Deep Learning Derived Traits for Phylogenetic Inference

Roberta Hunt

May 31, 2024
Version: Draft 0.1

University of Copenhagen
Department of Computer Science
Image Section

Computer Science

Deep Learning Derived Traits for Phylogenetic Inference

Roberta Hunt

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen

Chairperson	Stefan Sommer Department of Computer Science University of Copenhagen
External Reviewer	Sergei Tarasov The Tarasov Lab Finnish Museum of Natural History
External Reviewer	Emily A. Hartop Department of Natural History Norwegian University of Science and Technology
Supervisors	Kim Steenstrup Pedersen and François Lauze

May 31, 2024

Roberta Hunt

Deep Learning Derived Traits for Phylogenetic Inference

Computer Science, May 31, 2024

Chairperson: Stefan Sommer

External Reviewers: Sergei Tarasov and Emily A. Hartop

Supervisors: Kim Steenstrup Pedersen and François Lauze

University of Copenhagen

Image Section

Department of Computer Science

Universitetsparken 1

2100 Copenhagen

Abstract

Phylogenetics is a vast and important topic in biology, with far-reaching applications. Yet the state of the art phylogenetic inference process is currently time consuming and requires expert knowledge of the clades being analyzed. In this thesis we explore methods of applying deep learning to the problem of phylogenetic inference. The application of deep learning to phylogenetics is a broad topic, but our focus is on the automatic extraction of traits from images of insects and how they could be used in existing phylogenetic inference methods. In the first part of this thesis we show how traits extracted from images using deep metric learning techniques carry a phylogenetic signal and how these traits can be used in conjunction with genetic data. Combining the morphological traits with traditional genetic traits does have some advantages, for example, we can place species without genetic data on the tree, however the accuracy of the trees generated by adding the morphological traits shows that this area requires further study to be viable.

Eventually, in the interest of exploring explainability in the model, and directly combining deep learning with the phylogenetic inference methods, we moved towards working with simulated genetic data, as this allows us to work with simulated ground truth phylogenies and test methods more rigorously without the uncertainty of the empirical phylogeny. We present methods for directly optimizing deep learnt traits based on a phylogeny which could lead to further explainability of trait extraction algorithms in the future. Here we show how these methods can be used to extract traits, but the results are not yet comparable to traditional genetic inference methods.

Finally we discuss the results of this thesis and potential future areas of exploration.

Abstrakt

Fylogenetik er et stort og vigtigt emne inden for biologi, med vidtrækkende anvendelser. Alligevel er den nyeste fylogenetiske inferensproces i øjeblikket tidskrævende og kræver ekspertviden om de klader, der analyseres. I denne afhandling undersøger vi metoder til at anvende deep learning på problemet med fylogenetisk inferens. Anvendelsen af deep learning til fylogenetik er et bredt emne, men vores fokus er på den automatiske udvinding af karaktertræk fra billeder af insekter, og hvordan de kan bruges i eksisterende fylogenetiske inferensmetoder. I den første del af denne afhandling viser vi, hvordan karaktertræk udvundet fra billeder ved hjælp af deep metric learning teknikker bærer et fylogenetisk signal, og hvordan disse karaktertræk kan bruges sammen med genetiske data. At kombinere de morfologiske egenskaber med traditionelle genetiske egenskaber har nogle fordele, for eksempel kan vi placere arter uden genetiske data på træet, men nøjagtigheden af træerne genereret ved at tilføje de morfologiske egenskaber viser, at dette område kræver yderligere undersøgelse for at være fyldestgørende.

Herefter, for at udforske forklarligheden i modellen og kombinere deep learning med de fylogenetiske inferensmetoder, vil vi arbejde med simulerede genetiske data, da dette giver os mulighed for at undersøge simulerede ground truth fylogenier og testmetoder mere stringent uden usikkerhed om den empiriske fylogeni. Vi præsenterer metoder til direkte optimering af deep learnt egenskaber baseret på en fylogeni, som kan føre til yderligere forklaring af karaktertræksekstraktionsalgoritmer i fremtiden. Her viser vi, hvordan disse metoder kan bruges til at udtrække karaktertræk, men resultaterne er endnu ikke sammenlignelige med traditionelle genetiske inferensmetoder.

Til sidst diskuterer vi resultaterne af denne afhandling og mulige fremtidige forskningsideer.

Acknowledgement

This work would not have been possible without the assistance, consideration and patience of many others. First I'd like to thank my supervisor, Kim Steenstrup Pedersen, for being an excellent mentor during this time including being supportive and available for regular and valuable feedback. Second I'd like to thank Alexey Solodovnikov for first giving me a place to feel welcomed in the Natural History Museum of Denmark and making my stay there so engaging, including bringing me on beetle hunting trips, being so open minded about this topic and giving regular feedback during our collaboration. Third, I'd like to thank François Lauze for meeting with us and giving his mathematical insight into our struggles. Thank you to all the other members of the Image Section and Natural History Museum of Denmark and Natural History Museum of Berlin who made my time during research joyful and staved off the existential dread long enough to complete this work. In particular, Nora Forbes, José L. Reyes-Hernández, Thomas Alscher, Robin Bruneau, Jon Sporring, Lennard Hilgendorf, Chelsea Graham, Anders Illum and Sree Selvantharan. I'd like to thank Rudolf Meier and Emily Hartop for providing a space for me at the Natural History Museum in Berlin and setting up my working environment there. Next, but far from least, I'd like to thank my partner Esben Hansen, who despite strongly advising me against doing a PhD, nonetheless greatly supported me throughout. Finally to all my friends, loved ones, fellow kollektivists, larpers and family who have given space to my rants, low points and excitement during this project. Thank you.

Scientific Contributions

The following scientific papers are included in this thesis:

> **Roberta Hunt, Kim Steenstrup Pedersen**; Proceedings of the Asian Conference on Computer Vision (ACCV), 2022, pp. 2967-2983

> **Roberta Hunt, José L. Reyes-Hernández, Josh Jenkins Shaw, Alexey Solodovnikov and Kim Steenstrup Pedersen**; Integrating Deep Learnt Morphological Traits and Molecular Data for Total-Evidence Phylogenetics: Lessons from Digitized Collections, Systematic Biology, Revised Manuscript Under Review

The following datasets are included in this thesis:

> **Roberta Hunt, Kim Steenstrup Pedersen**; (2023) ’Rove-Tree-11 Dataset v1’. Available at: http://doi.org/10.17894/ucph.39619bba-4569-4415-9f25-d6a0ff64f0e3 (Accessed June 2024).

> **Roberta Hunt, José L. Reyes-Hernández, Josh Jenkins Shaw, Alexey Solodovnikov and Kim Steenstrup Pedersen**; (2024) ’Rove-Tree-11 Genetic Data’. Available at: https://erda.ku.dk/archives/78bc5d7c5746eca509bbd8fb2ea68205/published-archive.html (Accessed June 2024).

The following scientific work included in this thesis is not yet published elsewhere and has not yet been peer reviewed:

> **Roberta Hunt, Kim Steenstrup Pedersen**; Phylogenetic likelihood as a Loss Function: Integrating phylogenetics directly into the deep learning of traits. Chapter 5, this thesis.

Contents

Nomenclature

> *Individuals can be divided and conquered, but symbols. . . Symbols endure*
>
> — **Commander Klaes Ashford, The Expanse**

N	Number of species
L	Number of Latent Variables / Independent Characters / Traits
M	Number of Raw Characters / Size of Raw Input Data (ie, Number of Pixels if using images or Base Pairs if using Genetic Data)
S	Number of Examples per Species
τ	Tree Topology
q	Branch Lengths on Topology τ
$\mathbf{V}$	Variance-covariance matrix describing the tree topology and branch lengths. ('**T**' in Felsenstein [20])
$\mathbf{X}$	Input data $\in R^{NxSxM}$
$\mathbf{Y}$	Reconstruction output of autoencoder data $\in R^{NxSxM}$
$\mathbf{W}$	Matrix of independent character traits / latent variable per specimen $\in R^{NxSxL}$, assuming the same number of specimens per species.
$\mathbf{Z}$	Matrix of independent character traits / latent variable per species $\in R^{NxL}$
$\#\{S\}$	Cardinality, or size, of set S
AS	Align Score
nAS	normalized Align Score
nRF	normalized Robinson Foulds (metric)
RF	Robinson Foulds (metric)

Reading Guide

This work is separated into 7 chapters. The initial two provide introduction, motivation and background knowledge. Chapters 3 and 4 present papers and therefore the paper itself is pasted into the text, with further introduction to the work and analysis provided where enlightening. Chapter 5 presents a body of research not yet published elsewhere. It is recommended to read the chapters in order. The final two chapters discuss the work of this thesis and present ideas for future work.

Chapter 3 - Abbey Rove follows our first paper, "Rove-Tree-11" and covers some important initial steps to investigate how we can generate phylogenetic trees from images of digitized pinned insect specimens, including:

1. Assembling the Dataset - Publishing a dataset of images of rove beetles along with an associate phylogeny to be used in deep hierarchical metric learning research

2. Deep Metric Learning Techniques - Showing how existing deep metric learning can be used to extract traits from the hierarchical image dataset

3. Phylogenetic Inference and Comparison Methods - Showing how existing Bayesian phylogenetic inference methods can be applied to said traits

4. Tree Comparison Metrics - Showing that said traits carry a phylogenetic signal and comparing the resulting trees

Chapter 4 - Gattaca provides follows the second paper "Total-Evidence" and builds on the work in the first paper by:

1. Gathering Genetic Data - Assembly of a genetic dataset to complement the Rove-Tree-11 dataset

2. Exploring a total evidence approach - How can the deep learnt morphological traits be combined with genetic data for total evidence analysis

3. Use of Maximum Parsimony Inference - Due to software constraints, it was easiest to use maximum parsimony inference methods to complete the total evidence analysis as opposed to the bayesian methods used in Chapter 3.

Chapter 5 - What's under the likelihood - The third preliminary body of work explores some deeper questions of directly optimizing the phylogenetic inference using known phylogenetic functions or variants thereof. Here I show some ideas, preliminary results and hypotheses around the following topics:

1. Independance assumptions - Can we use deep learning to extract traits which obey the independance assumption of characters

2. Direct optimization - Can we use Felsenstein's likelihood or Blomberg's K to directly optimize deep learnt traits extracted from images which could eventually add explainability to these models

Chapter 6 - Discussion and Future Work Here we present some ideas for continuation of this work.

Finally Chapter 7 provides a conclusion of our research.

Audience This thesis is mainly written with someone like myself 3 years ago in mind: Someone with some computer science and deep learning background and little biological background, who is being introduced to the exciting research area of phylogenetics. With that in mind, many of the introductory biological topics will be too basic for the average bioinformatician, but may be completely new to the average computer scientist. I have tried to also provide some links to introductory material of the fundamentally related computer science topics.

References for the individual papers are attached at the end of the papers themselves since the papers are kept in their original published/preprint format.

1 Introduction

> *This book was written using 100% recycled words.*
>
> — **Terry Pratchett**
> Wyrd Sisters

The goal of this work is to explore how deep learning could be used to infer explainable evolutionary relationships from digitized images of specimens from museums.

1.1 Project Motivation

Phylogenetics as a field: Phylogenetics is the retrospective study of evolutionary relationships. Phylogenetic inference itself aims to answer surface level questions such as: How related are two species? At what point in time did these two species diverge? While on the surface (and certainly before I stepped into this field) these questions may appear innocuous, they are fundamental stepping stones to further important biological questions such as which biological traits have stronger evolutionary pressure to evolve (ie, the carcinization theory [56]). This can, for example, inform us about desirable evolutionary traits.

To put it simply, phylogenetics aims to infer evolutionary trees such as that shown in fig. 1.1. Such trees have several potential applications in biological research, a few concrete applications of phylogenetics are:

1. Biodiversity and Conservation: There are many potential measures of biodiversity. One of the easiest is simply genetic diversity [11], however, due to varying evolutionary rates in populations and intraspecific genetic diversity, this may not reflect endangered taxa and therefore phylogenetic biodiversity [48]. Phylogenetics is important to determining evolutionary distances which informs evolutionary diversity and determines where best to place conservation efforts to maintain the highest ancestral diversity [106].

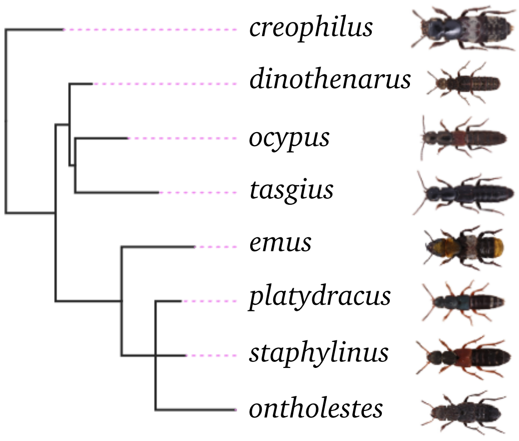


Fig. 1.1: Example phylogenetic tree. Lengths of horizontal branches represent calculated evolutionary distance. Polytomies (nodes with more than two children), such as the clade (*platydracus, staphylinus, ontholestes*) indicates uncertainty in the lineage. Many other ways of representing phylogenetic trees exist. Graphic made in part using the toytree package [18].

2. Virology: Phylogenetics is used in virology to study how infections spread by tracing infections and mapping them to the phylogeny of the virus [97].

3. Agriculture: Phylogenetics has applications in agriculture where it is used to help farmers understand the spread of disease and how to promote disease resistance in crops [57].

4. Forensics: Phylogenetics has been used in some court cases to prove transmission, for example in cases regarding HIV transmission [77].

5. Comparative Studies: Phylogenetics is necessary to account for the phylogenetic relationship in comparative studies of traits. This was famously used by Felsenstein to study the relationship between brain size and body weight in species [23].

The diversity of applications above highlights the wide-reaching real world impact of the field of phylogenetics.

How could deep learning contribute to phylogenetic research? The field of deep learning has expanded greatly in the last decade and continues to grow [45]. Deep learning has proven highly effective at learning complex relationships between datapoints, mirroring or surpassing expert diagnoses in many cases [68, 82].

There are three main areas where I see a strong potential for deep learning to contribute to the phylogenetic research community:

1. **Efficiency** The common methods of phylogenetic inference, such as maximum parsimony and bayesian inference, require iterative searches across tree space using random (or lightly informed) moves [19]. Some work has already shown promise in applying deep learning to suggesting potential tree moves and therefore increasing the efficiency of the tree search process [4].

2. **Scalability** Currently phylogenetic analysis is a meticulous process requiring taxonomic experts to identify phylogenetically important genes and morphological traits. This makes phylogenetic inference slow, sometimes taking weeks to run [24]. Deep learning methods however have shown potential to reach or surpass expert level classification accuracy [68] and vastly surpass them in scalability. Therefore it seems natural to assume that with the correct formulation, such methods may outperform experts at phylogenetic inference. Perhaps not in accuracy, but likely in scalability. The scalability factor is of particular importance when we consider that there are a estimated 5.5 million species of insects alone [88] with only 1 million currently described [84]. In order to understand and analyze life we need methods which can scale with the breadth of natural diversity. Current methods of phylogenetic inference can take weeks to analyze 10 000 sequences for a single gene [24]. Supertree methods, which combine smaller trees into larger ones, can be used to combine phylogenies from separate inference processes. However they are shown to have trouble reaching the accuracy of direct inference methods, and require their own hyperparameter tuning [101]. Therefore if deep learning could be used to compress the data to make inference methods on larger datasets feasible, or directly infer phylogenies for large dataset, this could vastly improve the scalability.

3. **Objectivity** The process of morphological trait choice is subjective. One study shows that subjective expert chosen morphological traits perform worse than objective morphometrics at species delimitation tasks [61]. Taxonomic experts have some biases due to their history with the clade in question. This bias is somewhat desirable, as it is based on years of research and understanding of the biological clade along with evolutionary relationships. Boster and Johnson [10] even showed that expert trait identification is more varied than novice trait identification, suggesting that expert trait choice is highly subjective, although this was in a study of fish similarity. However, if it is possible for us to embed evolutionary understanding into the deep learning models, or better yet, have the deep learning models learn these relationships independently, we can increase objectivity.

Given the above, I expect that deep learning methods may contribute greatly to the field of phylogenetics.

Why focus on morphological traits? In this work we choose to focus on morphological traits, even though the phylogenetic community has tended to focus on the genetic inference methods. The focus on genetic data happens for a variety of reasons. Firstly, genetic data is typically regarded as carrying a stronger phylogenetic signal [113], this is supported by systematic issues with morphological traits such as homologies, mimicry and convergent evolution [113]. Secondly, the high level of taxon-specific expertise required to score and defend morphological trait choice make morphological analyses increasingly cumbersome to biologists. Nevertheless, not only is it not possible to obtain genetic data from some species due to extinction, dna degradation, or issues with obtaining specimens [49], but this process is also usually destructive to the specimen [70]. Additionally morphological traits have been shown to improve phylogenetic inference when traits are chosen appropriately [49]. That said, the extent to which morphological traits should be used is debated [79]. We choose to focus on morphological traits, given the above, because we expect that deep learning can greatly increase the objectivity and scalability of morphological trait extraction, allowing automatic placements of species where genetic data is not possible to obtain.

Is habitus enough? Dorsal images of insects do not carry information about all the morphological traits that experts would use in phylogenetics. A fair criticism of this work is whether the ’habitus’, or overall external features of the organism, can provide enough evolutionary information to form a phylogeny. While information from the habitus is not the only information used in morphological traits, it is of importance, and takes us back to the historic phenetics analysis of Sokal and Sneath [85], who were among the first to generate phylogenies from morphological traits in a systematic manner. While we agree that including further internal morphological traits would be beneficial (as they have proven to be for traditional systematics), we see using external habitus information as a first step in this direction, and assume that methods developed in this area of research could be applied to more complex datasets, of for example, 3d scans of specimens which would include important morphological features.

1.2 Project and Research Questions

This work is part of the larger PHYLORAMA project, which is a collaborative project between the University of Copenhagen and the Natural History Museum of Denmark. Over the course of three phd projects PHYLORAMA aims to use state of the art computer science and biological methods to improve phylogenetic inference. One of the PhD projects focuses on traditional methods of phylogenetic reconstruction for the species-rich group of rove beetles, a focus which this research continues. Another

PhD project focuses on accurate and cheap 3D reconstruction of fossils preserved in amber. This would allow widespread accurate digitization of amber specimens, allowing them to be more easily and automatically integrated into the phylogenetic inference process. Overall, we hope PHYLORAMA will promote the digitization of museum specimens, and the use of advanced computer science techniques in the phylogenetic inference processes.

My thesis focuses on the automatic inference of phylogenies from images of pinned insects. As introduced above, we are interested in increasing the efficiency, scalability and objectivity of phylogenetic inference using deep learning on images. To do this, we first explore if we can use deep learning methods to extract phylogenetic traits from images, similar to classic phenetic methods. The extraction of these traits allows us to compress the phylogenetic data in a compact number of traits which can be efficiently run through existing inference methods. To understand how these methods compare to existing methods, we then want to gather genetic material and compare with existing genetic methods in so called total evidence analysis. Finally, given issues with explainability in deep learning models, we then want to explore how deep learning might be used to explain phylogenetic trees and add objectivity to trait extraction.

More specifically the research questions investigated in this work are:

1. How can deep learning best be used to extract phylogenetically relevant traits from dorsal images of pinned insect specimens?

   [Chapter 3] explores this by first assembling a dataset of pinned rove beetle images, then using existing deep metric learning methods to extract traits from images, using existing inference methods to infer phylogenetic trees from the extracted continuous traits and finally assessing the generated trees using existing metrics.

2. How could such traits be used in conjunction with genetic data for total evidence analysis?

   [Chapter 4] explores this by collaborating with biologists who gathered relevant genetic data and comparing trees inferred from deep learning morphological traits to trees inferred from molecular traits, and how they could be combined into total evidence analyses.

3. How can we add explainability to such models, so that deep learning methods can direct us towards phylogenetically relevant traits?

[Chapter 5] explores how we can use deep learning to explain traits related to a specific phylogeny and how we can improve the deep learning models. As a first step to this we first look at if we give the deep learning model the phylogenetic relationships, can it learn to extract important traits. To do this we use simulated genetic data to give certainty in the phylogeny and validity of evolutionary assumptions.

1.3 Potential Impact of Work

Given the above, it is our hope that the integration of deep learning into the phylogenetic inference process will ensure it is more streamlined, requires less expert knowledge, add objectivity and most importantly, make it feasible to complete the inference on larger sets of taxa. This could give us a much better overview of the natural history of our world and understanding of how to optimize our place in it.

1.4 Related Work

While Sapoval et al. [76] rightly assert that the application of deep learning in phylogenetics is in its infancy, research in this area is steadily growing. Here I give a brief reference to other attempts in this domain. Mo et al. [58] provide a more thorough overview of machine learning methods applied to phylogenetics, however they focus on methods for genetic data. The following section draws heavily from their summary paper.

Building upon Mo et al. [58], we can categorize the deep learning methods based on their goal as follows:

Topological Inference - Here, like with parts of this thesis, the goal is to directly infer the topology. One intuitive way to do this is to phrase it as a classification problem, however this quickly becomes computationally unrealistic as the number of unrooted bifurcating trees is $\frac{(2N-5)!}{2^{N-3}(N-3)!}$, where N is the number of taxa. Therefore with as few as 15 taxa the number of possible unrooted bifurcating trees is over 7 trillion. Thus researchers found alternative ways to approach the problem.

One way of doing so are so-called **quartet-based methods**, which predate modern deep learning, but are nonetheless intuitively appealing. In these methods we split the taxa into groups of 4 (a quartet). A quartet has 3 possible unrooted bifurcating tree topologies, therefore it's easy to frame as a deep learning classification problem, which is exactly what Suvorov et al. [91] and Zou et al. [112] did, each using

different approaches for input, dataset, and network. Solis-Lemus et al. [86] and Wang et al. [100] took a similar approach but using LSTMs and windowed input, allowed for variable length sequences which did not require alignment. Despite the success of neural networks at inferring quartets, there is still no way to merge quartets than the existing amalgamation algorithms, which provided better results on large trees when used with traditional methods and not deep learning[108]. There are also issues posed when trying to transfer these methods which were trained on simulated data to empirical datasets [108].

Another popular approach are **distance-based methods**, which also predate modern deep learning. These define a distance measure between the taxa, and merge leaves based on said distance according to some algorithm to build a tree. (see section 2.1 for more details on eg, neighbour joining). Therefore the goal of these methods is to correctly define a distance between taxa. To do this Cuthill et al. [16] used deep convolutional triplet networks to show that phenotypic distances of butterflies extracted using deep learning can carry a phylogenetic signal. This was separately confirmed by Kiel [42] who demonstrated this using classification loss on images of bivalves and by us in [37] (see Chapter 3) where we compared various loss functions. More recently Furusawa et al. [26] used variational autoencoders to show that this can be applied to projections of 3D models of primate mandibles. Adaïmé et al. [3] used a fusion model of extracted features to generate embeddings and finally phylogenies using bayesian inference. In 2020, Bhattacharjee and Bayzid [6] used autoencoders with a simple reconstruction loss to successfully impute values in incomplete distance matrices and generate resulting phylogenetic trees. In 2022, Nesterenko et al. [64] presented Phyloformer, which used a transformer based architecture to perform regression on the distance between two sequences using mean squared error loss. Jiang et al. [40] introduced DEPP, a neural network which can place new sequences on a given tree. Smith and Hahn [83] developed PhyloGAN which uses a discriminator network to distinguish between simulated and observed alignments in order to improve the simulated phylogeny the alignments are generated from.

Other related methods of applying deep learning to the problem of phylogenetic inference, which are not the focus of this thesis include:

1. Methods for **improving aspects of the topological inference process**, such as determining the degree of difficulty of a dataset [34], suggesting optimal tree moves [4], classifying quartets as Felsenstein or Farris [50], or improving the efficiency of calculating the tree probabilities [109].

2. **Branch length Inference** methods which assume the tree topology is known and the goal is to optimize the branch lengths. Suvorov and Schrider [92] used a regression model and given phylogeny to predict all branch lengths at once while Tao et al. [93] used a network to determine which branch-rate model was appropriate for a given phylogeny.

3. **Substitution Model Selection** methods which focus on choosing the appropriate substitution model to use in the inference process. Abadi et al. [1] and Burgstaller-Muehlbacher et al. [13] used summary statistics from, for example, the alignment process, fed into a neural network to predict which substitution model to use during inference.

4. **Discordance Quantification** methods which focus on quantifying the difference between inferred gene trees and inferred phylogenetic trees. Gene trees and phylogenetic trees may not match due to ancestral polymorphisms, horizontal gene transfer, gene duplication or loss, introgression, natural selection, genetic drift, or recombination. Both Rosenzweig et al. [74] and Zhang et al. [111] showed an improvement to traditional discordance quanitification using deep learning.

5. **Introgression detection** methods which focus is on using machine learning to detect genetic regions which cause discordance Schrider and Kern [78], Ray et al. [69], Gower et al. [33], Blischak et al. [7], Burbrink and Gehara [12], and Hibbins and Hahn [36] all used neural networks for this purpose, typically convolutional neural networks.

6. **Diversification rate inference** methods focus is on determining the net rate of change in biodiversity, such that $r_{\text{diversification}} = r_{\text{speciation}} - r_{\text{extinction}}$. To do this Voznica et al. [98] introduced a novel matrix method to encode the phylogenetic tree, called CBLV which Lambert et al. [47] then expanded upon.

7. **Viral Transmission** - Sun et al. [90] used a vectorized representation of the nodes and edges in the phylogenetic graph as input to predict pathogen transmission rates.

8. **Ancestral State Reconstruction** methods focus is on predicting the state of internal nodes, or rather, traits of ancestral species. Moreta et al. [59] used a deep generative protein model to complete ancestral sequence reconstruction.

1.4.1 Deep Learnt Morphological Traits for Use in Phylogenetic Analysis

The above related work is heavily focused on genetic data. Currently relatively few studies focus on deep learning methods of extracting phylogenetically relevant traits from morphological data. There are many reasons for this. One is that prior to Rove-Tree-11 [37] there was no standardized image dataset aimed at phylogenetic inference. Although Cuthill et al. [16], Kiel [42] and Furusawa et al. [26] used their own image datasets, they focused mainly on the biological results of their work, and not exploring the deep learning methodology itself.

Another reason that deep learning has been more frequently applied to genetic datasets for phylogenetic inference, is that empirical datasets are uncertain of the true phylogenetic tree. Genetic phylogenetic datasets on the other hand can be readily simulated, giving a known phylogenetic tree which is more reliable for proving methodological superiority. There is currently no reliable way to simulate morphological changes and thus generate a realistic simulated dataset with a known tree. On the other hand, methods of simulating genetic data are well established, with many software packages [51, 25] available. Although the extent to which these genetic simulations accurately represent reality is questionable [95].

2 Background

> *Life is like this dark tunnel. You may not always see the light at the end of the tunnel, but if you keep moving, you will come to a better place.*
>
> — **Uncle Iroh**
> Avatar the Last Airbender

In this chapter I briefly introduce the background knowledge pertinent for this work. I focus on the background knowledge which I did not have when beginning this project, and that I expect most computer scientists would not have prior to working with phylogenetics.

2.1 Phylogenetic Inference Methods

This section provides a simple overview of the different phylogenetic inference methods. For a more complete introduction please see Felsenstein's book "Inferring Phylogenies" [19].

Phylogenetic inference is the process of inferring a phylogenetic tree from observed traits (ie, morphological traits such as the presence of opposable thumbs, or genetic traits such as nucleotide sequences). Below is a brief introduction to popular methods.

2.1.1 Distance Based Methods

As mentioned in section 1.4, distance based methods simply define a distance matrix between the species, and use an algorithm to determine how to merge nodes based on these distances. One of the simplest and oldest distance based inference methods is neighbour-joining, where the two closest taxa (as determined by some measure of distance) are merged into a clade and their traits are aggregated into a new node [75]. This process repeats until all taxa are on the tree. Waterman et al. [102] showed that given a distance matrix which perfectly represents the topology, the perfect topology can always be inferred, and since these methods tend to be faster than iterative inference methods, they have a strong computational advantage. The difficulty, however, lies in calculating the distance accurately.

2.1.2 Iterative Methods

Unlike neighbour joining, the following methods score how well the traits match a given candidate tree. Unfortunately it is difficult to analytically optimize these scores in tree space. Therefore we rely on iterative updates with random initializations of the tree topology / sub tree topologies and branch lengths to find the best tree. To do this, in the simplest case, the tree is randomly initialized. Then random 'moves' are made on the tree (such as nearest neighbor interchange (NNI), Subtree pruning and regrafting (SPR), and tree bisection and reconnection (TBR)), and the loss is recalculated. Monte Carlo Markov Chain methods, such as the Metropolis Hasting sampling algorithm [35], are typically used to accept or reject tree moves. After the topology has been adjusted, the preferred branch lengths can be found, and the score of the new topology can be calculated. The move is then typically rejected or accepted based on the Metropolis Hastings algorithm, and the algorithm repeats until convergence or it reaches a set number of updates and burnins. Several computational improvements have been made throughout the years to this process, for example, once the optimal branch lengths of a specific clade are found, they are not recalculated as this is not necessary (see Felsenstein's pruning algorithm [20]).

Maximum Parsimony is an iterative method which calculates the 'parsimony' of the tree, or rather, attempts to find the tree that requires the fewest/smallest total changes along the branches (depending if we are dealing with discrete traits or continuous traits). For discrete traits this is also referred to as the "edit distance" in computer science. While it has been shown that the most parsimonious tree is not necessarily the most likely given the evolutionary process, maximum parsimony is still used today [103], and in some cases is still preferred over other methods [32].

Maximum Likelihood defines a probability model of observing the traits $\mathbf{Z}$ given the tree topology τ with branch lengths q, $p(\mathbf{Z}|\tau, q)$, and uses this to compute the likelihood of the topology and branch lengths based on the traits. Originally introduced by Cavalli-Sforza and Edwards [14] and made practically feasible by Felsenstein [21], this method is the precursor to Bayesian Inference methods, but is also used today in its own right.

Bayesian Inference builds upon the maximum likelihood approach, but using Bayes rule allows us to instead optimize $p(\tau, q, r|\mathbf{Z}) = \frac{p(\mathbf{Z}|\tau,q,r)p(\tau,q,r)}{p(\mathbf{Z})}$ by allowing us to put priors on the topology τ, branch lengths q, and perhaps more importantly the evolutionary rates r.

In maximum likelihood and Bayesian inference typically multiple trees are produced as output, and then probabilities of observing each branch can be calculated based on this output. Burn-in and bootstrapping are also commonly used in the probability calculation [55].

2.2 Deep Metric Learning

Deep metric learning, also known as representation learning, is a subfield of deep learning concerned with learning useful representations or embeddings of a dataset. Kaya and Bilge [41] amd Ghojogh et al. [31] provide an overview of deep metric learning methods. Deep metric learnt representations are typically in some way either more condensed, more disentangled, or otherwise advantageous to work with compared to the original data. In this thesis we mainly are interested in three areas of deep metric learning research:

Disentanglement is the process by which dependant variables are transformed into independent variables. A simple example is images of a single colored pixel on a black 20x20 pixel background, this could easily be represented by 5 variables, the x and y coordinates of the pixel, and the rgb values of the pixel. However, the original data has $20x20x3 = 1200$ variables. DL disentanglement algorithms attempt to use deep learning to find the independent variables and extract them automatically from the data. For a recent overview of disentangled representation learning methods see Wang et al. [99].

Trait Extraction is arguably what deep metric learning is all about, however, we are specifically interested in extracting phylogenetically relevant traits, as we know that due to convergent evolution, homologies, mimicry, not all traits (particularly morphological traits) carry a phylogenetic signal [44]. However, this field is just beginning and we return to this issue in Chapters 3 and 5.

Hierarchical Deep Metric Learning is an important area of study for us. Hierarchical data and categorization appear in many areas in our society. From coarse to fine-grained taxonomies of objects, to phylogenetic trees. The area of hierarchical deep metric learning concerns itself with either using a hierarchical representation to better complete a downstream task, or itself determining the hierarchy, as we are interested in. Badirli et al. [5] relevantly used a hierarchical Bayesian classifer in their model to complete zero-shot classification of insects using DNA and images.

Ge [28] modified the classic triplet loss function to infer hierarchies during training, and use these hierarchies to dynamically adjust the triplet margin.

2.3 Tree Comparison

In order to compare methods we need quantitative measures of how similar two trees are. Kuhner and Yamato [46] provide a thorough comparison of many different methods. Since most of this work focuses on topology-only phylogenies (ie, they do not contain branch length information), we focus on explaining two metrics which do not need branch lengths: The Robinson-Foulds metric (RF), the Align Score (AS) and the normalized versions of those (nRF and nAS).

2.3.1 Robinson-Foulds

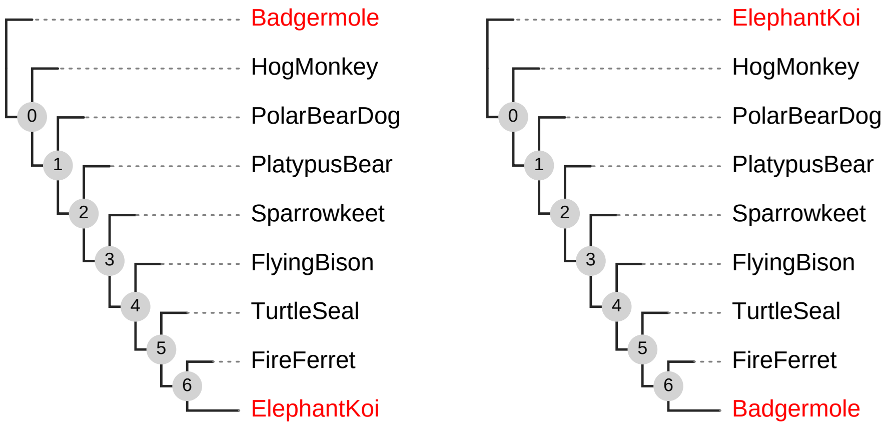
The Robinson-Foulds metric [73] is the most widely used metric in the biological community for comparing phylogenetics trees. To understand how it works, let's assume we are comparing two tree topologies τ_1 and τ_2 with sets of nodes n_{τ_1} and n_{τ_2}, with cardinalities J and K, respectively. Here a tree topology is defined as a directed acyclic graph for rooted trees, and an undirected connected acyclic graph for unrooted trees. Note that the number of internal nodes in each tree does not necessarily need to match ($J \neq K$). However, both trees must have the same set of leaves $L_{\tau_1} = L_{\tau_2} = L = \{L_1, L_2 ... L_N\}$.

First we need to recognize that each node i in a tree, $n_{\tau_1,i}$ can be said to separate the leaves of the tree into a partition of two disjoint sets, a and b, or $L_{\tau_1,i,a}$ and $L_{\tau_1,i,b}$. For the purpose of this calculation we can then say each node is defined as this leaf partition, ie, $n_{\tau_1,i} = \{L_{\tau_1,i,a}, L_{\tau_1,i,b}\}$ We can then define a set of partitions for each topology, representing all the splits in the tree . $P_{\tau_1} = \{n_{\tau_1,1}, n_{\tau_1,2} ... n_{\tau_1,J}\}$

Then, the RF score is defined as the number of nodes which exist in τ_1 but not in τ_2 plus the number of nodes which exist in τ_2 but not in τ_1:

$$RF = \#\{n_{\tau_1,i} \notin P_{\tau_2}\} + \#\{n_{\tau_2,i} \notin P_{\tau_1}\} \tag{2.1}$$

To normalize this value we can simply divide it by the total number of nodes in both trees.



(a) Tree 1, τ_1

(b) Tree 2, τ_2

Fig. 2.1: Toy trees based on species from the ATLA series for highlighting sensitivity of the RF score. The only difference between τ_1 and τ_2 is that the Badgermole and ElephantKoi have switched positions, and yet if we look at the nodes 0-6, we see that none of them have exact matches between the trees. Therefore these two trees have a RF score of 14 (7*2), and a maximum RF score of 14, so therefore an nRF of 1, despite being similar in many respects.

$$nRF = \frac{RF}{J+K} \tag{2.2}$$

The Robinson-Foulds metric, while attractive for its simplicity has some problems. For one, it does not account for branch lengths. For another, and most importantly for us, it can be very sensitive to small changes in a tree. The toy example in fig. 2.1 illustrates this.

2.3.2 Align Score

Due to the sensitivity issue highlighted in fig. 2.1 we prefer to measure methods against the Align Score [65], also sometimes referred to as a kind of generalized Robinson-Foulds, or the Nye Score.

The Align Score works on a similar principle to RF, except instead of giving a binary score to each match between trees, the Align Score first finds a matching between nodes in the trees, and then scores each match between 0 and 1 based on the intersection over union of the best match. Using the terminology above, the Align Score is calculated as follows:

First, they define the measure ’a’ between two sets of leaves $(L_{\tau_1,n_i,a}, L_{\tau_2,n_j,b})$ from the trees τ_1 and τ_2 as the proportion of shared leaves:

$a_{\tau_1 n_i a,\tau_2 n_j b} = \frac{L_{\tau_1,n_i,a} \cap L_{\tau_2,n_j,b}}{L_{\tau_1,n_i,a} \cup L_{\tau_2,n_j,b}}$

Given two nodes, $n_{\tau_1 i}$ and $n_{\tau_2 j}$, one from each tree, each of which partitions two sets of leaves, there are 4 sets of leaves $L_{\tau_1,n_i,a}, L_{\tau_1,n_i,b}, L_{\tau_2,n_j,a}, L_{\tau_2,n_j,b}$. There are therefore 4 possible ways to calculate a for two nodes: $(L_{\tau_1,n_i,a}, L_{\tau_2,n_j,a}), (L_{\tau_1,n_i,b}, L_{\tau_2,n_j,a}),$ $(L_{\tau_1,n_i,a}, L_{\tau_2,n_j,b}), (L_{\tau_1,n_i,b}, L_{\tau_2,n_j,b})$.

Since Nye et al. [65] are interested in finding the 'best' match, they calculate a similarity score for each node as:

$$s(n_i, n_j) = \max(\min(a_{\tau_1 n_i a,\tau_2 n_j a}, a_{\tau_1 n_i b,\tau_2 n_j b}), \min(a_{\tau_1 n_i a,\tau_2 n_j b}, a_{\tau_1 n_i b,\tau_2 n_j a})) \tag{2.3}$$

Using this similarity score, Nye et al. [65] build a KXJ matrix of similarity scores, and use linear sum assignment to find the matches which give the lowest overall 'cost', and therefore present the best overall matches between nodes. Here, if the number of nodes in the trees differ, the smallest number of node matches is made. If $J = 6, K = 4$, then only 4 matches are made.

Thankfully the Align Score is easy to normalize as follows:

$$nAS = \frac{AS}{\min(J, K)} \tag{2.4}$$

The trees in fig. 2.1 give an Align Score of 3.5 and a normalized Align Score of 0.5, compared to a Robinson Foulds score of 14 and a normalized Robinson Foulds score of 1. The normalized Align Score therefore seems more fair as the trees share many similarities.

2.4 Measuring Phylogenetic Signal

Several methods for measuring the phylogenetic signal of a trait exist. Münkemüller et al. [60] provides a thorough overview of these. The two referred to in this thesis are Abouheif's Cmean [2] and Blomberg's K [8] and they are defined as they are used. Briefly, Abouheif's Cmean calculates the autocorrelation of the trait arranged on the tree and is calculated based only on the topology. Blomberg's K instead compares the expected variance of the trait along the tree based on Brownian motion with the calculated variance and requires branch length information. Both use random shuffling of the trait values to show significance.

3 Abbey Rove: The Rove-Tree-11 Dataset

> *Reality leaves a lot to the imagination.*
>
> — **John Lennon**
> The Beatles

This chapter contains the Rove-Tree-11 paper, first published in Lecture Notes in Computer Science, vol 13845, pp 425–441, 2022, Reproduced with permission from Springer Nature:

> **Roberta Hunt and Kim Steenstrup Pedersen**. Rove-Tree-11: The Not-so-Wild Rover, a Hierarchically Structured Image Dataset for Deep Metric Learning Research. In: Wang, L., Gall, J., Chin, TJ., Sato, I., Chellappa, R. (eds) Computer Vision – ACCV 2022. ACCV 2022. Lecture Notes in Computer Science, vol 13845. Springer, Cham. `https://doi.org/10.1007/978-3-031-26348-4_25`

The following dataset is included in this chapter:

> **Roberta Hunt, Kim Steenstrup Pedersen**; (2023) ’Rove-Tree-11 Dataset v1’. Available at: `http://doi.org/10.17894/ucph.39619bba-4569-4415-9f25-d6a0ff64f0e3` (Accessed June 2024).

This paper was our first attempt to show how deep learning could be used to extract phylogenetically relevant morphological traits from images. The dataset properties and acquisition are explained in the paper. See the background section regarding recent research into deep metric learning, tree comparison methods, and measures of phylogenetic signal.

There are two important corrections that we would make to this paper if we did it now. The first is that we were in fact not the first group to extract morphological traits from images as claimed, Cuthill et al. [16] and Kiel [42] also explored this. However, as far as we know we are still the first whose goal was to test the methodology and present a dataset for benchmarking.

The second correction is that the use of the unnormalized Align Score can be problematic as it favors trees with polytomies. We only realized this after publishing this paper, which uses unnormalized values. That said, the overall conclusion of the Rove-Tree-11 paper that the phylogenetic signal is contained in the traits, is still valid as we have checked the scores in the second paper. However, ideally the comparison across loss functions at a species level would be redone to use the normalized Align Score, although we can also see from the second paper that this does not significantly affect the result (most loss functions produced similar nAS values).

Rove-Tree-11: The not-so-Wild Rover A hierarchically structured image dataset for deep metric learning research

Roberta Hunt[1] and Kim Steenstrup Pedersen[1,2]

[1] Department of Computer Science,
University of Copenhagen, Universitetsparken 1, 2100, Copenhagen, Denmark
{r.hunt,kimstp}@di.ku.dk
[2] Natural History Museum of Denmark, Øster Voldgade 5 - 7, 1350, Copenhagen,
Denmark kimstp@snm.ku.dk

Abstract. We present a new dataset of images of pinned insects from museum collections along with a ground truth phylogeny (a graph representing the relative evolutionary distance between species). The images include segmentations, and can be used for clustering and deep hierarchical metric learning. As far as we know, this is the first dataset released specifically for generating phylogenetic trees. We provide several benchmarks for deep metric learning using a selection of state-of-the-art methods.

Keywords: Phylogeny · Dataset · Tree · Hierarchy · Hierarchical Dataset · Rove · Staphylinidae · Phylogenetic Tree

1 Introduction

A phylogeny is a fundamental knowledge frame which hypothesizes how different species relate to each other [11]. A fully annotated phylogeny, i.e. a tree of life anchored in time scale, placed in the geographic context, and with a multitude of organismal traits mapped along the tree branches is an important tool in biology. It explains biodiversity changes over millennia or geological epochs, traces organismal movements in space and evolution of their properties, models populations response to climate change, navigates new species discovery and advises classification and taxonomy. An example phylogeny from our dataset is shown in fig. 1 along with some example images from the most abundant species in the dataset.

Traditionally biologists generate phylogenies [9,10] using genetic data or morphological features (relating to the shape or development of the organism, for example the head shape, or the pattern of the veins on the wings). Despite genetic data dominating phylogenetic research in recent years, morphological features extracted by visual inspection of specimens are still of use. Fossils, for example, contain no genetic data, but morphological features on the fossils can be used to relate them to existing biodiversity [26]. Occasionally morphological and genetic

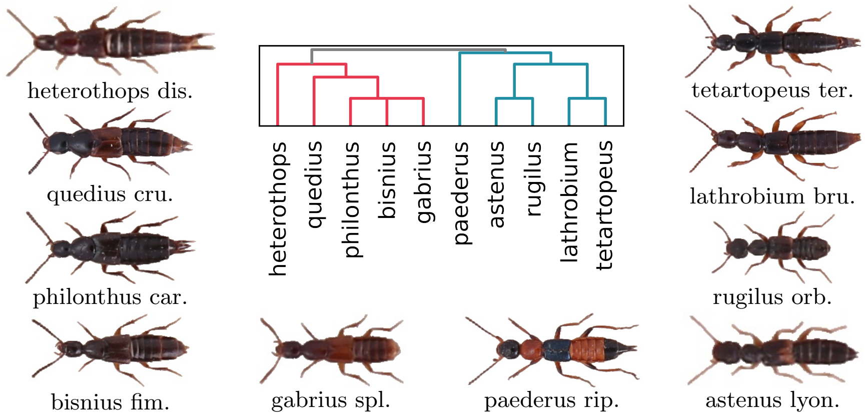


Fig. 1. Subset of phylogeny from the Rove-Tree-11 dataset, for the 10 genera with the most images in the dataset. Each leaf represents a genus. Genera which are closer together on the tree are more closely related, and nodes in the tree represent common ancestors. Nodes with more than two branches are considered not yet fully resolved. Many phylogenetic trees include estimations of time representing when the speciation event occurred (when the common ancestor split into two species). These dates are usually based on fossil evidence. This dated information is unfortunately not currently available for our ground truth tree. Example specimens from each genera are shown for reference.

data are even combined to generate a so called 'total-evidence' phylogeny [34]. Morphological features are also of importance for species/specimens which lack good quality genetic data. Much of phylogenetic research on insects is done from museum specimens captured many years ago. Often the DNA of such specimens has degraded and is no longer of use. Genetic extraction is also expensive, time consuming, and a destructive process which can require completely destroying the specimen, particularly in the case of small insects.

However, the traditional process of generating morphological features is slow, meticulous and introduces some aspects of subjectivity by the researcher performing the analysis. Typically a phylogenetic researcher would generate a matrix of discrete traits (although the use of continuous traits has recently been explored [35]) which they hypothesize are of use in distinguishing the species and are evolutionary important. With thousands of new species of insects discovered each year [1], it is difficult for phylogeneticists to keep up.

Deep metric learning [38,22] is a proven technique to generate informative embedding matrices from images, and we posit that it can be used to generate morphological embeddings which more objectively represent the morphological features of a specimen. In this dataset we are unfortunately only looking at one view of the insect, in our case, the dorsal view (the back), whereas biologists would ideally examine and compare all external and internal features of the insect. However, we hypothesize that this can be offset by the model's ability to learn minute details. Our intention is that these methods could eventually be improved and used as a tool for biologists to inform their decision making process. Additionally, many natural history museums worldwide [8,20] are digitizing their collections, including in many cases, taking images of millions of museum insects. The Natural History Museum of Denmark (NHMD) alone estimates they have over 3.5 million pinned and dried insect specimens spanning 100,000 described species [32] and is in the process of digitizing their collections [31]. The importance of such digitization efforts have been studied from a biology research perspective [17,36]. Thus, given the increased data availability, we predict that phylogenetic generation from images will become a growing field of research within computer vision and related areas of artificial intelligence.

Despite the rapidly growing availability of images of pinned insects from natural history museums, the ongoing push from the biological community to generate phylogenies, and the increasing ability of deep learning to learn complex shapes and relationships, few publicly available datasets exist targeting the generation of phylogenies from images using deep learning techniques. There are several reasons for this, as we will explore in more detail in sec. 2.1, when we compare with existing datasets. In brief, although the number of image analysis datasets is steadily growing, often the graphs which are included in the datasets are subjectively resolved (such as [5]) or the groupings they provide are too coarse-grained (such as [12]) or, particularly for biological datasets, the images are natural photos taken in the wild, meaning they are from various viewpoints and often obscured (such as [40],[41]). This makes it difficult for the model to learn which distinct morphological features are more related to those from others species. Typical morphology based phylogenies are generated from careful inspection and comparison of features, meaning we expect direct comparison to be very important for this task.

In this paper we present 'Rove-Tree-11', a dataset of 13,887 segmented dorsal images of rove beetles along with a ground truth phylogeny down to genus level[‡]. The species-level phylogeny is not included, because this level of information is not yet readily available. Our intention with releasing this data is that it can further research on deep hierarchical metric learning and computer vision solutions for building morphological phylogenies on interesting biological groups, leveraging the current digitization-wave that is gripping natural history museums worldwide.

The **contributions** of this paper are:

1. The release of a new hierarchically structured image dataset including segmentations and ground truth genus-level phylogeny
2. We provide baseline results on this dataset for the tasks of classification, clustering, and for predicting phylogenetic trees.

[‡] to genus-level means that each species within a genus is considered unresolved, or equally likely to be related to any other species within that genus.

2 Related Work

2.1 Comparison with Existing Datasets

Hierarchically structured data is often found in computer vision related tasks. Examples include cognitive synonym relations between object categories such as clothing items[27] and is especially found in tasks concerning nature. However, current datasets which present a ground truth hierarchical grouping of the data are not intended for morphological phylogenetic research, and therefore poorly suited to the task.

There are several natural history related image datasets which do, or could easily be adapted to, include a taxonomy (ie IP102 [44], CUB-200-2011 [41], iNaturalist [40], Mammal dataset [12], PlantCLEF 2021 [13] and ImageNet [6]). With the exception of PlantCLEF, these are however all 'in the wild' images and identification has typically been done by non-experts with the naked eye. The phylogenies are also usually superficial - including only a few levels, and typically based only on the current taxonomy, which is not fine-grained and not necessarily representative of the state of the art phylogenetic tree, as taxonomies have a longer review process[§]. In the case of PlantCLEF the majority of the training images are of herbaria sheets, and therefore not 'in-the-wild', however only a shallow taxonomy is provided with the PlantCLEF dataset. In the case of IP102, the hierarchical tree is grouped by the plant the insect parasitizes, and is not related to ancestral traits at all. With the exception of CUB-200-2011, iNaturalist and PlantCLEF, the species are also easily identified by a layman/amateur by the images alone, which is not necessarily the case in our dataset, where many of the identifications traditionally require a microscope or dissection. It is also often the case that the taxonomy is not properly updated until years after the phylogeny has been altered, particularly in the case of entomology where new species are discovered regularly, so using the most recent taxonomy may not actually represent the state-of-the-art knowledge of the evolution of the species. In the case of iNaturalist, the dataset does include a tree with the same number of levels as Rove-Tree-11, however, this depth begins from kingdom-level, whereas ours begins from family level (four taxonomic ranks lower on the taxonomic hierarchy), and represents the most recent phylogeny.

Additionally there are non-biological hierarchical datasets, such as DeepFashion [27], for which others have created their own hierarchy [5]. This hierarchy is however based on loose groupings of clothing items which are highly subjective. For example, the top-level groupings are: top, bottom, onepiece, outer and special, where special includes fashion items such as kaftan, robe, and onesie, which might morphologically be more related to coats, which are in the 'outer' cate-

[§] the taxonomy represents how the organism is classified - ie which class, order, family the organism belongs to, and is a non binary tree. The phylogeny represents how related different species are together, and would ideally be a binary tree. In an ideal world the taxonomy would be a congruent to the phylogeny, but in reality they tend to diverge as taxonomic revisions take longer

Table 1. Comparison of dataset properties. The table indicates number of images and categories, tree depth and whether or not the images are 'in the wild'. Tree depth is calculated as the maximum number of levels in the tree. For example, with iNaturalist this is 7 (corresponding to: kingdom, phylum, class, order, family, genus and species)

Dataset	No. Images	No. Cat.	Tree Depth	Wild?	year
Rove-Tree-11	13,887	215	11	No	2022
ImageNet [6]	14,197,122	21,841	2	yes	2018
IP102 [44]	75,222	102	4	yes	2019
CUB-200-2011 [41]	11,788	200	4 [2]	yes	2011
Cars196 [24]	16,185	196	1	yes	2013
INaturalist 2021 [40]	857,877	5,089	7	yes	2021
PlantCLEF 2021 [13]	330,772	997	3	mixed	2021
DeepFashion [27]	800,000	50	4 ¶	yes	2016

gory. This kind of subjective hierarchy can be useful in other applications, but not particularly for research on generating relationships based on morphology.

The Rove-Tree-11 dataset on the other hand is a well-curated museum collection, where the identification has been done by experts, often using a microscope, and the ground truth phylogeny is as up to date as possible. Additionally, because the images are of museum collections and not 'in-the-wild', the specimen is always fully visible, and the dataset has been curated to include only whole dorsal images. Whether dorsal-view images are sufficient to generate a phylogeny remains to be seen. Typically biologists would use features from all over the body, including ventral and sometimes internal organs. We hypothesize that dorsal view may be sufficient given the ability of deep learning models to learn patterns which are difficult for the human eye to distinguish. Additionally results from our classification experiments shown in table 3 suggest that distinguishing features can be learnt from the images, supporting our belief that phylogenies may be learnt from this dataset.

2.2 Related Methodologies

Classification Classification is one of the most developed fields in computer vision and deep learning, with numerous new state of the art architectures and methods discovered each year. However, there are some architectures which have gained widespread usage in recent years, which we will use to give baselines for this dataset. In particular, we will compare classification results using ResNet [16] and EfficientNet B0 [39]. ResNet is a series of models, introduced in 2015, which uses residual convolution blocks. EfficientNet was introduced in 2019 and is known for achieving high accuracies with few parameters. Classification is not the main focus of this dataset, but we provide classification results for comparison with similar datasets.

¶ hierarchy presented in [5]

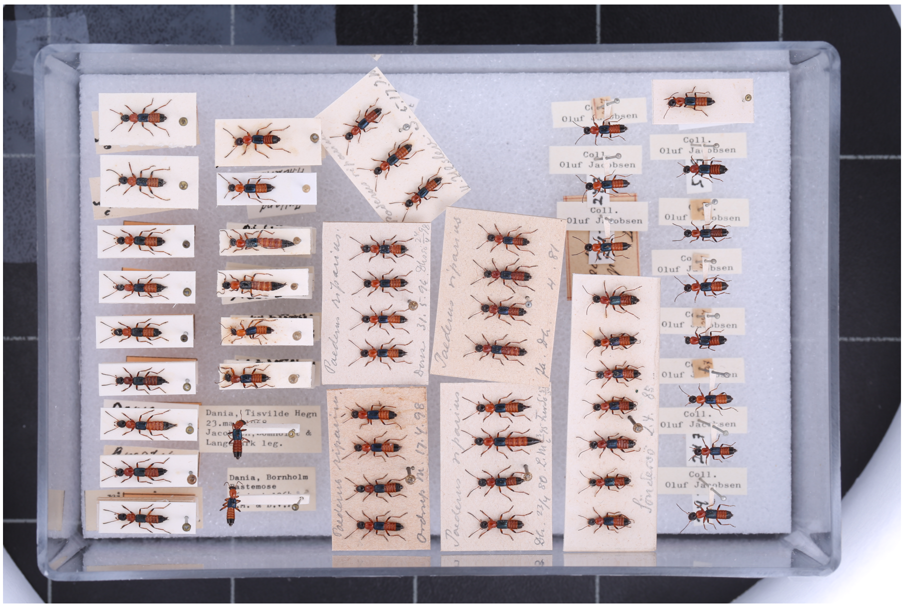


Fig. 2. Example image of museum unit tray from Stage 1 of image processing.

Deep Metric Learning The goal of deep metric learning (DML) is to learn an embedding of the data which represents the dataset and distances between datapoints meaningfully. This could be through clustering related data together, or through creating independence and interpretability in the variables. Recent research into deep metric learning can be split into three groups [38]. **Ranking-based** methods attempt to pull instances from the same class (positive examples) closer together in the embedding space, and typically push examples from other classes further away (eg, [15] [43]). **Classification-based methods**, such as ArcFace[7], work by modifying the discriminative classification task. Finally **Proxy Based** methods, such as Proxy NCA [29] compare each sample with a learned distribution for each class.

In this paper we demonstrate results for this dataset using seven deep metric learning methods; Five ranking-based losses: margin loss [43], triplet loss [43], contrastive loss [15], multisimilarity loss [42], lifted loss [45], one classification-based loss: arcface loss [7] and one proxy-based loss: proxynca [29]. With many state of the art methods and variations on these, choosing which to use is difficult. We chose these firstly because they are all used in [38] as benchmarks, making our results directly comparable. Of the 23 described in [38], we focus on seven which represented some of the better results and show a variety of methods. For a detailed description of each loss we refer the reader to [38].

During training DML models are typically evaluated not just on the loss, but also on a number of clustering metrics. In our case, to do this the dataset is evaluated using nearest neighbors Recall@1 (R1) and Normalized Mutual Information (NMI) after clustering using the k-means algorithm [28]. NMI is presented in our main results, and R1 in the supplemental material. NMI is a symmetric quantity measuring the overlap between clusters. A NMI of 1 indicates that the clusters are the same. Recall@1 is a measure of the % of results with a nearest neighbour in the same class. Both are described in further detail in [38].

Generating a Phylogeny from Embeddings In order to use this dataset for deep phylogenetic generation, we need methods to generate binary graphs from embedding spaces. We could treat this as a classification problem, however, with only one graph to generate, this dataset is not large enough to perform direct graph generation. Instead, the graph can be generated indirectly from the embedding space and compared with the ground truth. This is analogous to how biologists would traditionally generate phylogenetic trees for small datasets using morphological matrices. Biologists use maximum parsimony or bayesian methods [10] to find the best-fitting tree based on discrete characters (either morphological or genetic). However, the use of continuous characters in improving phylogeny generation has been recently explored [35]. Therefore if we assume our embedding space represents morphological features and is a morphological space, this could similarly be used to generate a phylogeny using the same continuous trait bayesian phylogenetic inference methods. We use RevBayes[19], a popular bayesian inference package to complete the analysis. Similar methods have been used to generate phylogenetic trees [23].

Phylogenetic Comparison The main purpose of this dataset is to allow exploration of methods for generating phylogenetic trees based on morphology. To do this, we need methods for comparing phylogenies. There are many standard methods of doing this in biology, a thorough comparison of them is provided in [25]. In brief, the metrics can be split into those which do and do not compare branch lengths. As branch lengths (i.e. evolutionary time) are not yet available in our ground truth phylogeny, we will focus on those which do not include branch lengths, called topology-only comparison methods. The most widely used of these is called the Robinson-Foulds (RF) metric, introduced in 1981 [37]. The RF metric defines the dissimilarity between two trees as the number of operations that would be required to turn one tree into another[*]. However, it has some notable disadvantages, including that apparently similar trees can have a disproportionately high RF score.

One of the more recently introduced metrics is called the Align Score [33]. The Align Score works in two stages. In the first stage, a 1:1 mapping of edges from each tree (T_1 and T_2) is assigned. This is done by calculating a similarity score $s(i, j)$ between the edges, i and j in T_1 and T_2 respectively, based on how similarly they partition the tree. More concretely, in tree T_1, edge i will partition the tree into two disjoint subsets P_{i0} and P_{i1}. The similarity scores can then by

[*] it is, however, different from the edit distance popular in computer science

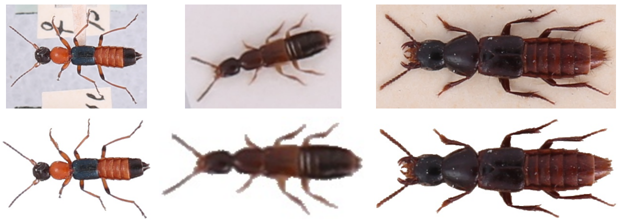


Fig. 3. Examples of specimen images before (above) and after (below) segmentation and rotation adjustment.

computed as:

$$s(i,j) = 1 - \max(\min(a_{00}, a_{11}), min(a_{01}, a_{10})) \tag{1}$$

where a_{rs} is the intersection over the union of the partitions:

$$a_{rs} = |\frac{P_{ir} \cap P_{js}}{P_{ir} \cup P_{js}}| \tag{2}$$

The munkres algorithm is then used to find the edge $j = f(i)$ that minimizes the assignment problem, and then the group with the minimum pairs are summed as follows to calculate the total align score for the two trees:

$$\sum_{i \in T_1} s(i, f(i)) \tag{3}$$

Unlike the RF score, for each set of partitions the align score calculates the similarity, $s(i,j)$, as a continuous variable instead of a binary value. That said, it has the disadvantage that the value is not normalized - a larger tree will likely have a larger align score, making the result difficult to interpret. Despite this, we choose to use it as it is a more accurate representation of the topological similarity between two trees[25].

3 An Overview of Rove-Tree-11

3.1 Image Collection

The images in the dataset were collected and prepared in 4 stages [14]:

Stage 1: Unit Tray Image Collection Rove-Tree-11 was collected by taking overview images of 619 unit trays from the entomology collection at Natural History Museum of Denmark, see fig. 2. A Canon EOS R5 mounted on a camera stand with a macro lens was used to take images of 5760 × 3840 pixels (px) resolution. Since the camera height and focus were kept fixed, the images can be related to physical distance as approx. 400 px per cm. Artificial lighting was used to minimize lighting variance over the images.

Table 2. Species-level classification results on segmented and unsegmented images. We can see that using segmentations drastically reduces the accuracy, indicating that the model is learning from the background and not the morphology of the beetle, as desired. Top-1 and Top-5 represent accuracies. Uncertainties represent 95% confidence intervals.

Model	Dataset	Top-1	Top-5
ResNet-18[16]	segmented	90.9 ± 1.2	99.1 ± 1.2
ResNet-18[16]	unsegmented	99.1 ± 0.3	99.9 ± 0.3

Stage 2: Bounding Box Identification and Sorting After image capture, bounding boxes for the individual specimens were then manually annotated using Inselect [18]. Images of 19,722 individual specimens were then sorted. Only dorsal views (views from the 'back' of the beetle) where the specimen was largely intact and limbs were mostly visible were included, resulting in images of 13,887 specimens in final dataset. See fig. 3 for examples of bounding boxes around specimens. Estimates of body rotation were also annotated in 45 degree increments which allows for coarse correction of the orientation of the crops.

Stage 3: Segmentation Segmentations were then generated through an iterative process. First 200 images were manually segmented. Then U-Net was trained on these 200 images and was used to generate predictions for the rest of the images. 3000 of these segmentations were considered good enough. U-Net was then retrained with these images, then rerun and new segmentations produced. The final segmentations were then manually corrected. Examples of segmentation masks and final segmented specimens can be seen in fig. 3 and fig. 4. The dataset is released with both the original crops and the segmentation masks, however, as we show in table 2, the segmentations are extremely important for phylogenetic analysis, as the background of the image is highly correlated with the species. This is because many of the same species were collected at the same time in the same place by the same person, meaning whether the specimen was glued to a card, the age and color of the card, could be correlated with the species, despite being unrelated to the phylogeny. The segmentations are not perfect. In particular they cut off some of the finer hairs on the body; It could therefore be the case that the segmentations are removing vital information which the model can use to complete classification. We consider this unlikely and suspect the model is instead learning from the backgrounds.

Stage 4: Rotation Adjustment Rotations were corrected by finding the principal axis of inertia of the segmentation masks, (see [21] for details). Since all the beetles are more or less oval shaped, the minimal axis of rotation of their masks tends to line up well with their heads and tails. Using this we further standardized the rotations of the segmentations. This process is shown in fig. 4.

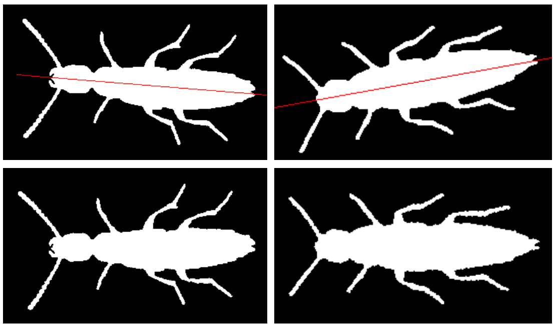


Fig. 4. Illustration of rotation adjustment algorithm. Example original masks (top) and rotated masks (bottom). The red line represents the principal axis of inertia found.

3.2 Preparation of phylogeny

A current genus-level phylogeny of the closely related subfamilies Staphylininae, Xantholininae and Paederinae is provided for the sample of genera used in our analysis. The full phylogeny is visualized in fig. 1 our supplementary material. A subset is shown in fig. 1. This phylogeny represents the current state of knowledge as it was pieced together from the most relevant recently published phylogenetic analyses, such as [47] for sister-group relationships among all three subfamilies and the backbone topology of Xantholininae and Staphylininae, [3] for the subtribe Staphylinina, [4] for the subtribe Philonthina and [46] for the subfamily Paederinae. Below genus-level the phylogeny is considered unresolved as we were unable to find species-level phylogenies for the 215 species included in Rove-Tree-11. A newick file of the phylogeny is provided with the dataset.

3.3 Dataset Statistics

In total, 13,887 images of beetles from the family Staphylinidae, commonly known as rove beetles, are included from 215 species - spanning 44 genera, 9 tribes and 3 subfamilies. Example images are shown in fig. 1.

The distribution of the dataset per genus is shown in fig. 7. A species-level distribution is provided in the supplementary material. From this we can see that the dataset is not evenly distributed, with the species with the highest number of specimens having 261 examples and the lowest having 2 with the genus Philonthus accounting for 24.8% of the dataset. This is due to the number of specimens the museum had in the unit trays that were accessed and imaged at the time, although the curators also includes samples of species which were easily distinguishable from each other, and examples which were hard and can only usually be determined by genital extraction by experts (ie Lathrobium geminum and Lathrobium elongatum. Examples from these two species are shown in fig. 5 to demonstate the difficulty of the task). The distribution of image sizes in the dataset is shown in fig. 6. The majority of the images (82%) are under 500×250 pixels.

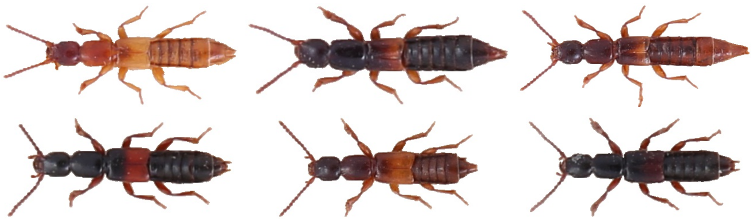


Fig. 5. Example images of Lathrobium geminum (top) and Lathrobium elongatum (bottom) from the dataset. Typically even experts need to dissect the specimen to complete the determination between these two species.

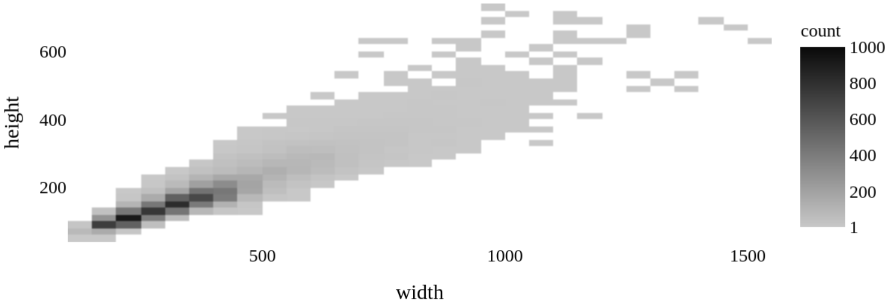


Fig. 6. Distribution of image sizes included in the dataset. The majority (82%) of images are under 500 × 250 pixels.

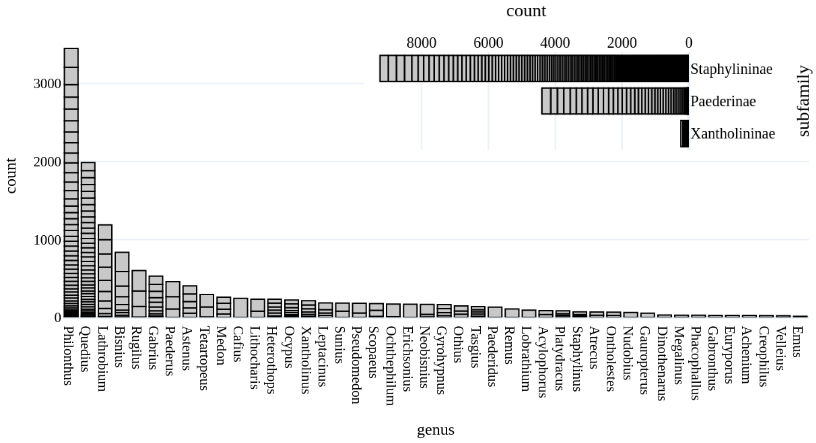


Fig. 7. Distribution of specimens per genus (bottom left) and per subfamily (top right). Each slice in the stacked bar chart represents a different species within that genus. Subfamily distribution is included as it is used to generate the validations and test sets for the clustering results in sec. 4.2. A full species level distribution is shown in the supplemental material.

Table 3. Classification results using deep learning architectures. Top-1 and Top-5 represent accuracies. Uncertainties represent 95% confidence intervals based on 3 runs.

Model	Params	Species		Genus	
		Top-1	Top-5	Top-1	Top-5
ResNet-18 [16]	11.4 M	90.9±1.2	99.2±1.2	98.9±0.3	100±0.3
ResNet-50 [16]	23.9 M	89.4±1.4	99.2±1.4	98.2±0.4	100±0.4
EfficientNet B0 [39]	5.3 M	91.9±1.8	99.3±1.8	99.1±0.2	100±0.2

4 Evaluation

Here we evaluate the dataset by performing benchmark experiments. As stated previously, the main purpose of this dataset is for deep metric learning on hierarchical phylogenetic relationships, so this is also the focus of the benchmarks, although we also provide benchmarks for the classification and clustering tasks. The same augmentations were applied to the dataset as for CUB200 and Cars176 and as in [38], with the exception that the RandomHorizontalFlip was changed to a RandomVerticalFlip, as this makes more sense for the Rove-Tree-11 dataset. Gradient accumulation was also used in some cases due to memory constraints on the available clusters. The details of which experiments this was applied to are provided in the codebase.

4.1 Classification

Results from classification experiments are provided in table 3. For these experiments the official pytorch implementations were used with default parameters: categorical cross entropy loss with an initial learning rate of 0.1, momentum of 0.9, weight decay of $1e-4$ and SGD optimizer. Training details are released with the code for this dataset. The only alterations from the defaults were to reduce the batch-size to 32 due to memory constraints and to alter the data augmentations, detailed in the code. A species-stratified train/val/test split of 70/15/15 was used. The split is provided with the code.

As shown in table 3, the models are able to achieve a top-1 species-level accuracy of 92% with no hyperparameter tuning, and a top-1 genus level of almost 100%. These results suggest that although this dataset could be used for classification tasks and might be useful as such for biologists, classification of this dataset is not particularly difficult, and this dataset is probably not ideal as a benchmark for classification in deep learning.

4.2 Clustering and Phylogenetic Results

In table 4, we present benchmark results of applying state of the art methods for deep metric learning to the Rove-Tree-11 dataset and comparing phylogenies generated using phylogenetic bayesian methods on the embedding space to the ground truth phylogeny as described in sec. 2. A more complete table showing R1 scores and Cars176 results, is provided for reference in the supplementary material (table 1). The 'Random' row represents the align score of a randomly generated tree with the 9 genera leaves included in the test set, against the ground truth tree based on 5 random initializations. Since the align score is not normalized, this random baseline is useful to gauge our results and represents an upper bound our models should achieve. Following best practice, as described in [30], the dataset was split into three groups for training, validation and testing. To properly test the ability of the model to generalize, the groups were split at subfamily level, so the train, validation and test sets should be as phylogenetically distinct as possible, in the sense that they belong to different parts of the phylogenetic tree. This results in 8534 training images from the subfamily Staphylininae, 4399 validation images from the subfamily Paederinae and 954 test images from the subfamily Xantholininae.

All results on Rove-Tree-11 were generated using implementations used in [38], modified to calculate the align score. A forked codebase is provided as a submodule in the github repository.

Based on the clustering results in table 4, we see that Rove-Tree-11 has similar NMI scores to CUB200, suggesting this dataset has a similar clustering difficulty to CUB200 and may be appropriate as a clustering benchmark. As with CUB200, the best models on Rove-Tree-11 are Triplet [43] and Multisimilarity [42]. We can also see that the align score results somewhat correspond with the NMI, with the best results being achieved with Triplet Loss. We can also see that the best test set align score of 4.0 is a marked improvement to the random align score baseline of 6.6, but still significantly far away from a perfect align score of 0, suggesting there is room for improvement. We find it surprising that the align score of the best model on the CUB200 dataset shows a 60% improvement to the random score, while on Rove-Tree-11 the improvement is only 40% on the test set and 51% on the validation set. This suggests that either CUB200 is an easier dataset to generate phylogenies from, or could be an artifact of the align score on trees of different depths (CUB200 has a depth of 4, while Rove-Tree-11 has a depth of 11). It is surprising that it could be an easier dataset, given that the images are in-the-wild, but this could also be due to phylogenetically close birds having similar backgrounds in the images (waterfaring birds might typically have ocean backgrounds, for example, and be more closely phylogenetically related). The phylogenetic tree produced by the best model is provided in the supplementary material along with the ground truth tree for visual inspection.

5 Conclusions

In this paper we present Rove-Tree-11, a novel dataset of segmented images of and research-grade classifications of rove beetles for researching methods for generating phylogenies from images. We provide an eleven-level fine-grained ground

Table 4. Benchmark clustering and Align-Score results on Rove-Tree-11 dataset. 'Random' represents the average align score of 5 randomly generated trees. This gives us a metric to compare our results with. A perfect align score would be 0. 95% confidence errors are provided based on 5 runs.

	CUB200		Rove-Tree-11			
	Test		Validation		Test	
Loss	NMI	Align	NMI	Align	NMI	Align
Random	-	21.9±0.2	-	15.8±0.9	-	6.6±0.5
Triplet	64.8±0.5	9.9±0.9	68.9±0.4	**7.8±1.1**	66.3±0.3	4.1±0.5
Margin	60.7±0.3	10.6±1.2	68.0±0.7	8.2±0.7	65.9±0.5	4.2±0.7
Lifted	34.8±3.0	15.9±2.0	55.0±0.6	10.5±0.7	56.0±1.1	4.9±0.8
Constrast.	59.0±1.0	11.0±1.2	66.7±0.5	8.5±1.0	65.4±0.5	4.5±0.6
Multisim.	**68.2±0.3**	**8.6±0.8**	**70.7±0.2**	8.2±0.4	**67.3±0.5**	**4.0±0.5**
ProxyNCA	66.8±0.4	9.8±0.8	67.5±0.7	9.0±0.8	65.5±0.3	4.2±0.4
Arcface	67.5±0.4	9.8±0.8	66.9±0.9	8.5±0.4	64.8±0.5	4.1±0.4

truth phylogeny for the 44 (train, validation and test) genera included in this dataset.

We start by demonstrating the importance of the provided segmentations as the model can learn from the background. We show benchmark results on this dataset for classification, deep metric learning methods and tree alignment. We further demonstrate that this dataset shows similar clustering results to the CUB200 dataset suggesting it may be appropriate as an alternative clustering benchmark. Finally, we demonstrate how this dataset can be used to generate and compare phylogenies based on the align score, and show that while it is possible to generate such trees, there is plenty of room for improvement and we hope this will be a growing field of research. Code and data are available (code: https://github.com/robertahunt/Rove-Tree-11, data: http://doi.org/10.17894/ucph.39619bba-4569-4415-9f25-d6a0ff64f0e3).

Ethical Concerns Models similar to those described, if applied to images of faces, could be used to generate family trees for humans. This could result in public images being used to infer familial relationships which could have a negative societal impact. The authors strongly discourage this form of misuse of the proposed methods.

Acknowledgements Many people have been involved in this project. First we would like to thank David Gutschenreiter, Søren Bech and André Fastrup who took the photos of the unit trays and completed the initial segmentations of the images as part of their theses. Next, we would like to thank Alexey Solodovnikov of the Natural History Museum of Denmark for providing the specimens, the ground truth phylogeny and guidance for all things entomological. Also, thanks Francois Lauze and the entire Phylorama team for their input the project.

References

1. Bakalar, N.: Nicholas. The New York Times (2014), https://www.nytimes.com/2014/05/27/science/welcoming-the-newly-discovered.html
2. Bameri, F., Pourreza, H.R., Taherinia, A.H., Aliabadian, M., Mortezapour, H.R., Abdilzadeh, R.: TMTCPT: The tree method based on the taxonomic categorization and the phylogenetic tree for fine-grained categorization. Biosystems **195**, 104137 (jul 2020). https://doi.org/10.1016/j.biosystems.2020.104137, https://doi.org/10.1016%2Fj.biosystems.2020.104137
3. Brunke, A., Smetana, A.: A new genus of staphylinina and a review of major lineages (staphylinidae: Staphylininae: Staphylinini). Systematics and Biodiversity **17**, 745–758 (11 2019). https://doi.org/10.1080/14772000.2019.1691082
4. Chani-Posse, M.R., Brunke, A.J., Chatzimanolis, S., Schillhammer, H., Solodovnikov, A.: Phylogeny of the hyper-diverse rove beetle subtribe philonthina with implications for classification of the tribe staphylinini (coleoptera: Staphylinidae). Cladistics **34**(1), 1–40 (2018). https://doi.org/https://doi.org/10.1111/cla.12188
5. Cho, H., Ahn, C., Min Yoo, K., Seol, J., Lee, S.g.: Leveraging class hierarchy in fashion classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops (Oct 2019)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Deng, J., Guo, J., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. CoRR **abs/1801.07698** (2018), http://arxiv.org/abs/1801.07698
8. DiSSCo: Distributed system of scientific collections. https://www.dissco.eu/ (July 2022)
9. Felsenstein, J.: Evolutionary trees from dna sequences: A maximum likelihood approach. Journal of Molecular Evolution **17**, 368–376 (1981)
10. Felsenstein, J.: Statistical inference of phylogenies. Journal of the Royal Statistical Society: Series A (General) **146**(3), 246–262 (May 1983). https://doi.org/https://doi.org/10.2307/2981654
11. Felsenstein, J.: Inferring phylogenies. Sinauer associates, Sunderland, MA (2003)
12. Fink, M., Ullman, S.: From aardvark to zorro: A benchmark for mammal image classification. Int. J. Comput. Vision **77**(1–3), 143–156 (may 2008). https://doi.org/10.1007/s11263-007-0066-8, https://doi.org/10.1007/s11263-007-0066-8
13. Goëau, H., Bonnet, P., Joly, A.: Overview of plantclef 2021: cross-domain plant identification. In: Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum. vol. 2936, pp. 1422–1436 (2021)
14. Gutschenreiter, D., Bech, S.: Deep-learning methods on taxonomic beetle data Automated segmentation and classification of beetles on genus and species level. Master's thesis, University of Copenhagen (2021)
15. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1735–1742 (2006). https://doi.org/10.1109/CVPR.2006.100
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)

17. Hedrick, B.P., Heberling, J.M., Meineke, E.K., Turner, K.G., Gassa, C.J., Park, D.S., Kennedy, J., Clarke, J.A., Cook, J.A., Blackburn, D.C., Edwards, S.V., Davis, C.C.: Digitization and the future of natural history collections. BioScience **70**(3), 243–251 (February 2020). https://doi.org/https://doi.org/10.1093/biosci/biz163
18. Hudson, L.N., Blagoderov, V., Heaton, A., Holtzhausen, P., Livermore, L., Price, B.W., van der Walt, S., Smith, V.S.: Inselect: Automating the digitization of natural history collections. PLOS ONE **10**(11), 1–15 (11 2015). https://doi.org/10.1371/journal.pone.0143402, https://doi.org/10.1371/journal.pone.0143402
19. Höhna, L., Heath, B., Lartillot, M., Huelsenbeck, R.: Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Systematic Biology **65**, 726–736 (2016)
20. iDigBio: Integrated digitized biocollections. https://www.idigbio.org/ (July 2022)
21. J. Peraire, S.W.: Lecture notes from mit course 16.07 dynamics, fall 2008. l26 - 3d rigid body dynamics: The inertia tensor (2008), https://ocw.mit.edu/courses/16-07-dynamics-fall-2009/dd277ec654440f4c2b5b07d6c286c3fd_MIT16_07F09_Lec26.pdf
22. KAYA, M., BİLGE, H.S.: Deep metric learning: A survey. Symmetry **11**(9) (2019). https://doi.org/10.3390/sym11091066, https://www.mdpi.com/2073-8994/11/9/1066
23. Kiel, S.: Assessing bivalve phylogeny using deep learning and computer vision approaches. bioRxiv (2021). https://doi.org/10.1101/2021.04.08.438943, https://www.biorxiv.org/content/early/2021/04/09/2021.04.08.438943
24. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13). Sydney, Australia (2013)
25. Kuhner, M.K., Yamato, J.: Practical Performance of Tree Comparison Metrics. Systematic Biology **64**(2), 205–214 (12 2014). https://doi.org/10.1093/sysbio/syu085
26. Lee, M., Palci, A.: Morphological phylogenetics in the genomic age. Current Biology **25**(19), R922–R929 (2015). https://doi.org/https://doi.org/10.1016/j.cub.2015.07.009, https://www.sciencedirect.com/science/article/pii/S096098221500812X
27. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
28. MacQueen, J.: Classification and analysis of multivariate observations. In: 5th Berkeley Symp. Math. Statist. Probability. pp. 281–297 (1967)
29. Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 360–368 (2017). https://doi.org/10.1109/ICCV.2017.47
30. Musgrave, K., Belongie, S., Lim, S.N.: A metric learning reality check. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 681–699. Springer International Publishing, Cham (2020)
31. Natural History Museum of Denmark: Digital nature: Giant grant makes the natural history collections of denmark accessible to everyone. Newsletter (2021)
32. Natural History Museum of Denmark: Entomology - Dry and Wet Collections. Homepage (2022)

33. Nye, T., Lio, P., Gilks, W.: A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics (Oxford, England) **22**, 117–9 (02 2006). https://doi.org/10.1093/bioinformatics/bti720
34. Orlov, I., Leschen, R.A., Żyła, D., Solodovnikov, A.: Total-evidence backbone phylogeny of aleocharinae (coleoptera: Staphylinidae). Cladistics **37**(4), 343–374 (2021). https://doi.org/https://doi.org/10.1111/cla.12444
35. Parins-Fukuchi, C.: Use of Continuous Traits Can Improve Morphological Phylogenetics. Systematic Biology **67**(2), 328–339 (09 2017). https://doi.org/10.1093/sysbio/syx072, https://doi.org/10.1093/sysbio/syx072
36. Popov, D., Roychoudhury, P., Hardy, H., Livermore, L., Norris, K.: The value of digitising natural history collections. Research Ideas and Outcomes **7**, e78844 (December 2021). https://doi.org/https://doi.org/10.3897/rio.7.e78844
37. Robinson, D., Foulds, L.: Comparison of phylogenetic trees. Mathematical Biosciences **53**(1), 131–147 (1981). https://doi.org/https://doi.org/10.1016/0025-5564(81)90043-2, https://www.sciencedirect.com/science/article/pii/0025556481900432
38. Roth, K., Milbich, T., Sinha, S., Gupta, P., Ommer, B., Cohen, J.P.: Revisiting training strategies and generalization performance in deep metric learning (2020)
39. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 6105–6114. PMLR (09–15 Jun 2019), https://proceedings.mlr.press/v97/tan19a.html
40. Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., Mac Aodha, O.: Benchmarking representation learning for natural world image collections. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12884–12893 (2021)
41. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
42. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5017–5025 (2019). https://doi.org/10.1109/CVPR.2019.00516
43. Wu, C.Y., Manmatha, R., Smola, A.J., Krähenbühl, P.: Sampling matters in deep embedding learning. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2859–2867 (2017). https://doi.org/10.1109/ICCV.2017.309
44. Wu, X., Zhan, C., Lai, Y., Cheng, M.M., Yang, J.: Ip102: A large-scale benchmark dataset for insect pest recognition. In: IEEE CVPR. pp. 8787–8796 (2019)
45. Yuan, Y., Chen, W., Yang, Y., Wang, Z.: In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation (2019)
46. Żyła, D., Bogri, A., Hansen, A., Jenkins Shaw, J., Kypke, J., Solodovnikov, A.: A new termitophilous genus of paederinae rove beetles (coleoptera, staphylinidae) from the neotropics and its phylogenetic position. Neotropical Entomology (02 2022). https://doi.org/10.1007/s13744-022-00946-x
47. Żyła, D., Solodovnikov, A.: Multilocus phylogeny defines a new classification of staphylininae (coleoptera, staphylinidae), a rove beetle group with high lineage diversity. Systematic Entomology **45**(1), 114–127 (2020). https://doi.org/https://doi.org/10.1111/syen.12382

Rove-Tree-11: Supplemental Material

Roberta Hunt[1]https://orcid.org/0000-0003-1963-3281 and Kim Steenstrup Pedersen[1,2]https://orcid.org/0000-0003-3713-0960

[1] Department of Computer Science,
University of Copenhagen, Universitetsparken 1, 2100, Copenhagen, Denmark
{r.hunt,kimstp}@di.ku.dk
[2] Natural History Museum of Denmark, Øster Voldgade 5 - 7, 1350, Copenhagen, Denmark kimstp@snm.ku.dk

1 Rove-Tree-11: Full Phylogenetic Trees

In fig. 1 we show the gold standard genus-level phylogeny for the train, validation and test sets. This phylogeny represents the current state of knowledge as it was pieced together from the most relevant recently published phylogenetic analyses, (from the main text): "such as [1] for sister-group relationships among all three subfamilies and the backbone topology of Xantholininae and Staphylininae, [2] for the subtribe Staphylinina, [3] for the subtribe Philonthina and [4] for the subfamily Paederinae. In this work we did not complete species-level phylogeny, so each species within the genera is assumed to be equally related."

In figs. 2 and 4 we show the species-level phylogenies produced by our model which performed best on the test set (multisimilarity, seed 2) on the validation set, and the test set, respectively. The best model gave an Align Score of 3.5 when compared against the gold standard (4.1 on average over 5 runs). The align score for each individual node is shown to provide some insight into the align score.

From figs. 2 and 4 we can easily see that most species and specimens in the dataset are mostly grouped close to those of the same genus, despite not telling the model these species are related. Which is encouraging. And using the align score as an indicator, we can see that many groups are well organized, however, there is still plenty of room for improvement.

With this visual comparison we can conclude that the model is learning some interesting phylogenetic features, but there is significant room for improvement, making this an interesting dataset for further research.

2 Species-level data distribution

We also show a species-level data distribution in fig. 6. From this we can see that the data is not uniformly distributed per species, with the largest group, rugilus orbiculatus having 262 specimens and the smallest, lathrobium castaneipenne, having 4. This of course may negatively affect our results and a uniform distribution would be preferred.

3 Expanded Results

Further results from experiments are shown in table 1 and expands on the results presented in the paper. Showing R1 scores and results on the Cars196 dataset for easy reference. It should be noted that the Cars196 data was taken directly from [5].

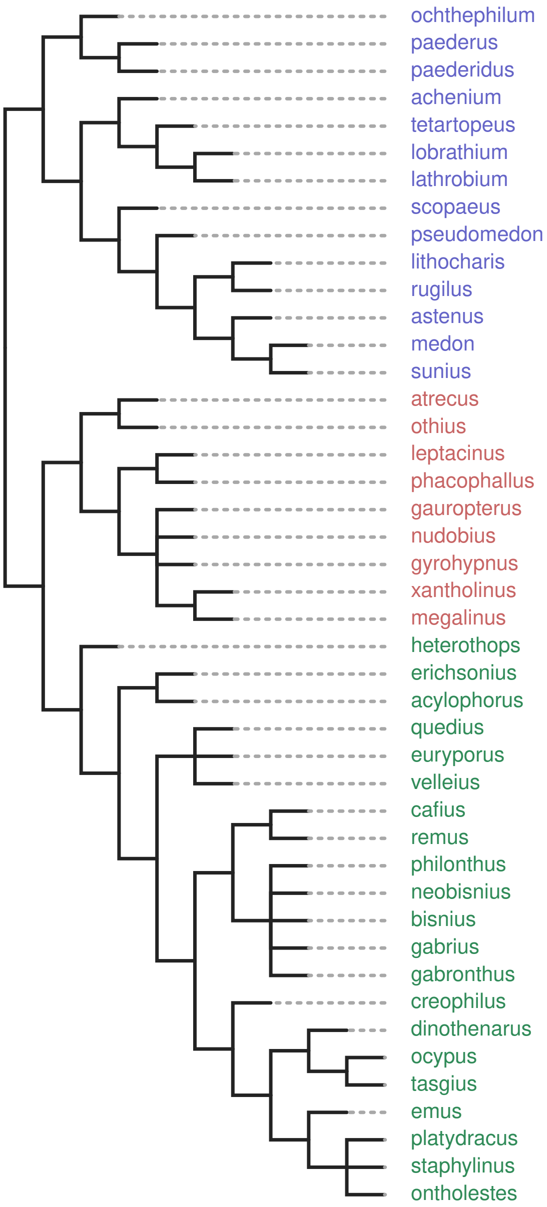


Fig. 1. Gold standard genera-level phylogenetic tree. Tree is split into training set (green), validation set (blue) and test set (red). Genera-level is used here instead of species level to make the tree more compact.

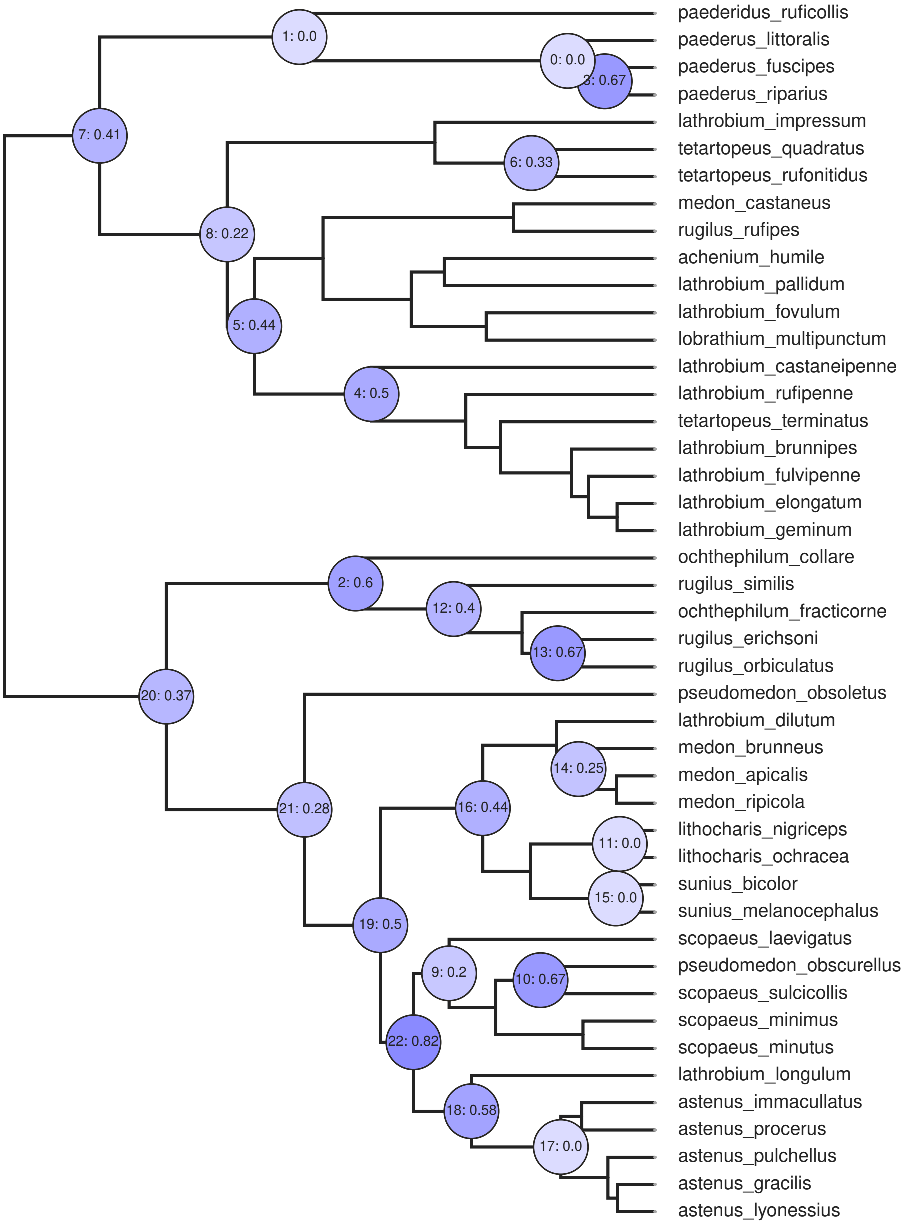


Fig. 2. Species-level phylogenetic tree produced by our best model for the validation set. The number on each node represents the align score of that node to it's matched node in the ground truth phylogeny (the format for this is [node number]:[align score]). Node numbers for the ground truth phylogeny on the validation set are provided in fig. 3. The shade of the node corresponds to the value of the align score - darker shades have higher (undesirable) align scores. The total score for this tree (from summing the score of each node) is 8.35.

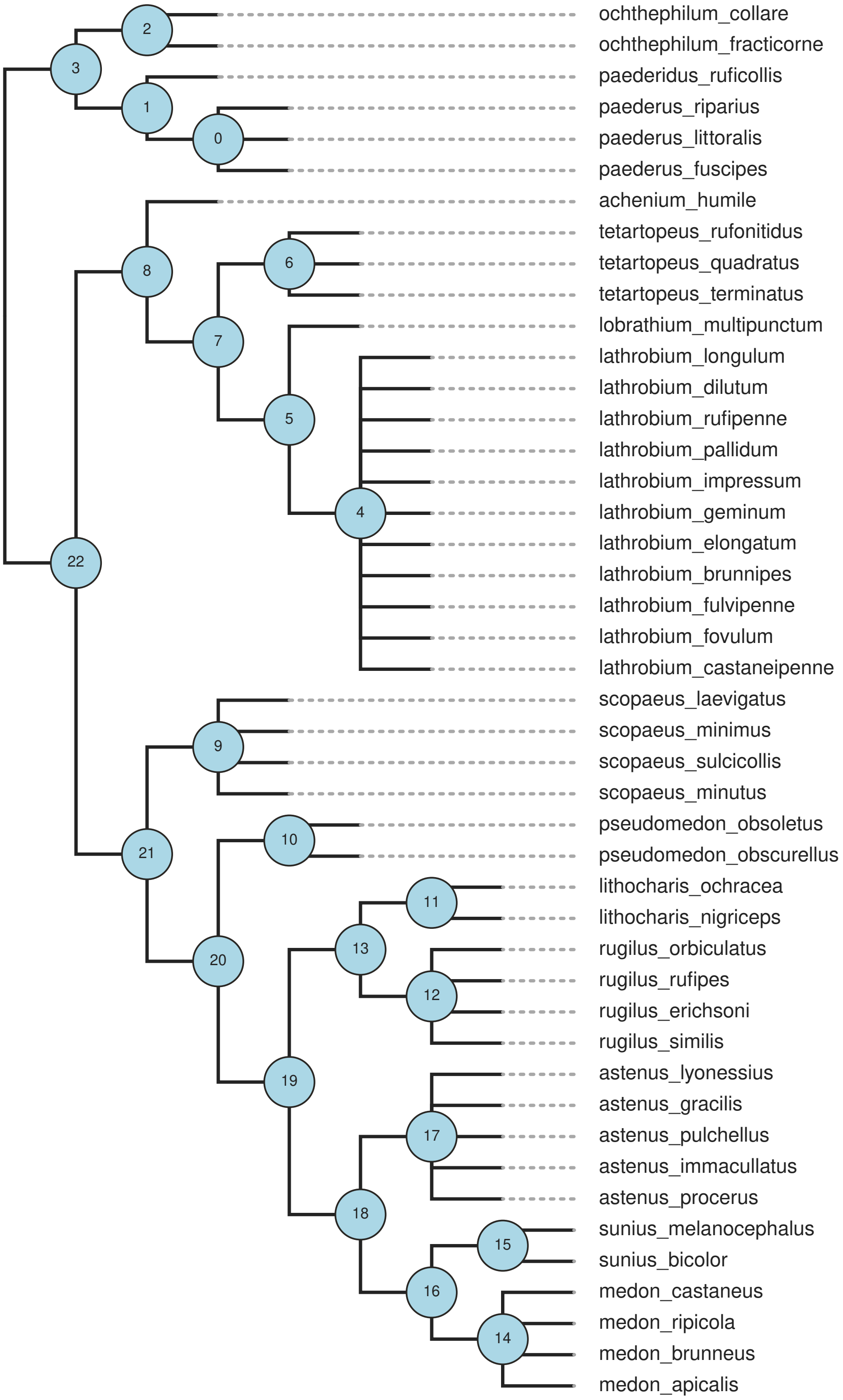


Fig. 3. Species-level ground truth phylogenetic tree for the validation set. The number on each node represents the number of that node, to link it to the align scores presented in fig. 2, which are somewhat synonymous to the most similar nodes in the other tree.

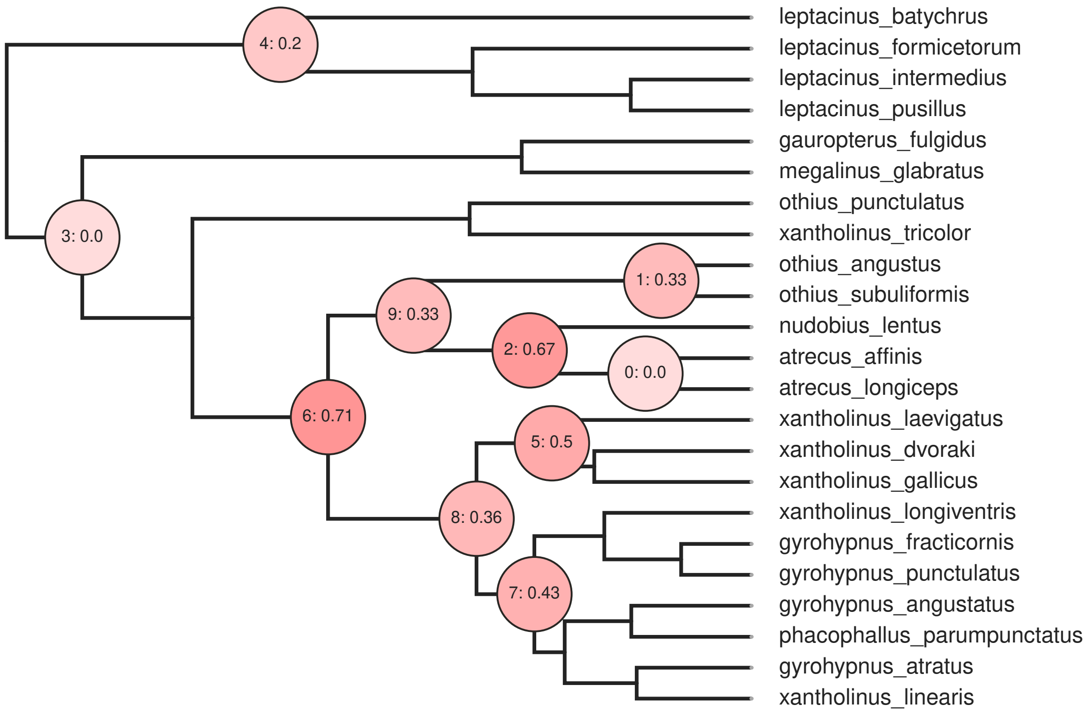


Fig. 4. Species-level phylogenetic tree produced by our best model for the test set. he number on each node represents the align score of that node to it's matched node in the ground truth phylogeny (the format for this is [node number]:[align score]). Node numbers for the ground truth phylogeny on the validation set are provided in fig. 5. The shade of the node corresponds to the value of the align score - darker shades have higher (undesirable) align scores. The total score for this tree (from summing the score of each node) is 8.35.

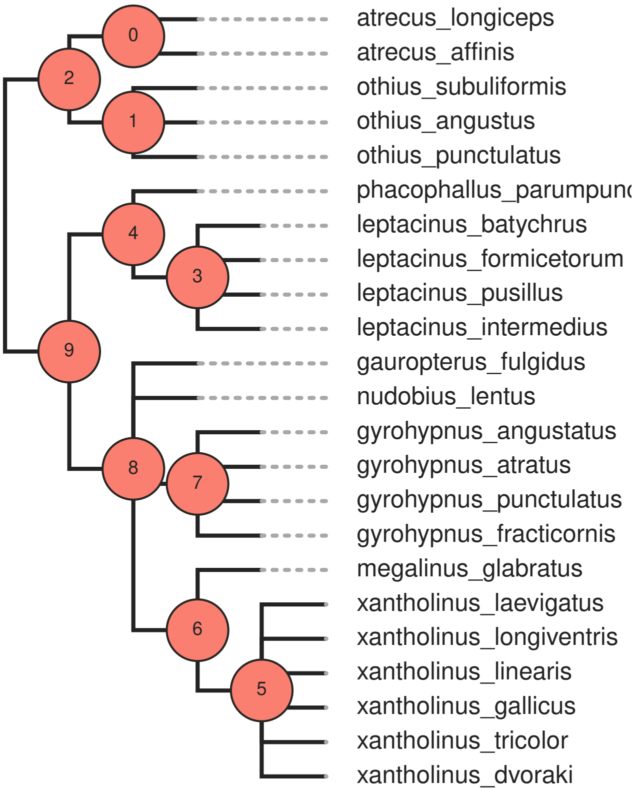


Fig. 5. Species-level ground truth phylogenetic tree for the test set. The number on each node represents the number of that node, to link it to the align scores presented in fig. 4, which are somewhat synonymous to the most similar nodes in the other tree.

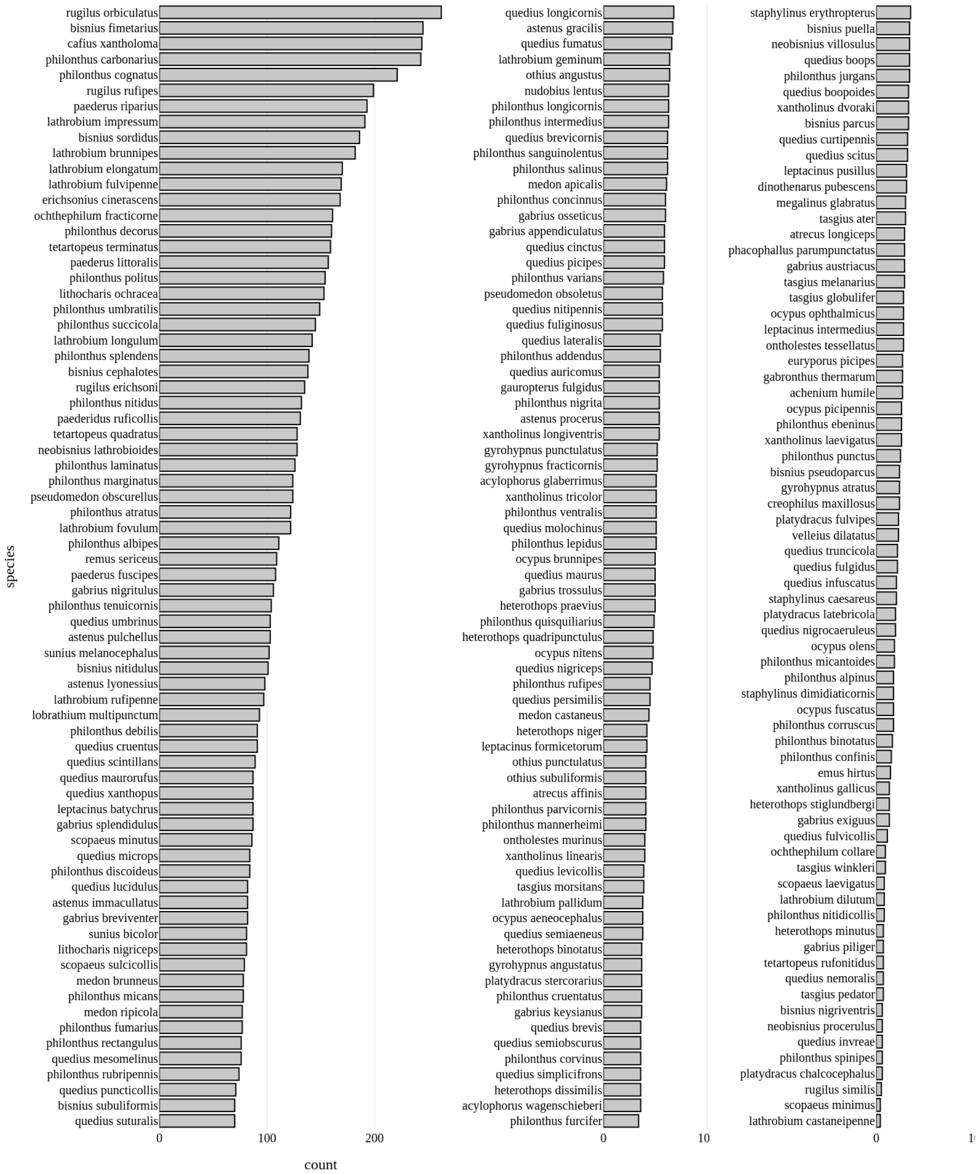


Fig. 6. Species-level data distribution. Here we can more clearly see that the images per species are not uniformly distributed, with Rugilus orbiculatus having 262 specimens and Lathrobium castaneipenne having 4.

Table 1. Benchmark clustering and Align-Score results on Rove-Tree-11 dataset. 'Random' represents the average align score of 5 randomly generated trees. This gives us a metric to compare our results with. A perfect align score would be 0. 95% confidence errors are provided based on 5 runs.

	Cars196		CUB200			Rove-Tree-11					
	Test		Test			Validation			Test		
Loss	NMI	R1	NMI	R1	Align	NMI	R1	Align	NMI	R1	Align
Random	-	-	-	-	21.9±0.2	-	-	15.8±0.9	-	-	6.6±0.5
Triplet	65.9±0.2	79.1±0.3	64.8±0.5	59.4±0.4	9.9±0.9	68.9±0.4	86.3±0.5	**7.8±1.1**	66.3±0.3	84.0±0.6	4.1±0.5
Margin	65.3±0.3	77.7±0.3	60.7±0.3	54.7±0.2	10.6±1.2	68.0±0.7	84.4±0.6	8.2±0.7	65.9±0.5	82.8±0.8	4.2±0.7
Lifted	72.2±0.4	63.8±0.4	34.8±3.0	23.5±5.2	15.9±2.0	55.0±0.6	63.3±1.2	10.5±0.7	56.0±1.1	58.8±0.8	4.9±0.8
Constrast.	64.0±0.1	75.8±0.4	59.0±1.0	53.6±0.8	11.0±1.2	66.7±0.5	82.3±0.6	8.5±1.0	65.4±0.5	80.8±1.4	4.5±0.6
Multisim.	**81.7±0.2**	69.4±0.4	**68.2±0.3**	62.9±0.6	**8.6±0.8**	**70.7±0.2**	**88.0±0.3**	8.2±0.4	**67.3±0.5**	**86.8±0.7**	**4.0±0.5**
ProxyNCA	78.5±0.6	65.8±0.2	66.8±0.4	**63.0±0.4**	9.8±0.8	67.5±0.7	83.7±0.8	9.0±0.8	65.5±0.3	82.1±0.9	4.2±0.4
Arcface	67.0±1.1	**79.2±1.0**	67.5±0.4	**63.0±0.7**	9.8±0.8	66.9±0.9	82.8±0.9	8.5±0.4	64.8±0.5	79.5±1.0	4.1±0.4

References

1. Żyła, D., Solodovnikov, A.: Multilocus phylogeny defines a new classification of staphylininae (coleoptera, staphylinidae), a rove beetle group with high lineage diversity. Systematic Entomology **45** (2020) 114–127
2. Brunke, A., Smetana, A.: A new genus of staphylinina and a review of major lineages (staphylinidae: Staphylininae: Staphylinini). Systematics and Biodiversity **17** (2019) 745–758
3. Chani-Posse, M.R., Brunke, A.J., Chatzimanolis, S., Schillhammer, H., Solodovnikov, A.: Phylogeny of the hyper-diverse rove beetle subtribe philonthina with implications for classification of the tribe staphylinini (coleoptera: Staphylinidae). Cladistics **34** (2018) 1–40
4. Żyła, D., Bogri, A., Hansen, A., Jenkins Shaw, J., Kypke, J., Solodovnikov, A.: A new termitophilous genus of paederinae rove beetles (coleoptera, staphylinidae) from the neotropics and its phylogenetic position. Neotropical Entomology (2022)
5. Roth, K., Milbich, T., Sinha, S., Gupta, P., Ommer, B., Cohen, J.P.: Revisiting training strategies and generalization performance in deep metric learning (2020)

4 Gattaca: Adding Genetic Data

> *Everything is gaussian if you just add enough numbers together*
>
> — **Kim Steenstrup Pedersen**

This chapter contains the following paper that continues the work from the previous chapter:

> **Roberta Hunt, José L. Reyes-Hernández, Josh Jenkins Shaw, Alexey Solodovnikov, Kim Steenstrup Pedersen**. „Integrating Deep Learnt Morphological Traits and Molecular Data for Total-Evidence Phylogenetics: Lessons from Digitized Collections“ Revised edition, resubmitted to: Systematic Biology, 2024, Under Review after major revision.

The following dataset is included in this thesis:

> **Roberta Hunt, José L. Reyes-Hernández, Josh Jenkins Shaw, Alexey Solodovnikov and Kim Steenstrup Pedersen**; (2024) 'Rove-Tree-11 Genes'. Available at: `https://erda.ku.dk/archives/78bc5d7c5746eca509bbd8fb2ea68205/published-archive.html` (Accessed June 2024).

All supplemental material for this preprint is available in a frozen archive in case any changes are made during the peer review process: `https://erda.ku.dk/archives/78bc5d7c5746eca509bbd8fb2ea68205/published-archive.html`. Example code and trained models are available in the supplemental material. Deep learning code is available from the Rove-Tree-11 code repository [37].

After completing the Rove-Tree-11 dataset, we wanted to explore how the deep learnt morphological traits might compare to and be combined with genetic traits and inference methods. We also wanted to present our ideas to systematic biologists to get their input and feedback.

Therefore we teamed up with the Coleoptera group at the Natural History Museum of Denmark, who gathered the genetic data and completed the sequence alignments, as well as advised on the inference process.

Unfortunately, but not unexpectedly, it was not possible to obtain genetic data for most of the individual species in the Rove-Tree-11 dataset. Therefore the analyses for this paper was completed at genus level instead of species level.

This paper is currently under review after resubmission.

Systematic Biology (2024), **0**, 0, *pp.* 1–39
doi:10.1093/sysbio/output

Integrating Deep Learnt Morphological Traits and Molecular Data for Total-Evidence Phylogenetics: Lessons from Digitized Collections

Roberta Hunt[1,*], José L. Reyes-Hernández[2], Josh Jenkins Shaw[2], Alexey Solodovnikov[2] and Kim Steenstrup Pedersen,[1,2]

[1] *Department of Computer Science, University of Copenhagen, Copenhagen, DK-2100, Denmark*
[2] *Natural History Museum of Denmark, Copenhagen, DK-2100, Denmark*

**r.hunt@di.ku.dk*

Abstract

Deep learning has previously shown success in learning morphological traits which carry a phylogenetic signal. In this paper we explore combining deep learnt morphological traits from digitized natural history collections with molecular data for total-evidence phylogenies and we reveal challenges. Deep-learnt morphological traits, while informative, underperform when used in isolation compared to molecular analyses. However, their use in total evidence analyses shows some promise. We explore these results and propose avenues for methodological enhancement.

Key words: Phylogenetics, deep representation learning, continuous morphological matrices, cladistics, total evidence analysis, image analysis, neural networks

Phylogenetics or reconstructing the Tree of Life at the fine detail is at the core of modern biology (Soltis and Soltis, 2003; Cavender-Bares et al., 2009; Yang and Rannala, 2012; Hassler et al., 2023). Although in recent decades phylogenetic inference has been a high throughput endeavor due to the availability of molecular data (Yang and Rannala, 2012), for more than two centuries it was a practice based on the visual study of the anatomical details of the organisms (Giribet, 2015). Despite the recent emphasis on molecular data in the phylogenetic analysis, morphology remains crucial (Giribet, 2015; Lee and Palci, 2015; de Almeida et al., 2023), particularly for including fossils and dating evolutionary events through a total evidence approach (Ronquist et al., 2012; Zhang et al., 2015; Pyron, 2015). Contrary to molecular data, the high throughput use of the morphological traits in statistical phylogenetics is strongly hindered by the complex nature of morphology and as a consequence lesser developed tools for its analysis (Tarasov, 2019). It is therefore noteworthy that recent research has shown that morphological traits automatically extracted from habitus images of the organisms using deep learning-based image analysis carry a phylogenetic signal (Furusawa et al., 2023; Cuthill et al., 2019; Kiel, 2021; Hunt and Pedersen, 2022). However, because the deep learnt traits are not sliced into homologous and non-homologous traits (characters), this signal is not amplified as such and could be blurred with non-phylogenetic clustering.

Manual, expert-based crafting of the morphological data matrix for phylogenetic analysis is so tedious and time consuming exactly because of the task of the formulating these traits (characters) and their states. This task is achieved by comparison of multiple species with various morphologies and, using the extensive knowledge of an expert about anatomy and its diversity of a group under investigation, only homologous character states are aligned for analysis among species. As shown in Cuthill et al. (2019), euclidean phenotypic distances calculated using a deep convolutional triplet network, captured the wing phylogeny pattern reflecting Müllerian mimicry and thus convergence between the interspecies co-mimics in the *Heliconius erato* and *H. melpomene* butterfly complex. Would they be able to capture other, less obvious (or less visual) pattern reflecting synapomorphic similarity and thus an overall species (not their single traits or genes) phylogeny? Lacking of the critical filter in the automatic capture of the phenotypic similarity in butterflies in this example and other organisms is a remaining problem to solve. Kiel (2021) well recognized the issue of convergent evolution that is “invisible” for the CNN he trained for assessing bivalve phylogeny from thousands of images of 75 bivalve families. Therefore, he trained two further CNNs on the same images but grouped by the orders and subclasses these families belonged to. He demonstrated that this improved the inferred phylogenetic relationships even for families, which these extra CNNs were not trained for. Furusawa et al. (2023) presented the morphologically regulated variational AutoEncoder (Morpho-VAE) for analyzing shapes of primate mandibles to overcome the difficulty of encoding shapes usually tediously done by comparing anatomically prominent landmarks. Their goals were not phylogenetic but they showed that their automatically extracted morphological features reflected the families to which the organisms belong, i.e. they were phylogenetically meaningful. These experiments show that further exploration of how to overcome the convergency problem in the automatic capture of phenotypic data for phylogenetic analysis are promising and needed.

Moreover, even in the expert-based data capture, apart from the prohibitively high time and expensive expertise consumption of this task, ultimately, it is an overall phylogeny itself that finally informs us which particular similarity is homology (synapomorphy)-based and which is homoplastic. Thus, to avoid subjectivity and increase efficiency of the phylogenetic analysis of the phenotypic data, developing automatic and rigorous methods of data capture is the way to go. If these methods can be improved and reconciled with the mass digitization initiatives currently ongoing within the major natural history collections and resulting in millions of specimen images being easily available (Baird, 2010; Smith and Blagoderov, 2012; Wilson et al., 2023), this could revolutionize phylogenetics by making phylogenetic research scalable and accessible. Therefore, it is of great interest to determine how this signal could be strengthened and how it might be integrated into existing methodologies. In this paper we explore this using a dataset of the habitus (here interpreted as the dorsal view of the entire beetle body) images of the rove beetles obtained from museum collections. We aim to determine how strong this signal is, and how the deep learnt continuous morphological traits that are made available from the images of the digitized specimens might be used for phylogenetic inference alone or in combination with molecular data. Before proceeding to our methods and results, we first introduce several important themes and considerations that determined our study design.

Is habitus enough? — Since Linnaeus (Linné et al., 1788), if not before, hierarchical grouping has been at the core of understanding the diversity of life on Earth with the evolutionary theory elevating the importance of constructing genealogical relationships to decipher the connections among both extant and extinct species. The 'Tree of life' became a universal metaphor depicting relationships between living and extinct organisms (Mindell, 2013; Hossfeld and Levit, 2016). Historically, in the absence of sophisticated optical and anatomical instruments, the habitus served as the primary means for comparing organismal phenotypes. However, evolutionary biology's advancements have highlighted the complexity of phenotypes, underscoring that the habitus represents merely a subset of an organism's morphological information. The maturing concepts of homology guided the division of a holistic phenotype into single morphological traits and their assessment as shared ancestry or, on the contrary, cases of convergence, a division crucial in phylogenetics (Wagner, 1989). In modern statistical phylogenetics, intuitive or even statistical phenetic assessment of the overall morphological similarity among organisms (Sokal and Sneath, 1963) gave way to the analysis of morphology partitioned into homological characters (Goloboff, 2022). A concept of a morphological trait and its homological states in phylogenetics is rather complex (Wiens, 2001) and raises questions about data automatically obtained from the specimen images. The suitability of utilizing images capturing the external morphology, or habitus, of organisms for phylogenetic analysis warrants initial examination.

The role of continuous traits in phylogenetics. — The first stochastic process model of the evolution of continuous trait data on a phylogeny was Brownian motion, proposed early in the history of statistical phylogenetics (Edwards, 1970; Felsenstein, 1985). Among the great diversity of morphological characters used in phylogenetics, continuous characters of shape are rare due to lack of tools for their proper coding and assessment. The majority of homological morphological traits used in the phylogenetic analysis are the so-called discrete characters (MacLeod and Forey, 2002). However, various studies suggest that continuous characters are very informative and even may be preferable over discrete characters (Parins-Fukuchi, 2017).

The growing role of computer vision in phylogenetics. — Advancements in computer vision and machine learning are enabling the collection of morphological data for phylogenetic analysis at a scale similar to phylogenomics. With digital images, algorithms can quickly convert these into numerical vectors that mimic DNA sequences, providing a new form of data for analysis. Numerical vectors automatically generated from habitus photos return us to an early trend of the morphology-based phylogenetics called phenetics which was abandoned early in the history of statistical phylogenetics for several important reasons (Jensen, 2009). However, phenetics played an important role as an early stepping stone towards modern phylogenetics and perhaps it was abandoned due to the lack of good tools to generate informative data at the time. Also, the idea that the overall similarity between taxa may have a phylogenetic signal if properly assessed, was never rejected and does have support, especially when dealing with lower taxonomic categories (Jensen, 2009).

The role of deep learning. — Deep learning, a subfield of machine learning, has proven to be an effective method to extract traits across various fields, from natural language processing (Otter et al., 2021) to image processing (Jiao and Zhao, 2019). In the field of entomology, deep learning has proven quite successful at classification and quanitification of mimicry (Valan et al., 2019; Kelly et al., 2021; Høye et al., 2021; MacLeod et al., 2022; Li Fan and Cui, 2022; Pichler and Hartig, 2023), and has already been applied to extracting morphological traits from images of various animals for phylogenetics (Cuthill et al., 2019; Kiel, 2021; Hunt and Pedersen, 2022; Furusawa et al., 2023). As the field of deep learning has grown over the last decade, there are thousands of model architectures, loss functions, and parameters to choose from, and testing them all is not feasible. Instead, this paper focuses on how simpler methods of boosting the phylogenetic signal might be completed. Therefore, for simplicity, we use only one well-known architecture, ResNet50 (He et al., 2016), and a few selected loss functions commonly used in the field of deep metric learning (Roth et al., 2020).

Deep metric learning, an advanced subfield of machine learning, focuses on understanding and quantifying the similarity or dissimilarity between data points in a high-dimensional space. By leveraging deep neural networks, this approach aims to learn a distance function that effectively maps data points such that similar items are brought closer together, while dissimilar items are pushed apart, in the embedding space. This methodology is particularly pertinent to phylogenetic trait extraction over traditional classification models due to its ability to capture nuanced relationships and continuous variations among biological species. Unlike classification models, which categorize data into discrete classes and often overlook the intricate relationships between them, deep metric learning accommodates the complexity and continuity of evolutionary traits. This makes it exceptionally suitable for phylogenetics, where the objective is to unravel the evolutionary distances and ancestral linkages among species. By focusing on relative distances rather than absolute categorizations, deep metric learning facilitates a more nuanced understanding of phylogenetic traits, enabling researchers to uncover subtle evolutionary patterns that classification models might miss. Loss functions utilized in deep metric learning can be categorized into three distinct types: ranking-based, classification-based, and proxy-based approaches. Ranking-based approaches modulate the latent space by endeavoring to minimize the distance between analogous images and maximize the separation between dissimilar images within this space. Classification-based approaches presuppose that executing a classification task will inherently organize the latent space into clusters. Conversely, proxy-based approaches create a distribution for each class and

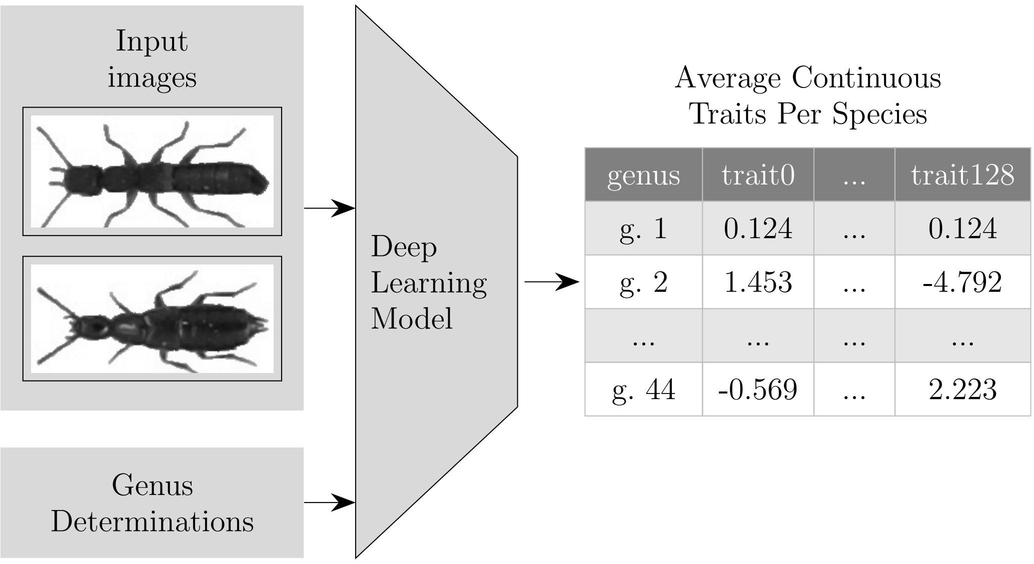
evaluate each data point relative to these distributions. To ensure a comprehensive understanding, our investigation encompasses loss functions from each of these categories. It is imperative to acknowledge that deep learning methodologies have been extensively applied to the elucidation of phylogenetic relationships using molecular traits. A detailed exposition of these applications is provided in Mo et al. (2024).

Methods of Phylogenetic Inference. — Studies such as the present one would ideally compare different statistical methods or approaches to phylogenetic reconstruction. Each phylogenetic method has different strengths and weaknesses and is highly dependent on the nature of the data, philosophical viewpoint of the practitioner and available computational resources (Yang and Rannala, 2012). Despite Bayesian Inference is a widely used method for inferring phylogenetic relationships based on molecular (Huelsenbeck et al., 2001) and morphological (Wright, 2019) data, we experienced some constraints in analyzing continuous trait data within a Bayesian framework. As far as we know, of all the Bayesian phylogenetic Inference software only RevBayes (Höhna et al., 2016) currently supports using continuous trait data, however it does not yet support missing values/taxa in said data (Zhang, 2022). Therefore we present our results using only Maximum Parsimony (Fitch, 1971).

Our contributions. — In this paper we show how deep learning can be used to extract continuous morphological traits that carry a phylogenetic signal, and how these traits can be used independently or combined with molecular data in a total-evidence framework. We compare and evaluate the results of these analyses across different methodological choices.

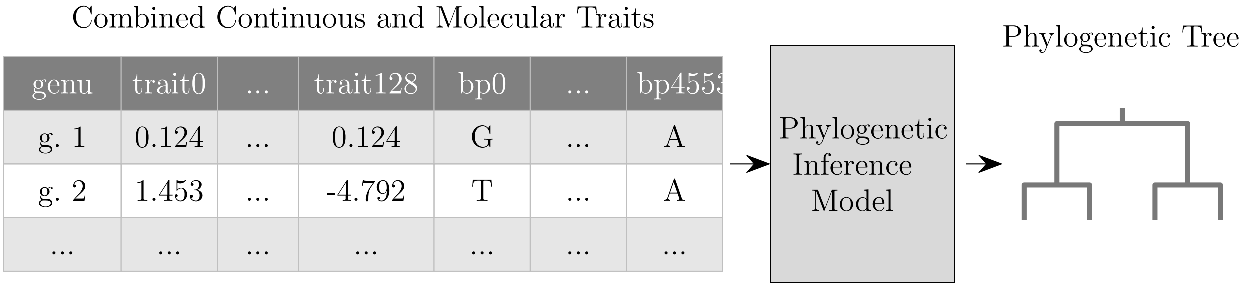
Materials and Methods

The analysis pipeline can be split into several steps — a graphical overview of the process is shown in figures 1 and 2. Each step is explained in the following sections.



genus	trait0	...	trait128
g. 1	0.124	...	0.124
g. 2	1.453	...	-4.792
...	...	...	...
g. 44	-0.569	...	2.223

Fig. 1. Pipeline for generation of continuous morphological traits per genus. A deep metric learning model is trained on the dataset images. It uses the species determinations to pull specimens from the same genus closer together in space. A vector of 128 traits is output from the model for each image. The average vector for each genus is then calculated.



genu	trait0	...	trait128	bp0	...	bp455
g. 1	0.124	...	0.124	G	...	A
g. 2	1.453	...	-4.792	T	...	A
...	...	...	...	...	...	...

Fig. 2. Pipeline of phylogenetic tree generation. First the continuous traits given by the deep learning model are combined with the molecular data. Both of these are then fed into the phylogenetic inference model which generates a tree.

Datasets for phylophenomics. — Deep learning methods require training datasets in order to learn representations of the data directly from the data. Few datasets for phylogenetic inference from images exist at present. iNaturalist (Van Horn et al., 2021), the butterfly dataset (Cuthill et al., 2019) and Rove-Tree-11 (Hunt and Pedersen, 2022) are the most notable of these. Other datasets are in-the-wild, making morphological analysis difficult, and/or only provide a shallow reference phylogeny, e.g. (Fink and Ullman, 2008; Wu et al., 2019; Goëau et al., 2021). However, iNaturalist has the disadvantage that the images are in-the-wild (images taken from non-standardized angles, lighting, backgrounds, etc.) making morphological trait extraction more difficult and both iNaturalist and the butterfly dataset have the problem that the phylogenies derived from such datasets are not particularly fine-grained. Therefore, here we focus our attention on the Rove-Tree-11 dataset which provides over 13,000 dorsal images of rove beetles and associated 11 level deep reference phylogeny. Additionally, none of these datasets come with associated DNA sequences. In gathering the DNA data for the Rove-Tree-11 dataset for this paper we hope to open many new research opportunities.

Dataset

Images and the reference phylogeny used in this analysis are from Rove-Tree-11 (Hunt and Pedersen, 2022). Rove-Tree-11 is a dataset of 13,887 segmented dorsal images of pinned beetles from the family Staphylinidae (rove beetles) housed in the Entomology collection at the Natural History Museum of Denmark which includes species labels and an associated reference phylogeny based on the state-of-the-art knowledge in the field. The reference phylogeny was generated by expert coleopterists at the Natural History Museum of Denmark by combining recently published phylogenies from (Chani-Posse et al., 2018; Brunke and Smetana, 2019; Żyła and Solodovnikov, 2020; Żyła et al., 2022).

The dataset was specifically released to explore deep-learning image based phylogenetic research, and as far as we know, it is the highest-depth publicly available dataset, which is why we focus our analysis on this dataset. Examples from the Rove-Tree-11 dataset can be seen in figure 3 along with a subset of the reference phylogeny. The distribution of the dataset on genus and sub-family level is shown in figure 4.

Molecular Data. — The initial Rove-Tree-11 dataset does not include molecular data, and indeed these would be difficult if not impossible to obtain for the exact specimens included in the dataset as they are pinned specimens, likely to have degraded DNA (Mandrioli, 2008). Therefore molecular data was gathered from GenBank to augment

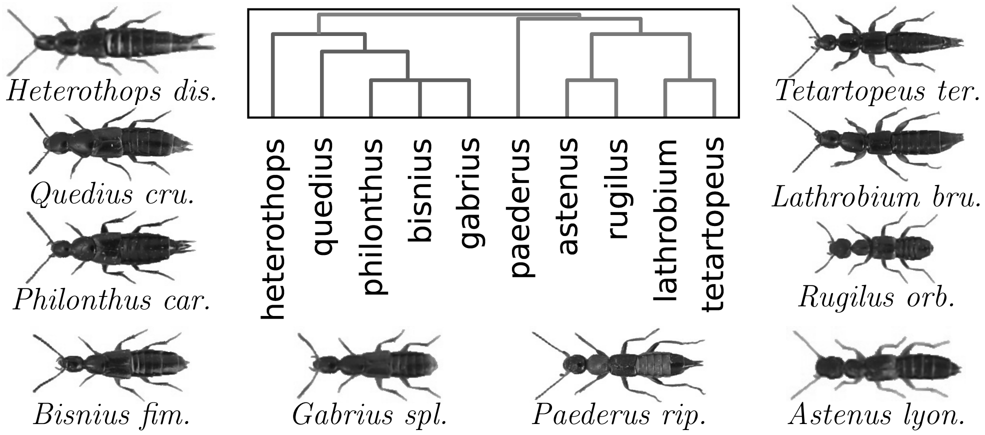


Fig. 3. Subset of the reference phylogeny from the Rove-Tree-11 dataset, for the 10 genera with the most images in the dataset. Each leaf represents a genus. Example specimens from each of the genera are shown in black and white for reference. Reproduced from Hunt and Pedersen (2022) with permission.

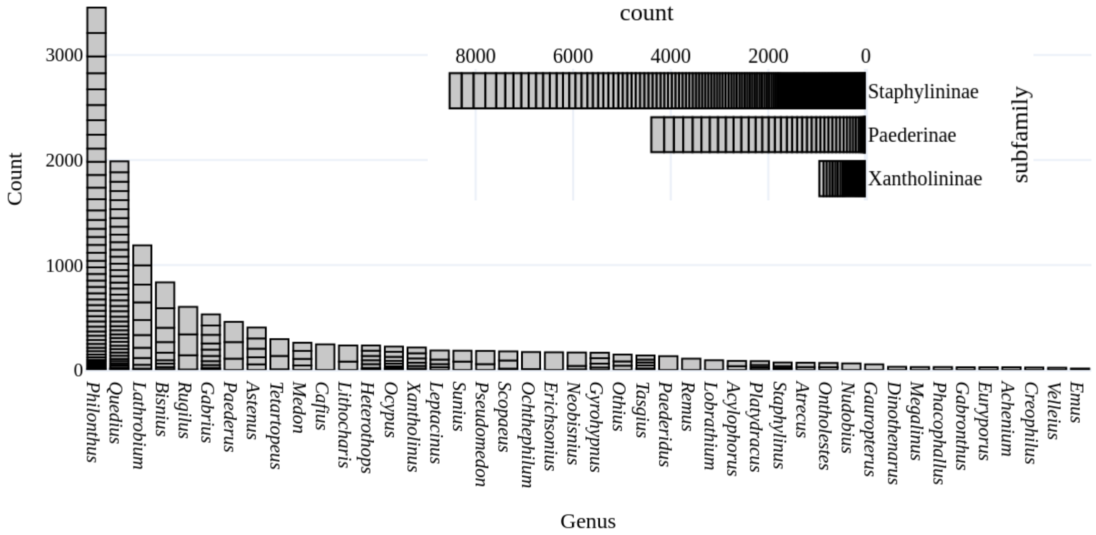


Fig. 4. Distribution of specimens per genus (bottom left) and per subfamily (top right). Each slice in the stacked bar chart represents a different species within that genus. Reproduced from Hunt and Pedersen (2022) with permission.

the Rove-Tree-11 images and phylogeny. Since molecular data is not available for all individual species, we focused on covering examples from every genus. Genus *Velleius* from the original Rove-Tree-11 is downgraded to a subgenus of *Quedius* and therefore it is not present in our analyses. Seven genes were used in order to cover the majority of the genera present in the dataset: carbamoyl- phosphate synthetase (cadA and cadC), topoisomerase I (topo), arginine kinase (argK), wingless (Wg), mitochondrial protein-encoding COI (COI) and nuclear ribosomal 28S (28S). These were chosen based on their wide use and availability in molecular phylogenies focused on Staphylininae, Paederinae, and Xantholininae (Chatzimanolis et al., 2010; Brunke et al., 2016; Schomann and Solodovnikov, 2017; Chani-Posse et al., 2018; Żyła and Solodovnikov, 2020; Jenkins Shaw et al., 2020). All molecular data used in this analysis is provided in the supplemental material. For protein-coding genes, the reading frame was found with Aliview (Larsson, 2014). Individual gene fragments were aligned using MAFFT 7 (Katoh and Standley, 2013). Trimming was performed on alignments using TrimAl 1.3 (Capella-Gutiérrez et al., 2009) with -gappyout setting. The alignments were concatenated with FASconCAT-G (Kück and Longo, 2014). The best partition scheme and model selection were obtained under the Bayesian inference criterion using PartitionFinder 2.1.1 (Lanfear et al., 2016). The following parameters were considered: "all models", the lengths of the branches were established so that they were "unlinked" and the search was established in the algorithm "greedy" (Lanfear et al., 2012).

Deep learning model

Following the methodology used by Hunt and Pedersen (2022) and Roth et al. (2020), we used a ResNet50 architecture (He et al., 2016) pretrained on the ImageNet dataset (Russakovsky et al., 2015) with 128 latent features and compared across various deep metric learning loss functions, described below. Models were trained for 50 epochs with the best checkpoint based on the validation results used to generate the continuous traits. For all training sessions, we fixed mini-batch size to 8 samples with gradient accumulation after 14 batches (an effective batch size of 112), a learning rate of 1e-5 with a step scheduler and a weight decay of 1e-4. Data augmentation was applied to improve the generalizability of the model. Details of the augmentations is available in the supplemental materials. Unlike Hunt and Pedersen (2022) we trained the deep learning model on genus-level, since molecular data is not available at species level and therefore the total evidence analysis is performed at genus-level.

Loss Functions. — Various deep metric learning loss functions were compared, based on those used by Hunt and Pedersen (2022) and Roth et al. (2020). We focused on six popular deep metric learning loss functions out of thousands available. These include four ranking-based losses: triplet (Wu et al., 2017), margin (Wu et al., 2017), contrastive (Hadsell et al., 2006), and multisimilarity (Wang et al., 2019). Lifted was not used contrary to Hunt and Pedersen (2022) because it did not appear to properly converge. Furthermore, we also explore one proxy-based loss: proxyNCA (Movshovitz-Attias et al., 2017) and one classification-based loss: arcface (Deng et al., 2018). We adopted a beta parameter of 0.6 for the margin loss, consistent with prior work by Hunt and Pedersen (2022). Distance-based batch mining was applied to margin, contrastive, and triplet losses. Additional details can be found in Roth et al. (2020).

Dataset split. — In order to train and evaluate the performance of deep learning methods a dataset is usually split into training, validation and test datasets. How the chosen dataset is split into these subsets can greatly affect the results obtained by deep learning algorithms (Tyagi and Mittal, 2020). The original Rove-Tree-11 dataset is divided into training, validation, and test sets based on sub-family classification: Staphylininae for training, Paederinae for validation, and Xantholininae for testing, termed the 'Clade' split. This arrangement aims to assess a deep learning model's capacity to learn phylogenetic relationships. However, we aim to investigate whether utilizing a typical stratified classification split could enhance the model's ability to discern phylogenetically significant features by exposing it to examples from all species during training. Hence, we introduce an additional dataset split called 'Stratified'. A stratified dataset split involves dividing the dataset into training, validation, and test sets in such a way that each set contains approximately the same percentage of samples of each target class as the original dataset, ensuring representative distribution of classes across each split.

For the stratified dataset split, we maintain the training, validation, and test set divisions but randomly allocate 70% of the images from each species to the training set, with the remaining 30% split equally between validation and test sets. This split conforms to standard practice in deep learning classification tasks. The new split's details are provided in a supplemental csv file for reproducibility purposes. No other model parameters were altered for this investigation beyond the adjustment in the training split.

Maximum Parsimony Analysis

Due to software limitations, as mentioned earlier, we exclusively employ maximum parsimony for our analyses. This decision stems from the constraint that RevBayes is the sole Bayesian program supporting continuous traits; unfortunately, combining it with missing molecular data yielded convergence issues, as noted in an open issue on the RevBayes GitHub repository (Zhang, 2022). We utilize TNT (Goloboff et al., 2008) for maximum parsimony analyses, employing random addition sequences with TBR branch swapping across 100 replicates, followed by generating a Nelsen strict consensus and conducting bootstrapping with 100 repetitions. In total evidence analyses with maximum parsimony, we enforce monophyly within genera to enable tree score calculation at the genus level. Example scripts are included in the supplemental material.

Quantitatively Comparing Trees

To assess the quality of trees obtained through various methods, we employ quantitative techniques for comparing them to a reference phylogenetic tree. A common metric in phylogenetics is the normalized Robinson-Foulds (nRF) score (Robinson and Foulds, 1981). While advantageous for comparing trees of different sizes and depths, it has limitations; it relies on binary matching of tree edges, potentially inflating scores for trees with few errors. Alternatively, the align score (Nye et al., 2006) computes the intersection over union for matched edges, providing a fairer representation of tree differences. However, it lacks normalization, making direct comparisons across trees of different sizes challenging. To address this, we propose normalizing the align score based on the upper bound of matched edges. Both metrics are utilized: the align score is adept at detecting subtle differences between trees, while the nRF score is widely accepted and easier to interpret. Since our reference tree lacks branch length information, other methods focusing on branch length differences are irrelevant. Further details and comparisons between the two metrics are provided in Kuhner and Yamato (2014).

Quanitifying Phylogenetic Signal in Traits

Numerous methods are available to quantify the phylogenetic signal in traits, with a comprehensive comparison provided in Münkemüller et al. (2012). Among these methods, Abouheif's Cmean (Abouheif, 1999) stands out as it does not rely on branch lengths and our phylogeny does not have branch lengths. This method evaluates the autocorrelation of trait values across the tree's leaves and tests its significance against a randomly permuted dataset. Specifically, the function computes the sum of squared differences between adjacent trait values along the ordered list of leaves, divided by the total difference. The equation for Abouheif's Cmean is shown in (1), where y_i represents the trait value for species i, y_{i+1} denotes the trait value for the neighboring species in the ordered list of leaves, and N indicates the total number of species. As a result, the function is normalized, and higher values indicate a stronger phylogenetic signal. The Cmean is defined as

$$Cmean(y) = 1 - \frac{\sum_{i=0}^{N-1}(y_{i+1} - y_i)^2}{2\sum_{i=0}^{N} y_i^2} \ . \tag{1}$$

Gene Ablations

To assess the impact of adding genetic data on continuous traits results, we conducted maximum parsimony analyses using all possible combinations of the seven genes (argK, cadA, cadC, COI, 28S, topo, Wg) employed in this study. This entails 127 total combinations. Due to computational constraints, we focused on a single model with the lowest normalized Align Score — Triplet loss trained on the Stratified dataset split with a random seed of 4—to investigate how altering the number of genes affects total evidence results.

Given that not all genes were available for all genera, the resulting tree sizes varied in the molecular-only ablation study. As the normalized Align Score remains somewhat sensitive to tree size despite normalization, total evidence trees with more genera may face a slight disadvantage. To mitigate this, when calculating the normalized Align Score for total evidence trees compared to the reference tree, we excluded any genera lacking molecular data.

Results

Deep Learnt Morphological Traits: Phylogenetic Signal and the Effect of the Dataset Split

To demonstrate that the deep learnt morphological traits have captured some phylogenetic signal, we report normalized Robinson Foulds (nRF) scores and normalized Align Scores (nAS) for trees inferred from the traits using maximum parsimony and varying dataset splits and loss functions (table 1) and cmean values for the traits of each model along with their p values (table 2). To gain some insight into the traits, a sample of gradcam saliency maps (Selvaraju et al., 2017) is shown in figure 5.

The positive examples (top) in figure 5 show that the deep learnt traits can focus on phylogenetically important morphological features. In this case the model appears to mainly look at the abdomen for this trait. However, the negative examples (bottom) show the model focussing on different areas in the same trait. *Rugilus* (f) shows the model focussing on the narrow neck, a distinctive *Ruglilus* trait. For *Nudobius* (e) this trait is looking at the majority of the beetle, showing that there is no direct relationship between

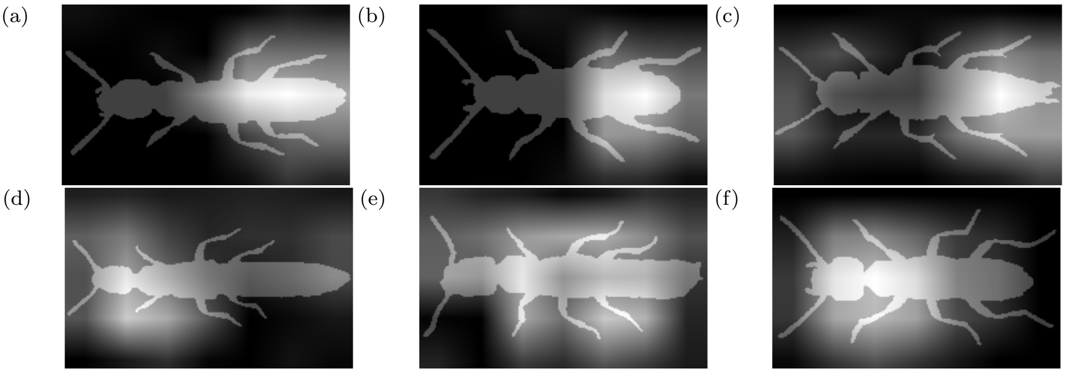


Fig. 5. Examples of gradcam saliency maps (trait 30 from the stratified dataset with triplet loss, seed 4). Saliency maps are shown superimposed on the mask of the beetle with brighter pixel values indicating higher influence on the latent variable. Saliencies in the top row are from genera (a) *Atrecus*, (b) *Lithocharis* and (c) *Ontholestes*, these show the model focusing on the abdomen for this trait. And (d) *Gyrohypnus*, (e) *Nudobius* and (f) *Rugilus* demonstrate some counter examples for this trait. In the case of *Rugilus* (f) the model focuses instead on the neck region which is distinctively small in the *Rugilus* genus.

the trait and the morphological area of interest, making the results difficult to interpret.

Table 1. *Tree Inference Results. Each row is the average of 5 runs. As a baseline for the scores, five randomly generated trees of this size gave a nAS*[a] *of* 0.702 ± 0.022 *and a nRF*[b] *score of* 0.993 ± 0.007. *Results are reported with 95% confidence intervals, using a student's t-distribution. Best results in bold. Results within confidence interval of the best score are underlined.*

			nAS			nRF	
Dataset Split	Loss Function		Average	Median		Average	Median
Clade	Arcface	0.596 ± 0.020	0.627 ± 0.022	0.625	0.947 ± 0.019	0.975 ± 0.017	0.969
Clade	Contrastive		0.629 ± 0.017	0.637		0.967 ± 0.005	0.969
Clade	Margin		0.615 ± 0.045	0.616		0.975 ± 0.017	0.970
Clade	Multisim.		$\mathbf{0.506 \pm 0.034}$	0.512		0.877 ± 0.043	0.897
Clade	Proxy		0.638 ± 0.009	0.634		0.987 ± 0.022	1.000
Clade	Triplet		0.560 ± 0.045	0.563		0.901 ± 0.065	0.903
Stratified	Arcface	0.595 ± 0.031	0.692 ± 0.042	0.708	0.913 ± 0.028	0.992 ± 0.022	1.000
Stratified	Contrastive		0.531 ± 0.024	0.538		$\mathbf{0.829 \pm 0.026}$	**0.815**
Stratified	Margin		0.596 ± 0.036	0.597		0.928 ± 0.045	0.925
Stratified	Multisim.		0.524 ± 0.076	**0.505**		0.845 ± 0.057	0.825
Stratified	Proxy		0.699 ± 0.009	0.699		1.000 ± 0.000	1.000
Stratified	Triplet		0.530 ± 0.067	0.532		0.886 ± 0.062	0.900

From table 1 we can make a few immediate observations about the tree inference. First, we can see that the dataset split has almost no influence over the average results, and the best overall model on average according to the normalized Align Score was using

[a] normalized Align Score
[b] normalized Robinson-Foulds

the cladistic dataset split. Next we see some preference for different loss functions. In particular contrastive, multisimilarity and triplet losses provide the best results. The normalized Align and Robinson Foulds scores also show a significant improvement to randomly generated trees in 10 out of 12 models, further indicating that the continuous traits can carry a phylogenetic signal, but also that model choice is important.

Table 2. *Phylogenetic Signal Quanitification Using Abouheif's Cmean. Each row is the average of the 128 traits of all 5 runs. Results are reported with 95% confidence intervals, using a student's t-distribution. Best results in bold. Results within confidence interval of the best score are underlined.*

						p value < 0.05		
		Average Cmean		Maximum Cmean		with Cmean		
Dataset Split	Loss Function	Cmean	p value	Cmean	p value	> 0.3	> 0.5	> 0.7
Clade	Arcface	0.107 ± 0.010	0.219 ± 0.019	0.480	0.001	9%	0%	0.0%
Clade	Contrastive	$\mathbf{0.288 \pm 0.014}$	0.067 ± 0.011	**0.794**	0.001	**46%**	**14%**	**0.6%**
Clade	Margin	0.188 ± 0.012	0.118 ± 0.015	0.610	0.001	21%	2%	0.0%
Clade	Multisim.	0.208 ± 0.013	0.118 ± 0.016	0.658	0.001	28%	4%	0.0%
Clade	Proxy	0.101 ± 0.010	0.226 ± 0.020	0.519	0.001	6%	0%	0.0%
Clade	Triplet	0.244 ± 0.012	0.076 ± 0.011	0.683	0.001	35%	5%	0.0%
Stratified	Arcface	-0.028 ± 0.007	0.511 ± 0.022	0.241	0.008	0%	0%	0.0%
Stratified	Contrastive	$\underline{0.281 \pm 0.012}$	0.050 ± 0.009	0.706	0.001	44%	8%	0.2%
Stratified	Margin	0.164 ± 0.011	0.138 ± 0.015	0.578	0.001	16%	1%	0.0%
Stratified	Multisim.	0.128 ± 0.010	0.179 ± 0.018	0.517	0.001	10%	0%	0.0%
Stratified	Proxy	-0.025 ± 0.008	0.498 ± 0.023	0.273	0.007	0%	0%	0.0%
Stratified	Triplet	0.232 ± 0.011	0.076 ± 0.012	0.645	0.001	31%	4%	0.0%

Abouheif's Cmean values in table 2 show relatively low average trait values, and a high variation in the maximum phylogenetic signal per trait in the different models. However, all maximum Cmeans have a significant p value (less than 0.05), indicating a strong and significant phylogenetic relationship obtained in some traits. The final three columns show the percent of traits which have a significant p value (less than 0.05) and a Cmean above a certain threshold, indicating the percentage of traits with a significantly strong phylogenetic signal. Of these, relatively few have a Cmean above 0.5. From this the Cladistic dataset split appears to have a positive influence on the phylogenetic signal, with the cladistic split, contrastive loss model significantly outperforming the other models. It is interesting to contrast this with the results in table 1 where the cladistic split, contrastive loss model performs worse than the average. This shows the strong influence of the metric on the results. When interpreting these results, we prefer to favor the align score as it directly measures the performance of phylogenetic inference from the traits, while Abouheif's Cmean, as an autocorrelation function is also influenced by the order the leaves are presented in the tree.

Adding Molecular Sequences: Total Evidence Analysis

Here we report results from the total evidence analysis where we combine both deep learnt morphological traits and molecular traits to complete the phylogenetic inference. We compare this combined result with using molecular data alone.

From table 3 we can make a couple of observations. First, the dataset split does significantly affect the results, with the clade split providing the best Align Score on average. This can be contrasted with the deep learnt traits only results in table 1, where the dataset split results were well within eachother's confidence intervals. Secondly, we can see that the results are on average slightly improved using total evidence compared with molecular traits only, however 10 out of 12 models have average nAS scores within the confidence interval of the molecular only model, suggesting we should be cautious in drawing conclusions. That said, only two models have average nRF scores within the molecular only model's confidence interval. This further highlights the discrepancies between these metrics, however, we believe the nAS is a more fair metric, for reasons explored in the methods section.

Qualitative Tree Comparison — In figure 6 we show (a) the reference phylogeny which we use as a gold standard, (b) the best molecular tree (c) the best total-evidence tree tree. Differences between them are labelled as groups 1-4 and examined as follows:

1. The total evidence tree figure 6(c) correctly places *Heterothops* and *Aclyophorus* in Staphylininae. while the molecular only tree figure 6(b) clades them as a sister group to Xantholininae and Staphylininae.

Table 3. *Results of Total Evidence Analysis - Combining Molecular data and Deep Learnt Morphological Traits. Confidence intervals are 95% level based on 5 runs. As a baseline comparison for the scores, five randomly generated trees of this size would give a nAS of* 0.702 ± 0.022 *and a nRF Score of* 0.993 ± 0.007*. Best results in bold. Results within confidence interval of the best score are underlined.*

				nAS			nRF	
Traits	Dataset Split	Loss Function		Average	Median		Average	Median
Molecular Only	-	-		0.141 ± 0.017	0.147		0.382 ± 0.040	0.405
Total Evidence	Clade	Arcface	0.128 ± 0.006	0.139 ± 0.021	0.138	0.320 ± 0.010	0.337 ± 0.024	0.333
Total Evidence	Clade	Contrastive		**0.119 ± 0.014**	0.119		0.311 ± 0.033	0.315
Total Evidence	Clade	Margin		0.127 ± 0.017	0.127		0.319 ± 0.025	0.315
Total Evidence	Clade	Multisim.		0.135 ± 0.024	0.136		0.316 ± 0.045	0.306
Total Evidence	Clade	Proxy		0.127 ± 0.013	0.126		0.324 ± 0.041	0.315
Total Evidence	Clade	Triplet		0.121 ± 0.023	**0.118**		**0.309 ± 0.031**	0.324
Total Evidence	Stratified	Arcface	0.147 ± 0.011	0.166 ± 0.037	0.160	0.339 ± 0.012	0.383 ± 0.028	0.389
Total Evidence	Stratified	Contrastive		0.141 ± 0.032	0.139		0.313 ± 0.019	**0.306**
Total Evidence	Stratified	Margin		0.162 ± 0.058	0.143		0.324 ± 0.011	0.324
Total Evidence	Stratified	Multisim.		0.135 ± 0.016	0.133		0.337 ± 0.026	0.333
Total Evidence	Stratified	Proxy		0.155 ± 0.028	0.149		0.365 ± 0.021	0.361
Total Evidence	Stratified	Triplet		0.121 ± 0.012	0.119		0.310 ± 0.017	0.315

2. The total evidence tree figure 6(c) places *Gauropterus* inside the majority of Xantholininae, as expected, while the molecular only tree figure 6(b) erroneously places *Gauropterus* outside the majority of Xantholininae.

3. The total evidence tree figure 6(c) pushes these clades to be unresolved, closely reflecting our current empirical data about these clades while the molecular-only tree figure 6(b) undesirably resolves these clades

4. The total evidence tree figure 6(c) erroneously nests *Quedius* and *Euryporus* inside the tribe Staphylinini, while the molecular only tree figure 6(b) correctly places them outside Staphylinini.

In total, three out of four controversial groups in this example are better placed by the total evidence tree.

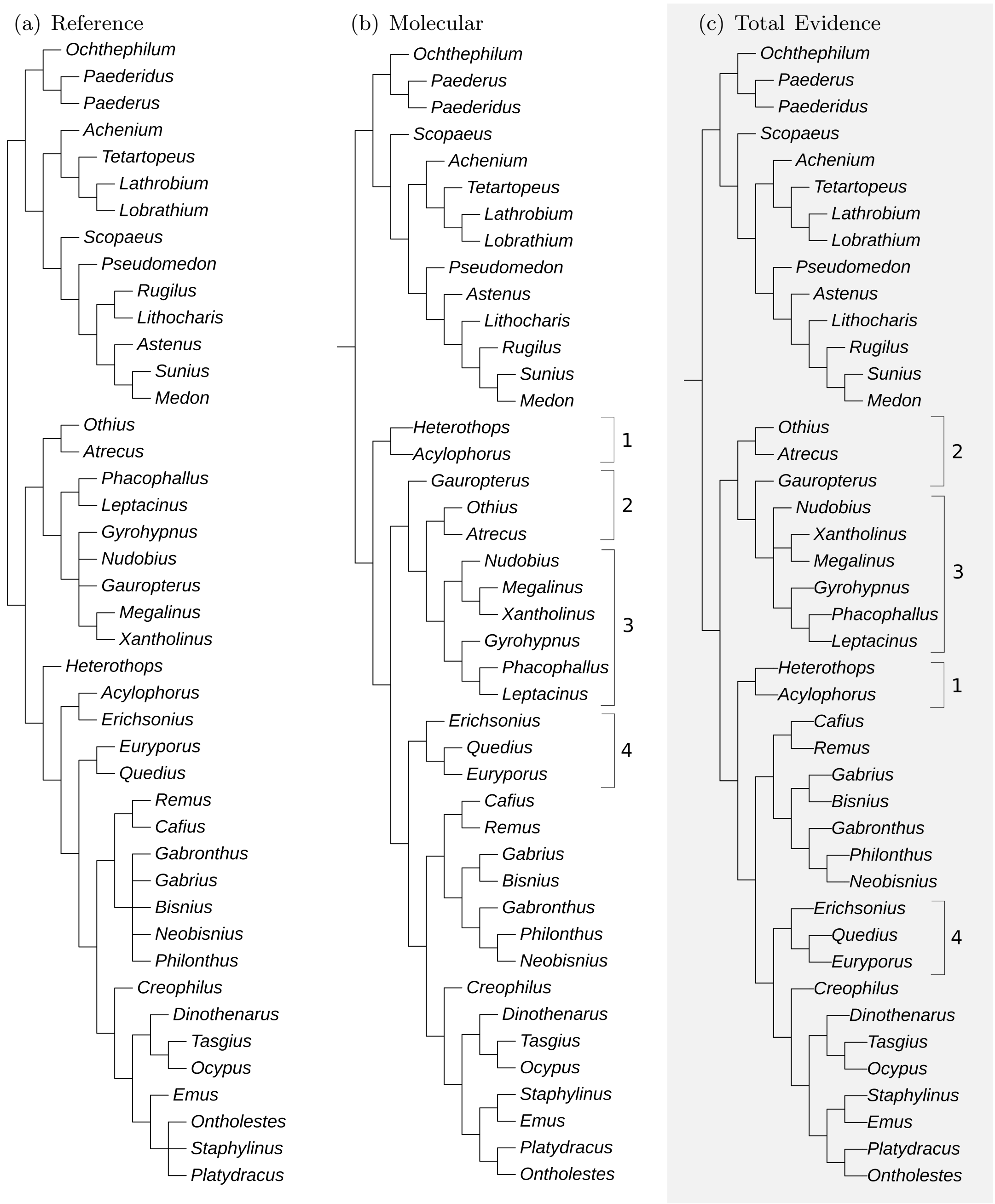


Fig. 6. Comparison of a) reference phylogeny, b) best total evidence tree and c) best molecular-only tree. Differences between best molecular-only and best total evidence tree highlighted by indicating the controversial groups 1,2,3,4 on both trees. Plots produced in part using iTOL (Letunic and Bork, 2021)

Varying amounts of Molecular Sequences: Gene Ablation Study

Figure 7 shows how changing the number of genes included in the analysis affects the results. We notice that the total evidence ablation results are slightly improved on

average compared to the molecular-only results. However, there is a large variation in the normalized Align Score. This variation is decreased as we add more genes, which also makes sense as the number of possible combinations decreases and the algorithm has more data to converge on.

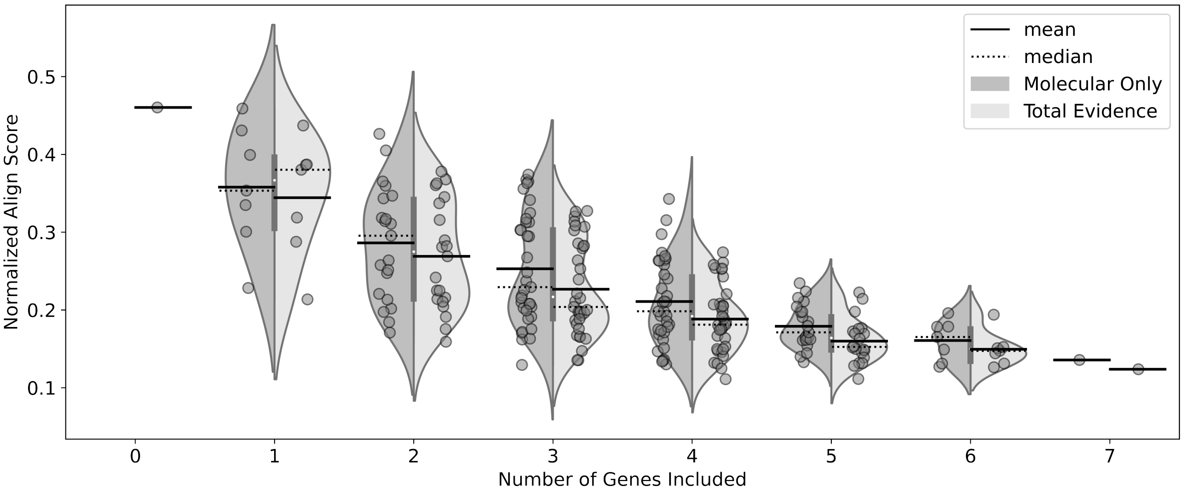


Fig. 7. Violin plot of the effect of including different gene combinations on the normalised Align Score. In each column the results to the left (darker) are for molecular-only ablations, and to the right (lighter) are for total evidence ablations. Each point represents an individual result.

Table 4. *Best subset of genes*[c] *given the number of genes. Best results in bold.*

	Molecular Only			Total Evidence		
No. Genes	Best Gene Combination	nAS	nRF	Best Gene Combination	nAS	nRF
0	-	-	-	-	0.461	0.846
1	28S	0.228	0.433	28S	0.214	0.420
2	28S, ArgK	0.171	0.362	28S, ArgK	0.159	0.408
3	28S, COI, wing	0.129	**0.275**	28S, ArgK, topo	0.135	0.286
4	28S, COI, topo, wing	0.130	0.314	28S, CADb, COI, wing	**0.111**	0.278
5	28S, ArgK, CADb, COI, wing	0.133	0.432	28S, CADb, COI, topo, wing	**0.111**	0.288
6	28S, ArgK, CADa, CADb, COI, topo	0.127	0.342	28S, ArgK, CADb, COI, topo, wing	0.126	0.324
7	All	0.136	0.378	All	0.124	0.315

In table 4 we see disagreement between the nRF and nAS scores, and in this case we follow the nAS scores as they are more stable. This indicates that the best model includes continuous deep learnt traits and the 4 genes 28S, CADb, COI, and Wg. On

[c] Genes used in this analysis: nuclear ribosomal 28S (28S), arginine kinase (ArgK), carbamoyl- phosphate synthetase (cadA and cadC), mitochondrial protein-encoding COI (COI) topoisomerase I (topo) and wingless (Wg)

average we can see both from table 4 and from figure 7 that the nAS in the total evidence results are generally lower and therefore better than they are in the molecular only results, further suggesting that deep learnt continuous traits can add to phylogenetic analyses.

Discussion

Deep Learnt Morphological Traits: Phylogenetic Signal

Table 1 shows that all of the models on average have a better align score than random trees, and 10 out of 12 models are outside of the confidence intervals for the random trees, indicating that the deep learnt traits can indeed carry a significant phylogenetic signal. However, that signal is still difficult to directly interpret as demonstrated by the saliency maps in figure 5. While applying saliency maps that are normally used in classification to latent variables is itself questionable, we can see that this deep learnt trait does appear to primarily focus on a single body part, however the saliency map can deviate from this, and while we can explain some of this through our knowledge of the distinctive traits (like the slim neck of *Rugilus*), others are difficult to interpret. We assume that this, in part, demonstrates the dependencies between these traits, and potentially demonstrates that further exploration is necessary into explaining these traits. We have put no constraints on the traits to be independent, and indeed we can see that many traits focus on the same area for example from the same genera. It could be that interpreting such models could be made easier by the use of disentangled networks. We discuss this further in the future work section.

The Effect of the Dataset Split

Tables 1 and 3 also provide insight into the dataset split. Both indicate a preference for the cladistic split, but in table 3 this preference is significant. This is interesting as we might expect to see the opposite - that the difference would be more pronounced when the molecular traits have no influence on the results. We are also surprised that the preference is not for the stratified split. We might expect that allowing the model to learn from all subfamilies would improve the model's ability to learn phylogenetically important traits for other groups. However, these results are encouraging, in the sense that they indicate that models trained on one clade may be highly successful in inferring relationships in unseen clades, reducing the necessity for retraining the model. However, further investigation into this phenomena on a variety of datasets would be required to firm up a conclusion on this aspect.

Loss Functions

When comparing across loss functions in table 1 and table 3 we see a clear preference for contrastive, multisimilarity and triplet losses. This is in line with results obtained in Hunt and Pedersen (2022). What is interesting is that the results from Abouheif's Cmean and the total evidence analysis in tables 2 and table 3 suggest a strong preference for the contrastive loss, which is not necessarily reflected in the results in table 1 where we infer the phylogeny directly from the morphological traits. A high value of Abouheif's Cmean is highly correlated to deeper phylogenetic relationships. Therefore one hypothesis could be that the contrastive loss is better at picking up deeper (tribe-level and above) phylogenetic relationships, but not so good at picking up more shallow relationships, making it result in traits which have a relatively high Cmean, but relatively poor overall tree. However, since molecular data is quite good for picking up shallow relationships, when combined with these morphological traits, this could result in a good total evidence tree. This is one potential explanation, but this would need to be investigated further and is outside the scope of this analysis.

Qualitative Tree Comparison

The total-evidence best tree topology (figure 6c) reveals improved placements of the *Heterothops* and *Gauropterus* inside Staphylininae, as well as it reflects the uncertainty in the unresolved position of the *Xantholinus* and *Megalinus* clades. However, it places the *Erichsonius*+(*Quedius*+*Europorus*) clade inside the tribe Staphylinini. Alternative phylogenetic placement of these groups suggests that deep learnt traits improved phylogenetic resolution at deeper nodes, in our case at the subfamily level, but failed with more terminal resolution, in our case at the tribal level.

Gene Ablation Study

From figure 7 and table 4 we can see that the total-evidence analysis with four genes (28S, CADb, COI and Wg) provides the best result of the ablation studies. From figure 7 we can see that there is significant variation in the results when a single gene is added, reinforcing the notion that gene choice is important for phylogenetic analyses. Figure 7 further demonstrates that good results can be obtained with relatively few genes, and that this result is not drastically affected by adding deep learnt morphological traits. We can further say that deep learnt morphological traits can indeed improve the phylogenetic analysis, with mean values below that of the molecular results, however this result should be used with caution as for the gene ablation study we used traits from the model which performed the best. Therefore a comparison of different models on the data of interest should be completed to make further conclusions from these analyses.

Conclusion

Quantitative morphological characteristics, extracted through the application of deep learning techniques applied to images of pinned insect specimens produced for mass collections digitization purposes, have been demonstrated to possess phylogenetic relevance. These traits, when integrated into molecular phylogenies, have the potential to augment the phylogenetic framework in a comprehensive total-evidence based approach, offering the possibility of incorporating species lacking molecular data into phylogenetic trees. However, the improvement of phylogenetic reconstructions by the inclusion of such morphological data derived through deep learning methodologies remains minimal. While this approach has shown promise, scaling up its implementation is still not feasible for two reasons. First, the phylogenetic signal of the deep learnt traits, at least in our dataset, was not too strong to justify further effort in gathering the same type of data in the same way. Second, the effort required for our image-based dataset compilation, even though obviously lesser than an effort by the expert to assemble a traditional morphological phylogenetic matrix, is still significant. Despite these challenges, as will be elaborated in the section on future research directions, we stress the opportunities for removal of both impediments. It is conceivable to amplify the phylogenetic signal automatically extracted from the collections specimen images via improved models, targeted loss functions, transfer learning. At the same time improved imaging, image processing and on-going optimization of these steps should enhance the data acquisition and thus large-scale non-destructive use of the digitized collections.

Future Work

Several promising directions exist within the burgeoning domain of deep metric learning, which could augment the phylogenetic signal gleaned from these models. Firstly, with natural history museums worldwide embarking on the digitization of their collections (Davies and et al, 2017; Hedrick et al., 2020; Popov et al., 2021; Ahlstrand, 2023; Johnson et al., 2023), an expansion of publicly accessible data to nearly complete sampling of big taxa is anticipated. Given that the performance of deep learning models is generally enhanced by training on more representative datasets (Sun et al., 2017), it is anticipated that the increased availability of image datasets in this domain will significantly enhance the efficacy of resultant models.

Secondly, the domain of deep learning is experiencing rapid evolution, characterized by the emergence of numerous sub-fields. It is still unclear which improvements to deep learning models may significantly increase the extracted phylogenetic signal. Two areas, in particular, hold significant promise. A growing body of research (Eddahmani et al., 2023) is investigating the enforcement of independence among continuous traits, which could lead to more explainable traits and enable their statistical dissociation. Another area of interest pertains to the application of variational autoencoders to encourage the continuous trait space to conform to specific distributions (e.g., a normal distribution), which has been demonstrated to improve clustering outcomes in various instances (Kingma and Welling, 2019). Invoking the central limit theorem, the normal distribution possesses intrinsic appeal in biological contexts and may facilitate a better structured trait space. More work also should be done into how to measure the phylogenetic signal in these models.

Thirdly, the integration of deep learning-derived molecular embeddings with morphological embeddings in phylogenetics remains an underexplored avenue, despite significant investment in the development of deep learning approaches for molecular traits (Mo et al., 2024).

Fourthly, there is a necessity for further investigation into the quantification and elucidation of deep learning-derived traits. The identification of loss functions capable of incentivizing models to discern phylogenetically pertinent traits remains elusive. Moreover, existing metrics for quantifying the informational content of traits, such as Abouheif's Cmean or align scores, lack intuitiveness. While saliency maps offer a generalized explanation of model behaviors, their applicability to latent variable models is not straightforward, and a comparison of different latent variable saliency methods should be completed.

Lastly, we advocate for the conceptualization of deep learning-derived traits as distributions rather than singular values. Morphological traits are traditionally represented as binary values across clades; however, we contend that recognizing each species as a continuous distribution of traits could hold substantial value. The employment of continuous variables in conjunction with Bayesian methodologies appears particularly conducive to this perspective.

Environmental Footprint

Using carbontracker (Anthony et al., 2020) we estimate that the training of each deep learning model in this paper used 1.32 kWh of power (based on one full model run), translating to 192g of CO_2. The full published deep learning results from this paper therefore produced an estimated 2.304 kg of CO_2. The carbon production from the phylogenetic models and from early experimentation is not included in this estimate.

Funding

This work is partially supported by the University of Copenhagen Data+ funding (Strategic funds 2023) under the project ”Expanding the Tree Life through a digital view of museum collections (PHYLORAMA)” and a Villum Experiment Grant under the project ”Baltic Amber Enigma”.

Acknowledgements

Many people have contributed to this work, but in particular, we would like to thank François Lauze for his advice throughout this process and David Gutschenreiter, Søren Bech and André Fastrup for their work gathering images and completing segmentations on the Rove-Tree-11 dataset. We would like to express our gratitude to ChatGPT for its assistance in refining the text.

Supplementary Material

The data underlying this article are available at `http://doi.org/10.17894/ucph.39619bba-4569-4415-9f25-d6a0ff64f0e3` for the Rove-Tree-11 dataset and in the article’s online supplementary material for the further molecular data and associated genbank accession numbers, example inference code, all generated trees, data augmentation details and stratified dataset split. All trained model runs and extracted continuous trait matrices are available in the following erda repository `https://erda.ku.dk/archives/b1b43e09d7686395c6ada1ccda0de365/published-archive.html`. The reference tree along with the full best molecular tree are shown in full in the Appendix and can be found along with the best total-evidence tree on TreeBASE at `http://purl.org/phylo/treebase/phylows/study/TB2:S30717`. The code used in this analysis is available on github `https://github.com/robertahunt/Revisiting_Deep_Metric_Learning_PyTorch`, commit a6654453c3b7785a17511255e02c468c53fe6f5d, forked from Roth et al. (2020).

References

Abouheif, E. (1999). A method for testing the assumption of phylogenetic independence in comparative data. *Evolutionary Ecology Research*, 1:895–909.

Ahlstrand, N. I. (2023). Digitization of the greenland vascular plant herbarium as a unique research infrastructure to study arctic climate change and inform nature management. *Collections*.

Anthony, L. F. W., Kanding, B., and Selvan, R. (2020). Carbontracker: Tracking and predicting the carbon footprint of training deep learning models. ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems. arXiv:2007.03051.

Baird, R. C. (2010). Leveraging the fullest potential of scientific collections through digitisation. *Biodiversity Informatics*, 7(2).

Brunke, A. and Smetana, A. (2019). A new genus of staphylinina and a review of major lineages (staphylinidae: Staphylininae: Staphylinini). *Systematics and Biodiversity*, 17:745–758.

Brunke, A. J., Chatzimanolis, S., Schillhammer, H., and Solodovnikov, A. (2016). Early evolution of the hyperdiverse rove beetle tribe staphylinini (coleoptera: Staphylinidae: Staphylininae) and a revision of its higher classification. *Cladistics*, 32(4):427–451.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973.

Cavender-Bares, J., Kozak, K. H., Fine, P. V. A., and Kembel, S. W. (2009). The merging of community ecology and phylogenetic biology. *Ecology Letters*, 12(7):693–715.

Chani-Posse, M. R., Brunke, A. J., Chatzimanolis, S., Schillhammer, H., and Solodovnikov, A. (2018). Phylogeny of the hyper-diverse rove beetle subtribe philonthina with implications for classification of the tribe staphylinini (coleoptera: Staphylinidae). *Cladistics*, 34(1):1–40.

Chatzimanolis, S., Cohen, I. M., Schomann, A., and Solodovnikov, A. (2010). Molecular phylogeny of the mega-diverse rove beetle tribe staphylinini (insecta, coleoptera, staphylinidae). *Zoologica Scripta*, 39(5):436–449.

Cuthill, J. F. H., Guttenberg, N., Ledger, S., Crowther, R., and Huertas, B. (2019). Deep learning on butterfly phenotypes tests evolution&#x2019;s oldest mathematical model. *Science Advances*, 5(8):eaaw4967.

Davies, T. G. and et al (2017). Open data and digital morphology. *Proceedings of the Royal Society B*, 284:20170194.

de Almeida, R. F., Cheek, M., Pellegrini, M. O., de Morais, I. L., Simão-Bianchini, R., Rattanakrajang, P., and Simões, A. R. G. (2023). Barking up the wrong tree: the importance of morphology in plant molecular phylogenetic studies. *bioRxiv*.

Deng, J., Guo, J., and Zafeiriou, S. (2018). Arcface: Additive angular margin loss for deep face recognition. *CoRR*, abs/1801.07698.

Eddahmani, I., Pham, C.-H., Napoléon, T., Badoc, I., Fouefack, J.-R., and El-Bouz, M. (2023). Unsupervised learning of disentangled representation via auto-encoding: A survey. *Sensors*, 23(4).

Edwards, A. W. (1970). Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society: Series B (Methodological)*, 32(2):155–164.

Felsenstein, J. (1985). Phylogenies from gene frequencies: a statistical problem. *Systematic zoology*, 34(3):300–311.

Fink, M. and Ullman, S. (2008). From aardvark to zorro: A benchmark for mammal image classification. *Int. J. Comput. Vision*, 77(1–3):143–156.

Fitch, W. M. (1971). Toward defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416.

Furusawa, C., Tsustumi, M., Koyabu, D., and Saito, N. (2023). A method for morphological feature extraction based on variational auto-encoder : an application to mandible shape.

Giribet, G. (2015). Morphology should not be forgotten in the era of genomics–a phylogenetic perspective. *Zoologischer Anzeiger - A Journal of Comparative Zoology*, 256:96–103. Special Issue: Proceedings of the 3rd International Congress on Invertebrate Morphology.

Goëau, H., Bonnet, P., and Joly, A. (2021). Overview of plantclef 2021: cross-domain plant identification. In *Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum*, volume 2936, pages 1422–1436.

Goloboff, P. A. (2022). *Refining phylogenetic analyses.* CRC Press.

Goloboff, P. A., Farris, J. S., and Nixon, K. C. (2008). Tnt, a free program for phylogenetic analysis. *Cladistics*, 24(5):774–786.

Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Hassler, G. W., Magee, A. F., Zhang, Z., Baele, G., Lemey, P., Ji, X., Fourment, M., and Suchard, M. A. (2023). Data integration in bayesian phylogenetics. *Annual Review of Statistics and Its Application*, 10(1):353–377.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of Computer Vision and Pattern Recognition - CVPR'16.*

Hedrick, B. P., Heberling, J. M., Meineke, E. K., Turner, K. G., Gassa, C. J., Park, D. S., Kennedy, J., Clarke, J. A., Cook, J. A., Blackburn, D. C., Edwards, S. V., and Davis, C. C. (2020). Digitization and the future of natural history collections. *BioScience*, 70(3):243–251.

Hossfeld, U. and Levit, G. S. (2016). 'tree of life' took root 150 years ago. *Nature*, 540(7631):38–38.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294(5550):2310–2314.

Hunt, R. and Pedersen, K. S. (2022). Rove-tree-11: The not-so-wild rover, a hierarchically structured image dataset for deep metric learning research. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 2967–2983.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. (2016). RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Systematic Biology*, 65(4):726–736.

Høye, T., Ärje, J., Bjerge, K., Liset Pryds Hansen, O., Iosifidis, A., Leese, F., Mann, H., Meissner, K., Melvad, C., and Raitoharju, J. (2021). Deep learning and computer vision

will transform entomology. *Proceedings of the National Academy of Sciences*, 118:e2002545117.

Jenkins Shaw, J., Żyła, D., and Solodovnikov, A. (2020). Molecular phylogeny illuminates amblyopinini (coleoptera: Staphylinidae) rove beetles as a target for systematic and evolutionary research. *Systematic Entomology*, 45(2):430–446.

Jensen, R. J. (2009). Phenetics: revolution, reform or natural consequence? *TAXON*, 58(1):50–60.

Jiao, L. and Zhao, J. (2019). A survey on the new generation of deep learning in image processing. *IEEE Access*, 7:172231–172263.

Johnson, K. R., Owens, I. F. P., and Group, T. G. C. (2023). A global approach for natural history museum collections. *Science*, 379(6638):1192–1194.

Katoh, K. and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, 30(4):772–780.

Kelly, M. B., McLean, D. J., Wild, Z. K., and Herberstein, M. E. (2021). Measuring mimicry: methods for quantifying visual similarity. *Animal Behaviour*, 178:115–126.

Kiel, S. (2021). Assessing bivalve phylogeny using deep learning and computer vision approaches. *bioRxiv*.

Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

Kuhner, M. K. and Yamato, J. (2014). Practical Performance of Tree Comparison Metrics. *Systematic Biology*, 64(2):205–214.

Kück, P. and Longo, G. (2014). Fasconcat-g: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in zoology*, 11:81.

Lanfear, R., Calcott, B., Ho, S. Y. W., and Guindon, S. (2012). PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. *Molecular Biology and Evolution*, 29(6):1695–1701.

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. (2016). PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. *Molecular Biology and Evolution*, 34(3):772–773.

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, 30(22):3276–3278.

Lee, M. and Palci, A. (2015). Morphological phylogenetics in the genomic age. *Current Biology*, 25(19):R922–R929.

Letunic, I. and Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, 49(W1):W293–W296.

Li Fan, Chunpeng Xu, E. A. J. and Cui, X. (2022). Quantifying plant mimesis in fossil insects using deep learning. *Historical Biology*, 34(5):907–916.

Linné, C. v., Gmelin, J. F., and Beer, G. E. (1788). *Systema naturae per regna tria naturae : secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis*, volume v. 1, pt. 1. Lipsiae [Leipzig], Impensis Georg. Emanuel. Beer, 1788-1793. https://www.biodiversitylibrary.org/bibliography/545.

MacLeod, N. and Forey, P. L. (2002). *Morphology, shape, and phylogeny*. CRC Press.

MacLeod, N., Price, B., and Stevens, Z. (2022). What you sample is what you get: ecomorphological variation in trithemis (odonata, libellulidae) dragonfly wings reconsidered. *BMC Ecology and Evolution*, 22(1):43.

Mandrioli, M. (2008). Insect collections and dna analyses: how to manage collections? *Museum Management and Curatorship*, 23(2):193–199.

Mindell, D. P. (2013). The Tree of Life: Metaphor, Model, and Heuristic Device. *Systematic Biology*, 62(3):479–489.

Mo, Y. K., Hahn, M. W., and Smith, M. L. (2024). Applications of machine learning in phylogenetics. page 13.

Movshovitz-Attias, Y., Toshev, A., Leung, T. K., Ioffe, S., and Singh, S. (2017). No fuss distance metric learning using proxies. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 360–368.

Münkemüller, T., Lavergne, S., Bzeznik, B., Dray, S., Jombart, T., Schiffers, K., and Thuiller, W. (2012). How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, 3(4):743–756.

Nye, T., Lio, P., and Gilks, W. (2006). A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics (Oxford, England)*, 22:117–9.

Otter, D. W., Medina, J. R., and Kalita, J. K. (2021). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624.

Parins-Fukuchi, C. (2017). Use of Continuous Traits Can Improve Morphological Phylogenetics. *Systematic Biology*, 67(2):328–339.

Pichler, M. and Hartig, F. (2023). Machine learning and deep learning—a review for ecologists. *Methods in Ecology and Evolution*, 14(4):994–1016.

Popov, D., Roychoudhury, P., Hardy, H., Livermore, L., and Norris, K. (2021). The value of digitising natural history collections. *Research Ideas and Outcomes*, 7:e78844.

Pyron, R. A. (2015). Post-molecular systematics and the future of phylogenetics. *Trends in Ecology & Evolution*, 30(7):384–389.

Robinson, D. and Foulds, L. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147.

Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D. L., and Rasnitsyn, A. P. (2012). A Total-Evidence Approach to Dating with Fossils, Applied to the Early Radiation of the Hymenoptera. *Systematic Biology*, 61(6):973–999.

Roth, K., Milbich, T., Sinha, S., Gupta, P., Ommer, B., and Cohen, J. P. (2020). Revisiting training strategies and generalization performance in deep metric learning.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Schomann, A. M. and Solodovnikov, A. (2017). Phylogenetic placement of the austral rove beetle genus hyperomma triggers changes in classification of paederinae (coleoptera: Staphylinidae). *Zoologica Scripta*, 46(3):336–347.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626. IEEE Computer Society.

Smith, V. and Blagoderov, V. (2012). Bringing collections out of the dark. *ZooKeys*, 209:1–6.

Sokal, R. and Sneath, P. (1963). *Principles of numerical taxonomy*, page 359. W. H. Freeman and Company, San Francisco.

Soltis, D. E. and Soltis, P. S. (2003). The Role of Phylogenetics in Comparative Genetics. *Plant Physiology*, 132(4):1790–1800.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Tarasov, S. (2019). Integration of Anatomy Ontologies and Evo-Devo Using Structured Markov Models Suggests a New Framework for Modeling Discrete Phenotypic Traits. *Systematic Biology*, 68(5):698–716.

Tyagi, S. and Mittal, S. (2020). Sampling approaches for imbalanced data classification problem in machine learning. In Singh, P. K., Kar, A. K., Singh, Y., Kolekar, M. H., and Tanwar, S., editors, *Proceedings of ICRIC 2019*, pages 209–221, Cham. Springer International Publishing.

Valan, M., Makonyi, K., Maki, A., Vondráček, D., and Ronquist, F. (2019). Automated Taxonomic Identification of Insects with Expert-Level Accuracy Using Effective Feature Transfer from Convolutional Networks. *Systematic Biology*, 68(6):876–895.

Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., and Mac Aodha, O. (2021). Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893.

Wagner, G. P. (1989). The biological homology concept. *Annual Review of Ecology and Systematics*, 20:51–69.

Wang, X., Han, X., Huang, W., Dong, D., and Scott, M. R. (2019). Multi-similarity loss with general pair weighting for deep metric learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5017–5025.

Wiens, J. J. (2001). Character Analysis in Morphological Phylogenetics: Problems and Solutions. *Systematic Biology*, 50(5):689–699.

Wilson, R. J., de Siqueira, A. F., Brooks, S. J., Price, B. W., Simon, L. M., van der Walt, S. J., and Fenberg, P. B. (2023). Applying computer vision to digitised natural history

collections for climate change research: Temperature-size responses in british butterflies. *Methods in Ecology and Evolution*, 14(2):372–384.

Wright, A. M. (2019). A Systematist's Guide to Estimating Bayesian Phylogenies From Morphological Data. *Insect Systematics and Diversity*, 3(3):2.

Wu, C.-Y., Manmatha, R., Smola, A. J., and Krähenbühl, P. (2017). Sampling matters in deep embedding learning. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2859–2867.

Wu, X., Zhan, C., Lai, Y., Cheng, M.-M., and Yang, J. (2019). Ip102: A large-scale benchmark dataset for insect pest recognition. In *IEEE CVPR*, pages 8787–8796.

Yang, Z. and Rannala, B. (2012). Molecular phylogenetics: Principles and practice. *Nature reviews. Genetics*, 13:303–14.

Zhang, C. (2022). [bug] states do not update using continuous characters with missing data. `https://github.com/revbayes/revbayes/issues/223`.

Zhang, C., Stadler, T., Klopfstein, S., Heath, T. A., and Ronquist, F. (2015). Total-Evidence Dating under the Fossilized Birth–Death Process. *Systematic Biology*, 65(2):228–249.

Żyła, D., Bogri, A., Hansen, A., Jenkins Shaw, J., Kypke, J., and Solodovnikov, A. (2022). A new termitophilous genus of paederinae rove beetles (coleoptera, staphylinidae) from the neotropics and its phylogenetic position. *Neotropical Entomology*.

Żyła, D. and Solodovnikov, A. (2020). Multilocus phylogeny defines a new classification of staphylininae (coleoptera, staphylinidae), a rove beetle group with high lineage diversity. *Systematic Entomology*, 45(1):114–127.

List of Figure Captions

5 What’s under the Likelihood?

> *Colon didn’t reply. I wish Captain Vimes were here, he thought. He wouldn’t have known what to do either, but he’s got a much better vocabulary to be baffled in.*
>
> — **Terry Pratchett**
> Guards! Guards!

After the work of the previous chapter, we can clearly identify a gap between the amount of phylogenetic signal the deep learnt morphological traits can extract and what can be inferred from genetic data. However, the cause of this gap is yet unknown. Do rove beetles simply have too many homologies to easily pick out the phylogenetic signal? Perhaps genetics will always give a stronger phylogenetic signal due to convergent evolution? Is there something more we should be doing in the deep learning model to extract traits with a higher signal?

To explore these questions we first choose a simulated dataset so we can rule out empirical dataset issues. Then we explore if enforcing some of the fundamental assumptions of phylogenetic inference could help the deep learning models extract meaningful traits. Finally we flip the problem, and ask ’if we give the model the hierarchy, could it learn to extract the most phylogenetically meaningful traits?’. If so, can these point us to phylogenetically important features in the original dataset?

This chapter stands apart from the previous two in two important ways: first, it is not yet peer reviewed or formed as a paper, but rather an exploration of prototypes. Secondly, it does not directly follow from the work with the Rove-Tree-11 dataset or morphological datasets. Rather we hope that eventually methods spawned from this area of research will be able to be applied to such a morphological dataset and improve automated morphological phylogenetic inference, however simulated genetic data gives us more control over assumptions and gives us certainty of the ground truth phylogenetic tree.

Code for experiments in this chapter is available on github (`https://github.com/robertahunt/PyTorch-VAE`). Code is forked from AntixK’s PyTorch-VAE framework [89].

5.1 Dataset

In order to identify if the issue is with the deep learning model, we need a dataset where we can directly compare phylogenetic inference results to our deep learnt trait extraction results. The easiest way to do this is using genetic data directly in our deep learning model. Unfortunately it is not feasible to obtain enough genetic data samples for the Rove-Tree-11 dataset that would allow deep learning from the genetic data alone. Therefore we started by looking at other empirical datasets which contain large numbers of image and associated genetic data. We presently know of two: INSECT [5] and BIOSCAN [30]. However the images in both are not from standard poses. Additionally neither contain a hierarchical representation aside from the taxonomy, and using a taxonomy as a phylogeny could conflate our problem. So finally we decide to use simulated genetic/phylogenetic data which gives us full certainty of the phylogeny and genetic sequence, at the potential cost of realistic sequences [95] and lack of specimen images and thereby lack of morphological traits.

Most simulated datasets enforce independence between traits, and one of our hypotheses is that deep learning models could be effective at disentanglement of dependant traits, leading to better inference overall. Thankfully some work has already been done on how the dependencies between traits affect the phylogenetic inference results using simulated data. Nasrallah et al. [63] used simulated data to show that dependencies between nucleotides negatively affected the phylogenetic results. Magee et al. [53] followed up on this in 2021 and showed that including dependant characters still improved phylogenetic results compared to removing them completely, however they are not as informative as independent characters.

Magee et al. [53] provide a repository for their code, so we choose to use their simulation setup to generate simulations with dependencies among nucleotides. We chose to base our study on their tunicates dataset which includes 50 species. The tunicates phylogenetic tree is shown in fig. 5.1. The tunicates tree is built from an empirical dataset of 18S genes from Tsagkogeorga et al. [96]. After the empirical tree is generated it is assumed to be ground truth, and traits extracted from it such as rate heterogeneity and stationary frequencies are used in the simulation step.

The **simulation model** used by us is directly from Magee et al. [53] which is based on that of Nasrallah and Huelsenbeck [62]. We briefly reiterate key points of their model here.

The model defines two kinds of sites: dependant and independent. Independent sites are modeled as evolving according to the classic GTR+G substitution model

[94]. The classic Generalized Time Reversible (GTR) model defines a symmetric rate matrix of transition probabilities from each nucleotide to another. It is symmetric in that it is assumed that the probability, for example, of A changing to G is the same as the probability of G changing to A. The +G means the model incorporates gamma distributed rate variation among sites. This simulates that some sites evolve slower than others. This is typically incorporated by discretizing the gamma distribution into a number of categories of rates, and assigning each site a category.

The dependant sites are modelled using a modified GTR+G model, where sites are paired together in doublets $\mathbf{x} = (x_1, x_2)$. The stationary frequencies of the doublets is defined as $\pi = (\pi_{AA}, \pi_{AC}, ...\pi_{TT})$. And the substitution matrix is denoted as $S \in R^{4x4}$. [62] define a set of Watson-Crick doublets, $W = \{AT, TA, CG, GC\}$. The probability of each doublet evolving into another doublet, $\mathbf{y} = (y_1, y_2)$ is defined by the instantaneous rate matrix $\mathbf{Q}$ with entries $\mathbf{Q_{x,y}}$ such that,

$$\mathbf{Q} = \begin{cases} \zeta \pi_y S_{x_1 y_1}, & \text{if single substitution of } x_1 \\ \zeta \pi_y S_{x_2 y_2}, & \text{if single substitution of } x_2 \\ \zeta \pi_y S_{x_1 y_1} S_{x_2 y_2} d, & \text{if double substitution and } \mathbf{x}, \mathbf{y} \in W \\ 0, & \text{if any other substitution} \\ -\sum_{x \neq y} \mathbf{Q_{x,y}}, & \text{if } \mathbf{x} = \mathbf{y} \end{cases} \tag{5.1}$$

Where ζ a scaling factor to ensure the rates stay overall similar between the independent sites and the dependant sites, and d is a scaling factor representing the overall dependency between doublets. The rate matrix and gamma shape parameter are shared between the independent and dependant sites.

5.1.1 Modelling Individuals

Normally in simulated phylogenetic studies, only one nucleotide sequence is generated per taxa. However, for the purposes of deep learning, you typically want many examples per species. In order to do this in the simulation software, and ensure that each specimen is related properly to the others in the tree, we expanded the tree so each species would have 200 leaves, each representing a single specimen. Another option would have been to run the normal simulation, and then run an intraspecies simulation for each species, generating mutations. However, this seems simpler and ensures that all mutation rates are shared along both the full tree and the individual specimens. We chose 200 specimens per species because based on experience this seems like a reasonable number of examples per class for deep learning classification algorithms. This means the total dataset has 10,000 sequences.

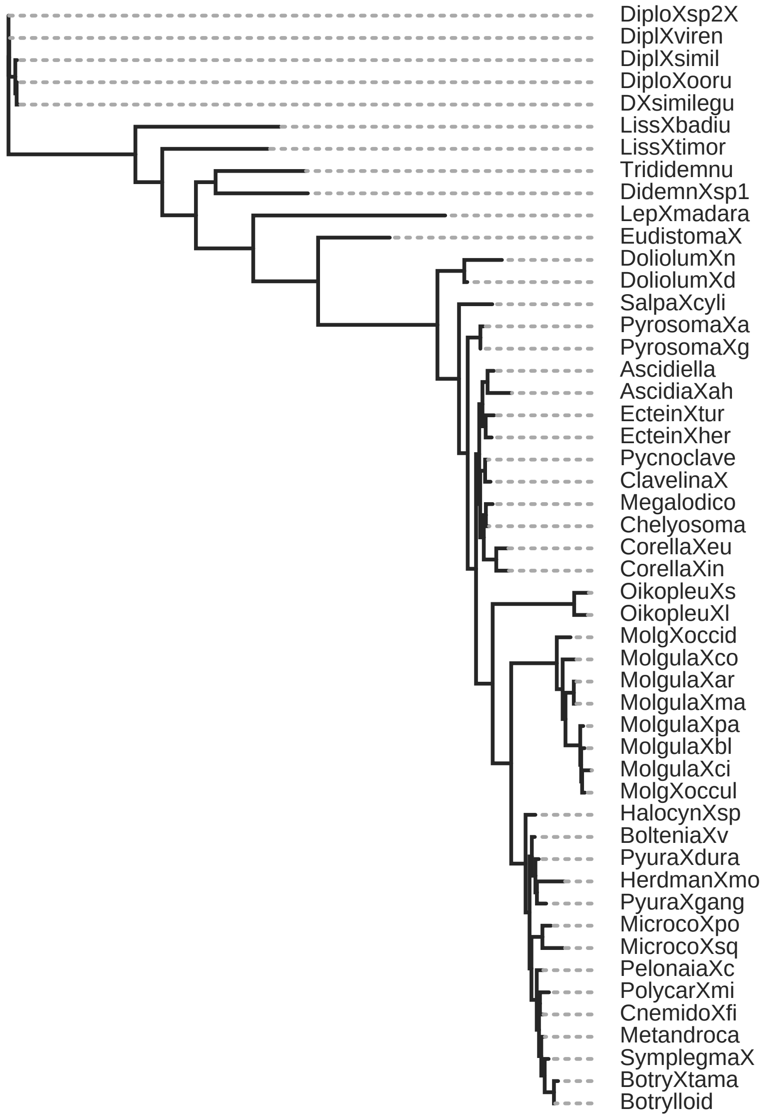


Fig. 5.1: Phylogenetic tree for tunicates dataset at species level [53]

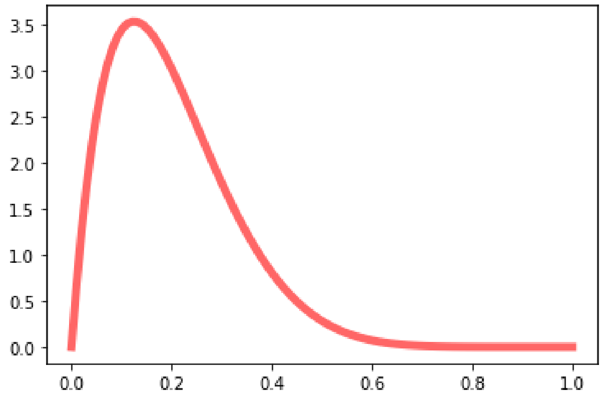


Fig. 5.2: PDF of beta distribution with $\alpha = 2, \beta = 8$

To add the branch lengthswefirst halved the species branch length to account for the specimen brancheswewould add.wethen sampled the specimen branch lengths from a beta distribution ($\alpha = 2, \beta = 8$) multiplied by half the species branch length. A plot of the distribution is shown for reference in fig. 5.2. The beta distribution seems like a reasonable choice as it always falls between 0 and 1, and with the chosen parameters is biased towards low values, representing little mutation per specimen. The resulting tree is provided as a nex file in the supplemental data.

For this study we chose to use made the following hyperpameter choices, and left the rest as the default from the original study [53]:

1. Number of independent characters: 200
2. Number of dependant characters: 200
3. Strength of dependency between traits (d variable in [53]): 1000

Hyperparameter Choices Magee et al.'s implementation [53] allows for specifying the level of dependency between traits and the number of traits. We chose the strength of the dependency to be high. For reference, in [53] they describe $d = 1000$ as an extreme case of epistatic dependency, and report that their empirical datasets had measured dependency values, d, between 0.5 and 8. We chose to use an extreme value in order to more easily see if our hypotheses have merit. We chose to use a 70/15/15 train/validation/test split.

5.2 Model Architecture and Preprocessing

Now that we have a dataset where we can rule out empirical errors, we can begin to explore how to improve the deep learning methodology. To do this, we first must choose a deep learning architecture. Since we are working with genetic data, it does not make sense to use the large image models from our previous work. Instead, we look at existing genetic models. Marin et al. [54] provide a good overview of these for DNA sequences and Gao et al. [27] provide a good overview for protein models. We want something relatively simple so we can quickly prototype the models, but it must be complex enough to be able to capture the dependencies between pairs in the sequence. We also know that we want to be able to measure the model's ability to capture the information contained in the dataset, which already leads us to some autoencoder based reconstruction loss. We further know that variational autoencoders have some appeal due to the Gaussian nature of Brownian motion, which the simpler continuous trait models are based on. Therefore, we decide to use the architecture from Riesselman et al. [72]. Riesselman et al. [72] show that this architecture is effective at predicting mutation rates in empirical protein sequence datasets, so therefore we expect it would be similarly effective at nucleotide substitutions. It is also a relatively simple autoencoder architecture, as shown in fig. 5.3. To preprocess the dataset we chose to use one-hot encoding, as was also used in Riesselman et al. [72]. This is fairly standard for dna sequences, although other methods do exist [58]. Models were trained for 200 epochs without early stopping. No data augmentation was used.

Autoencoders are a kind of unsupervised deep metric learning model where the goal is to learn a latent representation of the dataset by making the model learn to reconstruct each input datapoint. There are many variants of reconstruction loss functions. In this work we use L2 loss, described below.

Given an input datapoint $\mathbf{X_i}$ and a reconstruction produced by the model, $\mathbf{Y_i}$, and a batch size of B, the L2 reconstruction loss is

$$\text{Reconstruction Loss} = \frac{\sum_{i=0}^{B}(\mathbf{X_i} - \mathbf{Y_i})^2}{B}. \tag{5.2}$$

5.3 Phylogenetic Probability Model

Before looking at the methodology, we first review our understanding of the probability model behind the generative trait process. This is based partially on Felsenstein [19], and partially our own ideas.

5.3.1 The Generative Process

We assume that τ, the tree topology, q, the branch lengths, $\mathbf{X}$, the input data and $\mathbf{Z}$, the latent representations, or independent traits per species, are involved in the

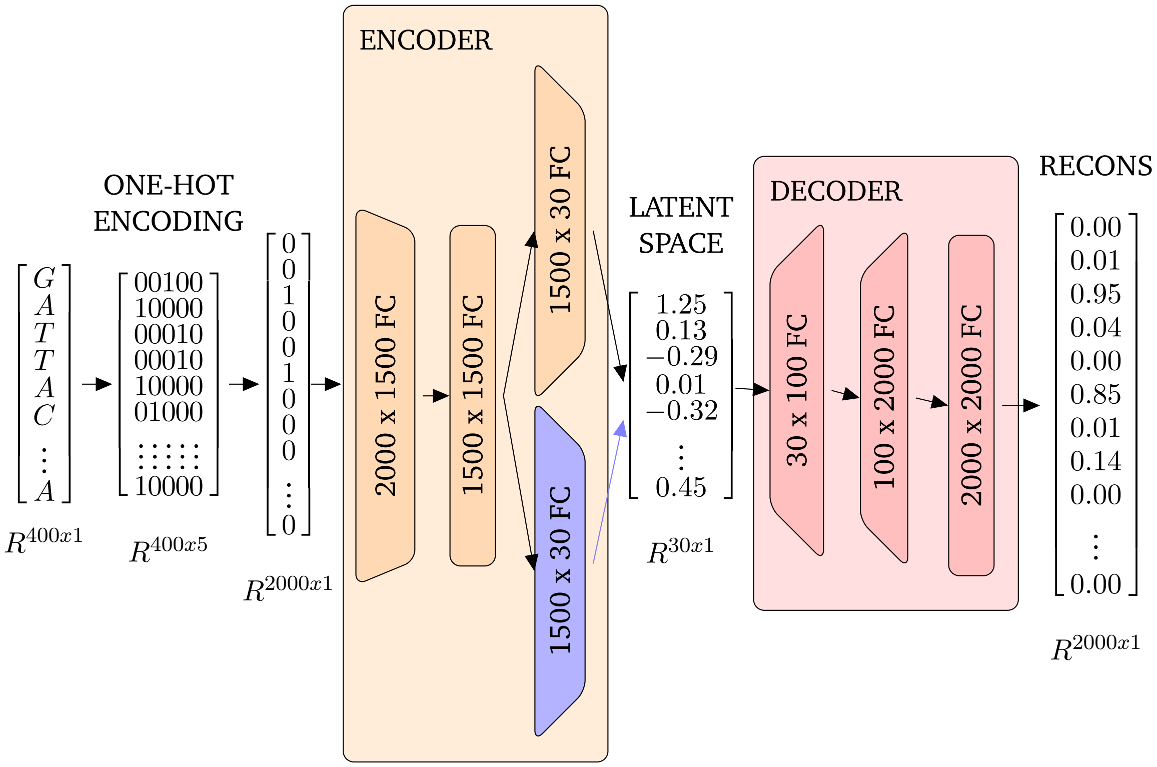


Fig. 5.3: The Deep Sequence architecture based on [72], including preprocessing steps. First the original input sequence of nucleotides is one-hot encoded, where each nucleotide (A,C,G,T) is mapped to a column, with the fifth column reserved for gaps. The input sequence is then flattened, and fed into the encoder (highlighted in orange). If a variational autoencoder is used, then both 1500x30 layers are used, the top in orange is used to produce the mean vector, and the bottom in blue is used to produce the variance vector. These two are then combined using the reparameterization trick to produce the latent vector. If a variational autoencoder is not used, then the blue steps do not occur, and the mean vector is used as the latent space directly. Finally, if reconstruction loss is used, the latent vector is then fed into the decoder (pink) to produce the reconstruction. FC stands for Fully Connected.

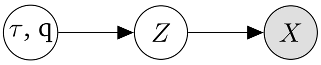


Fig. 5.4: Generative phylogenetic process. The ancestral topology τ and branching process q dictate the distribution of traits in each species, The individual specimens in each species are then assumed to be independantly sampled from the traits $\mathbf{Z}$, a non-linear process then transforms the independent traits $\mathbf{Z}$ into $\mathbf{X}$. White nodes refer to unobserved variables, grey nodes refer to observed variables.

generative process given by fig. 5.4. If we assume independent Brownian motion is responsible for the evolution process [20], then we can say that the distribution of the traits per species, $\mathbf{Z}$, is based on the tree topology τ and the branch lengths q and that $\mathbf{Z} \sim \mathcal{N}(\mu, \sigma^2)$ [20] where μ and σ^2 are functions of the topology τ and the branch lengths q. We can then say that there is some non-linear process which maps the independently evolved traits $\mathbf{Z}$ to the observed phenotypes / genotypes $\mathbf{X}$.

5.3.2 Probability Model

In phylogenetic inference we are often interested in modelling $p(\tau, q, \mathbf{X})$, the probability of a given tree topology, τ, with branch lengths, q and traits $\mathbf{X}$. Traditionally, in order to make the inference computationally feasible the phenotypes/genotypes $\mathbf{X}$ are assumed to be independent and the number of samples per species in $\mathbf{X}$ is assumed to be 1 ($S{=}1$), which is typically a consensus sequence in the case of genetic data, or the morphological traits of a type specimen for morphological characters. This greatly reduces the complexity of the problem and allows us to use the product rule for independent events to break the problem into the optimization of each trait, $\mathbf{X_i}$ individually, and then multiplying the result, such that, $p(\tau, q, \mathbf{X}) = \prod_{i=1}^{M} p(\tau, q, \mathbf{X_i})$. In order to complete the inference, this can be further factored using the definition of conditional probability

$$p(\tau, q, \mathbf{X}) = \prod_{i=1}^{M} p(\mathbf{X_i} | \tau, q) p(\tau, q) \tag{5.3}$$

.

However, we know that both morphological and genetic traits are not independant [63], and additionally that each species has a distribution of phenotypes / genotypes. To account for this we propose $\mathbf{Z}$, which gives the distribution of the hidden independent traits over each species. $\mathbf{Z}$ can then be mapped to $\mathbf{X}$ through some non-linear process allowing $\mathbf{X}$ to be dependent.

This change means we are now interested in modelling $p(\tau, q, \mathbf{X}, \mathbf{Z})$. Using the definition of conditional probability we can repetitively factor this to obtain

$$p(\tau, q, \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}|\mathbf{Z}, \tau, q)p(\mathbf{Z}|\tau, q)p(\tau, q) \tag{5.4}$$

which is consistent with the generative model in fig. 5.4.

Now, given that we assume these variables follow the generative process shown in fig. 5.4, we can say that $\mathbf{X}$ is conditionally independent from τ, q given $\mathbf{Z}$ and $\mathbf{Z}$ is conditionally independent from τ, q given $\mathbf{X}$. Equation (5.4) then simplifies to

$$p(\tau, q, \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z}|\tau, q)p(\tau, q) \tag{5.5}$$

.

Now, we can choose a model for each of the probabilities above as follows:

- $p(\tau, q)$ Since we have no information about which trees might be favourable, it seems natural to assume all topologies and branch lengths are equally likely, although this has been shown to produce some issues when assuming uniform priors on branch lengths [107].

- $p(\mathbf{Z}|\tau, q)$ If we assume Brownian motion, then it is natural to assume that $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ This is the typical optimization likelihood in phylogenetic inference and has been well studied. For simplicity we would like to use continuous traits and can therefore use Felsenstein's Brownian motion likelihood [20] (see equation (5.14))

- $p(\mathbf{X}|\mathbf{Z})$ We can assume that this is a generative neural network that can map from $\mathbf{Z}$ to $\mathbf{X}$

5.4 Methodology

With the dataset, architecture and a probability model, we are now ready to begin experimenting. Below we describe the methodology behind each of the models.

5.4.1 Inference Methods

The nucleotide inference used for the baseline model is completed using the analysis scripts provided by Magee et al. [53]. These set priors on the stationary frequencies and exchangeability rates. Together these specify the rate matrix, $\mathbf{Q}$ (see section 5.1). Site rates are modelled with a gamma distribution. The model is run twice, each with a burnin of 2,000 generations, and MCMC is run for 20,000 generations in total. Every 10th generation is logged, and the results used to form a probability distribution over the trees.

For the deep learnt traits, trees are inferred from the continuous traits using the RevBayes software following Parins-Fukuchi [66]. This could present some difficulties since said script was meant to be applied to continuous morphological traits. However, we compare the results to simple neighbour joining of average training data species traits to explore this. We also modify the Revbayes scripts to the same amount of runs, burnin and generations as the baseline model so the tree results are directly comparable.

5.4.2 Baseline Model - Direct Inference

We start with setting the baseline to compare our results against. To do this, we can use the analysis scripts from Magee et al. [53] described in section 5.4.1 to directly perform the inference on the DNA in Revbayes [38]. However, we have 200 sequences per species in our new dataset, giving us 10,000 sequences in total, which is computationally infeasible to run. Therefore we follow the standard procedure in biology and obtain consensus sequences for each of the 50 species instead. Concensus sequences were made using the ConcensusSequence function in the DECIPHER R package [104] with a threshold of 0.05 and a minInformation of 0.75. To calculate the distribution of RF scores from the trees we use the analysis scripts provided by [53].

5.4.3 Triplet Loss

To compare any of the below methods with the methods from our previous papers, we use the triplet loss. Triplet loss [15] is a relatively simple loss function that provided reasonable results in the Rove-Tree-11 paper. Triplet loss takes triplets of inputs from the batch: an anchor ($\mathbf{Z_a}$), a positive example ($\mathbf{Z_p}$) from the same species as the anchor, and a negative example ($\mathbf{Z_n}$) from a different species from the anchor. A simple way to think about triplet loss is as pull the positive example towards the anchor and push negative examples away. Mathematically it is defined as:

$$\text{Triplet Loss} = max(d(\mathbf{Z}_a, \mathbf{Z}_p) - d(\mathbf{Z}_a, \mathbf{Z}_n) + \alpha, 0) \tag{5.6}$$

Where d is a distance measure (Euclidean in this case), α is a margin representing how far the algorithm should push the clusters apart.

5.4.4 Hierarchical Triplet Loss

Hierarchical triplet loss was introduced by Ge et al. [29] as a method of learning hierarchical relationships from datasets. Their loss function is almost the same as that of normal triplet loss, except the margin changes between triplets depending on the distance along the hierarchy between datapoints. In Ge et al. [29] they learn the hierarchy dynamically from the data, however, since we would like to learn from the hierarchy, we attempt to use the loss directly.

Given a triplet of three data points, an anchor ($\mathbf{Z_a}$), a positive sample ($\mathbf{Z_p}$) and a negative sample ($\mathbf{Z_n}$), where if c_i represents the class label of datapoint $\mathbf{Z}_i$, $c_a = c_p \neq c_n$ and a distance function d which calculates the distance between two latent points, and α, which represents the triplet margin, the hierarchical triplet loss can be defined the same as in equation (5.6), except that the margin α is variable and defined to be the distance (sum of branch lengths) along the tree from c_a to c_n.

5.4.5 Disentanglement with Variational Autoencoders

Most phylogenetic inference methods rely on an assumption of independence between the characters / traits (either morphological traits, or genetic basepairs). This assumption significantly simplifies the calculation of the likelihood and makes it feasible with iterative methods. However, there are non-phylogenetically related dependencies between characters, particularly in genetics. The evolution of exons (genes which code for protein sequences) for example, is influenced by the protein folding structure which can produce dependencies across the entire gene [87]. Nasrallah et al. [63] showed that this can significantly affect the resulting phylogenetic inference if not handled. Many methods for detecting these dependencies have been explored [81, 62, 53]. However, we do not know of any methods, deep learning or otherwise for removing such dependencies. We suspect this is either due to a lack of searching with the correct jargon on my part, or because it is impossible to separate the non-phylogenetic dependencies from the phylogenetic ones without first knowing the phylogeny, which of course we will never be 100% certain about.

Brownian Motion Assumption: Many inference methods for continuous traits assume the evolution can be accurately approximated with Brownian motion. This has the attractive property that the resulting probability distribution over the traits is therefore Gaussian, which has some simplifying properties for the likelihood calculation, as shown by Felsenstein [20]. One trouble with deep metric learning models is that the distribution of datapoints in the latent space is typically not Gaussian by default, and in fact can lie on an exceedingly complex manifold Shao et al. [80].

It should be noted that Blomberg et al. [9] propose another method after finding issues with Brownian Motion models and the alternatively popular Ornstein-Uhlenbeck Process Models introduced in Felsenstein [22], we have not yet explored other more complex methods.

Deep Learning Model: Disentangled networks (networks which attempt to enforce independence between traits) present a growing body of scientific research in the deep learning community. A relevant survey paper of research can be found in Kaya and Bilge [41]. Variational autoencoders are one such method which we hypothesize have potential in this context since they both push for disentanglement of the traits and a Gaussian distribution in the latent space.

In order to see if variational autoencoders can disentangle the traits well enough, we compare results of trees obtained from the latent variables of a simple autoencoder architecture and a variational autoencoder architecture. This architecture is based on that used in Riesselman et al. [72] which was used for encoding protein sequences. The architectures are shown in fig. 5.3.

A thorough introduction to variational autoencoders is given in Kingma, Welling, et al. [43]. Here we briefly explain some main concepts used in this work.

VAEs are autoencoders (see section 5.2 for a brief intro to autoencoders) which define a prior, typically Gaussian, on the latent space. Instead of directly encoding the datapoint, the encoder learns to predict the parameters of a distribution for that datapoint, which are then sampled through the reparameterization trick and fed into the encoder. How well this overall distribution matches the prior distribution is calculated through the Kullback–Leibler (KL) divergence loss function, which effectively pulls the latent space towards a specified distribution. VAEs also typically have a reconstruction loss element, which pulls the latent space to represent the data.

For our variational autoencoder we chose to weight the KL divergence by 0.01 relative to the reconstruction loss (with a weight of 1 the model did not converge).

The KL divergence we use, between a Gaussian prior $p(x) \sim N(0,1)$ and a Gaussian posterior with diagonal covariance matrix $q(x) \sim N(\boldsymbol{\mu}_q, \boldsymbol{\sigma_q})$ is defined as:

$$\text{KL}(q||p) = \frac{1}{2}[\boldsymbol{\mu}_{\boldsymbol{q}}^{\boldsymbol{T}} \boldsymbol{\mu}_{\boldsymbol{q}} + \sum \boldsymbol{\sigma}_{\boldsymbol{q}} - k - log(\prod \boldsymbol{\sigma}_{\boldsymbol{q}})] \tag{5.7}$$

5.4.6 Blomberg's K

Blomberg's K (Blomberg et al. [8]) is a measure of phylogenetic signal. It compares the observed mean squared error and the expected mean squared error of continuous traits based on Brownian motion. A K value of 1 indicates they match perfectly, K values between 0 and 1 indicate lower phylogenetic signal, and above one indicates higher than expected phylogenetic signal.

Blomberg's K is described as the difference between the ratio of the observed mean squared error ratio, and the expected mean squared error ratio based on Brownian motion. It is calculated for a single trait (i) at a time. (Equations taken from [8])

$$K_i = \frac{MSER_{observed,i}}{MSER_{expected}} \tag{5.8}$$

$$MSER_{observed,i} = \frac{\frac{(\mathbf{X_{:i}}-\hat{\mu})^T(\mathbf{X_{:i}}-\hat{\mu})}{N-1}}{\frac{(\mathbf{X_{:i}}-\hat{\mu})^T\mathbf{V}^{-1}(\mathbf{X_{:i}}-\hat{\mu})}{N-1}} = \frac{(\mathbf{X_{:i}}-\hat{\mu})^T(\mathbf{X_{:i}}-\hat{\mu})}{(\mathbf{X_{:i}}-\hat{\mu})^T\mathbf{V}^{-1}(\mathbf{X_{:i}}-\hat{\mu})} \tag{5.9}$$

$$MSER_{expected} = \left(\frac{1}{N-1}\right)\left(tr(\mathbf{V}) - \frac{N}{\Sigma\Sigma\mathbf{V^{-1}}}\right) \tag{5.10}$$

Where $\mathbf{X_{:i}}$ is a vector of values of trait i for all taxa, N is the total number of taxa, $\mathbf{V} \in \mathbb{R}^{NxN}$ is the covariance matrix of the tree, specified from the branch lengths of the tree topology such that $\mathbf{V_{13}}$ is the length along the tree from the root to the latest common ancestor of species 1 and species 3 (see Felsenstein [20]) is the variance-covariance matrix generated from the tree topology, $tr()$ represents the trace function of a matrix, and $\hat{\boldsymbol{\mu}}$ is the phylogenetically corrected mean. The phylogenetically corrected mean can either be calculated iteratively using the independent contrasts methods [20] or through the following method, whichwecould not find an analytical reference for but is implemented in phylosig in the phytools package of R [71] as

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum \mathbf{V}^{-1}\mathbf{X}_i}{\sum \mathbf{V}^{-1}} \tag{5.11}$$

.

Importantly, the K value itself is not enough to prove phylogenetic signal, as Blomberg pointed out in Blomberg et al. [8]. A measure of significance has to be calculated for each trait to ensure the phylogenetic signal is not just due to randomness. To do this, the traits are randomly arranged on the tree a number of times, and the percentage of times the arrangement results in a K value higher than the calculated value, gives our p value. A typically assumption of 0.05 significance is used.

One way to formulate the above functions into loss functions is to attempt to minimize the mean squared error between the calculated loss and the expected loss:

$$\text{Blomberg K Loss} = \frac{\sum_{i=0}^{L}(MSER_{observed,i} - MSER_{expected})^2}{L} \tag{5.12}$$

This attempts to push the observed variance as close as possible to the expected variance, and since it is squared, penalizes large deviances. Another potential way to formulate the loss would be as $\max(1 - K, 0)$, however the extra penalty to large deviances is appealing.

equation (5.12), however, does not account for the test of significance. We can attempt to incorporate this using a shuffled version of the observed ratio, where the traits are shuffled relative to the tree, $MSER_{shuffled,i}$ as follows:

$$\begin{aligned}\text{Blomberg K Shuffled Loss} &= \frac{\sum_{i=0}^{L}(MSER_{observed,i} - MSER_{expected})^2}{L} + \\ &\frac{\sum_{i=0}^{L} max((MSER_{shuffled,i} - MSER_{observed}, 0))}{L}\end{aligned} \tag{5.13}$$

where L is the number of latent variables.

This should encourage significance of the resulting traits.

5.4.7 Optimizing the Likelihood

Felsenstein [20] calculated the likelihood of continuous traits given a tree and used this to infer a tree from the already extracted traits. We are interested in exploring if the likelihood could be directly optimized in a deep learning setting. Therefore, we want to explore if we can optimize the traits given a known tree. This is not our initial application, but it does have some merit. Imagine we have a good reference tree for a dataset, however, we wish to know which traits are most highly correlated with that tree. Statistical methods to do this do already exist, such as Blomberg's K described above, but they assume that the independent traits are already extracted, and they assume a small number of input variables so statistical significance is easy to prove. If we have an image, for example, with thousands of pixels, or a long nucleotide sequence with thousands of base-pairs, it becomes difficult for statistical methods to prove significance, although this can be accounted for using, for example, the Bonferroni correction [17].

However, this is where we believe deep learning methods can be highly effective. They have already been shown to be very successful at information compression, and indeed this is one of the major applications of the field of deep metric learning. Therefore the goal of this experiment is to determine how we can use deep learning to extract the most phylogenetically relevant traits from images or DNA sequences given a known tree.

Felsenstein [20] defines the probability of a single continuous trait evolving according to Brownian motion given a tree topology as,

$$p(\mathbf{Z_{:i}}|\mathbf{V},\hat{\mu},\sigma^2) = \frac{1}{(2\pi)^{\frac{N}{2}}|\sigma^2\mathbf{V}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{Z_{:i}}-\hat{\boldsymbol{\mu}}_i)^T(\sigma^2\mathbf{V}^{-1})(\mathbf{Z_{:i}}-\hat{\boldsymbol{\mu}}_i)} \tag{5.14}$$

.

where $\mathbf{Z_{:i}} \in \mathbb{R}^N$ is the column of the species trait matrix $\mathbf{Z} \in \mathbb{R}^{NxL}$ which contains trait i, $\mathbf{V} \in \mathbb{R}^{NxN}$ is the variance-covariance matrix generated from the tree topology described in section 5.4.6, σ^2 is the variance of the trait, N is the total number of species / leaves, L is the number of latent variables, and $\hat{\boldsymbol{\mu}}$ is the phylogenetically corrected mean as described in section 5.4.6.

This equation is essentially, $p(\mathbf{Z_{:i}}|\tau, q)$ in our probability model from section 5.3.2.

Given that the traits are assumed to be independent and that all traits have the same variance Felsenstein then uses the definition of independent probabilities to obtain

$$p(\mathbf{Z}|\mathbf{V}, \hat{\boldsymbol{\mu}}, \sigma^2) = \prod_{i}^{L} p(\mathbf{Z}_{:\mathbf{i}}|\mathbf{V}, \hat{\mu}, \sigma^2) = \frac{1}{(2\pi)^{\frac{NL}{2}} |\sigma^2 \mathbf{V}|^{\frac{L}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{L} (\mathbf{Z}_{:\mathbf{i}} - \hat{\boldsymbol{\mu}}_i)^T (\sigma \mathbf{V}^{-1})(\mathbf{Z}_{:\mathbf{i}} - \hat{\boldsymbol{\mu}}_i)} \tag{5.15}$$

.

If we assume each trait has its own variance, we can pull σ out of the determinant for each individual trait in equation (5.14) using the property that $|c\mathbf{A}| = c^N |\mathbf{A}|$ given $\mathbf{A} \in R^{NxN}$ [67] this equation instead becomes

$$p(\mathbf{Z}|\mathbf{V}, \hat{\boldsymbol{\mu}}, \boldsymbol{\sigma}^2) = \frac{1}{(2\pi)^{\frac{NL}{2}} |\mathbf{V}|^{\frac{L}{2}} \prod_{i}^{L} \boldsymbol{\sigma}_i^N} e^{-\frac{1}{2}\sum_{i=1}^{L} (\mathbf{Z}_{:\mathbf{i}} - \hat{\boldsymbol{\mu}}_i)^T (\boldsymbol{\sigma}_i^2 \mathbf{V}^{-1})(\mathbf{Z}_{:\mathbf{i}} - \hat{\boldsymbol{\mu}}_i)} \tag{5.16}$$

.

where $\boldsymbol{\sigma}^2$ is now a L length vector of the variances for each trait.

Felsenstein treated equation (5.15) as the likelihood function $f(\tau, q) = p(\mathbf{Z}|\tau, \mathbf{q})$ of the tree given the traits, however, since we want to optimize the traits, we can directly use the probability from equation (5.15). It seems advantageous to assume trait independent variances, so taking the log of equation (5.16) as the log probability is typically easier to optimize and provides the same maxima we get the following

$$\begin{aligned} log(p(\mathbf{Z}|\mathbf{V}, \hat{\boldsymbol{\mu}}, \boldsymbol{\sigma}^2)) = -\frac{NL}{2} log(2\pi) - \frac{L}{2} log(|\mathbf{V}|) - N \sum_{i=1}^{L} log(\boldsymbol{\sigma}_i) \\ - \frac{1}{2} \sum_{i=1}^{L} (\mathbf{Z}_{:\mathbf{i}} - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{\sigma}_i^2 \mathbf{V}^{-1})(\mathbf{Z}_{:\mathbf{i}} - \hat{\boldsymbol{\mu}}_i)^T \end{aligned} \tag{5.17}$$

.

We are interested in obtaining the traits with maximum probability (ie the maxima of this function), and in order to optimize with gradient descent, we then take the negative of equation (5.17):

$$\begin{aligned} -log(p(\mathbf{Z}|\mathbf{V}, \hat{\boldsymbol{\mu}}, \boldsymbol{\sigma}^2)) = \frac{NL}{2} log(2\pi) + \frac{L}{2} log(|\mathbf{V}|) + N \sum_{i=1}^{L} log(\boldsymbol{\sigma}_i) \\ + \frac{1}{2} \sum_{i=1}^{L} (\mathbf{Z}_{:\mathbf{i}} - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{\sigma}_i^2 \mathbf{V}^{-1})(\mathbf{Z}_{:\mathbf{i}} - \hat{\boldsymbol{\mu}}_i)^T \end{aligned} \tag{5.18}$$

Now we are almost ready to use equation (5.18) directly as a loss function in a deep learning algorithm.

We also note that we can attempt to specify the phylogenetically corrected mean, μ and the variance σ^2 to 0 and 1, respectively, giving:

$$- \, log(p(\mathbf{Z}|\mathbf{V}, \hat{\boldsymbol{\mu}}, \boldsymbol{\sigma}^2)) \;=\; \frac{NL}{2} log(2\pi) + \frac{L}{2} log(|\mathbf{V}|) + \frac{1}{2}\sum_{i=1}^{L} \mathbf{Z_{:i}} \mathbf{V}^{-1} \mathbf{Z_{:i}}^{T} \tag{5.19}$$

We first explore how this affects the results with the reconstruction loss directly, and then apply the best model to the following experiments. This is denoted as 'standard likelihood' in our results.

However, if we look at it we can see that if the trait matrix $\mathbf{Z}$ collapses to 0 this will give a trivial minimum. Therefore we need another term in our loss function to counteract this. The reconstruction loss from above is a clear choice, however there are others that come to mind to counteract the trivial solution of equation (5.18):

Variance Loss: This attempts to push the variance for each trait to be 1

$$\text{Variance Loss} = \frac{\sum_{i=1}^{N} (\sigma_i^2 - 1)^2}{N} \tag{5.20}$$

.

MSE $\hat{\mu}$ loss: This attempts to push the variance for each trait to be 1 and $\hat{\mu}$ to be 0.

$$\text{MSE } \hat{\mu} = \frac{\sum_{i=1}^{N} \hat{\mu}_i^2}{N} \tag{5.21}$$

Covariance Loss: This attempts to push the variance-covariance matrix to be closer to the expected covariance matrix $\mathbf{V}$,

$$\text{Covariance Loss} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (\boldsymbol{\Sigma}_{ij} - \mathbf{V}_{ij})^2}{N^2} \tag{5.22}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix calculated from the trait data $\boldsymbol{\Sigma} = \frac{(\mathbf{Z} - \hat{\boldsymbol{\mu}})^T (\mathbf{Z} - \hat{\boldsymbol{\mu}})}{N}$

5.4.8 Random Trees

Random tree results are shown in tab. 5.2. We include the random tree results to compare the deep learning models against as a way to show they are learning phylogenetically significant traits. These are generated the same way as in Chapters 3 and 4, by randomly generating a Nx2 trait matrix, then using single linkage neighbour joining to generate a hierarchy. For the results in this chapter we used 1000 random initializations to show significance.

5.5 Results

We train the deep learning model from 5.3 using the loss functions described above using the synthetic dataset from section 5.1. The results are shown in tab. 5.2 and tab. 5.1.

First we can see that none of the deep learnt models outperform direct genetic inference, despite some being given direct information about the hierarchy of interest. Second we can see that all models outperform even the best random tree after 1000 initializations (RF of 88). Next we see that of the different deep learning variations which use the hierarchical information, our likelihood-loss model appears to perform the best, and the Blomberg-K models appear to perform the worst. Finally we can see that the models which use the hierarchical information, while some outperform the non-hierarchical deep learnt traits methods, models which do not use the hierarchy still provide comparable, or better results. Using only reconstruction loss provides an average RF score of 57. Using only triplet and reconstruction losses gives us the second best deep learning model overall.

Looking at how the Bayesian inference results compare to neighbour joining using the average species traits, we can see that there is not a large difference, suggesting that the inference methodology itself is not the main issue here.

Of course the results in 5.2 should be taken lightly. They are result of only a single deep learning model initialization.

In tab. 5.1 we can see that the reconstruction accuracies are quite high (all above 98.9% for the validation set). To give a baseline comparison, constructing a sequence of the most frequent nucleotides at each loci gives a top-1 accuracy of 86.2%, demonstrating that the models are able to model the data.

	Top 1 Accuracy ↑	
Loss Function	Train	Val
Triplet (5.6) and Reconstruction (5.2)	99.6%	99.4%
Hierarchical Triplet (5.6) and Reconstruction (5.2)	99.3%	99.2%
Blomberg K w P (5.13) and Reconstruction (5.2)	98.9%	98.9%
Covariance (5.22) & Reconstruction (5.2)	99.3%	99.3%

Tab. 5.1: Top-1 accuracy of nucleotide reconstructions for different models.

			Bayesian Inference		NJ
Model	Uses Hierarchy?	Loss Functions	RF (min) ↓	RF (avg) ↓	RF ↓
Direct Inference	No	-	**8**	**23**	-
Autoencoder	No	Recons. (5.2)	44	56	64
VAE	No	KL Divergence (5.7) & Reconstruction (5.2)	52	57	64
Encoder	No	Triplet (5.6)	46	58	60
Autoencoder	No	Triplet (5.6) and Reconstruction (5.2)	32	44	52
Encoder	Yes	Hierarchical Triplet (5.6)	48	57	60
Autoencoder	Yes	Hierarchical Triplet (5.6) and Reconstruction (5.2)	50	59	64
Encoder	Yes	Blomberg K (5.12)	82	83	86
Encoder	Yes	Blomberg K Shuffled (5.13)	72	76	78
Autoencoder	Yes	Blomberg K Shuffled (5.13) and Reconstruction (5.2)	82	84	84
Encoder	Yes	Covariance (5.22)	48	55	60
AutoEncoder	Yes	Covariance (5.22) & Reconstruction (5.2)	32	43	48
Autoencoder	Yes	Likelihood (5.18) & Reconstruction (5.2)	28	52	58
Encoder	Yes	Likelihood (5.18) & Covariance (5.22)	44	55	60
Autoencoder	Yes	Standard Likelihood (5.19) & Reconstruction (5.2)	36	49	56
Encoder	Yes	Standard Likelihood (5.19) & Covariance (5.22)	48	55	52
Encoder	Yes	Standard Likelihood (5.19) & Variance (5.20)	52	57	54
Encoder	Yes	Standard Likelihood (5.19) & Variance (5.20) & MSE Mean (5.21)	66	67	64

Tab. 5.2: Results of various loss functions on phylogenetic inference of deep learning traits. Best models highlighted in bold. Second best models underlined. NJ is short for Neighbour Joining. For comparison randomly generated trees had a minimum RF of 88 and an average RF of 93 over 1000 random initializations.

5.6 Discussion

5.6.1 Direct Inference vs Deep Learning

Perhaps the most surprising result from this chapter is the large gap between the direct inference results and the deep learning results. The best tree from all the experiments, including those taking the hierarchy into account (RF of 28) does not improve upon the average direct inference model which gives an RF of 23.

Since deep learning models have proven that they are able to learn complex relationships and interactions such as substitution rates, transition matrices, mutation rates [72], it seems that we are missing some key elements which would push the models to incorporate these factors. That said, it is also hard to understand this, as we are directly using an architecture which worked well for predicting mutation rates in protein sequences [72]. This leads me to believe that either, 1) the models for learning the hierarchical relationships are simply not strong enough, or 2) there are some important assumptions of priors made in the inference models which are not reflected in the deep learning models, which requires a deeper understanding of the inference models and underlying assumptions for genetic data or 3) to have some healthy skepticism of my implementation. Testing this will require further experimentation to rule out experimental error, further work with existing hierarchical models and comparison with external results and further understanding of the genetic inference models which underline the phylogenetic inference.

5.6.2 Random Trees vs Deep Learning

Similar to the results of previous chapters, we can see that the deep learning models do still learn some of the phylogenetic relationships. Even with 1000 random initializations, the average tree RF score from all deep learning models tested still beats the best randomly initialized tree. This suggests that all of the deep learning models are able to learn some of the hierarchical relationships and push the latent space to represent these.

5.6.3 Bayesian Inference vs Neighbour Joining

If we compare the Bayesian inference results to the neighbour joining results we can see that they are similar (within 8 of the average inference result). This suggests that the main issue is not a major error in the Bayesian continuous traits inference, although it could be that a different kind of model, like the Ornstein–Uhlenbeck process would improve the results. Knowing that a perfect distance matrix will produce a perfect topology with neighbour joining leads me to believe that the error in these models is not simply an issue with the inference parameters.

5.6.4 Triplet vs Hierarchical Triplet

Surprisingly the Triplet and Hierarchical Triplet models show similar performance, despite the triplet model only being given information of which sequences are from the same species, and the hierarchical triplet model being given the full hierarchy. We suspect a grid search of the multiplying factor applied to the distances could provide better results. Other work with triplet losses has also shown that the choice of batch miner can greatly affect the results [105], so perhaps a more informed choice here would improve the results.

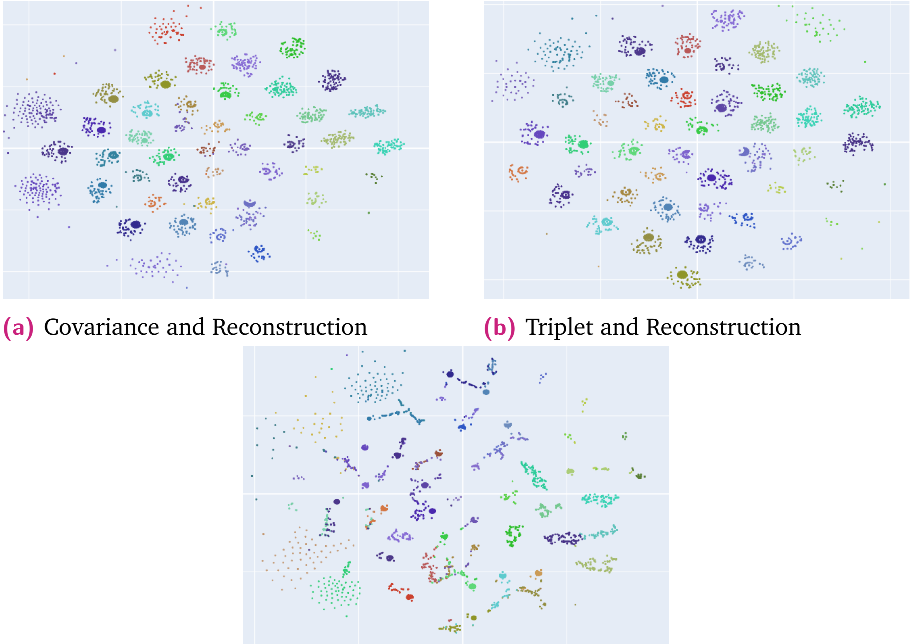
5.6.5 Architecture - Latent Space Size

One possible explanation could be that the latent space size of 30 variables is simply too small to encapsulate the variation in this data and the hierarchical relationships in 50 species. To explore this, we can first look at the accuracies of the reconstructions in tab. 5.1 which are all above 98.9%, suggesting that the models are able to learn most of the variation in this dataset. Next we can look at a t-SNE representation of the latent variables to see how they are distributed. For this we look both at the VAE latent variables shown in fig. 5.5, as the KL divergence directly affects the latent distribution, the triplet reconstruction model, as this performed best of the unsupervised models, and the covariance & reconstruction model, as this performed best of the supervised models.

From fig. 5.5 we can see that all models cluster the data well, which is not too surprising. Further we can see the triplet loss appears to keep the distances between groups rather consistent as expected. The variational autoencoder stretches the clusters into thin elongated structures in the t-SNE representation, most likely to fit them closer to a Gaussian distribution in $\mathbb{R}^{30x1}$, as expected.

5.7 Conclusion

In this work we have explored various deep learning methods for directly optimizing the hierarchical output of the deep learnt phylogenetic traits. Results show that there is much more work to be done in this area, as models underperform



(a) Covariance and Reconstruction

(b) Triplet and Reconstruction

(c) VAE

Fig. 5.5: t-SNE [52] plot of latent space for VAE, Covariance & Reconstruction, and Triplet & Reconstruction models. Each color represents a different species, and the coloring is consistent across all subplots.

classic inference methods. However they still show a significant improvement to random tree construction, indicating that they capture part of the phylogenetic signal. Further experiments exploring the effect of different simulated assumptions, such as the varying evolutionary rates should be explored to determine where the model is failing to capture evolutionary relationships. Once better methods are established, the use of these models to explain evolutionary important traits should be explored. Unfortunately we think the models need to be improved before this can be completed.

Discussion and Future Work 6

> *“Yes," said the skull. "Quit while you’re a head, that’s what I say."*
>
> — **Terry Pratchett**
> Soul Music

6.1 Discussion

The goal of this PhD project is to explore methods of applying deep learning to the problem of phylogenetic inference from images. Three research questions were identified and experiments conducted to explore these. Answers to each research question from this body of work follow.

1. How can deep learning best be used to extract phylogenetically relevant traits from dorsal images of pinned insect specimens?

   [Chapter 3] explores this. Here we introduce a novel dataset and show that deep metric learning methods can be used to extract phylogenetically relevant traits from images. These traits can then be directly used in existing phylogenetic inference methods. We have shown that said traits do carry a phylogenetic signal, but further work is necessary to improve the accuracy of trees inferred from these traits.

2. How could such traits be used in conjunction with genetic data for total evidence analysis?

   [Chapter 4] explores this. Here we show that deep learnt morphological traits underperform molecular traits in inferring trees for the Rove-Tree-11 dataset. These traits show some promise when combined with molecular traits in a total evidence analysis, however the additional effort to extract and verify the traits render such total evidence analyses cost-ineffective at this point in time.

3. How can we add explainability to such models, so that deep learning methods can direct us towards phylogenetically relevant traits?

   [Chapter 5] explores this. Here we show using simulated genetic data that known likelihood and phylogenetic signal functions can be adapted to deep learning loss functions which can then directly optimize traits based on a given tree. However, this study shows that our implementation of such optimization is not yet capable of extracting all the evolutionarily relevant information from the original traits and underperforms current molecular inference methods. We suggest further simulated experiments which quantify the effect of various evolutionary mechanisms on the trait extraction.

Overall, we explored the research questions of interest. We showed that there is potential in this area of research, however there are many unanswered questions and possibilities to explore further.

6.2 Future Work

The results of the previous chapter highlight challenges in this field. We believe there is much potential in the field of deep learnt phylogenetic relationships, as indicated by the increase in recent research in this area [58]. However, a lot of work needs to be done before we can reach that potential. Below are my thoughts on avenues that should be explored further.

Probability Modelling Further work on understanding the underlying evolutionary processes and how to implement them effectively in deep learning models is one of the areas that seems most promising. A further development of the probability model itself is necessary here, as well as methods to effectively include them directly into the deep learning process. Zhang and Matsen IV [110] and the subsplit networks they work with present a particularly interesting avenue of research.

Comparison with other protein/dna models Sequence models have developed much further since DeepSequence, the architecture which Chapter 5 utilized, was published. Transformer models such as [39] have shown promise in various genetic applications, but to my knowledge have yet to be applied to phylogenetics. The ability of such models to capture long range dependencies could provide the key.

Simulated Rove-Tree-11 DNA Since the simulation model used in Magee et al. [53] and Chapter 5 is based on an empirical dataset, it is promising that the same simulations could be done with the Rove-Tree-11 dataset to further explore ar-

eas of improvement in phylogenetic inference, and offer a comparison between morphological trait extraction and genetic extraction.

Disentangling Traits - We are interested in further exploration of the possibility to disentangle dependant traits while preserving evolutionary trait relationships, and if it is possible, how this affects the resulting inference methods.

Simulation Experiments and Model Improvement - An exploration of how different simulated parameters such as the variable evolutionary rates of the model affects the deep learning model's ability to infer the phylogeny should be explored, and might provide insight into new approaches. I envision this including changing the variables in the simulation itself and measuring the effect on the resulting inference from the deep learnt traits.

Explanatory Models We believe a strong path to understanding how we can improve the models, is understanding how we can push them to extract phylogenetically explanatory traits. In particular, if we can show that these models can learn to find synapomorphies[1] automatically, this would be an interesting step forward.

[1] traits shared exclusively in evolutionarily-related clades, such as mobile shoulder joints in apes

7 Conclusion

> *By the end of my PhD I could swing a sledgehammer.*
>
> — **Jocelyn Bell Burnell**

The intention of this PhD project is to explore how deep learning can be used to generate phylogenies from digitized images of pinned insect collections. To do this we show that first, in our Rove-Tree-11 chapter, it is possible to extract phylogenetically relevant traits from images automatically using deep learning and that current deep metric learning loss functions produce similar results. While this provides hope, we also show in the Gattaca chapter that these deep learnt morphological traits are still currently inferior to genetic data and inference methods, although they show some promise in total evidence analyses. Finally we explore methodological improvements through novel loss functions for the extraction of traits from simulated genetic data. In this we also find that traditional genetic inference methods outperform deep learning, even when deep learnt methods are directly trained on the phylogenetic tree. This suggests that there is a large potential for improvement in deep learnt phylogenetic methods, probably requiring a stronger model of the underlying mechanisms of evolution.

Glossary

allele a version of many unique versions of a gene that exists in a population. 121

base pair A genetic base pair from a DNA string, ie A,C,G or T. Used interchangeably in this work with nucleotide. 120

carcinization The tendency for crustaceans to independently evolve into crab-like morphologies. 1

clade a monophyletic group; A group of taxa which share a common ancestor. Ie, given some set of species, A, the subset $M \subset A$ is monophyletic if there exists a node in the phylogenetic tree topology τ that splits A into two disjoint sets, M' (the complement of M) and M. 3, 120

convergent evolution A genetic base pair from a DNA string, ie A,C,G or T. Used interchangeably in this work with nucleotide. 4

diversification rate the net rate of change in biodiversity, defined as the rate of speciation (new species formation) minus the rate of extinction. 8

dorsal The 'back' of an organism. Usually associated with the spine in vertibrates. As opposed to ventral. 4

embedding Typically used by deep learning practitioners to refer to a vector of learnt latent variables of the original data, extracted through an encoder neural network. Used interchangeably in this work with trait, morphological trait, continuous trait and latent variable vector. 120

exon a protein coding region of a gene or genome. 99

gene duplication or loss when a gene in an organism is copied into another part of the genome, or is lost entirely from the genome. 8

gene tree A tree representing how a gene evolved. This may or may not coincide with the phylogenetic tree. 8

group a subset of taxa; Ie, given some set of taxa, A, any subset $B \subset A$ can be considered a group. 119

habitus Overall outer appearance of an organism. 4

homologies Traits which are the same across different groups, for example, some species of beetle may have the same antennae shape. 4

horizontal gene transfer when a gene is transferred from one organism to another non-sexually. Common in bacteria. 8

introgression the process of gene transfer between related species through sexual reproduction of hybrids with ancestral species. 8

latent variable Typically used by deep learning practitioners to refer to a learnt variable of the original data, extracted through an encoder neural network. Used interchangeably in this work with trait, morphological trait, continuous trait and embedding. 121

mimicry The biological phenomena where distantly related species evolve to look alike, for example *Emus hirtus*, a species of rove beetle which mimics a bee in coloration and fuzziness. 4

monophyletic a set of species derived from a single common ancestor. See also clade. 119

nucleotide A basic building block of DNA. Used interchangeably in this work with base pair. 119

phenotypic relating to the phenotype. A phenotype is an observable trait or characteristic of an organism. ie, shape, size, color, bird call. 7

phylogenetic tree A tree representing the evolution of taxa and their ancestors. 8

polymorphism when more than one allele exists in a population for a given gene. 8

recombination the process of gene exchange between chromosomes during meiosis, in the process of sexual reproduction. 8

taxa short for taxonomic group. In this thesis typically refers to any taxonomic group at the leaves of a tree, for example a genus or species. 6

trait Typically used by biologists to refer to a feature of a species, such as size, shape, color. Used interchangeably in this work with latent variable, morphological trait, continuous trait, embedding. 3, 119, 120

ventral The ’front’ of the organism, usually associated with the belly or abdomen. 119

Bibliography

[1] Shiran Abadi, Oren Avram, Saharon Rosset, Tal Pupko, and Itay Mayrose. „ModelTeller: model selection for optimal phylogenetic reconstruction using machine learning". In: *Molecular biology and evolution* 37.11 (2020), pp. 3338–3352 (cit. on p. 8).

[2] Ehab Abouheif. „A method for testing the assumption of phylogenetic independence in comparative data". In: *Evolutionary Ecology Research* 1 (1999), pp. 895–909 (cit. on p. 16).

[3] Marc-Élie Adaïmé, Shu Kong, and Surangi W Punyasena. „Deep learning approaches to the phylogenetic placement of extinct pollen morphotypes". In: *PNAS Nexus* 3.1 (Dec. 2023), pgad419. eprint: `https://academic.oup.com/pnasnexus/article-pdf/3/1/pgad419/55380858/pgad419\_supplementary\_data.pdf` (cit. on p. 7).

[4] Dana Azouri, Shiran Abadi, Yishay Mansour, Itay Mayrose, and Tal Pupko. „Harnessing machine learning to guide phylogenetic-tree search algorithms". In: *Nature communications* 12.1 (2021), p. 1983 (cit. on pp. 3, 7).

[5] Sarkhan Badirli, Zeynep Akata, George Mohler, Christine Picard, and Mehmet M Dundar. „Fine-Grained Zero-Shot Learning with DNA as Side Information". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 19352–19362 (cit. on pp. 13, 90).

[6] Ananya Bhattacharjee and Md Shamsuzzoha Bayzid. „Machine learning based imputation techniques for estimating phylogenetic trees from incomplete distance matrices". In: *BMC genomics* 21.1 (2020), p. 497 (cit. on p. 7).

[7] Paul D Blischak, Michael S Barker, and Ryan N Gutenkunst. „Chromosome-scale inference of hybrid speciation and admixture with convolutional neural networks". In: *Molecular Ecology Resources* 21.8 (2021), pp. 2676–2688 (cit. on p. 8).

[8] Simon P Blomberg, Theodore Garland Jr, and Anthony R Ives. „Testing for phylogenetic signal in comparative data: behavioral traits are more labile". In: *Evolution* 57.4 (2003), pp. 717–745 (cit. on pp. 16, 101, 102).

[9] Simone P. Blomberg, Suren I. Rathnayake, and Cheyenne M. Moreau. „Beyond Brownian Motion and the Ornstein-Uhlenbeck Process: Stochastic Diffusion Models for the Evolution of Quantitative Characters". In: *The American Naturalist* 195.2 (2020). PMID: 32017624, pp. 145–165. eprint: `https://doi.org/10.1086/706339` (cit. on p. 100).

[10] James S Boster and Jeffrey C Johnson. „Form or function: A comparison of expert and novice judgments of similarity among fish“. In: *American Anthropologist* 91.4 (1989), pp. 866–889 (cit. on p. 3).

[11] TM Brooks, A Cuttelod, DP Faith, et al. „Why and how might genetic and phylogenetic diversity be reflected in the identification of key biodiversity areas?“ In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1662 (2015), p. 20140019 (cit. on p. 1).

[12] Frank T Burbrink and Marcelo Gehara. „The biogeography of deep time phylogenetic reticulation“. In: *Systematic biology* 67.5 (2018), pp. 743–755 (cit. on p. 8).

[13] Sebastian Burgstaller-Muehlbacher, Stephen M Crotty, Heiko A Schmidt, et al. „ModelRevelator: Fast phylogenetic model estimation via deep learning“. In: *Molecular Phylogenetics and Evolution* 188 (2023), p. 107905 (cit. on p. 8).

[14] Luigi L Cavalli-Sforza and Anthony WF Edwards. „Phylogenetic analysis. Models and estimation procedures“. In: *American journal of human genetics* 19.3 Pt 1 (1967), p. 233 (cit. on p. 12).

[15] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. „Large scale online learning of image similarity through ranking.“ In: *Journal of Machine Learning Research* 11.3 (2010) (cit. on p. 98).

[16] Jennifer F. Hoyal Cuthill, Nicholas Guttenberg, Sophie Ledger, Robyn Crowther, and Blanca Huertas. „Deep learning on butterfly phenotypes tests evolution&#x2019;s oldest mathematical model“. In: *Science Advances* 5.8 (2019), eaaw4967. eprint: `https://www.science.org/doi/pdf/10.1126/sciadv.aaw4967` (cit. on pp. 7, 9, 17).

[17] Olive Jean Dunn. „Multiple comparisons among means“. In: *Journal of the American statistical association* 56.293 (1961), pp. 52–64 (cit. on p. 103).

[18] Deren A. R. Eaton. „Toytree: A minimalist tree visualization and manipulation library for Python“. In: *Methods in Ecology and Evolution* 11.1 (2020), pp. 187–191. eprint: `https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13313` (cit. on p. 2).

[19] J. Felsenstein. *Inferring Phylogenies*. Sinauer, 2003 (cit. on pp. 3, 11, 94).

[20] Joseph Felsenstein. „Maximum-likelihood estimation of evolutionary trees from continuous characters.“ In: *American journal of human genetics* 25.5 (1973), p. 471 (cit. on pp. xv, 12, 96, 97, 100, 101, 103).

[21] Joseph Felsenstein. „Maximum-likelihood estimation of evolutionary trees from continuous characters.“ In: *American journal of human genetics* 25.5 (1973), p. 471 (cit. on p. 12).

[22] Joseph Felsenstein. „Phylogenies and quantitative characters“. In: *Annual Review of Ecology and Systematics* 19.1 (1988), pp. 445–471 (cit. on p. 100).

[23] Joseph Felsenstein. „Phylogenies and the Comparative Method“. In: *The American Naturalist* 125.1 (1985), pp. 1–15 (cit. on p. 2).

[24] Alexander A Fisher, Gabriel W Hassler, Xiang Ji, et al. „Scalable bayesian phylogenetics“. In: *Philosophical Transactions of the Royal Society B* 377.1861 (2022), p. 20210242 (cit. on p. 3).

[25] William Fletcher and Ziheng Yang. „INDELible: a flexible simulator of biological sequence evolution“. In: *Molecular biology and evolution* 26.8 (2009), pp. 1879–1888 (cit. on p. 9).

[26] Chikara Furusawa, Masato Tsustumi, Daisuke Koyabu, and Nen Saito. „A method for morphological feature extraction based on variational auto-encoder : an application to mandible shape“. In: (Jan. 2023) (cit. on pp. 7, 9).

[27] Wenhao Gao, Sai Pooja Mahajan, Jeremias Sulam, and Jeffrey J Gray. „Deep learning in protein structural modeling and design“. In: *Patterns* 1.9 (2020) (cit. on p. 94).

[28] Weifeng Ge. „Deep metric learning with hierarchical triplet loss“. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 269–285 (cit. on p. 14).

[29] Weifeng Ge, Weilin Huang, Dengke Dong, and Matthew R. Scott. „Deep Metric Learning with Hierarchical Triplet Loss“. In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss. Cham: Springer International Publishing, 2018, pp. 272–288 (cit. on p. 99).

[30] Zahra Gharaee, ZeMing Gong, Nicholas Pellegrino, et al. „A Step Towards Worldwide Biodiversity Assessment: The BIOSCAN-1M Insect Dataset“. In: *Advances in Neural Information Processing Systems* 36 (2024) (cit. on p. 90).

[31] Benyamin Ghojogh, Mark Crowley, Fakhri Karray, and Ali Ghodsi. „Deep metric learning“. In: *Elements of Dimensionality Reduction and Manifold Learning*. Springer, 2023, pp. 531–562 (cit. on p. 13).

[32] Pablo A. Goloboff, Ambrosio Torres, and J. Salvador Arias. „Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology“. In: *Cladistics* 34.4 (2018), pp. 407–437. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/cla.12205` (cit. on p. 12).

[33] Graham Gower, Pablo Iáñez Picazo, Matteo Fumagalli, and Fernando Racimo. „Detecting adaptive introgression in human evolution using convolutional neural networks“. In: *Elife* 10 (2021), e64669 (cit. on p. 8).

[34] Julia Haag, Dimitri Höhler, Ben Bettisworth, and Alexandros Stamatakis. „From easy to hopeless—predicting the difficulty of phylogenetic analyses“. In: *Molecular Biology and Evolution* 39.12 (2022), msac254 (cit. on p. 7).

[35] W Keith Hastings. „Monte Carlo sampling methods using Markov chains and their applications“. In: (1970) (cit. on p. 12).

[36] Mark S Hibbins and Matthew W Hahn. „Distinguishing between histories of speciation and introgression using genomic data“. In: *BioRxiv* (2022), pp. 2022–09 (cit. on p. 8).

[37] Roberta Hunt and Kim Steenstrup Pedersen. „Rove-Tree-11: The not-so-Wild Rover, A hierarchically structured image dataset for deep metric learning research“. In: *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Dec. 2022, pp. 2967–2983 (cit. on pp. 7, 9, 47).

[38] Sebastian Höhna, Michael J. Landis, Tracy A. Heath, et al. „RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language“. In: *Systematic Biology* 65.4 (May 2016), pp. 726–736. eprint: `https://academic.oup.com/sysbio/article-pdf/65/4/726/16636545/syw021.pdf` (cit. on p. 98).

[39] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. „DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome“. In: *Bioinformatics* 37.15 (2021), pp. 2112–2120 (cit. on p. 114).

[40] Yueyu Jiang, Metin Balaban, Qiyun Zhu, and Siavash Mirarab. „DEPP: deep learning enables extending species trees using single genes“. In: *Systematic biology* 72.1 (2023), pp. 17–34 (cit. on p. 7).

[41] Mahmut Kaya and Hasan Sakir Bilge. „Deep Metric Learning: A Survey“. In: *Symmetry* 11.9 (2019) (cit. on pp. 13, 100).

[42] Steffen Kiel. „Assessing bivalve phylogeny using Deep Learning and Computer Vision approaches“. In: *bioRxiv* (2021). eprint: `https://www.biorxiv.org/content/early/2021/04/09/2021.04.08.438943.full.pdf` (cit. on pp. 7, 9, 17).

[43] Diederik P Kingma, Max Welling, et al. „An introduction to variational autoencoders“. In: *Foundations and Trends® in Machine Learning* 12.4 (2019), pp. 307–392 (cit. on p. 100).

[44] Christian Peter Klingenberg and Nelly A Gidaszewski. „Testing and quantifying phylogenetic signals and homoplasy in morphometric data“. In: *Systematic biology* 59.3 (2010), pp. 245–261 (cit. on p. 13).

[45] Mario Krenn, Lorenzo Buffoni, Bruno Coutinho, et al. „Predicting the Future of AI with AI: High-quality link prediction in an exponentially growing knowledge network“. In: (Sept. 2022) (cit. on p. 2).

[46] Mary K. Kuhner and Jon Yamato. „Practical Performance of Tree Comparison Metrics“. In: *Systematic Biology* 64.2 (Dec. 2014), pp. 205–214. eprint: `https://academic.oup.com/sysbio/article-pdf/64/2/205/24586546/syu085.pdf` (cit. on p. 14).

[47] Sophia Lambert, Jakub Voznica, and Hélène Morlon. „Deep learning from phylogenies for diversification analyses“. In: *Systematic Biology* 72.6 (2023), pp. 1262–1279 (cit. on p. 8).

[48] Christopher Lean and James Maclaurin. „The value of phylogenetic diversity“. In: *Biodiversity Conservation and Phylogenetic Systematics: Preserving our evolutionary heritage in an extinction crisis* (2016), pp. 19–37 (cit. on p. 1).

[49] Michael S.Y. Lee and Alessandro Palci. „Morphological Phylogenetics in the Genomic Age“. In: *Current Biology* 25.19 (2015), R922–R929 (cit. on p. 4).

[50] Alina F Leuchtenberger, Stephen M Crotty, Tamara Drucks, et al. „Distinguishing Felsenstein zone from Farris zone using neural networks“. In: *Molecular biology and evolution* 37.12 (2020), pp. 3632–3641 (cit. on p. 7).

[51] Nhan Ly-Trong, Suha Naser-Khdour, Robert Lanfear, and Bui Quang Minh. „AliSim: A Fast and Versatile Phylogenetic Sequence Simulator for the Genomic Era“. In: *Molecular Biology and Evolution* 39.5 (May 2022), msac092. eprint: `https://academic.oup.com/mbe/article-pdf/39/5/msac092/43899863/msac092.pdf` (cit. on p. 9).

[52] Laurens van der Maaten and Geoffrey Hinton. „Visualizing Data using t-SNE“. In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605 (cit. on p. 111).

[53] Andrew F Magee, Sarah K Hilton, and William S DeWitt. „Robustness of Phylogenetic Inference to Model Misspecification Caused by Pairwise Epistasis". In: *Molecular Biology and Evolution* 38.10 (May 2021), pp. 4603–4615. eprint: `https://academic.oup.com/mbe/article-pdf/38/10/4603/40449488/msab163.pdf` (cit. on pp. 90, 92, 93, 98, 99, 114).

[54] Frederikke Isa Marin, Felix Teufel, Marc Horlacher, et al. „BEND: Benchmarking DNA Language Models on Biologically Meaningful Tasks". In: *The Twelfth International Conference on Learning Representations*. 2023 (cit. on p. 94).

[55] Bob Mau, Michael A Newton, and Bret Larget. „Bayesian phylogenetic inference via Markov chain Monte Carlo methods". In: *Biometrics* 55.1 (1999), pp. 1–12 (cit. on p. 13).

[56] Patsy A McLaughlin and Rafael Lemaitre. „Carcinization in the Anomura-fact or fiction? I. Evidence from adult morphology". In: *Contributions to Zoology* 67.2 (1997), pp. 79–123 (cit. on p. 1).

[57] Uriel D. Menalled, Richard G. Smith, Stephane Cordeau, et al. „Phylogenetic relatedness can influence cover crop-based weed suppression". In: *Scientific Reports* 13.1 (Oct. 2023), p. 17323 (cit. on p. 2).

[58] Yu K. Mo, Matthew W. Hahn, and Megan L. Smith. „Applications of Machine Learning in Phylogenetics". In: (Feb. 2024), p. 13 (cit. on pp. 6, 94, 114).

[59] Lys Sanz Moreta, Ola Rønning, Ahmad Salim Al-Sibahi, et al. „Ancestral protein sequence reconstruction using a tree-structured Ornstein-Uhlenbeck variational autoencoder". In: *International Conference on Learning Representations*. 2021 (cit. on p. 8).

[60] Tamara Münkemüller, Sébastien Lavergne, Bruno Bzeznik, et al. „How to measure and test phylogenetic signal". In: *Methods in Ecology and Evolution* 3.4 (2012), pp. 743–756 (cit. on p. 16).

[61] Marko Mutanen and Etheresia Pretorius. „Subjective visual evaluation vs. traditional and geometric morphometrics in species delimitation: a comparison of moth genitalia". In: *Systematic Entomology* 32.2 (2007), pp. 371–386 (cit. on p. 3).

[62] Chris A. Nasrallah and John P. Huelsenbeck. „A Phylogenetic Model for the Detection of Epistatic Interactions". In: *Molecular Biology and Evolution* 30.9 (June 2013), pp. 2197–2208. eprint: `https://academic.oup.com/mbe/article-pdf/30/9/2197/13173187/mst108.pdf` (cit. on pp. 90, 91, 99).

[63] Chris A. Nasrallah, David H. Mathews, and John P. Huelsenbeck. „Quantifying the Impact of Dependent Evolution among Sites in Phylogenetic Inference". In: *Systematic Biology* 60.1 (Nov. 2010), pp. 60–73. eprint: `https://academic.oup.com/sysbio/article-pdf/60/1/60/24208042/syq074.pdf` (cit. on pp. 90, 96, 99).

[64] Luca Nesterenko, Bastien Boussau, and Laurent Jacob. „Phyloformer: towards fast and accurate phylogeny estimation with self-attention networks". In: *bioRxiv* (2022), pp. 2022–06 (cit. on p. 7).

[65] Tom Nye, Pietro Lio, and Walter Gilks. „A novel algorithm and web-based tool for comparing two alternative phylogenetic trees". In: *Bioinformatics (Oxford, England)* 22 (Feb. 2006), pp. 117–9 (cit. on pp. 15, 16).

[66] Caroline Parins-Fukuchi. „Use of Continuous Traits Can Improve Morphological Phylogenetics“. In: *Systematic Biology* 67.2 (Sept. 2017), pp. 328–339. eprint: `https://academic.oup.com/sysbio/article-pdf/67/2/328/24105567/syx072.pdf` (cit. on p. 98).

[67] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. „The matrix cookbook“. In: *Technical University of Denmark* 7.15 (2008), p. 510 (cit. on p. 104).

[68] Alexander Popkov, Fedor Konstantinov, Vladimir Neimorovets, and Alexey Solodovnikov. „Machine learning for expert-level image-based identification of very similar species in the hyperdiverse plant bug family Miridae (Hemiptera: Heteroptera)“. In: *Systematic Entomology* 47.3 (2022). Publisher Copyright: © 2022 Royal Entomological Society., pp. 487–503 (cit. on pp. 2, 3).

[69] Dylan D Ray, Lex Flagel, and Daniel R Schrider. „IntroUNET: identifying introgressed alleles via semantic segmentation“. In: *PLoS genetics* 20.2 (2024), e1010657 (cit. on p. 8).

[70] Enrique Rayo, Gabriel F Ulrich, Niklaus Zemp, et al. „Minimally destructive hDNA extraction method for retrospective genetics of pinned historical Lepidoptera specimens“. In: *Scientific Reports* 14.1 (2024), p. 12875 (cit. on p. 4).

[71] Liam J. Revell. „phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things).“ In: *PeerJ* 12 (2024), e16505 (cit. on p. 101).

[72] Adam J. Riesselman, John B. Ingraham, and Debora S. Marks. „Deep generative models of genetic variation capture the effects of mutations“. In: *Nature Methods* 15.10 (Oct. 2018), pp. 816–822 (cit. on pp. 94, 95, 100, 109).

[73] D.F. Robinson and L.R. Foulds. „Comparison of phylogenetic trees“. In: *Mathematical Biosciences* 53.1 (1981), pp. 131–147 (cit. on p. 14).

[74] Benjamin Rosenzweig, Andrew Kern, and Matthew Hahn. „Accurate detection of incomplete lineage sorting via supervised machine learning“. In: *bioRxiv* (2022), pp. 2022–11 (cit. on p. 8).

[75] N Saitou and M Nei. „The neighbor-joining method: a new method for reconstructing phylogenetic trees.“ In: *Molecular Biology and Evolution* 4.4 (July 1987), pp. 406–425. eprint: `https://academic.oup.com/mbe/article-pdf/4/4/406/11167444/7sait.pdf` (cit. on p. 11).

[76] Nicolae Sapoval, Amirali Aghazadeh, Michael G. Nute, et al. „Current progress and open challenges for applying deep learning across the biosciences“. In: *Nature Communications* 13.1 (Apr. 2022), p. 1728 (cit. on p. 6).

[77] Diane I. Scaduto, Jeremy M. Brown, Wade C. Haaland, et al. „Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences“. In: *Proceedings of the National Academy of Sciences* 107.50 (2010), pp. 21242–21247. eprint: `https://www.pnas.org/doi/pdf/10.1073/pnas.1015673107` (cit. on p. 2).

[78] Daniel R Schrider and Andrew D Kern. „Supervised machine learning for population genetics: a new paradigm“. In: *Trends in Genetics* 34.4 (2018), pp. 301–312 (cit. on p. 8).

[79] Robert W. Scotland, Richard G. Olmstead, and Jonathan R. Bennett. „Phylogeny Reconstruction: The Role of Morphology“. In: *Systematic Biology* 52.4 (2003), pp. 539–548 (cit. on p. 4).

[80] Hang Shao, Abhishek Kumar, and P. Thomas Fletcher. „The Riemannian Geometry of Deep Generative Models“. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2018 (cit. on p. 100).

[81] Beth Shapiro, Andrew Rambaut, Oliver G. Pybus, and Edward C. Holmes. „A Phylogenetic Method for Detecting Positive Epistasis in Gene Sequences and Its Application to RNA Virus Evolution“. In: *Molecular Biology and Evolution* 23.9 (June 2006), pp. 1724–1730. eprint: `https://academic.oup.com/mbe/article-pdf/23/9/1724/13414246/msl037.pdf` (cit. on p. 99).

[82] Jiayi Shen, Casper JP Zhang, Bangsheng Jiang, et al. „Artificial intelligence versus clinicians in disease diagnosis: systematic review“. In: *JMIR medical informatics* 7.3 (2019), e10010 (cit. on p. 2).

[83] Megan L Smith and Matthew W Hahn. „Phylogenetic inference using generative adversarial networks“. In: *Bioinformatics* 39.9 (Sept. 2023), btad543. eprint: `https://academic.oup.com/bioinformatics/article-pdf/39/9/btad543/51546716/btad543.pdf` (cit. on p. 7).

[84] Royal Entomological Society. *Facts and Figures*. `https://www.royensoc.co.uk/understanding-insects/facts-and-figures/` (cit. on p. 3).

[85] R.R. Sokal and P.H.A. Sneath. „Principles of numerical taxonomy“. In: San Francisco: W. H. Freeman and Company, 1963, p. 359 (cit. on p. 4).

[86] Claudia Solis-Lemus, Shengwen Yang, and Leonardo Zepeda-Nunez. „Accurate phylogenetic inference with a symmetry-preserving neural network model“. In: *arXiv preprint arXiv:2201.04663* (2022) (cit. on p. 7).

[87] Tyler N Starr and Joseph W Thornton. „Epistasis in protein evolution“. In: *Protein science* 25.7 (2016), pp. 1204–1218 (cit. on p. 99).

[88] Nigel E. Stork. „How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth?“ In: *Annual Review of Entomology* 63.Volume 63, 2018 (2018), pp. 31–45 (cit. on p. 3).

[89] A.K Subramanian. *PyTorch-VAE*. `https://github.com/AntixK/PyTorch-VAE`. 2020 (cit. on p. 89).

[90] Chaoyue Sun, Ruogu Fang, Marco Salemi, Mattia Prosperi, and Brittany Rife Magalis. „DeepDynaForecast: Phylogenetic-informed graph deep learning for epidemic transmission dynamic prediction“. In: *PLOS Computational Biology* 20.4 (2024), e1011351 (cit. on p. 8).

[91] Anton Suvorov, Joshua Hochuli, and Daniel R Schrider. „Accurate inference of tree topologies from multiple sequence alignments using deep learning“. In: *Systematic biology* 69.2 (2020), pp. 221–233 (cit. on p. 6).

[92] Anton Suvorov and Daniel R Schrider. „Reliable estimation of tree branch lengths using deep neural networks“. In: *bioRxiv* (2022), pp. 2022–11 (cit. on p. 8).

[93] Qiqing Tao, Koichiro Tamura, Fabia U. Battistuzzi, and Sudhir Kumar. „A machine learning method for detecting autocorrelation of evolutionary rates in large phylogenies“. In: *Molecular biology and evolution* 36.4 (2019), pp. 811–824 (cit. on p. 8).

[94] S Tarvaré. „Some probabilistic and statistical problems in the analysis of DNA sequences“. In: *Some mathematical question in biology-DNA sequence analysis* (1986) (cit. on p. 91).

[95] Johanna Trost, Julia Haag, Dimitri Höhler, et al. „Simulations of sequence evolution: how (un) realistic they are and why“. In: *Molecular biology and evolution* 41.1 (2024), msad277 (cit. on pp. 9, 90).

[96] Georgia Tsagkogeorga, Xavier Turon, Russell R Hopcroft, et al. „An updated 18S rRNA phylogeny of tunicates based on mixture and secondary structure models“. In: *BMC evolutionary biology* 9 (2009), pp. 1–16 (cit. on p. 90).

[97] Yu. A. Vakulenko, A. N. Lukashev, and A. A. Deviatkin. „The use of statistical phylogenetics in virology“. In: *Russian Journal of Infection and Immunity* 11.1 (2020), pp. 42–56 (cit. on p. 2).

[98] Jakub Voznica, Anna Zhukova, Veronika Boskova, et al. „Deep learning from phylogenies to uncover the epidemiological dynamics of outbreaks“. In: *Nature Communications* 13.1 (2022), p. 3896 (cit. on p. 8).

[99] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. „Disentangled representation learning“. In: *arXiv preprint arXiv:2211.11695* (2022) (cit. on p. 13).

[100] Zhicheng Wang, Jinnan Sun, Yuan Gao, et al. „Fusang: a framework for phylogenetic tree inference via deep learning“. In: *Nucleic Acids Research* 51.20 (Oct. 2023), pp. 10909–10923. eprint: `https://academic.oup.com/nar/article-pdf/51/20/10909/53174297/gkad805.pdf` (cit. on p. 7).

[101] Tandy Warnow. „Supertree construction: opportunities and challenges“. In: *arXiv preprint arXiv:1805.03530* (2018) (cit. on p. 3).

[102] Michael S Waterman, Temple F Smith, Mona Singh, and William A Beyer. „Additive evolutionary trees“. In: *Journal of theoretical Biology* 64.2 (1977), pp. 199–213 (cit. on p. 11).

[103] John J Wiens and Maria R Servedio. „Phylogenetic analysis and intraspecific variation: performance of parsimony, likelihood, and distance methods“. In: *Systematic Biology* 47.2 (1998), pp. 228–253 (cit. on p. 12).

[104] Erik S Wright. „Using DECIPHER v2. 0 to analyze big biological sequence data in R.“ In: *R Journal* 8.1 (2016) (cit. on p. 98).

[105] Hong Xuan, Abby Stylianou, and Robert Pless. „Improved embeddings with easy positive triplet mining“. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020, pp. 2474–2482 (cit. on p. 110).

[106] Haoran Xue. „Phylogenetics and its Application in Biodiversity Conservation“. In: *Molecular Genetics and Genomics Tools in Biodiversity Conservation*. Ed. by Ashwani Kumar, Baharul Choudhury, Selvadurai Dayanandan, and Mohammed Latif Khan. Singapore: Springer Nature Singapore, 2022, pp. 1–16 (cit. on p. 1).

[107] Ziheng Yang and Bruce Rannala. „Branch-Length Prior Influences Bayesian Posterior Probability of Phylogeny“. In: *Systematic Biology* 54.3 (June 2005), pp. 455–470. eprint: `https://academic.oup.com/sysbio/article-pdf/54/3/455/26544339/10635150590945313.pdf` (cit. on p. 97).

[108]Paul Zaharias, Martin Grosshauser, and Tandy Warnow. „Re-evaluating deep neural networks for phylogeny estimation: the issue of taxon sampling“. In: *Journal of Computational Biology* 29.1 (2022), pp. 74–89 (cit. on p. 7).

[109]Cheng Zhang and Frederick A. Matsen IV. „Variational Bayesian Phylogenetic Inference“. In: *International Conference on Learning Representations*. 2019 (cit. on p. 7).

[110]Cheng Zhang and Frederick A Matsen IV. „Variational Bayesian phylogenetic inference“. In: *International Conference on Learning Representations*. 2018 (cit. on p. 114).

[111]Yubo Zhang, Qingjie Zhu, Yi Shao, et al. „Inferring historical introgression with deep learning“. In: *Systematic Biology* 72.5 (2023), pp. 1013–1038 (cit. on p. 8).

[112]Zhengting Zou, Hongjiu Zhang, Yuanfang Guan, and Jianzhi Zhang. „Deep residual neural networks resolve quartet molecular phylogenies“. In: *Molecular biology and evolution* 37.5 (2020), pp. 1495–1507 (cit. on p. 6).

[113]Zhengting Zou and Jianzhi Zhang. „Morphological and molecular convergences in mammalian phylogenetics“. In: *Nature Communications* 7.1 (2016), p. 12758 (cit. on p. 4).

List of Figures

List of Tables

Colophon

This thesis was typeset with $\LaTeX\,2_\varepsilon$. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.

Declaration

I swear by my pretty floral bonnet

— **Malcolm Reynolds, Firefly**

I swear that I have completed this work solely and only with the help of the references mentioned and my collaborators.

Copenhagen, May 31, 2024

Roberta Hunt