs are passed to two GNN modules (”GNN pool1” and ”GNN embed1”). The pooling group reduces the graph size, while the embedding group captures the node features. The filtered data from each GraphSAGE operator are concatenated (”concat1”) and then passed to a DiffPool layer (”diff-pool1”), which finalizes the first coarsening operation. The second coarsening operation is applied in the same way as the first (as represented by ”GNN pool1”, ”GNN embed2”, ”concat2”), and the outputs from the second DiffPool layer (”diff-pool2”) are passed to the final (fifth) GNN module (”GNN embed3”). After the final GNN module, the outputs are concatenated (”concat3”) and transformed by a global mean pooling operation (red ball ”M”) to create a final graph representation. This graph representation is passed to a readout layer group (”readout” as represented by light blue boxes) consisting of two linear layers to perform graph-level regression which ultimately outputs a vector of n predicted parameters (”pred” as represented by a purple box). Only the first linear layer is followed by GELU (see the orange band of the first linear layer). See Appendix E for the detailed description and technical details.

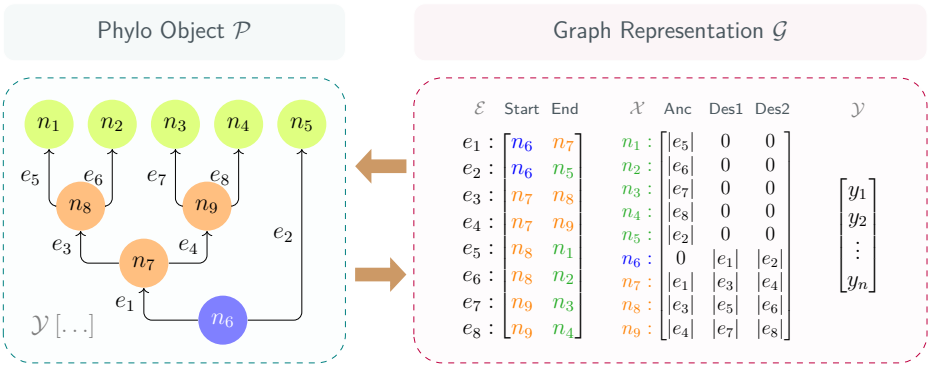
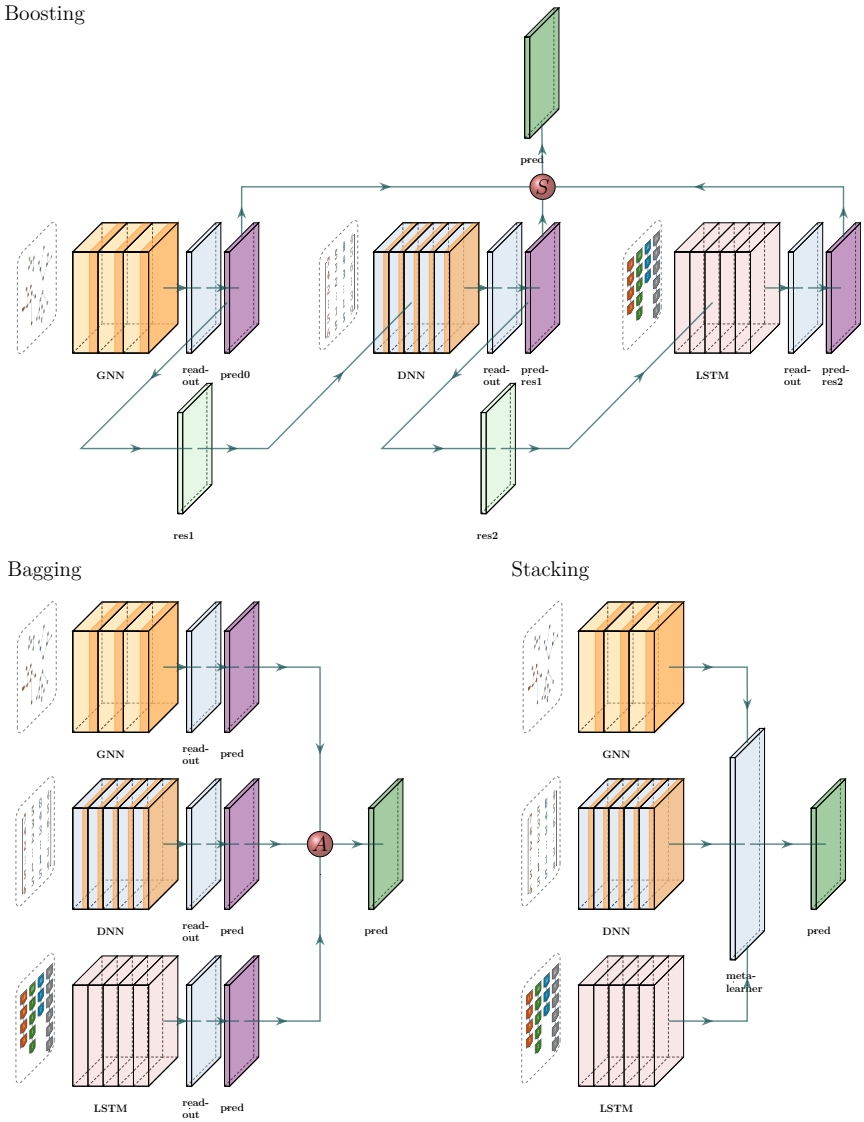


Figure 3.2: Illustration of data transformation between a "phylo" object and its graph representation. The left panel shows a visualization of a "phylo" object. The blue circle represents the root node, orange circles represent the internal nodes and green circles represent the tip nodes. Arrows represent directed edges between each pair of the nodes. The right panel shows the transformed graph data structure. The adjacency list is denoted as $\mathcal{E}$. Each row of the adjacency list represents one edge, the first column represents the starting node and the second column represents the end node. Note that the adjacency list is transposed (in the example into $\mathcal{E}^{2\times 8}$) after converting to a tensor. The node feature matrix is denoted as $\mathcal{X}$. Each row of the node feature matrix represents the features contained in one node, the first column represents the distance from the node to its direct ancestor node, the second and the third columns represent the distances from the node to its two descendants. In the node feature matrix, the distances from a node to non-existing nodes (e.g. the tip nodes have no descendants, and the root node has no ancestor) are represented by zeros. The node and edge labels before the colons (including the colons) are placed here for visual assistance. After transformation, we use graph-level attributes $\mathcal{Y}$ to store the parameters used to generate the "phylo" object. The node labels are given by $n_1, n_2, n_3, \ldots, n_9$, the edge labels are given by $e_1, e_2, e_3, \ldots, e_8$, the edge lengths are given by $|e_1|, |e_2|, |e_3|, \ldots, |e_8|$. The generating parameters are given by a vector $[y_1, y_2, \ldots, y_n]$ where n is the number of parameters



(Caption on next page.)

Figure 3.3: *(Figure on previous page.)* Illustration of the ensemble learning strategies to combine the graph neural network (GNN), the dense neural network (DNN) and the long short-term memory recurrent neural network (LSTM). The neural networks are largely simplified. With boosting, GNN, DNN and LSTM were trained sequentially to iteratively correct residuals. For example, DNN is trained to predict the residuals of GNN predictions. Subsequently, LSTM is trained to predict the residuals of the residuals after DNN corrected the GNN predictions. The final prediction comes from the initial prediction by GNN minus two learned residual terms by DNN and LSTM. With bagging, we trained GNN, DNN and LSTM independently, translated their original outputs to parameter predictions and then aggregated the predictions. With stacking, we trained GNN, DNN and LSTM simultaneously but without readout. We directly concatenated the outputs from GNN, DNN and LSTM and then used a meta-learner to make predictions from the outputs. With bagging, we trained GNN, DNN and LSTM independently ("GNN", "DNN" and "LSTM" blocks of boxes), translated their outputs to parameter predictions through their own readout layers (three "readout" boxes next to the neural networks and three "pred" boxes next to the readout layers) and then aggregated the predictions (red ball "A"). With stacking, we trained GNN, DNN and LSTM simultaneously ("GNN", "DNN" and "LSTM" blocks of boxes) but without their own readout layers. We combined the features from DNN, LSTM and GNN and fed to a meta-learner ("meta-learner") comprising linear neural network layers to output parameter predictions. With boosting, there can be different pathways. In our illustration, GNN, DNN and LSTM were trained sequentially to iteratively correct residuals. First, the GNN is trained from the graphs to make the initial predictions (see "GNN", "readout" and then "pred0") and from predicted and ground truth values of the parameters we computed the residuals ("res1"); second, the DNN is trained to predict these residuals from the summary statistics (see "DNN", "readout" and then "pred-res1"), learning to correct the GNN's errors; lastly, the LSTM is trained to predict the residuals of the residuals (see "LSTM", "readout" and then "pred-res2"), which is the initial predictions minus the predicted residuals by the DNN, from branching times, to further improve the predictive accuracy. Finally, we subtracted the two residual terms from the initial predictions (red ball "S") to make the corrected predictions. See Appendix F for a detailed explanation.

3.3 Results

3.3.1 Performance Analysis

We evaluated the performances of various neural networks, both individually and in combination through ensemble strategies, in predicting parameters from simulated DDD phylogenies. These predictions were compared against best-case and naive-case MLE results using the Simplex optimizer.

Among all the methods, we consider boosting GNN with LSTM as the most robust method based on the goodness of fit (see Figure 3.4, Figure 3.5, Figure 3.6 and Figure 3.7), the mean absolute errors (see Appendix I, Figure 3.16), and robustness (see rows named Boost BT in Figure 3.8 and Figure 3.17). Both neural networks and MLE approaches generally struggle with small phylogenies (see Figure 3.4, Figure 3.5 and Figure 3.6 with larger errors for the small phylogenies represented by the yellow points, see also Appendix I, Figure 3.16). Performance improves significantly on medium and large phylogenies for both neural network and MLE approaches.

The MLE implementation sometimes fails to find an optimal solution, partly due to numerical overflow issues. In our visualizations, failed MLE estimations are indicated by small squares spreading along the x-axis to avoid misinterpretation. MLE tended to give small or near-zero estimates, particularly for the carrying capacity. This phenomenon is more prominent when starting optimization from a random point. For all figures showing the MLE error, the ideal situation is that all the data points lie near the horizontal black two-dash reference lines (at which the error is 0) and do not spread along or near the purple dotted reference lines (which suggests near-zero MLE estimations). See the last panel at the bottom right for explanation; then also refer to the left two panels in the last row (the MLE results) in Figure 3.4, Figure 3.5 and Figure 3.6 with the help of the explanation panel. Some MLE estimates are removed from the visualizations due to disagreements of results between different MLE settings, see Appendix G for details. Therefore, the MLE performances shown are both visually and statistically better than that in the complete results, particularly for the best case, as the disagreements occur mainly when MLE struggles to recover parameters accurately.

Neural networks often return values closer to the parameter space's mid-points, a result of making "safer" predictions that minimize loss compared to random guesses. Note that the y - axis in Figure 3.4, Figure 3.5, Figure 3.6 and Figure 3.7 represents estimation error, which is why the red dashed lines indicating the mid-points cross $y = 0$ at $x = (lowerbound + upperbound)/2$. For example, for speciation rate λ in Figure 3.4, we observe $y = 0$ when $x = (0.1 + 4.0)/2 = 2.05$). Consequently, neural networks usually overestimate at low true values and underestimate at high true values (see Figure 3.4, Figure 3.5 and Figure 3.6). These errors are mitigated or partially corrected when the neural networks are trained in tandem through boosting strategies, e.g. boosting GNN results with DNN or LSTM or both (see the panels of Boost SS, Boost BT and Boost SS+BT in Figure 3.4, Figure 3.5 and Figure 3.6). This happens particularly for large phylogenies (the blue data points in Appendix I, Figure 3.16) when the underlying true carrying capacity (K) is large, or for small phylogenies (the yellow data points) when the underlying true speciation rate (λ) is small.

However, boosting strategies can introduce their own challenges. When boosting GNN results first with LSTM and then with DNN, the DNN failed to identify a general pattern of errors from LSTM results. This led to overfitting on the training dataset at the second epoch of the training session (the total loss in the validation dataset started to increase and became much larger than the total loss in the training dataset), which, in turn, resulted in poor performance on the testing dataset (see the panels named Boost BT+SS in Figure 3.4, Figure 3.5, Figure 3.6 and Appendix I, Figure 3.16).

Upon further analysis of the residuals, we observed that inaccuracies in the predictions were largely influenced by the size of the phylogeny (Figure 3.4, Figure 3.5 and Figure 3.6). For neural network approaches, the prediction errors for speciation rate, extinction rate, and carrying capacity tended to increase as the size of the phylogeny decreased, especially in phylogenies with fewer than 200 nodes. Systematic error was also identified in the estimation of carrying capacity: neural networks generally overestimated this parameter in smaller phylogenies and underestimated it in larger ones. Boosting strategies were effective in mitigating or partially correcting systematic errors, and enhancing prediction accuracy, particularly for carrying capacity (see the rows of Boost SS, Boost BT and Boost SS+BT in Appendix I, Figure 3.16).

We calculated the strength of the carrying capacity effect using the formula $1/K' = (\lambda - \mu)/K$, where λ represents the true speciation rate, μ the true extinction rate, K the true carrying capacity, and K' the diversity at which speciation becomes zero for linear negative diversity-dependence [82]. A larger $(\lambda - \mu)/K$ value corresponds to smaller K' and therefore a stronger carrying capacity effect. In the case of smaller phylogenies, neural networks tended to underestimate speciation and extinction rates while overestimating carrying capacity when the carrying capacity effect is weak, and the reverse is observed when the effect is strong. In contrast, MLE tends to overestimate speciation and extinction rates while underestimating carrying capacity under conditions of weak carrying capacity effect, with the reverse occurring under strong effects, except for the carrying capacity which is always underestimated (see Appendix I, Figure 3.16). Neural network methods tend to underestimate the carrying capacity effect. This phenomenon can be mitigated by the boosting strategies, especially the Boost BT method, which achieved similar performance to the best-case MLE estimates (see Figure 3.7).

Unlike GNN and LSTM, DNN cannot by itself reliably recover speciation and extinction rates from the summary statistics of the phylogenies, with its predictions mostly clustering around the mid-points of the parameter space (around the red dashed lines in the DNN panels in Figure 3.4, Figure 3.5 and Figure 3.6). The overall accuracy of the carrying capacities recovered by the DNN is also lower than the accuracy recovered by the other approaches (see the row named DNN in Appendix I, Figure 3.16).

Among all ensemble learning strategies, boosting consistently outperformed both bagging and stacking in enhancing prediction accuracy compared to using neural networks independently, as can be seen, for instance, by the lower mean absolute prediction errors in Appendix I, Figure 3.16. Boosting strategies also exhibited better performance in recovering the true values of the carrying capacity effect (see Figure 3.7). The most effective neural network approaches overall matched or even surpassed the results of MLE while

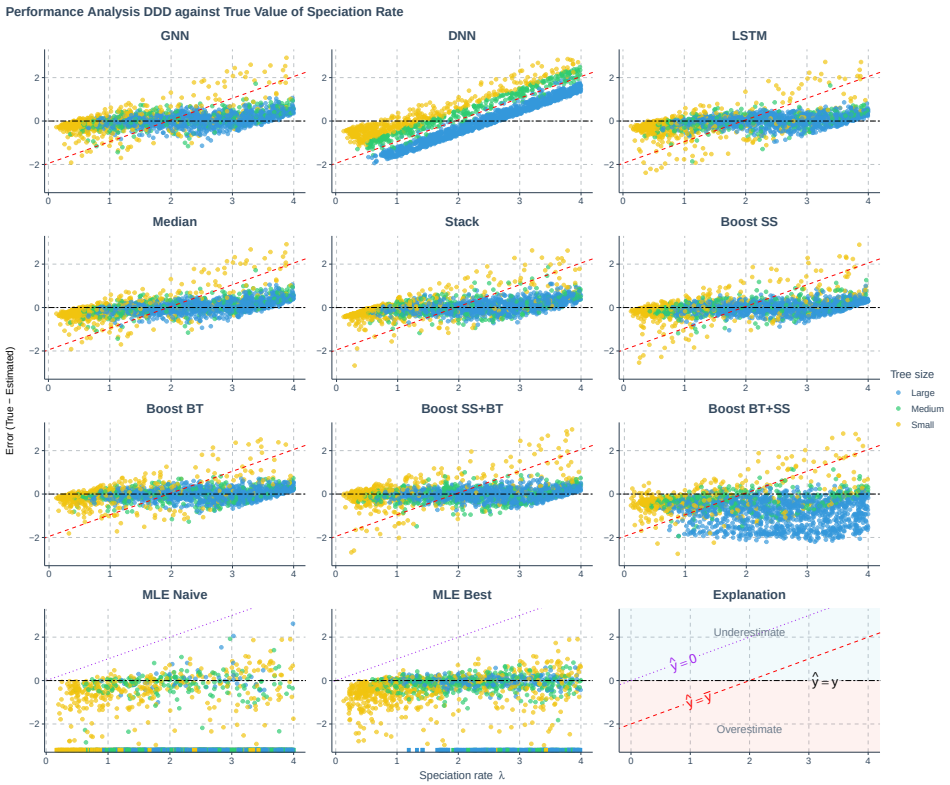


Figure 3.4: Prediction error of various methods applied to phylogenies simulated under a diversity-dependent diversification scenario, against true values of the speciation rate. The errors shown (y-axis) are the differences between the true parameters (x-axis) used to simulate the phylogenies and the values predicted or estimated by each method. Each panel represents an estimation method. Phylogenies are categorized based on their size: yellow for small phylogenies with fewer than 200 nodes (including root, internal, and tip nodes), green for medium-sized phylogenies with 200 to 500 nodes, and blue for large phylogenies with more than 500 nodes, refer to Appendix I, Figure 3.19 for how the tree sizes are distributed. GNN: Predictions obtained by the graph neural network using the phylogenies. DNN: Predictions by the dense neural network using summary statistics. LSTM: Predictions by the long short-term memory recurrent neural network using branching times. Median: Bagging strategy that takes the median value of the predictions from GNN, DNN, and LSTM. Stack: Stacking strategy that utilizes a meta-learner to integrate results from GNN, DNN, and LSTM. Boost SS: Boosting strategy that corrects GNN results using DNN. Boost BT: Boosting strategy that corrects GNN results using LSTM. Boost SS+BT: Sequential correction of GNN errors first using DNN, followed by LSTM. Boost BT+SS: Sequential correction of GNN errors first using LSTM, followed by DNN. MLE Naive: Maximum Likelihood Estimation results using a random starting point within the parameter space of the training dataset for each parameter's optimization. MLE Best: MLE results using the true parameter values as the starting points for optimization. Red dashed lines in panels representing neural network results indicate the mid-points of the parameter spaces ($\hat{y} = \bar{y}$ where $\hat{y}$ denotes a estimated parameter and $\bar{y}$ denotes the mid-point of the parameter space). Data points close to purple dotted lines ($\hat{y} = 0$) in MLE result panels indicate near-zero estimates. Black two-dash lines indicate accurate estimates ($\hat{y} = y$ where y denotes the true parameter value). In the MLE result panels, small squares spreading along the x-axis signify optimization failures. Due to significantly lower accuracy, other aggregation methods from the bagging strategy are not displayed on the plot. λ: Speciation rate.

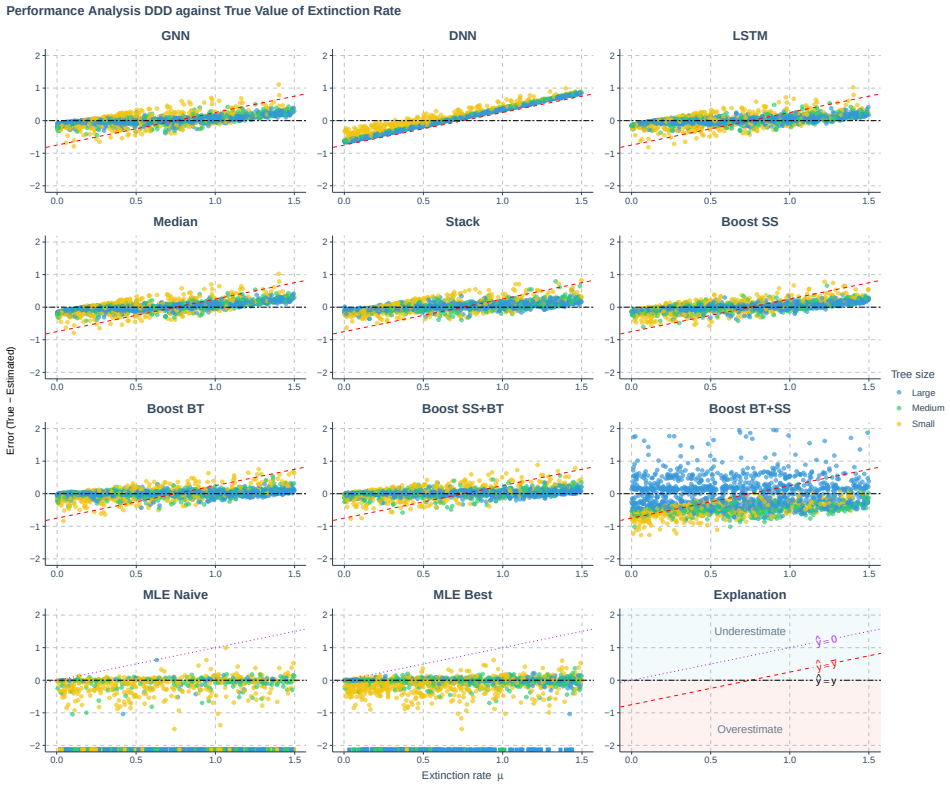


Figure 3.5: Prediction error of various methods applied to phylogenies simulated under a diversity-dependent diversification scenario, against true values of the extinction rate. The errors shown (y-axis) are the differences between the true parameters (x-axis) used to simulate the phylogenies and the values predicted or estimated by each method. Each panel represents an estimation method. Phylogenies are categorized based on their size: yellow for small phylogenies with fewer than 200 nodes (including root, internal, and tip nodes), green for medium-sized phylogenies with 200 to 500 nodes, and blue for large phylogenies with more than 500 nodes. GNN: Predictions obtained by the graph neural network using the phylogenies. DNN: Predictions by the dense neural network using summary statistics. LSTM: Predictions by the long short-term memory recurrent neural network using branching times. Median: Bagging strategy that takes the median value of the predictions from GNN, DNN, and LSTM. Stack: Stacking strategy that utilizes a meta-learner to integrate results from GNN, DNN, and LSTM. Boost SS: Boosting strategy that corrects GNN results using DNN. Boost BT: Boosting strategy that corrects GNN results using LSTM. Boost SS+BT: Sequential correction of GNN errors first using DNN, followed by LSTM. Boost BT+SS: Sequential correction of GNN errors first using LSTM, followed by DNN. MLE Naive: Maximum Likelihood Estimation results using a random starting point within the parameter space of the training dataset for each parameter's optimization. MLE Best: MLE results using the true parameter values as the starting points for optimization. Red dashed lines in panels representing neural network results indicate the mid-points of the parameter spaces ($\hat{y} = \bar{y}$ where $\hat{y}$ denotes a estimated parameter and $\bar{y}$ denotes the mid-point of the parameter space). Data points close to purple dotted lines ($\hat{y} = 0$) in MLE result panels indicate near-zero estimates. Black two-dash lines indicate accurate estimates ($\hat{y} = y$ where y denotes the true parameter value). In the MLE result panels, small squares spreading along the x-axis signify optimization failures. Due to significantly lower accuracy, other aggregation methods from the bagging strategy are not displayed on the plot. μ: extinction rate.

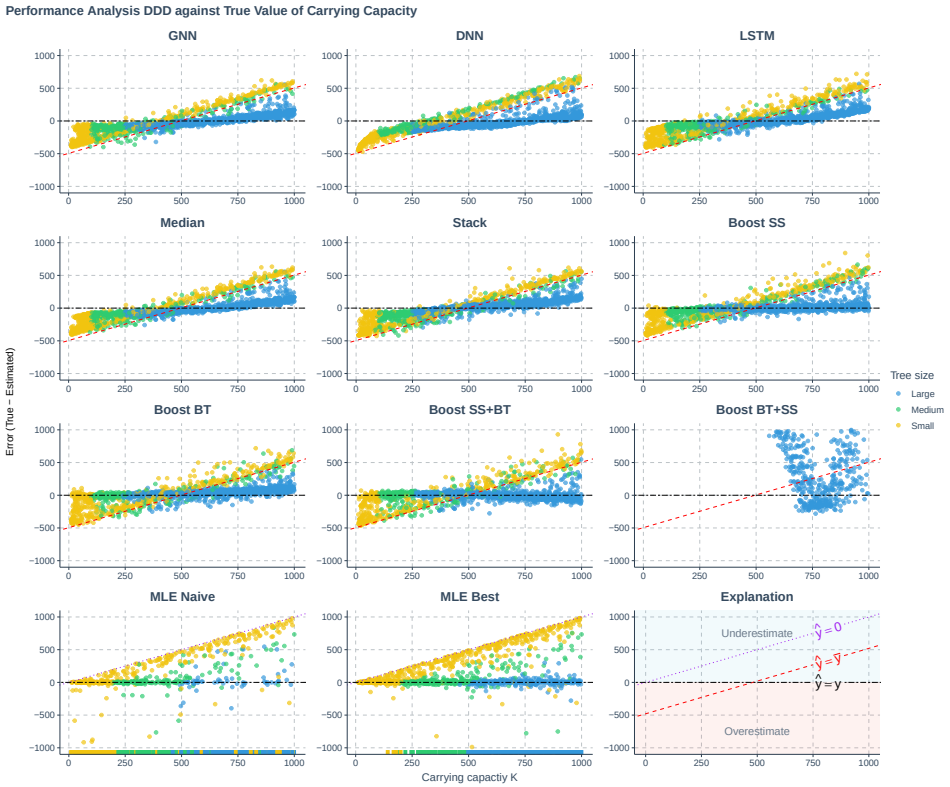


Figure 3.6: Prediction error of various methods applied to phylogenies simulated under a diversity-dependent diversification scenario, against true values of the carrying capacity. The errors shown (y-axis) are the differences between the true parameters (x-axis) used to simulate the phylogenies and the values predicted or estimated by each method. Each panel represents an estimation method. Phylogenies are categorized based on their size: yellow for small phylogenies with fewer than 200 nodes (including root, internal, and tip nodes), green for medium-sized phylogenies with 200 to 500 nodes, and blue for large phylogenies with more than 500 nodes. GNN: Predictions obtained by the graph neural network using the phylogenies. DNN: Predictions by the dense neural network using summary statistics. LSTM: Predictions by the long short-term memory recurrent neural network using branching times. Median: Bagging strategy that takes the median value of the predictions from GNN, DNN, and LSTM. Stack: Stacking strategy that utilizes a meta-learner to integrate results from GNN, DNN, and LSTM. Boost SS: Boosting strategy that corrects GNN results using DNN. Boost BT: Boosting strategy that corrects GNN results using LSTM. Boost SS+BT: Sequential correction of GNN errors first using DNN, followed by LSTM. Boost BT+SS: Sequential correction of GNN errors first using LSTM, followed by DNN. MLE Naive: Maximum Likelihood Estimation results using a random starting point within the parameter space of the training dataset for each parameter's optimization. MLE Best: MLE results using the true parameter values as the starting points for optimization. Red dashed lines in panels representing neural network results indicate the mid-points of the parameter spaces ($\hat{y} = \bar{y}$ where $\hat{y}$ denotes a estimated parameter and $\bar{y}$ denotes the mid-point of the parameter space). Data points close to purple dotted lines ($\hat{y} = 0$) in MLE result panels indicate near-zero estimates. Black two-dash lines indicate accurate estimates ($\hat{y} = y$ where y denotes the true parameter value). In the MLE result panels, small squares spreading along the x-axis signify optimization failures. Due to significantly lower accuracy, other aggregation methods from the bagging strategy are not displayed on the plot. K: carrying capacity.

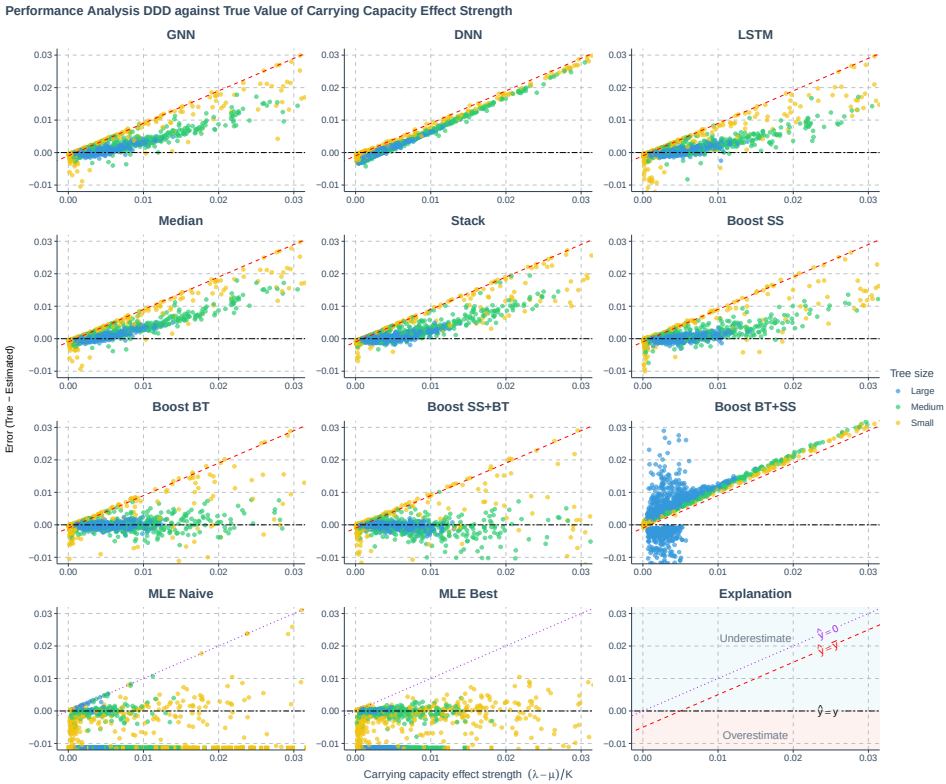


Figure 3.7: Prediction error of estimated carrying capacity effect computed from estimated values of speciation rate, extinction rate and carrying capacity using various methods applied to phylogenies simulated under a diversity-dependent diversification scenario, plotted against the true carrying capacity effect computed from true parameters. The errors shown (y-axis) are the differences between the values of true carrying capacity effect (x-axis) used to simulate the phylogenies and the values predicted or estimated by each method. Each panel represents an estimation method. Phylogenies are categorized based on their size: yellow for small phylogenies with fewer than 200 nodes (including root, internal, and tip nodes), green for medium-sized phylogenies with 200 to 500 nodes, and blue for large phylogenies with more than 500 nodes. GNN: Predictions obtained by the graph neural network using the phylogenies. DNN: Predictions by the dense neural network using summary statistics. LSTM: Predictions by the long short-term memory recurrent neural network using branching times. Stack: Stacking strategy that utilizes a meta-learner to integrate results from GNN, DNN, and LSTM. Boost SS: Boosting strategy that corrects GNN results using DNN. Boost BT: Boosting strategy that corrects GNN results using LSTM. Boost SS+BT: Sequential correction of GNN errors first using DNN, followed by LSTM. MLE Naive: Maximum Likelihood Estimation results using random starting points for parameter optimization. MLE Best: MLE results using the true parameter values as the starting points for optimization. Red dashed lines in panels representing neural network results indicate the mid-points of the parameter spaces ($\hat{y} = \bar{y}$ where $\hat{y}$ denotes an estimated parameter and $\bar{y}$ denotes the mid-point of the parameter space). Data points close to purple dotted lines ($\hat{y} = 0$) in MLE result panels indicate near-zero estimates. Black two-dash lines indicate accurate estimates ($\hat{y} = y$ where y denotes the true parameter value). In the MLE result panels, small squares spreading along the x-axis signify optimization failures. λ: Speciation rate. μ: extinction rate. K: carrying capacity.

exhibiting no bias, even on smaller phylogenies. Overall, sequential boosting of GNN results first with DNN and then with LSTM (Boost SS+BT) led to best performance in terms of prediction accuracy, except for estimating carrying capacity on large phylogenies with around 2000 nodes (see the row named Boost SS+BT in Appendix I, Figure 3.16 this strategy led to overestimation on very large phylogenies). However, Boost SS+BT led to more overestimation on the true values of the carrying capacity effect, as compared to the strategy boosting the GNN results with only LSTM (Boost BT, see Figure 3.7).

In general, boosting methods (except for Boost BT + SS) can yield more accurate and less biased estimates across tree sizes, matching or outperforming MLE. In particular, the Boost SS+BT strategy achieves the highest overall accuracy by improving inference of carrying capacity effects.

The neural network architecture by Voznica et al. [144] performed similarly and exhibited similar patterns as our approaches. CNN1D better recovered the carrying capacity effect strength than than GNN, DNN and LSTM alone, but lagged behind our boosting approaches–except for Boost BT+SS–in overall parameter prediction accuracy (see Figure 3.32, Figure 3.28, Figure 3.29, Figure 3.30 and Figure 3.31 in Appendix L).

3.3.2 Robustness Analysis

As a proxy for robustness of each method, we used the mean absolute errors of the parameters estimated from sets of phylogenies simulated under identical true parameters. Our analysis indicates that the robustness of the methods against phylogenetic heterogeneity (e.g., phylogenies of very different sizes, topologies and other characteristics) depends on the values of the underlying true parameters. We observed that the strength of the carrying capacity effect critically influences robustness. Generally, a weaker carrying capacity effect (associated to a smaller value of $(\lambda - \mu)/K$) tends to diminish the robustness of both MLE and neural network methods across all parameters: speciation rate, extinction rate, and carrying capacity (as can be seen in Figure 3.8 and Appendix I, Figure 3.17 and Figure 3.18, by observing the increase of error along with the darkening background colors from light pink to dark blue).

When the carrying capacity effect is weak, neural network methods typically exhibit greater robustness in estimating speciation and extinction rates compared to the best-case MLE results (see Figure 3.8 and Appendix I, Figure 3.18). When the carrying capacity effect is exceptionally strong, the best-case MLE results can outperform neural networks particularly when estimating carrying capacity. Naive-case MLE results consistently show less robustness compared to all neural network methods.

A higher extinction rate generally decreases the robustness of all methods in estimating any parameter. A higher speciation rate enhances the robustness of carrying capacity estimates across all methods, although its impact on the robustness of speciation and extinction rate estimates is not consistent. A higher carrying capacity generally decreases the robustness of all methods in estimating carrying capacity.

Note that MLE naive-case results often contained more extreme estimations than best-case results, consequently, the exclusion of extreme values could lead to a wrong impression in the figures that when the carrying capacity effect is weak, the naive-case MLE is more

robust than the best-case MLE. This is particularly prominent for the speciation rate. The exclusion of these extreme values is crucial, however, as they are rare and their magnitude can obstruct meaningful interpretation and comparison.

We find that DNN alone (estimating parameters from summary statistics) shows the worst robustness among all the methods and LSTM alone (estimating parameters from branching times) shows the greatest robustness overall. Among all the estimation methods, the MLE best-case achieved the greatest possible robustness in estimating the extinction rate and the carrying capacity while GNN alone (estimating parameters from phylogenies) achieved the greatest possible robustness in estimating the speciation rate. Among the neural network methods, GNN alone achieved the greatest possible robustness in estimating the speciation rate and the carrying capacity while Boost BT (boosting GNN estimates with LSTM) achieved the the greatest possible robustness in estimating the extinction rate. See Figure 3.8 and Appendix I, Figure 3.17 and Figure 3.18 for details.

3.3.3 Misspecification Analysis

Under model misspecification, in the naive cases, MLE estimates of the carrying capacity (DDD) based on trees generated under BD K scale closely with the expected total number of nodes $\mathbb{E}[N_{\text{nodes}}(t)]$ and are generally larger than the total number of nodes of the underlying trees. Under the naive case, 5.0% of K estimates fell in $[2N_{\text{nodes}}, 5N_{\text{nodes}}]$, 1.9% in $[5N_{\text{nodes}}, 20N_{\text{nodes}}]$, 5.4% in $[20N_{\text{nodes}}, \infty)$ and 2.6% are estimated as ∞. In the best-case scenario (K initialized at ∞), MLE always converged to $K = \infty$.

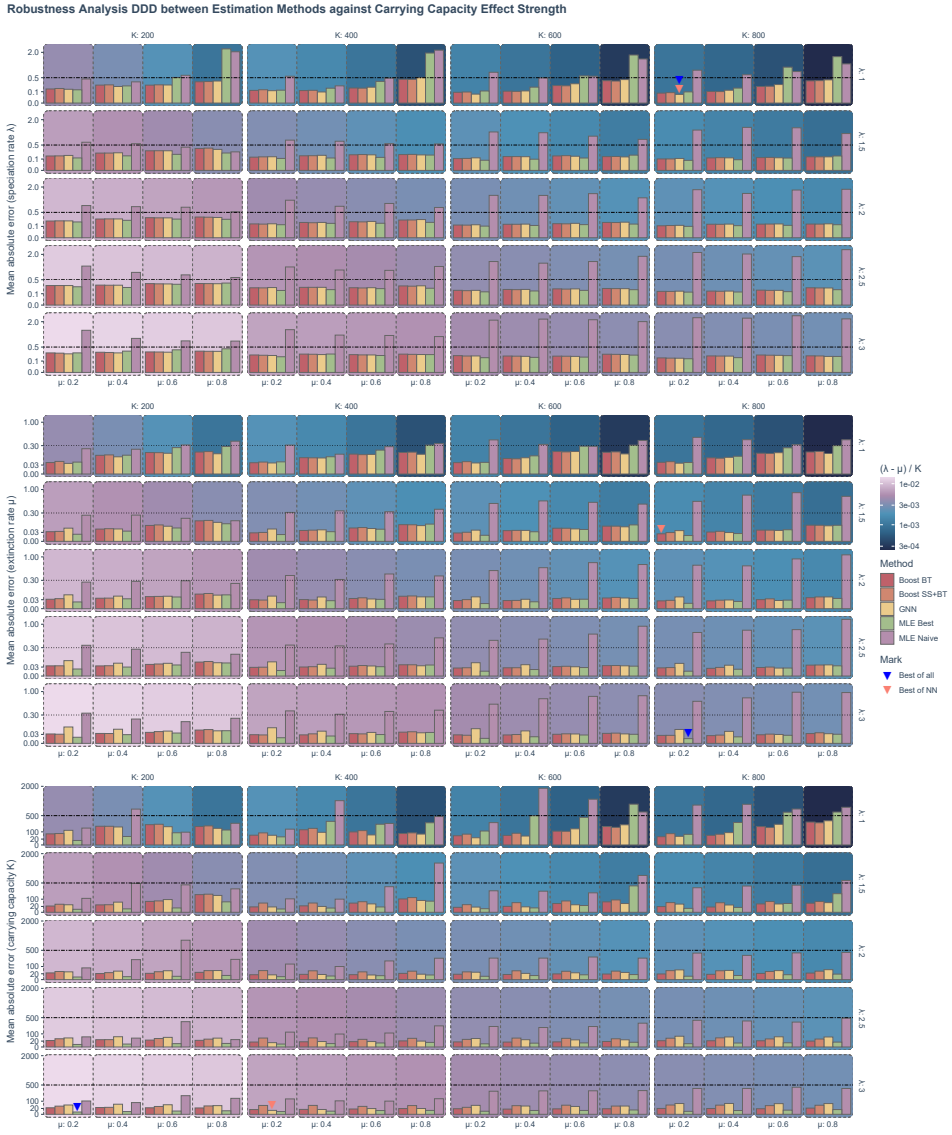
Neural network estimates of K reflect both the center of the training range (mean of sampled K values) and tree size (N_{tips}, roughly half the total number of nodes). On medium- to large-sized trees, predictions concentrate near N_{tips}, and never exceed the training upper bound of 1000 when using the GNN alone. The Boost BT ensemble approach occasionally yields slightly larger estimates than 1000 on large trees, but exhibits the same size-dependent bias toward N_{tips}.

See Figure 3.9 for details.

3.3.4 Empirical Data

For the empirical data sets, MLE estimates of carrying capacity are typically lower than those of the neural networks, especially in smaller phylogenies (Figure 3.10). However, as the size of the phylogenies increases, MLE estimates tend to converge towards those produced by neural networks. Similarly, MLE estimates of net diversification rate also align more closely with neural network estimates in larger phylogenies Figure 3.10.

MLE generally provides a broader range of estimates on all the parameters except for carrying capacity on small phylogenies. Neural networks provide a broader range of carrying capacity estimates on small phylogenies and less frequently produce zero or near-zero estimates for extinction rates which we often observe for MLE. We also observed that MLE sometimes produces very high values (ranging from 10,000 to infinity) for carrying capacity on empirical trees, see Figure 3.10 for the comparison between MLE and neural networks on empirical tree parameter estimation.



(Caption on next page.)

Figure 3.8: *(Figure on previous page.)* The robustness (mean absolute error) of neural network and maximum likelihood estimation was assessed for 80 sets of phylogenies, each containing 1000 trees randomly simulated under a diversity-dependent diversification scenario, employing identical parameter settings but varying in size, topology, and structure. The robustness results of the speciation rate, the extinction rate and the carrying capacity are shown from top to bottom. For each panel group associated to a parameter, each panel contains the robustness of different estimation methods (the MLE and neural networks) under a combination of parameters indicated by the facet strip labels. Each facet column represents the robustness under a specific carrying capacity (K) setting used in the simulation of the phylogenies. Each facet row represents a specific speciation rate (λ). Each group of the bars represents a specific extinction rate (μ) as shown by the x-axis. The background color of a panel represents the carrying capacity effect strength (calculated as $(\lambda - \mu)/K$ and visualized in "log10" scale), from bottom-left to top-right, the carrying capacity effect strength decreases. The color of a bar represents the associated estimation method. Boosting BT: Graph neural network with long short-term memory recurrent neural network correcting its residuals using branching times. Boosting SS + BT: Graph neural network with dense neural network and long short-term memory recurrent neural network correcting residuals sequentially using summary statistics and branching times. GNN: Graph neural network. MLE Best: Maximum likelihood estimation using true parameters as the starting points. MLE Naive: Maximum likelihood estimation using a random value as the starting point of optimization for each parameter. X-axis: Represents extinction rate (μ) settings. Y-axis: Represents the mean absolute error in a square-root transformed scale. Some bars are marked; for each parameter, the blue triangle represents the greatest possible robustness achieved among all the estimation methods, the red triangle represents the greatest possible robustness achieved among the neural network methods.

Generally, neural network estimates of all the parameters under the DDD scenario are close to the center (mean) of the distribution generated by the bootstrapping method. See Figure 3.15 in Appendix H for details.

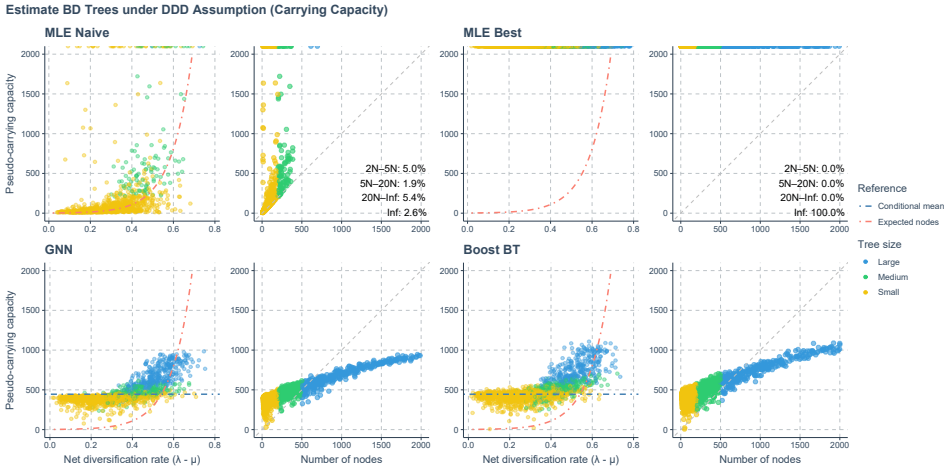


Figure 3.9: Results of applying MLE and neural networks for DDD trees to trees generated under BD (model misspecification). The predicted carrying capacities by estimation method (y-axis) are plotted against real net diversification rates $\lambda - \mu$ and numbers of observed total nodes of the trees N_{nodes} (x-axis). There are four groups of panels. Each panel group represents an estimation method. Phylogenies are categorized based on their size: yellow for small phylogenies with fewer than 200 nodes (including root, internal, and tip nodes), green for medium-sized phylogenies with 200 to 500 nodes, and blue for large phylogenies with more than 500 nodes. MLE Naive: Maximum Likelihood Estimation results with starting point set to true speciation and extinction rates and carrying capacity $K = 10000$. MLE Best: MLE results with starting point set to true speciation and extinction rates and carrying capacity $K = \infty$). GNN: Predictions obtained by the graph neural network using the phylogenies. Boost BT: Boosting strategy that corrects GNN results using LSTM. Red dot-dashed lines in panels referencing the expected total number of nodes of the trees under the true BD process given net diversification rate and simulation time. Blue dot-dashed lines referencing the mid-points of the parameter space of carrying capacity when training the neural networks. Gray dashed lines indicate where the number of nodes of a tree is equal to its estimated pseudo-carrying capacity. In the MLE result panels, predicted pseudo-carrying capacities larger than 2000 but not ∞ are not shown. The actual infinite predictions are displayed as points attaching to the top panel borders. Percentages shown at the bottom-right corners indicate the proportions of predicted pseudo-carrying capacities lying between certain value ranges: (1) between 2 N_{nodes} and 5 N_{nodes}, (2) between 5 N_{nodes} and 20 N_{nodes}, (3) between 20 N_{nodes} and ∞, and (4) ∞.

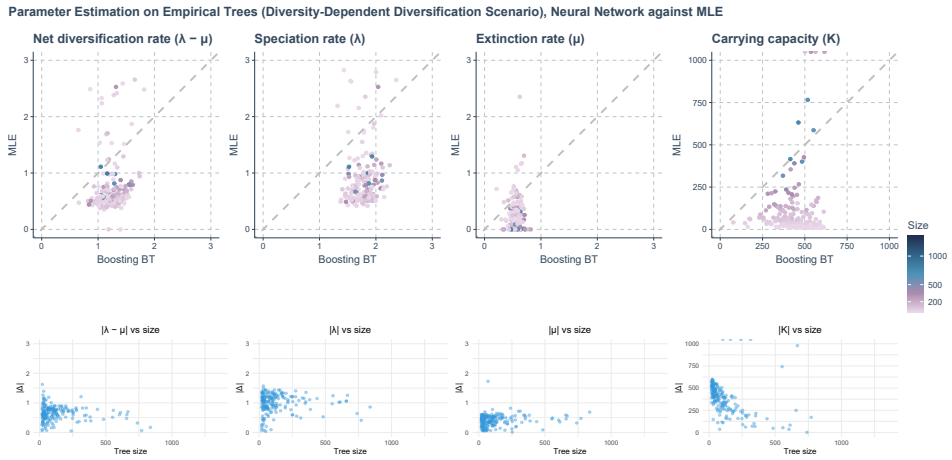


Figure 3.10: Comparison of the estimations of Maximum Likelihood Estimation (MLE) and neural network methods (specifically, Boosting BT, which refers to using graph neural network to make first predictions and then using a long short-term memory recurrent neural network to correct for residuals) on empirical trees under a diversity-dependent diversification scenario. Each column, arranged from left to right, focuses on a specific parameter being estimated. In the first row, x-axis represents the estimated values of the neural network. Y-axis represents estimated values from MLE. A gray dashed line is included in each panel to indicate where the estimations from the neural network and MLE are exactly the same. The color of the points varies from purple to blue, with the gradient representing the size of the phylogenies measured by the total number of nodes (including root, internal, and tip nodes). In the second row, the change of absolute differences between MLE and neural network predictions along tree sizes are presented. X-axis represents tree size, y-axis represents absolute differences.

3.3.5 Other Scenarios

Birth–Death Scenario

Neural network methods outperformed MLE in accuracy on smaller phylogenies, under the BD scenario in the simulated dataset (see Appendix J, Figure 3.20 and Figure 3.21). Both MLE and neural network methods give less accurate estimates on small phylogenies; this is more prominent for the MLE estimates.

Net diversification rate strongly affects predictive accuracy across all methods: larger net diversification rates typically yield larger trees and result in lower prediction errors. Extinction-to-speciation ratio has minimal impact on neural network accuracy and only a weak effect on MLE performance–higher ratios slightly increase prediction errors (see Appendix J, Figure 3.22 and Figure 3.23).

On empirical phylogenies, similar to the DDD scenario, neural network methods seldom produce zero-estimation of extinction rate, unlike MLE, which often produces zero or near-zero estimates for the extinction rate. Neural networks tend to give estimates within the parameter space of the training dataset. They predict conservative speciation and extinction rates yet are highly consistent with MLE estimation on the net diversification rate. The

consistency of prediction increases on larger empirical phylogenies. See Figure 3.11 for details.

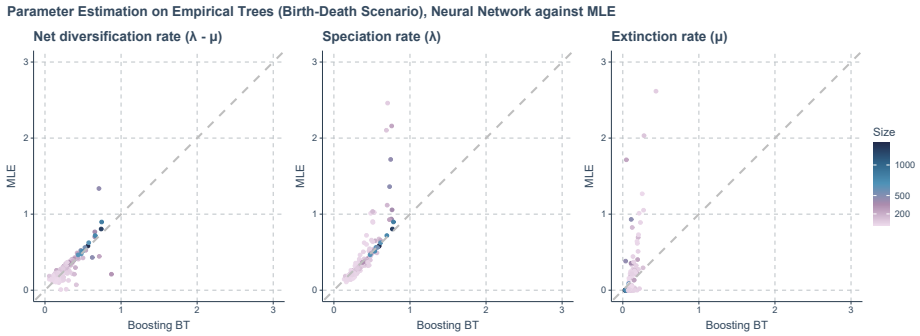


Figure 3.11: Comparing the estimations of Maximum Likelihood Estimation (MLE) and neural network methods (specifically, Boosting BT, which refers to using graph neural network to make first predictions and then using long short-term memory recurrent neural network to correct for residuals) on empirical trees under a birth–death scenario. Each panel, arranged from left to right, focuses on a specific parameter being estimated. X-axis: Represents the estimated values of the neural network. Y-axis: Represents estimated the values from MLE. A gray dashed line is included in each panel to indicate where the estimations from the neural network and MLE are exactly the same. The color of the points varies from purple to blue, with the gradient representing the size of the phylogenies measured by the total number of nodes (including root, internal, and tip nodes).

Protracted Birth–Death Scenario

Both maximum likelihood estimation (MLE) and neural network methods did not perform well on estimating parameters under the PBD scenario; MLE estimates were generally less accurate, but neural networks also failed to predict the parameters as all the parameter estimates are close to the mid-points of corresponding parameter spaces (Appendix K, Figure 3.24,). However, there are exceptions: neural networks seem to perform better on the speciation rate of the incipient species (λ_3) and on the mean duration of speciation (τ) when the true value is between 0 and 2.

MLE estimates become significantly inaccurate as phylogenies become smaller (Appendix K, Figure 3.25); it is also noticeable that MLE estimates of the speciation completion rate (λ_2) are very inaccurate, especially when the phylogenies are large. A general pattern is that both MLE and neural network methods achieve more accurate estimates on phylogenies with higher true values of the mean duration of speciation (Appendix K, Figure 3.26).

3.4 Discussion

We have developed an ensemble learning based neural network approach that matches and sometimes outperforms the accuracy and robustness of maximum likelihood estimation (MLE) for estimating phylogenetic tree parameters. Our approach leverages different classes of neural networks by learning from the phylogenies, their branching times and their summary statistics simultaneously.

When trained, our neural networks can compute estimates faster than MLE on larger phylogenies as computation time is less affected by increases in phylogeny size. We considered boosting strategies most effective in eliminating systematic prediction errors in neural network estimates. Among them, Boost BT (which corrects GNN results using LSTM) achieved overall best performance which is comparable to, or even surpassing, the best-case MLE, in terms of accuracy and robustness. We observed that generally the performance of the naive MLE was second-worst (Boost BT+SS was worst). Interestingly, some phylogenies, such as small trees and those shaped by relatively weak effects, pose significant challenges to both MLE and neural network methods.

Previous neural network methods applied to phylogenies have experimented with various architectures such as CNN1D, GNN, and LSTM [126, 143, 144]. The deep learning architectures we employed differ from those used in prior studies, making direct comparisons challenging. However, the comparison we performed against CNN1D reveals that these neural network architectures we have tested shared similar patterns. Additionally, while previous research focused on birth–death [126] and trait-state-dependent models [143, 209], our approach is novel in its application to models such as diversity-dependent diversification (DDD) and protracted birth–death (PBD) from a neural network perspective. Despite these differences, our findings align with recent studies in underscoring the potential of neural networks to infer diversification processes, offering a viable alternative to mathematically complex methods. Future research could attempt conformalized prediction approaches to quantify uncertainty in the NNs, as these have been shown to be quite efficient [223].

3.4.1 Rethinking Neural Networks

Although performance was equal or potentially better than MLE, the neural network approach is not without its shortcomings. The neural networks often defaulted to predicting values close to the mid-point of the true parameter space of the training dataset, indicating that they struggle to extract meaningful features from the dataset. This predictive performance in the absence of information is similar to that of Bayesian analysis, in which the posterior is equal to the posterior when the likelihood function is flat. The similarity also depends on the choice of loss function, for example when the mean-squared-error (MSE) is chosen, the neural-network regression setting is similar to maximum-likelihood estimation with independent and Gaussian observation noise. Under these assumptions the network converges on the conditional mean of the target variable. In our study we adopted the Huber loss, which retains this connection while introducing heavier tails to reduce the impact of outliers, thereby behaving much like MSE [224, 225]. This conservative prediction strategy minimizes overall error compared to random guessing. Examples can be found in the GNN predictions of carrying capacity from simulated DDD trees, the DNN estimates of the speciation and extinction rates (Figure 3.4, Figure 3.5 and Figure 3.6, but see Appendix O for a detailed investigation of possible under-performance of DNN on the summary statistics) and most neural network predictions of PBD related parameters (Appendix K, Figure 3.24), especially for smaller phylogenies.

This behavior, while effective in reducing apparent error metrics, can skew our understanding of a neural network's performance particularly when the focal parameter space is

relatively narrow. Neural networks may consistently show smaller overall error compared to MLE, because the latter has no prior knowledge of the limits of the parameter space, which would lead to a false impression of better accuracy of the neural networks. We therefore recommend performing case-specific residual analyses on the neural network predictions and the MLE estimates, which are often overlooked or over-simplified. This behavior also dominates neural network predictions on small trees under model misspecification, where neural networks trained on DDD trees try to recover carrying capacity K from BD trees (see Figure 3.9). If robust prior bounds on the admissible parameter space are available, however, neural networks can leverage the knowledge and achieve better performance.

To mitigate the impact of training datasets that may insufficiently represent the parameter space, we can first train the neural networks using a relatively narrow training dataset and generate parameter predictions from empirical data; then retrain the networks with a broader training dataset that covers a larger or different portion of the parameter space. By comparing predictions of the neural networks from the two training datasets on the same empirical data, we can evaluate whether and how strongly our predictions depend on the parameter ranges represented in the training data. This procedure is similar to a prior sensitivity analysis in Bayesian inference. Generally, we recommend to train the neural networks with as large a dataset (sample size) and as broad a parameter space as possible, unless there is prior knowledge of the parameter space being estimated.

Improving neural network predictions that are close to the mean is unlikely to be achieved by increasing the amount of training data: we did not observe major performance improvement when changing the size of the datasets (from 1,000 to 100,000 phylogenies per dataset). Instead, one might consider increasing the complexity of the network architecture, such as increasing their depth or adapting the scale of the hidden nodes [226], but note that this can only work if there is additional signal in the data (as discussed in the next section) and this will typically require more data.

Although potentially beneficial, increasing the depth can also harm predictive power. In particular for GNNs, increasing their depth may lead to ”over-smoothing” and ”over-squashing”. Over-smoothing causes node features to become increasingly similar as more layers are added [227], leading to a loss of distinct node embeddings across different clusters. Over-squashing involves the compression of expansive node information through bottleneck edges into a fixed-size vector, which is problematic in graphs with large diameters and long-range dependencies [228], e.g.phylogenies. Both issues degrade node representations and distort information flow, making deeper GNNs potentially less effective than shallower ones [229]. Moreover, over-smoothing and over-squashing are intrinsically linked, creating a trade-off that cannot be easily resolved [230].

In our analyses, we observed that increasing the number of GraphSAGE layers beyond three in the differentiable pooling architecture destabilized the training process and reduced the accuracy of estimates on validation datasets, introducing more outliers. We therefore opted to maintain two layers throughout our study. We explored newer algorithms designed to mitigate deep GNN issues [227, 231, 232], but found that these deeper architectures performed worse than our differentiable pooling approach with fewer layers.

For DNN and LSTM, we also experimented with more complex architectures, different activation functions and various hyper-parameter optimizations but failed to achieve better performance.

3.4.2 Fundamental Problems with Phylogenies

The lack of improvement when changing the amount of training data or the network architecture suggests that the real challenges of estimating parameters might not lie in the architecture of the networks, but might instead be attributed to underlying weak or absent phylogenetic signals. Whenever this is the case, we expect similarities in inaccuracies of both MLE and neural network approaches. This occurs, for example, for the carrying capacity when it is high and thus has a weak effect (measured by $(\lambda - \mu)/K$). Here, the phylogeny is typically not near the carrying capacity, allowing the number of species to grow (almost) unbounded. This may result in carrying capacity estimates that are arbitrarily high, especially in the MLE methods. The PBD scenario is known to present difficulties in reliably recovering parameters with MLE [123] and we find similar poor performance with neural networks. A second case where accurate parameter estimation is complicated occurs when extinction processes erase critical information [87], as observed in the decline of estimation robustness associated with increasing extinction rate (see Figure 3.8 for a comparison of errors across different values of μ: the prediction errors tend to increase with μ, particularly when μ constitutes a larger proportion relative to λ).

More generally, small phylogenies tend to contain less information than large ones. In our results we see that estimation accuracy and robustness decline with decreasing size of the phylogenies. This trend is observed across both MLE and neural network methods. In the BD and PBD scenarios, where datasets have greater variability in phylogeny sizes, poor estimations for small trees could be explained by both low information content, or under-representation of such trees. After re-balancing tree sizes (the first supplementary Appendix M, the same patterns occurred and they are therefore unlikely to be a result of under- or over-representation of different phylogeny sizes, but instead reflect low information content of small trees (compare Figure 3.20 and Figure 3.33). Empirical phylogenies usually offer only one single tree per one exact process, so formal goodness-of-fit tests have low power regardless of the estimator. In such single-realization settings, the safest strategy is to compare several estimators, if available; large discrepancies are a warning sign of misspecification.

3.4.3 Confronting the Empirical Phylogenies

The processes of evolution within natural systems are often unknown. Determining the "true parameters" of an empirical phylogeny is challenging, even when they meet theoretical assumptions, making it difficult to evaluate which tool provides more accurate estimates. Therefore, choosing the right tool is crucial.

With neural networks, it is possible that the true parameter value is not part of the assumed parameter range for simulating the training data. In such cases, neural network accuracy decreases notably, as shown in our second supplementary analysis (Appendix N). We also noticed that when comparing the estimates of MLE and neural network methods on the empirical phylogenies (see Figure 3.10, Figure 3.11), MLE estimates spread wider than

the neural networks (e.g. our BD training dataset comprises phylogenies simulated using speciation rate between 0 and 0.8 and extinction rate between 0 and 0.72, where our neural networks never predict speciation rates larger than 0.8 or extinction rates larger than 0.72, see Figure 3.11, similar results can be found under the DDD scenario in Figure 3.10). Expanding the training dataset's parameter space can resolve the generalization issue (we expanded our training datasets several times in the experiments), but this approach requires significantly more computational resources for both simulation and training of the neural networks.

Our supplementary study on generalizability and data completeness (explained in Appendix N) also reveals that neural networks tend to provide more accurate estimates of speciation and extinction rates from complete phylogenies (with both extant and extinct tips) than from phylogenies with only extant species under the BD scenario. This increase in accuracy was not observed in the DDD scenario. While complete phylogenies offer a broader picture and more contextual information, obtaining them is challenging because it is nearly impossible to account for all extinct species.

Our analyses indicate that GNN is more robust but more prone to systematic errors (GNN achieved the greatest possible robustness in estimating the speciation rate and carrying capacity among neural network methods). We show that using GNN as a base and other neural networks like LSTM to enhance GNN might effectively combine the advantages of different methods and information sources, thus strengthening overall generalization ability. Our boosting methods (e.g. Boost BT) perform the best in this context.

In conclusion, when applied with caution, we expect that neural network methods can be applied to diversification scenarios where MLE is absent or non-tractable, as our best-performing neural network method showed comparable or even better performance to the best-case MLE. Our neural networks particularly perform better than MLE in terms of accuracy and robustness on small phylogenies and can be significantly faster when estimating very large phylogenies. Thus, if properly trained, neural network methods may substitute for or at least cross-reference with MLE estimates where they exist.

3.5 Appendix

A) Protocol Transforming Phylogenies

Phylogenetic trees are usually stored in the ”phylo” data format in R. This data format is not directly compatible with GNN implementations. To facilitate graph convolutional operations, we transformed phylogeny from a ”phylo” object into three major components: adjacency list, node feature matrix and graph attributes (see Figure 3.2). The adjacency list contains information on the connectivity between nodes and tips, the node feature matrix contains distances between nodes and tips, and the graph-level attributes include the ground truth model parameters used to generate the phylogenies. These components are stored in separate tensors. In machine learning, a tensor is a mathematical object that generalizes scalars, vectors, and matrices to higher dimensions, allowing complex operations to be performed efficiently on multi-dimensional arrays.

Adjacency List

In the context of a phylogenetic tree, tip nodes usually represent taxonomic units such as species, while root nodes and internal nodes represent the points where two taxonomic units depart from each other. An edge in a phylogenetic tree represents the hierarchical connection between two nodes (the ancestor and the descendant), and as such describes the evolutionary relatedness between taxa. Each root node, internal node and tip node in an R ”phylo” object is indexed sequentially, each edge is also sequentially indexed independently of node indices. The sub-list ”edge” of a ”phylo” object contains the adjacency list of a phylogenetic tree which describes the relationships between nodes. Each row of the adjacency list represents an edge, the first column contains the index (or numbering) of the ancestor node, and the second column contains the index of the descendant node.

This data structure effectively captures the tree’s branching pattern, showing how each taxon (or node) is connected to others. The adjacency list in ”phylo” object uses a ”1-based” indexing in R, we therefore element-wise deduct 1 from the list to convert it into ”0-based” indexing which is compatible with the python environment.

We output the converted adjacency list within the ”phylo” object as the adjacency list $\mathscr{E}$ of the graph representation, in PyTorch Geometric, which is conventionally named as ”data.edge_index”. We store $\mathscr{E}$ as a ”torch.long” long integer type tensor and transpose it such that it has shape $[2, num\_edges]$, where ”num_edges” is the number of edges in the ”phylo” object. This tensor has two dimensions. This way, the connections between nodes in the transformed graph are all single-directional, from the ancestor nodes to their descendants (if any). Training the GNN with graphs of non-directed edges gives no performance advantage, according to our tests in phylogenetic tree parameter estimation tasks. Single-directional data structure can save GPU memory and reduce the computation complexity.

Node Feature Matrix

In a ”phylo” object in R, the ”edge.length” sub-list defines the lengths of the edges in the phylogenetic tree. In a phylogenetic context, these lengths often correspond to evolutionary distances, time, or genetic change. ”edge.length” is a numeric vector where each element

corresponds to the length of the edge as defined in the adjacency list. The order of lengths in the "edge.length" vector aligns with the order of edges in the adjacency list.

For each tree, we aggregate information contained in "edge.length" to a node feature matrix. Each row of the matrix represents features contained in a node. The first column contains the edge length from a node to its direct ancestor node, the second and the third columns contain the edge lengths from a node to its two daughter nodes. We pad the row of the root node with an 0 in the first column as it has no ancestor. We also pad the rows of the tip nodes with two 0s in the second and the third columns as they have no descendants. The row order of feature matrix aligns with the order of edges in the adjacency list.

We output the node feature matrix of each tree as the node feature matrix $\mathcal{X}$ of the graph representation, in PyTorch Geometric, this is conventionally named as "data.x". We store $\mathcal{X}$ as a "torch.float" floating point type tensor, it has shape $[num\_nodes, num\_node\_features]$, where "num_nodes" is the number of nodes (including tip nodes) in the "phylo" object and "num_node_features" in our case is 3, i.e. the phylogenetic distances from a node to its ancestor (if any) and two descendants (if any). This tensor has two dimensions. We do not store the phylogenetic distance information in edge features because GCN operators will eventually pass and aggregate the edge features into each of the node. Our data structure is simpler and so is the GNN architecture.

Graph-Level Attributes as Training Targets

We store all the parameters used to simulate a tree (ground truth values) in the graph-level attributes $\mathcal{Y}$. These can have arbitrary length, which should be consistent with the number of the parameters to be estimated (the three diversification scenarios, BD, DDD and PBD, have different number of parameters). We store graph-level attributes as a "torch.float" floating point type tensor with length of the number of parameters we want to predict for each type of the phylogenetic tree. In PyTorch Geometric, graph-level attributes can be named as "data.y". The graph-level attributes are used as training targets to compute loss (see Appendix D for the definition of loss).

B) Protocol Transforming Summary Statistics

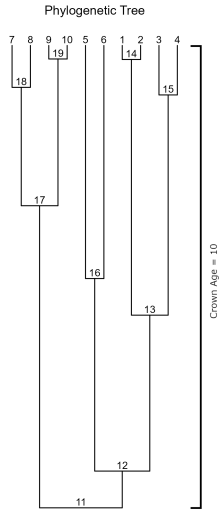
The summary statistics of a phylogeny are represented by a 1D vector, so the protocol for DNN is straightforward: we convert the vector into a tensor containing floating type data, with the shape [num_stats], where "num_stats" denotes the total number of statistics. This tensor has only one dimension. This conversion guarantees that each tensor is associated with its respective tree, with all contained statistics maintaining their original order. Within the PyTorch Geometric framework, these statistics are encapsulated as "data.stats" for each tree. When using DNN alone to estimate parameters from the summary statistics, the ground truth values of the parameters of the trees are stored in the same way as the graph-level attributes, as model training targets. When using DNN with other neural networks (e.g. in stacking and boosting strategies), they share the same ground truth values which are the graph-level attributes.

C) Protocol Transforming Branching Times

To address the varying lengths in branching times across different phylogenetic trees, we standardize these sequences by padding them to match the length of the longest branching time sequence. This is achieved by appending zeros to the shorter sequences until they match the predefined maximum length. The padded sequences are stored in tensors containing floating type data. As the original branching times do not contain zero values, this padding strategy allows us to distinguish between original data and padding. Consequently, we can pass masks of the sequences to the LSTM, which indicates the positions of the paddings, making LSTM concentrate only on the informative portions of the sequences, thereby optimizing its performance. When using LSTM alone to estimate parameters from the branching times, the ground truth values of the parameters of the trees are stored in the same way as the graph-level attributes, as model training targets. When using LSTM with other neural networks (e.g. in stacking and boosting strategies), they share the same ground truth values which are the graph-level attributes.

An example of data components extracted and computed from a phylogenetic tree is illustrated in Figure 3.12.

(Figure on next page.)



Adjacency List

Start	End
11	12
12	13
13	14
14	1
14	2
13	15
15	3
15	4
12	16
16	5
16	6
11	17
17	18
18	7
18	8
17	19
19	9
19	10

Feature Matrix

	Anc	Des1	Des2
1	0.152	0	0
2	0.152	0	0
3	0.535	0	0
4	0.535	0	0
5	2.52	0	0
6	2.52	0	0
7	0.453	0	0
8	0.453	0	0
9	0.146	0	0
10	0.146	0	0
11	0	0.397	3.267
12	0.397	1.686	2.082
13	1.686	2.765	2.382
14	2.765	0.152	0.152
15	2.382	0.535	0.535
16	2.082	2.52	2.52
17	3.267	1.28	1.587
18	1.28	0.453	0.453
19	1.587	0.146	0.146

Branching Times

	Time
11	5
12	4.603
13	2.917
16	2.52
17	1.733
15	0.535
18	0.453
14	0.152
19	0.146

Summary Stats (first 16)

Stat	Value
area_per_pair	5.378
average_leaf_depth	3.4
avg_ladder	0
avg_vert_depth	2.632
b1	6.333
b2	3.25
beta	10
blum	6.004
cherries	5
colless	4
colless_corr	0.111
colless_quad	8
crown_age	5
diameter	7
double_cherries	2
eigen_centrality	0.436
…	…

Figure 3.12: Example of a simulated phylogenetic tree and its derived input data for the neural networks. The first panel is a simulated tree with extant taxa (tips) under the diversity-dependent diversification scenario, its nodes (including root, internal and tip nodes) are labeled. The second panel shows the corresponding adjacency list, in which each parent-child relationship is presented in a row by an "Start → End" format. The third panel shows the corresponding feature matrix, with one row per node and three columns ("Anc", "Des1", "Des2") corresponding to the branch - length values toward the parent and up to two descendant nodes, the row indices are only for visual assistance, not included in the input data. The fourth panel shows branching times for each internal node, listing node ages, the row indices are not included. The fifth panel shows a subset of the first 16 summary statistics computed, the final row ("...") indicates continuation of the full statistic vector, the "Stat" column is not included in the input data.

D) Total Loss

Total loss comprises three key components: Huber loss, link prediction loss and entropy of regularization. Huber loss was used for optimizing regression accuracy while the remaining components focused on alleviating a possible issue where GNN can be hard to train, if incorporating the differentiable pooling method [140].

The Huber loss [233] for vectors y and $\hat{y}$, each with n elements, computed as the average loss across all elements, is given by:

$$L_\delta(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2 & \text{for } |y_i - \hat{y}_i| \le \delta, \\ \delta(|y_i - \hat{y}_i| - \frac{1}{2}\delta) & \text{otherwise,} \end{cases} \tag{3.2}$$

where $\mathbf{y}$ is the true value vector comprising the ground truth parameters used for simulating a phylogenetic tree, $\hat{\mathbf{y}}$ is the predicted value vector comprising the parameter predictions, y_i and $\hat{y}_i$ are the i-th elements of $\mathbf{y}$ and $\hat{\mathbf{y}}$ respectively, n is the number of elements in the vectors $\mathbf{y}$ and $\hat{\mathbf{y}}$ and δ is the threshold parameter that defines the transition from squared to linear loss (here loss refers to the difference between ground truth and predicted values). In our research, we set $\delta = 0.8$ for all the training sessions, making the neural networks more sensitive to smaller errors and more robust to outliers .

The total loss L is given by

$$L = L_\delta(\mathbf{y}, \hat{\mathbf{y}}) + L_{\text{LP}} + L_{\text{E}}, \tag{3.3}$$

where L_{LP} is the link prediction loss and L_{E} is the entropy of regularization, see Ying et al. [140] for their definitions.

E) Neural Network Architecture

For the graph neural network, we used GraphSAGE [234], a sample-and-aggregate graph convolutional neural network, to capture a graph-level representation. GraphSAGE has achieved strong performance of learning from large graphs. We use graph neural network (GNN) to refer to the graph neural network approach which incorporates GraphSAGE.

GNN is mainly assembled from five GNN modules (see Figure 3.1-C for five blocks of boxes in yellow and orange colors). Each module comprises the same number of GraphSAGE operators [234], where the number of layers (GraphSAGE operators, as illustrated by the number of combined boxes within each GNN modules in Figure 3.1-C) $N_{\mathscr{L}} = 1, 2, \ldots, 6$. Each GNN operator is accompanied by a Batch Normalization for 1D Inputs (BatchNorm1d, not shown in Figure 3.1) operator [235] and then a Gaussian Error Linear Units (GELU, as illustrated by the orange bands within the yellow boxes in Figure 3.1-C) activation function [236]. The GraphSAGE operators facilitate the convolution operation over graphs, capturing both local node features and their neighborhood information. The BatchNorm1d operator is commonly employed in neural networks to stabilize and accelerate the training process. The GELU activation layer is used for introducing non-linearity into the data. Learned features from all the GraphSAGE operators within a module are collected and concatenated. Larger $N_{\mathscr{L}}$ will result in the GNN modules to aggregate information into each node from its more distantly connected neighbors. According to our experiments, the optimal case is $N_{\mathscr{L}} = 2$, all figures and results relating to GNN were reported on the optimal case.

The graph-learning process also involves graph coarsening operations. We incorporated the differentiable pooling (DiffPool hereafter) technique to better learn hierarchical representations of the graphs. DiffPool can aggregate graph nodes into clusters after each operation. It facilitates graph coarsening and captures intricate hierarchical structure, which makes it particularly suitable for graph-level tasks [140]. In the first coarsening operation, the graph data inputs are passed to two GNN modules (pooling and embedding, see Figure 3.1-C for the blocks marked as "GNN pool1" and "GNN embed1"). The pooling group reduces the graph size, while the embedding group captures the node features. The filtered data from each GraphSAGE operator are concatenated (see Figure 3.1-C for the blocks of boxes marked as "concat1") then passed to a DiffPool layer (see Figure 3.1-C for the red box marked as "diff-pool1"), which finalizes the first coarsening operation. The second coarsening operation is applied in the same way as the first (as represented by "GNN pool1", "GNN embed2", "concat2" in Figure 3.1-C), and the outputs from the second DiffPool layer ("diff-pool2" in Figure 3.1-C) are passed to the final (fifth) GNN module ("GNN embed3" in Figure 3.1-C). The nodes in a graph are dynamically clustered and reduced after each coarsening operation. The coarsening ratio at each operation is determined by a pre-set DiffPool pooling ratio. Let $N_{\text{coarsened}}$ represent the number of nodes in the coarsened graph and N_{original} the number of nodes in the original graph. The DiffPool pooling ratio ρ_{pool} is given by $\rho_{\text{pool}} = \frac{N_{\text{coarsened}}}{N_{\text{original}}}$. Throughout the study, we used a manually optimized value $\rho_{\text{pool}} = 0.25$. This is a manually optimized hyper-parameter.

After the final GNN module, the outputs are concatenated ("concat3" in Figure 3.1-C) and transformed by a global mean pooling operation (red ball "M" in Figure 3.1-C) to create a

final graph representation. This graph representation is passed to a readout layer group ("readout" as represented by light blue boxes in Figure 3.1-C) consisting of two linear layers to perform graph-level regression which ultimately outputs a vector of n predicted parameters ("pred" as represented by a purple box in Figure 3.1-C). Only the first linear layer is followed by GELU (see the orange band of the first linear layer). All the linear layers incorporate dropout operations with a pre-set dropout ratio to prevent over-fitting and to utilize as many neuron connections as possible. Let ρ_{dropout} represent the probability p of disabling a connection between an input node and a hidden node of a linear layer in each epoch. The dropout ratio ρ_{dropout} is simply given by $\rho_{\text{dropout}} = p$. Throughout the study, we used a commonly picked value $\rho_{\text{dropout}} = 0.5$. This is a manually optimized hyper-parameter.

DNN's major component is a stack comprises 5 linear layers ("DNN stack" in Figure 3.1-A), each followed by a BatchNorm1D (not shown in figure) and a GELU (the orange band within the boxes). All the linear layers within the stack incorporate dropout operations with $\rho_{\text{dropout}} = 0.5$. Learned features from all the linear layers within the stacks are collected and concatenated ("concat" in Figure 3.1-A). A single linear readout layer ("readout" in Figure 3.1-A) outputs n predicted parameters ("pred" in Figure 3.1-A). According to our experiments, stacking more linear layers gives no substantial improvement to the performance.

LSTM's major component is a stack of 5 LSTM recurrent neural network layers ("LSTM stack" in Figure 3.1-B). The final hidden state from the last recurrent neural network layer is processed by a linear layer with $\rho_{\text{dropout}} = 0.5$ accompanied by a GELU ("linear" in Figure 3.1-B), then passed to a single linear readout layer ("readout" in Figure 3.1-B) that outputs n predicted parameters ("pred" in Figure 3.1-B). According to our experiments, stacking more recurrent neural network layers provides no substantial improvement to the performance.

The hyper-parameters not mentioned are set by their default values. The dimensions of the boxes do not map to any hyper-parameter settings, they are set for the best visual effect. The values below the boxes indicate their respective number of hidden neurons, their input and output neurons are not shown in the figure, they can be found in the configuration files in our GitHub repository `eveGNN` [221].

F) Ensemble Learning

With bagging, we trained GNN, DNN and LSTM independently (”GNN”, ”DNN” and ”LSTM” blocks of boxes in Figure 3.3-Bagging), translated their outputs to parameter predictions through their own readout layers (three ”readout” boxes next to the neural networks and three ”pred” boxes next to the readout layers in Figure 3.3-Bagging) and then aggregated the predictions (red ball ”A” in Figure 3.3-Bagging). We experimented with four aggregation methods: taking the mean, median, max and min values among the three predictions. We also recorded the individual predictions without aggregation.

With stacking, we trained GNN, DNN and LSTM simultaneously (”GNN”, ”DNN” and ”LSTM” blocks of boxes in Figure 3.3-Stacking) but without their own readout layers. We combined the features from DNN, the LSTM’s final hidden state, and GNN’s graph representation and fed to a meta-learner (”meta-learner” in Figure 3.3-Stacking) comprising linear neural network layers that learns to best readout parameter predictions from these combined outputs.

With boosting, there can be different pathways. In our illustration, GNN, DNN and LSTM were trained sequentially to iteratively correct residuals. For example, firstly, the GNN (”GNN” in Figure 3.3-Boosting) is trained from the graphs to make the initial predictions (”readout” and then ”pred0” in Figure 3.3-Boosting) and from predicted and ground truth values of the parameters we computed the residuals (”res1” in Figure 3.3-Boosting); secondly, the DNN (”DNN” in Figure 3.3-Boosting) is trained to predict these residuals from the summary statistics (”readout” and then ”pred-res1” in Figure 3.3-Boosting), learning to correct the GNN’s errors; lastly, the LSTM (”LSTM” in Figure 3.3-Boosting) is trained to predict the residuals of the residuals (”readout” and then ”pred-res2” in Figure 3.3-Boosting), which is the initial predictions minus the predicted residuals by the DNN, from branching times, to further improve the predictive accuracy. Finally, we subtracted the two residual terms from the initial predictions (red ball ”S” in Figure 3.3-Boosting) to make the corrected predictions.

G) Comparison between MLE Optimizers

On the phylogenies from the diversity-dependent diversification (DDD) dataset, we compared between three approaches: "Simplex", "Subplex" and "DEoptim". Simplex is a derivative-free optimization method that uses a simplex of solutions to iteratively explore and adjust within the parameter space, suitable for non-smooth objective functions but potentially slow for high-dimensional problems [237]. Subplex is an enhancement of the Simplex method, Subplex breaks high-dimensional optimization into smaller subproblems, each optimized using Simplex techniques, providing improved efficiency and effectiveness in complex parameter landscapes [238]. DEoptim (Differential Evolution) is a more recent population-based algorithm that applies evolutionary strategies such as mutation, crossover, and selection to efficiently navigate and optimize multimodal and complex objective functions [239].

All three MLE methods encountered consistent optimization challenges, likely due to numerical issues related to machine precision limits or unexpected negative values during matrix operations. From a random sample of 2000 DDD phylogenies, the completion rates for each method were as follows: Simplex achieved 1966 completions from true parameter starts and 1910 from random starts; Subplex completed 1681 from true starts and 1612 from random starts; DEoptim finished 1122 from true starts and 999 from random starts. It is more difficult to estimate parameters from random starts, comparing to true starts.

For all the three optimization approaches, -1 will be returned as a parameter estimation if the likelihood becomes too small in the searching process. This means that the algorithm cannot find optima given the initial starting point of the parameters. It is highly possible that the unfinished estimations consisted of inaccurate or even -1 values. The comparison between MLE optimizers can be skewed due to less completion rate of the Subplex and DEoptim results.

In instances where optimization starting points were randomly set, a significant number of outcomes were trapped at local optima, failing to achieve global optima and often leading to inaccurate parameter estimates. This issue was less prevalent when starting points were the true parameters. For visual reference, see Figure 3.13 and Figure 3.14. Notably, in DEoptim's best-case scenarios, estimation accuracy deteriorated significantly on larger phylogenies, as shown in the last row of Figure 3.14.

In the best-case scenarios, all MLE methods tended to yield more accurate estimates on larger phylogenies, while in naive cases, larger phylogenies posed challenges. However, all MLE methods generally performed better with larger trees, and all displayed similar trends of bias. We calculated the strength of the carrying capacity effect with the formula $(\lambda - \mu)/K$, where λ is the true speciation rate, μ is the true extinction rate, and K is the true carrying capacity.

Subplex was the fastest among the tested algorithms, Simplex and DEoptim were slower. Simplex, although slower, completed the most computations and did not show a definitive performance disadvantage compared to Subplex or DEoptim. For this reason, in all comparisons between MLE and neural network methods, we consistently used results from the Simplex optimizer due to data coverage and reliability.

With the Simplex optimizer, we also compared the estimates between using two integration methods: one is matrix exponentiation ("analytical") and the other is a numerical integrator ("odeint::runge_kutta_cash_karp54"). To ensure that the MLE estimates shown in the figures are reliable, we always adopted the results of the numerical integrator and excluded the data points on which the discrepancy of estimates between two integration methods is larger than 10% relatively. See Table 3.2 for details.

Table 3.2: The distribution of relative differences of estimated parameters between using matrix exponentiation ("analytical") and numerical integrator ("odeint::runge_kutta_cash_karp54") with the Simplex optimizer. The "Case" column indicates the maximum-likelihood estimation scenarios, "Best" represents using true speciation rate, true extinction rate and $K = \infty$ as starting points, "Naive" represents using true speciation rate, true extinction rate and $K = 10000$ as starting points. The "Parameter" column indicates each of the estimated parameters, "K" is the carrying capacity, "λ" is the speciation rate and "μ" is the extinction rate. The rest of the columns indicate the ranges of measured relative differences between estimates of the two integration methods. The percentages under these columns indicate the proportions of data having the relative discrepancies smaller than the values indicated by the column names.

Case	Parameter	$\leqslant$ 10%	$\leqslant$ 1%	$\leqslant$ 0.1%	$\leqslant$ 0.01%
Best	K	89.5%	78.6%	73.2%	64.6%
Best	λ	83.9%	78.1%	73.7%	63.2%
Best	μ	83.1%	77.1%	72.6%	61.5%
Naive	K	76.9%	68.9%	64.3%	56.0%
Naive	λ	71.8%	67.3%	64.1%	54.3%
Naive	μ	71.0%	67.0%	62.9%	52.6%

(Figures on next page.)

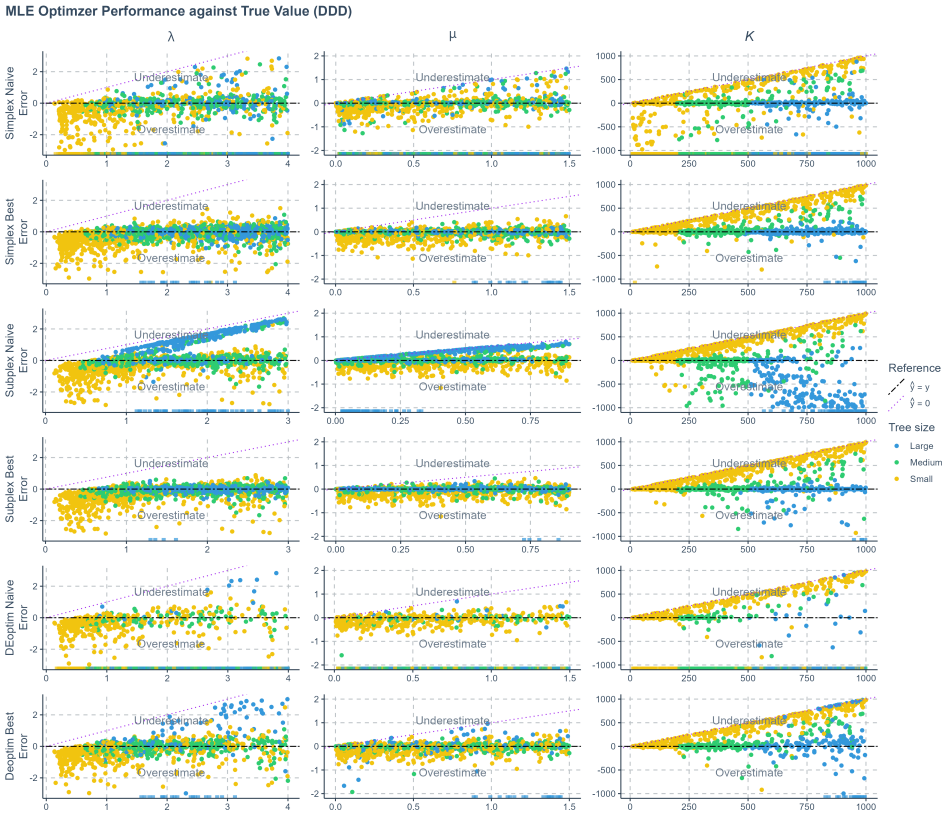


Figure 3.13: Error of maximum likelihood estimation using Simplex, Subplex and DEoptim optimzers applied to phylogenies simulated under a diversity-dependent diversification scenario, against true values. For each optimzer there were two cases. The best case (Best) refers to using the true parameter values as the starting points for the searching process; the naive case (Naive) refers to using randomly sampled values from the true parameter space as the starting points. The errors shown (y-axis) are the differences between the true parameters (x-axis) used to simulate the phylogenies and the values estimated by each method. Each row represents a method, and each column corresponds to the results for one specific parameter. Phylogenies are categorized based on their size: yellow for small phylogenies with fewer than 200 nodes (including root, internal, and tip nodes), green for medium-sized phylogenies with 200 to 500 nodes, and blue for large phylogenies with more than 500 nodes. Data points close to purple dotted lines ($\hat{y} = 0$) in MLE result panels indicate near-zero estimates. Black two-dash lines indicate accurate estimates ($\hat{y} = y$ where y denotes the true parameter value). Small squares spreading along the x-axis signify optimization failures. Extremely deviating estimates are not shown in the figure. λ: Speciation rate. μ: extinction rate. K: carrying capacity.

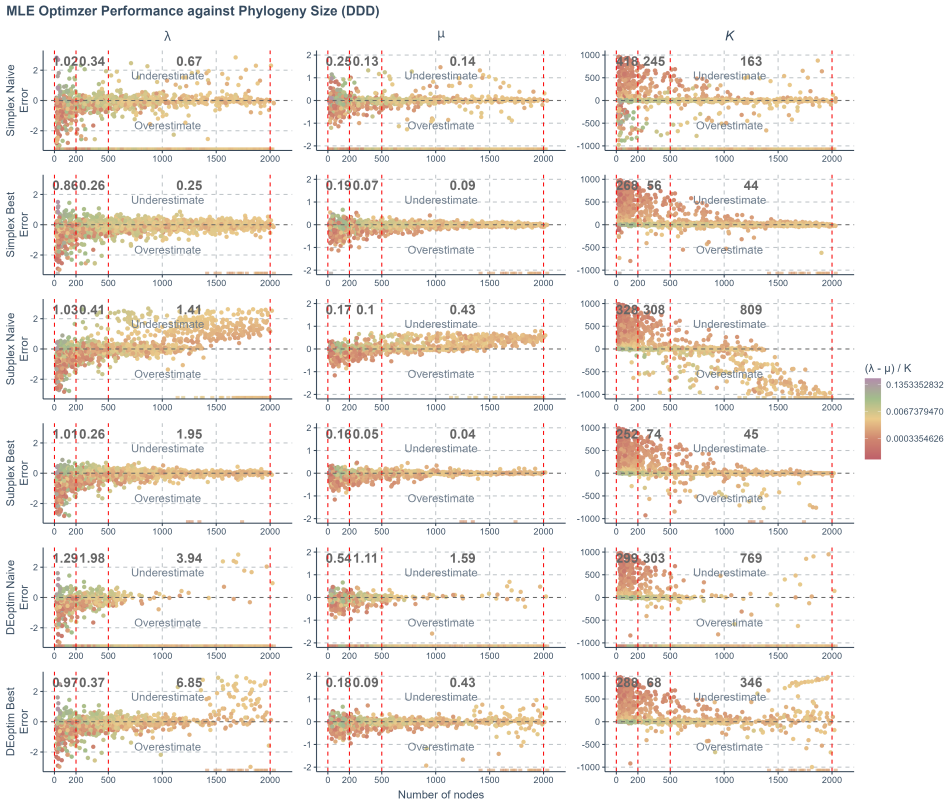


Figure 3.14: Error of maximum likelihood estimation using Simplex, Subplex and DEoptim optimzers applied to phylogenies simulated under a diversity-dependent diversification scenario, against the total number of nodes (including root, internal, and tip nodes) in the phylogenies. For each optimzer there were two cases. The best case (Best) refers to using the true parameter values as the starting points for the searching process; the naive case (Naive) refers to using randomly sampled values from the true parameter space as the starting points. The errors shown (y-axis) are the differences between the true parameters used to simulate the phylogenies and the values estimated by each method. Each row represents a method, and each column corresponds to the results for one specific parameter. Phylogenies are categorized based on their size into three sectors within each panel, separated by four vertical red dashed lines. From left to right, the sectors are: small phylogenies with fewer than 200 nodes, medium-sized phylogenies with 200 to 500 nodes, and large phylogenies with more than 500 nodes. The values shown in black within each sector are the mean absolute prediction errors of all data points in the sectors. Color coding: The color of the data points illustrates the strength of the carrying capacity effect, calculated as $(\lambda - \mu)/K$. The color gradient transitions from red to purple, indicating increasing strength of the effect. This scale is transformed using log10 for clearer visual differentiation. Small squares spreading along the x-axis signify optimization failures. Extremely deviating estimates are not shown in the figure. X-axis: Size of the phylogenies. Y-axis: Error. λ: Speciation rate. μ: extinction rate. K: carrying capacity.

H) Estimation Uncertainty for Empirical Trees

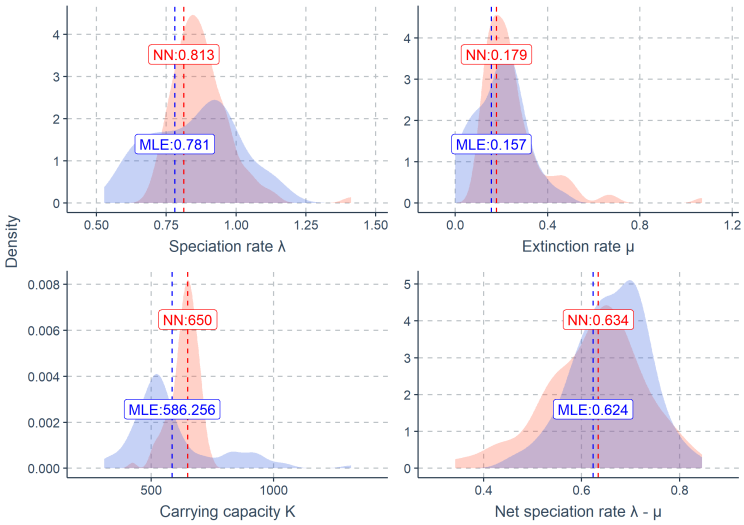
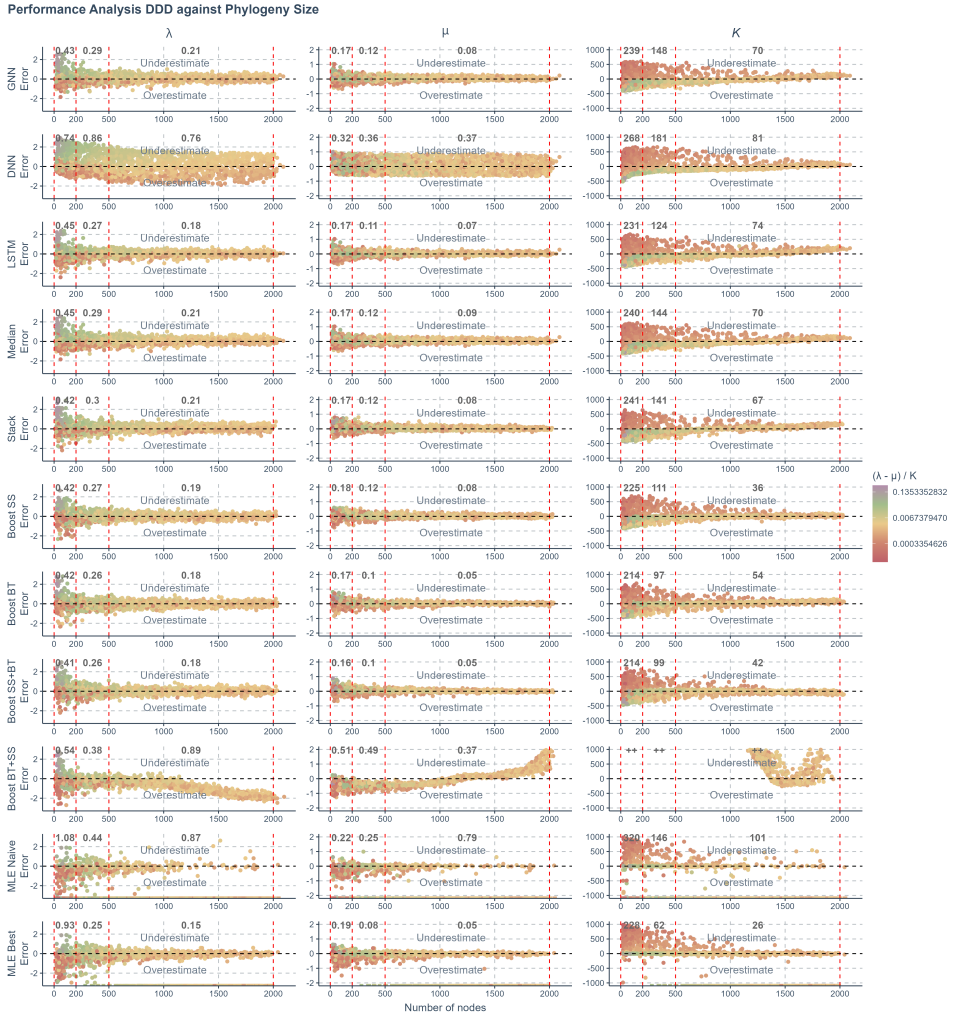


Figure 3.15: The neural network estimation uncertainty for a bird phylogeny (Furnariidae). The parameters are estimated using a pre-trained neural network (Boost BT, boosting strategy that corrects GNN results using LSTM) under a diversity-dependent diversification scenario. For reference, maximum likelihood estimation (MLE) is also used to estimate the same parameters. Each panel shows one parameter's estimates using neural network and MLE methods with their uncertainties. The red dashed lines with red numbers indicate the estimates by the neural network method. The blue dashed lines with blue numbers indicate the estimates by the MLE method. Each pink area indicates the density distribution of a neural network estimate from 1000 bootstrap-simulated phylogenies, showing the uncertainty of neural network. Each blue area indicates the density distribution of an MLE estimate from the same set of simulated phylogenies, showing the uncertainty of MLE. X-axis: Parameter (Estimate) values. Y-axis: Density. λ: Speciation rate. μ: Extinction rate. K: Carrying capacity. $\lambda - \mu$: Net speciation rate.

The rest of the figures of neural network estimation uncertainty on the empirical phylogenies under the DDD scenario can be found at:

https://github.com/EvoLandEco/eveGNN/tree/master/uncertainty

I) Results under the Diversity-Dependent Diversification Scenario



(Caption on next page.)

Figure 3.16: *(Figure on previous page.)* The prediction error of various methods applied to phylogenies simulated under a diversity-dependent diversification scenario, against the total number of nodes (including root, internal, and tip nodes) in the phylogenies. The errors shown are the differences between the true parameters used to simulate the phylogenies and the values predicted or estimated by each method. Each row represents a method, and each column corresponds to the results for one specific parameter. Phylogenies are categorized based on their size into three sectors within each panel, separated by four vertical red dashed lines. From left to right, the sectors are: small phylogenies with fewer than 200 nodes, medium-sized phylogenies with 200 to 500 nodes, and large phylogenies with more than 500 nodes. The values shown in black within each sector are the mean absolute prediction errors of all data points in the sectors. Color coding: The color of the data points illustrates the strength of the carrying capacity effect, calculated as $(\lambda - \mu)/K$. The color gradient transitions from red to purple, indicating increasing strength of the effect. This scale is transformed using log10 for clearer visual differentiation. GNN: Predictions obtained by the graph neural network using the phylogenies transformed to graph format. DNN: Predictions by the dense neural network using summary statistics. LSTM: Predictions by the long short-term memory recurrent neural network using branching times. Median: Bagging strategy that takes the median value of the predictions from GNN, DNN, and LSTM. Stack: Stacking strategy that utilizes a meta-learner to integrate results from GNN, DNN, and LSTM. Boost SS: Boosting strategy that corrects GNN results using DNN. Boost BT: Boosting strategy that corrects GNN results using LSTM. Boost SS+BT: Sequential correction of GNN errors first using DNN, followed by LSTM. Boost BT+SS: Sequential correction of GNN errors first using LSTM, followed by DNN. MLE Naive: Maximum Likelihood Estimation results using random starting points for parameter optimization. MLE Best: MLE results using the true parameter values as the starting points for optimization. In the MLE result panels, small squares spreading along the x-axis signify optimization failures. Due to significantly lower accuracy, other aggregation methods from the bagging strategy are not displayed on the plot. X-axis: Size of the phylogenies. Y-axis: Error. λ: Speciation rate. μ: Extinction rate. K: Carrying capacity.

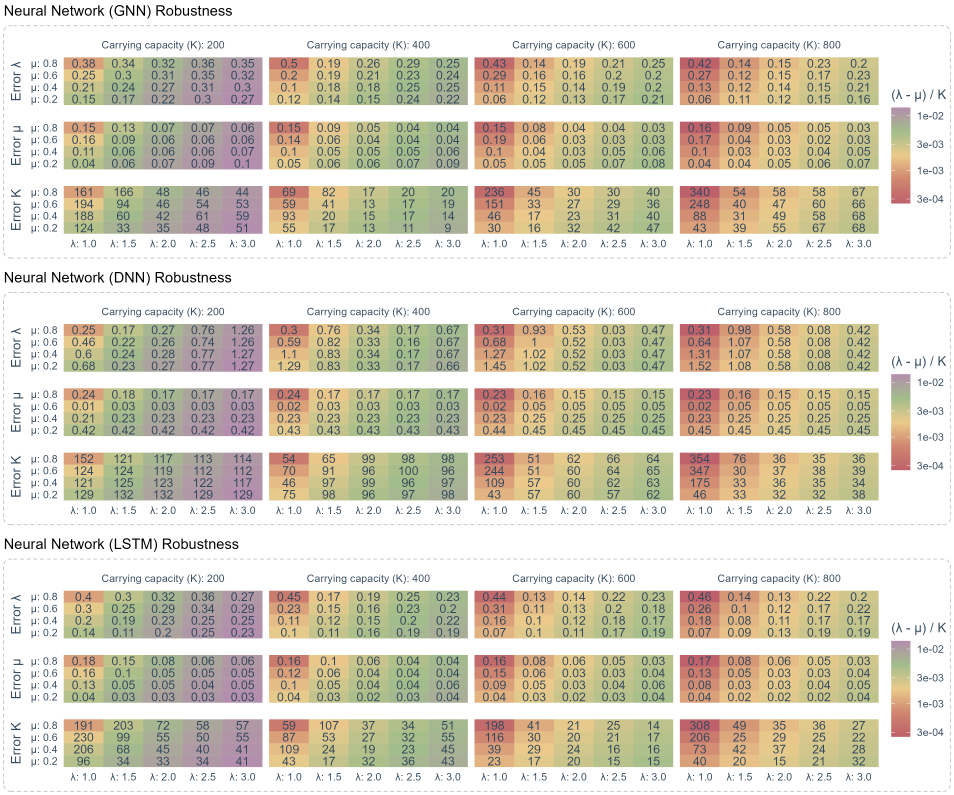
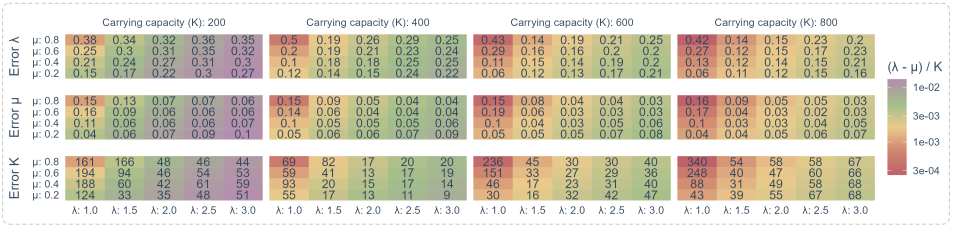
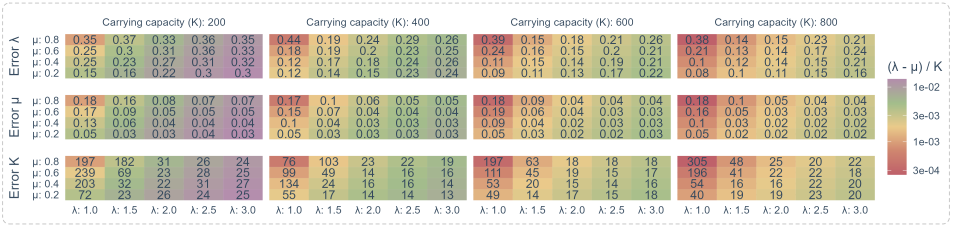


Figure 3.17: Robustness comparison between graph neural network (GNN), dense neural network (DNN) and long short-term memory recurrent neural network (LSTM) when operating independently. Same structure as the previous figure.

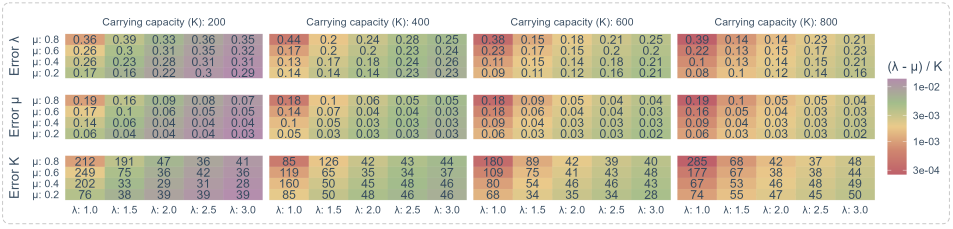
Neural Network (GNN) Robustness



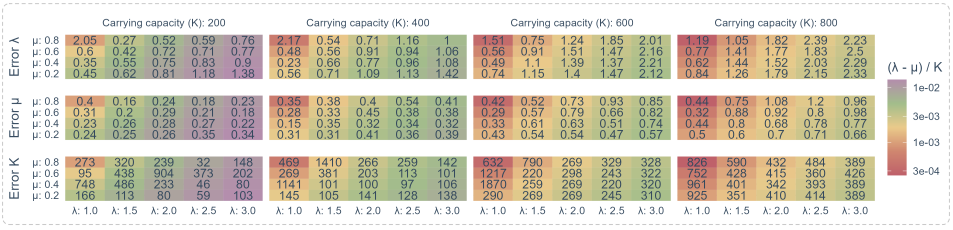
Neural Network (Boosting BT) Robustness



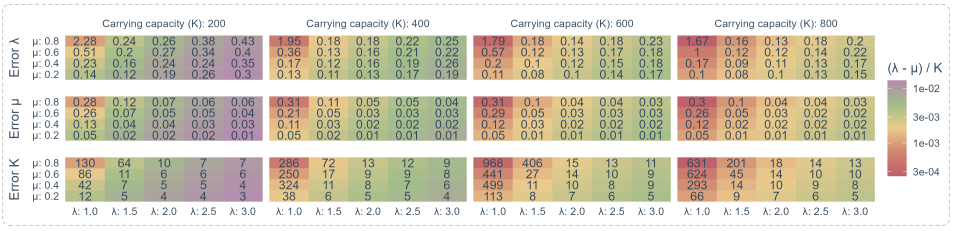
Neural Network (Boosting SS + BT) Robustness



Maximum Likelihood Estimation (Naive Case) Robustness



Maximum Likelihood Estimation (Best Case) Robustness



(Caption on next page.)

Figure 3.18: *(Figure on previous page.)* The robustness of neural network and maximum likelihood estimation was assessed on 80 sets of phylogenies, each containing 1000 trees randomly simulated under a diversity-dependent diversification scenario, employing identical parameter settings but varied in size, topology, and structure. Each segment delineated by dashed lines corresponds to distinct methods. Each column within a segment is associated with a specific carrying capacity (K) used in the simulation of the phylogenies. Each row within a segment details the mean absolute errors between the true and estimated values of a specific parameter, with parameter names labeled on the left side of each row. GNN: Graph neural network. Boosting BT: Graph neural network with long short-term memory recurrent neural network correcting its residuals using branching times. Boosting SS + BT: Graph neural network with dense neural network and long short-term memory recurrent neural network correcting residuals sequentially using summary statistics and branching times. Naive Case: Maximum likelihood estimation using random initial parameter as the starting point. Best Case: Maximum likelihood estimation using true parameter as the starting point. X-axis: Represents the true speciation rate (λ) used to simulate phylogenies. Y-axis: Represents the true extinction rate (μ) used to simulate phylogenies. Cell Content: The numbers displayed within each heatmap cell indicate the mean absolute error for a parameter, given the specific λ, μ and K settings. Color Coding: The background color of each cell illustrates the strength of the carrying capacity effect, calculated as $(\lambda - \mu)/K$. The color gradient transitions from red to purple, indicating increasing strength of the effect. This scale is transformed using log10 for clearer visual differentiation. Note that the numerical values within the cells are not mapped to the background colors. For a detailed reference to the effect strength values corresponding to the background colors, refer to the figure legends.

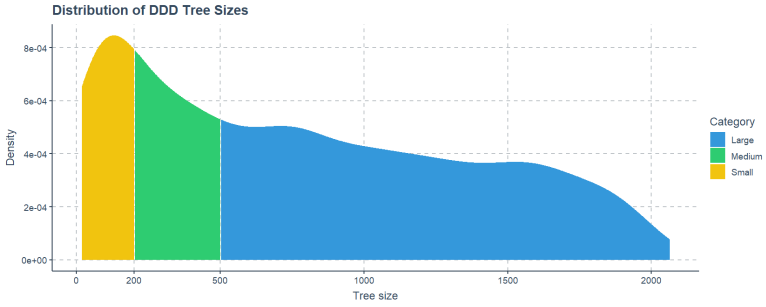


Figure 3.19: The density distribution of phylogeny sizes under the diversity-dependent diversification (DDD) scenario. The colors of the areas under the density curve indicate the three categories used in our analyses. Yellow area: Small-sized phylogenies with less than 200 nodes (approx. 100 tips). Greed area: Medium-sized phylogenies with more than 200 nodes and less than 500 nodes (approx. 250 tips). Blue area: Large-sized phylogenies with more than 500 nodes.

J) Results under the Birth–Death Scenario

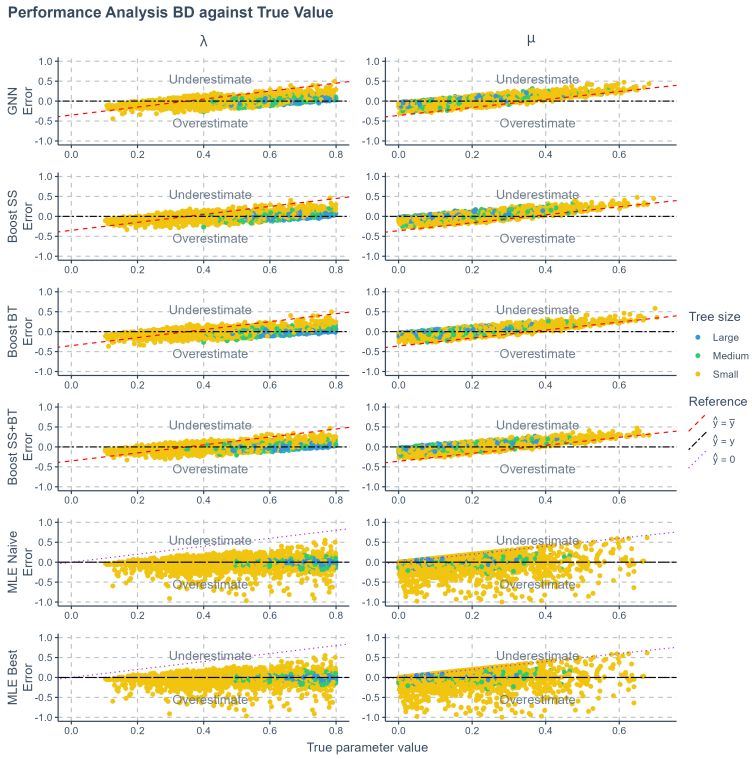


Figure 3.20: The prediction error (absolute error) of various methods applied to phylogenies simulated under a birth–death scenario, against true values. The errors shown are the differences between the true parameters used to simulate the phylogenies and the values predicted or estimated by each method. Each row represents a method, and each column corresponds to the results for one specific parameter. Phylogenies are categorized based on their size: yellow for small phylogenies with fewer than 200 nodes (including root, internal, and tip nodes), green for medium-sized phylogenies with 200 to 500 nodes, and blue for large phylogenies with more than 500 nodes. GNN: Predictions obtained by the graph neural network using the phylogenies. Boost SS: Boosting strategy that corrects GNN results using DNN. Boost BT: Boosting strategy that corrects GNN results using LSTM. Boost SS+BT: Sequential correction of GNN errors first using DNN, followed by LSTM. MLE Naive: Maximum Likelihood Estimation results using random starting points for parameter optimization. MLE Best: MLE results using the true parameter values as the starting points for optimization. Red dashed lines in panels representing neural network results indicate the mid-points of the parameter spaces ($\hat{y} = \bar{y}$ where $\hat{y}$ denotes a estimated parameter and $\bar{y}$ denotes the mid-point of the parameter space). Data points close to purple dotted lines ($\hat{y} = 0$) in MLE result panels indicate near-zero estimates. Black two-dash lines indicate accurate estimates ($\hat{y} = y$ where y denotes the true parameter value). X-axis: True parameter values. Y-axis: Error, or difference between true and predicted values. λ: Speciation rate. μ: Extinction rate.

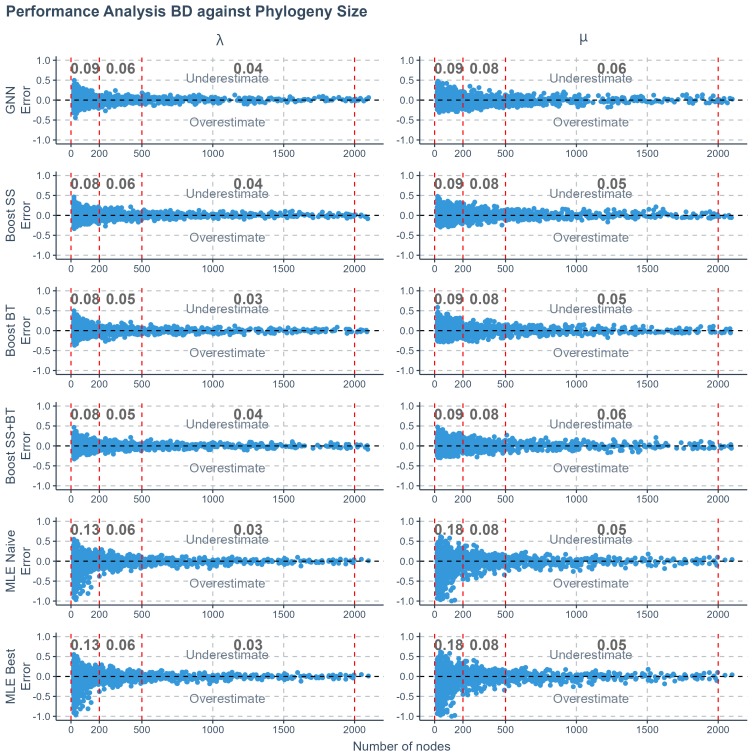


Figure 3.21: The prediction error (absolute error) of various methods applied to phylogenies simulated under a birth–death scenario, against the total number of nodes in the phylogenies. The errors shown are the differences between the true parameters used to simulate the phylogenies and the values predicted or estimated by each method. Each row represents a method, and each column corresponds to the results for one specific parameter. Phylogenies are categorized based on their size into three sectors within each panel, separated by four vertical red dashed lines. From left to right, the sectors are: small phylogenies with fewer than 200 nodes (including root, internal, and tip nodes), medium-sized phylogenies with 200 to 500 nodes, and large phylogenies with more than 500 nodes. GNN: Predictions obtained by the graph neural network using the phylogenies. Boost SS: Boosting strategy that corrects GNN results using DNN. Boost BT: Boosting strategy that corrects GNN results using LSTM. Boost SS+BT: Sequential correction of GNN errors first using DNN, followed by LSTM. MLE Naive: Maximum Likelihood Estimation results using random starting points for parameter optimization. MLE Best: MLE results using the true parameter values as the starting points for optimization. X-axis: Size of the phylogenies. Y-axis: Error. λ: Speciation rate. μ: Extinction rate.

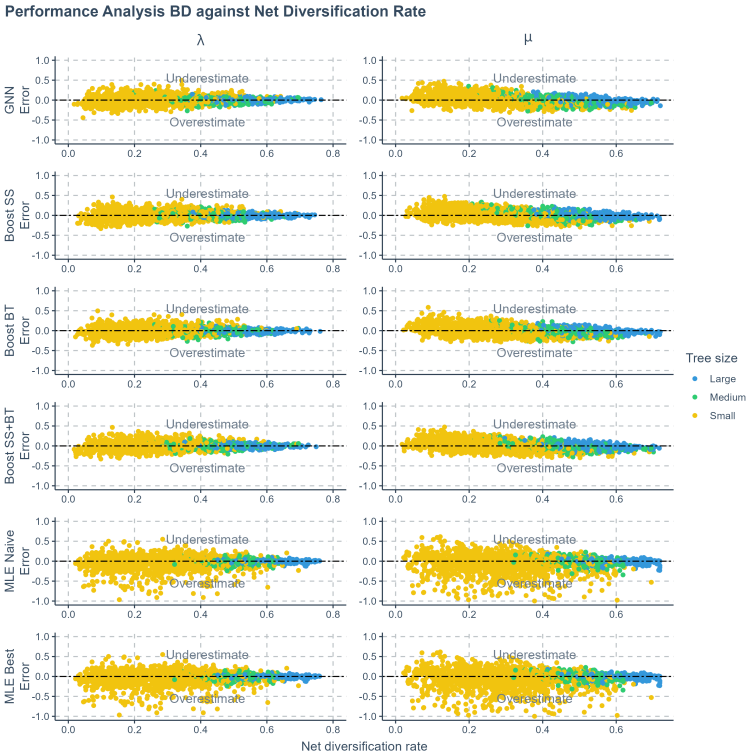


Figure 3.22: The prediction error (absolute error) of various methods applied to phylogenies simulated under a birth–death scenario, against net diversification rate ($\lambda - \mu$). The errors shown are the differences between the true parameters used to simulate the phylogenies and the values predicted or estimated by each method. Each row represents a method, and each column corresponds to the results for one specific parameter. Phylogenies are categorized based on their size: yellow for small phylogenies with fewer than 200 nodes (including root, internal, and tip nodes), green for medium-sized phylogenies with 200 to 500 nodes, and blue for large phylogenies with more than 500 nodes. GNN: Predictions obtained by the graph neural network using the phylogenies. Boost SS: Boosting strategy that corrects GNN results using DNN. Boost BT: Boosting strategy that corrects GNN results using LSTM. Boost SS+BT: Sequential correction of GNN errors first using DNN, followed by LSTM. MLE Naive: Maximum Likelihood Estimation results using random starting points for parameter optimization. MLE Best: MLE results using the true parameter values as the starting points for optimization. Black two-dash lines indicate accurate estimate. X-axis: True parameter values. Y-axis: Error, or difference between true and predicted values. λ: Speciation rate. μ: Extinction rate.

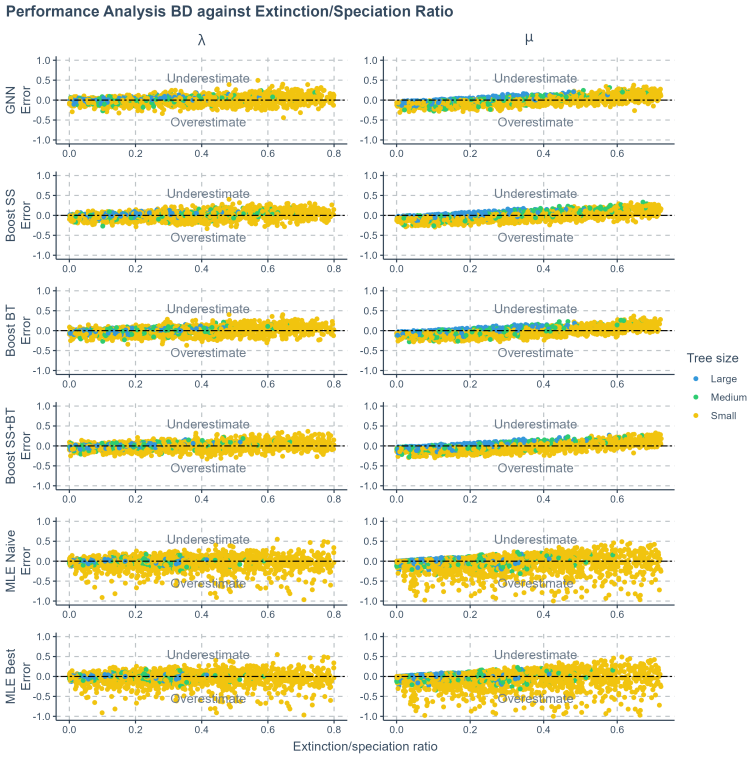


Figure 3.23: The prediction error (absolute error) of various methods applied to phylogenies simulated under a birth–death scenario, against extinction/speciation ratio (μ/λ). The errors shown are the differences between the true parameters used to simulate the phylogenies and the values predicted or estimated by each method. Each row represents a method, and each column corresponds to the results for one specific parameter. Phylogenies are categorized based on their size: yellow for small phylogenies with fewer than 200 nodes (including root, internal, and tip nodes), green for medium-sized phylogenies with 200 to 500 nodes, and blue for large phylogenies with more than 500 nodes. GNN: Predictions obtained by the graph neural network using the phylogenies. Boost SS: Boosting strategy that corrects GNN results using DNN. Boost BT: Boosting strategy that corrects GNN results using LSTM. Boost SS+BT: Sequential correction of GNN errors first using DNN, followed by LSTM. MLE Naive: Maximum Likelihood Estimation results using random starting points for parameter optimization. MLE Best: MLE results using the true parameter values as the starting points for optimization. Black two-dash lines indicate accurate estimates. X-axis: True parameter values. Y-axis: Error, or difference between true and predicted values. λ: Speciation rate. μ: Extinction rate.

K) Results under the Protracted Birth–Death Scenario

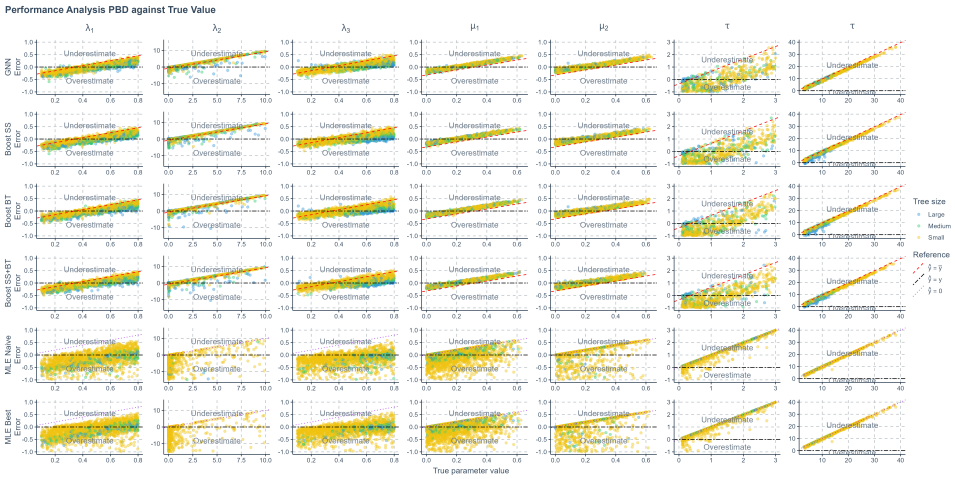


Figure 3.24: The prediction error (absolute error) of various methods applied to phylogenies simulated under a protracted birth–death scenario, against true values. The errors shown are the differences between the true parameters used to simulate the phylogenies and the values predicted or estimated by each method. Each row represents a method, and each column corresponds to the results for one specific parameter. Phylogenies are categorized based on their size: yellow for small phylogenies with fewer than 200 nodes (including root, internal, and tip nodes), green for medium-sized phylogenies with 200 to 500 nodes, and blue for large phylogenies with more than 500 nodes. GNN: Predictions obtained by the graph neural network using the phylogenies. Boost SS: Boosting strategy that corrects GNN results using DNN. Boost BT: Boosting strategy that corrects GNN results using LSTM. Boost SS+BT: Sequential correction of GNN errors first using DNN, followed by LSTM. MLE Naive: Maximum Likelihood Estimation results using random starting points for parameter optimization. MLE Best: MLE results using the true parameter values as the starting points for optimization. Red dashed lines in panels representing neural network results indicate the mid-points of the parameter spaces. Purple dotted lines in MLE result panels signify where estimated values are 0. X-axis: True parameter values. Y-axis: Error, or difference between true and predicted values. λ_1: Speciation initiation rate of the good species. λ_2: Speciation completion rate. λ_3: Speciation initiation rate of the incipient species. μ_1: Extinction rate of the good species. μ_2: Extinction rate of the incipient species. τ: Expected duration of speciation.

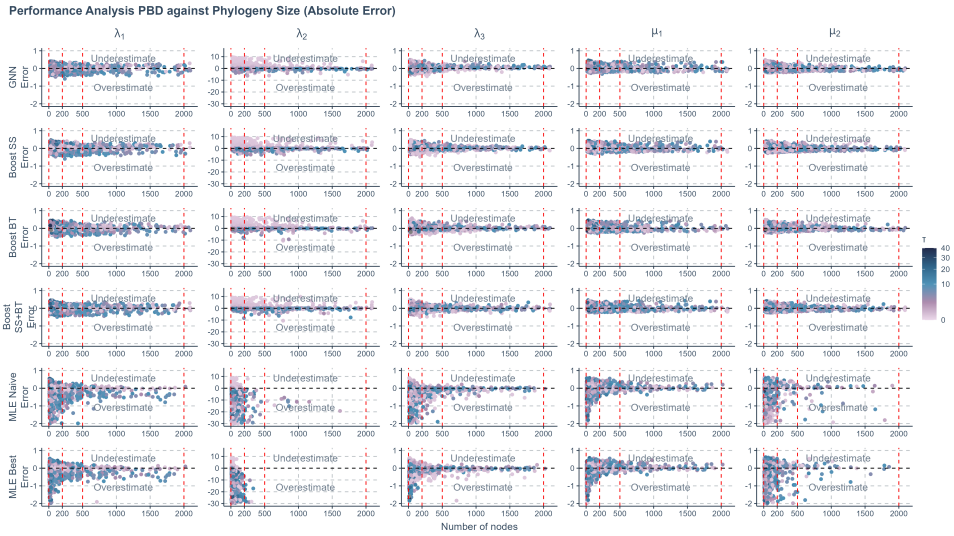


Figure 3.25: The prediction error (absolute error) of various methods applied to phylogenies simulated under a protracted birth–death scenario, against the total number of nodes in the phylogenies. The errors shown are the differences between the true parameters used to simulate the phylogenies and the values predicted or estimated by each method. Each row represents a method, and each column corresponds to the results for one specific parameter. Phylogenies are categorized based on their size into three sectors within each panel, separated by four vertical red dashed lines. From left to right, the sectors are: small phylogenies with fewer than 200 nodes (including root, internal, and tip nodes), medium-sized phylogenies with 200 to 500 nodes, and large phylogenies with more than 500 nodes. Color Coding: The color of the data points illustrates the expected duration of speciation. The color gradient transitions from light purple to dark blue, indicating increasing value of the duration. This scale is transformed using square root for clearer visual differentiation. GNN: Predictions obtained by the graph neural network using the phylogenies. Boost SS: Boosting strategy that corrects GNN results using DNN. Boost BT: Boosting strategy that corrects GNN results using LSTM. Boost SS+BT: Sequential correction of GNN errors first using DNN, followed by LSTM. MLE Naive: Maximum Likelihood Estimation results using random starting points for parameter optimization. MLE Best: MLE results using the true parameter values as the starting points for optimization. X-axis: Size of the phylogenies. Y-axis: Error. λ_1: Speciation initiation rate of the good species. λ_2: Speciation completion rate. λ_3: Speciation initiation rate of the incipient species. μ_1: Extinction rate of the good species. μ_2: Extinction rate of the incipient species. τ: Expected duration of speciation.

Figure 3.26: The prediction error (absolute error) of various methods applied to phylogenies simulated under a protracted birth–death scenario, against the true mean duration of speciation. The errors shown are the differences between the true parameters used to simulate the phylogenies and the values predicted or estimated by each method. Each row represents a method, and each column corresponds to the results for one specific parameter. Color Coding: The color of the data points illustrates the total number of nodes of the phylogenies. The color gradient transitions from light purple to dark blue, indicating increasing value of the node number. This scale is transformed using square root for clearer visual differentiation. GNN: Predictions obtained by the graph neural network using the phylogenies. Boost SS: Boosting strategy that corrects GNN results using DNN. Boost BT: Boosting strategy that corrects GNN results using LSTM. Boost SS+BT: Sequential correction of GNN errors first using DNN, followed by LSTM. MLE Naive: Maximum Likelihood Estimation results using random starting points for parameter optimization. MLE Best: MLE results using the true parameter values as the starting points for optimization. X-axis: Size of the phylogenies. Y-axis: Error. λ_1: Speciation initiation rate of the good species. λ_2: Speciation completion rate. λ_3: Speciation initiation rate of the incipient species. μ_1: Extinction rate of the good species. μ_2: Extinction rate of the incipient species. τ: Expected duration of speciation.

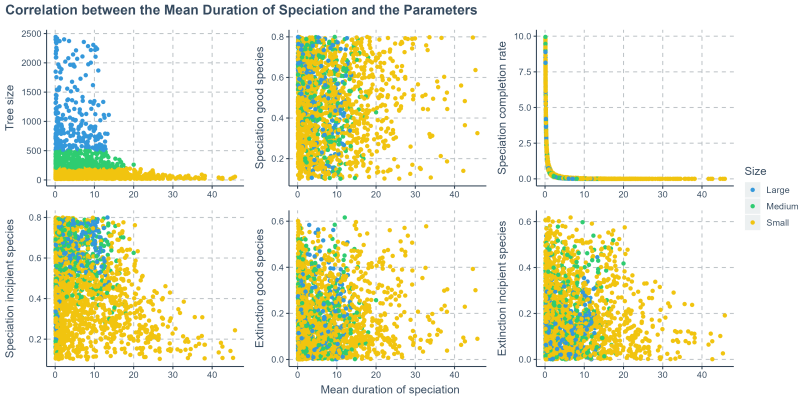


Figure 3.27: The correlation between the mean duration of speciation and the true parameter values under the protracted birth–death diversification scenario. Phylogenies are categorized based on their size: yellow for small phylogenies with fewer than 200 nodes (including root, internal, and tip nodes), green for medium-sized phylogenies with 200 to 500 nodes, and blue for large phylogenies with more than 500 nodes. X-axis: Mean duration of speciation. Y-axis: Tree size.

L) Comparison between Our Methods and Existing Methods

To benchmark our approaches against a convolutional architecture, we replaced the "Median" bagging panel in all performance figures related to the DDD scenario with a "CNN1D" panel reflecting the implementation of Voznica et al. [144]. This allows a direct side - by - side comparison between our neural networks and one of the best - performing 1D - CNNs in the literature.

Overall, the architecture by Voznica et al. [144] performed similarly and exhibited similar patterns as our approaches. CNN1D better recovered the carrying capacity effect strength than than GNN, DNN and LSTM alone, but lagged behind our boosting approaches–except for Boost BT+SS–in overall parameter prediction accuracy.

See the figures below for details.

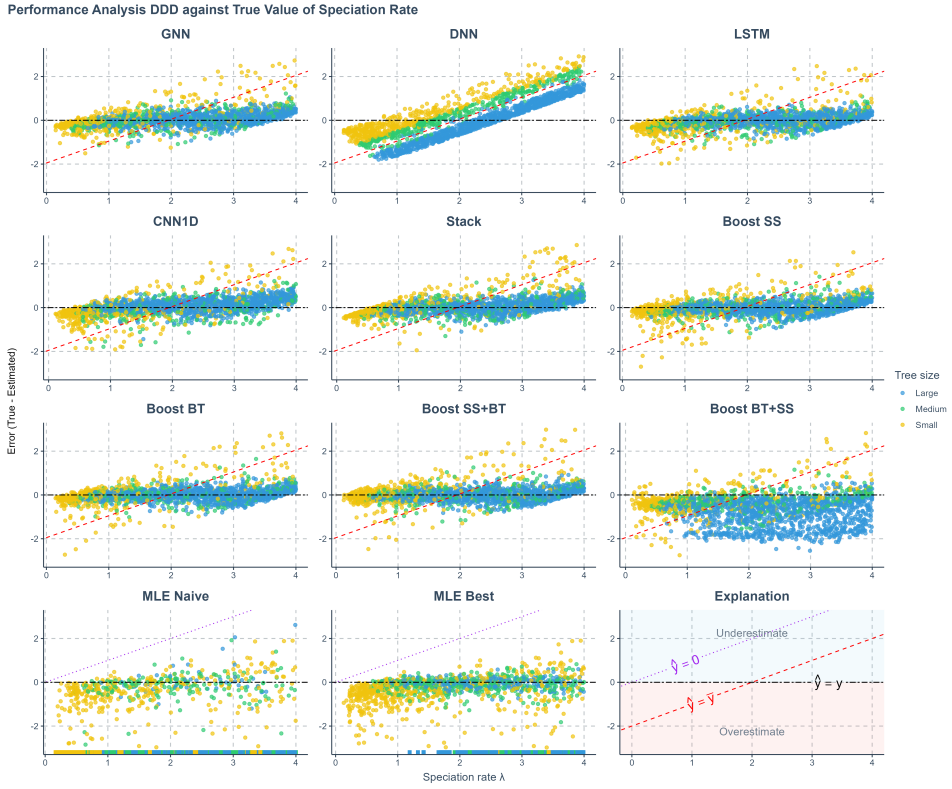


Figure 3.28: Prediction error of speciation rate plotted against true speciation rate under a diversity-dependent diversification scenario. Compared to the original figure in the results section, the "Median" bagging approach panel is replaced with "CNN1D" established by Voznica et al. [144].

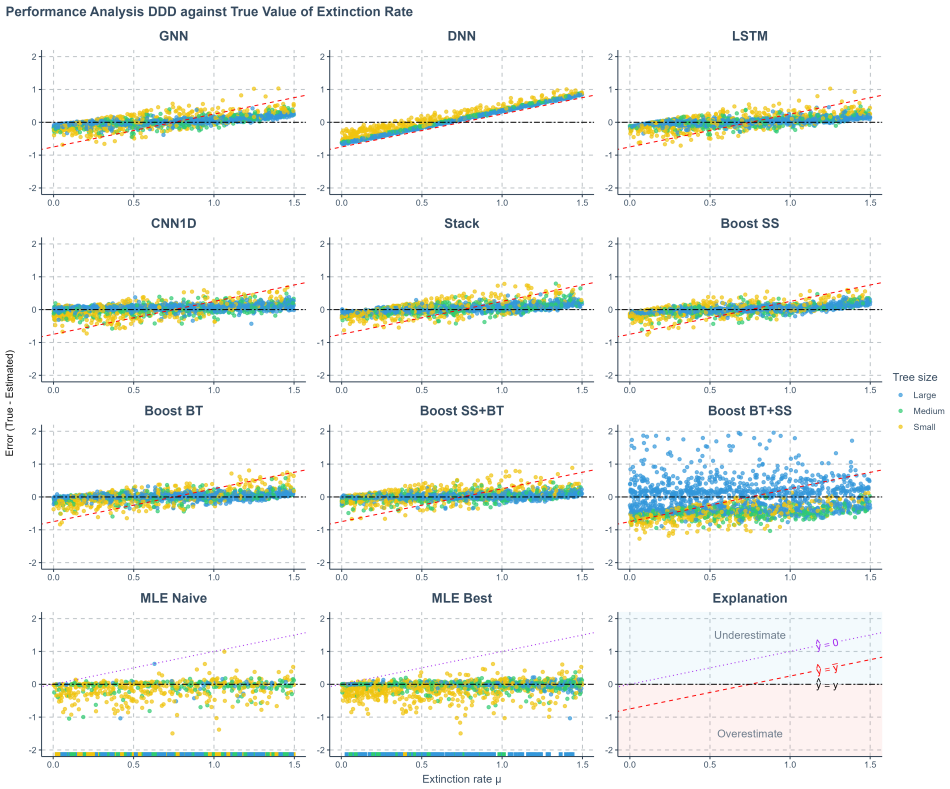


Figure 3.29: Prediction error of extinction rate plotted against true extinction rate under a diversity-dependent diversification scenario. Compared to the original figure in the results section, the "Median" bagging approach panel is replaced with "CNN1D" established by Voznica et al. [144].

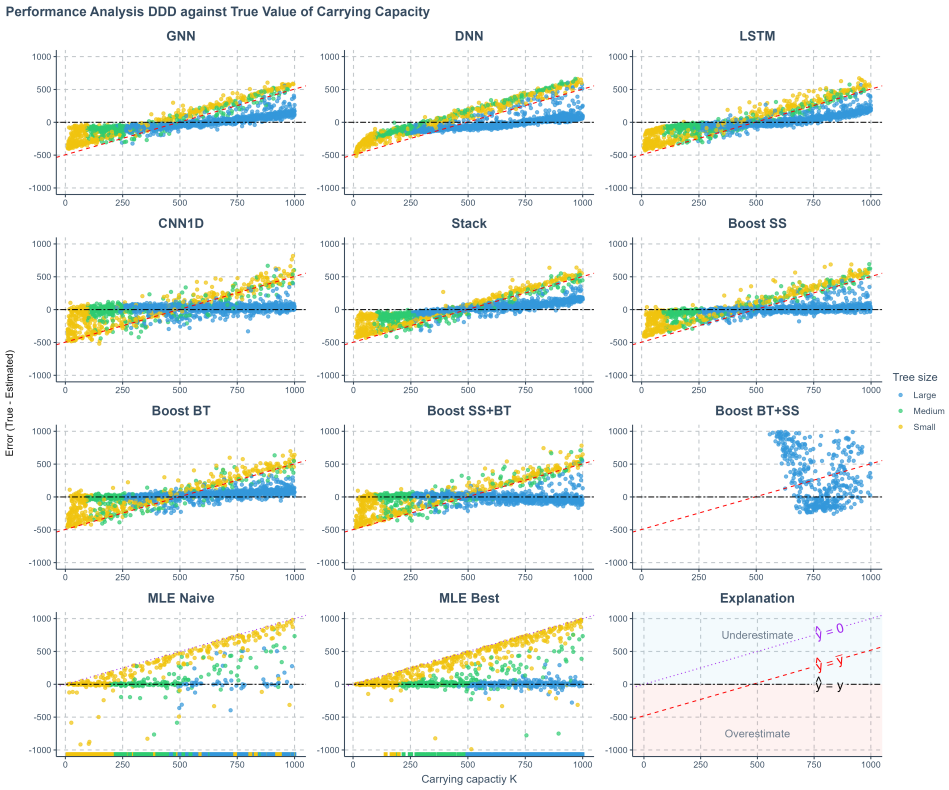


Figure 3.30: Prediction error of carrying capacity plotted against true carrying capacity under a diversity-dependent diversification scenario. Compared to the original figure in the results section, the ”Median” bagging approach panel is replaced with ”CNN1D” established by Voznica et al. [144].

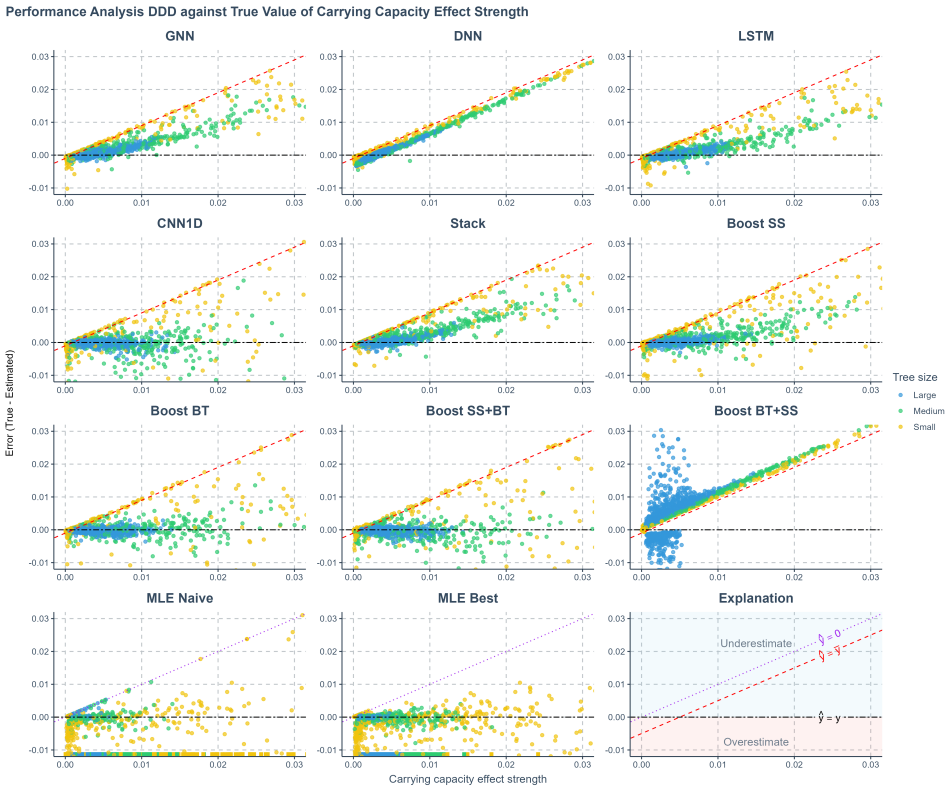


Figure 3.31: Prediction error of effect strength of carrying capacity plotted against true effect strength of carrying capacity under a diversity-dependent diversification scenario. Compared to the original figure in the results section, the "Median" bagging approach panel is replaced with "CNN1D" established by Voznica et al. [144].

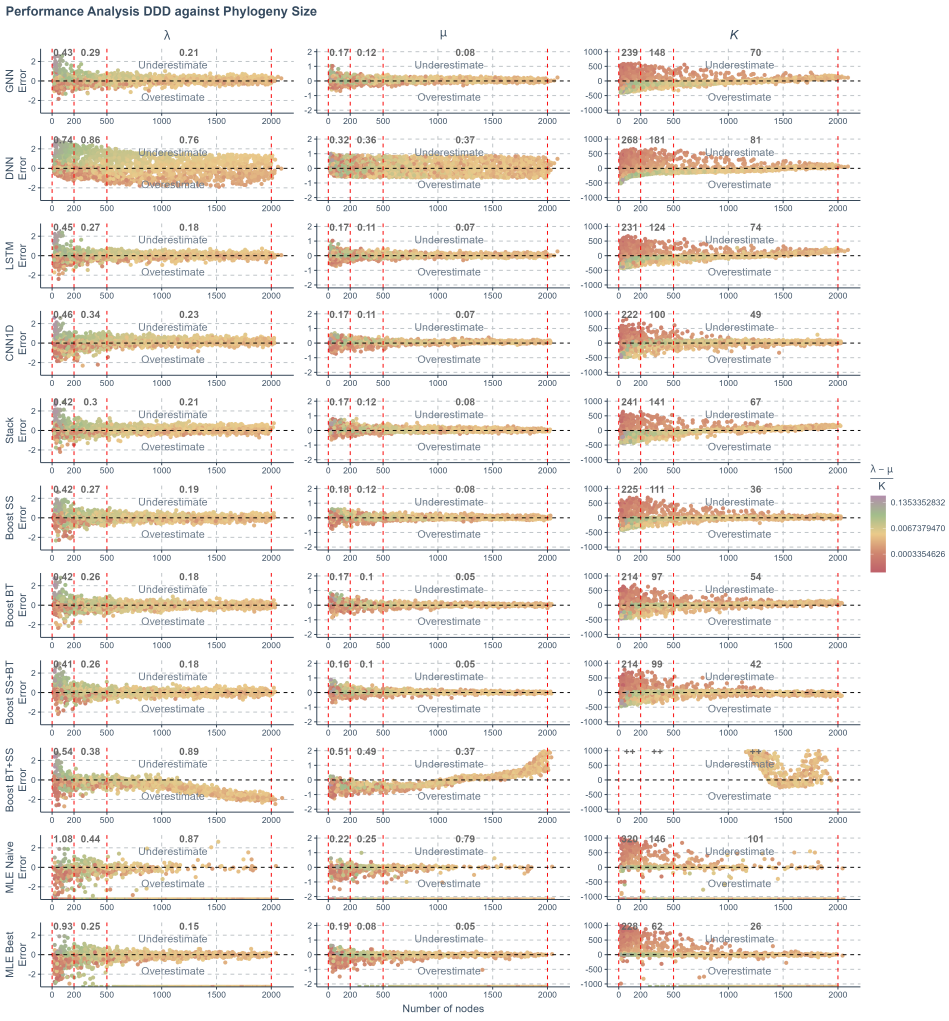


Figure 3.32: Prediction error of estimated parameters on trees of different sizes under a diversity-dependent diversification scenario. Compared to the original figure in the results section, the ”Median” bagging approach panel is replaced with ”CNN1D” established by Voznica et al. [144].

M) Dataset Re-balancing

Birth–death processes without carrying capacity effect may have larger variance of tree size than the DDD trees. A skew in the frequency of tree size across datasets may appear that leads to a non-representative sample.

To address this issue, we re-balanced the BD dataset by creating 10 bins, each designated to hold phylogenies within specific size ranges, spanning from 10 to 2000 nodes in increments of 200 nodes per bin (the first bin accepts phylogeny of sizes 10 to 200). We randomly simulated phylogenies using parameters sampled from the same space as the BD training dataset and allocated them to these bins according to their sizes, continuing this process until each bin reached its target capacity of 10,000 phylogenies. This method leads to a more equal representation of phylogenies of each size range, reducing size-based sampling bias. The filled bins were subsequently combined to form a re-balanced dataset, which in total has 100,000 phylogenies.

To compare with the original BD dataset, we trained neural networks on the re-balanced dataset, and validated neural network performance on an additional testing dataset (10,000 phylogenies simulated using the same parameter space). We computed MLE estimates on 2,000 randomly sampled phylogenies from the testing dataset. See Figure 3.33 for the results.

(Figure on next page.)

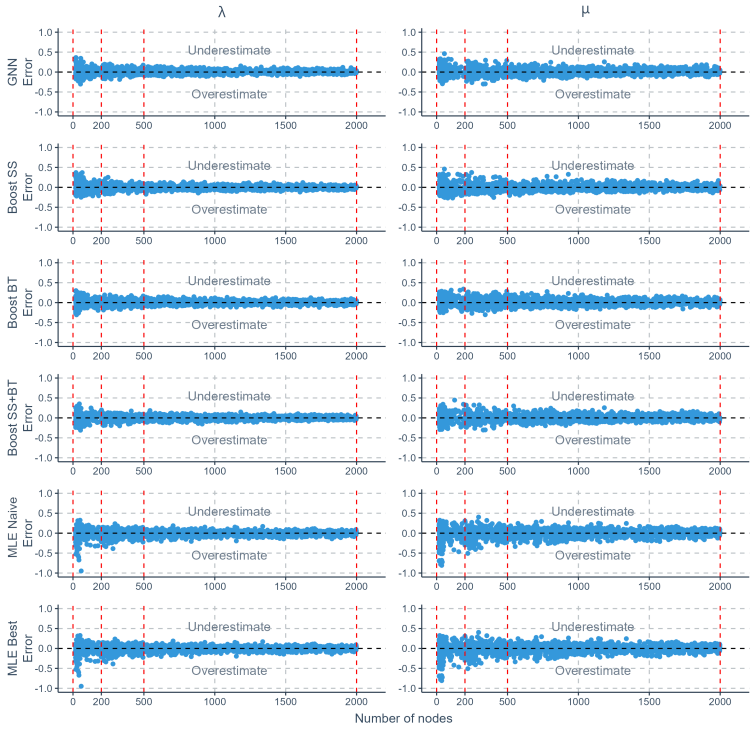


Figure 3.33: The prediction error (absolute error) of various methods applied to re-balanced phylogenies simulated under a birth–death scenario, against the total number of nodes in the phylogenies. The errors shown are the differences between the true parameters used to simulate the phylogenies and the values predicted or estimated by each method. Each row represents a method, and each column corresponds to the results for one specific parameter. Phylogenies are categorized based on their size into three sectors within each panel, separated by four vertical red dashed lines. From left to right, the sectors are: small phylogenies with fewer than 200 nodes (including root, internal, and tip nodes), medium-sized phylogenies with 200 to 500 nodes, and large phylogenies with more than 500 nodes. This scale is transformed using square root for clearer visual differentiation. GNN: Predictions obtained by the graph neural network using the phylogenies. Boost SS: Boosting strategy that corrects GNN results using DNN. Boost BT: Boosting strategy that corrects GNN results using LSTM. Boost SS+BT: Sequential correction of GNN errors first using DNN, followed by LSTM. MLE Naive: Maximum Likelihood Estimation results using random starting points for parameter optimization. MLE Best: MLE results using the true parameter values as the starting points for optimization. X-axis: Size of the phylogenies. Y-axis: Error. λ: Speciation rate. μ: Extinction rate.

N) Data outside the Training Space and Complete Phylogeny

We simulated additional datasets to explore the generalization ability of the neural networks when facing data with true parameters completely outside the training space, as well as to compare neural network performances between extant and complete phylogenies. Each simulated dataset was divided into in-distribution and out-of-distribution datasets. We used the in-distribution datasets for training and testing and the out-of-distribution datasets for evaluating the generalization ability of the trained neural networks. Validating trained neural networks on the out-of-distribution datasets can provide insights into whether their performances are tailored to the peculiarities of the already seen data and whether they are robust to new, unseen phylogenies. For each tree we kept two versions: tree of all species (TAS) and tree of extant species (TES). See Table 3.3 for the parameter settings of the additional datasets, see Table 3.4 for the criteria of in-distribution and out-of-distribution dataset separation. To conserve GPU memory, the parameter space for additional datasets was deliberately kept smaller, given that the TAS dataset inherently contains far more information than the TES.

Table 3.3: List of simulated tree datasets. The type column specifies which function is used to generate the trees. The age column specifies the crown age of the trees. The N column specifies the number of trees in the dataset. The rest of the columns specify the lower (a) and the upper (b) bounds of the initial parameters for the tree simulations, all the parameters are sampled from $U(a, b)$ except for λ_1 of the protracted birth–death scenario. λ_1 is computed as $\lambda_1 = 10^e$ where e is sampled from $U(-3, 1)$. U denotes uniform distribution. List A shows the parameter distributions of the birth–death trees and the diversity-dependent-diversification trees, λ: intrinsic speciation rate/birth rate; μ: intrinsic extinction rate/death rate; K: carrying capacity. List B shows the parameter distributions of the protracted birth–death trees, λ_1: speciation-initiation rate of good species; λ_2: speciation-completion rate; λ_3: speciation-initiation rate of incipient species; μ_1: extinction rate of good species; μ_2: extinction rate of incipient species. *In diversity-dependent-diversification simulations, the maximum extinction rate is capped at 1.5 if $0.9\lambda > 1.5$.

A: Parameter settings for BD and DDD trees

Type	Age	N	λ_0		μ_0		K	
			a	b	a	b	a	b
BD	10	60k	0.1	0.6	0.0	$0.9\lambda_0$	-	-
DDD	10	100k	0.1	3.0	0.0	$0.9\lambda_0$*	10	1000

B: Parameter settings for PBD trees

Type	Age	N	b_1		λ_1		b_2		μ_1		μ_2	
			a	b	a	b	a	b	a	b	a	b
PBD	10	100k	0.1	0.8	0.001	10	0.1	0.8	0.0	$0.8b_1$	0.0	$0.8b_2$

Table 3.4: Criteria for in-distribution (in-sample) and out-of-distribution (out-of-sample) dataset separation. Trees generated from each model are separated into left out-of-sample group, in-sample group and right out-of-sample group, based on the parameter ranges. The Model column shows the model of a parameter; the Parameter column shows the corresponding parameter; the Left Out column shows the criteria for the left out-of-sample group; the In Sample column shows the criteria for the in-sample group; the Right Out column shows the criteria of the right out-of-sample group. λ: intrinsic speciation rate/birth rate; μ: intrinsic extinction rate/death rate; K: carrying capacity. List B shows the parameter distributions of the protracted birth–death trees, λ_1: speciation-initiation rate of good species; λ_2: speciation-completion rate; λ_3: speciation-initiation rate of incipient species; μ_1: extinction rate of good species; μ_2: extinction rate of incipient species.

Model	Parameter	Left Out	In	Right Out
BD	λ_0	$[0.10, 0.18)$	$[0.18, 0.52]$	$(0.52, 0.60]$
BD	μ_0	$[0.00, 0.08)$	$[0.08, 0.46]$	$(0.46, 0.54]$
DDD	λ_0	$[0.00, 0.30)$	$[0.30, 2.70]$	$(2.70, 3.00]$
DDD	μ_0	$[0.00, 0.10)$	$[0.10, 0.80]$	$(0.80, 0.90]$
DDD	K	$[10, 100)$	$[100, 900]$	$(900, 1000]$
PBD	b_1	$[0.10, 0.18)$	$[0.18, 0.72]$	$(0.72, 0.80]$
PBD	λ_1	$[0.001, 0.002)$	$[0.002, 5]$	$(5, 10]$
PBD	b_2	$[0.10, 0.18)$	$[0.18, 0.72]$	$(0.72, 0.80]$
PBD	μ_1	$[0.00, 0.06)$	$[0.06, 0.58]$	$(0.58, 0.64]$
PBD	μ_2	$[0.00, 0.06)$	$[0.06, 0.58]$	$(0.58, 0.64]$

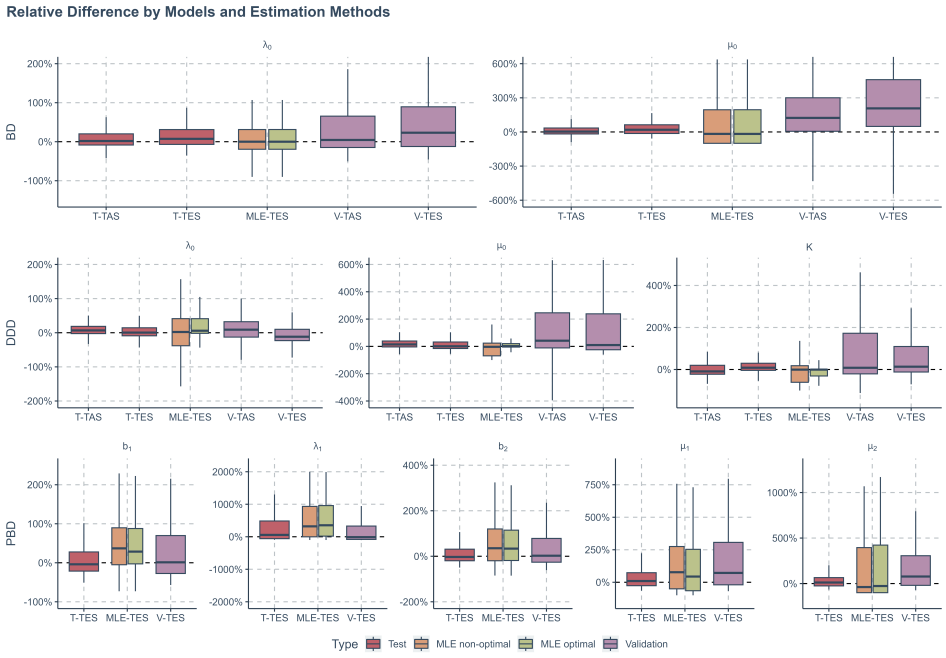


Figure 3.34: Comparisons of relative differences (in percentage) between ground true and estimated parameter values. From top to bottom, the panels in each row present the relative differences of trees generated by a specific diversification process. From left to right, the panels in each column present the relative differences of a specific parameter used when simulating the trees. Within each panel, each box represents a specific method for parameter estimation on a specific data set (as described in x-axis labels). Red boxes represent parameter estimation by using only graph neural network (GNN) on the in-sample datasets (Test in figure), yellow boxes represent the non-optimal maximum likelihood estimation (MLE) method on the complete datasets (direct outputs from simulation, without any separation), green boxes represent the optimal MLE method on the complete datasets, purple boxes represent parameter estimation by GNN on the out-of-sample datasets. BD - birth–death trees; DDD - diversity-dependent-diversification trees; PBD - protracted birth–death trees. λ - birth rate/intrinsic speciation rate; μ death rate/intrinsic extinction rate; K - carrying capacity; λ_1 - speciation rate of good species; λ_2 - speciation-completion rate; λ_3 - speciation rate of incipient species; μ_1 - extinction rate of good species; μ_2 extinction rate of incipient species. T-TAS - GNN parameter estimation on full trees (with extinct lineages) in the in-sample data set; T-TES - GNN parameter estimation on extant trees (without extinct lineages) in the in-sample dataset; MLE-TES - MLE parameter estimation on extant trees in the complete dataset; V-TAS - GNN parameter estimation on full trees in the out-of-sample dataset; V-TES - GNN parameter estimation on extant trees in the out-of-sample dataset.

O) Learning from Summary Statistics

To analyze the relationships between summary statistics and the true parameters used in the diversity-dependent diversification simulations, we computed Pearson correlation indices for each summary statistic against the true values of the three parameters. The absolute values of the correlations are visualized as a heatmap, detailed in Figure 3.35.

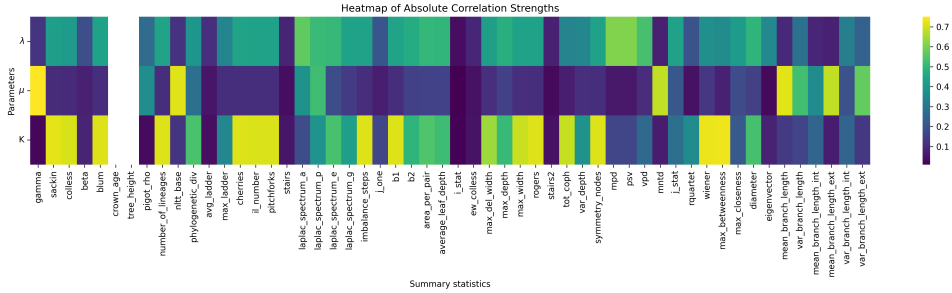


Figure 3.35: Heatmap of absolute correlation strengths between true parameters and summary statistics from simulated trees under a diversity-dependent diversification scenario. Each column corresponds to a specific summary statistic, while each row corresponds to a true parameter that was used to simulate the phylogenies. The true parameters are denoted as follows: λ for speciation rate, μ for extinction rate, and K for carrying capacity. The color gradient, ranging from dark purple to yellow, represents the increasing values of the absolute Pearson correlations between the summary statistics and the true parameters, See Appendix Q for the details of the statistics.

The heatmap analysis reveals that a large number of summary statistics showed a high correlation with carrying capacity, whereas strong correlations with speciation and extinction rates are considerably less frequent. Generally, the correlation strength between the true parameters and extinction rate is markedly lower. These findings align with the observed performance of DNN, which generally yields good estimates for carrying capacity but is much less effective in accurately predicting speciation rates and particularly poor at estimating extinction rates.

We further evaluated the DNN's performance by benchmarking it against a range of non-neural-network regression techniques using the same summary statistics. These included multivariate linear regression, Ridge regression, Lasso regression, random forest regression, and gradient boosting regression. Our findings indicated that while these traditional regression and machine learning methods generally outperformed the DNN, they still did not match the performance of our more complex neural networks.

Despite these findings, DNNs have shown efficacy in enhancing the performance of other neural networks, particularly through the prediction of residuals using summary statistics. DNN might not be the optimal choice for estimating parameters from summary statistics alone, but they can be valuable in ensemble learning strategies.

P) Meta Information of the Selected Empirical Trees

Family	Tree	Ntip
Amphibia	Caecilidae	31
Amphibia	Hynobiidae	46
Amphibia	Salamandridae	42
Amphibia	Plethodontidae	278
Amphibia	Pipidae	23
Amphibia	Eleutherodactylidae	145
Amphibia	Ranidae	218
Bird	Tyrannidae	419
Bird	Thraupidae	370
Bird	Psittacidae	330
Bird	Trochilidae	334
Bird	Columbidae	306
Bird	Furnariidae	302
Bird	Muscicapidae	279
Bird	Accipitridae	242
Bird	Picidae	223
Bird	Thamnophilidae	219
Bird	Fringillidae	194
Bird	Strigidae	191
Bird	Turdidae	170
Bird	Meliphagidae	177
Bird	Phasianidae	176
Bird	Emberizidae	163
Bird	Anatidae	157
Bird	Cisticolidae	142
Bird	Pycnonotidae	124
Bird	Rallidae	125
Bird	Cuculidae	138
Bird	Estrildidae	140
Bird	Nectariniidae	127
Bird	Leiothrichidae	127
Bird	Corvidae	120
Bird	Zosteropidae	120
Bird	Sturnidae	109
Bird	Parulidae	109
Bird	Ploceidae	108
Bird	Icteridae	102
Bird	Apodidae	99
Bird	Laridae	99
Bird	Alaudidae	91
Bird	Monarchidae	87
Bird	Scolopacidae	89

Bird	Caprimulgidae	88
Bird	Alcedinidae	91
Bird	Campephagidae	80
Bird	Procellariidae	81
Bird	Hirundinidae	83
Bird	Troglodytidae	79
Bird	Phylloscopidae	71
Bird	Ardeidae	61
Bird	Pellorneidae	66
Bird	Sylviidae	62
Bird	Cardinalidae	68
Bird	Charadriidae	64
Bird	Falconidae	64
Bird	Motacillidae	62
Bird	Acanthizidae	63
Bird	Cotingidae	65
Bird	Vireonidae	58
Bird	Acrocephalidae	52
Bird	Bucerotidae	55
Bird	Paridae	53
Bird	Pachycephalidae	50
Bird	Locustellidae	53
Bird	Rhinocryptidae	53
Bird	Timaliidae	55
Bird	Cracidae	50
Bird	Pipridae	52
Bird	Grallariidae	49
Bird	Malaconotidae	46
Bird	Passeridae	48
Bird	Rhipiduridae	42
Bird	Dicaeidae	45
Bird	Tinamidae	47
Bird	Petroicidae	44
Bird	Ramphastidae	35
Bird	Tityridae	41
Bird	Trogonidae	42
Bird	Lybiidae	41
Bird	Paradisaeidae	40
Bird	Phalacrocoracidae	33
Bird	Bucconidae	35
Bird	Threskiornithidae	34
Bird	Mimidae	34
Bird	Odontophoridae	34
Bird	Oriolidae	30
Bird	Laniidae	29

Bird	Pittidae	31
Bird	Platysteiridae	30
Bird	Cettiidae	32
Bird	Megalaimidae	28
Bird	Maluridae	27
Bird	Sittidae	24
Bird	Meropidae	26
Bird	Dicruridae	24
Bird	Otididae	25
Bird	Alcidae	23
Bird	Hydrobatidae	22
Bird	Musophagidae	23
Bird	Megapodiidae	21
Bird	Cacatuidae	21
Bird	Diomedeidae	21
Bird	Vangidae	21
CrocoTurtle	Crocodylia	25
CrocoTurtle	Testudines	233
Mammal	Vespertilionidae	386
Mammal	Soricidae	329
Mammal	Sciuridae	276
Mammal	Pteropodidae	174
Mammal	Phyllostomidae	150
Mammal	Bovidae	138
Mammal	Cercopithecidae	127
Mammal	Molossidae	98
Mammal	Didelphidae	84
Mammal	Hipposideridae	74
Mammal	Rhinolophidae	73
Mammal	Echimyidae	69
Mammal	Dasyuridae	63
Mammal	Mustelidae	59
Mammal	Heteromyidae	58
Mammal	Leporidae	58
Mammal	Macropodidae	56
Mammal	Nesomyidae	55
Mammal	Ctenomyidae	51
Mammal	Dipodidae	51
Mammal	Emballonuridae	49
Mammal	Cebidae	48
Mammal	Cervidae	45
Mammal	Felidae	40
Mammal	Talpidae	39
Mammal	Geomyidae	38
Mammal	Pitheciidae	37

Mammal	Canidae	34
Mammal	Delphinidae	34
Mammal	Viverridae	34
Mammal	Herpestidae	33
Mammal	Spalacidae	31
Mammal	Ochotonidae	28
Mammal	Gliridae	27
Mammal	Tenrecidae	25
Mammal	Atelidae	24
Mammal	Erinaceidae	22
Mammal	Phalangeridae	22
Squamate	Xantusiidae	26
Squamate	Gerrhosauridae	28
Squamate	Cordylidae	42
Squamate	Varanidae	53
Squamate	Chamaeleonidae	142
Squamate	Iguanidae	31
Squamate	Phrynosomatidae	114
Squamate	Pythonidae	26
Squamate	Viperidae	209

Q) List of Summary Statistics

Summary Statistics	
Gamma	Area Per Pair (aPP)
Sackin	Average Leaf Depth (aLD)
Colless	I Statistic
Aldous' Beta Statistic	ewColless
Blum	Max Delta Width (maxDelW)
Crown Age	Maximum of Depth
Tree Height	Variance of Depth
Pigot's Rho	Maximum Width
Number of Lineages	Rogers
nLTT with Empty Tree	Total Cophenetic Distance
Phylogenetic Diversity	Symmetry Nodes
AvgLadder Index	Mean Pairwise Distance (mpd)
Cherries	Variance Pairwise Distance (vpd)
ILnumber	Phylogenetic Species Variability (psv)
Pitchforks	Mean Nearest Taxon Distance (mntd)
Stairs	J Statistic of Entropy
Stairs2	Rquartet Index
Laplacian Spectrum Asymmetry	Laplacian Spectrum Log Eigen
Laplacian Spectrum Peakedness	Laplacian Spectrum Eigengap
Number of Nodes	Wiener Index
B1	Max Betweenness
B2	Max Closeness
Diameter (no branch lengths)	Maximum Eigenvector Value
Mean Branch Length	Variance of Branch Length
Mean External Branch Length	Variance of External Branch Length
Mean Internal Branch Length	Variance of Internal Branch Length
Number of Imbalancing Steps	J_One Statistic

R) Computational Costs

In our experiments, generating 100,000 phylogenies took 3–8 hours, depending on the scenario (BD, DDD, or PBD). MLE on BD models ran in about 2 hours for all 100,000 trees (with parallelization), whereas fitting DDD models could take up to 24 hours per tree and often failed. Neural-network costs split into training and inference: training on 100,000 trees for 100 epochs with boosting approaches can exceed 48 hours on high-end GPUs (e.g., NVIDIA A100), though simpler architectures (DNN, CNN1D and LSTM) finish in $\approx$ 2 hours. Once trained, even our most complex network predicts parameters for 100,000 trees in $\approx$ 45 minutes (including up to 30 minutes of data loading and preprocessing) under GPU acceleration.

4

Neural Recoverability and Complex Diversification Models

Reconstructing the forces that shaped macroevolutionary histories from extant phylogenies is fundamentally challenging: richly parameterized diversification models are often only weakly identifiable; different evolutionary mechanisms can yield nearly indistinguishable tree shapes. Here we use a model with evolutionary relatedness dependence to evaluate how much information about such forces can be recovered from simulated trees. We train graph neural networks and long short-term memory classifiers to distinguish three scenarios of feedback of diversity on diversification: effect of phylogenetic diversity, evolutionary distinctiveness, and nearest-neighbor distance. We also train a suite of regression networks to estimate the underlying diversification parameters. We then analyze classification performance, regression errors, and their dependence on tree size and on the strength and sign of richness and relatedness effects. Across network architectures and complexity levels, scenario classification is only moderately accurate and strongly asymmetric as revealed by the confusion matrix. Trees generated under an effect of nearest-neighbor distance on diversification tend to be correctly classified, whereas those with an effect of evolutionary distinctiveness are frequently misclassified. Regression networks systematically shrink predictions toward the empirical mean, especially for complex models, suggesting broad regions of parameter space with low identifiability. Strong global dependence of diversification rates on diversity further erodes recoverability by driving large variations in tree size that mask the subtler signatures of relatedness effects. In contrast, sufficiently strong speciation-relatedness effects can carve out narrow regions of parameter space in which scenarios and parameters become practically recoverable. Together, our results underscore the need for additional data or constraints when using flexible diversification models for macroevolutionary inference.

Qin, T., van Benthem, K., Valente, L.†, & Etienne, R.† Identifying evolutionary relatedness effects on diversification from phylogenies using neural networks. Preprint. † indicates joint senior authors.

4.1 Introduction

Time - calibrated phylogenies are now central tools for studying macroevolutionary dynamics. From such trees, we seek to reconstruct how speciation and extinction rates varied through time and across lineages, and how ecological limits and biotic interactions have shaped present-day diversity [167, 168]. Early work typically assumed constant diversification rates, which implies exponential lineage accumulation [78, 81]. However, reconstructed phylogenies for many clades show pronounced slowdowns in lineages-through-time curves, inconsistent with simple constant-rate models and motivating a broad family of more flexible diversification models, including time-dependent, diversity-dependent, and protracted speciation models [82, 84, 85, 89, 90]. A common mechanistic theme in these models is that speciation rates decline as clades fill ecological or niche space, with species richness acting as a proxy for ecological limits [96, 97].

At the same time, there is growing recognition that extant phylogenies carry only limited information about past diversification histories. Nee et al. [79] already showed that a constant-rate birth–death model is equivalent to a pure birth model with temporally declining speciation rate. This property allowed them to compute the probability of a phylogeny given the constant-rate birth–death model, but also implies that there is inherent indistinguishability of these models based on phylogenetic information alone. This result was generalized by Louca and Pennell [153], who showed that, under general birth–death processes, an infinite number of distinct speciation-extinction trajectories can produce exactly the same distribution of extant time trees, forming large "congruence classes" of observationally equivalent models. Even for more constrained model families, branching patterns alone often cannot discriminate between alternative mechanisms. For instance, Pannetier et al. [240] demonstrated that diversity-dependent and purely time-dependent diversification models that share the same expected diversity - through - time curve are essentially indistinguishable from extant trees using standard likelihood methods. These results highlight that increasing model flexibility does not automatically yield more informative inferences.

The recently-developed eve model of Qin et al. [166] augments diversity-dependent diversification by allowing speciation and extinction rates to depend linearly on both species richness and a measure of evolutionary relatedness (ER) for each lineage. The rationale for this is that ER may have an effect in diversification rates that is independent from species diversity, for instance, if closely related species are more likely to compete for niches. ER can be defined at different phylogenetic scales, for example using clade-wide phylogenetic diversity (PD, total branch length in the tree) or more localized, lineage-specific metrics such as evolutionary distinctiveness (ED, average phylogenetic distance to all other species in the clade) or nearest-neighbor distance (NND, phylogenetic distance to the most closely related species). Depending on the signs and magnitudes of the richness and ER effects, eve can generate a wide range of branching patterns and tree shapes, and previous work has suggested that the PD, ED and NND variants can produce near identical tree-shape signatures for moderate effect sizes, in terms of classic phylogenetic summary statistics.

The complexity and state dependence of eve make likelihood-based inference of parameters of the model analytically intractable, encouraging a shift toward simulation-based

approaches. Neural networks and other deep-learning methods have recently been proposed to estimate diversification parameters and classify models from phylogenetic trees [126, 143, 144, 147]. These methods can learn from graph representations of trees, branching times and summary statistics, and for relatively simple birth–death or diversity-dependent models they can achieve performance comparable to, or potentially exceeding, maximum likelihood estimation. However, when applied to models already known to suffer from non-identifiability, such as protracted speciation [123], neural networks tend to converge to conservative predictions that reflect the empirical mean of the training distribution rather than the true parameters. In such cases, the limiting factor appears to be the information content of the trees, even after extensive sweeps over neural network architectures, depths, and hyperparameter settings, rather than the expressive power of the estimator [147]. This suggests that models as flexible as diversity-dependent diversification or protracted speciation may already lie close to the practical limits of what can be identified from extant trees, regardless of whether one uses maximum likelihood or neural networks, and thus there is little reason to expect neural networks to reliably recover all parameters of an even more complex model such as eve.

In this study we use neural networks as means to explore the putative limits of information content of phylogenies regarding diversification processes in complex models. We simulate large collections of extant trees under the three ER scenarios (PD, ED, NND), three levels of model complexity (2, 4 or 6 free parameters), and a broad range of effect sizes for the effects of species richness and ER. We then train graph neural networks (GNNs) and long short-term memory (LSTM) networks as classifiers to distinguish PD, ED and NND, and as regressors (in an ensemble GNN+LSTM architecture as developed by Qin et al. [147]) to recover the underlying parameters from tree representations and branching times. Performance is evaluated as a function of tree size, true parameter values, model complexity and regressor/classifier misspecification, using simulated tree datasets.

By asking when these neural network learners succeed or fail, we obtain a map of the practical recoverability of eve under a range of scenarios. High classification accuracy or precise parameter recovery marks regions of parameter space where the three ER scenarios generate distinct, possibly information-rich tree patterns. Systematic misclassification and regression predictions collapsing toward the conditional mean correlate to non-recoverability, suggesting that different parameter combinations or ER mechanisms give rise to overlapping tree shapes that cannot be distinguished. Small trees provide fewer branching events and shorter histories, and thus are expected to fall into low-information, more non-identifiable settings, analogous to sample size in time-series hidden Markov models [241].

Our goals are therefore twofold. First, we quantify how tree size (number of nodes of the trees), effect size (magnitude of the values of parameters controlling the effects of species richness and ER on diversification) and model complexity (number of effective parameters) interact to determine the recoverability of eve parameters and the discriminability of PD, ED and NND from extant trees. Second, we interpret the characteristic failure modes of our networks, e.g., conservative regression, confusion between ER scenarios, and differing sensitivity to global (PD) versus local (ED/NND) forces, to gain insight into the expected behaviors for parameter estimation using deep learning on complex diversification models.

4.2 Methods

4.2.1 Software and Hardware

We used `PyTorch 1.12.1` [212], `PyTorch Geometric 2.3.1` [213], `Python 3.7.1` and `CUDA 12.2.2` [211] for the neural networks and `R 4.2.1` [214] for data processing, simulation and visualization. All the computationally heavy tasks were performed on the Hábrók high-performance computing cluster of the University of Groningen. Our neural networks were trained, optimized and evaluated on the NVIDIA A100 and V100 tensor core GPUs of the Hábrók cluster.

4.2.2 Simulation Approaches

We used the `R` package `evesim` [166] — an efficient `C++` implementation of the `eve` model — to simulate phylogenetic trees. We kept only extant lineages for each of the trees. The parameter settings used for simulation were chosen to limit the maximum number of extant lineages to maintain a manageable memory and computational demand for both simulation and neural network training. We filtered out simulated trees containing more than 1500 lineages (terminal tips), due to limited available hardware resources. Large trees were rare in our settings, typically less than 10 trees were removed per complete run. The sizes of the simulated trees may vary per each complete run. We recorded how many trees were removed in this way.

In the `eve` diversification model, there are three evolutionary scenarios, each of which represents a unique relatedness metric affecting macroevolutionary trajectories: PD (Phylogenetic Diversity), ED (Evolutionary Distinctiveness) and NND (Nearest Neighbor Distance). Each of the evolutionary scenarios are characterized by six parameters that shape the phylogenies: intrinsic speciation rate rate λ_0; intrinsic extinction rate rate μ_0; effect size of species richness on speciation β_N; effect size of evolutionary relatedness on speciation β_Φ; effect size of species richness on extinction γ_N; effect size of evolutionary relatedness on extinction γ_Φ.

Let N_t be species richness at time t and $\Phi_{i,t}$ the evolutionary relatedness score for lineage i at time t (we write Φ_t when it does not depend on lineage). Then we have

$$\lambda_i(t) = \lambda_0 + \beta_N N_t + \beta_\Phi \Phi_{i,t}, \tag{4.1}$$

$$\mu_i(t) = \mu_0 + \gamma_N N_t + \gamma_\Phi \Phi_{i,t}. \tag{4.2}$$

Under the PD scenario, $\Phi_{i,t} \equiv \Phi_t$ for all lineages i. Consequently, $\lambda_i(t) = \lambda_j(t)$ and $\mu_i(t) = \mu_j(t)$ for all i, j. Under ED and NND scenarios, $\Phi_{i,t}$ depends on i, so $\lambda_i(t)$ and $\mu_i(t)$ may differ across lineages at the same t. See Qin et al. [166] for more details.

By changing the number of parameters involved in simulation, we generated datasets of three different levels of model complexity for each of the three evolutionary scenarios. The low-complexity datasets were the results of simulation using λ_0 and μ_0; the medium-complexity datasets were the results of simulation using two more parameters β_N and β_Φ; the high-complexity datasets were the results of simulation using two additional parameters γ_N and γ_Φ. This staged parameterization helps disentangle the contributions of time-varying speciation and time-varying extinction by progressively introducing separate

Table 4.1: Parameter settings for the simulated tree datasets using different model settings. The "Complexity" column specifies which level of model complexity the datasets belong to, with regard to the number of parameters. Each complexity level is crossed with all three evolutionary scenarios as described in the methods. All simulated trees have identical crown age of 10 time units. For each complexity level and evolutionary scenario combination, 100,000 trees were simulated. The sub-columns under each of the parameters specify the lower (*min*) and the upper (*max*) bounds of the parameter space; all parameters are sampled from $U(min, max)$, where U denotes uniform distribution. λ_0: intrinsic speciation rate rate; μ_0: intrinsic extinction rate rate; β_N: effect size of species richness on speciation; β_Φ: effect size of evolutionary relatedness on speciation; γ_N: effect size of species richness on extinction; γ_Φ: effect size of evolutionary relatedness on extinction. For each parameter, the symbol "–" indicates that this parameter is always 0, thus disabling its impact on the simulation.

Complexity	λ_0		μ_0		β_N		β_Φ		γ_N		γ_Φ	
	min	*max*	*min*	*max*	*min*	*max*	*min*	*max*	*min*	*max*	*min*	*max*
Low	+0.200	+0.600	0.000	$+0.800\lambda_0$	–	–	–	–	–	–	–	–
Medium	+0.200	+0.600	0.000	$+0.800\lambda_0$	-0.050	0.000	-0.050	+0.050	–	–	–	–
High	+0.200	+0.600	0.000	$+0.800\lambda_0$	-0.050	0.000	-0.050	+0.050	0.000	+0.050	-0.002	+0.002

sets of parameters that modulate each process and comparing their incremental effects on the simulated trees.

Per complexity level and per scenario, we randomly sampled the parameters from uniform distributions. The upper bound for the extinction rates were proportionally dependent on the drawn speciation rate to avoid cases where extinction rates could be larger than speciation rates, because in such cases the whole tree likely goes extinct. We simulated 100,000 trees (before size-filtering) per scenario per complexity and split later for neural network training (90%) and validation (10%). For each scenario-complexity combination, we additionally simulated 10,000 trees for the purpose of performance test. The number of simulated trees is bounded by time limits and available hardware resources of the computing cluster. See Table 4.1 for detailed parameter settings and Figure 4.25 for the distribution of parameters of successfully generated and retained trees.

We assumed an identical crown age of 10 time units ($t = 10$) for all phylogenies. The choice of crown age is arbitrary because we can rescale the crown age arbitrarily, as long as we rescale the generating parameters accordingly. See also Figure 4.1 for an illustration of simulation settings.

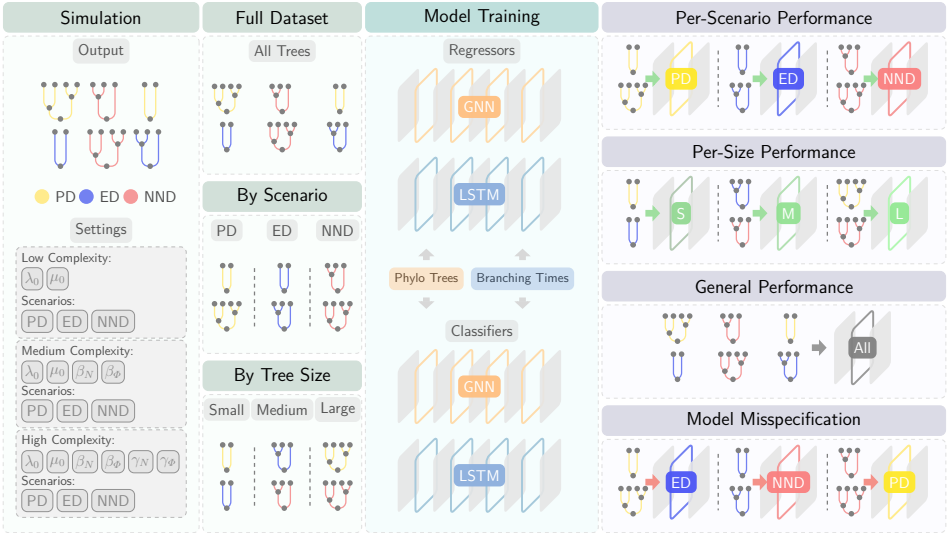


Figure 4.1: General overview of the workflow: phylogenetic trees are simulated under PD (phylogenetic diversity), ED (evolutionary distinctiveness), and NND (nearest-neighbor distance) scenarios across low/medium/high complexity parameterizations, then assembled into the full dataset, or stratified into per-scenario and per-tree-size datasets. These datasets were used to train neural network classifiers/regressors using information such as tree representations and branching times. Neural networks were evaluated per scenario, per size, overall, and under model misspecification.

4.2.3 Classification

Network Architecture and Training

The `eve` simulations involve three evolutionary scenarios. To investigate whether neural networks can identify these scenarios, we trained graph neural network (GNN) and long short-term memory (LSTM) classifiers independently on the simulated datasets. In short, in our settings, GNNs learn representations by iterative neighborhood aggregation over the graph topology [139], whereas LSTMs are gated recurrent networks that summarize sequential inputs via a memory cell [137]. For the GNN, the full phylogeny was represented as a graph, with nodes corresponding to taxa and edges representing evolutionary relationships. For the LSTM, we transformed the phylogeny into a sequence of branching times, treating these as time-series data. Input data were prepared using `PyTorch` utilities and our custom functions in the codebase `eveGNN` [221].

Detailed base architectures and hyperparameter settings of these neural networks are described in Qin et al. [147]. This design has shown good regression performance on phylogenetic trees [147]. Here, instead of outputting parameter estimates, for each input phylogeny a classifier produces a three-class prediction, which we write as a probability vector

$$\hat{\mathbf{y}} = \mathbf{p} = \left[p_{\mathrm{PD}},\ p_{\mathrm{ED}},\ p_{\mathrm{NND}}\right], \tag{4.3}$$

with $p_{\mathrm{PD}} + p_{\mathrm{ED}} + p_{\mathrm{NND}} = 1$. Each component represents the predicted probability that the input tree was generated under the corresponding scenario.

During training, we optimized the classifiers using cross-entropy loss, which directly compares the predicted probabilities $\hat{\mathbf{y}}$ to the true class labels $\mathbf{y}$. Because the classification task requires phylogenies from all three evolutionary scenarios, the classifiers were trained on the full dataset that pools trees generated under PD, ED, and NND.

For the GNN classifiers, two GNN-specific loss terms designed to facilitate graph-structured learning were added to the cross-entropy loss term. The full specification of the classification loss is given in Appendix A. For optimization, we employed the AdamW (adaptive moment estimation with decoupled weight decay) algorithm [219] to iteratively minimize the total loss and update network weights.

To balance computational efficiency and GPU memory usage, training was conducted using mini-batches of size 64, where each mini-batch contained data from 64 simulated phylogenies. During training, model checkpoints were saved at multiple epochs. We compared training and validation losses across epochs and selected the checkpoint that achieved stable generalization performance, balancing underfitting and overfitting.

Performance Analysis

We evaluated the trained classifiers using confusion matrices, standard classification metrics, distributional comparisons of generating parameters, empirical accuracy surfaces, and probability calibration diagnostics. Here we illustrate the main ideas, precise definitions and formulas for all classification performance measures are given in Appendix H.

To summarize scenario-level performance, we computed confusion matrices for each classifier and for different groupings of the test data. For every tree, the predicted class

was taken to be the scenario with the largest posterior probability among {PD, ED, NND}, ties (equal probabilities) were broken at random but were rare in practice. The resulting 3×3 confusion matrices have rows corresponding to the true scenario and columns to the predicted scenario, so that each entry C_{rc} counts how many trees with true class r were predicted as class c. For visualization, we row-normalized each confusion matrix so that the entries in each row r sum to one, i.e. $\tilde{C}_{rc} = C_{rc} / \sum_{c'} C_{rc'}$. The diagonal elements then correspond directly to the class-wise recall (true positive rate) for each scenario, whereas the off-diagonal elements give the conditional misclassification rates. We constructed two sets of contrasts based on these matrices: (i) a comparison between GNN and LSTM architectures, and (ii) a comparison between medium- and high-complexity versions of the `eve` model (four versus six free parameters), fixing the architecture to GNN. For each contrast and each cell (r, c) we computed the change in row-normalized recall between methods A and B,

$$\Delta p_{rc} = 100\big(\tilde{C}^{(B)}_{rc} - \tilde{C}^{(A)}_{rc}\big), \tag{4.4}$$

expressed in percentage points.

In addition to confusion matrices, we computed overall and per-scenario accuracy, precision, recall (sensitivity), and F1-score for each classifier. The exact formulas and notation used are provided in Appendix H. To study how classification performance varies across parameter space, we also sliced the results into consecutive ranges based on the true phylogenetic parameters and the simulated tree sizes, as detailed in Appendix J. These ranges allowed us to examine how accuracy and other metrics depend on effect sizes and tree size.

To further investigate how classification success depends on the true diversification parameters, we treated the correctness of the prediction for each tree as a binary response ($\texttt{correct} \in \{0, 1\}$) and compared the distributions of the generating parameters between correctly and incorrectly classified trees. First, we summarized the marginal distributions of each parameter $\{\lambda_0, \mu_0, \beta_N, \beta_\Phi, \gamma_N, \gamma_\Phi\}$ and of tree size ($|\mathcal{T}|$) using ridge–line density plots (i.e., vertically stacked and slightly overlapping density curves on a shared horizontal axis), stratified by architecture (GNN vs. LSTM) and true scenario (PD, ED, NND). For each scenario-architecture-parameter combination we carried out a two-sample Kolmogorov-Smirnov (KS) test comparing the parameter values for correctly versus incorrectly classified trees, and adjusted the resulting p-values across tests by the Benjamini-Hochberg procedure. The KS statistic provides a nonparametric measure of the maximum discrepancy between empirical cumulative distributions and is sensitive to differences of distributions in both location and shape.

Second, we quantified a global measure of dependence between correctness and the full parameter vector $(\lambda_0, \mu_0, \beta_N, \beta_\Phi, \gamma_N, \gamma_\Phi, |\mathcal{T}|, \lambda_0 - \mu_0)$ while controlling for scenario and architecture. We used partial distance correlation, with the true class and model encoded as conditioning variables. This test evaluates the null hypothesis that, given scenario and architecture, the parameter vector is independent of the correctness indicator. The resulting p-value is reported as an overall measure of residual parameter dependence.

To evaluate how well the class probabilities output by our classifiers reflect true predictive uncertainty, we performed a calibration analysis based on reliability diagrams and expected calibration error (ECE). For each phylogeny in the test sets we recorded the predicted

scenario ($\hat{y} \in \{\text{PD}, \text{ED}, \text{NND}\}$), the corresponding maximum predicted probability $\hat{p} = \max\{\hat{p}_{\text{PD}}, \hat{p}_{\text{ED}}, \hat{p}_{\text{NND}}\}$, and an indicator of correctness $\mathbb{I}(\hat{y} = y)$, where y denotes the true scenario. Following standard practice, we partitioned the confidence interval $[0, 1]$ into $M = 10$ equally wide bins $B_m = [(m-1)/M, m/M)$ and, for every analysis stratum, computed for each bin the average confidence

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{i:\, \hat{p}_i \in B_m} \hat{p}_i \tag{4.5}$$

and the empirical accuracy

$$\text{acc}(B_m) = \frac{1}{|B_m|} \sum_{i:\, \hat{p}_i \in B_m} \mathbb{I}(\hat{y}_i = y_i), \tag{4.6}$$

where $|B_m|$ is the number of test trees whose confidence falls in bin B_m. To summarize miscalibration with a single scalar, we used the expected calibration error

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right|, \tag{4.7}$$

where N is the total number of test trees in the stratum; lower ECE indicates better agreement between predicted confidences and observed accuracies. We computed reliability curves and ECE in several complementary strata (overall, per scenario, per tree size group, and across model complexities) and also after partitioning trees into ranges of β_N and β_Φ.

4.2.4 Regression

Network Architecture and Training

Like classification, we also used GNN and LSTM for the regression tasks. Additionally, we combined GNN and LSTM using a sequential "boosting" method to leverage all available information and reduce prediction error. In this approach, the GNN first provides an initial parameter estimate; the LSTM then predicts the residual errors of the GNN, and the final estimate is obtained by adding this correction. Details of the neural network regressors and the "boosting" method are described and discussed in Qin et al. [147]. For each level of complexity, the vector of predicted variables $\hat{\theta}$ comprises all nonzero parameters used in simulation in the model (two, four, or six for the low-, medium-, and high-complexity settings, respectively, as described above).

For the regression analyses, the simulated tree datasets were reorganized into subsets tailored to different questions. Across all complexity levels, we partitioned the data according to (i) evolutionary scenario in the `eve` model, yielding three subsets that contain only phylogenies generated under PD, ED, or NND, respectively, for studying per-scenario and regressor-misspecification performance, and (ii) tree size (i.e., the total number of nodes), yielding three consecutive ranges: small trees ($n \in [0, 200]$), medium trees ($n \in [201, 500]$), and large trees ($n > 500$), for studying size-specific performance. In addition to these subsets, we retained the full dataset that contains all simulated phylogenies for assessing overall performance (see also Figure 4.1).

For the LSTM regressors, we used the Huber loss [233] to quantify prediction error during training. For the GNN regressors and the boosting method, two GNN-specific loss terms designed to facilitate graph-structured learning were added to the Huber loss term. The full specification of the regression loss is given in Appendix A. For optimization and training, similar approaches were used as in the classification tasks.

For each complexity level, the above subsetting allowed us to train regressors separately on (i) the complete training dataset, (ii) scenario-specific training datasets, and (iii) size-specific training datasets. For each neural network architecture, this yields a total of seven regression models per complexity level. The structure of the regression analyses and data partitions is summarized in Figure 4.1.

Performance Analysis

Regression performance was evaluated using residual diagnostics and correlation-based measures. Here we provide a brief overview; formal definitions and equations are given in Appendix I. For each trained regressor, we computed residuals as the differences between the true parameters and the predicted parameters, $\hat{\theta} - \theta$, and summarized these residuals as a function of (i) the true net diversity-independent part of the diversification rate $\lambda_0 - \mu_0$, (ii) tree size (the total number of nodes, including root, internal, and tip nodes), and (iii) the four evolutionary relatedness effect sizes β_N, β_Φ, γ_N, and γ_Φ. As a reference for expected tree sizes under a simple birth–death process, we also computed the expectation and an approximate confidence interval for tree size (Appendix K). To quantify how closely the predictions follow the conditional mean of the training data, we calculated four additional metrics: the coefficient of determination, a slope difference measure, the distance correlation, and Spearman's rank correlation. These metrics characterize, respectively, linear fit, systematic shrinkage toward the mean, overall dependence (including nonlinear structure), and monotonic association between predictions and true parameters.

The regression analyses focused on four aspects of performance: overall performance across all trees, per-size performance, per-scenario performance, and performance under regressor misspecification. Performance under regressor misspecification was assessed by applying scenario-specific regressors (trained on trees from only one scenario) to the complete testing datasets that combine trees from all three scenarios. Because the testing sets contain trees generated under all scenarios, this cross-application reveals how strongly regressor performance degrades when the assumed scenario does not match the true generative process. Our pipeline first classifies trees into scenarios and then applies a scenario-specific regressor to estimate parameters. Quantifying how sensitive the regressors are to upstream classification errors therefore provides insight into how the full pipeline may perform in practice. Figure 4.1 provides an overview of the analysis steps.

4.3 Results

4.3.1 Classification

Overall, the classifiers recover only a moderate amount of information about the underlying diversification scenario. Across architectures and model complexities, F1 scores and scenario-specific recalls rarely approach one and are often closer to 0.5, especially for ED trees, indicating that misclassification is common. In the following sections, we examine

when misclassification becomes less common, treating this primarily as a diagnostic of practical scenario recoverability from extant trees.

Confusion Matrices

The row-normalized confusion matrices in Figure 4.2 show that all classifiers recover substantial information about the underlying diversification scenario. The errors are strongly structured and non-symmetric across scenarios. For both architectures, when trained on all complexities simultaneously (see the two panels in the top row of Figure 4.2), NND trees are easiest to classify and ED trees are consistently the most difficult. For example, under the GNN trained on all complexities (top-left panel), around 72% of NND trees are correctly identified, whereas only about 56% of PD trees and 21% of ED trees fall on the diagonal. Large proportions of misclassified ED trees are predicted as NND, potentially indicating that the ED and NND scenarios generate overlapping tree patterns. PD trees are also more frequently confused with NND, whereas the reverse error (NND mislabeled as PD) is comparatively rare.

Comparing architectures on the same test set (top row in Figure 4.2), the LSTM trades some PD and NND recall for better separation of the ED scenario. The change-in-recall panel (top right in Figure 4.2) shows that, across all true scenarios, the LSTM systematically shifts probability mass away from predicting NND and toward predicting ED (positive changes in the ED column and negative changes in the NND column). In other words, the LSTM architecture partially recovers information that distinguishes ED from NND, but at the cost of slightly more confusion between PD and ED and a modest reduction in NND accuracy. This trade-off suggests that even with a different representation of the trees, the ED scenario remains only weakly recoverable from PD and NND.

The bottom row of Figure 4.2 examines how model complexity affects scenario-level recoverability when using the GNN classifier. For trees generated under the medium-complexity (four-parameter) `eve` model (bottom-left panel), PD recall is high (about 79%), NND recall is moderate (about 54%), and ED recall remains low (about 21%). When complexity is increased to the full six-parameter model (bottom-middle panel), the change panel (bottom right) reveals that, across all true scenarios, the high-complexity setting induces a shift toward predicting NND.

Taken together, these results indicate that scenario information in `eve` trees is unevenly distributed and, unsurprisingly, degrades as the number of free parameters increases. At the scenario level, increased model flexibility in `eve` potentially leads to practical non-recoverability between evolutionary relatedness mechanisms.

Accuracy Metrics

The sliced accuracy curves in Figure 4.3 and Figure 4.4 show that, once we condition on scenario and complexity, differences between the GNN and LSTM architectures are modest compared with the effects of the underlying diversification parameters. Across most panels the two architectures track each other closely, and all three accuracy measures (F1, precision, recall) exhibit qualitatively similar trends. The main determinants of classification performance are the strength and direction of the diversification effects and, to a lesser extent, tree size and model complexity.

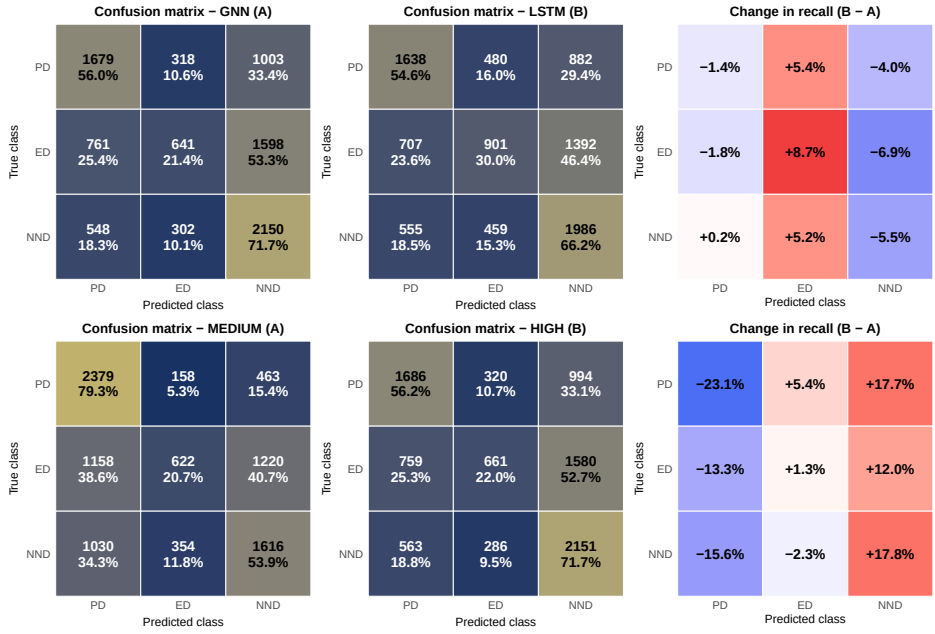


Figure 4.2: Row-normalized confusion matrices and changes in scenario-specific recall. **Top row**: comparison between classifier architectures on trees of *all complexities*. The left and middle panels show 3 × 3 confusion matrices for the GNN (A) and LSTM (B) classifiers, respectively, evaluated on the same test trees and pooled across all complexity levels (low, medium, high). Rows correspond to the true evolutionary scenario (PD, ED, NND) and columns to the predicted scenario. The y-axis order is such that the diagonal from top left to bottom right corresponds to matching true and predicted classes (PD, ED, NND). Each tile shows the absolute number of trees and the row-wise percentage of trees in that cell; colors encode the row-normalized value $\tilde{C}_{rc}$, so the diagonal entries represent the per-scenario recall (true positive rate) and off-diagonal entries represent conditional misclassification rates. The right-hand panel displays the change in recall (in percentage points) between methods B and A for each cell, $\Delta p_{rc} = 100(\tilde{C}_{rc}^{(B)} - \tilde{C}_{rc}^{(A)})$. Red tiles indicate that method B predicts that combination of true and predicted class more frequently than method A, and blue tiles indicate the opposite. **Bottom row**: analogous comparison *between complexity levels* of the eve model. Both confusion matrices are obtained with the GNN classifier, but the left panel (A) uses only trees simulated under the medium-complexity (four-parameter) settings, whereas the middle panel (B) uses only trees from the high-complexity (six-parameter) settings. The right-hand panel again shows changes in recall $\Delta p_{rc} = 100(\tilde{C}_{rc}^{(B)} - \tilde{C}_{rc}^{(A)})$ in the same format as above. Together, these panels reveal which scenarios are systematically confused with one another, how this pattern depends on architecture, and how increasing the number of free parameters in the eve model shifts errors towards particular predicted scenarios.

Effect sizes, as proxied by the true parameter values or composite quantities, play a central role. An increase in the true net diversification rate ($\lambda_0 - \mu_0$) is generally associated with higher F1 scores, for PD trees (Figure 4.3, top row, left most panel). However, this trend flattens or even reverses for ED and NND trees once $\lambda_0 - \mu_0$ exceeds roughly 0.4. Larger absolute richness effects on speciation and extinction ($|\beta_N|$ and $|\gamma_N|$) consistently depress F1 scores and recall across all scenarios, indicating that strong global diversity dependence tends to obscure differences between PD, ED and NND (Figure 4.3, middle row). By contrast, increasing the absolute magnitude of the evolutionary relatedness effect on speciation ($|\beta_\Phi|$) improves classification performance for PD and NND trees, with the highest scores observed for large positive β_Φ. For ED trees, performance improves as β_Φ moves from negative to positive values. No robust, monotonic effect on any of the three metrics is observed for the extinction-relatedness parameter γ_Φ; its influence on scenario recoverability appears weak compared to the speciation components (β_N and β_Φ).

Tree size exerts an additional, mostly positive influence on performance. For PD and ED trees, F1, precision and recall generally increase with the number of nodes, reflecting the fact that larger trees contain more branching events and are thus more informative. The pattern for NND trees is less straightforward: accuracy curves are flatter and in some slices even decline in the largest size bins. This is likely due to a combination of smaller typical sizes for NND trees and reduced sample sizes in the upper quantiles (NND trees are generally much smaller in size), so the apparent downturn should be interpreted cautiously. Overall, the consistent improvement for PD and ED with increasing size supports the view that small phylogenies fall into a largely non-recoverable part of parameter space.

In order to assess the impact of model complexity, we only compare medium- and high-complexity versions of the eve model because parameters other than speciation and extinction rates (λ_0 and μ_0) are set to zero (and thus not in effect) under the low complexity setting. Similarly, we do not study species richness and evolutionary relatedness effects on extinction rate (γ_N and γ_Φ) because these parameters are only in effect under the high complexity setting. We find that diversification model complexity has scenario-specific effects (Figure 4.4). For PD trees, moving from the four-parameter to the six-parameter setting generally leads to marginally higher precision but substantially lower recall, resulting in a net reduction in F1 (but these patterns are not observed in the tree size effect panel). Thus, under the high-complexity model the classifier is potentially more conservative: PD predictions are more often correct when they are made, but many true PD trees are reclassified as ED or NND. For ED and NND trees, there are no clear general trends.

Parameter Space Associated with Correct vs. Incorrect Classifications

To assess which parts of parameter space are associated with successful scenario classification, we compared the generating parameters of correctly and incorrectly classified trees (Figure 4.5). For each combination of true scenario (PD, ED, NND) and classifier architecture (GNN vs. LSTM), ridge–line densities show how the marginal distributions of λ_0, μ_0, β_N, β_Φ, γ_N and γ_Φ differ between correct and incorrect predictions. We observe that correctly classified trees tend to occupy more extreme regions of parameter space for β_Φ and less negative (smaller absolute value) regions of β_N, although NND trees are an exception. This observation further verifies the confounding effect of β_N and the important role of the absolute effect sizes of β_Φ.

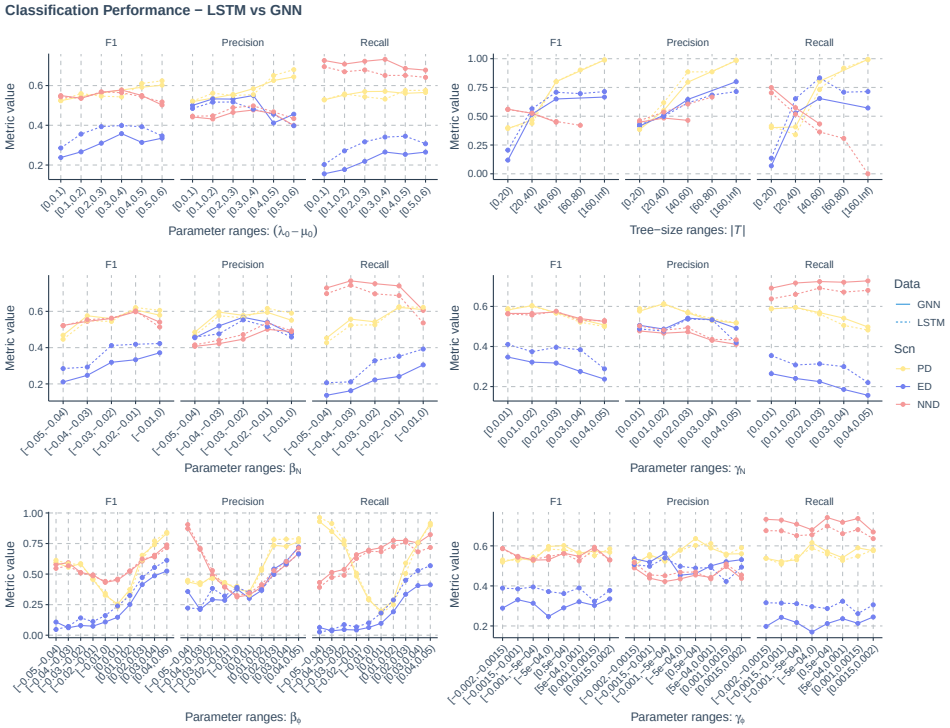


Figure 4.3: Comparison of trends of neural network classification accuracies (changing along true value slices of the parameters) between GNN (graph neural networks) and LSTM (long short-term memory recurrent networks) performances, indicated by solid and dashed lines, and between three evolutionary scenarios (true scenario under which the trees were generated). Light yellow lines represent performances of trees generated under the phylogenetic diversity (PD) scenario, dark blue lines stand for performances of trees under the evolutionary distinctiveness (ED) scenario, and red lines stand for performances of trees under the nearest neighbor distance (NND) scenario. X-axis: true value slices of the parameters. Y-axis: classification performance metrics, as shown by the column facet strips. See Appendix H for detailed explanation of the three classification performance metrics.

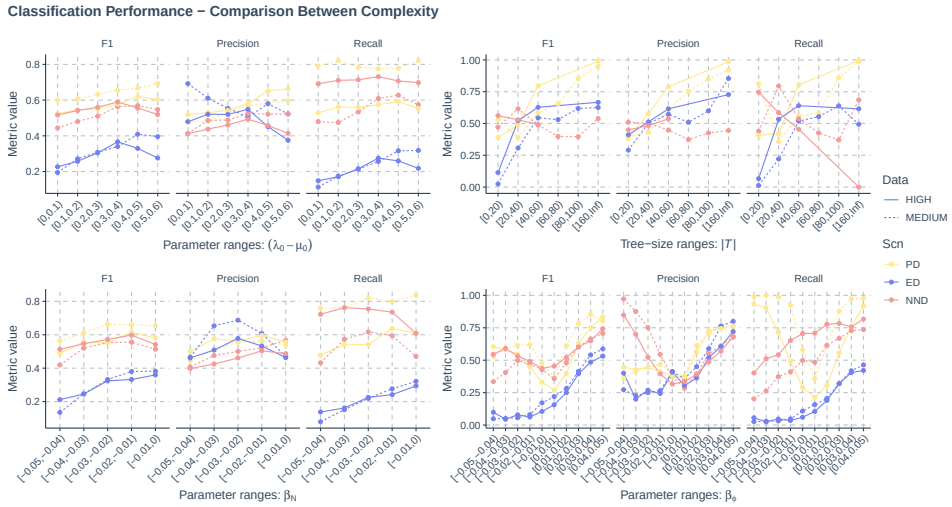


Figure 4.4: Comparison of trends of neural network classification accuracies (changing along true value slices of the parameters) between medium and high complexity levels (which indicate the number of parameters used to generate the trees) as shown by solid and dashed lines. The low complexity scenario was not considered because parameters other than speciation and extinction rates are not in effect in that scenario. The comparisons were also made between three evolutionary scenarios (true scenario under which the trees were generated). Light yellow lines stand for performances of trees generated under the phylogenetic diversity (PD) scenario, dark blue lines stand for performances of trees under the evolutionary distinctiveness (ED) scenario, and red lines stand for performances of trees under the nearest neighbor distance (NND) scenario. X-axis: true value slices of the parameters. Y-axis: classification performance metrics, as shown by the column facet strips. See Appendix H for detailed explanation of the three classification performance metrics.

The one-dimensional Kolmogorov–Smirnov tests reveal that the strongest and most systematic separations between correct and incorrect classifications occur for the richness and relatedness effects on speciation, β_N and β_Φ. For all scenario-architecture combinations, the KS statistics for these parameters are moderate to large and the Benjamini–Hochberg adjusted p-values are mostly below 0.01, indicating that the distributions of β_N and β_Φ differ markedly between correctly and incorrectly classified trees. For the other parameters, the patterns show only weak or sporadic differences. For our classifiers, the speciation components (β_N and β_Φ) — particularly the interaction between species richness and evolutionary relatedness — carry more discriminative information than the extinction components (γ_N and γ_Φ) or the baseline rates (λ_0 and μ_0).

Partial distance correlation analysis further indicates a significant global dependence between the eight-dimensional parameter vector $(\lambda_0, \mu_0, \beta_N, \beta_\Phi, \gamma_N, \gamma_\Phi, |\mathcal{T}|, \lambda_0 - \mu_0)$ and the correctness indicator even after conditioning on scenario and architecture (Figure 4.5, subtitle). Together, these findings imply that classification errors are not purely random but are concentrated in regions of parameter space where effect sizes and tree sizes jointly yield weak or ambiguous signals. Strong richness and relatedness effects carve out parameter space in which the PD, ED and NND scenarios become practically non-recoverable.

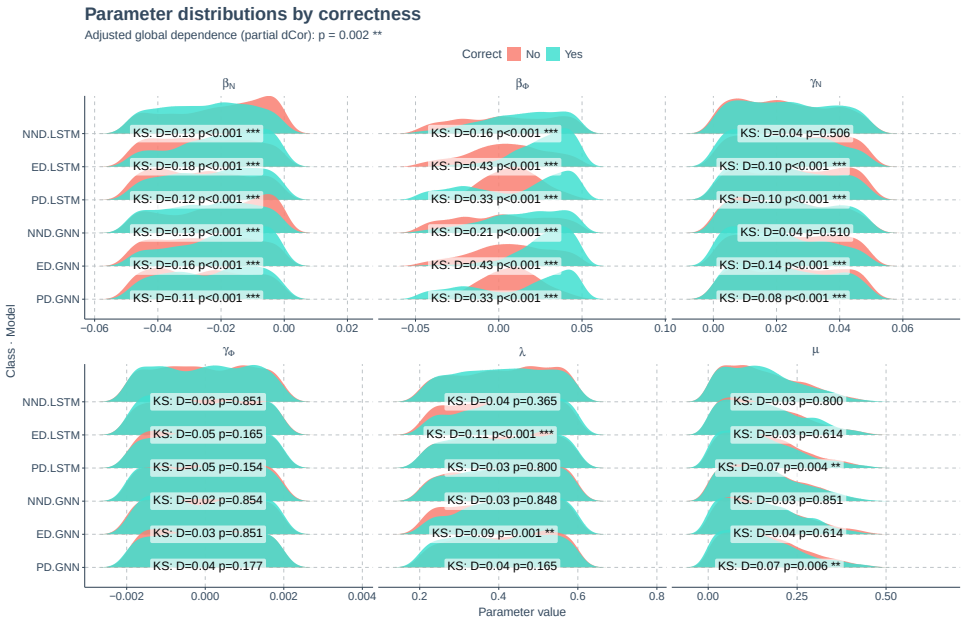


Figure 4.5: Parameter distributions for correct vs. incorrect classifications. Each panel shows ridge–line kernel density estimates of a generating parameter ($\lambda_0, \mu_0, \beta_N, \beta_\Phi, \gamma_N, \gamma_\Phi$; columns) for a given combination of true scenario (rows within each panel: PD, ED, NND) and classifier architecture (GNN vs. LSTM). Within each ridge–line, turquoise densities correspond to trees that were correctly classified and salmon densities to misclassified trees. On top of the each ridge–lines, we report the two-sample Kolmogorov–Smirnov statistic and Benjamini–Hochberg adjusted p-value comparing the parameter distributions between correct and incorrect trees for that scenario-architecture-parameter combination. The adjusted global dependence between the eight-dimensional parameter vector $(\lambda_0, \mu_0, \beta_N, \beta_\Phi, \gamma_N, \gamma_\Phi, |\mathcal{T}|, \lambda_0 - \mu_0)$ and the correctness indicator after conditioning on scenario and architecture is significant ($p = 0.002$). Together, these summaries highlight which parameters and scenarios exhibit systematic differences between correctly and incorrectly classified trees, and where classification errors appear compatible with a lack of parameter signal.

Calibration of Scenario Probabilities

Reliability diagrams for the GNN classifier (Figure 4.6) show that, when predictions are pooled across all trees, the network is systematically overconfident. In the overall panel, the reliability curve lies below the diagonal for most confidence bins. Predictions with nominal confidence between 0.5 and 0.8 attain observed accuracies closer to 0.4-0.6, and even the highest-confidence bin ($\hat{p} \approx 1$) reaches an accuracy of only about 0.8. The corresponding expected calibration error (ECE) is therefore non-negligible.

Stratifying by true diversification scenario (Figure 4.6, first row, middle panel) reveals that ED is strongly overconfident across all confidence bins, with observed accuracies that are far below the reported confidences and an ECE that is much larger than for PD or NND. This aligns with the confusion matrices (Figure 4.2), where ED is the hardest scenario to identify. The patterns for PD and NND are asymmetric and non-monotonic.

Calibration differences between medium- and high-complexity eve models are relatively

minor (see Figure 4.6, top row, third column). The two complexity-specific curves almost coincide and are both below the diagonal. The increase of free parameters potentially only changes which class is predicted rather than how well probabilities match empirical frequencies. Calibration on all size groups show overconfidence. ECE exhibit no clear trend or difference for different complexities and tree sizes.

Calibration stratified by parameter space indicates that miscalibration depends more strongly on the relatedness effect β_Φ than on the richness effect β_N. Across β_N ranges the reliability curves are fairly similar and moderately overconfident, with modest increase in ECE as $|\beta_N|$ increases. The β_Φ panel shows larger between-bin variation. ECE generally decrease as $|\beta_\Phi|$ decreases. These results suggest that parameter settings in which β_Φ induces weak or ambiguous changes in tree shape are precisely where both classification performance and probability calibration deteriorate, reflecting a lack of practical recoverability for the `eve` scenarios in those regions of parameter space. This is also true when increasingly negative β_N strengthens richness-driven confounding effect by shrinking the differences in tree summary statistics among PD, ED, and NND, thereby making the scenarios less distinguishable and degrading classification performance, as observed in Qin et al. [166].

4.3.2 Regression

Prediction accuracy for the net diversification rate ($\lambda_0 - \mu_0$) improved with increasing true net diversification rate (see top facet rows of all panels in Figure 4.7), only at the low complexity level. Similar improvement was not observed at medium and high complexity levels, not only for the net diversification rate, but for all the parameters examined. See also the figures in Appendix B.

Training neural network regressors on partial datasets—either sliced by scenario or by tree size—did not yield obvious performance gains compared to training on complete datasets (see the performance differences between the left two panels and the right two panels of Figure 4.7. See also Figure 4.8). Further quantitative analyses even showed that training the regressors on partial datasets impaired the neural networks' generalization ability, with predictions aligning closer to the midpoint of the generative parameter space of the training datasets (see Figure 4.8 and compare between metrics in the same color group). When trained on complete datasets, the neural networks generally predict better on larger trees for $\lambda_0 - \mu_0$, β_N, γ_N and γ_Φ.

Increasing model complexity degrades performance in estimating the net diversification rate. In more complex scenarios, the expected correlation between larger tree sizes and higher accuracy diminishes, with predictions converging toward the sample mean of the parameters (points close to the red dashed line in Figure 4.7 and Figure 4.8; results for other parameters are in Appendix B and Appendix D).

In the low-complexity scenario — a simple birth–death process — net diversification rate estimates ($\lambda_0 - \mu_0$) appear more accurate when the actual tree sizes were near the mean of the generated sample distribution rather than the theoretical expectation conditioned on the true net diversification rate and crown age (see the left panel of Figure 4.9 as well as those in Figure 4.26, Figure 4.27, Figure 4.28 and Figure 4.29 in Appendix F). In medium- and

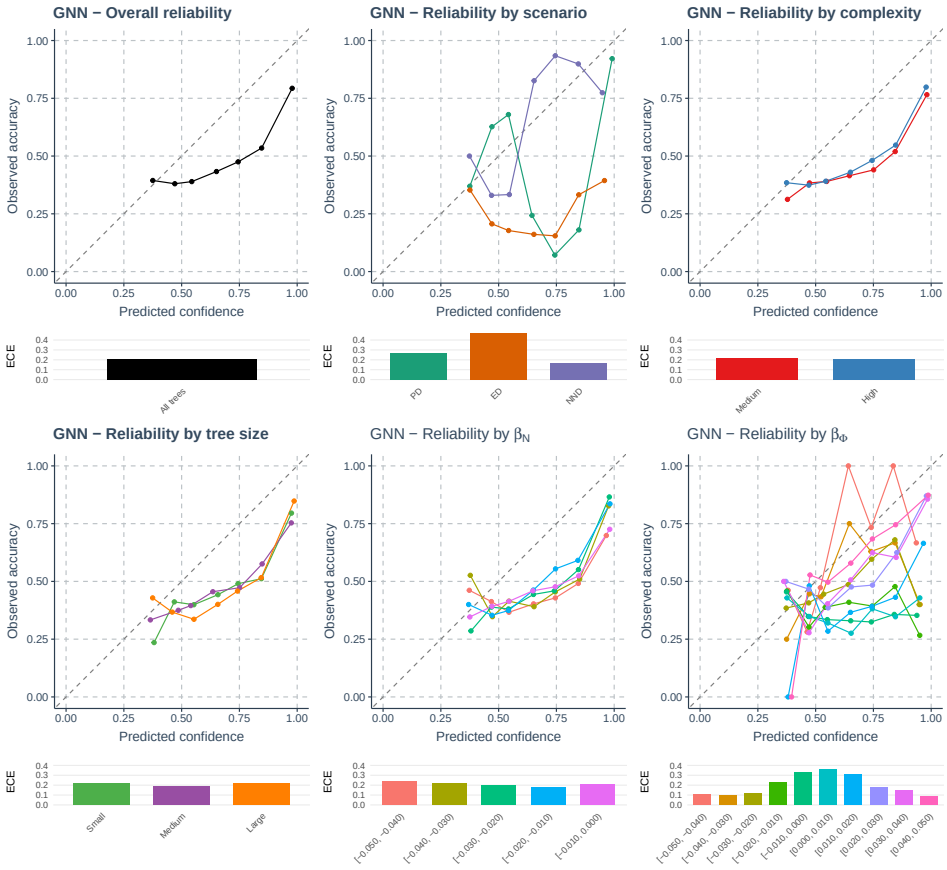


Figure 4.6: Calibration of the GNN classifier. Each main panel shows a reliability diagram in which the observed accuracy (y-axis) is plotted against the predicted confidence (x-axis; maximum class probability) in ten confidence bins; the dashed diagonal indicates perfect calibration. Top row: overall calibration across all test trees (left); calibration stratified by true diversification scenario (PD, ED, NND; middle); and calibration for GNNs trained on medium- and high-complexity eve models (right). Bottom row: calibration stratified by tree size (small, medium, large; left), by ranges of the richness effect on speciation β_N (middle), and by ranges of the relatedness effect on speciation β_Φ (right). Below each reliability panel, a horizontal bar plot reports the expected calibration error (ECE) for each curve, using matching colors; smaller ECE values indicate better calibration.

high-complexity scenarios, although theoretical expectations are unavailable, we observed that net diversification rate estimates, as well as those for other parameters (β_N, β_Φ, γ_N, and γ_Φ), were likewise more accurate when tree sizes approximated the sample's arithmetic mean. Moreover, tree-size distributions shifted systematically with model complexity: higher-complexity simulations generally produced smaller trees, with the reduction in tree size becoming increasingly pronounced from PD to ED to NND.

Additionally, larger tree sizes corresponded to more accurate estimates of both the net diversification rate ($\lambda_0 - \mu_0$, see Figure 4.7 and Figure 4.9) and the effect of species richness on extinction (γ_N, see Figure 4.14 in Appendix B and Figure 4.28 in Appendix F), whereas this trend was less evident for all the other parameters (β_N, β_Φ, and γ_Φ, see the other figures in Appendix B and Appendix F). When a parameter exhibited a weak correlation with tree size, increasing the tree size did not lead to substantially improved estimation accuracy. This is particularly true for the parameters of higher complexity levels (see the performances between panels within the same rows in Figure 4.9 and Appendix F). In all cases, extremely small trees (with total number of nodes less than 50) were associated with very poor estimates.

In the regressor misspecification test, all eve model scenarios are identical under low complexity setting because there is no parameters other than the baseline rates (λ_0 and μ_0) in effect. Misspecification mainly affects the estimates of all parameters when trees are of medium or large size. The neural networks are always conservative by making predictions that are close to the sample mean of the training data on small trees. Regardless of which evolutionary scenario the neural networks were trained on, they produce much worse estimates when recovering parameters from trees generated under the PD scenario (where phylogenetic diversity controls the processes globally). As the real underlying scenarios of the trees to be tested shifting from global to local evolutionary forces, the estimates by the neural networks trained on all scenarios become more stable. See Figure 4.10 and Figure 4.30, Figure 4.31, Figure 4.32 and Figure 4.33 in Appendix G for details.

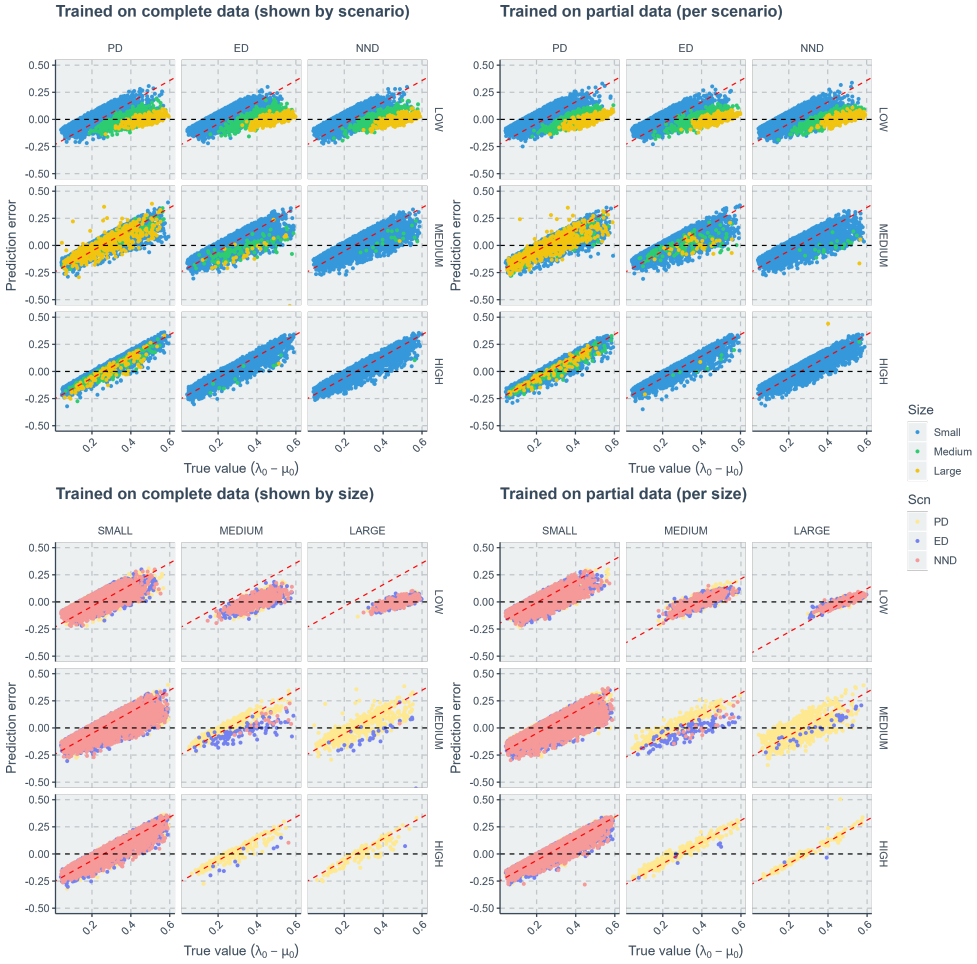


Figure 4.7: Comparison of neural network regression errors along true values of the net diversification rate ($\lambda_0 - \mu_0$). The left two panels show the results when the neural networks were trained on complete datasets. The right two panels show the results when the neural networks were trained on partial datasets (training datasets were sliced either by evolutionary scenarios or by tree size groups while in visualization we slice by both). Within each panel, each facet column indicates results of either an evolutionary scenario or a tree size group. Each facet row indicates results of a diversification model complexity. When columns represent scenario groups, blue points correspond to large trees, green points to medium trees, and yellow points to large trees. When columns represent tree size groups, light yellow points correspond to performances of trees generated under the phylogenetic diversity scenario, dark blue points to performances of trees under the evolutionary distinctiveness scenario, and red points to performances of trees under the nearest neighbor distance scenario. X-axis: true value of the net diversification rate ($\lambda_0 - \mu_0$). Y-axis: prediction error (absolute difference between true value and predicted value).

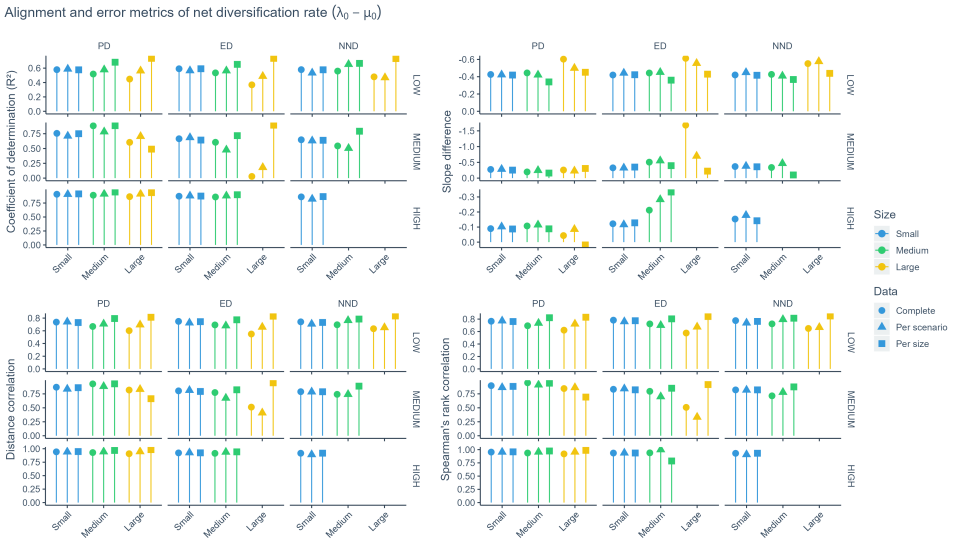


Figure 4.8: Comparison of the degree to which neural network regression estimates align with the conditional mean. There are four metrics, each describing a facet of the alignment, see Appendix I for details and interpretations. Within each panel, each facet column indicates results of a evolutionary scenario (PD for phylogenetic diversity, ED for evolutionary distinctiveness, NND for nearest neighbor distance); each facet row indicates results of a diversification model differing in complexity (as indicated by the number of parameters used to generate the trees). Yellow bars stand for metrics of small trees, green bars stand for medium trees and blue bars stand for large trees. At the top of the bars, shapes represent how the datasets were sliced to train the neural networks. Circles indicate complete datasets, triangles indicate slices by evolutionary scenario and squares indicate slices by tree size group. X-axis: size group of the trees. Y-axis: value of alignment metrics. As model complexity increases, the simulations rarely yield medium or large trees, so those size groups may be missing in the corresponding panels.

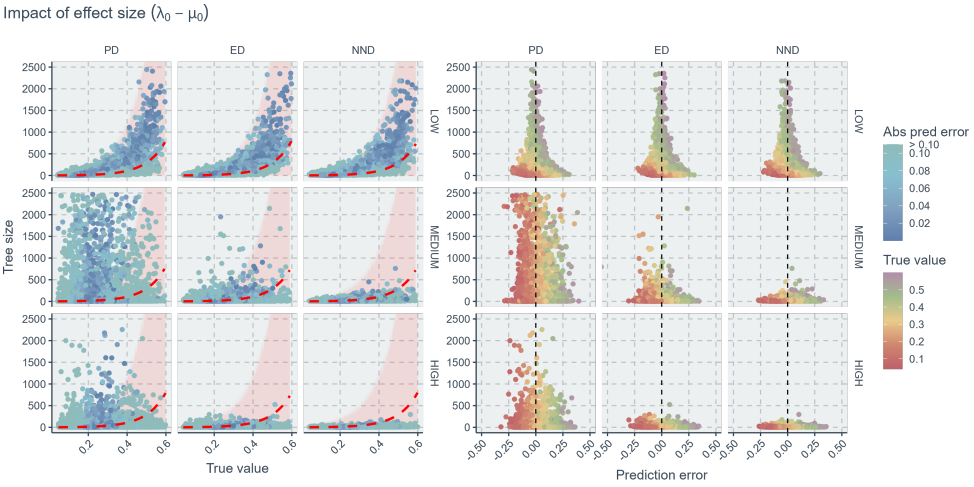


Figure 4.9: Relationships of tree size to true parameter values and prediction errors of the net diversification rate ($\lambda_0 - \mu_0$). Within each panel, each column indicates results of an evolutionary scenario (PD for phylogenetic diversity, ED for evolutionary distinctiveness, NND for nearest neighbor distance); each row indicates results of a diversification model complexity (which indicates the number of parameters used to generate the trees). In the left panels, colors indicate the values of absolute prediction errors; as the blue points become darker, the errors become smaller thus the predictions more accurate. The red dashed lines represent expected tree size with respect to true net diversification rate and a fixed crown age 10. The pink ribbons represent possible variations of tree sizes due to stochasticity. In the right panels, the colors indicate the values of true net diversification rate. Points falling close to the vertical black dashed lines are accurate predictions (zero error). X-axis (left panel): true value of the net diversification rate ($\lambda_0 - \mu_0$). X-axis (right panel): prediction error (absolute difference between true value and predicted value). Y-axis: tree size.

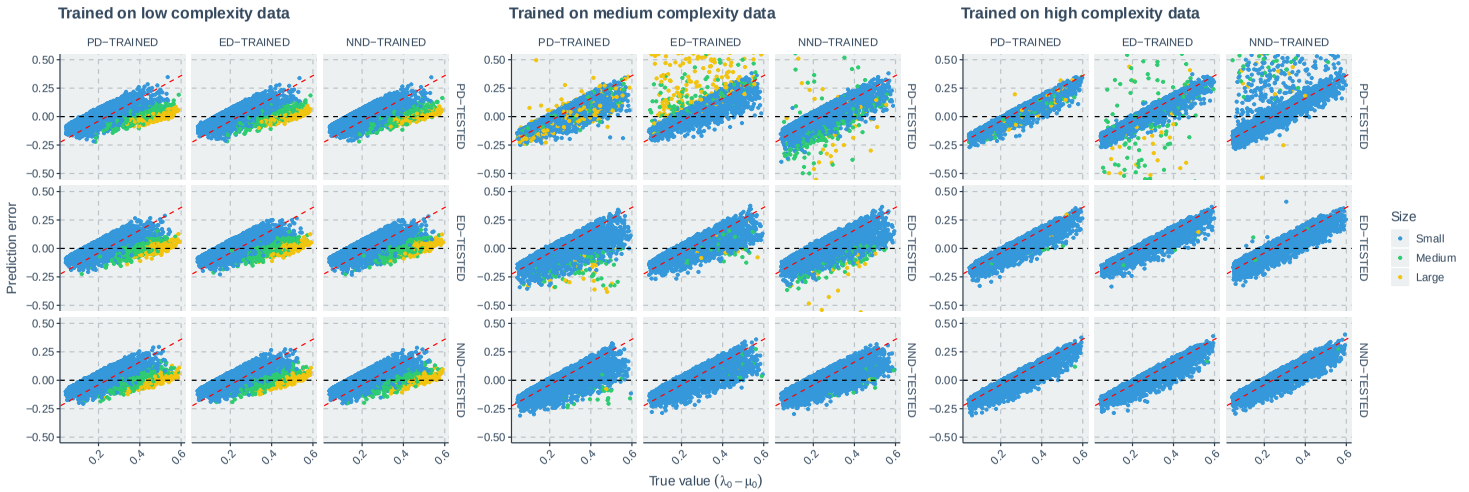


Figure 4.10: Comparison of neural network regression errors under misspecification. By misspecification, neural networks trained on trees under different evolutionary scenarios are used to estimate parameters on all trees, including those having different (misspecified) evolutionary scenarios. The three panels show the prediction error of three model complexity levels (which indicate the number of parameters used to generate the trees). Within each panel, each column indicates results of using neural networks trained on 3 evolutionary scenarios (see column facet strips, e.g. PD-TRAINED represents neural networks trained on trees generated under the phylogenetic diversity scenario) to estimate parameters on tree dataset generated under a particular scenario (see row facet strips, e.g. PD-TESTED indicate that neural network predictions were made on PD trees). Yellow points stand for results of small trees, green points stand for medium trees and blue points stand for large trees. X-axis: true value of the net diversification rate ($\lambda_0 - \mu_0$). Y-axis: prediction error (absolute difference between true value and predicted value).

4.4 Discussion

4.4.1 Conservative Predictions as Indicators of Limited Information

Our regression experiments show that, for most eve parameters and especially in the medium- and high-complexity settings, neural networks often return predictions close to the empirical mean of the training distribution (e.g., Figure 4.7, Figure 4.8, Appendix B, Appendix D). This pattern is strongest for the ER parameters (β_N, β_Φ, γ_N, γ_Φ), where slopes of the error-versus-truth relationships and alignment metrics indicate that predictions behave almost like constant conditional means. Under standard loss functions such as squared error or Huber loss, this behavior is expected whenever the inputs carry little information about the target: in low-signal situations, the risk-minimizing solution is to predict the mean of the conditional distribution rather than to chase noisy fluctuations (a similar situation was found in Qin et al. [147]). In our setting, the conservative predictions therefore point to parts of eve parameter space where extant trees potentially do not provide enough information for the networks to recover individual parameter values. Practically, this means that point estimates in those parts of parameter space should be interpreted as "typical values compatible with the data" rather than as precise predictions of the true parameters. Inference workflows that rely on such predictors should therefore report broad uncertainty, treat estimates as lower-dimensional summaries (e.g. of net diversification) or explicitly acknowledge that the available data support many alternative diversification histories, consistent with analytical results on non-identifiability of birth–death models from extant trees as discussed in the literature [153, 240].

4.4.2 Tree Size, Effect Sizes and Limits of Parameter Recovery

Across both classification and regression analyses, tree size and the strength of diversification effects emerge as the main determinants of how much information the trees potentially contain about the underlying process. For classification, F1 scores, precision and recall generally increase with tree size for PD and ED trees (Figure 4.3 and Figure 4.4). Very small trees are generally associated with very low scenario accuracy regardless of parameter values, which aligns with classical findings that tests of phylogenetic signal and model-based comparative methods have little power on small trees [242, 243], analogous to non-identifiability caused by small sample size in time-series hidden Markov models [241]. The pattern for NND trees is less clear because these trees are typically small and occupy a narrow range of sizes, so even the largest NND trees in our simulations contain fewer branching events than PD and ED trees of comparable parameter values. For regression, larger trees help most clearly for estimates of net diversification rate and γ_N, with errors decreasing with size when model complexity is low and, to a lesser extent, at medium complexity (Figure 4.7, Figure 4.9, Figure 4.14). However, once the full six-parameter eve model is used, size-related improvements become modest or minimal and predictions for several parameters collapse back towards the empirical mean. Importantly, the trees that yield the most accurate estimates are not those with extreme sizes but those whose sizes are close to the typical sizes produced by the simulation models (Figure 4.9 and Figure 4.29), suggesting that the networks effectively learn a mapping from size and a few coarse topological summaries to parameter averages, with limited ability to exploit more subtle shape information. After all, note that under our low-complexity setting (effectively

a constant-rate birth–death process), the unlabeled tree topology is known to be independent of the speciation and extinction rates and therefore carries no information about the generative parameter [244, 245]. It is not surprising that our GNN-based approaches bring no substantial gain (see also [147]). Together, these results indicate that working with larger clades of organisms helps only up to a point: larger trees improve recovery of a small number of parameters that tightly control overall diversification pace, but they do little for parameters that mainly modulate subtle ER patterns once diversity dependence and stochasticity are included.

Effect sizes play an equally important role. Across scenarios and architectures, strong species richness dependence ($|\beta_N|$ and $|\gamma_N|$) consistently reduces classification accuracy and degrades regression performance. When richness effects are large, tree size reacts sensitively (we would expect smaller trees), and the resulting variation in size can overwhelm the weaker signatures of ER effects, making it harder for the networks to distinguish PD, ED and NND or to disentangle the contributions of β_Φ and γ_Φ. In contrast, strong relatedness effects on speciation ($|\beta_\Phi|$) improve scenario separability: PD and NND trees with large positive β_Φ are classified more accurately, and correct trees in these scenarios preferentially occupy extreme β_Φ values while avoiding strongly negative β_N (Figure 4.5). For ED trees, shifting β_Φ from negative to positive values gradually improves performance, which aligns with a shift from antagonistic to reinforcing interactions between richness and relatedness on speciation (see [166] for how ED trees are generated). The combination of parameter-sliced accuracy curves, ridge-line distributions and statistical tests therefore provides an empirical map of the parts of parameter space where `eve` parameters and scenarios potentially leave strong enough imprints on tree shape to be recoverable, and the parts where their effects are washed out by stochasticity and richness dependence.

4.4.3 Scenario Overlap and Redundancy in Complex Models

Our scenario-level analyses show that, even when neural networks perform better than chance, PD, ED and NND are far from being cleanly separated in tree space. Confusion matrices reveal that NND trees are generally easiest to classify, while ED trees are consistently the most difficult (Figure 4.2). Many ED trees are mislabeled as NND, and PD trees are more often confused with NND than vice versa, indicating that the local ER mechanisms of ED and the neighborhood-based mechanism of NND frequently generate tree shapes that are compatible with the PD scenario. Increasing `eve` complexity exacerbates this problem: when the six-parameter model is used, misclassifications shift towards NND for all true scenarios, and scenario-specific recalls decline even though the networks are trained and tested on data generated exactly from the fitted model. The ridge-line summaries potentially also support the idea, as correctly and incorrectly classified trees often arise from overlapping ranges of net diversification and ER effect sizes, with clear separations only in parameter-size combinations where β_Φ is strong and tree size is moderate to large (see Figure 4.5). It can also be seen that regression and misspecification experiments tell a similar story. Even when the networks are trained on all scenarios simultaneously, estimates for ER parameters remain conservative and close to empirical means for most of the parameter space (Figure 4.8, Figure 4.10, Appendix G). We propose to visualize scenario redundancy as overlapping "clouds" of trees in representation space: the networks can detect broad differences between scenarios in parts of parameter space where ER effects

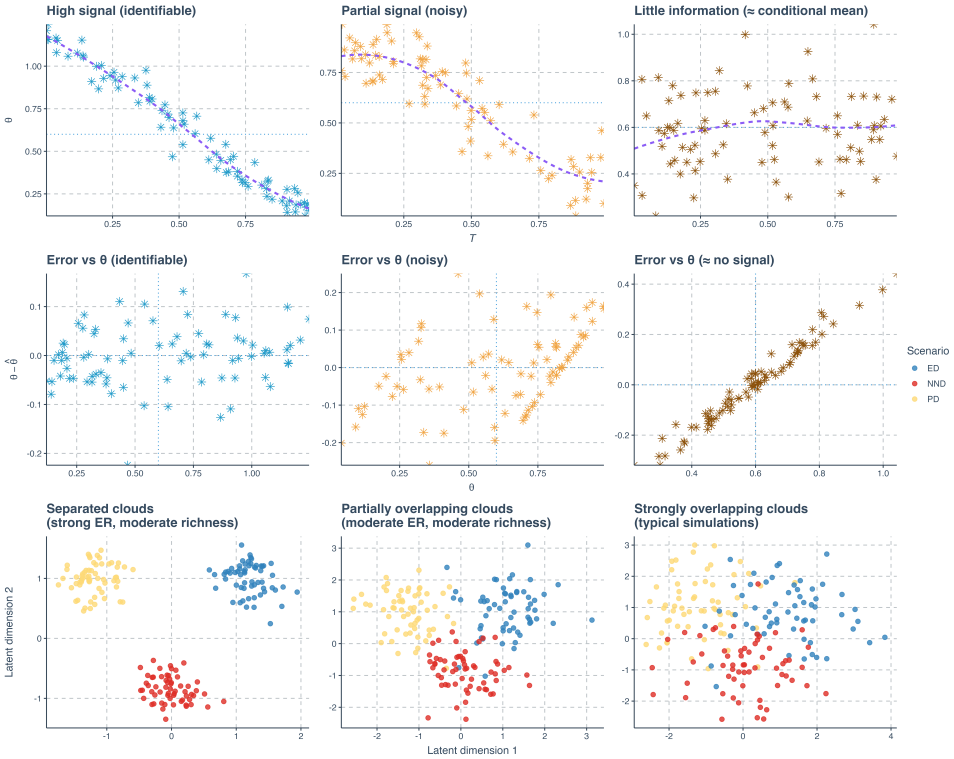


Figure 4.11: Illustration of how increasing model complexity (flexibility) reduces the information that tree summaries carry about diversification parameters and scenarios. **Top row:** for three hypothetical settings, points show simulated pairs of a tree-derived summary $\mathcal{T}$ and a parameter value θ, the violet dashed curve is a LOESS (locally estimated scatterplot smoothing) estimate of the conditional mean $\mathbb{E}[\theta \mid \mathcal{T}]$, and the blue horizontal line marks the unconditional mean $\mathbb{E}[\theta]$. From left to right, the relationship between $\mathcal{T}$ and θ weakens: a clear trend (high signal) is followed by a noisy trend (partial signal) and finally by an almost flat curve concentrated around the mean (little information, with many distinct parameter values producing similar summaries). **Middle row:** corresponding errors in the hypothetical estimation of parameters, where each point shows $\theta - \hat{\theta}$ against the true θ; as signal decreases, errors become more structured and eventually align with a straight line, indicating that predictions collapse toward an almost constant conditional mean. **Bottom row:** schematic two-dimensional "representation spaces" under two hypothetical latent dimensions for three diversification scenarios (PD, ED, NND) under analogous conditions. When ER effects are strong and richness dependence is moderate (left), the three scenarios occupy well-separated clouds; with weaker or more noisy signal (middle), the clouds partially overlap; in parameter ranges similar to our typical simulations (right), the clouds strongly overlap, illustrating that the networks can still detect broad differences among scenarios in some parts of parameter space, but cannot reliably assign individual trees to a unique scenario when their representations become nearly indistinguishable.

are strong and diversity dependence is moderate, but they cannot reliably assign trees to a unique scenario in the more typical ranges of parameters explored by our simulations.

We further illustrate how overlapping clusters of trees in a hypothetical representation space can obscure scenario-specific signals and drive conservative, mean-regressing predictions by our classifiers and regressors as we observed, in Figure 4.11. This schematic is generated from toy simulations.

4.4.4 Practical Lessons and Recommendations

Although our goal was not to build production-ready predictors, the results provide several practical lessons for using neural networks with complex diversification models. First, both classification and regression outputs are most trustworthy in specific combinations of parameter values and tree sizes. Scenario labels from our classifiers are most informative for large trees (roughly more than 200 nodes) with strong positive relatedness effects on speciation and moderate richness dependence; in those cases F1 scores approach 0.8-0.9 for some scenarios, confusion matrices are dominated by the diagonal (but not for ED), and parameter distributions of correct and incorrect trees differ. In contrast, for small trees, for weak or strongly negative β_Φ, or when $|\beta_N|$ and $|\gamma_N|$ are large, our results indicate that extant trees alone do not contain enough information to support confident scenario assignments. In applied settings with complex diversification models, we therefore recommend treating our classifier outputs (or other potential neural classifiers alike) as qualitative "scenario hints" outside of the high-information parts of parameter space, and avoiding strong biological interpretations of small differences in scenario probabilities when F1 scores are low.

Second, for parameter estimation, our regressors are primarily useful for tracking broad trends in net diversification and a few extinction-related effects, not for precise recovery of the full eve parameter vector. Estimates for $\lambda_0 - \mu_0$ and γ_N improve with tree size and, at low complexity, can capture coarse gradients across the parameter space; however, predictions for β_N, β_Φ and γ_Φ are often shrunk towards empirical means and show limited correlation with the truth. In practice, this means that neural networks trained on simulated eve trees could be used only as fast approximators for low-dimensional summaries, such as net diversification rate or composite ER indices, but should not be relied on for full parameter recovery unless extensive validation shows that the empirical trees lie in a high-information part of the simulated space.

Third, our calibration analysis highlights that miscalibration is most severe for the hardest scenario in terms of recoverability (ED) and for parameter ranges where relatedness effects are weak or ambiguous. Across most strata the GNN is overconfident, with predicted probabilities overstating actual accuracies by 10-30 percentage points. For downstream applications that require probabilistic statements, this suggests incorporating explicit calibration steps (e.g. temperature scaling or isotonic regression; [246–248]) trained on held-out simulated data, or moving to approaches that output predictive intervals or sets—for example, conformal prediction [249–251] or Bayesian neural architectures [149, 150, 252]—so that lack of information is expressed as wide uncertainty rather than overconfident point predictions. Such approaches cannot create information that is not present in the trees, but they can reduce the risk of over-interpretation.

Finally, our results underscore that attempts to fit highly flexible diversification models to single empirical trees should be paired with careful simulation-based diagnostics. Before estimating parameters or comparing complex scenarios on real data, one can run our workflow (or similar simulation pipelines) under plausible parameter ranges and tree sizes and check whether the empirical tree falls into a region where classification and regression errors are acceptably low. If not, simpler models tailored to specific hypotheses (e.g., omitting certain scenarios, constraining parameters to a lower-dimension, or focusing a subset of parameters) may be more appropriate. Complementary data sources (e.g., fossils, traits, spatial information or replicated trees across clades) are likely necessary to break the many-to-one mappings between diversification histories and extant trees [153].

4.5 Appendix

A) Total Loss

In regression tasks, total loss comprises three key components: Huber loss, link prediction loss and entropy of regularization. Huber loss was used for optimizing regression accuracy while the remaining components focused on alleviating a possible issue where GNN can be hard to train, if incorporating the differentiable pooling method [140].

The Huber loss [233] for vectors y and $\hat{y}$, each with n elements, computed as the average loss across all elements, is given by:

$$L_\delta(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2 & \text{for } |y_i - \hat{y}_i| \le \delta, \\ \delta(|y_i - \hat{y}_i| - \frac{1}{2}\delta) & \text{otherwise,} \end{cases} \tag{4.8}$$

where $\mathbf{y}$ is the true value vector comprising the ground truth parameters used for simulating a phylogenetic tree, $\hat{\mathbf{y}}$ is the predicted value vector comprising the parameter predictions, y_i and $\hat{y}_i$ are the i-th elements of $\mathbf{y}$ and $\hat{\mathbf{y}}$ respectively, n is the number of elements in the vectors $\mathbf{y}$ and $\hat{\mathbf{y}}$ and δ is the threshold parameter that defines the transition from squared to linear loss (here loss refers to the difference between ground truth and predicted values). In our research, we set $\delta = 0.8$ for all the training sessions, making the neural networks more sensitive to smaller errors and more robust to outliers .

The total loss function L_1 in regression is given by

$$L_1 = L_\delta(\mathbf{y}, \hat{\mathbf{y}}) + L_{\text{LP}} + L_{\text{E}}, \tag{4.9}$$

where L_{LP} is the link prediction loss and L_{E} is the entropy of regularization, see Ying et al. [140] for their definitions.

In classification tasks, we replaced the Huber loss component with cross-entropy loss for the purpose of multi-class classification. It measures the difference between the true class labels and the predicted probabilities, penalizing confident but incorrect predictions more heavily. For a set of n instances, the cross-entropy loss between the true labels $\mathbf{y}$ and the predicted probabilities $\hat{\mathbf{y}}$ is defined as:

$$L_{\text{CE}}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{n} \sum_{i=1}^{n} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}), \tag{4.10}$$

where n is the number of instances in the dataset, C is the total number of classes, $y_{i,c}$ is a binary indicator (0 or 1) if class label c is the correct classification for instance i, $\hat{y}_{i,c}$ is the predicted probability that instance i belongs to class c and $\log(\hat{y}_{i,c})$ is the natural logarithm of the predicted probability.

The total loss L_2 in classification is given by

$$L_2 = L_{\text{CE}}(\mathbf{y}, \hat{\mathbf{y}}) + L_{\text{LP}} + L_{\text{E}}, \tag{4.11}$$

where L_{LP} is the link prediction loss and L_{E} is the entropy of regularization.

B) Complete and Partial Data Training

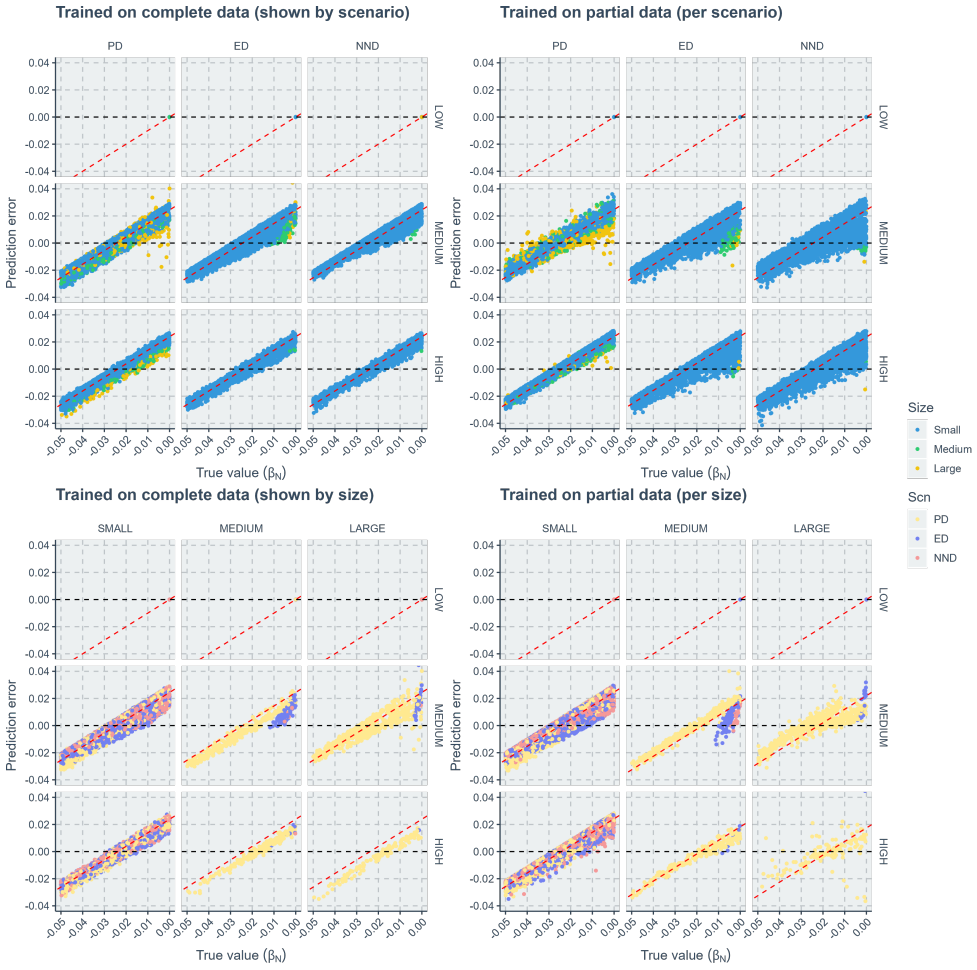


Figure 4.12: Comparison of neural network regression errors along true values of the species richness effect size on speciation (β_N). The left two panels show the results when the neural networks were trained on complete datasets. The right two panels show the results when the neural networks were trained on partial datasets (training datasets were sliced either by evolutionary scenarios or by tree size groups while in visualization we slice by both). Within each panel, each facet column indicates results of either an evolutionary scenario or a tree size group. Each facet row indicates results of a diversification model complexity. When columns represent scenario groups, blue points correspond to large trees, green points to medium trees, and yellow points to large trees. When columns represent tree size groups, light yellow points correspond to performances of trees generated under the phylogenetic diversity scenario, dark blue points to performances of trees under the evolutionary distinctiveness scenario, and red points to performances of trees under the nearest neighbor distance scenario. X-axis: true value of the species richness effect size on speciation (β_N). Y-axis: prediction error (absolute difference between true value and predicted value).

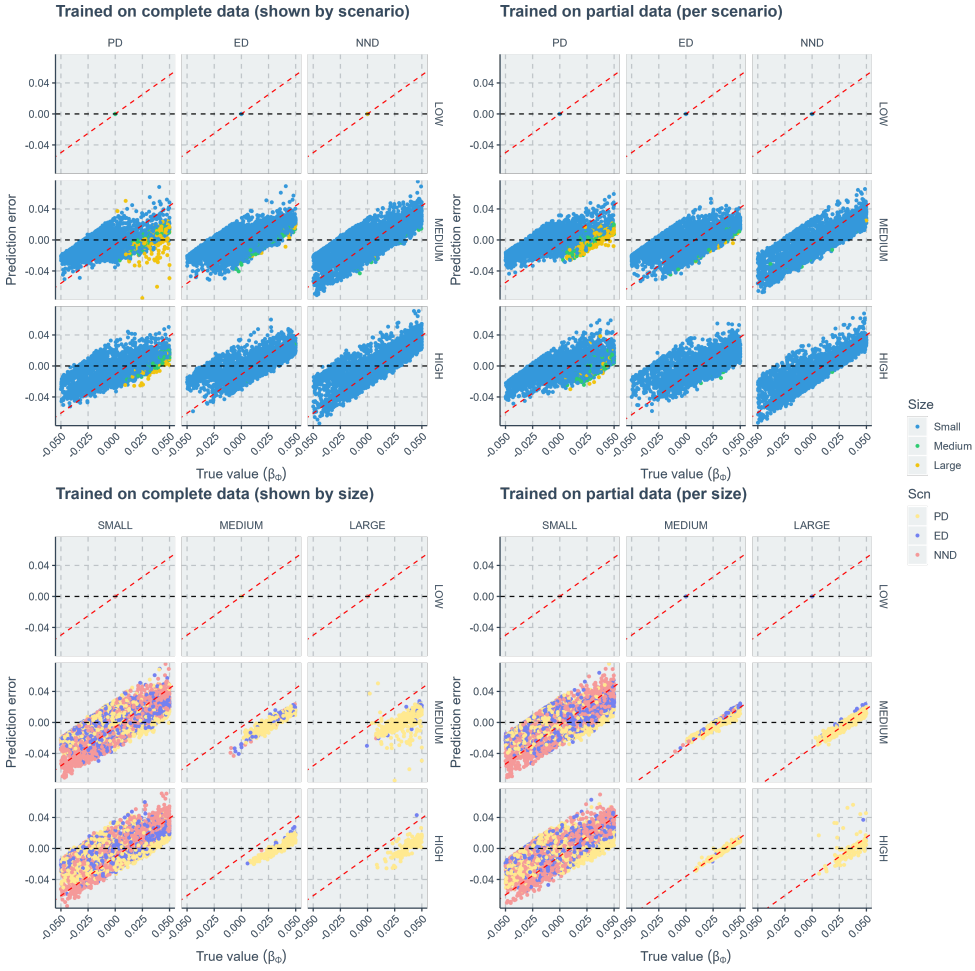


Figure 4.13: Comparison of neural network regression errors along true values of the evolutionary relatedness effect size on speciation (β_{Φ}). The left two panels show the results when the neural networks were trained on complete datasets. The right two panels show the results when the neural networks were trained on partial datasets (training datasets were sliced either by evolutionary scenarios or by tree size groups while in visualization we slice by both). Within each panel, each facet column indicates results of either an evolutionary scenario or a tree size group. Each facet row indicates results of a diversification model complexity. When columns represent scenario groups, blue points correspond to large trees, green points to medium trees, and yellow points to large trees. When columns represent tree size groups, light yellow points correspond to performances of trees generated under the phylogenetic diversity scenario, dark blue points to performances of trees under the evolutionary distinctiveness scenario, and red points to performances of trees under the nearest neighbor distance scenario. X-axis: true value of the evolutionary relatedness effect size on speciation (β_{Φ}). Y-axis: prediction error (absolute difference between true value and predicted value).

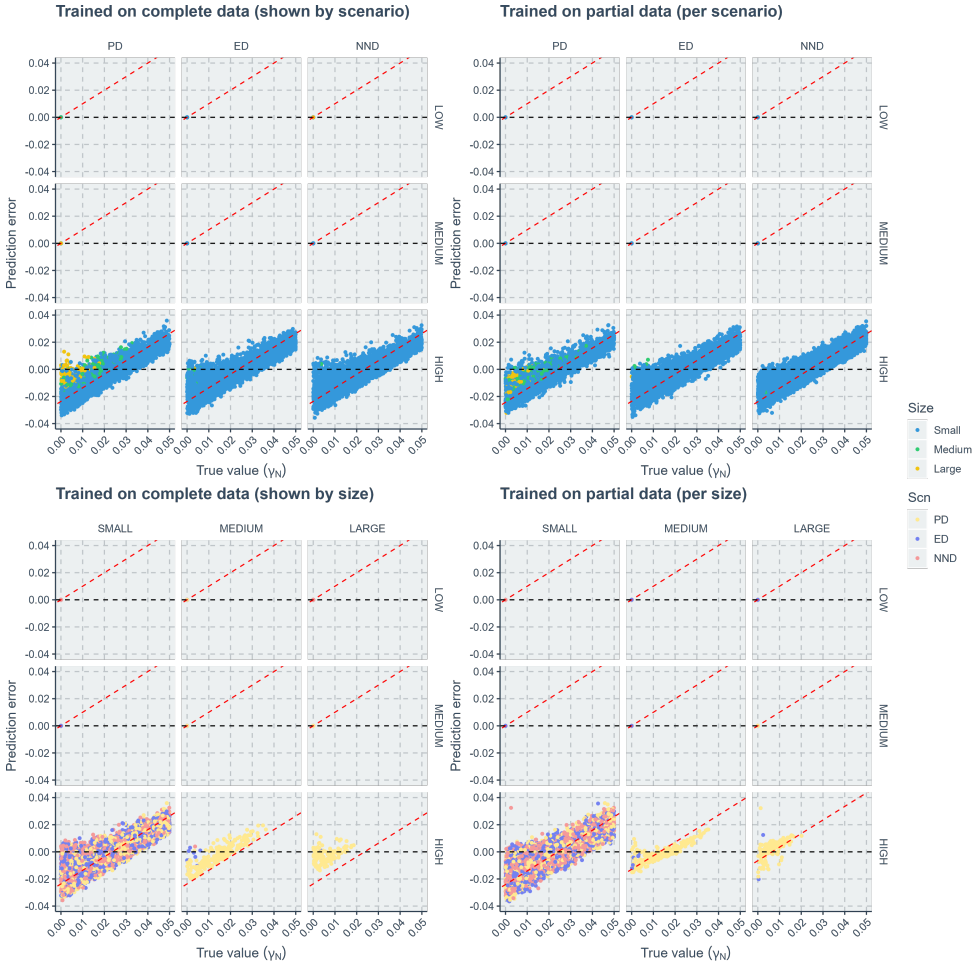


Figure 4.14: Comparison of neural network regression errors along true values of the species richness effect size on extinction (γ_N). The left two panels show the results when the neural networks were trained on complete datasets. The right two panels show the results when the neural networks were trained on partial datasets (training datasets were sliced either by evolutionary scenarios or by tree size groups while in visualization we slice by both). Within each panel, each facet column indicates results of either an evolutionary scenario or a tree size group. Each facet row indicates results of a diversification model complexity. When columns represent scenario groups, blue points correspond to large trees, green points to medium trees, and yellow points to large trees. When columns represent tree size groups, light yellow points correspond to performances of trees generated under the phylogenetic diversity scenario, dark blue points to performances of trees under the evolutionary distinctiveness scenario, and red points to performances of trees under the nearest neighbor distance scenario. X-axis: true value of the species richness effect size on extinction (γ_N). Y-axis: prediction error (absolute difference between true value and predicted value).

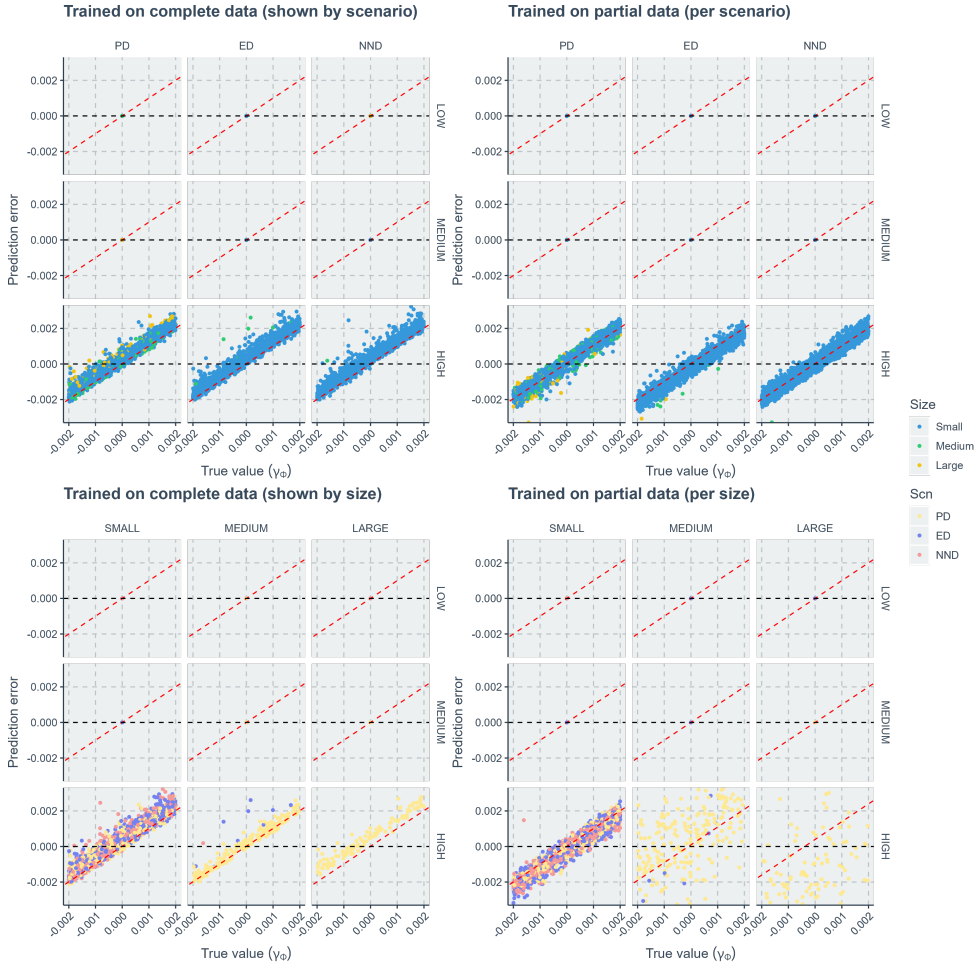


Figure 4.15: Comparison of neural network regression errors along true values of the evolutionary relatedness effect size on extinction (γ_Φ). The left two panels show the results when the neural networks were trained on complete datasets. The right two panels show the results when the neural networks were trained on partial datasets (training datasets were sliced either by evolutionary scenarios or by tree size groups while in visualization we slice by both). Within each panel, each facet column indicates results of either an evolutionary scenario or a tree size group. Each facet row indicates results of a diversification model complexity. When columns represent scenario groups, blue points correspond to large trees, green points to medium trees, and yellow points to large trees. When columns represent tree size groups, light yellow points correspond to performances of trees generated under the phylogenetic diversity scenario, dark blue points to performances of trees under the evolutionary distinctiveness scenario, and red points to performances of trees under the nearest neighbor distance scenario. X-axis: true value of the evolutionary relatedness effect size on extinction (γ_Φ). Y-axis: prediction error (absolute difference between true value and predicted value).

C) Contour Plots of the Point Estimates

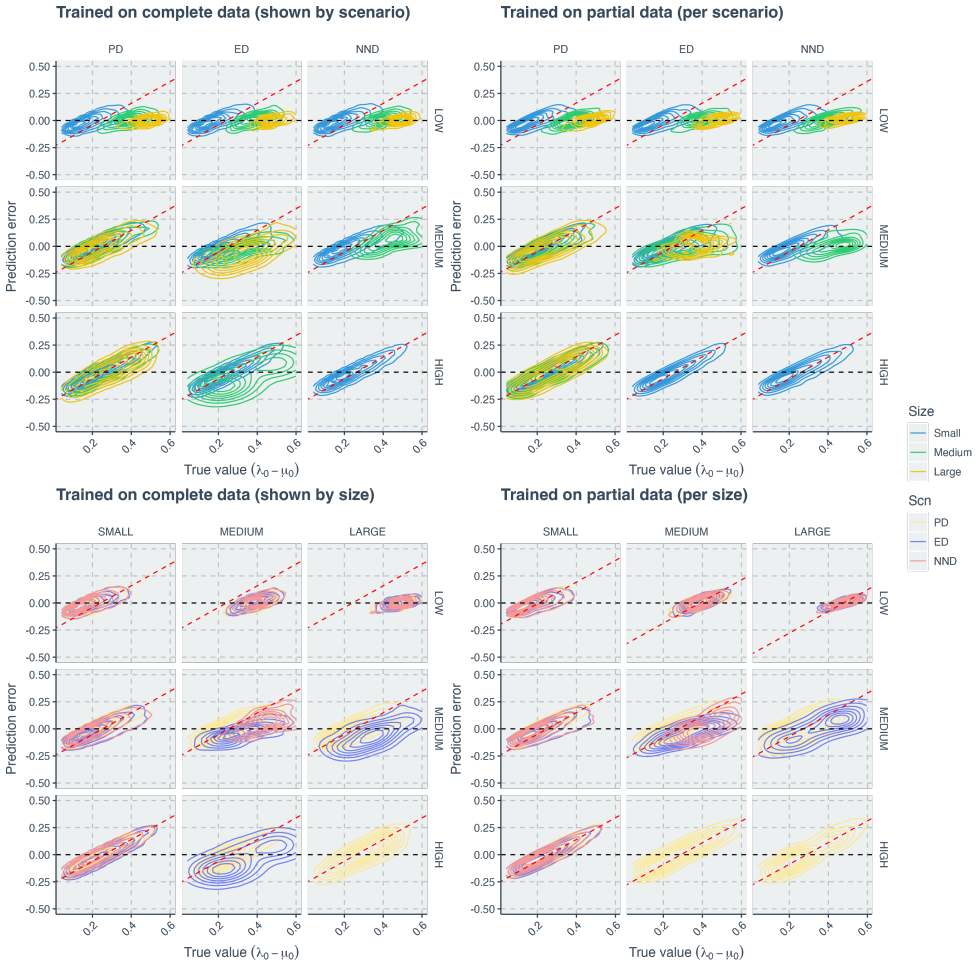


Figure 4.16: Comparison of neural network regression errors along true values of the net diversification rate ($\lambda_0 - \mu_0$), using contour plot instead of point cloud. The left two panels show the results when the neural networks were trained on complete datasets. The right two panels show the results when the neural networks were trained on partial datasets (training datasets were sliced either by evolutionary scenarios or by tree size groups while in visualization we slice by both). Within each panel, each facet column indicates results of either an evolutionary scenario or a tree size group. Each facet row indicates results of a diversification model complexity. When columns represent scenario groups, blue contours correspond to large trees, green contours to medium trees, and yellow contours to large trees. When columns represent tree size groups, light yellow contours correspond to performances of trees generated under the phylogenetic diversity scenario, dark blue contours to performances of trees under the evolutionary distinctiveness scenario, and red contours to performances of trees under the nearest neighbor distance scenario. X-axis: true value of the net diversification rate ($\lambda_0 - \mu_0$). Y-axis: prediction error (absolute difference between true value and predicted value).

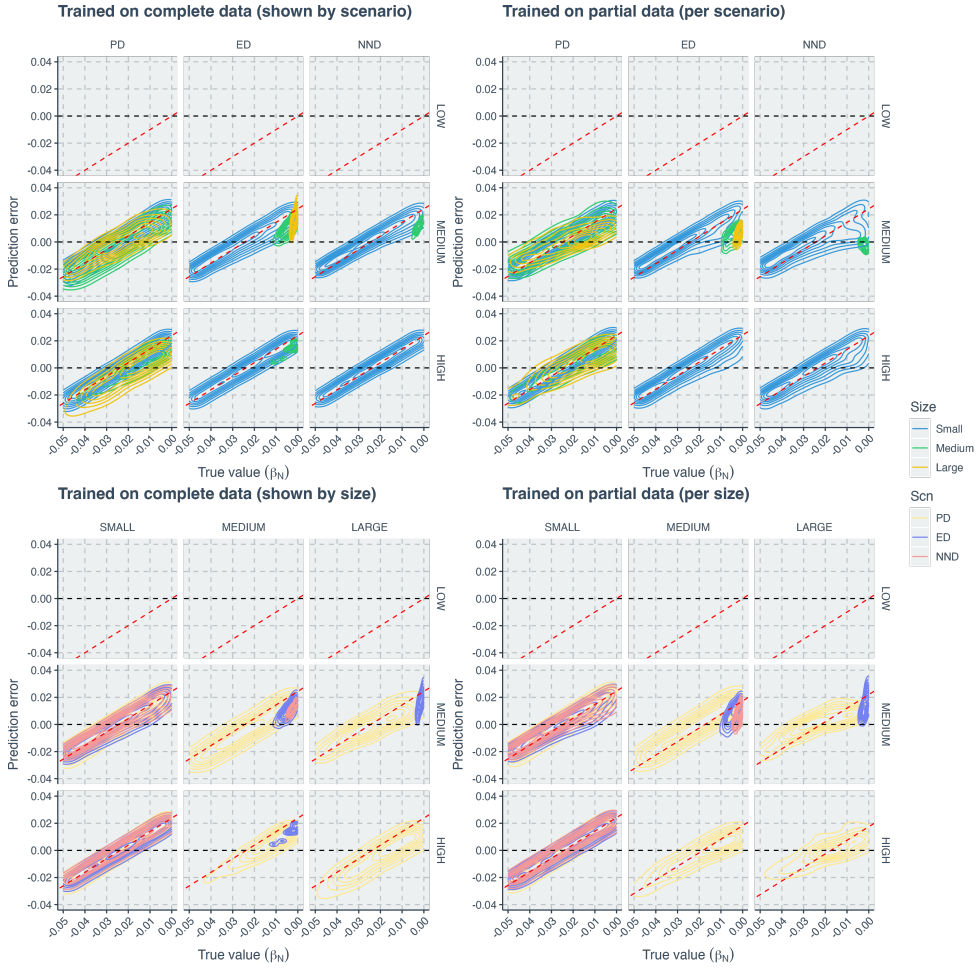


Figure 4.17: Comparison of neural network regression errors along true values of the species richness effect size on speciation (β_N), using contour plot instead of point cloud. The left two panels show the results when the neural networks were trained on complete datasets. The right two panels show the results when the neural networks were trained on partial datasets (training datasets were sliced either by evolutionary scenarios or by tree size groups while in visualization we slice by both). Within each panel, each facet column indicates results of either an evolutionary scenario or a tree size group. Each facet row indicates results of a diversification model complexity. When columns represent scenario groups, blue contours correspond to large trees, green contours to medium trees, and yellow contours to large trees. When columns represent tree size groups, light yellow contours correspond to performances of trees generated under the phylogenetic diversity scenario, dark blue contours to performances of trees under the evolutionary distinctiveness scenario, and red contours to performances of trees under the nearest neighbor distance scenario. X-axis: true value of the species richness effect size on speciation (β_N). Y-axis: prediction error (absolute difference between true value and predicted value).

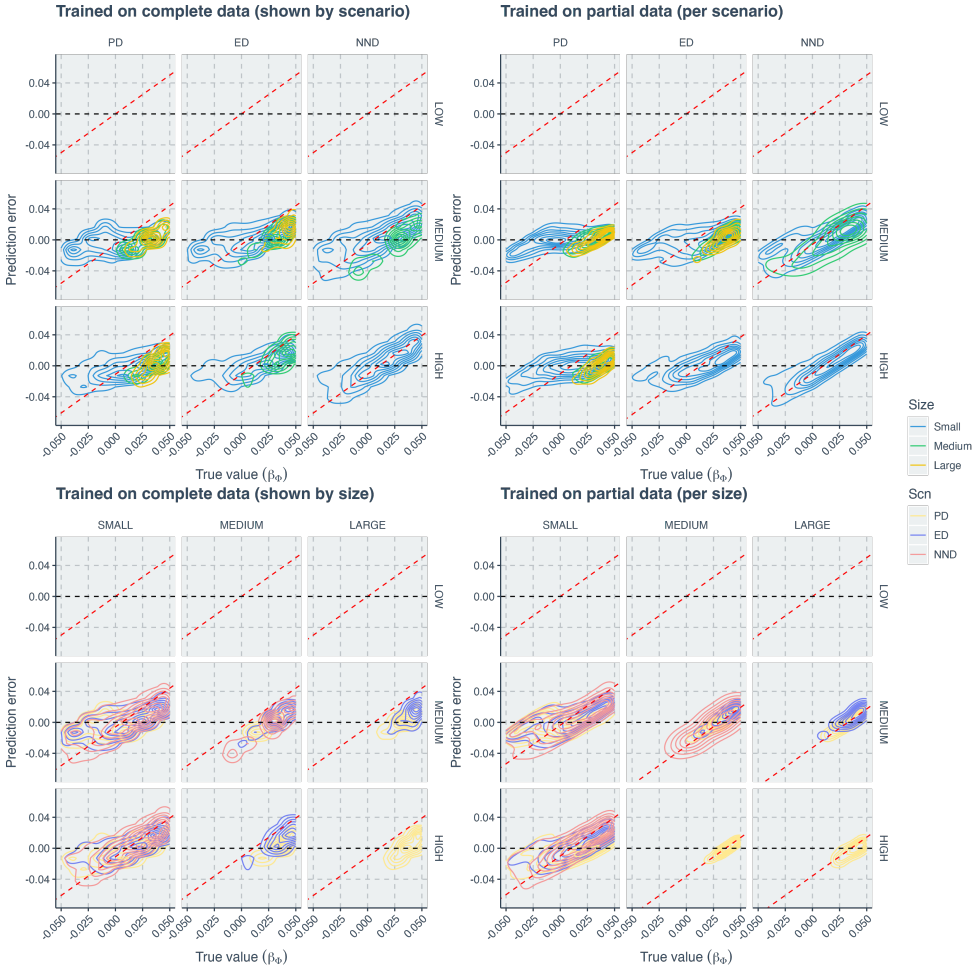


Figure 4.18: Comparison of neural network regression errors along true values of the evolutionary relatedness effect size on speciation (β_{Φ}), using contour plot instead of point cloud. The left two panels show the results when the neural networks were trained on complete datasets. The right two panels show the results when the neural networks were trained on partial datasets (training datasets were sliced either by evolutionary scenarios or by tree size groups while in visualization we slice by both). Within each panel, each facet column indicates results of either an evolutionary scenario or a tree size group. Each facet row indicates results of a diversification model complexity. When columns represent scenario groups, blue contours correspond to large trees, green contours to medium trees, and yellow contours to large trees. When columns represent tree size groups, light yellow contours correspond to performances of trees generated under the phylogenetic diversity scenario, dark blue contours to performances of trees under the evolutionary distinctiveness scenario, and red contours to performances of trees under the nearest neighbor distance scenario. X-axis: true value of the evolutionary relatedness effect size on speciation (β_{Φ}). Y-axis: prediction error (absolute difference between true value and predicted value).

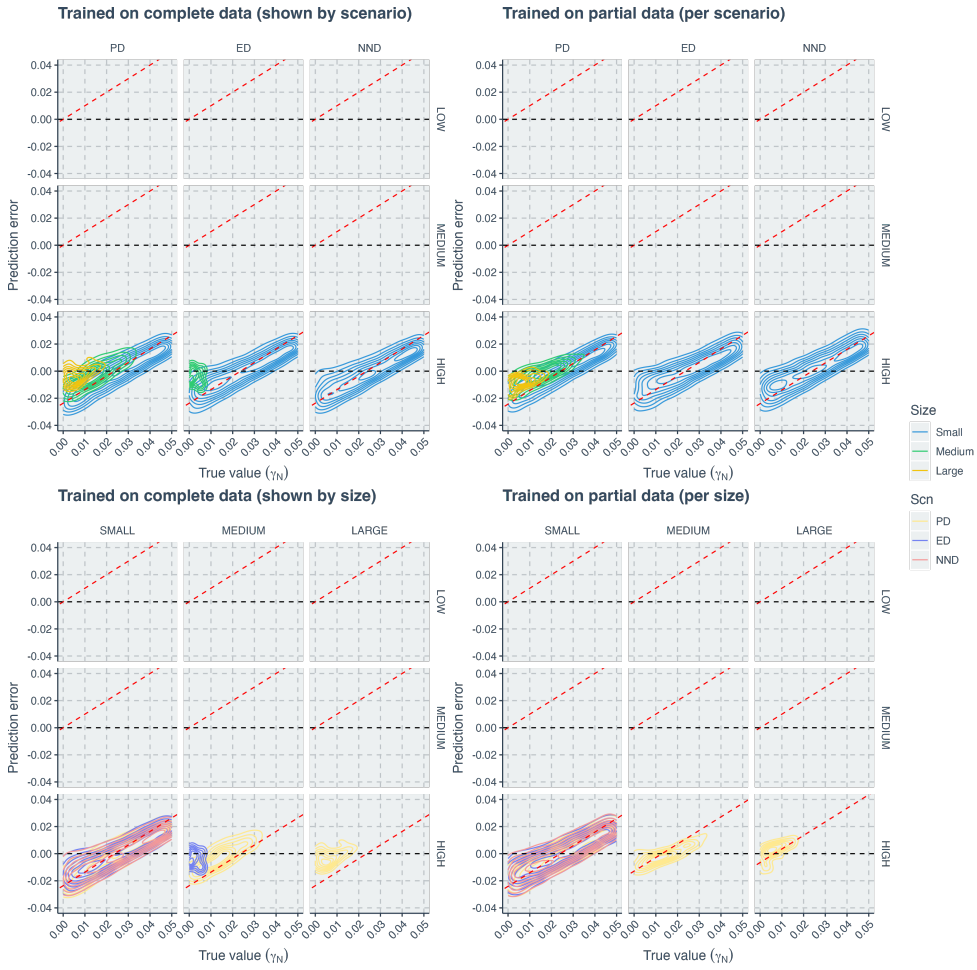


Figure 4.19: Comparison of neural network regression errors along true values of the species richness effect size on extinction (γ_N), using contour plot instead of point cloud. The left two panels show the results when the neural networks were trained on complete datasets. The right two panels show the results when the neural networks were trained on partial datasets (training datasets were sliced either by evolutionary scenarios or by tree size groups while in visualization we slice by both). Within each panel, each facet column indicates results of either an evolutionary scenario or a tree size group. Each facet row indicates results of a diversification model complexity. When columns represent scenario groups, blue contours correspond to large trees, green contours to medium trees, and yellow contours to large trees. When columns represent tree size groups, light yellow contours correspond to performances of trees generated under the phylogenetic diversity scenario, dark blue contours to performances of trees under the evolutionary distinctiveness scenario, and red contours to performances of trees under the nearest neighbor distance scenario. X-axis: true value of the species richness effect size on extinction (γ_N). Y-axis: prediction error (absolute difference between true value and predicted value).

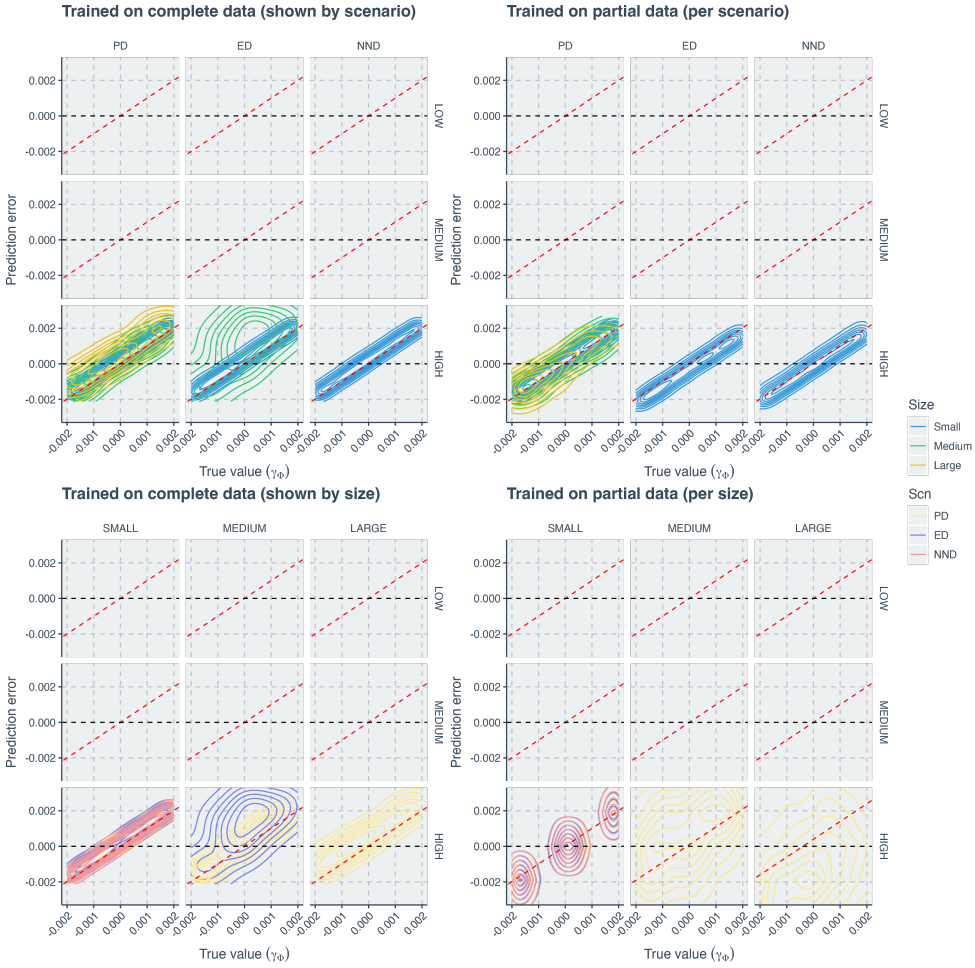


Figure 4.20: Comparison of neural network regression errors along true values of the evolutionary relatedness effect size on extinction (γ_Φ), using contour plot instead of point cloud. The left two panels show the results when the neural networks were trained on complete datasets. The right two panels show the results when the neural networks were trained on partial datasets (training datasets were sliced either by evolutionary scenarios or by tree size groups while in visualization we slice by both). Within each panel, each facet column indicates results of either an evolutionary scenario or a tree size group. Each facet row indicates results of a diversification model complexity. When columns represent scenario groups, blue contours correspond to large trees, green contours to medium trees, and yellow contours to large trees. When columns represent tree size groups, light yellow contours correspond to performances of trees generated under the phylogenetic diversity scenario, dark blue contours to performances of trees under the evolutionary distinctiveness scenario, and red contours to performances of trees under the nearest neighbor distance scenario. X-axis: true value of the evolutionary relatedness effect size on extinction (γ_Φ). Y-axis: prediction error (absolute difference between true value and predicted value).

D) Alignment with Conditional Mean

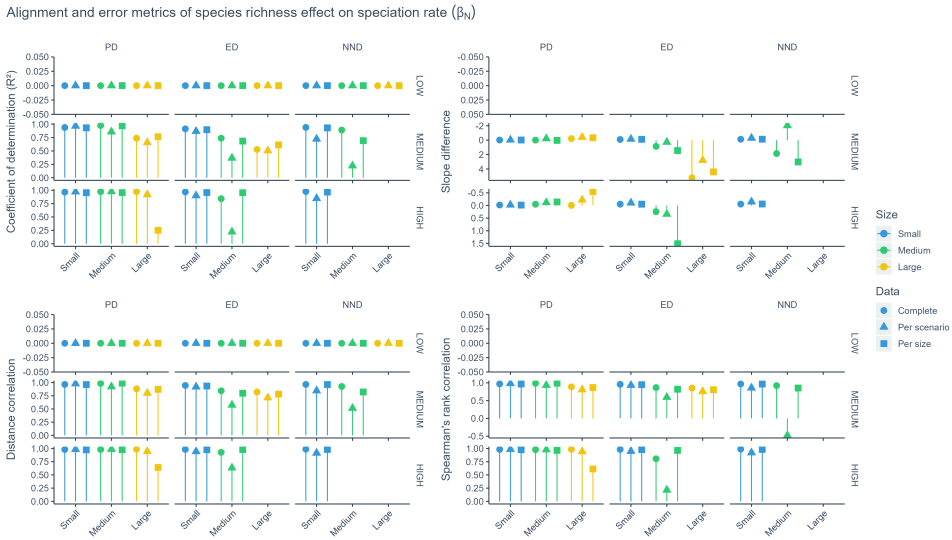


Figure 4.21: Comparison of the degree to which neural network regression estimates align with the conditional mean. There are four metrics, each describing a facet of the alignment, see Appendix I for details and interpretations. Within each panel, each facet column indicates results of a evolutionary scenario (PD for phylogenetic diversity, ED for evolutionary distinctiveness, NND for nearest neighbor distance); each facet row indicates results of a diversification model differing in complexity (as indicated by the number of parameters used to generate the trees). Yellow bars stand for metrics of small trees, green bars stand for medium trees and blue bars stand for large trees. At the top of the bars, shapes represent how the datasets were sliced to train the neural networks. Circles indicate complete datasets, triangles indicate slices by evolutionary scenario and squares indicate slices by tree size group. X-axis: size group of the trees. Y-axis: value of alignment metrics.

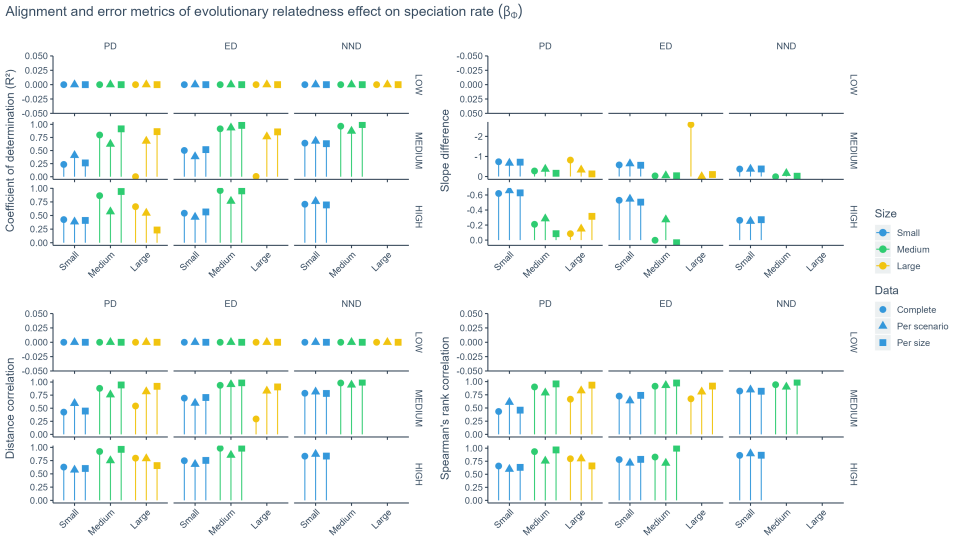


Figure 4.22: Comparison of the degree to which neural network regression estimates align with the conditional mean. There are four metrics, each describing a facet of the alignment, see Appendix I for details and interpretations. Within each panel, each facet column indicates results of a evolutionary scenario (PD for phylogenetic diversity, ED for evolutionary distinctiveness, NND for nearest neighbor distance); each facet row indicates results of a diversification model differing in complexity (as indicated by the number of parameters used to generate the trees). Yellow bars stand for metrics of small trees, green bars stand for medium trees and blue bars stand for large trees. At the top of the bars, shapes represent how the datasets were sliced to train the neural networks. Circles indicate complete datasets, triangles indicate slices by evolutionary scenario and squares indicate slices by tree size group. X-axis: size group of the trees. Y-axis: value of alignment metrics.

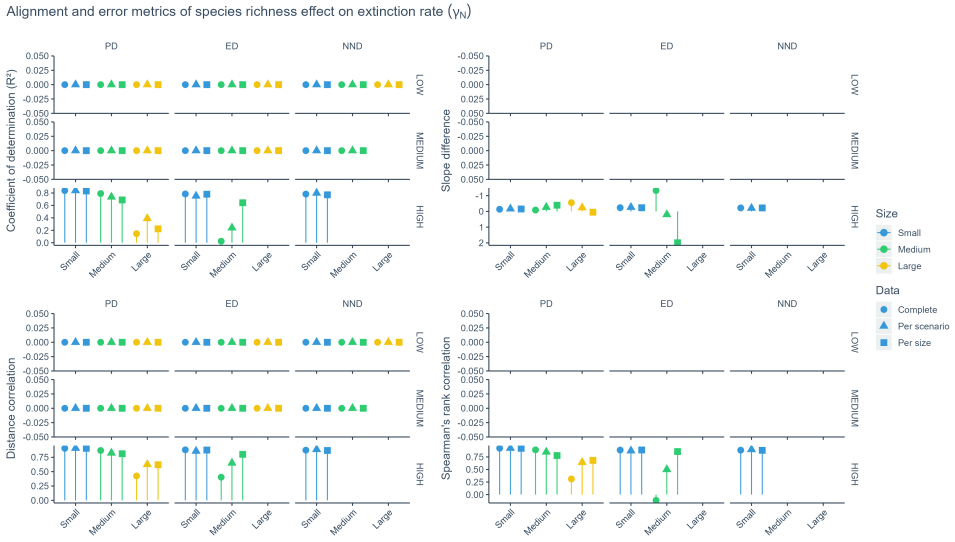


Figure 4.23: Comparison of the degree to which neural network regression estimates align with the conditional mean. There are four metrics, each describing a facet of the alignment, see Appendix I for details and interpretations. Within each panel, each facet column indicates results of a evolutionary scenario (PD for phylogenetic diversity, ED for evolutionary distinctiveness, NND for nearest neighbor distance); each facet row indicates results of a diversification model differing in complexity (as indicated by the number of parameters used to generate the trees). Yellow bars stand for metrics of small trees, green bars stand for medium trees and blue bars stand for large trees. At the top of the bars, shapes represent how the datasets were sliced to train the neural networks. Circles indicate complete datasets, triangles indicate slices by evolutionary scenario and squares indicate slices by tree size group. X-axis: size group of the trees. Y-axis: value of alignment metrics.

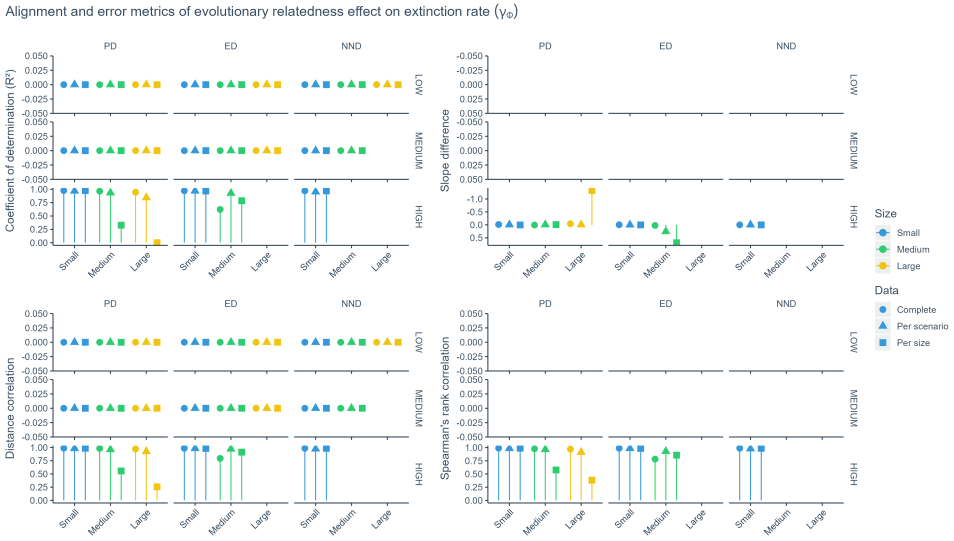


Figure 4.24: Comparison of the degree to which neural network regression estimates align with the conditional mean. There are four metrics, each describing a facet of the alignment, see Appendix I for details and interpretations. Within each panel, each facet column indicates results of a evolutionary scenario (PD for phylogenetic diversity, ED for evolutionary distinctiveness, NND for nearest neighbor distance); each facet row indicates results of a diversification model differing in complexity (as indicated by the number of parameters used to generate the trees). Yellow bars stand for metrics of small trees, green bars stand for medium trees and blue bars stand for large trees. At the top of the bars, shapes represent how the datasets were sliced to train the neural networks. Circles indicate complete datasets, triangles indicate slices by evolutionary scenario and squares indicate slices by tree size group. X-axis: size group of the trees. Y-axis: value of alignment metrics.

E) Distribution of Parameters

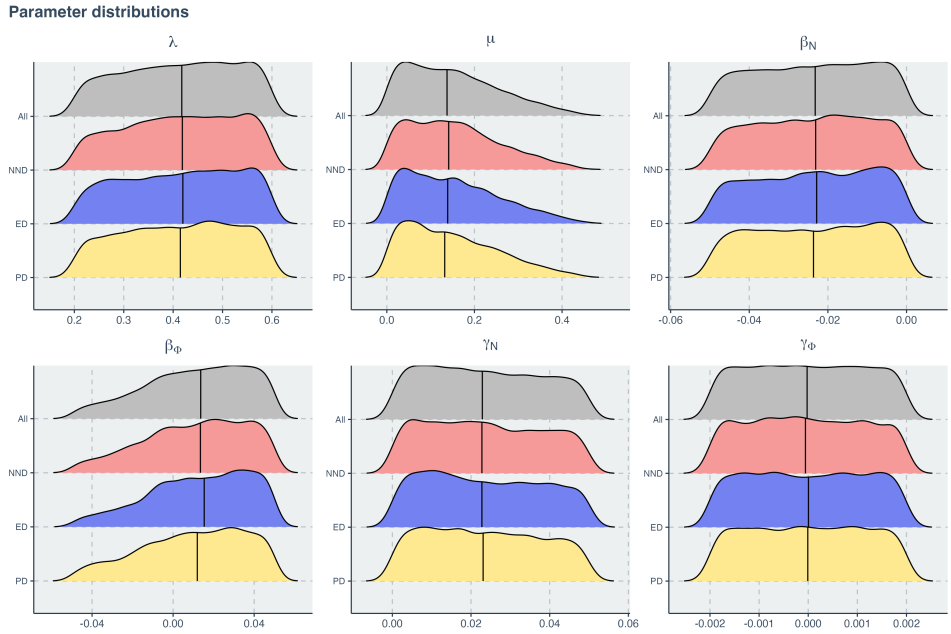


Figure 4.25: Generation parameter densities of phylogenetic trees in the training and validation dataset, before splitting. The densities of the trees in the test dataset are similar. Each panel displays the density of one tree generation parameter. Within panel, each ridge–line displays one angle view. The first ridge–line shows density of all trees. The second shows only NND (nearest-neighbor distance scenario) trees. The third shows only ED (evolutionary distinctiveness scenario) trees. The fourth shows only PD (phylogenetic diversity) trees.

F) Effect Size and Tree Size

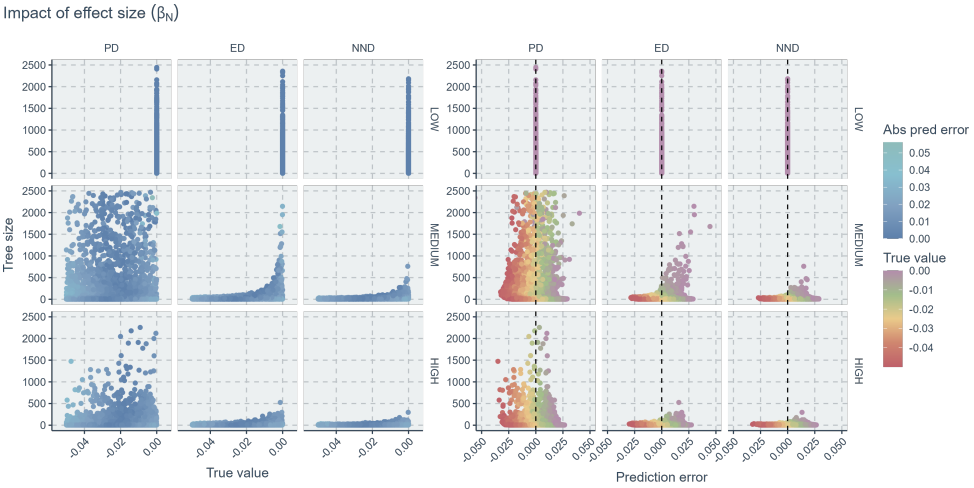


Figure 4.26: Relationships of tree size to true parameter values and prediction errors of the species richness effect size on speciation (β_N). Within each panel, each column indicates results of an evolutionary scenario (PD for phylogenetic diversity, ED for evolutionary distinctiveness, NND for nearest neighbor distance); each row indicates results of a diversification model complexity (which indicates the number of parameters used to generate the trees). In the left panels, colors indicate the values of absolute prediction errors; as the blue points become darker, the errors become smaller thus the predictions more accurate. In the right panels, the colors indicate the values of true parameter. Points falling close to the vertical black dashed lines are accurate predictions (zero error). X-axis (left panel): true parameter values. X-axis (right panel): prediction error (absolute difference between true value and predicted value). Y-axis: tree size.

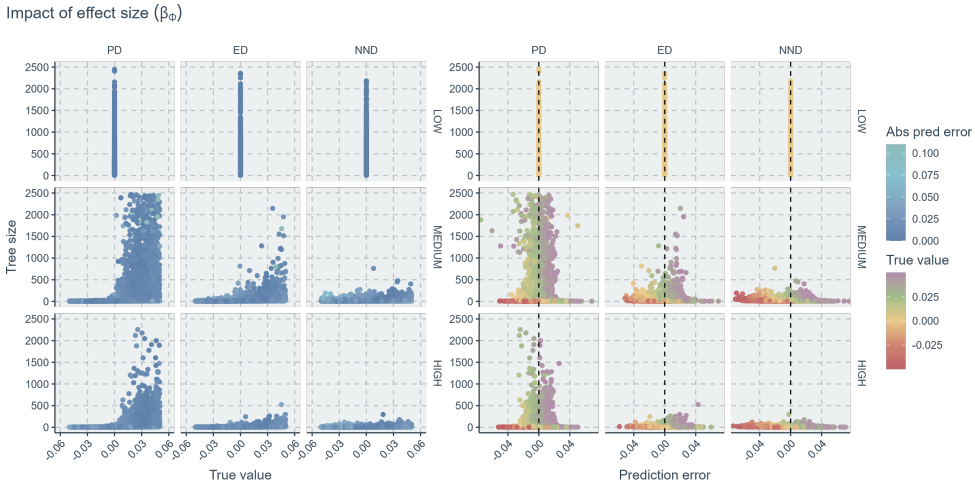


Figure 4.27: Relationships of tree size to true parameter values and prediction errors of the evolutionary relatedness effect size on speciation (β_Φ). Within each panel, each column indicates results of an evolutionary scenario (PD for phylogenetic diversity, ED for evolutionary distinctiveness, NND for nearest neighbor distance); each row indicates results of a diversification model complexity (which indicates the number of parameters used to generate the trees). In the left panels, colors indicate the values of absolute prediction errors; as the blue points become darker, the errors become smaller thus the predictions more accurate. In the right panels, the colors indicate the values of true parameter. Points falling close to the vertical black dashed lines are accurate predictions (zero error). X-axis (left panel): true parameter values. X-axis (right panel): prediction error (absolute difference between true value and predicted value). Y-axis: tree size.

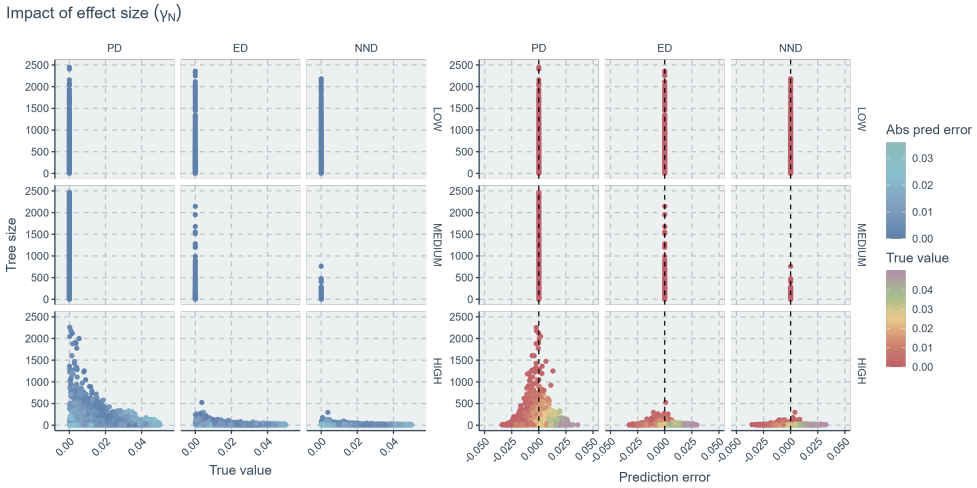


Figure 4.28: Relationships of tree size to true parameter values and prediction errors of the species richness effect size on extinction (γ_N). Within each panel, each column indicates results of an evolutionary scenario (PD for phylogenetic diversity, ED for evolutionary distinctiveness, NND for nearest neighbor distance); each row indicates results of a diversification model complexity (which indicates the number of parameters used to generate the trees). In the left panels, colors indicate the values of absolute prediction errors; as the blue points become darker, the errors become smaller thus the predictions more accurate. In the right panels, the colors indicate the values of true parameter. Points falling close to the vertical black dashed lines are accurate predictions (zero error). X-axis (left panel): true parameter values. X-axis (right panel): prediction error (absolute difference between true value and predicted value). Y-axis: tree size.

Figure 4.29: Relationships of tree size to true parameter values and prediction errors of the evolutionary relatedness effect size on extinction (γ_Φ). Within each panel, each column indicates results of an evolutionary scenario (PD for phylogenetic diversity, ED for evolutionary distinctiveness, NND for nearest neighbor distance); each row indicates results of a diversification model complexity (which indicates the number of parameters used to generate the trees). In the left panels, colors indicate the values of absolute prediction errors; as the blue points become darker, the errors become smaller thus the predictions more accurate. In the right panels, the colors indicate the values of true parameter. Points falling close to the vertical black dashed lines are accurate predictions (zero error). X-axis (left panel): true parameter values. X-axis (right panel): prediction error (absolute difference between true value and predicted value). Y-axis: tree size.

G) Regressor Misspecification

This appendix reports misspecification tests for the neural-network regressors. Networks trained under one evolutionary scenario are evaluated on trees generated under the same or different scenarios, and the resulting absolute prediction errors are shown for four key parameters controlling richness and evolutionary-relatedness effects on speciation and extinction (β_N, β_Φ, γ_N, γ_Φ). Each figure is organized by model complexity (panels), training scenario (columns), testing scenario (rows), and tree size (point colors), highlighting how robust parameter recovery is to scenario mismatch.

(Figures on next page.)

Figure 4.30: Comparison of neural network regression errors under misspecification. By misspecification, neural networks trained on trees under different evolutionary scenarios are used to estimate parameters on all trees, including those having different (misspecified) evolutionary scenarios. The three panels show the prediction error of three model complexity levels (which indicate the number of parameters used to generate the trees). Within each panel, each column indicates results of using neural networks trained on 3 evolutionary scenarios (see column facet strips, e.g. PD-TRAINED represents neural networks trained on trees generated under the phylogenetic diversity scenario) to estimate parameters on tree dataset generated under a particular scenario (see row facet strips, e.g. PD-TESTED indicate that neural network predictions were made on PD trees). Yellow points stand for results of small trees, green points stand for medium trees and blue points stand for large trees. X-axis: true value of the species richness effect size on speciation (β_N). Y-axis: prediction error (absolute difference between true value and predicted value).

Figure 4.31: Comparison of neural network regression errors under misspecification. By misspecification, neural networks trained on trees under different evolutionary scenarios are used to estimate parameters on all trees, including those having different (misspecified) evolutionary scenarios. The three panels show the prediction error of three model complexity levels (which indicate the number of parameters used to generate the trees). Within each panel, each column indicates results of using neural networks trained on 3 evolutionary scenarios (see column facet strips, e.g. PD-TRAINED represents neural networks trained on trees generated under the phylogenetic diversity scenario) to estimate parameters on tree dataset generated under a particular scenario (see row facet strips, e.g. PD-TESTED indicate that neural network predictions were made on PD trees). Yellow points stand for results of small trees, green points stand for medium trees and blue points stand for large trees. X-axis: true value of the evolutionary relatedness effect size on speciation (β_Φ). Y-axis: prediction error (absolute difference between true value and predicted value).

Figure 4.32: Comparison of neural network regression errors under misspecification. By misspecification, neural networks trained on trees under different evolutionary scenarios are used to estimate parameters on all trees, including those having different (misspecified) evolutionary scenarios. The three panels show the prediction error of three model complexity levels (which indicate the number of parameters used to generate the trees). Within each panel, each column indicates results of using neural networks trained on 3 evolutionary scenarios (see column facet strips, e.g. PD-TRAINED represents neural networks trained on trees generated under the phylogenetic diversity scenario) to estimate parameters on tree dataset generated under a particular scenario (see row facet strips, e.g. PD-TESTED indicate that neural network predictions were made on PD trees). Yellow points stand for results of small trees, green points stand for medium trees and blue points stand for large trees. X-axis: true value of the species richness effect size on extinction (γ_N). Y-axis: prediction error (absolute difference between true value and predicted value).

Figure 4.33: Comparison of neural network regression errors under misspecification. By misspecification, neural networks trained on trees under different evolutionary scenarios are used to estimate parameters on all trees, including those having different (misspecified) evolutionary scenarios. The three panels show the prediction error of three model complexity levels (which indicate the number of parameters used to generate the trees). Within each panel, each column indicates results of using neural networks trained on 3 evolutionary scenarios (see column facet strips, e.g. PD-TRAINED represents neural networks trained on trees generated under the phylogenetic diversity scenario) to estimate parameters on tree dataset generated under a particular scenario (see row facet strips, e.g. PD-TESTED indicate that neural network predictions were made on PD trees). Yellow points stand for results of small trees, green points stand for medium trees and blue points stand for large trees. X-axis: true value of the evolutionary relatedness effect size on extinction (γ_{Φ}). Y-axis: prediction error (absolute difference between true value and predicted value).

H) Details of Classification Performance Metrics

This appendix summarizes the notation and formulas used for the classification metrics reported in the main text. Let $y_i \in \{\mathrm{PD}, \mathrm{ED}, \mathrm{NND}\}$ denote the true class of tree i, and $\hat{y}_i$ the predicted class (with the largest probability) returned by a classifier. For a given target class k, we define the usual entries of the 2×2 contingency table as

$$\mathrm{TP}_k = \sum_i \mathbb{I}(y_i = k, \hat{y}_i = k), \quad \mathrm{FP}_k = \sum_i \mathbb{I}(y_i \neq k, \hat{y}_i = k), \tag{4.12}$$

$$\mathrm{FN}_k = \sum_i \mathbb{I}(y_i = k, \hat{y}_i \neq k), \quad \mathrm{TN}_k = \sum_i \mathbb{I}(y_i \neq k, \hat{y}_i \neq k), \tag{4.13}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The class-wise precision, recall, and F1-score are then given by

$$\mathrm{Prec}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}, \qquad \mathrm{Rec}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}, \tag{4.14}$$

$$\mathrm{F1}_k = \frac{2\,\mathrm{Prec}_k\,\mathrm{Rec}_k}{\mathrm{Prec}_k + \mathrm{Rec}_k}, \tag{4.15}$$

with the convention that ratios with zero denominators are treated as undefined and omitted from macro-averages. Overall (micro-averaged) accuracy is defined as

$$\mathrm{Acc} = \frac{1}{N} \sum_i \mathbb{I}(y_i = \hat{y}_i), \tag{4.16}$$

where N is the total number of test trees. Unless stated otherwise, we report macro-averaged precision, recall, and F1-score obtained by taking the unweighted arithmetic mean of the class-wise scores.

I) Definitions and Interpretations of the Alignment Metrics

Let x denote the true value of the quantity under investigation and let

$$y = x - \hat{x} \tag{4.17}$$

represent the prediction error, where $\hat{x}$ is the neural network prediction. In each subgroup, a linear model is fitted via

$$y = \beta_0 + \beta_1 x. \tag{4.18}$$

The coefficient of determination, denoted by R^2, is used as a metric of prediction alignment to the conditional mean. It is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{4.19}$$

where $\hat{y}_i$ are the fitted values from the regression and $\bar{y}$ is the mean of the observed y values. A high R^2 indicates that a large proportion of the variability in the prediction error is explained by the true values. A low R^2 implies that the neural network is capturing additional patterns beyond this baseline.

The slope difference is defined as

$$\Delta_{\text{slope}} = \hat{\beta}_1 - 1, \tag{4.20}$$

where $\hat{\beta}_1$ is the estimated slope from the regression. When predictions are simply the conditional mean, one would expect a near-unit slope; thus, a slope difference close to zero indicates strong alignment with the conditional mean. A large absolute value of Δ_{slope} suggests that the fitted relationship deviates substantially from a unit slope, implying that the neural network is capturing additional signal beyond the conditional mean.

Distance correlation is employed to quantify the overall dependence between x and y without assuming any specific functional form. It is defined as

$$\text{dCor}(x, y) = \frac{\text{dCov}(x, y)}{\sqrt{\text{dCov}(x, x)\,\text{dCov}(y, y)}}, \tag{4.21}$$

where dCov denotes the distance covariance. A high distance correlation indicates a strong association between the true values and the prediction error. A low distance correlation, on the other hand, suggests that the neural network is exploiting additional information.

Spearman's rank correlation coefficient provides a robust, nonparametric measure of the monotonic association between x and y. When the network simply reports the conditional mean, the prediction error y should vary monotonically with x (for example, as $y = x - c$ for some constant c), resulting in a high absolute Spearman correlation (close to 1). A Spearman correlation near zero, however, indicates that the prediction error does not follow a monotonic trend with respect to x; this lack of an association implies that the network's predictions capture additional patterns or signals beyond the conditional mean.

J) Parameter and Tree Size Slices

Table 4.2: Parameter and tree size ranges used for slicing classification performance results.

Parameter or Size	Ranges
λ_0	[0.2, 0.3), [0.3, 0.4), [0.4, 0.5), [0.5, 0.6)
μ_0	[0.00, 0.12), [0.12, 0.24), [0.24, 0.36), [0.36, 0.48)
β_N	[-0.05, -0.04), [-0.04, -0.03), [-0.03, -0.02), [-0.02, -0.01), [-0.01, 0.00)
β_Φ	[-0.05, -0.04), [-0.04, -0.03), [-0.03, -0.02), [-0.02, -0.01), [-0.01, 0.00), [0.00, 0.01), [0.01, 0.02), [0.02, 0.03), [0.03, 0.04), [0.04, 0.05)
γ_N	[0.00, 0.01), [0.01, 0.02), [0.02, 0.03), [0.03, 0.04), [0.04, 0.05)
γ_Φ	[-0.002, -0.001), [-0.001, 0.000), [0.000, 0.001), [0.001, 0.002)
Tree Size	[0, 20), [20, 40), [40, 60), [60, 80), [80, 100), [100, 120), [120, 140), [140, 160), [160, $+\infty$)

K) Mean, Variance, and Approximate Confidence Interval

Consider a linear birth–death process $\{N(t)\}_{t\geq 0}$ with $N(0) = 1$. Each lineage gives birth at rate λ and dies at rate μ, independently across lineages. Let $r = \lambda - \mu$.

Mean

Let $M(t) = \mathbb{E}[N(t)]$. Conditional on $N(t) = n$, the total birth rate is $n\lambda$ (increasing the count by 1) and the total death rate is $n\mu$ (decreasing the count by 1). Taking expectations yields

$$\frac{\mathrm{d}}{\mathrm{d}t}M(t) = \mathbb{E}[\lambda N(t) - \mu N(t)] = (\lambda - \mu)M(t) = rM(t), \qquad M(0) = 1. \tag{4.22}$$

Hence,

$$\boxed{\mathbb{E}[N(t)] = M(t) = e^{rt}.} \tag{4.23}$$

Second Moment and Variance

Write $S(t) = \mathbb{E}[N(t)^2]$. Using the standard jump-moment calculation, conditional on $N(t) = n$,

$$N \longrightarrow n+1 \text{ at rate } n\lambda, \qquad N \longrightarrow n-1 \text{ at rate } n\mu. \tag{4.24}$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}S(t) = \mathbb{E}\left[n\lambda\big((n+1)^2 - n^2\big) + n\mu\big((n-1)^2 - n^2\big)\right]. \tag{4.25}$$

Since $(n+1)^2 - n^2 = 2n + 1$ and $(n-1)^2 - n^2 = -2n + 1$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}S(t) = \lambda\,\mathbb{E}[2N(t)^2 + N(t)] + \mu\,\mathbb{E}[-2N(t)^2 + N(t)] = 2rS(t) + (\lambda + \mu)M(t). \tag{4.26}$$

With $M(t) = e^{rt}$ and $S(0) = \mathbb{E}[N(0)^2] = 1$, we solve the linear ODE

$$S'(t) - 2rS(t) = (\lambda + \mu)e^{rt}. \tag{4.27}$$

Case $\lambda \neq \mu$ (i.e. $r \neq 0$). Using the integrating factor e^{-2rt},

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(S(t)e^{-2rt}\big) = (\lambda + \mu)e^{-rt}, \tag{4.28}$$

so

$$S(t) = \frac{2\lambda}{r}e^{2rt} - \frac{\lambda + \mu}{r}e^{rt}. \tag{4.29}$$

Hence,

$$\mathrm{Var}[N(t)] = S(t) - M(t)^2 = \frac{\lambda + \mu}{r}\,e^{rt}\big(e^{rt} - 1\big). \tag{4.30}$$

That is,

$$\boxed{\mathrm{Var}[N(t)] = \frac{\lambda + \mu}{\lambda - \mu}\,e^{(\lambda-\mu)t}\Big(e^{(\lambda-\mu)t} - 1\Big), \qquad \lambda \neq \mu.} \tag{4.31}$$

Critical case $\lambda = \mu$ (i.e. $r = 0$). Then $M(t) \equiv 1$, and the moment equation reduces to $S'(t) = 2\lambda M(t) = 2\lambda$ with $S(0) = 1$, giving

$$S(t) = 1 + 2\lambda t, \qquad \boxed{\mathrm{Var}[N(t)] = S(t) - M(t)^2 = 2\lambda t.} \tag{4.32}$$

Approximate Normal Confidence Interval for $N(t)$
For sufficiently large expected counts (typically when $\lambda > \mu$ and t is not too small), one may use a crude normal approximation

$$N(t) \approx \mathcal{N}\big(M(t), \mathrm{Var}[N(t)]\big). \tag{4.33}$$

Let $z_{1-\alpha/2}$ denote the standard normal critical value (e.g. $z_{0.975} \approx 1.96$). An approximate two-sided $(1 - \alpha)$ interval is

$$\boxed{M(t) \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}[N(t)]}.} \tag{4.34}$$

5

Conclusion

This thesis set out to understand what phylogenies of extant species can and cannot tell us about the processes that generated them. Across three chapters, I focused on two intertwined themes: (i) how ecological limits and diversity feedback can be expressed not only through species richness but also through evolutionary relatedness, and (ii) how far modern inference tools—from likelihood-based estimators to neural networks—can reliably recover diversification mechanisms and parameters from the information encoded in reconstructed trees. Taken together, these chapters support three overarching conclusions about diversification inference from extant phylogenies:

1) **Evolutionary relatedness offers a flexible extension of ecological limits models, but its effects are scale-dependent.** Evolutionary relatedness can modulate interactions among lineages in ways that are not captured by richness alone. Yet "relatedness dependence" is not a single hypothesis: lineage-specific and clade-wide effects generate qualitatively different macroevolutionary patterns, and they should be treated as distinct modeling assumptions rather than interchangeable parameterizations.

2) **Inference is constrained less by the choice of estimator than by the information content of trees.** Likelihood-based estimators and neural networks can both perform well when the process potentially leaves a clear imprint, and both fail when the imprint is weak. Tree size and effect strength set hard limits: when multiple mechanisms yield nearly indistinguishable trees, improvement on the estimators may not fully compensate for missing information.

3) **Model flexibility comes with an identifiability cost that manifests as conservative, mean-regressing predictions.** As models become more richly parameterized, the mapping from parameters and scenarios to observed tree patterns flattens over much of parameter space. This produces systematic shrinkage in regression and asymmetric confusion in classification, providing a practical warning sign: apparent predictive accuracy can hide broad non-identifiability.

Bibliography

[1] C. A. Augier. *Essai d'une Nouvelle Classification Des Végétaux, Conforme Àl'ordre Que La Nature Paroìt Avoir Suivi Dans Le Règne Végétal ; d'où Résulte Une Méthode Qui Conduit à La Connoissance Des Plantes & de Leurs Rapports Naturels.* Bruyset Ainé et Comp., 1801.

[2] N. P. Hellström, G. André, and M. Philippe. Life and works of Augustin Augier de Favas (1758–1825), author of "Arbre botanique" (1801). *Archives of Natural History*, 44(1):43–62, April 2017. ISSN 0260-9541. doi: 10.3366/anh.2017.0413.

[3] J.-B. d. M. de Lamarck. *Philosophie zoologique.* F. Savy, November 1873. ISBN 978-1-139-10380-0. doi: 10.1017/cbo9781139103800.

[4] J. D. Archibald. Edward Hitchcock's pre-Darwinian (1840) "Tree of Life". *Journal of the History of Biology*, 42(3):561–592, August 2009. ISSN 1573-0387. doi: 10.1007/s10739-008-9163-y.

[5] R. Chambers. *Vestiges of the Natural History of Creation.* J. Churchill, 1853. doi: 10.5962/bhl.title.33033.

[6] H. G. Bronn. *Untersuchungen über die Entwickelungs-Gesetze der organischen Welt während der Bildungs-Zeit unserer Erd-Oberfläche.* E. Schweizerbart, 1858.

[7] D. P. Mindell. The tree of life: Metaphor, model, and heuristic device. *Systematic Biology*, 62(3):479–489, May 2013. ISSN 1063-5157. doi: 10.1093/sysbio/sys115.

[8] U. Hossfeld and G. S. Levit. 'Tree of life'took root 150 years ago. *Nature*, 540(7631): 38–38, November 2016. ISSN 1476-4687. doi: 10.1038/540038a.

[9] E. Haeckel. *Generelle Morphologie Der Organismen: Allgemeine Grundzüge Der Organischen Formen-Wissenschaft, Mechanisch Begründet Durch Die von Charles Darwin Reformierte Descendenz-Theorie. Band 1: Allgemeine Anatomie. Band 2: Allgemeine Entwicklungsgeschichte.* DE GRUYTER, Berlin, New York, January 1866. ISBN 978-3-11-084828-1. doi: 10.1515/9783110848281.

[10] I. Irisarri, D. Baurain, H. Brinkmann, F. Delsuc, J.-Y. Sire, A. Kupfer, J. Petersen, M. Jarek, A. Meyer, M. Vences, and H. Philippe. Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nature Ecology & Evolution*, 1(9):1370–1378, September 2017. ISSN 2397-334X. doi: 10.1038/s41559-017-0240-5.

[11] C. A. Long, R. R. Sokal, and P. H. A. Sneath. *Principles of Numerical Taxonomy*, volume 46. Oxford University Press (OUP), San Francisco, February 1965. doi: 10.2307/1377831.

[12] K. de Queiroz and D. A. Good. Phenetic clustering in biology: A critique. *The Quarterly Review of Biology*, 72(1):3–30, March 1997. ISSN 0033-5770. doi: 10.1086/419656.

[13] E. Mayr. Taxonomy, evolutionary. In S. Brenner and J. H. Miller, editors, *Encyclopedia of Genetics*, pages 1934–1937. Academic Press, New York, January 2001. ISBN 978-0-12-227080-2. doi: 10.1006/rwgn.2001.1466.

[14] P. H. Harvey. Phylogeny and systematics. In N. J. Smelser and P. B. Baltes, editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 11405–11411. Pergamon, Oxford, January 2001. ISBN 978-0-08-043076-8. doi: 10.1016/B0-08-043076-7/03133-8.

[15] C. R. Woese, O. Kandler, and M. L. Wheelis. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579, June 1990. ISSN 1091-6490. doi: 10.1073/pnas.87.12.4576.

[16] G. M. Cooper, M. Brudno, N. C. S. Program, E. D. Green, S. Batzoglou, and A. Sidow. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Research*, 13(5):813–820, May 2003. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.1064503.

[17] G. J. Slater, L. J. Harmon, and M. E. Alfaro. Integrating fossils with molecular phylogenies improves inference of trait evolution. *Evolution*, 66(12):3931–3944, December 2012. ISSN 0014-3820. doi: 10.1111/j.1558-5646.2012.01723.x.

[18] G. J. Slater and L. J. Harmon. Unifying fossils and phylogenies for comparative analyses of diversification and trait evolution. *Methods in Ecology and Evolution*, 4(8):699–702, August 2013. ISSN 2041-210X. doi: 10.1111/2041-210X.12091.

[19] J. F. Parham, P. C. J. Donoghue, C. J. Bell, T. D. Calway, J. J. Head, P. A. Holroyd, J. G. Inoue, R. B. Irmis, W. G. Joyce, D. T. Ksepka, J. S. L. Patané, N. D. Smith, J. E. Tarver, M. van Tuinen, Z. Yang, K. D. Angielczyk, J. M. Greenwood, C. A. Hipsley, L. Jacobs, P. J. Makovicky, J. Müller, K. T. Smith, J. M. Theodor, R. C. M. Warnock, and M. J. Benton. Best practices for justifying fossil calibrations. *Systematic Biology*, 61(2):346–359, March 2012. ISSN 1063-5157. doi: 10.1093/sysbio/syr107.

[20] J. J. Wiens. Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic Biology*, 52(4):528–538, August 2003. ISSN 1063-5157. doi: 10.1080/10635150390218330.

[21] H. Yang. Ancient DNA from pleistocene fossils: Preservation, recovery, and utility of ancient genetic information for quaternary research. *Quaternary Science Reviews*, 16(10):1145–1161, January 1997. ISSN 0277-3791. doi: 10.1016/S0277-3791(97)00006-1.

[22] J. Moore and P. Willmer. Convergent evolution in invertebrates. *Biological Reviews*, 72(1):1–60, February 1997. ISSN 1469-185X, 1464-7931. doi: 10.1017/S0006323196004926.

[23] C. G. Kurland, B. Canback, and O. G. Berg. Horizontal gene transfer: A critical view. *Proceedings of the National Academy of Sciences*, 100(17):9658–9662, August 2003. ISSN 1091-6490. doi: 10.1073/pnas.1632870100.

[24] M. P. Speed and K. Arbuckle. Quantification provides a conceptual basis for convergent evolution. *Biological Reviews*, 92(2):815–829, March 2017. ISSN 1469-185X. doi: 10.1111/brv.12257.

[25] N. Galtier. A model of horizontal gene transfer and the bacterial phylogeny problem. *Systematic Biology*, 56(4):633–642, August 2007. ISSN 1063-5157. doi: 10.1080/10635150701546231.

[26] H. Morlon, T. L. Parsons, and J. B. Plotkin. Reconciling molecular phylogenies with the fossil record. *Proceedings of the National Academy of Sciences*, 108(39): 16327–16332, September 2011. ISSN 1091-6490. doi: 10.1073/pnas.1102543108.

[27] T. A. Heath, J. P. Huelsenbeck, and T. Stadler. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences*, 111(29), July 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1319091111.

[28] A. Gavryushkina, D. Welch, T. Stadler, and A. J. Drummond. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS computational biology*, 10(12):e1003919, December 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003919.

[29] P. J. Wagner and J. D. Marcot. Probabilistic phylogenetic inference in the fossil record: Current and future applications. *The Paleontological Society Papers*, 16: 189–211, October 2010. ISSN 2399-7575. doi: 10.1017/s108933260000187x.

[30] C. R. Marshall. Using the fossil record to evaluate timetree timescales. *Frontiers in Genetics*, 10:1049, November 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.01049.

[31] W. Maddison. Reconstructing character evolution on polytomous cladograms. *Cladistics : the international journal of the Willi Hennig Society*, 5(4):365–377, December 1989. ISSN 0748-3007, 1096-0031. doi: 10.1111/j.1096-0031.1989.tb00569.x.

[32] A. Purvis and T. Garland. Polytomies in comparative analyses of continuous characters. *Systematic Biology*, 42(4):569–575, December 1993. ISSN 1076-836X. doi: 10.2307/2992489.

[33] R. D. Page and E. C. Holmes. *Molecular Evolution: A Phylogenetic Approach*. John Wiley & Sons, 2009.

[34] A. J. Drummond and R. R. Bouckaert. *Bayesian Evolutionary Analysis with BEAST*. Cambridge University Press, August 2015. ISBN 978-1-139-09511-2. doi: 10.1017/cbo9781139095112.

[35] M. Kellis. *Computational Biology - Genomes, Networks, and Evolution*, volume 26.2. MIT OpenCourseWare, Massachusetts Institute of Technology, October 2020.

[36] N. E. Robinson and A. B. Robinson. Molecular clocks. *Proceedings of the National Academy of Sciences*, 98(3):944–949, January 2001. ISSN 1091-6490. doi: 10.1073/pnas.98.3.944.

[37] J. F. Gillooly, A. P. Allen, G. B. West, and J. H. Brown. Metabolic rate calibrates the molecular clock: Reconciling molecular and fossil estimates of evolutionary divergence, arXiv, April 2004. doi: 10.48550/arXiv.q-bio/0404027.

[38] Y.-X. Fu. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. *Molecular Biology and Evolution*, 18(4):620–626, April 2001. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a003842.

[39] L. Bromham, X. Hua, R. Lanfear, and P. F. Cowman. Exploring the relationships between mutation rates, life history, genome size, environment, and species richness in flowering plants. *The American Naturalist*, 185(4):507–524, April 2015. ISSN 0003-0147. doi: 10.1086/680052.

[40] H. Liu and J. Zhang. Yeast spontaneous mutation rate and spectrum vary with environment. *Current Biology*, 29(10):1584–1591.e3, May 2019. ISSN 0960-9822. doi: 10.1016/j.cub.2019.03.054.

[41] T. Lepage, D. Bryant, H. Philippe, and N. Lartillot. A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution*, 24(12):2669–2680, December 2007. ISSN 0737-4038. doi: 10.1093/molbev/msm193.

[42] G. Baele, W. L. S. Li, A. J. Drummond, M. A. Suchard, and P. Lemey. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Molecular Biology and Evolution*, 30(2):239–243, February 2013. ISSN 0737-4038. doi: 10.1093/molbev/mss243.

[43] S. Y. W. Ho and S. Duchêne. Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular Ecology*, 23(24):5947–5965, October 2014. ISSN 1365-294X. doi: 10.1111/mec.12953.

[44] B. Kolaczkowski and J. W. Thornton. A mixed branch length model of heterotachy improves phylogenetic accuracy. *Molecular Biology and Evolution*, 25(6):1054–1066, June 2008. ISSN 0737-4038. doi: 10.1093/molbev/msn042.

[45] B. T. Grenfell, O. G. Pybus, J. R. Gog, J. L. N. Wood, J. M. Daly, J. A. Mumford, and E. C. Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656):327–332, January 2004. ISSN 0036-8075. doi: 10.1126/science.1090727.

[46] O. G. Pybus and A. Rambaut. Evolutionary analysis of the dynamics of viral infectious disease. *Nature Reviews Genetics*, 10(8):540–550, August 2009. ISSN 1471-0064. doi: 10.1038/nrg2583.

[47] T. Stadler, D. Kühnert, S. Bonhoeffer, and A. J. Drummond. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV).

Proceedings of the National Academy of Sciences, 110(1):228–233, January 2013. ISSN 1091-6490. doi: 10.1073/pnas.1207965110.

[48] D. P. Faith. Conservation evaluation and phylogenetic diversity. *Biological conservation*, 61(1):1–10, 1992. ISSN 0006-3207. doi: 10.1016/0006-3207(92)91201-3.

[49] N. J. Isaac, S. T. Turvey, B. Collen, C. Waterman, and J. E. Baillie. Mammals on the EDGE: Conservation priorities based on threat and phylogeny. *PloS one*, 2(3):e296, March 2007. ISSN 1932-6203. doi: 10.1371/journal.pone.0000296.

[50] C. O. Webb. Exploring the phylogenetic structure of ecological communities: An example for rain forest trees. *The American Naturalist*, 156(2):145–155, August 2000. ISSN 0003-0147. doi: 10.1086/303378.

[51] C. O. Webb, D. D. Ackerly, M. A. McPeek, and M. J. Donoghue. Phylogenies and community ecology. *Annual review of ecology and systematics*, 33(1):475–505, November 2002. ISSN 0066-4162. doi: 10.1146/annurev.ecolsys.33.010802.150448.

[52] S. Y. Strauss, C. O. Webb, and N. Salamin. Exotic taxa less related to native species are more invasive. *Proceedings of the National Academy of Sciences*, 103(15):5841–5845, April 2006. ISSN 1091-6490. doi: 10.1073/pnas.0508073103.

[53] J. Felsenstein. Phylogenies and the comparative method. *The American Naturalist*, 125(1):1–15, January 1985. ISSN 0003-0147. doi: 10.1086/284325.

[54] Stadler, Tanja, Magnus, Carsten, Vaughan, Timothy G., Barido-Sottani, Joëlle, Bošková, Veronika, Huisman, Jana, and Pečerska, Jūlija. *Decoding Genomes: From Sequences to Phylodynamics*. ETH Zurich, 2024. doi: 10.3929/ethz-b-000664449.

[55] A. Rzhetsky and M. Nei. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. *Journal of Molecular Evolution*, 35(4):367–375, October 1992. ISSN 1432-1432. doi: 10.1007/BF00161174.

[56] M. R. E. Symonds and S. P. Blomberg. A primer on phylogenetic generalised least squares. In *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*, pages 105–130. Springer, Berlin, Heidelberg, 2014. ISBN 978-3-662-43550-2. doi: 10.1007/978-3-662-43550-2_5.

[57] J. Clavel, L. Aristide, and H. Morlon. A penalized likelihood framework for high-dimensional phylogenetic comparative methods and an application to new-world monkeys brain evolution. *Systematic Biology*, 68(1):93–116, January 2019. ISSN 1063-5157. doi: 10.1093/sysbio/syy045.

[58] D. C. Adams and M. L. Collyer. Multivariate phylogenetic comparative methods: Evaluations, comparisons, and recommendations. *Systematic Biology*, 67(1):14–31, January 2018. ISSN 1063-5157. doi: 10.1093/sysbio/syx055.

[59] J. A. Fuentes-G., P. D. Polly, and E. P. Martins. A Bayesian extension of phylogenetic generalized least squares: Incorporating uncertainty in the comparative study of trait relationships and evolutionary rates. *Evolution*, 74(2):311–325, February 2020. ISSN 0014-3820. doi: 10.1111/evo.13899.

[60] T. Janzen and R. S. Etienne. Phylogenetic tree statistics: A systematic overview using the new R package 'treestats'. *Molecular Phylogenetics and Evolution*, 200:108168, November 2024. ISSN 1055-7903. doi: 10.1016/j.ympev.2024.108168.

[61] M. J. Sackin. "Good" and "Bad" phenograms. *Systematic Biology*, 21(2):225–226, July 1972. ISSN 1063-5157. doi: 10.1093/sysbio/21.2.225.

[62] O. G. Pybus and P. H. Harvey. Testing macro–evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society B: Biological Sciences*, 267 (1459):2267–2272, November 2000. ISSN 1471-2954. doi: 10.1098/rspb.2000.1278.

[63] E. Lewitus and H. Morlon. Characterizing and comparing phylogenies from their Laplacian spectrum. *Systematic Biology*, 65(3):495–507, May 2016. ISSN 1063-5157. doi: 10.1093/sysbio/syv116.

[64] L. Chindelevitch, M. Hayati, A. F. Poon, and C. Colijn. Network science inspires novel tree shape statistics. *Plos one*, 16(12):e0259877, December 2021. ISSN 1932-6203. doi: 10.1371/journal.pone.0259877.

[65] D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267, February 2006. ISSN 0737-4038. doi: 10.1093/molbev/msj030.

[66] D. Bryant and V. Moulton. Neighbor-net: An agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution*, 21(2):255–265, August 2004. ISSN 1537-1719. doi: 10.1093/molbev/msh018.

[67] R. A. L. Elworth, H. A. Ogilvie, J. Zhu, and L. Nakhleh. Advances in computational methods for phylogenetic networks in the presence of hybridization. In T. Warnow, editor, *Bioinformatics and Phylogenetics*, volume 29, pages 317–360. Springer International Publishing, Cham, 2019. ISBN 978-3-030-10836-6 978-3-030-10837-3. doi: 10.1007/978-3-030-10837-3_13.

[68] C. Solís-Lemus, P. Bastide, and C. Ané. PhyloNetworks: A package for phylogenetic networks. *Molecular biology and evolution*, 34(12):3292–3298, September 2017. ISSN 1537-1719. doi: 10.1093/molbev/msx235.

[69] M. S. Hibbins and M. W. Hahn. Phylogenomic approaches to detecting and characterizing introgression. *Genetics*, 220(2):iyab173, November 2021. ISSN 1943-2631. doi: 10.1093/genetics/iyab173.

[70] R. C. Griffiths and P. Marjoram. An ancestral recombination graph. *Progress in population genetics and human evolution*, 87:257, 1997. ISSN 0940-6573. doi: 10.1007/978-1-4757-2609-1_16.

[71] A. L. Lewanski, M. C. Grundler, and G. S. Bradburd. The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics. *Plos Genetics*, 20(1):e1011110, January 2024. ISSN 1553-7404. doi: 10.1371/journal.pgen.1011110.

[72] W. F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284(5423): 2124–2128, June 1999. ISSN 1095-9203. doi: 10.1126/science.284.5423.2124.

[73] E. Bapteste, L. van Iersel, A. Janke, S. Kelchner, S. Kelk, J. O. McInerney, D. A. Morrison, L. Nakhleh, M. Steel, and L. Stougie. Networks: Expanding evolutionary thinking. *Trends in Genetics*, 29(8):439–441, August 2013. ISSN 0168-9525. doi: 10.1016/j.tig.2013.05.007.

[74] C. R. Linder and L. H. Rieseberg. Reconstructing patterns of reticulate evolution in plants. *American Journal of Botany*, 91(10):1700–1708, October 2004. ISSN 0002-9122, 1537-2197. doi: 10.3732/ajb.91.10.1700.

[75] K. Marhold and J. Lihová. Polyploidy, hybridization and reticulate evolution: Lessons from the Brassicaceae. *Plant Systematics and Evolution*, 259(2-4):143–174, July 2006. ISSN 0378-2697, 1615-6110. doi: 10.1007/s00606-006-0417-x.

[76] F. Racimo, S. Sankararaman, R. Nielsen, and E. Huerta-Sánchez. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16(6):359–371, May 2015. ISSN 1471-0064. doi: 10.1038/nrg3936.

[77] S. H. Martin and C. D. Jiggins. Interpreting the genomic landscape of introgression. *Current opinion in genetics & development*, 47:69–74, December 2017. ISSN 0959-437X. doi: 10.1016/j.gde.2017.08.007.

[78] S. Nee, E. C. Holmes, R. M. May, and P. H. Harvey. Extinction rates can be estimated from molecular phylogenies. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1307):77–82, April 1994. ISSN 0962-8436. doi: 10.1098/rstb.1994.0054.

[79] S. Nee, R. M. May, and P. H. Harvey. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1309):305–311, May 1994. ISSN 1471-2970. doi: 10.1098/rstb.1994.0068.

[80] S. Nee. Birth-death models in macroevolution. *Annual Review of Ecology, Evolution, and Systematics*, 37(1):1–17, December 2006. ISSN 1545-2069. doi: 10.1146/annurev.ecolsys.37.091305.110035.

[81] T. Kubo and Y. Iwasa. Inferring the rates of branching and extinction from molecular phylogenies. *Evolution*, 49(4):694–704, August 1995. ISSN 0014-3820. doi: 10.2307/2410323.

[82] R. S. Etienne, B. Haegeman, T. Stadler, T. Aze, P. N. Pearson, A. Purvis, and A. B. Phillimore. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the Royal Society B: Biological Sciences*, 279 (1732):1300–1309, October 2012. ISSN 0962-8452. doi: 10.1098/rspb.2011.1439.

[83] A. Purvis. Phylogenetic approaches to the study of extinction. *Annual Review of Ecology, Evolution, and Systematics*, 39(1):301–319, December 2008. ISSN 1545-2069. doi: 10.1146/annurev.ecolsys.063008.102010.

[84] A. B. Phillimore and T. D. Price. Density-dependent cladogenesis in birds. *PLoS Biology*, 6(3):e71, March 2008. ISSN 1545-7885. doi: 10.1371/journal.pbio.0060071.

[85] D. Moen and H. Morlon. Why does diversification slow down? *Trends in Ecology and Evolution*, 29(4):190–197, April 2014. ISSN 0169-5347. doi: 10.1016/j.tree.2014.01.010.

[86] R. Aguilée, F. Gascuel, A. Lambert, and R. Ferriere. Clade diversification dynamics and the biotic and abiotic controls of speciation and extinction rates. *Nature Communications*, 9(1):1–13, August 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-05419-7.

[87] S. Louca and M. W. Pennell. Why extinction estimates from extant phylogenies are so often zero. *Current Biology*, 31(14):3168–3173. e4, January 2021. ISSN 0960-9822. doi: 10.1101/2021.01.04.425256.

[88] T. Stadler. Inferring speciation and extinction processes from extant species data. *Proceedings of the National Academy of Sciences*, 108(39):16145–16146, September 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1113242108.

[89] R. S. Etienne and J. Rosindell. Prolonging the past counteracts the pull of the present: Protracted speciation can explain observed slowdowns in diversification. *Systematic Biology*, 61(2):204, March 2012. ISSN 1076-836X. doi: 10.1093/sysbio/syr091.

[90] J. Rosindell, S. J. Cornell, S. P. Hubbell, and R. S. Etienne. Protracted speciation revitalizes the neutral theory of biodiversity. *Ecology Letters*, 13(6):716–727, May 2010. ISSN 1461-0248. doi: 10.1111/j.1461-0248.2010.01463.x.

[91] M. T. Ghiselin and J. W. Valentine. *Evolutionary Paleoecology of the Marine Biosphere*, volume 24. Prentice Hall, Englewood Cliffs, New Jersey, September 1975. ISBN 0-13-293720-4. doi: 10.2307/2412725.

[92] J. J. Sepkoski. A kinetic model of Phanerozoic taxonomic diversity I. Analysis of marine orders. *Paleobiology*, 4(3):223–251, 1978. ISSN 0094-8373. doi: 10.1017/s0094837300005972.

[93] J. J. Wiens. The causes of species richness patterns across space, time, and clades and the role of "ecological limits". *The Quarterly review of biology*, 86(2):75–96, June 2011. ISSN 0033-5770. doi: 10.1086/659883.

[94] N. Cusimano and S. S. Renner. Slowdowns in diversification rates from real phylogenies may not be real. *Systematic Biology*, 59(4):458–464, July 2010. ISSN 1063-5157. doi: 10.1093/sysbio/syq032.

[95] S. Höhna, T. Stadler, F. Ronquist, and T. Britton. Inferring speciation and extinction rates under different sampling schemes. *Molecular Biology and Evolution*, 28(9): 2577–2589, September 2011. ISSN 0737-4038. doi: 10.1093/molbev/msr095.

[96] D. S. Srivastava, M. W. Cadotte, A. A. M. Macdonald, R. G. Marushia, and N. Mirotchnick. Phylogenetic diversity and the functioning of ecosystems. *Ecology Letters*, 15 (7):637–648, May 2012. ISSN 1461-0248. doi: 10.1111/j.1461-0248.2012.01795.x.

[97] A. Kondratyeva, P. Grandcolas, and S. Pavoine. Reconciling the concepts and measures of diversity, rarity and originality in ecology and evolution. *Biological Reviews*, 94(4):1317–1337, March 2019. ISSN 1464-7931. doi: 10.1111/brv.12504.

[98] M. R. Helmus, T. J. Bland, C. K. Williams, and A. R. Ives. Phylogenetic measures of biodiversity. *The American Naturalist*, 169(3):E68, March 2007. ISSN 1537-5323. doi: 10.1086/511334.

[99] M. W. Cadotte, R. Dinnage, and D. Tilman. Phylogenetic diversity promotes ecosystem stability. *Ecology*, 93(sp8):S223–S233, August 2012. ISSN 1939-9170. doi: 10.1890/11-0426.1.

[100] A. L. Pigot, T. Bregman, C. Sheard, B. Daly, R. S. Etienne, and J. A. Tobias. Quantifying species contributions to ecosystem processes: A global assessment of functional trait and phylogenetic metrics across avian seed-dispersal networks. *Proceedings of the Royal Society B: Biological Sciences*, 283(1844):20161597, December 2016. ISSN 0962-8452. doi: 10.1098/rspb.2016.1597.

[101] Y.-L. Zheng, J. H. Burns, Z.-Y. Liao, Y.-p. Li, J. Yang, Y.-j. Chen, J.-l. Zhang, and Y.-g. Zheng. Species composition, functional and phylogenetic distances correlate with success of invasive Chromolaena odorata in an experimental test. *Ecology Letters*, 21 (8):1211–1220, May 2018. ISSN 1461-0248. doi: 10.1111/ele.13090.

[102] J. J. Henn, S. Yelenik, and E. Damschen. Environmental gradients influence differences in leaf functional traits between native and non-native plants. *Oecologia*, 191(2): 397–409, September 2019. ISSN 0029-8549. doi: 10.1007/s00442-019-04498-7.

[103] T.-J. Qin, J. Zhou, Y. Sun, H. Müller-Schärer, F.-L. Luo, B.-C. Dong, H.-L. Li, and F.-H. Yu. Phylogenetic diversity is a better predictor of wetland community resistance to Alternanthera philoxeroides invasion than species richness. *Plant Biology*, 22(4): 591–599, March 2020. ISSN 1438-8677. doi: 10.1111/plb.13101.

[104] L. Gallien, F. Mazel, S. Lavergne, J. Renaud, R. Douzet, and W. Thuiller. Contrasting the effects of environment, dispersal and biotic interactions to explain the distribution of invasive plants in alpine communities. *Biological invasions*, 17(5):1407–1423, May 2015. ISSN 1387-3547. doi: 10.1007/s10530-014-0803-1.

[105] C. Ma, S.-p. Li, Z. Pu, J. Tan, M. Liu, J. Zhou, H. Li, and L. Jiang. Different effects of invader–native phylogenetic relatedness on invasion success and impact: A meta-analysis of Darwin's naturalization hypothesis. *Proceedings of the Royal Society B: Biological Sciences*, 283(1838):20160663, September 2016. ISSN 0962-8452. doi: 10.1098/rspb.2016.0663.

[106] D. S. Park, X. Feng, B. S. Maitner, K. C. Ernst, and B. J. Enquist. Darwin's naturalization conundrum can be explained by spatial scale. *Proceedings of the National Academy*

of Sciences, 117(20):10904–10910, May 2020. ISSN 1091-6490. doi: 10.1073/pnas.1918100117.

[107] D. L. Rabosky and R. E. Glor. Equilibrium speciation dynamics in a model adaptive radiation of island lizards. *Proceedings of the National Academy of Sciences, USA*, 107 (51):22178–22183, December 2010. ISSN 1091-6490. doi: 10.1073/pnas.1007606107.

[108] M. Foote, R. A. Cooper, J. S. Crampton, and P. M. Sadler. Diversity-dependent evolutionary rates in early palaeozoic zooplankton. *Proceedings of the Royal Society B: Biological Sciences*, 285(1873):20180122, February 2018. ISSN 1471-2954. doi: 10.1098/rspb.2018.0122.

[109] M. M. Pires, D. Silvestro, and T. B. Quental. Interactions within and between clades shaped the diversification of terrestrial carnivores. *Evolution*, 71(7):1855–1864, June 2017. ISSN 1558-5646. doi: 10.1111/evo.13269.

[110] R. S. Etienne, B. Haegeman, Á. Dugo-Cota, C. Vilà, A. Gonzalez-Voyer, and L. Valente. The phylogenetic limits to diversity-dependent diversification. *Systematic Biology*, 72(2):433–445, March 2023. ISSN 1063-5157. doi: 10.1093/sysbio/syac074.

[111] A. Mooers, O. Gascuel, T. Stadler, H. Li, and M. Steel. Branch lengths on birth–death trees and the expected loss of phylogenetic diversity. *Systematic Biology*, 61(2): 195–203, March 2012. ISSN 1063-5157. doi: 10.1093/sysbio/syr090.

[112] W. P. Maddison, P. E. Midford, and S. P. Otto. Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, 56(5):701–710, October 2007. ISSN 1063-5157. doi: 10.1080/10635150701607033.

[113] R. G. FitzJohn. Diversitree: Comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution*, 3(6):1084–1092, December 2012. ISSN 2041-210X, 2041-210X. doi: 10.1111/j.2041-210X.2012.00234.x.

[114] L. M. Valente, A. B. Phillimore, and R. S. Etienne. Equilibrium and non-equilibrium dynamics simultaneously operate in the Galápagos islands. *Ecology Letters*, 18(8): 844–852, June 2015. ISSN 1461-0248. doi: 10.1111/ele.12461.

[115] D. S. Caetano, B. C. O'Meara, and J. M. Beaulieu. Hidden state models improve state-dependent diversification approaches, including biogeographical models. *Evolution*, 72(11):2308–2324, October 2018. ISSN 0014-3820. doi: 10.1111/evo.13602.

[116] L. Herrera-Alsina, P. van Els, and R. S. Etienne. Detecting the dependence of diversification on multiple traits from phylogenetic trees and trait data. *Systematic Biology*, 68(2):317–328, September 2019. ISSN 1076-836X. doi: 10.1093/sysbio/syy057.

[117] S. M. Kidwell and K. W. Flessa. The quality of the fossil record: Populations, species, and communities. *Annual Review of Earth and Planetary Sciences*, 24(1):433–464, May 1996. ISSN 0084-6597, 1545-4495. doi: 10.1146/annurev.earth.24.1.433.

[118] S. Nee. Inferring speciation rates from phylogenies. *Evolution*, 55(4):661–668, May 2001. ISSN 1558-5646. doi: 10.1111/j.0014-3820.2001.tb00801.x.

[119] H. Morlon. Phylogenetic approaches for studying diversification. *Ecology Letters*, 17 (4):508–525, February 2014. ISSN 1461-0248. doi: 10.1111/ele.12251.

[120] T. Stadler. Simulating trees with a fixed number of extant species. *Systematic Biology*, 60(5):676–684, October 2011. ISSN 1063-5157. doi: 10.1093/sysbio/syr029.

[121] R. S. Etienne, A. L. Pigot, and A. B. Phillimore. How reliably can we infer diversity-dependent diversification from phylogenies? *Methods in Ecology and Evolution*, 7(9): 1092–1099, May 2016. ISSN 2041-210X. doi: 10.1111/2041-210X.12565.

[122] S. Xie, L. Valente, and R. S. Etienne. Identifying summary statistics for approximate Bayesian computation in a phylogenetic island biogeography model, bioRxiv, October 2023. doi: 10.1101/2023.10.13.562305.

[123] R. S. Etienne, H. Morlon, and A. Lambert. Estimating the duration of speciation from phylogenies. *Evolution*, 68(8):2430–2440, August 2014. ISSN 0014-3820. doi: 10.1111/evo.12433.

[124] H. K. Alexander, A. Lambert, and T. Stadler. Quantifying age-dependent extinction from species phylogenies. *Systematic Biology*, 65(1):35–50, January 2016. ISSN 1063-5157. doi: 10.1093/sysbio/syv065.

[125] T. Janzen, S. Höhna, and R. S. Etienne. Approximate Bayesian computation of diversification rates from molecular phylogenies: Introducing a new efficient summary statistic, the nLTT. *Methods in Ecology and Evolution*, 6(5):566–575, March 2015. ISSN 2041-210X. doi: 10.1111/2041-210X.12350.

[126] S. Lambert, J. Voznica, and H. Morlon. Deep learning from phylogenies for diversification analyses. *Systematic Biology*, 72(6):1262–1279, November 2023. ISSN 1063-5157. doi: 10.1093/sysbio/syad044.

[127] D. L. Rabosky. Extinction rates should not be estimated from molecular phylogenies. *Evolution*, 64(6):1816–1824, June 2010. ISSN 0014-3820. doi: 10.1111/j.1558-5646.2009.00926.x.

[128] E. J. Ward. A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. *Ecological Modelling*, 211(1):1–10, February 2008. ISSN 0304-3800. doi: 10.1016/j.ecolmodel.2007.10.030.

[129] M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, December 2002. ISSN 1943-2631. doi: 10.1093/genetics/162.4.2025.

[130] M. A. Beaumont. Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):379–406, December 2010. ISSN 1543-592X, 1545-2069. doi: 10.1146/annurev-ecolsys-102209-144621.

[131] D. L. Rabosky. Heritability of extinction rates links diversification patterns in molecular phylogenies and fossils. *Systematic Biology*, 58(6):629–640, December 2009. ISSN 1063-5157. doi: 10.1093/sysbio/syp069.

[132] F. Bokma. Time, species, and separating their effects on trait variance in clades. *Systematic Biology*, 59(5):602–607, October 2010. ISSN 1063-5157. doi: 10.1093/sysbio/syq029.

[133] N. Kutsukake and H. Innan. Simulation-based likelihood approach for evolutionary models of phenotypic traits on phylogeny. *Evolution*, 67(2):355–367, February 2013. ISSN 0014-3820. doi: 10.1111/j.1558-5646.2012.01775.x.

[134] S. Xie, L. Valente, and R. S. Etienne. Can we ignore trait-dependent colonization and diversification in island biogeography? *Evolution*, 77(3):670–681, March 2023. ISSN 0014-3820. doi: 10.1093/evolut/qpad006.

[135] C. C. Aggarwal. *Neural Networks and Deep Learning: A Textbook*. Springer, 2018. ISBN 978-3-031-29642-0. doi: 10.1007/978-3-031-29642-0.

[136] H. Zhu, M. Akrout, B. Zheng, A. Pelegris, A. Jayarajan, A. Phanishayee, B. Schroeder, and G. Pekhimenko. Benchmarking and analyzing deep neural network training. In *2018 IEEE International Symposium on Workload Characterization (IISWC)*, pages 88–100. IEEE, September 2018. doi: 10.1109/IISWC.2018.8573476.

[137] H. Sak, A. Senior, and F. Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition, arXiv, 2014. doi: 10.48550/ARXIV.1402.1128.

[138] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee. Recent advances in recurrent neural networks, arXiv, 2017. doi: 10.48550/ARXIV.1801.01078.

[139] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks, arXiv.org, 2016. doi: 10.48550/ARXIV.1609.02907.

[140] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec. Hierarchical graph representation learning with differentiable pooling. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[141] G. Li, C. Xiong, A. Thabet, and B. Ghanem. DeeperGCN: All you need to train deeper gcns, arXiv, 2020. doi: 10.48550/ARXIV.2006.07739.

[142] L. Rampášek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf, and D. Beaini. Recipe for a general, powerful, scalable graph transformer, arXiv, 2022. doi: 10.48550/ARXIV.2205.12454.

[143] I. Lajaaiti, S. Lambert, J. Voznica, H. Morlon, and F. Hartig. A comparison of deep learning architectures for inferring parameters of diversification models from extant phylogenies, bioRxiv, March 2023. doi: 10.1101/2023.03.03.530992.

[144] J. Voznica, A. Zhukova, V. Boskova, E. Saulnier, F. Lemoine, M. Moslonka-Lefebvre, and O. Gascuel. Deep learning from phylogenies to uncover the epidemiological dynamics of outbreaks. *Nature Communications*, 13(1):3896, July 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-31511-0.

[145] D. Moi and C. Dessimoz. Reconstructing protein interactions across time using phylogeny-aware graph neural networks, bioRxiv, July 2022. doi: 10.1101/2022.07.21.501014.

[146] D. Reiman, A. A. Metwally, J. Sun, and Y. Dai. PopPhy-CNN: A phylogenetic tree embedded architecture for convolutional neural networks to predict host phenotype from metagenomic data. *IEEE Journal of Biomedical and Health Informatics*, 24(10): 2993–3001, October 2020. ISSN 2168-2208. doi: 10.1109/JBHI.2020.2993761.

[147] T. Qin, K. J. van Benthem, L. Valente, and R. S. Etienne. Parameter estimation from phylogenetic trees using neural networks and ensemble learning. *Systematic Biology*, page syaf060, September 2025. ISSN 1063-5157. doi: 10.1093/sysbio/syaf060.

[148] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, arXiv.org, 2014. doi: 10.48550/ARXIV.1412.6980.

[149] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.

[150] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[151] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, 27, 2014.

[152] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204. PMLR, 2015.

[153] S. Louca and M. W. Pennell. Extant timetrees are consistent with a myriad of diversification histories. *Nature*, 580(7804):502–505, April 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2176-1.

[154] B. Legried and J. Terhorst. A class of identifiable phylogenetic birth–death models. *Proceedings of the National Academy of Sciences*, 119(35):e2119513119, August 2022. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2119513119.

[155] B. Legried and J. Terhorst. Identifiability and inference of phylogenetic birth–death models. *Journal of Theoretical Biology*, 568:111520, July 2023. ISSN 0022-5193. doi: 10.1016/j.jtbi.2023.111520.

[156] S. Höhna, B. T. Kopperud, and A. F. Magee. CRABS: Congruent rate analyses in birth–death scenarios. *Methods in Ecology and Evolution*, 13(12):2709–2718, December 2022. ISSN 2041-210X, 2041-210X. doi: 10.1111/2041-210X.13997.

[157] D. Silvestro, J. Schnitzler, and G. Zizka. A Bayesian framework to estimate diversification rates and their variation through time and space. *BMC Evolutionary Biology*, 11(1):311, 2011. ISSN 1471-2148. doi: 10.1186/1471-2148-11-311.

[158] C. Zhang, T. Stadler, S. Klopfstein, T. A. Heath, and F. Ronquist. Total-evidence dating under the fossilized birth–death process. *Systematic Biology*, 65(2):228–249, October 2016. ISSN 1076-836X. doi: 10.1093/sysbio/syv080.

[159] T. Hauffe, J. L. Cantalapiedra, and D. Silvestro. Trait-mediated speciation and human-driven extinctions in proboscideans revealed by unsupervised Bayesian neural networks. *Science Advances*, 10(30):eadl2643, July 2024. ISSN 2375-2548. doi: 10.1126/sciadv.adl2643.

[160] A. F. Magee, S. Höhna, T. I. Vasylyeva, A. D. Leaché, and V. N. Minin. Locally adaptive Bayesian birth-death model successfully detects slow and rapid rate shifts. *PLoS computational biology*, 16(10):e1007999, October 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007999.

[161] G. Wu, K. Wu, R. E. Halling, E. Horak, J. Xu, G.-M. Li, S. Lee, L. Pecoraro, R. F. Arzu, and S. T. Ndolo Ebika. The rapid diversification of boletales is linked to early eocene and mid-miocene climatic optima. *bioRxiv : the preprint server for biology*, pages 2023–10, October 2023. doi: 10.1101/2023.10.24.563795.

[162] S. Kong, C. Solís-Lemus, and G. P. Tiley. Phylogenetic networks empower biodiversity research. *Proceedings of the National Academy of Sciences*, 122(31):e2410934122, August 2025. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2410934122.

[163] C. Zhang and F. A. Matsen IV. A variational approach to bayesian phylogenetic inference. *Journal of Machine Learning Research*, 25(145):1–56, 2024.

[164] K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, May 2020. ISSN 1091-6490. doi: 10.1073/pnas.1912789117.

[165] J.-M. Lueckmann, J. Boelts, D. Greenberg, P. Goncalves, and J. Macke. Benchmarking simulation-based inference. In *International Conference on Artificial Intelligence and Statistics*, pages 343–351. PMLR, 2021.

[166] T. Qin, L. Valente, and R. S. Etienne. Impact of evolutionary relatedness on species diversification and tree shape. *Journal of Theoretical Biology*, 598:111992, February 2025. ISSN 0022-5193. doi: 10.1016/j.jtbi.2024.111992.

[167] S. Nee, A. O. Mooers, and P. H. Harvey. Tempo and mode of evolution revealed from molecular phylogenies. *Proceedings of the National Academy of Sciences, USA*, 89(17): 8322–8326, September 1992. ISSN 0027-8424. doi: 10.1073/pnas.89.17.8322.

[168] T. B. Quental and C. R. Marshall. Diversity dynamics: Molecular phylogenies need the fossil record. *Trends in Ecology and Evolution*, 25(8):434–441, August 2010. ISSN 0169-5347. doi: 10.1016/j.tree.2010.05.002.

[169] M. M. Mayfield and J. M. Levine. Opposing effects of competitive exclusion on the phylogenetic structure of communities. *Ecology Letters*, 13(9):1085–1093, June 2010. ISSN 1461-0248. doi: 10.1111/j.1461-0248.2010.01509.x.

[170] J. HilleRisLambers, P. Adler, W. Harpole, J. Levine, and M. Mayfield. Rethinking community assembly through the lens of coexistence theory. *Annual Review of Ecology, Evolution, and Systematics*, 43(1):227–248, December 2012. ISSN 1543-592X, 1545-2069. doi: 10.1146/annurev-ecolsys-110411-160411.

[171] P. Gerhold, J. F. Cahill, M. Winter, I. V. Bartish, and A. Prinzing. Phylogenetic patterns are not proxies of community assembly mechanisms (they are far better). *Functional Ecology*, 29(5):600–614, March 2015. ISSN 1365-2435. doi: 10.1111/1365-2435.12425.

[172] C. M. Tucker, T. J. Davies, M. W. Cadotte, and W. D. Pearse. On the relationship between phylogenetic diversity and trait diversity. *Ecology*, 99(6):1473–1479, May 2018. ISSN 0012-9658. doi: 10.1002/ecy.2349.

[173] F. Mazel, M. W. Pennell, M. W. Cadotte, S. Diaz, G. V. Dalla Riva, R. Grenyer, F. Leprieur, A. O. Mooers, D. Mouillot, C. M. Tucker, and W. D. Pearse. Prioritizing phylogenetic diversity captures functional diversity unreliably. *Nature Communications*, 9(1):2888, July 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-05126-3.

[174] P. Venail, K. Gross, T. H. Oakley, A. Narwani, E. Allan, P. Flombaum, F. Isbell, J. Joshi, P. B. Reich, and D. Tilman. Species richness, but not phylogenetic diversity, influences community biomass production and temporal stability in a re-examination of 16 grassland biodiversity studies. *Functional Ecology*, 29(5):615–626, March 2015. ISSN 1365-2435. doi: 10.1111/1365-2435.12432.

[175] M. R. Pie, R. Divieso, and F. S. Caron. Clade density and the evolution of diversity-dependent diversification. *Nature Communications*, 14(1):4576, July 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-39629-5.

[176] R. E. Ricklefs. Evolutionary diversification, coevolution between populations and their antagonists, and the filling of niche space. *Proceedings of the National Academy of Sciences, USA*, 107(4):1265–1272, January 2010. ISSN 0027-8424. doi: 10.1073/pnas.0913626107.

[177] M. Rillo and R. Etienne. Diversity-dependent diversification. *Oxford Bibliographies Online*, February 2022. doi: 10.1093/OBO/9780199941728-0141.

[178] H. Hildenbrandt and T. Qin. evesim: Evolutionary relatedness dependent diversification simulation powered by the Rcpp backend SimTable, CRAN, September 2024. doi: 10.32614/CRAN.package.evesim.

[179] M. W. Cadotte and T. J. Davies. Rarest of the rare: Advances in combining evolutionary distinctiveness and scarcity to inform conservation at biogeographical scales. *Diversity and Distributions*, 16(3):376–385, April 2010. ISSN 1366-9516. doi: 10.1111/j.1472-4642.2010.00650.x.

[180] D. W. Redding, K. Hartmann, A. Mimoto, D. Bokal, M. DeVos, and A. O. Mooers. Evolutionarily distinctive species often capture more phylogenetic diversity than expected. *Journal of theoretical biology*, 251(4):606–615, April 2008. ISSN 0022-5193. doi: 10.1016/j.jtbi.2007.12.006.

[181] J. Rosindell, K. Manson, R. Gumbs, W. D. Pearse, and M. Steel. Phylogenetic biodiversity metrics should account for both accumulation and attrition of evolutionary heritage. *Systematic Biology*, 73(1):158–182, January 2024. ISSN 1063-5157. doi: 10.1093/sysbio/syad072.

[182] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4): 403–434, December 1976. ISSN 0021-9991. doi: 10.1016/0021-9991(76)90041-3.

[183] T. Qin. eve: Evolution emulator, phylogenetically dependent, Zenodo, May 2023. doi: 10.5281/zenodo.7991231.

[184] J. Lemant, C. Le Sueur, V. Manojlović, and R. Noble. Robust, universal tree balance indices. *Systematic biology*, 71(5):1210–1224, April 2022. ISSN 1076-836X. doi: 10.1093/sysbio/syac027.

[185] J. S. Rogers. Central moments and probability distributions of three measures of phylogenetic tree imbalance. *Systematic Biology*, 45(1):99–110, March 1996. ISSN 1063-5157. doi: 10.2307/2413515.

[186] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, January 2000. ISSN 0169-7439. doi: 10.1016/s0169-7439(99)00047-7.

[187] D. L. Rabosky. Ecological limits and diversification rate: Alternative paradigms to explain the variation in species richness among clades and regions. *Ecology Letters*, 12(8):735–743, July 2009. ISSN 1461-0248. doi: 10.1111/j.1461-0248.2009.01333.x.

[188] G. G. Mittelbach, D. W. Schemske, H. V. Cornell, A. P. Allen, J. M. Brown, M. B. Bush, S. P. Harrison, A. H. Hurlbert, N. Knowlton, H. A. Lessios, C. M. McCain, A. R. McCune, L. A. McDade, M. A. McPeek, T. J. Near, T. D. Price, R. E. Ricklefs, K. Roy, D. F. Sax, D. Schluter, J. M. Sobel, and M. Turelli. Evolution and the latitudinal diversity gradient: Speciation, extinction and biogeography. *Ecology Letters*, 10(4): 315–331, February 2007. ISSN 1461-0248. doi: 10.1111/j.1461-0248.2007.01020.x.

[189] S. A. H. Geritz, E. van der Meijden, and J. A. J. Metz. Evolutionary dynamics of seed size and seedling competitive ability. *Theoretical Population Biology*, 55(3):324–343, June 1999. ISSN 0040-5809. doi: 10.1006/tpbi.1998.1409.

[190] D. I. Bolnick and B. M. Fitzpatrick. Sympatric speciation: Models and empirical evidence. *Annual Review of Ecology, Evolution, and Systematics*, 38(1):459–487, December 2007. ISSN 1545-2069. doi: 10.1146/annurev.ecolsys.38.091206.095804.

[191] M. P. Thakur and A. J. Wright. Environmental filtering, niche construction, and trait variability: The missing discussion. *Trends in Ecology and Evolution*, 32(12):884–886, December 2017. ISSN 0169-5347. doi: 10.1016/j.tree.2017.09.014.

[192] K. H. Kozak and JohnJ. Wiens. Does niche conservatism promote speciation? A case study in North American salamanders. *Evolution*, 60(12):2604–2621, December 2006. ISSN 1558-5646. doi: 10.1111/j.0014-3820.2006.tb01893.x.

[193] R. A. Pyron, G. C. Costa, M. A. Patten, and F. T. Burbrink. Phylogenetic niche conservatism and the evolutionary basis of ecological speciation. *Biological Reviews*, 90(4):1248–1262, November 2015. ISSN 1469-185X. doi: 10.1111/brv.12154.

[194] N. R. Owen, R. Gumbs, C. L. Gray, and D. P. Faith. Global conservation of phylogenetic diversity captures more than just functional diversity. *Nature Communications*, 10 (1):859, February 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08600-8.

[195] E. Stam. Does imbalance in phylogenies reflect only bias? *Evolution*, 56(6):1292–1295, June 2002. ISSN 1558-5646. doi: 10.1111/j.0014-3820.2002.tb01440.x.

[196] M. G. B. Blum and O. François. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. *Systematic Biology*, 55(4):685–691, August 2006. ISSN 1063-5157. doi: 10.1080/10635150600889625.

[197] J. Bengtsson. Interspecific competition increases local extinction rate in a metapopulation system. *Nature*, 340(6236):713–715, August 1989. ISSN 1476-4687. doi: 10.1038/340713a0.

[198] E. M. Dangremond, E. A. Pardini, and T. M. Knight. Apparent competition with an invasive plant hastens the extinction of an endangered lupine. *Ecology*, 91(8): 2261–2271, August 2010. ISSN 1939-9170. doi: 10.1890/09-0418.1.

[199] A. Timmermann. Quantifying the potential causes of Neanderthal extinction: Abrupt climate change versus competition and interbreeding. *Quaternary Science Reviews*, 238:106331, June 2020. ISSN 0277-3791. doi: 10.1016/j.quascirev.2020.106331.

[200] T. Qin, K. J. van Benthem, L. Valente, and R. S. Etienne. Performance and robustness of parameter estimation from phylogenetic trees using neural networks, bioRxiv, August 2024. doi: 10.1101/2024.08.02.606350.

[201] E. P. Martins and T. F. Hansen. Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, 149(4):646–667, April 1997. ISSN 0003-0147. doi: 10.1086/286013.

[202] P. J. Wagner. Phylogenetic analyses and the fossil record: Tests and inferences, hypotheses and models. *Paleobiology*, 26(S4):341–371, January 2000. ISSN 0094-8373, 1938-5331. doi: 10.1017/S0094837300027007.

[203] J. Hey. Using phylogenetic trees to study speciation and extinction. *Evolution*, 46(3): 627–640, June 1992. ISSN 1558-5646. doi: 10.1111/j.1558-5646.1992.tb02071.x.

[204] J. Dopazo and J. M. Carazo. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *Journal of Molecular Evolution*, 44(2):226–233, February 1997. ISSN 1432-1432. doi: 10.1007/PL00006139.

[205] Q. Tao, K. Tamura, F. U. Battistuzzi, and S. Kumar. A machine learning method for detecting autocorrelation of evolutionary rates in large phylogenies. *Molecular Biology and Evolution*, 36(4):811–824, April 2019. ISSN 0737-4038. doi: 10.1093/molbev/msz014.

[206] A. Suvorov, J. Hochuli, and D. R. Schrider. Accurate inference of tree topologies from multiple sequence alignments using deep learning. *Systematic Biology*, 69(2):221–233, March 2020. ISSN 1063-5157. doi: 10.1093/sysbio/syz060.

[207] Z. Zou, H. Zhang, Y. Guan, and J. Zhang. Deep residual neural networks resolve quartet molecular phylogenies. *Molecular Biology and Evolution*, 37(5):1495–1507, May 2020. ISSN 0737-4038. doi: 10.1093/molbev/msz307.

[208] A. Suvorov and D. R. Schrider. Reliable estimation of tree branch lengths using deep neural networks. *PLOS Computational Biology*, 20(8):e1012337, August 2024. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1012337.

[209] A. Thompson, B. J. Liebeskind, E. J. Scully, and M. J. Landis. Deep learning and likelihood approaches for viral phylogeography converge on the same answers whether the inference model is right or wrong. *Systematic Biology*, 73(1):183–206, January 2024. ISSN 1063-5157. doi: 10.1093/sysbio/syad074.

[210] T. Qin. EvoNN: Neural networks for evolution, Zenodo, May 2024. doi: 10.5281/zenodo.18193471.

[211] D. Luebke. CUDA: Scalable parallel programming for high-performance scientific computing. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 836–838. IEEE, May 2008. doi: 10.1109/isbi.2008.4541126.

[212] S. Imambi, K. B. Prakash, and G. R. Kanagachidambaresan. PyTorch. In K. B. Prakash and G. R. Kanagachidambaresan, editors, *Programming with TensorFlow*, pages 87–104. Springer International Publishing, Cham, 2021. ISBN 978-3-030-57076-7 978-3-030-57077-4. doi: 10.1007/978-3-030-57077-4_10.

[213] M. Fey and J. E. Lenssen. Fast graph representation learning with PyTorch geometric, arXiv.org, 2019. doi: 10.48550/ARXIV.1903.02428.

[214] R Core Team. R: A language and environment for statistical computing, 2025.

[215] E. Paradis and K. Schliep. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics (Oxford, England)*, 35(3):526–528, February 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty633.

[216] T. Stadler and F. Bokma. Estimating speciation and extinction rates for phylogenies of higher taxa. *Systematic Biology*, 62(2):220–230, March 2013. ISSN 1063-5157. doi: 10.1093/sysbio/sys087.

[217] L. F. Henao Diaz, L. J. Harmon, M. T. C. Sugawara, E. T. Miller, and M. W. Pennell. Macroevolutionary diversification rates show time dependency. *Proceedings of the National Academy of Sciences*, 116(15):7403–7408, April 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1818058116.

[218] M. Graczyk, T. Lasota, B. Trawiński, and K. Trawiński. Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, N. T. Nguyen, M. T. Le, and J. Świątek, editors, *Intelligent Information and Database Systems*, volume 5991, pages 340–350. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-12100-5 978-3-642-12101-2. doi: 10.1007/978-3-642-12101-2_35.

[219] I. Loshchilov and F. Hutter. Decoupled weight decay regularization, arXiv, 2017. doi: 10.48550/ARXIV.1711.05101.

[220] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright. Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on Optimization*, 9(1):112–147, January 1998. ISSN 1095-7189. doi: 10.1137/S1052623496303470.

[221] T. Qin. eveGNN: Codebase for phylogenetic tree parameter estimation with neural networks, Zenodo, November 2023. doi: 10.5281/zenodo.18193452.

[222] F. L. Condamine, J. Rolland, and H. Morlon. Assessing the causes of diversification slowdowns: Temperature-dependent and diversity-dependent models receive equivalent support. *Ecology Letters*, 22(11):1900–1912, September 2019. ISSN 1461-0248. doi: 10.1111/ele.13382.

[223] Y. Romano, E. Patterson, and E. Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[224] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury, Thomson Learning Inc., Pacific Grove, CA,, April 2002. ISBN 978-1-003-45628-5. doi: 10.1201/9781003456285.

[225] C. M. Bishop and N. M. Nasrabadi. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, NY, 2006. ISBN 978-0-387-31073-2.

[226] W. Zhang, Z. Sheng, Y. Jiang, Y. Xia, J. Gao, Z. Yang, and B. Cui. Evaluating deep graph neural networks, arXiv, 2021. doi: 10.48550/ARXIV.2108.00955.

[227] Q. Li, Z. Han, and X.-M. Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32 of *Proceedings of the AAAI Conference on Artificial Intelligence*.

Association for the Advancement of Artificial Intelligence (AAAI), April 2018. doi: 10.1609/aaai.v32i1.11604.

[228] U. Alon and E. Yahav. On the bottleneck of graph neural networks and its practical implications, arXiv, 2020. doi: 10.48550/ARXIV.2006.05205.

[229] V. P. Dwivedi, L. Rampášek, M. Galkin, A. Parviz, G. Wolf, A. T. Luu, and D. Beaini. Long range graph benchmark. *Advances in Neural Information Processing Systems*, 35:22326–22340, 2022.

[230] J. H. Giraldo, K. Skianis, T. Bouwmans, and F. D. Malliaros. On the trade-off between over-smoothing and over-squashing in deep graph neural networks. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, Cikm '23, pages 566–576, Birmingham United Kingdom, October 2023. ACM. ISBN 979-8-4007-0124-5. doi: 10.1145/3583780.3614997.

[231] M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li. Simple and deep graph convolutional networks, arXiv, 2020. doi: 10.48550/ARXIV.2007.02133.

[232] A. Gravina, D. Bacciu, and C. Gallicchio. Anti-symmetric DGN: A stable architecture for deep graph networks. September 2022.

[233] P. J. Huber. Robust estimation of a location parameter. In *Breakthroughs in Statistics*, pages 492–518. Springer New York, New York, NY, 1992. ISBN 978-0-387-94039-7 978-1-4612-4380-9. doi: 10.1007/978-1-4612-4380-9_35.

[234] W. L. Hamilton, R. Ying, and J. Leskovec. Inductive representation learning on large graphs, arXiv, 2017. doi: 10.48550/ARXIV.1706.02216.

[235] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv.org, 2015. doi: 10.48550/ARXIV.1502.03167.

[236] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus), arXiv, 2016. doi: 10.48550/ARXIV.1606.08415.

[237] S. L. Morgan and S. N. Deming. Simplex optimization of analytical chemical methods. *Analytical Chemistry*, 46(9):1170–1181, August 1974. ISSN 0003-2700, 1520-6882. doi: 10.1021/ac60345a035.

[238] T. H. Rowan. *Functional Stability Analysis of Numerical Algorithms*. PhD thesis, The University of Texas at Austin, 1990.

[239] D. Ardia, K. Boudt, P. Carl, K. Mullen, and B. G. Peterson. Differential evolution with DEOptim: An application to non-convex portfolio optimization, April 2010.

[240] T. Pannetier, C. Martinez, L. Bunnefeld, and R. S. Etienne. Branching patterns in phylogenies cannot distinguish diversity-dependent diversification from time-dependent diversification. *Evolution*, 75(1):25–38, January 2021. ISSN 0014-3820. doi: 10.1111/evo.14124.

[241] D. J. Cole. Parameter redundancy and identifiability in hidden Markov models. *METRON*, 77(2):105–118, August 2019. ISSN 2281-695X. doi: 10.1007/s40300-019-00156-3.

[242] S. P. Blomberg, T. Garland, and A. R. Ives. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution*, 57(4):717–745, April 2003. ISSN 0014-3820, 1558-5646. doi: 10.1111/j.0014-3820.2003.tb00285.x.

[243] J. P. Townsend, Z. Su, and Y. I. Tekle. Phylogenetic signal and noise: Predicting the power of a data set to resolve phylogeny. *Systematic Biology*, 61(5):835, October 2012. ISSN 1063-5157. doi: 10.1093/sysbio/sys036.

[244] M. Steel and A. McKenzie. Properties of phylogenetic trees generated by Yule-type speciation models. *Mathematical biosciences*, 170(1):91–112, March 2001. ISSN 0025-5564. doi: 10.1016/s0025-5564(00)00061-4.

[245] T. Gernhard. The conditioned reconstructed process. *Journal of theoretical biology*, 253(4):769–778, August 2008. ISSN 0022-5193. doi: 10.1016/j.jtbi.2008.04.005.

[246] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD02, pages 694–699, Edmonton Alberta Canada, July 2002. ACM. ISBN 978-1-58113-567-1. doi: 10.1145/775047.775151.

[247] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*, Icml '05, pages 625–632, Bonn, Germany, 2005. ACM Press. ISBN 978-1-59593-180-1. doi: 10.1145/1102351.1102430.

[248] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[249] G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[250] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, July 2018. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2017.1307116.

[251] A. N. Angelopoulos and S. Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, arXiv, December 2022. doi: 10.48550/arXiv.2107.07511.

[252] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1050–1059. PMLR, June 2016.

[253] S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, February 1997. ISSN 1943-2631. doi: 10.1093/genetics/145.2.505.

[254] S. Kiranyaz, T. Ince, and M. Gabbouj. Real-time patient-specific ECG classification by 1-D convolutional neural networks. *IEEE Transactions on Biomedical Engineering*, 63(3):664–675, March 2016. ISSN 1558-2531. doi: 10.1109/TBME.2015.2468589.

[255] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. ISSN 1476-4687. doi: 10.1038/323533a0.

[256] Kolmogorov–Smirnov test. In *The Concise Encyclopedia of Statistics*, pages 283–287. Springer, New York, NY, 2008. ISBN 978-0-387-32833-1. doi: 10.1007/978-0-387-32833-1_214.

[257] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A: Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, January 1922. ISSN 0264-3952. doi: 10.1098/rsta.1922.0009.

Acronyms and Terms

ABC Approximate Bayesian computation, a simulation-based Bayesian inference approach that approximates the posterior distribution by comparing summary statistics computed from simulated data to those from observed data, without requiring an explicit likelihood [129, 130, 253].

BD Birth–death model, a class of stochastic branching-process models in which lineages speciate (birth) at rate λ and go extinct (death) at rate μ. BD models form a standard baseline for modelling diversification and generating phylogenetic trees [79].

CNN1D One-dimensional convolutional neural network, a convolutional architecture that applies 1D filters along an ordered axis (e.g., time or sequence position) to learn local patterns [254].

DDD Diversity-dependent diversification, a class of birth–death models in which diversification rates (typically speciation and/or extinction) depend on diversity, often represented by the number of extant lineages. DDD models are commonly used to capture ecological limits such as niche saturation or carrying capacity effects [82].

DNN Dense neural network, a feed-forward neural network composed of fully connected layers, commonly used for regression or classification tasks [255].

ECE Expected calibration error, a metric of probabilistic calibration that measures the average discrepancy between predicted probabilities and observed frequencies (typically computed by binning predictions) [248].

ED Evolutionary distinctiveness, a measure of how unique or distinct a species is in terms of its evolutionary history. ED is computed by summing the pairwise distances between each lineage and all other lineages, divided by the number of lineages minus one. A species with high ED has fewer close relatives and represents a larger amount of independent evolutionary history than a species with low ED [179].

ER Evolutionary relatedness, a measure of evolutionary relationships between species. In our study, ER can be measured as PD, ED or NND (see definitions below).

Gamma Gamma statistic, a measure of internal node distribution. The value of gamma is calculated based on the distribution of node heights. A negative gamma value suggests a decrease in diversification rate over time, while a positive gamma value indicates an increase in diversification rate [62].

GNN Graph neural network, a neural-network architecture designed to operate on graph-structured data by iteratively aggregating information from neighboring nodes and/or edges [139].

J One J One balance index, a measure of the degree of balance of a phylogeny. It is defined for trees with any degree distribution, and enables meaningful comparison of trees with different numbers of tips [184].

KS Kolmogorov–Smirnov test, a nonparametric test that quantifies the maximum difference between two empirical cumulative distribution functions [256].

LSTM Long short-term memory network, a recurrent neural network architecture that uses gating mechanisms to capture long-range dependencies in sequential data [137].

MBL Mean branch length, a measure of average evolutionary change or divergence represented in a phylogenetic tree. It is calculated by summing the lengths of all the branches in a phylogenetic tree and dividing by the total number of branches [51].

MLE Maximum likelihood estimation, a parameter estimation approach that chooses the parameter values that maximize the likelihood of the observed data under a specified statistical model [257].

MPD Mean pairwise distance, a measure of the average pairwise phylogenetic distance between species in a clade. It calculates the average of all pairwise distances between species in a phylogenetic tree, providing a measure of the overall phylogenetic spread [51].

NND Nearest neighbor distance, a measure of how distinct a species is in terms of its evolutionary history with respect to the most closely related species. NND is computed by taking the branch length distance on a phylogenetic tree from one species to its nearest neighbor on the tree. While ED measures distinctiveness on a global phylogenetic level, NND measures it on a local phylogenetic level [50].

PBD Protracted birth–death model, a birth–death framework that explicitly models speciation as a process taking time to complete (e.g., an incipient stage followed by completion). This allows some incipient lineages to go extinct before becoming fully recognized species and can reduce the ”pull of the present” in reconstructed phylogenies [90].

PD Phylogenetic diversity, a measure of biodiversity that takes into account the evolutionary relationships between species. Specifically, PD is the minimum total length of all the phylogenetic branches required to span a given set of taxa on the phylogenetic tree. It provides insight into the evolutionary history represented by a group of species [48].

Rogers Rogers balance index, a measure of ”patristic distance”, which represents the sum of branch lengths between two taxa in a phylogenetic tree which reflects the total amount of change along the evolutionary path between two taxa [185].

Speciation rate evenness A measure of how evenly the speciation rates are distributed on the tips of a phylogeny, using an approach originally proposed by Helmus et al. [98], which is similar in concept to Pielo's evenness index in community ecology. Its value is between 0 and 1. Values less than 1 represent increasing unevenness of speciation rates. It is also sensitive to tree topology: if the tips of a tree have the same speciation rates, the evenness value will be 1 (maximum) only when the tree is star-like.

SR Species richness, a measure of the number of different species present in a particular area or ecological community. It is a simple count of species.

Curriculum Vitæ

Tianjian Qin

Research Profile

Theoretical biologist specializing in data-driven mathematical modeling of complex biological and epidemiological systems. My work combines stochastic processes, phylogenetic and phylodynamic methods, graph and network theory, and cutting-edge AI technologies to study diversification and infectious disease dynamics.

I am increasingly focused on infectious disease epidemiology in animal populations, leveraging:

- **Machine learning and deep learning** for pattern discovery, prediction, and model comparison;
- **Stochastic birth–death models** for diversification and epidemic dynamics;
- **Network- and graph-based models** for contact and movement structures;
- **Bioinformatics and geographic information systems** to link sequence, spatial, and epidemiological data;
- **High-performance computing and big-data frameworks** for large-scale simulations and inference;
- **Open, reproducible software and interactive web applications** to support research, policy, and teaching.

Research Interests

- Contact and livestock trade networks
- Stochastic epidemic and diversification processes
- Statistics and visualizations on dynamic networks
- Bayesian and likelihood-free inference
- AI augumented infectious disease modeling
- Open science, reproducible pipelines, and interactive teaching tools

Education

PhD in Evolutionary Life Sciences (Theoretical Biology) *2019 – 2026*

University of Groningen, Groningen Institute for Evolutionary Life Sciences (GELIFES), Netherlands

Supervisors: *Prof. Dr. Rampal Etienne, Dr. Luis Valente & Dr. Koen van Benthem*

Research on stochastic diversification processes and neural network–based inference from phylogenetic trees. Development of the `eve` model (evolutionary relatedness–dependent diversification) and simulation/inference pipelines in `R`, `C++`, and `Python`.

MSc in Ecology (Wetland Ecology and Invasion Biology) *2016 – 2019*

Beijing Forestry University, School of Nature Conservation, China

Supervisors: *Prof. Dr. Hongli Li, Prof. Dr. Feihai Yu*

Field and experimental work on wetland plant communities and biological invasions. Combined phylogenetic diversity metrics, community composition, and trait data to investigate resistance to invasion by *Alternanthera philoxeroides* and other invasive species.

BSc in Marine Biology *2012 – 2016*

Nanjing Normal University, School of Life Sciences, China

Broad foundation in ecology and evolution, with emphasis on marine biodiversity, fieldwork, quantitative ecology, and programming.

Academic and Research Experience

Post-Doctoral Researcher – Infectious Disease Modeling *Ongoing*

Wageningen University & Research, Netherlands

- Developing network-based models and non-parametric samplers of livestock trade and contact networks to construct digital twins of animal movement systems for disease-spread simulation and policy-making.
- Applying spectral graph theory, stochastic processes, and other mathematical tools to reveal complex temporal network structures.
- Developing neural network approaches for modeling spatial point patterns of the farms of the Netherlands.
- Developing `Transformer`-based approaches for dynamic livestock trade network contact prediction.
- Developing a `Python` library for advanced dynamic network statistics and visualizations.

- Developing `HerdLink`, an interactive web-based application for exploring and assessing the livestock trade networks of the Netherlands.
- Developing `R` package `HandelR`, an automated scheduler for managing research data and running scripts to support open science and reproducible research.
- Exploring integration of phylogenetic/diversification tools with phylodynamic thinking for zoonotic and livestock-associated pathogens.
- Supporting approximate Bayesian computation calibration and HPC configuration.

Doctoral Researcher – Theoretical & Computational Biology *2019 – 2026*

University of Groningen, Netherlands

- Traveled to 43 countries, gaining first-hand experience of diverse ecosystems and cultures.
- Developed stochastic birth–death models (`eve`) to investigate how evolutionary relatedness shapes speciation and extinction at multiple phylogenetic scales.
- Developed novel statistics, simulation algorithms and visualization tools for time–varying stochastic branching processes.
- Implemented high-performance simulation pipelines in `R`, `C++` and `Python`, generating large ensembles of phylogenetic trees under complex diversification scenarios.
- Designed neural network inference frameworks (`EvoNN`) for parameter estimation and scenario classification from phylogenetic trees.
- Applied approximate Bayesian computation, likelihood-based methods, and deep learning to assess identifiability, redundancy and robustness of diversification models.
- Extensive experience working on supercomputers and in `Unix/Linux` shell scripting.
- Emphasized open science: released simulation tools as open-source software with thorough documentation and reproducible workflows.
- Collaborated on the `C++` based `R` package `treestats` for efficient computing of summary statistics on phylogenies.
- Collaborated on the `R` package `DDD` (Diversity-Dependent Diversification) for implementing reliable maximum-likelihood estimation optimizers.
- Collaborated on the *Brassicaceae* Tree of Life project and developed `JavaScript`-based phylogenetic manipulation tool `miniape`.

Researcher in Ecology & Invasion Biology (MSc) *2016 – 2019*

Beijing Forestry University, China

- Conducted extensive field surveys across 18 provinces in China to quantify wetland plant communities and invasion patterns.
- Developed experience with GIS tools by modeling invasive species occurrences at country scale.
- Established automated pipelines of leaf image analysis and phylogenetic tree reconstruction for the research group.
- Analyzed how phylogenetic diversity and species richness affect community resistance to invasion along latitudinal gradient.
- Performed greenhouse experiments manipulating artificial plant communities to study functional and phylogenetic interactions among native species and *A. philoxeroides*.

Grants, Scholarships, and Awards

Grants

- **Implementatie-impuls COVID-19 programma (ZonMw)** (Mar 2026 – Dec 2026)
  Co-applicant. Project: *SSS-mod Paraatheidspakket bij toekomstige respiratoire uitbraken.*
- **Small Compute Applications (NWO)** (2025 – 2026)
  Received to support high-performance computing for data-intensive modeling work.

Scholarships

- **CSC–RUG Joint Scholarship** (2019 – 2023)
  Competitive scholarship to support PhD research in theoretical biology.
- **First-Class Academic Performance Scholarship (Graduate)** (2018 – 2019)
  In recognition of outstanding performance within the graduate cohort.
- **First-Class Academic Performance Scholarship (Graduate)** (2017 – 2018)
  In recognition of outstanding performance within the graduate cohort.

Awards

- **Denise Kirschner Best Student Paper Prize**, Journal of Theoretical Biology (2025)
  Inaugural best graduate student paper prize for: Qin, T., Valente, L., & Etienne, R.S. (2025).
- **Graduate Academic Innovation Award**, Beijing Forestry University (2018)
  Awarded for innovative graduate research.

Professional Skills

Quantitative and Modeling Skills

- Stochastic birth–death and branching processes.
- Network and graph-based modeling.
- Machine learning and deep learning.
- Approximate Bayesian computation, likelihood-based inference, Markov chain Monte Carlo.
- Phylogenetic and phylodynamic analysis; tree and network statistics.
- Geospatial analysis and statistics; spatial point pattern modeling.

Programming and Technical Skills

- **R**: package development, simulation, data analysis, visualization.
- **C/C++**: high-performance simulation engines, R/Python integration.
- **Python**: ML pipelines, automation scripts, graph/network analysis.
- **Web**: HTML, CSS, JavaScript, D3.js for interactive teaching and outreach tools.
- **HPC & Linux**: Environment configuration, bash scripting, job scheduling, pipeline automation.
- **Git & CI/CD**: Git/GitLab workflows, GitHub Actions for automated testing and pipelines.
- **Server**: deploying and maintaining websites and online game servers on cloud-based Linux servers.
- LaTeX and visualization tools (TikZ, PGFPlots) for scientific writing and figures.

International and Interdisciplinary Experience

- Fieldwork across multiple provinces in China during MSc training, including wetland biodiversity and invasion surveys.
- Research and collaboration experience at the interface of ecology, evolution, statistics, and computer science.
- Comfortable working in multicultural academic environments and interdisciplinary teams.

Languages

- **Chinese**: native

- **English**: proficient (IELTS Certificate C1, 2019).
- **Dutch**: intermediate (DUO *inburgeringsdiploma* A2, 2025).
- **German**: reading and basic communication (Summer Camp A2.2, 2013).

Personal Interests

- Photography and digital storytelling.
- Travel and cultural exploration.
- Continuous learning at the interface of biology, mathematics, and computer science.

List of Publications

Peer-Reviewed Articles

1. **Qin, T.**, van Benthem, K., Valente, L.†, & Etienne, R.† (2025). Parameter estimation from phylogenetic trees using neural networks and ensemble learning. *Systematic Biology*.

2. **Qin, T.**, Valente, L.†, & Etienne, R.† (2025). Impact of evolutionary relatedness on species diversification and tree shape. *Journal of Theoretical Biology*.

3. Sun, K., Liu, X.-S., **Qin, T.-J.**, Jiang, F., Cai, J.-F., Shen, Y.-L., A, S.-H., & Li, H.-L. (2021). Relative abundance of invasive plants more effectively explains the response of wetland communities to different invasion degrees than phylogenetic evenness. *Journal of Plant Ecology*.

4. **Qin, T.**, Zhou, J., Sun, Y., Müller-Schärer, H., Luo, F., Dong, B., Li, H., & Yu, F.-H. (2020). Phylogenetic diversity is a better predictor of wetland community resistance to *Alternanthera philoxeroides* invasion than species richness. *Plant Biology*.

5. Wan, J.-Z., Wang, M.-Z., **Qin, T.**, Bu, X.-Q., Li, H.-L., & Yu, F.-H. (2019). Spatial environmental heterogeneity may be the driver of functional trait variation in *Hydrocotyle vulgaris* (Araliaceae), an aquatic plant invader. *Aquatic Biology*.

6. **Qin, T.-J.*** , Guan, Y.-T.* , Quan, H., Dong, B.-C., Luo, F.-L., Zhang, M.-X., Li, H.-L., & Yu, F.-H. (2019). Growth traits of the exotic plant *Hydrocotyle vulgaris* and the evenness of resident plant communities are mediated by community age, not species diversity. *Weed Research*.

7. **Qin, T.-J.**, Guan, Y.-T., Zhang, M.-X., Li, H.-L., & Yu, F.-H. (2018). Sediment type and nitrogen deposition affect the relationship between *Alternanthera philoxeroides* and experimental wetland plant communities. *Marine and Freshwater Research*.

8. Liu, L., Guan, Y.-T., **Qin, T.-J.**, Wang, Y.-Y., Li, H.-L., & Zhi, Y.-B. (2018). Effects of water regime on the growth of the submerged macrophyte *Ceratophyllum demersum* at different densities. *Journal of Freshwater Ecology*.

Preprints and Under Review

- **Qin, T.**, van Benthem, K., Valente, L.†, & Etienne, R.† Identifying evolutionary relatedness effects on diversification from phylogenies using neural networks. Preprint.

Manuscripts in Preparation

- **Qin, T.**, Atamer Balkan, B., Schmid, B., & ten Bosch, Q. Towards better modeling and evaluation of synthetic livestock movement networks. Manuscript in preparation.

* *indicates authors who contributed equally to this work;* † *indicates joint senior authors.*