Applications of Machine Learning in Phylogenetics

Yu K. Mo[1], Matthew W. Hahn[1,2], and Megan L. Smith[2,3,*]

[1]Department of Computer Science and [2]Department of Biology, Indiana University, Bloomington, IN 47405
[3]Department of Biological Sciences, Mississippi State University, Starkville, MS 39762
[*]Corresponding Author: msmith@biology.msstate.edu

Abstract

Machine learning has increasingly been applied to a wide range of questions in phylogenetic inference. Supervised machine learning approaches that rely on simulated training data have been used to infer tree topologies and branch lengths, to select substitution models, and to perform downstream inferences of introgression and diversification. Here, we review how researchers have used several promising machine learning approaches to make phylogenetic inferences. Despite the promise of these methods, several barriers prevent supervised machine learning from reaching its full potential in phylogenetics. We discuss these barriers and potential paths forward. In the future, we expect that the application of careful network designs and data encodings will allow supervised machine learning to accommodate the complex processes that continue to confound traditional phylogenetic methods.

1 Introduction

Phylogenetics aims to elucidate the evolutionary relationships among species. In recent decades, owing to rapid growth in the availability of genomic data, phylogenetic analysis has been able to use hundreds to thousands of loci (Delsuc et al., 2005). Using whole genomes, or even near-whole genomes, may allow for a more comprehensive view of the evolutionary events shaping species (Scornavacca et al., 2020). However, the accuracy of inference may be compromised when using such large datasets, as even small biases can be magnified many-fold. Biases in phylogenetics are often due to unmodeled heterogeneity in the evolutionary process, including heterogeneity across time, sites, genes, or lineages (Kapli et al., 2020). These processes may arise either individually or in combination, presenting challenges in subsequent analyses.

Recently, machine learning techniques have been used across fields, demonstrating impressive power in uncovering intricate relationships from data that contains extensive heterogeneity. Notable examples include successful applications in image classification (Krizhevsky et al., 2017), language models (Devlin et al., 2019), protein structure prediction (Jumper et al., 2021), and population genetics (Schrider & Kern, 2018). Machine learning is comprised of two fundamental paradigms—supervised and unsupervised approaches. Supervised learning relies on the availability of labeled training data, where the true underlying state or value of the data is known. In phylogenetics and related fields, large amounts of labeled training data are generally unavailable, so simulations are often used to generate such data. The primary objective of supervised machine learning is to learn a function that can map input data to appropriate outputs. Within supervised learning, there are two primary tasks: classification and regression. While classification aims to predict discrete labels or categories, regression predicts continuous-valued outputs. In contrast, unsupervised learning operates without the need for labeled data, focusing instead on discerning underlying structures or patterns in the input data. Unsupervised approaches include tasks such as clustering and dimensionality reduction. Notably, deep learning is a specialized subset of machine learning that leverages neural networks (NNs) with many layers (hence "deep"). Some NN architectures are adept at automatically extracting hierarchical features from raw data, obviating the need for manual feature engineering—a significant advantage over traditional machine learning methods.

In the context of phylogenetics, machine learning algorithms are extremely flexible, both with regards to the structuring of input data, and the data used for training. Furthermore, machine learning approaches can learn complex relationships from input data without calculating likelihoods. This facilitates the application of machine learning to complex models, especially scenarios in which standard likelihood and Bayesian inference may be intractable. Given the lack of analytical phylogenetic solutions that can be reasonably applied to large genomic datasets, machine learning offers the promise of moving beyond conventional methods.

Despite the promise that machine learning in general has for addressing many biological problems, there is uncertainty about its superiority over conventional approaches in many applications to phylogenetics. While a growing number of papers have applied machine learning to multiple problems in the field, researchers have not yet seen a clear advantage to such approaches. Here, we review recent applications of machine learning to different tasks in phylogenetics (Table 1), examining their limitations and strengths. We attempt to provide a general overview of the types of machine learning approaches that have been used—and those that could be used—in the hope that future work will bring the promise of machine learning to fruition.

2 Tree Reconstruction

Reconstructing evolutionary relationships among taxa is a central goal in evolutionary biology. A phylogenetic tree is composed of two primary components: a topology and a set of branch lengths. The topology serves as a representation of the hierarchical evolutionary relationships among species. The branch lengths represent evolutionary change, measured either in absolute time, in the number of nucleotide substitutions, or in other units. This section reviews machine learning approaches for inferring both components of phylogenetic trees.

2.1 Topology inference

Perhaps the most natural framing of the problem of topology inference is to use supervised machine learning approaches for classification, since the goal is to predict a discrete output (topology) from sequence data. Recall that supervised machine learning approaches require labeled training data, which are generally unavailable in phylogenetics. Because of this, in most phylogenetic applications simulations are performed under each model of interest prior to inference, and these simulated data are used to train the machine learning network. When the goal is topology inference, the model space includes, at a minimum, the number of possible tree topologies. With as few as ten taxa, there are more than two million unrooted topologies, making it infeasible to use such approaches to infer tree topologies for even moderate numbers of taxa. The challenges associated with a large state-space of topologies are not unique to machine learning approaches: even conventional methods have difficulties in inferring trees for large numbers of species (Felsenstein, 1978b; Roch, 2006). To circumvent this problem, researchers have used three different types of approaches in order to apply machine learning to phylogenetic inference (Figure 1). Here we review these approaches and the specific models that have been used.

2.1.1 Quartet-based methods

The first machine learning approaches in phylogenetics used quartet-based methods. In general, quartet-based methods involve extracting sets of four taxa from the full dataset, building trees for each set of four taxa, and then constructing a phylogeny from these quartet trees using one of several quartet amalgamation approaches, such as quartet puzzling (Bryant & Steel, 2001; Reaz et al., 2014; Snir & Satish, 2012). Because there are only three possible topologies for an unrooted quartet, such approaches are not plagued by the need to consider a very large state-space of topologies. Quartet-based methods therefore provide efficient inference algorithms that are scalable to very large datasets.

Several supervised learning approaches have been used to infer quartet trees. Suvorov et al. (2020) used a convolutional neural network (CNN) that takes integer-encoded nucleotide alignments as input. Machine learning algorithms generally require that input data are numerical, and integer-encoding can be used to represent categorical variables. In this application, each nucleotide was encoded as an integer between 0 and 3, with gaps encoded as 4, and each alignment was represented as a matrix in which rows correspond to sequences and columns correspond to sites in the alignment. The topology associated with each alignment was an integer-encoded class label. Training data were simulated under a wide range of branch lengths, several substitution models, with site heterogeneity, and with or without gaps. In the absence of gaps, the CNN generally performed as well as or better than traditional approaches. On datasets that included gaps, the CNN substantially outperformed traditional approaches, likely because it better utilized this significant source of phylogenetic signal. The CNN initially exhibited reduced accuracy in some zones of branch length space (e.g., the Felsentstein zone; (Felsenstein, 1978a)). However, when more training data were included from these regions the CNN was able to outperform other approaches, highlighting the importance of carefully considering where to put effort in training such models.

In a similar approach, Zou et al. (2020) used a residual neural network, which takes as input one-hot encoded amino acid sequences. One-hot encoding is an alternative to integer-encoding for representing categorical variables as numeric input. In this application, each site was represented by twenty channels, with each channel corresponding to an amino acid. For an individual site, the channel corresponding to the amino acid present in the position is set to one, while all other channels are set to zero. One-hot encoding may be more appropriate than integer-encoding, since it avoids implicit ordered relationships among states. In Zou et al.'s approach, models were trained on amino acid sequences simulated on large, random trees, which were then pruned to subsets of four taxa. Both site and time heterogeneity were included in the simulations; additionally, the training data intentionally included a sizable proportion of trees susceptible to long-branch attraction, to ensure that a large number of difficult examples were included. When benchmarked against existing inference approaches, the residual network predictors consistently delivered better results with less computational time (not including training time), especially when dealing with several cases that confound existing methods—such as long branch attraction and heterotachy. By combining their approach with a quartet amalgamation approach, these authors were able to infer larger species trees with moderate accuracy.

Both of the methods described above treat alignments as images. While this approach to representing data has been found to be powerful in population genetics (Flagel et al., 2019), there are several limitations in the context of phylogenetics. For example, when inferring relationships among taxa, we would like the order in which sequences are included in the model to be irrelevant (a property referred to as "permutation equivariant"). However, most network architectures do not perform in this way. Zou et al. (2020) accommodated this behavior by including all permutations of the alignment when training, but such an approach increases the compute time and memory needed to train a neural network. Solís-Lemus et al. (2023) address this issue using a symmetry-preserving long short-term memory (LSTM) recurrent neural network (RNN). By avoiding the need to include permutations of the training alignments, they substantially improved compute times and memory usage compared to Zou et al. (2020). These approaches have also been limited in the ease with which they can be applied to empirical datasets both due to limitations in the lengths of alignments than can be considered and the lack of a user-friendly pipeline. Fusang (Wang et al., 2023) addresses these issues by using a sliding window approach to accommodate variable alignment lengths and developing an easy-to-use pipeline. Fusang takes as input an alignment including no more than 40 sequences, infers quartet topologies, and then uses a stepwise addition algorithm with beam search to infer larger trees from quartet trees.

Even though NNs can be very efficient for inferring quartet trees, considering larger trees remains prohibitive—the approaches described above still must rely on quartet-amalgamation approaches to build larger trees. Additionally, as with all supervised machine learning, accuracy is likely limited in cases where the training data is not reflective of real data. Zaharias et al. (2022) explored these limitations by comparing the networks from Zou et al. (2020) to standard approaches on larger trees and on test datasets with higher rates of nucleotide evolution and/or shorter alignment lengths. They found that the neural networks only outperformed traditional approaches when the goal was to infer a quartet tree from relatively long amino acid sequences simulated under model conditions very similar to those used for training. Furthermore, when larger trees were considered, traditional approaches outperformed the combination of neural networks and quartet amalgamation. Machine learning approaches are therefore severely limited by their inability to directly infer trees from larger numbers of taxa, as well as by the specifics of the data used in training.

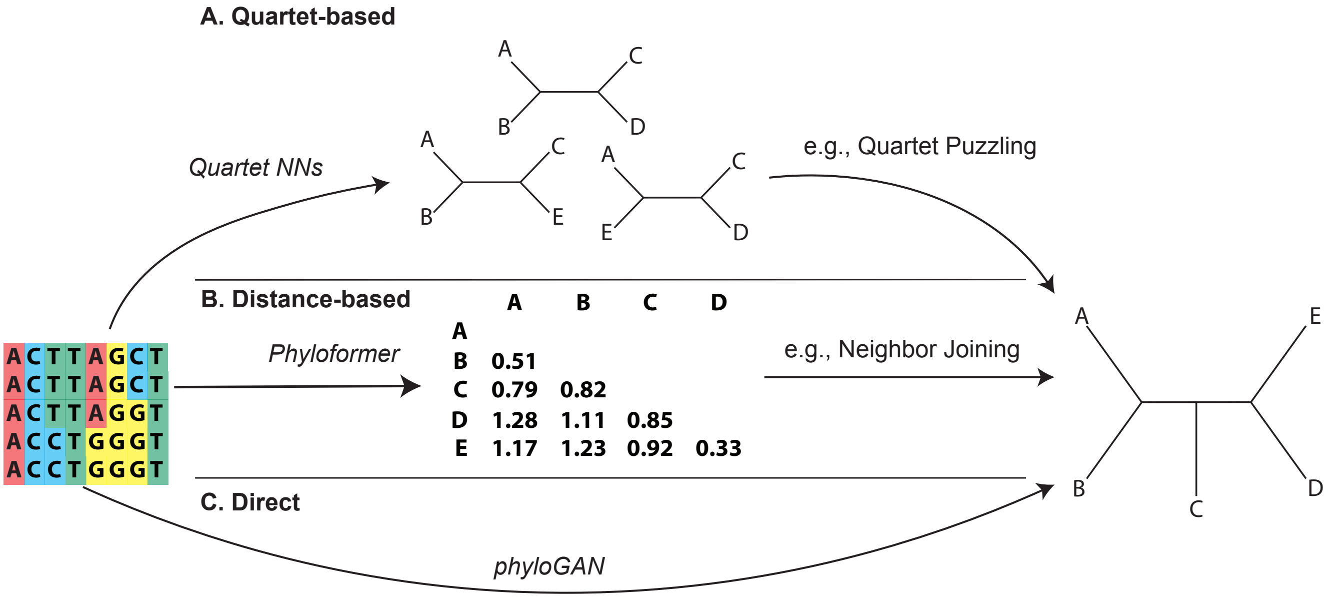

Figure 1: Methods for topology inference using machine learning. A. Quartet-based methods infer one of the three topologies possible with unrooted quartets. Trees from each quartet are inferred with NNs; a collection of such trees are then fed into existing quartet amalgamation algorithms (e.g. Quartet Puzzling) to infer a larger phylogeny. B. Distance-based methods estimate pairwise distances using NNs (e.g. Phyloformer). Distances are combined using standard methods (e.g. Neighbor Joining) to reconstruct trees. C. Direct methods infer a tree directly from an alignment using NNs (e.g. phyloGAN).

2.1.2 Distance-based methods

Rather than using machine learning to directly infer trees from sequence alignments, it is possible to instead infer evolutionary distances, which can then be used as input to standard distance-based approaches. Although often scoffed at by modern phylogeneticists, distance-based approaches such as neighbor joining (Saitou & Nei, 1987) are in fact guaranteed to infer the correct tree in most of parameter space, as long as distances are accurately inferred. In addition, they are much more accurate than maximum likelihood in the presence of high amounts of incomplete lineage sorting (Liu & Edwards, 2009; Mendes & Hahn, 2018). Therefore, it makes sense to apply machine learning to the task of accurately inferring distances.

Nesterenko et al.(2022) developed Phyloformer, which uses self-attention networks to infer evolutionary distances for up to 100 species. Their model encapsulates alignment in a pairwise way, introducing a representation for each pair with the attention mechanism. The process entails an iterative sharing of information, first across sites within each pair (referred to as site-level attention) and subsequently across pairs within each site (termed pair-level attention). Such an approach is permutation-equivariant, and accommodates alignments of varying sizes. After inferring distances, these authors used neighbor joining for tree construction. Their approach outperformed traditional distance-based approaches, and was competitive with (and much faster than) maximum likelihood when training and testing data included similar numbers of species. However, Phyloformer does not always compare favorably to standard methods, especially on trees with more than twenty leaves.

In a related approach, Bhattacharjee and Bayzid (2020) used autoencoders and matrix factorization to impute missing values in distance matrices. Alternatively, Jiang et al. (2023) use a CNN for phylogenetic placement—placing sequences from individual genes onto trees that may have been inferred using different genomic regions. In this case they inferred evolutionary distances for these new sequences, and then used a distance-based algorithm to place the new sequences on the tree (Balaban et al., 2022). Inferring evolutionary distances reframes phylogenetic inference as a regression problem, rather than as a classification problem. This reframing makes it possible to scale machine learning approaches to larger trees.

2.1.3 Direct methods

In maximum likelihood and Bayesian approaches to phylogenetic inference, the large number of possible topologies is accommodated by using heuristic searches to explore tree space; such approaches could also be used for direct inference of tree topologies from sequence data in machine learning contexts. Generative adversarial networks (GANs) consist of a generator, which aims to produce realistic data, and a discriminator, which aims to distinguish real and fake data (Goodfellow et al., 2020). Recently, Smith and Hahn (2023) proposed phyloGAN. phyloGAN consists of a generator, which generates topologies and branch lengths, and a CNN-based discriminator, which attempts to distinguish alignments simulated under these topologies and branch lengths from empirical (real) alignments. Ideally, at the end of training, it should be virtually impossible to distinguish simulated and empirical alignments. Once this level of accuracy is achieved, the topology that underpins the simulated data is considered to be the inferred topology. phyloGAN was tested on up to fifteen species, and a version incorporating gene tree heterogeneity was tested on six species. While phyloGAN worked well with small numbers of species (up to ten), it was computationally intensive, and several metrics indicated issues during training. Additionally, since phyloGAN performs a heuristic exploration of tree space, it must be trained anew for each empirical dataset, and thus many of the potential computational benefits of machine learning approaches are not realized. Future work may explore alternative approaches for heuristically exploring model spaces using machine learning frameworks, including approaches covered in the next section.

2.1.4 Improving steps in topology inference

Machine learning approaches have been used to assist standard phylogenetic approaches for topology inference. For example, machine learning approaches have been used to improve heuristic searches for tree topologies. Azouri et al. (2021) used a random forest (RF) regressor to predict likelihood scores for subtree-prune-regraft (SPR) moves, a standard and important step in heuristic tree searches. Given a starting topology, their network could accurately predict the change in likelihood associated with different SPR moves, which suggests that such an approach could be used to limit search space and therefore to reduce the computational requirements for heuristic searches. In a follow-up paper, Azouri et al. (2023) used reinforcement learning as an alternative to traditional heuristic search algorithms. By allowing for suboptimal moves that, nonetheless, improved the final outcome of the search, this approach out-competed greedy search strategies.

Machine learning approaches have also been used to guide researchers in their decisions about which standard approaches to use for topological inference. Leuchtenberger et al. (2020) developed a feed-forward neural network to classify alignments as belonging to the Farris (Siddall, 1998) or Felstenstein zone (Felsenstein, 1978a; Huelsenbeck & Hillis, 1993). They based their choice to use maximum parsimony (in the Farris Zone) or maximum likelihood (in the Felsenstein zone) on the predictions of this neural network. Using this approach resulted in higher overall accuracy compared to always using either maximum parsimony or maximum likelihood. In a follow-up paper, Leuchtenberger and von Haeseler (2024) simplified this neural network to develop a simple, more interpretable classifier, illustrating how subsequent investigations into complex networks can yield theoretical insights. In a similar application, Haag et al. (2022) developed a random forest regressor, Pythia, to predict the difficulty of inferring a tree from a particular alignment. They suggested that the predicted level of difficulty be used to guide decisions regarding analysis design, including potentially collecting more data prior to analyses for difficult alignments.

2.2 Branch length inference

In addition to a tree topology, most researchers are also interested in inferring the branch lengths of a tree. However, few studies have successfully inferred branch lengths using machine learning. While it may seem that this regression problem should be easier than the classification problem of inferring topologies, the size of the output vector depends on the number of edges in the tree—there are $2n-2$ branches in a rooted tree with n tips. The dependence on the number of tips complicates the use of machine learning approaches.

Suvorov and Schrider (2022) employed both a CNN and a multilayer perceptron (MLP) to infer branch lengths on fixed tree topologies with four or eight taxa. For the CNN-based approach, they adapted a previously proposed architecture (Suvorov et al., 2020). Instead of a classification task, the model was restructured for regression, aiming to predict all branch lengths simultaneously. Meanwhile, the MLP was fed with feature vectors derived from site pattern frequencies present within each alignment. Notably, the predictions generated by their models showed slightly superior accuracy compared to maximum likelihood estimates. Despite these promising results, there remains a degree of skepticism regarding the scalability of machine learning to infer branch lengths, especially when considering more species. Nevertheless, the flexibility of machine learning approaches with respect to the types of input data that can be considered offers many interesting possibilities. For instance, in the future such methods could facilitate the integration of heterogeneous fossil data in estimating time-calibrated trees.

As with topological inference, machine learning approaches can also be used to guide researchers in decisions about which approaches may be most appropriate for inferring branch lengths. For example, Tao et al. (2019) used a logistic regression model to predict whether rates of molecular evolution are autocorrelated in inferred phylogenies. Their approach, CorrTest, can be used to determine whether an independent branch-rate model or an autocorrelated branch-rate model should be used to estimate divergence times.

3 Other kinds of phylogenetic inferences

In addition to phylogenetic tree inference, machine learning approaches have been applied to both upstream and downstream tasks in phylogenetics. Prior to tree inference using many approaches (e.g., Bayesian inference, maximum likelihood, neighbor joining) it is necessary to infer a sequence substitution model. After tree inference, researchers are often interested in detecting and quantifying discordance, testing for introgression, and inferring macroevolutionary parameters. Below, we review some recent machine learning approaches to these upstream and downstream tasks.

3.1 Substitution models

It is crucial to select a suitable substitution model for accurate phylogenetic inference from sequence data, as it has long been known that misspecified models can lead to inaccurate estimates of trees (Buckley, 2002; Sanderson, 2002) and branch lengths (Abadi et al., 2019). Existing methods for model selection infer the model that provides the best fit to the data, using one of several criteria. Popular criteria include likelihood ratio tests (LRTs), Akaike information criteria (AIC), corrected AIC (AICc), Bayesian information criteria (BIC), and decision theory (DT). However, these criteria rely on assumptions that are often not met in phylogenetics, and there is a lack of consensus regarding which criteria are the most appropriate (Abadi et al., 2019). Additionally, substitution model choice tends to impact branch length estimates more-so than topology inference (Abadi et al., 2019), but no criteria to-date have been designed to select the model best-suited for branch length inference. Finally, using these criteria to perform substitution model selection is computationally expensive, as it requires computation of the likelihood. Here we discuss two recent machine learning approaches that attempt to address these gaps.

ModelTeller (Abadi et al., 2020) is a machine learning approach that uses an RF regressor to rank 24 potential substitution models according to their accuracy in downstream branch length inference. Features fed into the model included over 50 summary statistics that can be broadly categorized into four primary groups: features inherent to the alignment, features drawn from an approximated tree inferred through a distance-based method, parameters inferred under a parameter-rich substitution model, and sequence similarity within certain subsets. ModelTeller's primary distinction compared to traditional approaches lies in selecting a substitution model that improves accuracy in branch length inference. This leads to improved performance in terms of the accuracy of branch length estimates under the models selected using ModelTeller compared to models selected using more standard approaches, particularly on datasets simulated under realistic models. Additionally, ModelTeller was substantially faster than standard methods.

A later model, ModelRevelator (Burgstaller-Muehlbacher et al., 2023) aims to infer the correct generating model of nucleotide substitution using two neural networks. The first network, NNmodelfinder, takes as input a set of statistics calculated from pairwise alignments and predicts the best substitution model from a set of six possible models. The second network, NNalphafind, takes as input base composition profiles and predicts whether a site homogeneous model is appropriate or not. If a site homogeneous model is not appropriate, then NNalphafind estimates the α parameter of a model with Γ-distributed rate heterogeneity among sites. Used together, these networks can predict the best substitution model for a given sequence alignment, whether rate heterogeneity should be included, and, when rate heterogeneity is included, the α parameter to use in downstream inference. ModelRevelator performed comparably to maximum likelihood combined with substitution model selection under BIC as implemented in IQ-TREE (Minh et al., 2020), with substantially reduced computation times on large alignments.

Both ModelTeller and ModelRevelator are designed to select a substitution model that is suitable for inference; however, each uses different criteria for assessing suitability. ModelTeller is particularly focused on identifying a model that results in the most accurate estimates of branch lengths. The primary objective of ModelRevelator is to select the best substitution model and estimate the α parameter when the best model includes rate heterogeneity.

3.2 Levels of discordance

Gene tree topologies often differ from the species tree topology due to several biological factors, including incomplete lineage sorting, introgression, and gene duplication and loss (Maddison, 1997). Two recent studies used deep learning to estimate the amount of discordance in phylogenetic datasets (Rosenzweig et al., 2022; Zhang et al., 2023). Rosenzweig et al. (2022) used several approaches, including a deep neural network (DNN), to estimate the amount of discordance in four-taxon datasets using a set of summary statistics calculated from alignments and inferred gene trees. Estimates from their DNN were more accurate than relying on inferred gene trees alone to estimate discordance, particularly when branch lengths were long. In addition to their network for estimating the amount of discordance, they introduced a network for inferring the quartet species tree topology from the same set of statistics. Similarly, Zhang et al. (2023) used CNNs to estimate the proportion of all different possible topologies for four and five-taxon datasets from multiple sequence alignments. Their CNN, called ERICA, was able to accurately infer topology proportions. The authors then used these inferred proportions to try to infer introgression and to identify potentially introgressed genomic windows. The ability of these approaches to estimate the proportions of quartet topologies more accurately than standard pipelines—which rely on inferred gene trees alone—offers promise for improving many quartet-based methods for species tree inference, as these generally assume that quartet frequencies are accurately estimated from input gene trees (Mirarab & Warnow, 2015).

3.3 Introgression

Most machine learning approaches for studying introgression have focused on population-scale data, rather than phylogenetic data. For example, Schrider et al. (2018) used ExtraTrees classifiers to detect introgressed regions between closely related species, while Ray et al. (2023) used a CNN and image segmentation for a similar task. Similarly, Gower et al. (2021) developed a CNN to detect adaptive introgression given data from three closely related populations or species. Several recent papers have also addressed introgression from a phylogenetic perspective using machine learning.

Two recent studies used supervised machine learning to determine whether there was evidence for reticulation in a dataset. Blischak et al. (2021) used a CNN to detect various types of reticulation in four-taxon trees, including hybrid speciation and introgression. Their CNN took as input mean and minimum values of d_{XY} (a measure of sequence divergence) between sets of populations. They compared HyDe-CNN to an RF classifier trained on several phylogenetic statistics for detecting introgression and found that HyDe-CNN had increased power. In a similar approach, Burbrink and Gehara (2018) trained a neural network to distinguish a bifurcating species tree from models including reticulation between two parent clades and one clade with a putative reticulate history. As input, their network takes pairwise distances between all sequences in the phylogeny (11 sequences from three clades). Their network had moderate power to distinguish among models with and without reticulations. When applied to their empirical data, the model supported a reticulate history for a clade in which reticulation was also inferred using SNaQ (Solís-Lemus & Ané, 2016). Most recently, Hibbins and Hahn (2022) used supervised machine learning to distinguish speciation and introgression histories. Under many regions of parameter space, gene trees and site patterns matching the introgression history can become more common than those matching the species tree, challenging many traditional approaches to species tree inference. By using several summary statistics calculated from gene trees, Hibbins and Hahn were able to accurately infer the speciation history for rooted three-taxon trees, even in regions of parameter space where traditional approaches fail. While powerful, these approaches have primarily focused on four or fewer taxa. Future work may expand machine learning approaches to study introgression on larger trees.

3.4 Diversification rates

In addition to the kinds of inferences described above, recent studies have attempted to use inferred phylogenies for downstream inference of diversification rates. One challenge in any such analysis is determining the optimal way to encode phylogenetic trees. To address this issue, Voznica et al. (2022) introduced the compact bijective ladderized vector (CBLV), an encoding of phylogenetic trees that can be used as input into a CNN. They trained a CNN that took as input the CBLV to infer parameters of phylodynamic birth-death models and to perform model selection. They compared the performance of this CNN to a feed-forward neural network trained on summary statistics calculated from phylogenetic trees. Both networks were able to accurately infer parameters and distinguish among phylodynamic models. Lambert et al. (2023) used similar networks to infer speciation and turnover rates under a constant rate birth-death (CRBD) model and to infer the parameters of a binary state speciation and extinction (BiSSE) model. Lajaaiti et al. (2023) compared these networks to several other networks for inferring diversification parameters. They trained an additional CNN and RNN on lineage through time (LTT) plots. They also trained a graph neural network (GNN) that took phylogenies encoded as graphs directly as input. Under the CRBD model, the RNN and CNN trained on LTT plots outperformed the network trained on CBLV encodings. However, these same networks performed poorly under the BiSSE model, likely because the LTT plots did not include additional information about tip states, which was included in the other networks. Perhaps surprisingly, the GNN performed poorly across both models. These approaches highlight the importance of carefully choosing network architectures and data encodings for the task at hand.

4 Discussion

Recent progress has revealed the promise of machine learning in phylogenetics. However, inferences have often been limited to relatively small trees and relatively limited regions of parameter space. Moving forward, careful considerations of training datasets, network architectures, and data encodings will facilitate the use of machine learning to address fundamental challenges in phylogenetic inference.

Supervised machine learning requires a labeled training set. In the context of phylogenetics, however, we do not have labels for many real-world examples—we therefore have to simulate data. Despite attempts to simulate realistic data across a wide range of parameter space, biases will inevitably creep in. For example, training data generated under one substitution model may not generalize to empirical datasets that evolved under a different model. Importantly, this challenge is not specific to machine learning, and likelihood-based approaches may also fail due to model misspecification. The relative robustness of machine learning approaches and likelihood-based approaches to misspecified models remains unclear, with recent work suggesting similar impacts of model violations (Thompson et al., 2024). Just as it is important to evaluate the robustness of likelihood-based approaches to prevalent model misspecifications, it is important to evaluate the robustness of machine learning approaches to misspecifications of the model(s) used to simulate training data. Because of the flexibility of machine learning approaches, one approach to avoiding such biases would be to generate synthetic training data across increasingly large sets of models and parameters. However, this is computationally costly, and even when researchers attempt to consider a broad range of relevant parameters, there will inevitably be mismatches between training and empirical data, potentially leading to poor generalization to unseen data. To develop more robust networks, widely used techniques such as dropout, regularization, and ensemble methods can be employed. Alternatively, noise can be added to training data to improve generalization (as is done with image augmentation). In the context of phylogenetics, adding noise could involve masking regions of the alignment during training. Alternatively, techniques from domain adaptation have emerged as promising solutions. Domain adaptation aims to develop networks that are robust to differences between the distribution of training data and the distribution of target or empirical data. Mo and Siepel (2024) used domain adaptation to make more accurate inferences of recombination rates and selection coefficients in the presence of domain differences. Their approach used adversarial domain-invariant feature extraction, which incorporates an additional layer to prevent the model from extracting features that differ between the training and target data. Such an approach promotes the extraction of domain-invariant features, and could be used to make robust inferences in phylogenetics.

A major intended advantage of machine learning is that, once trained, models can be applied to new datasets with minimal computational expenses. Even though a trained model makes inferences almost instantaneously, training remains computationally expensive. Ideally, trained networks would be applicable across a wide range of empirical datasets, but this is limited by the details of the training data used and the choice of network architectures. Specifically, many network architectures (e.g., most CNNs) are not invariant to dataset size. In other words, only datasets with the exact dimensions of the training data can be analyzed. However, in phylogenetics, datasets may vary in size due to different alignment lengths or different numbers of taxa. This challenge has been addressed in population genetics through padding (Flagel et al., 2019), and by designing appropriate network architectures that are size invariant (Sanchez et al., 2021). Approaches that treat alignments as images in phylogenetics have often not considered alignments of variable sizes. However, Suvorov et al. (2020) used padding to accommodate simulated alignments that vary in length due to indels; since their model was only applicable to quartets, it did not consider variation in the number of taxa. Similarly, Wang et al. (2023) used a sliding window approach to accommodate variable alignment lengths. Approaches that rely on summary statistics can generally accommodate variable alignment lengths and numbers of taxa, as long as the statistics themselves do not change in dimensionality (Abadi et al., 2020; Burgstaller-Muehlbacher et al., 2023). Alternatively, Nesterenko et al. (2022) accommodated variable input sizes in Phyloformer through a carefully designed network, rather than through any manipulation of the input data. Moving forward, designing machine learning approaches that can be applied to alignments varying in size should be a central goal. To facilitate the reuse of networks in new empirical systems, techniques from transfer learning could also be used. Specifically, supervised transfer learning can be useful when limited training data are available from a new domain. For example, a network that has already been trained on data from one domain can be reused in a related, but distinct, domain. Supervised transfer learning and limited simulations in the new domain can be used to generate a robust network with reduced computational expenses compared to training the network from scratch. Combined, these approaches may facilitate more efficient uses of supervised machine learning in phylogenetic contexts.

Another major consideration is how to encode input data. Most commonly, encoded alignments (Suvorov & Schrider, 2022; Suvorov et al., 2020; Zou et al., 2020), or summary statistics (Abadi et al., 2020; Burgstaller-Muehlbacher et al., 2023) have been used as input. When using encoded alignments, a primary challenge is scalability to longer alignments or more taxa. This is especially pertinent as available genomic data continues to grow. Encoded alignments can also pose challenges to network reusability, as discussed above. Alternatively, the input can be represented with summary statistics that are explanatory features drawn from alignments and trees for the task at hand. However, selecting a good set of features relies on prior knowledge, and the choice of statistics can heavily impact inference. Alternative strategies for representing alignments have been proposed, using attention mechanisms (Burgstaller-Muehlbacher et al., 2023; Nesterenko et al., 2022; Rao et al., 2021) or language models (Lupo et al., 2022). Such approaches can lead to networks that can accept variable input sizes, and are capable of incorporating relationships among sites and lineages simultaneously. It is also essential to develop a suitable representation for phylogenetic trees. Several efforts in this direction have been made, from explanatory summary statistics (Voznica et al., 2022), to embeddings such as the CBLV (Voznica et al., 2022), to graphical representations in GNNs (Lajaaiti et al., 2023). While early uses are promising, these encodings have only been explored for a small set of inferential tasks, and it is unclear which encodings will prove most useful over a wider range of questions.

The promise of supervised machine learning is to efficiently consider a wide range of the complex processes that complicate phylogenetic inference. To date, most machine learning approaches for tree inference have largely not addressed heterogeneity introduced by incomplete lineage sorting (ILS), gene duplication and loss, and introgression (though several exceptions have been described here). While standard phylogenetic approaches also have trouble modeling this heterogeneity, machine learning shows potential to include multiple of these processes at once. For example, if machine learning approaches can be used to more accurately infer quartet frequencies in the presence of these processes (as demonstrated in the case of ILS by (Rosenzweig et al., 2022; Zhang et al., 2023)) then the accuracy of phylogenetic trees could be improved. Moving forward, we expect that creative network architectures, data encodings, and task designs will facilitate the use of machine learning to improve phylogenetic inferences in the presence of complex processes that cannot be accommodated by standard approaches.

4.1 Acknowledgements

This work was supported by a National Science Foundation grant to M.W.H. (DEB-1936187).

References

Abadi, S., Avram, O., Rosset, S., Pupko, T., & Mayrose, I. (2020). ModelTeller: Model selection for optimal phylogenetic reconstruction using machine learning. *Molecular Biology and Evolution*, *37*(11), 3338–3352.

Abadi, S., Azouri, D., Pupko, T., & Mayrose, I. (2019). Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications*, *10*(1), 934.

Azouri, D., Abadi, S., Mansour, Y., Mayrose, I., & Pupko, T. (2021). Harnessing machine learning to guide phylogenetic-tree search algorithms. *Nature Communications*, *12*(1), 1983.

Azouri, D., Granit, O., Alburquerque, M., Mansour, Y., Pupko, T., & Mayrose, I. (2023). The tree reconstruction game: Phylogenetic reconstruction using reinforcement learning. *arXiv*. https://doi.org/10.48550/arXiv.2303.06695

Balaban, M., Jiang, Y., Roush, D., Zhu, Q., & Mirarab, S. (2022). Fast and accurate distance-based phylogenetic placement using divide and conquer. *Molecular Ecology Resources*, *22*(3), 1213–1227.

Bhattacharjee, A., & Bayzid, M. S. (2020). Machine learning based imputation techniques for estimating phylogenetic trees from incomplete distance matrices. *BMC Genomics*, *21*(1), 497.

Blischak, P. D., Barker, M. S., & Gutenkunst, R. N. (2021). Chromosome-scale inference of hybrid speciation and admixture with convolutional neural networks. *Molecular Ecology Resources*, *21*(8), 2676–2688.

Bryant, D., & Steel, M. (2001). Constructing optimal trees from quartets. *Journal of Algorithms*, *38*(1), 237–259.

Buckley, T. R. (2002). Model misspecification and probabilistic tests of topology: Evidence from empirical data sets. *Systematic Biology*, *51*(3), 509–523.

Burbrink, F. T., & Gehara, M. (2018). The biogeography of deep time phylogenetic reticulation. *Systematic Biology*, *67*(5), 743–755.

Burgstaller-Muehlbacher, S., Crotty, S. M., Schmidt, H. A., Reden, F., Drucks, T., & von Haeseler, A. (2023). ModelRevelator: Fast phylogenetic model estimation via deep learning. *Molecular Phylogenetics and Evolution*, *188*, 107905.

Delsuc, F., Brinkmann, H., & Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviewes Genetics*, *6*(5), 361–375.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Felsenstein, J. (1978a). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, *27*(4), 401–410.

Felsenstein, J. (1978b). The number of evolutionary trees. *Systematic Zoology*, *27*(1), 27–33.

Flagel, L., Brandvain, Y., & Schrider, D. R. (2019). The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Molecular Biology and Evolution*, *36*(2), 220–238.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144.

Gower, G., Picazo, P. I., Fumagalli, M., & Racimo, F. (2021). Detecting adaptive introgression in human evolution using convolutional neural networks. *eLife*, *10*, e64669.

Haag, J., Höhler, D., Bettisworth, B., & Stamatakis, A. (2022). From Easy to Hopeless—Predicting the Difficulty of Phylogenetic Analyses. *Molecular Biology and Evolution*, *39*(12), msac254.

Hibbins, M. S., & Hahn, M. W. (2022). Distinguishing between histories of speciation and introgression using genomic data. *bioRxiv*. https://doi.org/10.1101/2022.09.07.506990

Huelsenbeck, J., & Hillis, D. (1993). Success of phylogenetic methods in the four-taxon case. *Systematic Biology*, *42*(3), 247–264.

Jiang, Y., Blaban, M., Zhu, Q., & Mirarab, S. (2023). DEPP: Deep learning enables extending species trees using single genes. *Systematic Biology*, *72*(1), 17–34.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

Kapli, P., Yang, Z., & Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, *21*(7), 428–444.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90.

Lajaaiti, I., Lambert, S., Voznica, J., Morlon, H., & Hartig, F. (2023). A comparison of deep learning architectures for inferring parameters of diversification models from extant phylogenies. *bioRxiv*. https://doi.org/10.1101/2023.03.03.530992

Lambert, S., Voznica, J., & Morlon, H. (2023). Deep learning from phylogenies for diversification analyses. *Systematic Biology*, syad044.

Leuchtenberger, A. F., Crotty, S. M., Drucks, T., Schmidt, H. A., Burgstaller-Muehlbacher, S., & von Haeseler, A. (2020). Distinguishing Felsenstein zone from Farris zone using neural networks. *Molecular Biology and Evolution*, *37*(12), 3632–3641.

Leuchtenberger, A. F., & von Haeseler, A. (2024). Learning from an artificial neural network in phylogenetics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. https://doi.org/10.1109/TCBB.2024.3352268

Liu, L., & Edwards, S. V. (2009). Phylogenetic analysis in the anomaly zone. *Systematic Biology*, *58*(4), 452–460.

Lupo, U., Sgarbossa, D., & Bitbol, A.-F. (2022). Protein language models trained on multiple sequence alignments learn phylogenetic relationships. *Nature Communications*, *13*(1), 6298.

Maddison, W. (1997). Gene trees in species trees. *Systematic Biology*, *46*(3), 523–536.

Mendes, F. K., & Hahn, M. W. (2018). Why concatenation fails near the anomaly zone. *Systematic Biology*, *67*(1), 158–169.

Minh, B., Schmidt, H., Chernomor, O., Schrempf, D., Woodhams, M., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, *37*(5), 1530–1534.

Mirarab, S., & Warnow, T. (2015). ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, *31*(12), i44–i52.

Mo, Z., & Siepel, A. (2024). Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data. *PLOS Genetics*, *19*(11), e1011032.

Nesterenko, L., Boussau, B., & Jacob, L. (2022). Phyloformer: Towards fast and accurate phylogeny estimation with self-attention networks. *bioRxiv*. https://doi.org/10.1101/2022.06.24.496975

Rao, R. M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., & Rives, A. (2021). MSA transformer. *International Conference on Machine Learning*, 8844–8856.

Ray, D. D., Flagel, L., & Schrider, D. R. (2023). IntroUNET: Identifying introgressed alleles via semantic segmentation. *bioRxiv*. https://doi.org/10.1101/2023.02.07.527435

Reaz, R., Bayzid, M. S., & Rahman, M. S. (2014). Accurate phylogenetic tree reconstruction from quartets: A heuristic approach. *PLOS ONE*, *9*(8), e104008.

Roch, S. (2006). A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *3*(1), 92–94.

Rosenzweig, B. K., Kern, A. D., & Hahn, M. W. (2022). Accurate detection of incomplete lineage sorting via supervised machine learning. *bioRxiv*. https://doi.org/10.1101/2022.11.09.515828

Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425.

Sanchez, T., Cury, J., Charpiat, G., & Jay, F. (2021). Deep learning for population size history inference: Design, comparison and combination with approximate bayesian computation. *Molecular Ecology Resources*, *21*(8), 2645–2660.

Sanderson, M. J. (2002). Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Molecular Biology and Evolution*, *19*(1), 101–109.

Schrider, D. R., Ayroles, J., Matute, D. R., & Kern, A. D. (2018). Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLoS Genetics*, *14*(4), e1007341.

Schrider, D. R., & Kern, A. D. (2018). Supervised machine learning for population genetics: A new paradigm. *Trends in Genetics*, *34*(4), 301–312.

Scornavacca, C., Delsuc, F., & Galtier, N. (2020). *Phylogenomics in the genomic era*. Open access book. https://hal.inria.fr/PGE

Siddall, M. (1998). Success of parsimony in the four-taxon case: Long-branch repulsion by likelihood in the Farris zone. *Cladistics*, *14*(3), 209–220.

Smith, M. L., & Hahn, M. W. (2023). Phylogenetic inference using generative adversarial networks. *Bioinformatics*, *39*(9), btad543.

Snir, S., & Satish, R. (2012). Quartet MaxCut: A fast algorithm for amalgamating quartet trees. *Molecular Phylogenetics and Evolution*, *62*(1), 1–8.

Solís-Lemus, C., & Ané, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genetics*, *12*(3), e1005896.

Solís-Lemus, C., Yang, S., & Leonardo, Z.-N. (2023). Accurate phylogenetic inference with a symmetry-preserving neural network model. *arXiv*. https://doi.org/10.48550/arXiv.2201.04663

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

Suvorov, A., & Schrider, D. R. (2022). Reliable estimation of tree branch lengths using deep neural networks. *bioRxiv*. https://doi.org/10.1101/2022.11.07.515518

Tao, Q., Tamura, K., U. Battistuzzi, F., & Kumar, S. (2019). A Machine Learning Method for Detecting Autocorrelation of Evolutionary Rates in Large Phylogenies. *Molecular Biology and Evolution*, *36*(4), 811–824.

Thompson, A., Liebeskind, B., Skully, E. J., & Landis, M. (2024). Deep learning and likelihood approaches for viral phylogeography converge on the same answers whether the inference model is right or wrong. *Systematic Biology*, syad074.

Voznica, J., Zhukova, A., Boskova, V., Saulnier, E., Lemoine, F., Moslonka-Lefebvre, M., & Gascuel, O. (2022). Deep learning from phylogenies to uncover the epidemiological dynamics of outbreaks. *Nature Communications*, *13*(1), 3896.

Wang, Z., Sun, J., Gao, Y., Xue, Y., Zhang, Y., Li, K., Zhang, W., Zhang, C., Zu, J., & Zhang, L. (2023). Fusang: A framework for phylogenetic tree inference via deep learning. *Nucleic Acids Research*, *51*(20), 10909–10923.

Zaharias, P., Grosshauser, M., & Warnow, T. (2022). Re-evaluating deep neural networks for phylogeny estimation: The issue of taxon sampling. *Journal of Computational Biology*, *29*(1), 74–89.

Zhang, Y., Zhu, Q., Shao, Y., Jiang, Y., Ouyang, Y., Zhang, L., & Zhang, W. (2023). Inferring historical introgression with deep learning. *Systematic Biology*, syad033.

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

Table 1: Recent machine learning applications in phylogenetics

Purpose	Method type	Algorithm/ architecture	Input/alignment format	Encoding	Output	Reference
Topology inference	classification	CNN	Nucleotide	Integer	Quartet topology	Suvorov et al., 2020
	classification	Residual NN	Amino acid	One-hot		PhyDL (Zou et al., 2020)
	classification	LSTM	Amino acid	Integer + Embedding		Solís-Lemus et al., 2023
	classification	CNN	Nucleotide	Integer	Tree topology	Fusang (Wang et al., 2023)
	regression	Transformer	Amino acid	One-hot	Pairwise evolutionary distances	Phyloformer (Nesterenko et al., 2022)
	regression	Matrix Factorization Autoencoder	Distance matrix with missing entries	None	An imputed distance matrix	Bhattacharjee & Bayzid, 2020
	regression	CNN	Reference tree and sequences from reference and query species	One-hot	Distances between the query and all backbone sequences	Jiang et al., 2023
	generative	GAN	Nucleotide	Integer	Tree topology	phyloGAN (Smith & Hahn, 2023)
Improving steps in topology inference	regression	Random forest	Phylogeny	Summary statistics	Ranking of possible SPR moves	Azouri et al., 2021
		Reinforcement learning	Nucleotide		Tree topology	The Phylogenetic Game (Azouri et al., 2023)
	classification	MLP	Nucleotide	Site pattern frequencies	Classification of alignment as Felsenstein- or Farris-type	F-zoneNN (Leuchtenberger et al., 2020)
	regression	Random forest	Nucleotide, amino acid, or morphological data	Summary statistics	The degree of difficulty of a phylogenetic dataset	Haag et al., 2022
Branch length inference	regression	MLP	Nucleotide	Site pattern frequencies	Branch lengths	Suvorov & Schrider, 2022
		CNN		Integer		
	classification	Logistic regression	Phylogeny	Summary statistics	Whether an independent branch-rates model should be rejected in favor of an autocorrelated model	CorrTest (Tao et al., 2019)
Substitution model selection	regression	Random forest	Nucleotide	Summary statistics	Ranking of substitution models based on their predicted performance in branch length estimation	ModelTeller (Abadi et al., 2020)
	classification	Residual NN	Nucleotide	Summary statistics	Model of sequence evolution	NNmodelfind (Burgstaller-Muehlbacher et al., 2023)
	classification and regression	Bidirectional LSTM	Nucleotide	Summary statistics	Whether rate heterogeneity should be considered, and if so an estimation of the shape parameter	NNalphafind (Burgstaller-Muehlbacher et al., 2023)
Discordance detection	regression	Linear regression Ensemble MLP	Nucleotide	Summary statistics	The amount of biological discordance in a set of gene trees	ml4ils (Rosenzweig et al., 2022)
	regression	CNN	Nucleotide	One-hot	The proportion of each possible topology for four- or five-taxon trees	ERICA (Zhang et al., 2023)
Introgression detection	classification	Extra-Trees classifier	Nucleotide	Summary statistics	Classification of a genomic region as introgressed or not	FILET (Schrider et al., 2018)
	classification	CNN (U-Net)	biallelic SNP matrix	Integer	Classification of alleles as introgressed or not	IntroUNET (Ray et al., 2023)
	classification	CNN	biallelic SNP matrix	Counts of minor alleles per haplotype per window	Classification of regions experiencing adaptive introgression	Genomatnn (Gower et al., 2021)
	classification	CNN	Nucleotide	Summary statistics	Best scenario of hybridization and admixture	HyDe-CNN (Blishak et al., 2021)
	classification	MLP	Nucleotide	Summary statistics	Best scenario of hybridization and admixture	Burbrink & Gehara, 2018
	classification	Various machine learning algorithms	Gene trees in coalescent units	Summary statistics	Distinguishing the speciation history from the introgression history	Hibbins & Hahn, 2022
Diversification rate inference	classification and regression	MLP CNN	Phylogeny	Summary statistics Vectorized representation	One of three possible phylodynamic models or estimates of phylodynamic model parameters	PhyloDeep (Voznica et al., 2022)
	regression	MLP CNN	Phylogeny with or without binary traits on tips	Summary statistics Vectorized representation	Estimates of diversification model parameters	Lambert et al., 2023
	regression	Various neural networks		Summary statistics, Vectorized representations, Graphs		Lajaaiti et al., 2023