Towards the next generation of species delimitation methods: an overview of Machine Learning applications

Matheus M. A. Salles[a*], Fabricius M. C. B. Domingos[a]

[a]Departamento de Zoologia, Universidade Federal do Paraná, Curitiba 81531-980, Brazil

[*]Corresponding author: matheus.salles@ufpr.br

ABSTRACT

Species delimitation is the process of distinguishing between populations of the same species and distinct species of a particular group of organisms. Various methods exist for inferring species limits, whether based on morphological, molecular, or other types of data. In the case of methods based on DNA sequences, most of them are rooted in the coalescent theory. However, coalescence-based models have limitations, especially regarding complex evolutionary scenarios, large datasets, and varying genetic data types. In this context, machine learning (ML) can be considered as a promising analytical tool, and provides an effective way to explore dataset structures when species-level divergences are hypothesized. In this review, we examine the use of ML in species delimitation and provide an overview and critical appraisal of existing workflows. We also provide simple explanations on how the main types of ML approaches operate, which should help uninitiated researchers and students interested in the field. Our review suggests that while current ML methods designed to infer species limits are analytically powerful, they also present specific limitations and should not be considered as definitive alternatives to coalescent methods for species delimitation. On the other hand, such variability might also represent an advantage, highlighting the flexibility of ML algorithms. Future enterprises should consider the constraints related to the use of simulated data, as in other model-based methods relying on simulations. We also propose best practices for the use of ML methods in species delimitation, offering insights into potential future applications. We expect that the proposed guidelines will be useful for enhancing the accessibility, effectiveness, and objectivity of ML in species delimitation.

Key words: bioinformatics, molecular data, speciation, phylogenetics, artificial intelligence, deep learning.

1. Introduction

1.1. Inferring species limits

Species represent fundamental entities across all biological disciplines. Consequently, the review, categorization, and characterization of taxa within this level constitute a pivotal aspect of biodiversity research (Bortolus, 2008; Vink et al., 2012; Ely et al., 2017). The process of identifying, characterizing, and defining a species is data-intensive and entails various practical dimensions. This complexity arises from managing extensive biological data and dealing with a range of theoretical elements, from the establishment of homologies, to taxon-specific traits, and the very philosophical notion of species. Furthermore, conceptual issues surrounding the definition of species concepts still attract debates among taxonomists and evolutionary biologists (Pante et al., 2015; Zachos, 2016). These discussions reach the realms of philosophy, because a multitude of data and methodologies will probably not fully solve many fundamental questions surrounding the nature of species (Zachos, 2016; Wilkins et al., 2022), or the ‘species ontology’ (what a species really is or represents). A complete resolution on this subject remains elusive, as it intertwines the empirical evidence biologists are able to extract from nature with philosophical definitions surrounding species concepts (Pigliucci, 2003).

One of the most popular modern definitions is the 'Biological Species Concept' (de Queiroz, 2005a; Zachos, 2016), which defines species as interbreeding populations reproductively isolated from others (Mayr, 1969; 1996; 2000). Yet, many challenges to this concept emerged throughout the years as empirical data clearly shows that the history of life on Earth does not fit into a bifurcating process (Edwards et al., 2016; Mallet et al., 2016), and a clear delineation of reproductive barriers is hindered by instances of asexual reproduction, natural hybridization and gene flow (Arnold, 1992; Shurtliff, 2013; Gompert et al., 2017). Hence, taxonomists and evolutionary biologists must recognize that multiple species definitions will coexist in the practice of species delimitation, and these are usually chosen based on the biological context of the organisms under study.

Another important concept, the General Lineage Concept (GLC), diverges from many others by prioritizing the recognition of independently evolving lineages over specific biological criteria such as reproduction or morphology (de Queiroz, 1998; 1999; 2007). According to the GLC, a species is defined as an independently evolving metapopulation lineage, emphasizing each species' unique evolutionary identity across time and space (de Queiroz, 2007). While unique morphological, ecological, or any other biological trait might be considered relevant in supporting the investigation of the speciation process, they are not mandatory criteria for species definition under the GLC perspective, but rather additional evidence supporting lineage separation (de Queiroz, 2007). Thus, this concept accounts for the contingent nature of the speciation process, where different biological properties may support species limits in varying degrees. It also emphasizes the need for multiple lines of evidence to corroborate hypotheses of species divergence, aligning with Integrative Taxonomy approaches (Wiens & Penkrot, 2002; Dayrat, 2005; Padial et al., 2010; Fujita et al., 2012).

The GLC also provides a theoretical distinction between the 'species ontology problem' (what a species is) and the 'delimitation problem' (how to operationally distinguish among putative species) (de Queiroz, 2007). Interestingly, while a clear relationship exists between these components, namely the species concept and species delimitation, historically, a significant part of the scientific efforts has focused on the former (see Sites Jr and Marshall, 2004; Wiens, 2007; de Queiroz, 2011; Hausdorf, 2011). The development of theoretical considerations related to species delimitation, in particular that based on molecular data, occurred mainly in the last two decades, accompanied by the introduction of new criteria and statistical methods (Lukhtanov, 2019; Rannala and Yang, 2020). Historically, identifying species limits, and describing new species, have primarily relied on morphological data (Wiens, 2007; Rannala, 2015; Rannala and Yang, 2020). However, morphological traits can be influenced by environmental factors, leading to convergence or divergence without necessarily reflecting genetic or evolutionary relationships between lineages (Price et al., 2003; Wake et al., 2011; Jarvis et al., 2014). Thus, genomic data has emerged as a crucial tool for inferring species limits, offering a more objective approach for species delimitation (Fujita et al., 2012), while complementing traditional morphological methods (Jörger and Schrödl, 2013).

Modern species delimitation methods (SDMs) aiming at identifying evolutionary units (Tautz et al., 2003; Vogler and Monaghan, 2007) mostly operate with molecular data under the principles of Coalescent Theory, notably, the multispecies coalescent (MSC; Rannala and Yang, 2003; Degnan and Rosenberg, 2009). The MSC analytical framework addresses various evolutionary assumptions while also managing different types of problems, such as conflicts among different gene trees, **incomplete lineage sorting** (terms in bold are defined in the Glossary, available in Appendix A), and errors in phylogenetic inference (Knowles & Carstens, 2007; Carstens et al., 2013; Jacobs et al., 2018). The use of modern SDMs has also grown due to advancements in statistical frameworks for phylogenetic inference (Edwards, 2009; O'Meara, 2012), along with Molecular Biology tools (e.g., next-generation sequencing (NGS); Slatko et al., 2018) and Bioinformatics (Searls, 2010).

Nonetheless, using SDMs with genetic data may fail to distinguish population structure from species-level divergence (Sukumaran and Knowles, 2017), and may also be affected by other issues associated with the reliance on the MSC model (Rannala and Yang, 2003; Degnan and Rosenberg, 2009; Edwards, 2009; Fujita et al., 2012). Some methods also have their functionality and performance compromised in scenarios when there is introgression between groups that constitute potential species (Rannala and Yang, 2010; Leaché et al., 2014; Jackson et al., 2017), and are more reliable in situations where gene flow ceases immediately after population divergence (Fujita et al., 2012; Smith and Carstens, 2020). Also, simulations have shown that ignoring gene flow leads the

MSC to overestimate **population sizes** and underestimate divergence times (e.g., Leaché et al., 2014). Hence, the effectiveness of the MSC framework is limited, to some extent, when additional processes influence divergence during speciation (Smith and Carstens, 2020). In any case, different SDMs have varying capabilities to address difficult evolutionary scenarios, and while such methods may introduce biases in certain situations, they are not inherently useless.

1.2. Machine learning, evolutionary biology, and species delimitation

Machine learning (ML), a branch of artificial intelligence (AI) known for its computational efficiency and predictive accuracy, has recently gained popularity in Evolutionary Biology mainly due to its ability to analyze and process large, complex, and high-dimensional datasets (Chicco, 2017; Borowiec et al., 2022; Fountain-Jones et al., 2021; Greener et al., 2021; Morimoto et al., 2021). In general terms, ML can be defined as a group of computational programs that can learn through experience (E) with respect to a class of tasks (T), and an evaluation measure (P), if its performance on the tasks of T, evaluated by P, increases with E (Mitchell, 1997). Many ML algorithms are known to be extremely useful in various aspects of biology. This includes photo-based species identification (Wäldchen and Mäder 2018), morphology-based species delimitation and description (Domingos et al., 2014; Breitman et al., 2018), biodiversity monitoring (McClure et al., 2020), behavioural studies (Valletta et al., 2017; Wang, 2019), DNA sequencing (Libbrecht and Noble, 2015; Liu, 2019), population genetics (Sheehan and Song 2016; Schrider and Kern, 2018; Fonseca & Carstens, 2024), ecology (Christin et al., 2019; Scalon et al., 2020; Pichler et al., 2020; Lürig et al., 2021; Silva et al., 2024), medicine (Sidey-Gibbons and

Sidey-Gibbons, 2019), microbiology (Qu et al., 2019), and more (see Borowiec et al., 2022; Fountain-Jones et al., 2021; Morimoto et al., 2021).

Therefore, its potential in evolutionary biology, and particularly in species delimitation, is evident (Karbstein et al., 2023). Specific examples can already be found in studies involving model selection in demography and phylogeography (Pudlo et al., 2016; Fonseca et al., 2021), speciation (Blischak et al., 2021), phylogenetics (Suvorov et al., 2020; Solis-Lemus et al., 2022 preprint; Smith & Hahn, 2023; Zaharias et al., 2022; Mo et al., 2024), and species delimitation (Pei et al., 2018; Derkarabetian et al., 2019; Smith & Carstens, 2020; Pyron et al., 2023), with the last one forming the primary focus of this review.

In the following sections, we provide an overview of ML applications in the context of species delimitation, with an emphasis on those that operate using molecular data.

2. Current ML applications for species delimitation

In the same way that there are two primary categories of ML, namely supervised and unsupervised learning (SML and UML, respectively), species delimitation methods can also be broadly categorized into two main groups: discovery and validation (see Carstens et al., 2013; Rannala, 2015). Discovery approaches involve grouping samples without prior information (Pons et al., 2006; O'Meara, 2010; Huelsenbeck et al., 2011), while validation approaches require researchers to first assign the samples to potential lineages (species hypotheses) before testing them (Flouri et al., 2018; Sukumaran et al., 2021). This draws a conceptual parallel between traditional discovery approaches and UML methods, and between validation methods and supervised algorithms (Fig.

1). Also, it is important to note that ML methods are likelihood-free species delimitation approaches, offering several advantages over **likelihood-based approaches**, such as eliminating the need for complex statistical calculations, making them computationally efficient and suitable for analyzing large datasets with many taxa.

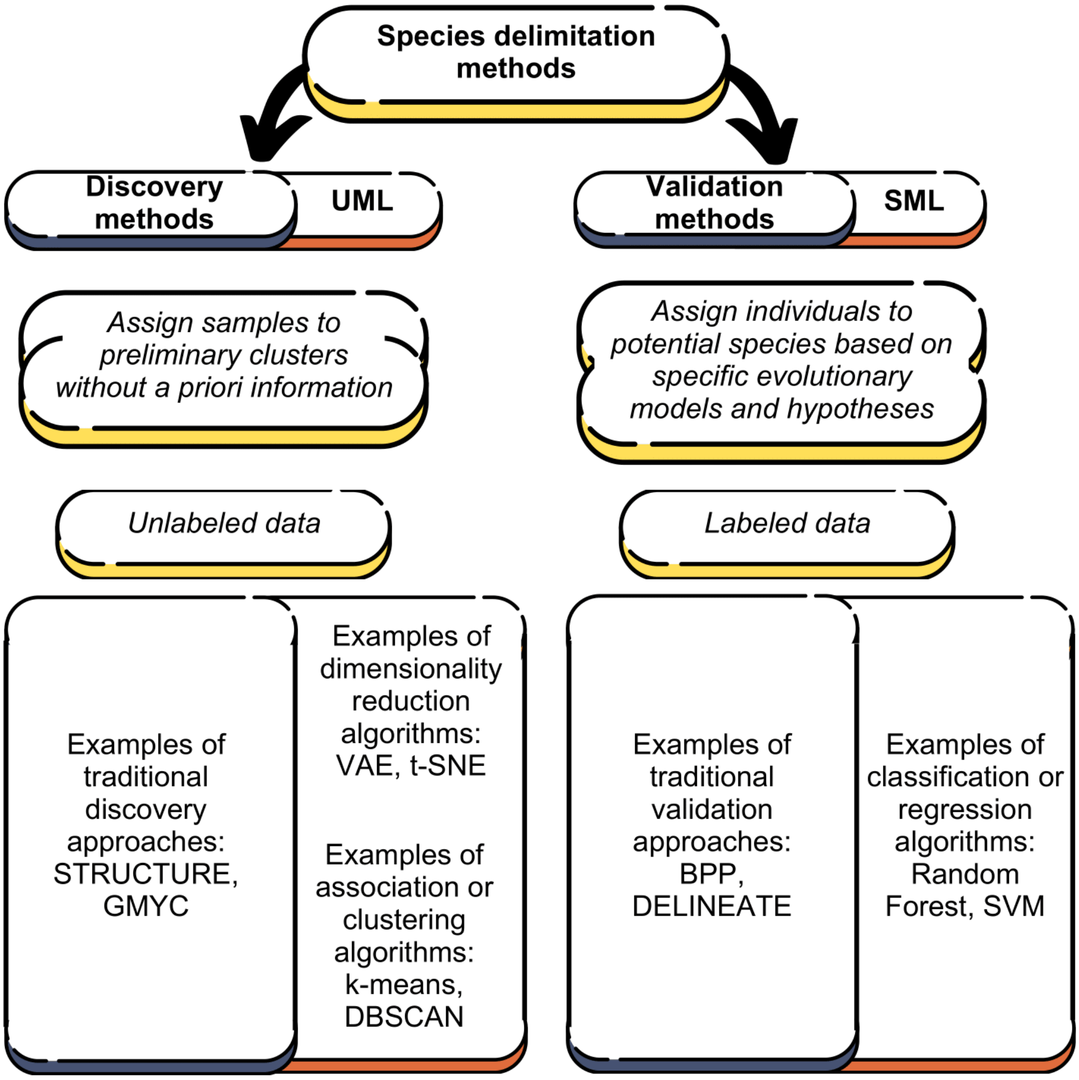

Species delimitation methods
Discovery methods
UML
Validation methods
SML
Assign samples to preliminary clusters without a priori information
Assign individuals to potential species based on specific evolutionary models and hypotheses
Unlabeled data
Labeled data
Examples of traditional discovery approaches: STRUCTURE, GMYC
Examples of dimensionality reduction algorithms: VAE, t-SNE
Examples of association or clustering algorithms: k-means, DBSCAN
Examples of traditional validation approaches: BPP, DELINEATE
Examples of classification or regression algorithms: Random Forest, SVM

Fig. 1. Comparative diagram categorizing species delimitation methods and machine learning algorithms, along with some of their key characteristics. Species delimitation methods can be broadly categorized as discovery and validation methods, akin to unsupervised and supervised machine learning algorithms, respectively.

Below, we present a comprehensive overview of recently applied ML methods in the domain of molecular species delimitation, emphasizing their computational attributes and underlying assumptions. Our selection process involved a thorough search across scientific literature repositories, databases, and online journals, with a specific emphasis on studies featuring ML methods and workflows explicitly designed for species limits inference. We prioritized research projects that either introduced novel methodologies (see Table 1) or enhanced and tested existing techniques in this context (Table A.1 in Appendix B). In our selection process, we focused exclusively on projects directly dedicated to species delimitation, despite the abundant literature on ML within related fields such as demography, population genetics, and phylogeography. Additionally, our emphasis is on methods designed for analyzing DNA sequence data. The categorized methods include SML, UML, and **deep learning**. While the backend processes may differ among such ML categories, their main goal when it comes to species delimitation usually remains the same: to analyze a given set of test data and classify it into distinct outcomes that define the species represented within the data.

Some studies applied ML techniques using other types of data rather than molecular information, such as morphology or ecology, for species delimitation and integrative taxonomy. A brief exploratory section regarding these particular studies can be found in Appendix B.

Table 1. List of proposed ML applications specifically designed to work on inferences about species limits.

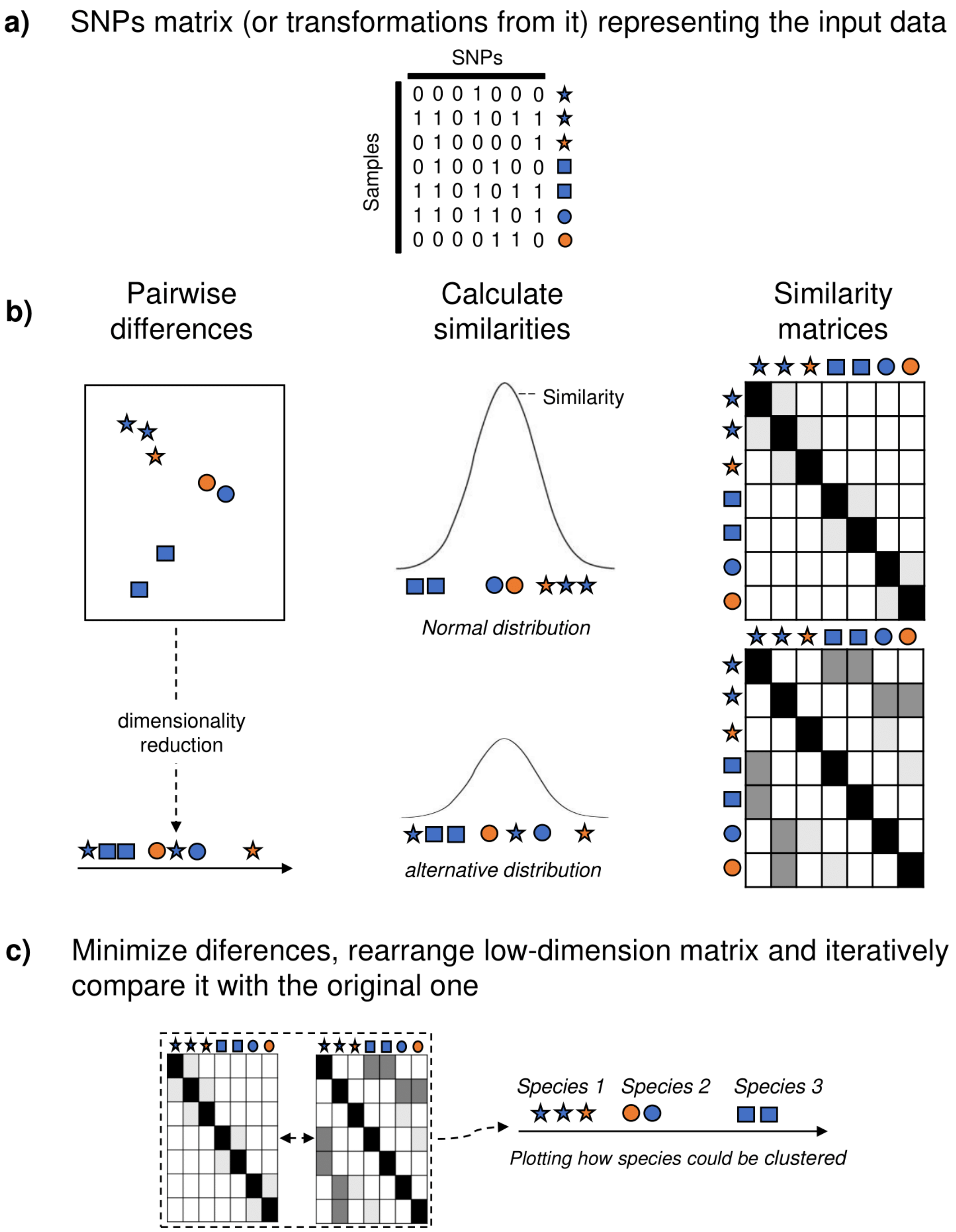
Reference	Languages	Category	Algorithms	Simulator	Input	Data representation
CLADES: A Classification-based Machine Learning Method for Species Delimitation from Population Genetic Data (Pei et al., 2018)[1]	python	SML	Support vector machines	MCcoal	Multiple sequence alignment (MSA) or SNP matrix	Population genetics summary statistics
A demonstration of unsupervised machine learning in species delimitation (Derkarabetian et al., 2019)[2]	R/python	UML	Variational autoencoders and t-Distributed Stochastic Neighbor Embedding	NA	SNP data matrix	One-hot-encoding of the SNP data matrix and *axis* from a discriminant analysis of principal components
Process-based species delimitation leads to identification of more biologically relevant species (Smith & Carstens, 2020)[3]	python	SML	Random forest	fastsimcoal	SNP data matrix	Folded multi-dimensional SFS
Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system (Perez et al., 2021)[4]	python	Deep learning	Convolutional neural networks	ms	SNP data matrix	Matrices (as images), with genotypes encoded as higher or lower frequency states
Speciation Hypotheses from Phylogeographic Delimitation Yield an Integrative Taxonomy for Seal Salamanders (*Desmognathus monticola*) (Pyron et al., 2023)[5]	R	UML	Self-organizing maps (SOMs)	NA	SNP data matrix	SNP matrix, in which the rows are individual specimens, the columns are the 2-4 possible states at each SNP locus, and the entries are the frequency of that state

Online repositories where it is possible to find more information about the currently existing platforms. [1] https://github.com/pjweggy/CLADES; [2] https://www.sciencedirect.com/science/article/abs/pii/S1055790319301721; [3] https://github.com/meganlsmith/delimitR; [4] https://github.com/manolofperez/CNN_spDelimitation_Piloso; [5] https://github.com/kyleaoconnell22/Pyron_et_al_UML_sp_delim/tree/main

2.1 Discovery and unsupervised methods

Unsupervised machine learning (UML) relies solely on the inherent data structure to find patterns within the data, whether by clustering similar data points together, reducing the dimensionality of the data while retaining essential information, or by identifying unusual patterns or outliers, which may indicate errors or novel phenomena (Hastie et al., 2009; Libbrecht and Noble, 2015; Dike et al., 2018). Consequently, UML algorithms operate without predefined assumptions about the dataset underlying structure, population parameters, species numbers, or sample categorization, making them particularly suitable for species delimitation where no prior hypotheses are put forward.

In terms of delimiting species, clustering or dimensionality reduction UML algorithms are generally employed (Fig. 2). Clustering methods group input data into subsets, where samples with high similarities are placed in the same cluster and exhibit less similarity with samples in other clusters. Dimensionality reduction focuses on compressing data to identify a smaller distinct set of variables that could capture essential features of the original data, while minimizing information loss. Thus, UML dimensionality reduction may provide intuitive data visualization and accommodate various data types (Libbrecht and Noble, 2015), being particularly effective when coalescent methods tend to oversplit potential species (Derkarabetian et al., 2019). In sum, UML algorithms enable the simultaneous use of diverse data types, mainly by extracting and condensing the necessary information to identify limits between biological groups (Pyron, 2023; Pyron et al., 2023).


a)
SNPs matrix (or transformations from it) representing the input data
SNPs
Samples
b)
Pairwise differences
Calculate similarities
Similarity matrices
Similarity
Normal distribution
dimensionality reduction
alternative distribution
c)
Minimize diferences, rearrange low-dimension matrix and iteratively compare it with the original one
Species 1
Species 2
Species 3
Plotting how species could be clustered

0	0	0	1	0	0	0	★
1	1	0	1	0	1	1	★
0	1	0	0	0	0	1	★
0	1	0	0	1	0	0	■
1	1	0	1	0	1	1	■
1	1	0	1	1	0	1	●
0	0	0	0	1	1	0	●

Fig. 2. Diagram outlining a potential UML workflow for species delimitation, utilizing the t-SNE algorithm (inspired by Derkarabetian et al., 2019). a) Data representation is the initial step, and it varies depending on the chosen ML tool, which may work with sequence data, SNP matrices, or population genetics metrics extracted from them. b) t-SNE, as a dimensionality reduction technique, iteratively finds a lower-dimensional representation of the original data. It identifies local similarity spaces between sample pairs by analyzing Gaussian and lower-dimensional distributions, such as the Cauchy or t-student with one degree of freedom. c) The algorithm's goal is to align the new similarity matrix with the original data by iteratively moving data points closer to their nearest neighbors in the higher-dimensional space and away from more distant ones. This process continues until the maximum number of iterations is reached or no further improvements can be made, resulting in the proper grouping of samples based on their similarities (e.g., individuals or populations assigned to a species based on the chosen data representation).

Derkarabetian et al. (2019) conducted a study to assess the performance of UML and deep learning methods for species delimitation. Their research highlighted the effectiveness of variational autoencoder (VAE) and t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithms for accurately identifying species clusters. In the case of VAE, single-nucleotide polymorphism (SNP) matrices were converted via 'one-hot coding', where nucleotides were transformed into binary variables (e.g., A = [1, 0, 0, 0]; C = [0, 1, 0, 0], and so on), including ambiguous bases (e.g., M = [0.5, 0.5, 0.0, 0.0]). This VAE approach employed multiple layers of encoding to compress high-dimensional input data, followed by the reconstruction of data through successive decoding layers. The latent variables, represented as a normal distribution with mean (μ) and standard deviation (σ), offered a two-dimensional depiction of the SNP matrix, facilitating a clear visualization that accounted for the uncertainty surrounding groupings due to standard deviations among samples and groups. In the case of t-SNE, data derived from a principal component analysis (PCA) was used as input variables, followed by clustering tests using the output from the UML algorithms. Both approaches yielded more readily interpretable outcomes compared to other methods assessed by the authors, revealing distinct species groupings in a two-dimensional space (Derkarabetian et al., 2019). Notably, the identified groups in this study corresponded to those of an integrative taxonomy approach considered by the researchers in their comparisons, suggesting that the limits identified by UML algorithms might correspond to species-level divergence rather than population structure (Derkarabetian et al., 2019).

Pyron et al. (2023) introduced a novel UML approach designed for delineating species limits from extensive genomic datasets, primarily grounded in **self-organizing maps (SOMs)**. This approach produces discrete outcomes rather than continuous ones, grouping genotypes based on similarity, and is posited as more advantageous than prior workflows. Additionally, the authors propose determining the number of species by analyzing the degree of grid occupancy in the SOM output. This quantification establishes how many units, representing distinct clusters of genotypes, have been effectively mapped from the original SNP matrix. Subsequently, the method estimates the cumulative distances from each sample to its immediate neighbors. To effectively separate these candidate species, Pyron et al. (2023) recommend performing cluster analyses, such as k-means. The determination of the optimal number of **classes** or species in the dataset is achieved by selecting the value that maximizes the sequential reduction in the weighted sum of squares from k to k + 1. Also, we highlight that this technique is rooted in the assessment of similarity rather than dissimilarity. An extension of this method has been proposed in the form of a SuperSOM approach, incorporating the possibility of using several trait classes simultaneously, such as alleles, morphological and ecological variables (Pyron, 2023).

2.2. Validation and supervised methods

While UML approaches are powerful and widely applicable, there are situations where supervised machine learning (SML) will offer analytical advantages. Unlike UML, a workflow for applying any SML method to population genetic data generally include data simulation for various evolutionary scenarios, encoding both simulated and observed genetic data into **feature vectors**, **training** the algorithm, assessing its predictive performance through accuracy estimates, and applying it to new observed data points (Fig. 3). Thus, the use of simulated genetic data based on known evolutionary models is essential, given the scarcity of adequately sized datasets with high-confidence labels in Evolutionary Biology.

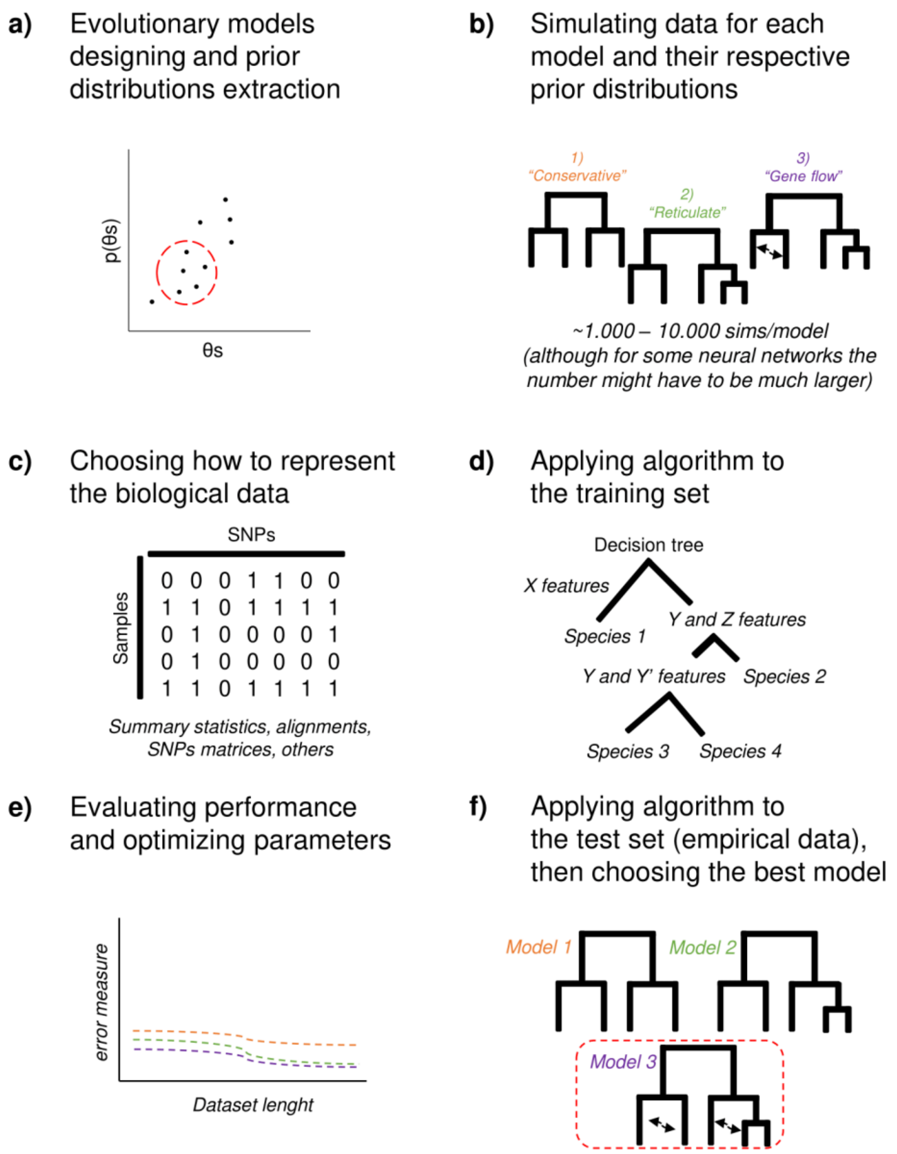

a) Evolutionary models designing and prior distributions extraction
p(θs)
θs
b) Simulating data for each model and their respective prior distributions
1) "Conservative"
2) "Reticulate"
3) "Gene flow"
~1.000 – 10.000 sims/model
(although for some neural networks the number might have to be much larger)
c) Choosing how to represent the biological data
SNPs
Samples
0 0 0 1 1 0 0
1 1 0 1 1 1 1
0 1 0 0 0 0 1
0 1 0 0 0 0 0
1 1 0 1 1 1 1
Summary statistics, alignments, SNPs matrices, others
d) Applying algorithm to the training set
Decision tree
X features
Species 1
Y and Z features
Y and Y' features
Species 2
Species 3
Species 4
e) Evaluating performance and optimizing parameters
error measure
Dataset lenght
f) Applying algorithm to the test set (empirical data), then choosing the best model
Model 1
Model 2
Model 3

Fig. 3. Diagram illustrating a potential SML workflow for species delimitation (inspired by the work of Smith and Carstens, 2020). a) The initial step involves designing priors for the evolutionary models considered in the study. b) Simulated data is generated for each model, typically ranging from 1,000 to 10,000 simulations per model, using relevant simulation software. c) The data is represented according to the requirements of the chosen ML tool. d) Following data simulation and representation, ML model training begins, involving various preliminary steps like data pre-processing, dataset division, feature selection, and algorithm choice. e) Model performance (both in terms of biological accuracy and computationally) is assessed using statistical metrics, allowing for retraining and adjustment based on the results. f) Once the model is adequately trained and evaluated, it can be used to predict species categories for new data, which can be either newly simulated data or empirical data consistent with the model's proposal, determining how many species exist in that particular biological system.

The process of training and applying ML algorithms is influenced by the assumptions of the underlying evolutionary processes, such as population size, selection strength, and gene flow. Thus, the reliability of results obtained from SML methods rely on the resemblance between the training data (typically simulated) and the actual biological data. Anyhow, SML algorithms generally demand a significantly smaller amount of simulated data compared to other methods based on simulations, such as **Approximate Bayesian Computation (ABC)**, resulting in reduced computational effort (e.g., a few thousand simulated datasets versus hundreds of thousands of simulations per scenario in most ABC approaches; Csilléry et al., 2010; Pudlo et al., 2016; Raynal et al., 2019).

CLADES (Pei et al., 2018), for example, is a SML approach designed for species delimitation, utilizing **classification models** trained and evaluated on *multilocus* sequence data. Notably, this study introduced the application of **support vector machines (SVM)** for species delimitation. For model training, datasets at the population level were simulated, with and without gene flow. Within this framework, species delimitation is framed as a classification task, where the goal is to classify pairs of populations as either belonging to the same or different species. Each training sample was represented as a list of summary statistics, and a SVM **regression** is calculated, through iterative training, to minimize the misclassification cost. Subsequently, the SVM classifier computed the probability of the training samples belonging to each potential grouping.

The training dataset was simulated based on a two-species model (A and B) where both species diverged at time τ with identical population size parameters (θA = θB = θ). Each species further consisted of two populations that recently split at time τp. **Migration** between species A and B was allowed at a rate of M = Nm migrants per generation, with m representing the migration rate per generation. The MCcoal software (Rannala and Yang, 2003) was used to simulate multilocus sequence data of length L under various parameter combinations for training. For each possible parameter combination (θ, τ, M), sequences were simulated for 100 loci with a length of L = 100Kbp for all populations. For each locus, 40 sequences were sampled, with 10 sequences per population. Additionally, symmetrical migration between species A and B was assumed before the populations of the species split at time τp. All training samples were combined to train a global classifier, enabling it to adapt to various values of θ and M instead of assuming fixed parameters. Longer loci improved

CLADES' efficiency, and it was robust to different modeling structures, accommodating various demographic events and evolutionary parameters.

Smith and Carstens (2020) introduced delimitR, a SML approach designed to conduct species delimitation in a model selection task; delimitR employs the multidimensional **site frequency spectrum** (mSFS) with a **binning** strategy as a predictor variable for a **Random Forest (RF)** classifier. Working with data summarized through the mSFS, delimitR facilitates the evaluation of models that vary in terms of lineage numbers. In essence, this framework aims to discriminate between various divergence models compatible with virtually any species concept, as asserted by the authors. Given its supervised nature, delimitR demands researchers to define reasonable priors, such as divergence times or migration rates, and to make decisions about the inclusion of models within the set (Smith and Carstens, 2020). For each model, Smith and Carstens (2020) simulated 10,000 mSFS. A RF classifier was constructed using 1,000 **decision trees** to accommodate the extensive number of models. delimitR's performance improved with larger SNP matrices and increasing divergence times. Compared to ABC methods, delimitR showed lower error rates, even though the detection of migration becomes challenging in cases of recent divergence between lineages (Smith and Carstens, 2020). The authors acknowledge that further research is needed to elucidate the association between the model space, number of parameters, and delimitation accuracy.

2.3. Deep learning

Artificial neural networks (ANNs) are increasingly employed in Evolutionary Biology, often referred to as 'deep learning' (Sheehan and Song, 2016). Deep learning techniques have found success in various fields in the Biological Sciences (Angermueller et al., 2016; Sheehan and Song, 2016; Schrider and Kern, 2018). However, its adoption in Evolutionary Biology is relatively recent (see Angermueller et al., 2016; Sheehan and Song, 2016; Blischak et al., 2021; Yelmen and Jay, 2023). The popularity of ANNs can be attributed to their highly flexible data input and output structure, allowing networks trained for one task to be repurposed for another by modifying their final **layers**, for instance, through **transfer learning** approaches. This versatility enables the resolution of intricate tasks that might prove challenging for **shallow learning** algorithms. Conversely, deep learning often demands meticulous and more specific fine-tuning compared to shallow learning methods.

The fundamental stages involved in creating a supervised shallow learning framework for species delimitation can be paralleled with the primary phases found in a deep learning workflow. These encompass data simulation and representation, **model** training and optimization, all the way to predicting the relevant categories from empirical data (Fig. 3). For a detailed description of how neural networks work, and their general structure, see Sheehan and Song (2016), Borowiec et al. (2022), and Korfmann et al. (2023).

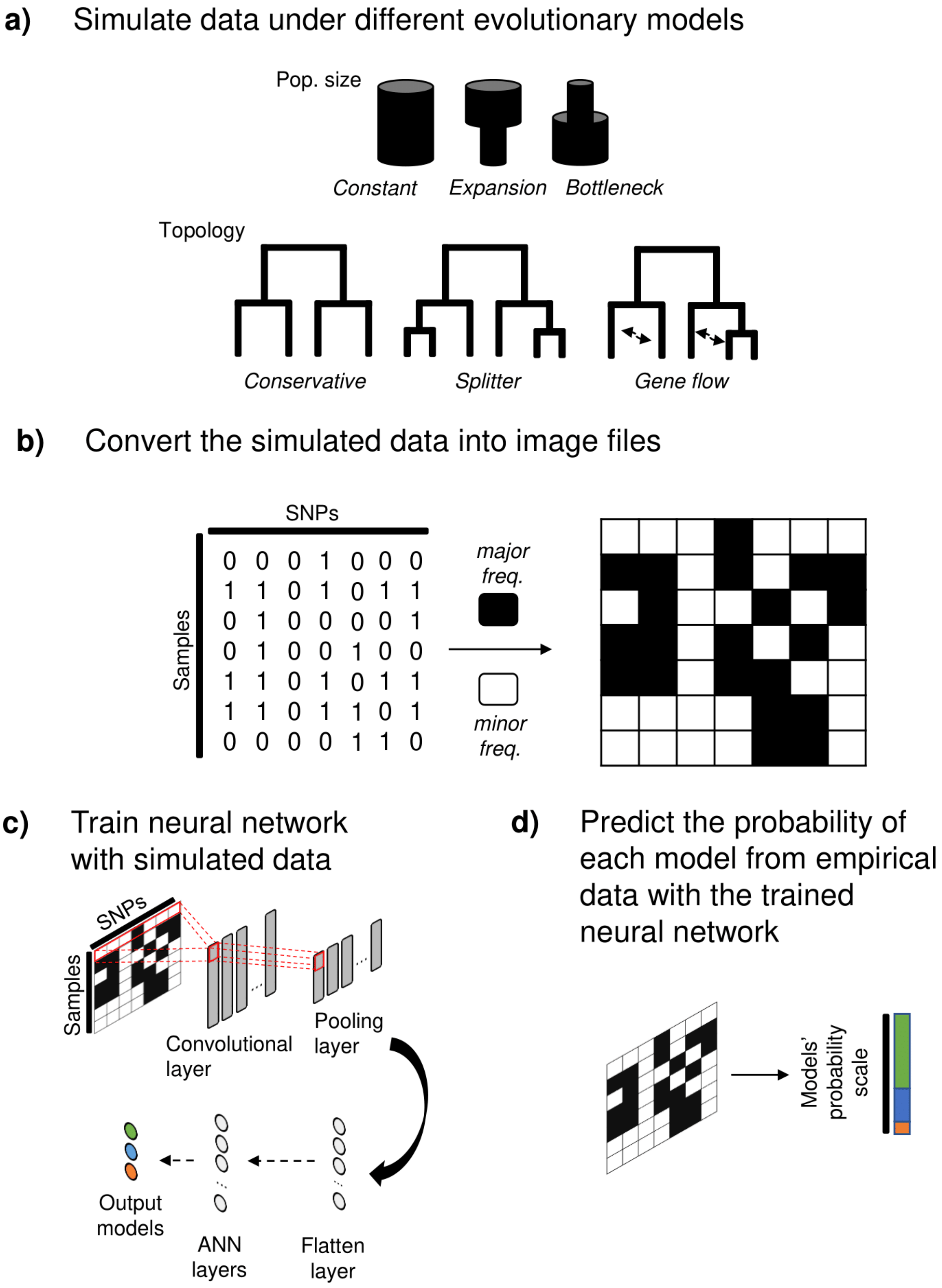


Fig. 4. Diagram illustrating a potential deep learning workflow applied in the context of species delimitation, using CNNs (inspired by Perez et al., 2021). a) The process typically begins with the simulation of biological data under various evolutionary models, considering factors like topology, population size, gene flow, and more, similar to SML. b) Next, data representation is crucial. For CNNs, SNP matrices are often converted into arrays or image files, where pixel contrast reflects differences in minor and major frequencies between samples. c) With the simulated and properly represented data, the network training phase can commence. The parameter configuration and network architecture may vary, depending on the specific study's requirements. d) Once each model is trained and its performance is rigorously evaluated, the final stage of the workflow involves predicting categories for new data. This can include using new simulated data with slight parametric modifications, still within the trained model's limits, as well as empirical data whose evolutionary history aligns with the proposed model. In both cases, the goal is to determine which delimitation model best applies to the biological system being investigated.

Perez et al. (2021) propose a species delimitation approach that accommodates the integration of coalescence-based methods with model selection using **convolutional neural networks (CNNs)**. Briefly, this approach can combine models from coalescent analyses, such as using BPP (Flouri et al., 2018; 2020), allowing for the comparison of different evolutionary scenarios. Thus, it allows for the test of species limits by integrating data from various sources, including the possibility of incorporating knowledge from both genetic analyses using coalescence-based methods and morphological hypotheses reflecting diverse taxonomic arrangements. The initial steps involve simulating genetic data for each delimitation hypothesis, with the study encompassing 10,000 simulations per model, and transforming them into images. These images of simulated data are used to train a neural network capable of recognizing simulations generated from each model. Then, each species hypothesis probability can be predicted through CNNs using a **test set**. In the same study, the authors conducted a comparison between their model selection approach and ABC using empirical data. It is worth noting that while CNNs used 10,000 simulations per model, ABC required 100,000 simulations per model. The CNNs consistently demonstrated superior performance in distinguishing between the simulated demographic scenarios, outperforming ABC in all cases, with fewer simulations and faster execution times (Perez et al., 2021).

2.4 How has machine learning changed our approach to delimit species so far?

To date, relatively few studies (<20, also see Appendix B) have specifically explored ML techniques for species delimitation, particularly when focusing on molecular data. Among these, only five introduced novel ML approaches for species delimitation, providing comprehensive details from initial simulations to statistical performance evaluations (Pei et al., 2018; Derkarabetian et al., 2019; Smith and Carstens, 2020; Perez et al., 2021; Pyron et al., 2023). These approaches, and also other ML frameworks applied in phylogeography and demographic inferences, are often advocated by the researchers and developers themselves on the following arguments: i) challenges and limitations associated with the assumptions of coalescent methods (Derkarabetian et al., 2019; Smith and Carstens, 2020; Blischak et al., 2021; Martin et al., 2021; Derkarabetian et al., 2022); ii) ML computational efficiency and the capacity of handling complex evolutionary models (Pei et al., 2018; Martin et al., 2021; Perez et al., 2021; Derkarabetian et al., 2022; Pyron et al., 2023); and iii) ML acting as a likelihood-free approach, enabling the consideration of models where likelihood computation would be intractable (Smith and Carstens, 2020; Martin et al., 2021; Perez et al., 2021; Sanchez et al., 2020). While ML algorithms are often used similarly to simulation-based approaches like ABC, additional steps are generally incorporated, such as: i) selecting a more informative subset of summary statistics based on specific criteria (Smith and Carstens, 2020; Martin et al., 2021), and ii) handling larger or more complex genetic datasets compared to what Bayesian methods can do in a reasonable amount of time (Ghirotto et al., 2021; Smith and Carstens, 2020; Collin et al., 2021).

2.5. What types of species ML methods might be detecting?

A significant part of the studies we analyzed were philosophically based on species concepts grounded on evolutionary or genealogical independence criteria. This might stem from our focus on workflows using molecular data, which generally aims at identifying lineages and genetic clusters characterized by significant levels of genetic divergence and restricted amounts of gene flow. Also, some studies specifically model parameters like migration, which make them in line with concepts focused on reproductive criteria. While evolutionary and genealogical independence evidence (or reproductive criteria) may have their limitations in investigating species limits, results generated by ML methods in this context can still serve as hypotheses for further investigations (e.g., Fujita et al., 2012), aligning with the GLC perspective (de Queiroz, 1998; 1999; 2005b).

In this context, it is reasonable to assert that ML-based delimitation methods, just as coalescence-based methods, might not always be identifying species *per se*, but rather: i) incompletely separated (or incipient) species, which may eventually be classified as distinct (Burbrink et al., 2021), or even as 'subspecies' (de Queiroz, 2020); or ii) population or phylogeographic variation (Rosenblum et al., 2012; Sukumaran et al., 2021). Consequently, while ML methods hold increasing promise for species boundaries inference, it is necessary to evaluate the extent to which the ML methods could effectively discern evolutionary independence among metapopulation lineages. So far, there are no definitive coalescent-based solutions to differentiate between population structure and species (Sukumaran & Knowles, 2017; Leaché et al., 2019). Thus, while model-based evolutionary lineage structure detected through ML can be biologically relevant for species delimitation, additional data and an evolutionary process-based perspective are crucial to discern the nature of the inferred biological entities (Smith & Carstens, 2020; Sukumaran et al., 2021).

Inferring species limits from molecular data and integrating phenotypic data can be a solution in some cases, but robust species delimitation still requires mechanistic hypotheses about the speciation process itself (Padial & De la Riva, 2021; Pyron et al., 2023; Pyron et al., 2024), because distinguishing between population structure and actively diverging or collapsing species require explicit hypotheses and quantifiable tests (Sukumaran & Knowles, 2017; Derkarabetian et al., 2019; Huang, 2020; Pyron et al., 2024). Just as phenotypic, ecological, or other biological attributes are not mandatory criteria for designating an evolutionary lineage as a species (de Queiroz, 2007; Pyron et al., 2023), genetic or genealogical groupings identified using ML-based delimitation methods can be similarly interpreted. Within this context, while the primary criterion for recognizing a species can still be evolutionary independence, other characteristics may serve as secondary evidence of divergence and could be also analyzed using ML frameworks.

Due to its great versatility in handling diverse data types, ML future applications to infer species limits may also focus on evaluating which of the different biological properties could be most effectively integrated into the species hypotheses testing process. They may be useful in discerning between patterns of population structure and species-level divergence, especially through the integration of distinct traits, such as genomic divergence, gene flow, ecological adaptation, and phenotypic differentiation (Freedman et al., 2023; Prates et al., 2023; Pyron et al., 2024). Again, this approach aligns with de Queiroz's GLC (1998; 1999; 2005), providing a deeper understanding of the speciation processes through multiple biological perspectives.

Only a few detailed ML pipelines have been proposed so far to explore the relationships between evolutionary models and divergence scenarios in terms of distinct characteristics, whether genetic, phenotypic, geographic or ecological. For example, Yang et al. (2022) introduced a CNN method that successfully integrates morphological and molecular data for species identification. Pyron (2023), on the other hand, implemented a UML method using SOMs for learning high-dimensional associations between observations (e.g., specimens) across a wide set of input features (e.g., genetics, geography, environment, and phenotype). Future methodologies could further explore this integration of multiple sources of information, both regarding species delimitation and integrative taxonomy.

3. Advantages, limitations and future perspectives

3.1. Strengths and benefits of using ML to delimit species

In general, ML methods applied to infer species limits offer some advantages over coalescent or traditional simulation-based methods. Despite particular constraints, ML algorithms can perform as well as or even outperform (in terms of biological accuracy) traditional model selection tools and likelihood-based species delimitation methods (Pei et al., 2018; Smith and Carstens, 2020; Perez et al., 2021; Derkarabetian et al., 2022). Moreover, being likelihood-free, they are computationally more efficient and generally can be trained on models that are at times too intricate for formal statistical estimators (Pei et al., 2018; Kuzenkov et al., 2020; Smith and Carstens, 2020; Suvorov et al., 2020; Martin et al., 2021; Perez et al., 2021). Some of these algorithms have proven to be highly efficient in complex evolutionary scenarios, including situations involving gene flow or population size fluctuations (Pei et al., 2018; Perez et al., 2021). This efficiency does not compromise the ability to distinguish between different models (Smith et al., 2017), and even simple SML methods provide high selection accuracy when comparing multiple models in a single analysis (Gehara et al., 2020 preprint).

Specifically, when it comes to deep learning, a major advantage is their capacity to automatically extract information from alignments (commonly treated as images), as opposed to relying on summary statistics typically required by other ML methods. This facilitates accurate and efficient classification or regression tasks, as observed in studies by Sanchez et al. (2020), Fonseca et al. (2021), Perez et al. (2021), and Borowiec et al. (2022), thus holding promise in future species delimitation studies. Besides, especially in supervised approaches, which often use explicit speciation models to validate species (e.g., Smith and Carstens, 2020), ML enables a more in-depth exploration of the speciation and phylogeographic processes that underlie the formation of independent evolutionary lineages. Thus, given that properly sampled genomic datasets can offer sufficient data for analyzing complex evolutionary models, ML might serve a dual role: providing primary evidence for examining species limits patterns, and assisting in the investigation and reconstruction of the evolutionary processes responsible for these patterns.

3.2. Constraints regarding ML and species delimitation

Certain algorithms, especially those in SML or deep learning, can be overly specialized. Modern ML methods are proficient at interpolating within the observed range of values in the training data, even in cases where specific values have not been encountered before, being adaptive and not solely reliant on memorizing specific training instances. Even so, because such algorithms are typically trained on simulated data with specific values of evolutionary parameters, such as θ and M, their performance might be compromised when applied far outside the training parameter space (Schrider and Kern, 2018; Borowiec et al., 2022). Besides, ML algorithms have some degree of **inductive bias** (Hüllermeier et al., 2013). Therefore, exploring in further details the association between training capacity and predictive power should be a priority for future studies.

Methods relying on a substantial volume of simulated data across diverse evolutionary scenarios need to consider the careful design of prior distributions to simulate models that closely resemble the real biological system under investigation. This challenge becomes more pronounced for non-model organisms, where data availability may severely limit the quality of parameter estimates (Tagu et al., 2014; Fonseca et al., 2016; Cerca et al., 2021; Jorna et al., 2021). Nonetheless, these simulation problems are not exclusive to ML-based workflows, as model selection frameworks such as ABC also employ simulated data (Beaumont et al., 2002; Bertorelle et al., 2010). All model-based methods depend on the specified models and its parameters, whether they are used for simulations or for direct likelihood estimation. Thus, traditional species delimitation methods that do not require simulations remain important alternatives for addressing delimitation challenges, in particular when there is no clear reference for simulations. Coalescence-based inferential methods are also limited in terms of their coverage of different evolutionary scenarios, but they possess optimality and iterability properties that span a reasonable portion of the parameter space, albeit at a considerable computational cost (e.g., Flouri et al., 2018;

Sukumaran et al., 2021). Nevertheless, methods not reliant on simulations can also be sensitive to model misspecification, as the MSC deals with assumptions that may not be appropriate for many biological systems.

Either wat, it may be unfeasible to simulate data or train an ML algorithm across an entire parameter space, especially in complex evolutionary models (Rannala and Yang, 2020). Limited information is available regarding the asymptotic statistical performance of most ML methods applied for species delimitation, and important phenomena may be entirely missing from the simulations (e.g., background selection, Mo and Siepel (2023), or missing data Arnab et al. (2023)). Thus, such models may never be comprehensive enough, have limitations in representing real data, and demand substantial computational resources (Arenas, 2012; Mangul et al., 2019a; Zaharias et al., 2022). This leads to an inherent challenge in avoiding some degree of misspecification in the training data, even considering the variety of powerful genetic data simulators currently available, such as SLiM (Messer, 2013), discoal (Kern and Schrider, 2016), msprime (Baumdicker et al., 2021), and fastsimcoal2 (Excoffier et al., 2021).

Another crucial perspective to consider is that numerous studies, whether focusing on species delimitation, demography, or population genetics, incorporate ML for inferences based on summary statistics (Pei et al., 2018; Smith and Carstens, 2020; Collin et al., 2021; Ghirotto et al., 2021). There are methodologies tailored for handling data derived from SNP matrices (Derkarabetian et al., 2019; Sanchez et al., 2020; Smith and Carstens, 2020; Blischak et al., 2021; Fonseca et al., 2021; Martin et al., 2021; Perez et al., 2021) or raw sequence data (Pei et al., 2018; Ghirotto et al., 2021), and only a few pipelines offer extensibility to various genetic markers (e.g., Collin et al., 2021). Notably, deep learning techniques are valuable tools in this context, offering the capability to analyze both raw genetic data and summary statistics (Korfmann et al., 2023).

While summary statistics can also be derived from the original genetic data and are valuable for distinguishing between simulated models, not all of them may be suitable for making inferences about species limits. The practical implementation of such statistics on the detection of specific evolutionary processes often encounters confounding factors that can mimic similar effects on gene histories (Flagel et al., 2019). For example, Tajima's D is a statistic sensitive to both positive selection and changes in population size (Simonsen et al., 1995). Moreover, since different studies often employ their specific set of summary statistics, comparing the results of ML applications is not always straightforward, or feasible, without acknowledging the significant nuances tied to the biological context considered in each approach. Thus, the tendency of some ML algorithms to rely on specific representations of data rather than the original dataset can be seen as a drawback in certain scenarios. Unless we precisely know which type of data is truly sufficient to represent the target data, an approach solely based on a particular set of summary statistics can inevitably result in a degree of information loss (Rannala and Yang, 2020).

An alternative to learning from summary statistics is to consider the alignment itself as input, as demonstrated in the CNNs approach introduced by Perez et al. (2021). Along with other deep learning techniques, CNNs implicitly enable dimensionality reduction while capturing structures within the input data. Thus, comparing different ML approaches might be misleading due to the variability in the biological foundations employed in each workflow. In other words, it is not always reasonable to strictly compare results produced by different ML approaches, as they are generally trained on specific parameters and data representation.

3.3. Possible avenues and prospects for future studies

Regarding ML, one approach to mitigate the effects of misspecification during simulation involves designing or using a simulator that enforces greater compatibility between simulated and actual data. Generative adversarial networks (GANs), a type of deep learning algorithm commonly used for creating synthetic images and voices (Chadha et al., 2021), have shown promise in this regard (see Callier, 2022; Wang et al., 2021). GANs operate with two networks, the generator and the discriminator, trained together (Goodfellow et al., 2014). While the generator generates simulated data, the discriminator distinguishes between real and fake data. Over the course of training, the generator network becomes more powerful at producing realistic **examples**, and the discriminator network becomes more skilled at distinguishing between real and synthetic data. Once training is complete, the generator network can be utilized to generate new examples that are indistinguishable from real data, providing a reliable way to work with **labeled data**. Researchers have already assessed the utility of GANs in various fields, including genomics, phylogenetics, and population genetics (Booker et al., 2023; Nesterenko et al., 2022 preprint; Yelmen and Jay, 2023). Smith and Hahn (2023), for instance, introduced phyloGAN, a workflow that takes a concatenated alignment (or a set of alignments) as input and infers a phylogenetic tree, potentially accounting for gene tree heterogeneity.

While such approaches perform effectively in relatively straightforward scenarios, challenges still emerge as the complexity of evolutionary model space increases. This complexity might stem from more variables in evolutionary models or larger trees and alignments, resulting in potential issues related to accuracy and execution time (Nesterenko et al., 2022 preprint; Smith and Hahn, 2023). Consequently, it is important to recognize that applications of GANs in the field of Evolutionary Biology are still in the early stages of development. To fully harness the potential of this tool in species delimitation, further efforts are required to refine estimates of genetic population parameters (e.g., Wang et al., 2021). Future advancements in GANs within the realm of evolutionary biology should focus, for instance, on enhancing the efficiency of exploring parameter spaces, reducing computational training times, and accommodating more complex models (Smith and Hahn, 2023).

Besides, issues related to potential errors in data simulation can be likened to a "domain adaptation" problem, where a model trained on one data distribution is applied to a dataset originating from a different distribution (Farahani et al., 2021; Mo and Siepel, 2023). A classic illustration of domain adaptation is found in image classification. Consider a situation in which a recognition model needs to identify different dog breeds from photographs ("target domain"), but there is an abundance of labeled training data available only in cartoon drawings of dogs ("source domain"). In such cases, a ML model must be trained on one dataset with the expectation of performing well on another, even in the presence of systematic differences between the two distributions.

Recent approaches typically involve learning a "domain-invariant" data representation through a feature extractor neural network. This is accomplished by minimizing domain disparities (Rozantsev et al., 2018), utilizing adversarial networks (Ganin and Lempitsky, 2015; Liu and Tuzel, 2016; Bousmalis et al., 2017), or employing auxiliary reconstruction tasks (Ghifary et al., 2016). Domain adaptation techniques have found applications in fields such as genomics (Cochran et al., 2022) and population genetics (Mo and Siepel, 2023), particularly as an unsupervised domain adaptation problem. Through extensive simulation studies, Mo and Siepel (2023) convincingly demonstrated that their domain-adapted models significantly outperformed standard networks across various simulation misspecification scenarios. This outcome underscores the potential of domain adaptation techniques as a promising avenue for developing more robust deep learning models in the realm of population genetic inference (Mo and Siepel, 2023), potentially including species delimitation.

In addition to the limitations regarding simulations and training models in specific parameter spaces, there is the issue associated with the manipulation of data attributes and different types of input data. This becomes even more relevant considering that ML techniques are lauded for their adaptability, especially considering transfer learning frameworks. A neural network initially trained for a specific task can be repurposed for different learning contexts with the simple modification of some of its layers, even though reusing trained models can be very challenging due to differences in data dimensionality (Sanchez et al., 2020). As an example, a deep learning **architecture** originally trained for inferring historical population sizes can be repurposed for classifying demographic scenarios (Pan and Yang, 2010). Also, deep learning methods used for phylogeographic model selection (Fonseca et al., 2021) could be easily applied to species limits issues with minimal adaptations.

4. Optimizing the use of ML in the context of species delimitation

4.1. Enhancing Species Delimitation through accessible and purpose-built ML

The introduction of new ML approaches will increasingly enhance researchers' ability to make biologically precise decisions, especially when these methods are purpose-built, from conception to implementation, for the specific task of delimiting evolutionary lineages. A critical step in any species delimitation study is to select the appropriate methods to be employed, considering the available data and putative evolutionary scenarios. With a multitude of possibilities in the modern Evolutionary Biology toolkit, the ideal choice should not only consider an appropriate fit with the biological problem under investigation, but also a statistical evaluation and performance optimization (Greener et al., 2021; Morimoto et al., 2021), under various diversification scenarios, while estimating historical parameters like divergence time, population size, and migration rate. It is important to assess in which specific evolutionary scenarios coalescent methods might exhibit strong limitations, and whether a new ML workflow might outperform others in terms of performance. Thus, a comprehensive analysis of the methods characteristics, advantages, disadvantages, and overall performance compared to existing SDMs is desired.

Such evaluation should also encompass both the algorithm's biological predictions and computational performance. Comparisons should be performed considering the inherent properties of the used ML algorithms, such as how the workflows manipulate the data attributes, and the different types of input and output data. In nearly all studies using ML methods to infer species limits, at least a minimal approach to estimating error or noise is typically employed (Pei et al., 2018; Smith & Carstens, 2020; Martin et al., 2021; Derkarabetian et al., 2022). For example, it is common for researchers to evaluate the ML model's performance using genetic datasets of varying sizes, or alignments of different dimensions. Then, the quantity and quality of data clearly influence the effectiveness of ML applications, as analyses conducted on larger, well-filtered datasets consistently yield better results (Pei et al., 2018; Smith & Carstens, 2020; Martin et al., 2021; Derkarebetian, et al., 2022). This effect is pronounced in UML approaches, as they tend to be more susceptible to data-related issues (Martin et al., 2021).

From a practical perspective, evaluating the suitability of an ML tool for species delimitation also involves assessing its accessibility, particularly when compared to traditional SDMs. To promote the widespread adoption of ML tools in species delimitation, it is crucial to ensure that analyses are accessible and reproducible. For example, a thorough description of the ML method, but without a detailed reference to the dataset, can lead to significant issues within the workflow (Chicco, 2017; Greener et al., 2021). The same rationale extends to the availability of the trained models. For example, Derkarebetian et al. (2022) assessed a ML approach's capability to delimit cryptic species, and constructed a "customized" training dataset from a well-studied lineage with biological characteristics akin to their focal taxon. In cases like these, where a specific ML classifier has been designed and trained with a particular dataset based on a specific evolutionary model's parameters, it is also important to ensure both the dataset and the classifier are meticulously described and made accessible to the public. Such precautions minimize the need to construct entirely new workflows for each study, involving tasks such as data simulation, model training, and the selection of evaluation metrics, enabling researchers to evaluate and enhance the method without needing to start from scratch (Greener et al., 2021; Heil et al., 2021).

Additionally, ML's ability to efficiently compare a wide range of models using large datasets in less computational time provides an important advantage over traditional model comparison approaches. Nonetheless, access to adequate computing resources remains a challenge for many researchers in species delimitation and various scientific disciplines (Veretnik et al., 2008; Truong et al., 2012; Helmy et al., 2016; Mangul et al., 2019b). Then, efforts to provide resources like graphics processing units, cloud storage, and computational clusters are all crucial steps toward making ML more accessible and inclusive for scientists across diverse domains of knowledge, including species delimitation.

4.2. Combining analytical frameworks to investigate complex delimitation models

All models, while inherently limited in representing the underlying nature of species diversification and, hence, of the current species limits among the tested entities, will be more or less useful depending on their effectiveness in extracting relevant evolutionary information from the available data. Accordingly, in some systems, certain methods should be prioritized based on the processes driving divergence, and using multiple methods with similar biases might not always enhance biological interpretability. For instance, Smith and Cartens (2020) argue that traditional methods like BPP can accurately infer the number of species but may overlook significant processes, such as secondary contact, something that ML workflows like delimitR could be more efficient in dealing with. In this context, the choice on which species delimitation method to use should be done before hypothesis-testing, considering the nature of the available data, and possibly prior relevant biological information regarding the evolution of the organisms.

One approach that would greatly benefit from the combination of coalescence-based methods and ML algorithms, and that could shape the future direction of genetic-based species delimitation, involves the empirical validation of speciation-based models, which can provide a nuanced understanding of the speciation process. Different speciation-based delimitation models, whether relying on ML, coalescence, or a combination of both, can be employed to capture different facets of the evolutionary divergence process, and to test different increasingly complex scenarios, with model validation serving as the means to articulate expert knowledge and the available statistical tools for hypothesis testing. In sum, while currently no universally superior species delimitation method exists, ML algorithms offer promising prospects for their integration into systematic protocols tailored for species delimitation.

5. Conclusions

Relatively few studies have explored ML techniques for species delimitation using molecular data so far. They are generally employed due to coalescence-based methods' specific assumptions and limitations. Besides, they are computationally efficient, can be easily integrated with traditional methods, and clearly provides a concrete and robust way to explore dataset structures when species-level divergences are hypothesized. The flexibility of ML-based methods allows them to accommodate complex evolutionary scenarios. Furthermore, likelihood-free approaches such as ML can provide more accurate estimates of species limits and population parameters, particularly in cases where traditional methods may struggle to converge or produce biased results.

Both ML approaches and coalescence-based methods provide a wide array of choices, necessitating careful selection considering multiple factors. Particularly, ML algorithms offer promising prospects but require thorough evaluation, comparison, and adaptation to specific biological problems. Besides, selecting an appropriate ML method for species delimitation should prioritize suitability for specific data and research questions over popularity. This assessment includes biological predictions, computational performance, and comparisons to existing methods, even considering that comparing existing methods can be challenging.

Some specific challenges can be highlighted regarding the utilization of ML frameworks to infer species limits. For example, overly specialized algorithms might perform well within observed ranges of evolutionary parameters but can struggle outside the training space. This gains importance as ML applications in Evolutionary Biology rely heavily on simulated data. Besides, model specialization for simulated data can hinder generalizability and transferability across different studies or data types. To address this issue, there are some potential solutions and emerging approaches. For example, GANs enable the creation of more realistic simulated data, and domain adaptation techniques to transfer knowledge across datasets with systematic differences. Another challenge relies on handling data derived from distinct genetic markers, hindering the comparison of different ML approaches.

Just as some coalescence-based methods, ML-based delimitation methods may not always discern species, but might identify incompletely separated species or ephemeral population variations. Therefore, ML should be progressively developed and used alongside traditional methods to enhance objectivity and robustness in species delimitation processes, combining the strengths of distinct analytical structures for hypothesis testing. This approach may allow for the accurate estimation of the speciation process, facilitating a clearer differentiation between population structure and evolutionary independence. Also, future applications of ML methods in species delimitation may focus on integrating various biological properties into species hypothesis testing. Finally, there is potential in utilizing ML methods in Integrative Taxonomy approaches, as combining morphological, ecological, and molecular data, is crucial for robust species delimitation and may benefit from the flexibility of these AI-based approaches. As these conditions are increasingly met, ML is poised to become an integral part of the toolkit used by scientists not only in the field of species delimitation, but for various Evolutionary Biology applications worldwide.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank André Luiz Gomes de Carvalho, Fernanda de Pinho Werneck and Renato José Pires Machado for their helpful suggestions in earlier versions of the text. We extend our gratitude to Daniel R. Schrider for critically reviewing the first draft of this manuscript. This work was supported by the Brazilian federal institution "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior" (CAPES) through a PhD scholarship to MMAS.

References

References identified with an asterisk (*) are cited only within the Appendices.

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. 2016. Deep learning for computational biology. Molecular Systems Biology 12.

Arenas, M. 2012. Simulation of molecular data under diverse evolutionary scenarios. PLoS computational biology 8.

Arnab, S.P., Amin, M. R., & DeGiorgio, M. 2023. Uncovering footprints of natural selection through spectral analysis of genomic summary statistics. Molecular Biology and Evolution, 40.

Arnold, M. L. 1992. Natural hybridization as an evolutionary process. Annual review of Ecology and Systematics, *23*, 237–261.

Baumdicker, F., et al. 2021. Efficient ancestry and mutation simulation with msprime 1.0. Genetics 220. doi:10.1093/ genetics/iyab229

Beaumont, M.A., Zhang, W., Balding, D.J. 2002. Approximate Bayesian computation in population genetics. Genetics 162, 2025–2035. doi:10.1093/genetics/162.4.2025

Bertorelle, G., Benazzo, A., Mona, S. 2010. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol Ecol. 19, 2609–2625. doi:10.1111/j.1365-294X.2010.04690.x

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, 2676–2688. https://doi.org/10.1111/1755-0998.13355

Booker, W.W., Ray, D.D., & Schrider, D.R. 2023. This population does not exist: learning the distribution of evolutionary histories with generative adversarial networks. Genetics, 224(2), iyad063.

Borowiec, M.L., Dikow, R.B., Frandsen, P.B., McKeeken, A., Valentini, G., & White, A.E. 2022. Deep learning as a tool for ecology and evolution. Methods in Ecology and Evolution 13, 1640–1660.

Bortolus, A. 2008. Error cascades in the biological sciences: the unwanted consequences of using bad taxonomy in ecology. AMBIO: A journal of the human environment 37, 114–118.

Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., & Krishnan, D. 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 3722–3731.

Breitman, M.F., Domingos, F.M., Bagley, J.C., Wiederhecker, H.C., Ferrari, T.B., Cavalcante, V.H., ... & Colli, G.R. 2018. A new species of *Enyalius* (Squamata, Leiosauridae) endemic to the Brazilian Cerrado. Herpetologica 74, 355–369.

Burbrink, F.T., & Ruane, S. 2021. Contemporary philosophy and methods for studying speciation and delimiting species. Ichthyology & Herpetology 109, 874–894.

Callier, V. 2022. Machine learning in evolutionary studies comes of age. Proceedings of the National Academy of Sciences 119.

Carstens, B. C., Pelletier, T. A., Reid, N. M., & Satler, J. D. 2013. How to fail at species delimitation. Molecular Ecology 22, 4369–4383.

Cerca, J., Maurstad, M.F., Rochette, N. C., Rivera-Colón, A.G., Rayamajhi, N., Catchen, J.M., & Struck, T H. 2021. Removing the bad apples: A simple bioinformatic method to improve loci-recovery in de novo RADseq data for non-model organisms. Methods in Ecology and Evolution 12, 805–817.

Chadha, A., Kumar, V., Kashyap, S., & Gupta, M. 2021. Deepfake: An overview. In Proceedings of Second International Conference on Computing, Communications, and Cyber-Security, pp. 557-566. Springer, Singapore.

Chicco, D. 2017. Ten quick tips for machine learning in computational biology. BioData Mining 10, 1–17. https://doi.org/10.1186/s13040-017-0155-3

Christin, S., Hervet, É., & Lecomte, N. 2019. Applications for deep learning in ecology. Methods in Ecology and Evolution 10, 1632–1644.

Cochran, K., Srivastava, D., Shrikumar, A., Balsubramani, A., Hardison, R.C., Kundaje, A., Mahony, S. 2022. Domain adaptive neural networks improve cross-species prediction of transcription factor binding. Genome Res. 32, 512–523.

Collin, F.D., Durif, G., Raynal, L., Lombaert, E., Gautier, M., Vitalis, R., Marin, J.M., & Estoup, A. 2021. Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC Random Forest. Molecular Ecology Resources 21, 2598–2613. https://doi.org/10.1111/1755-0998.13413.

Csilléry, K., Blum, M.G., Gaggiotti, O.E., & François, O. 2010. Approximate Bayesian computation (ABC) in practice. Trends in Ecology & Evolution 25, 410–418.

Dayrat, B. 2005. Towards integrative taxonomy. Biological journal of the Linnean society, 85, 407–417.

Degnan, J. H. & Rosenberg, N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24, 332–340.

Derkarabetian, S., Castillo, S., Koo, P.K., Ovchinnikov, S., & Hedin M. 2019. A demonstration of unsupervised machine learning in species delimitation. Molecular Phylogenetics and Evolution 139. https://doi.org/10.1016/j.ympev.2019.106562

Derkarabetian, S., Starrett, J., & Hedin, M. 2022. Using natural history to guide supervised machine learning for cryptic species delimitation with genetic data. Frontiers in Zoology 19, 1–15.

Dike, H.U., Zhou, Y., Deveerasetty, K.K., & Wu, Q. 2018. Unsupervised learning based on artificial neural network: A review. In 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS), pp. 322-327.

Domingos, F.M., Bosque, R.J., Cassimiro, J., Colli, G.R., Rodrigues, M.T., Santos, M.G., & Beheregaray, L. B. 2014. Out of the deep: cryptic speciation in a Neotropical gecko (Squamata, Phyllodactylidae) revealed by species delimitation methods. Molecular Phylogenetics and Evolution 80, 113–124.

*Duan, L., Han, L.N., Liu, B., Leostrin, A., Harris, A.J., Wang, L., Arslan, E., Ertuğrul, K., Knyazev, M., Hantemirova, E., Wen, J., & Chen, H.F. 2023. Species delimitation of the liquorice tribe (Leguminosae: Glycyrrhizeae) based on phylogenomic and machine learning analyses. Journal of Systematics and Evolution 61, 22–41. https://doi.org/10.1111/jse.12902.

Edwards, S. V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63, 1–19.

Edwards, S.V., Xi, Z., Janke, A., Faircloth, B.C., McCormack, J.E., Glenn, T.C., ... & Davis, C.C. 2016. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Molecular Phylogenetics and Evolution 94, 447–462.

Edwards, S. V., Potter, S., Schmitt, C. J., Bragg, J. G., & Moritz, C. 2016. Reticulation, divergence, and the phylogeography–phylogenetics continuum. Proceedings of the National Academy of Sciences, 113, 8025–8032.

Ely, C.V., de Loreto Bordignon, S.A., Trevisan, R., & Boldrini, I.I. 2017. Implications of poor taxonomy in conservation. Journal for Nature Conservation 36, 10–13.

Excoffier, L. et al. 2021. fastsimcoal2: demographic inference under complex evolutionary scenarios. Bioinformatics 37, 4882–4885. doi:10.1093/bioinformatics/btab468.

*Fan, X.K., Wu, J., Comes, H.P., Feng, Y., Wang, T., Yang, S.Z., Iwasaki, T., Zhu, H., Jiang, Y., Lee, J., & Li, P. 2023. Phylogenomic, morphological, and niche differentiation analyses unveil species delimitation and evolutionary history of endangered maples in Acer series Campestria (Sapindaceae). Journal of Systematics and Evolution 61, 284–298. https://doi.org/10.1111/jse.12919.

Farahani, A., Voghoei, S., Rasheed, K., & Arabnia, H.R. 2021. A brief review of domain adaptation. Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020, 877–894.

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

Flouri, T., Jiao, X., Rannala, B., Yang, Z. 2018. Species Tree Inference with BPP using Genomic Sequences and the Multispecies Coalescent. Molecular Biology and Evolution 35, 2585–2593. doi:10.1093/molbev/msy147.

2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. Molecular Biology and Evolution 37, 1211–1223.

Fonseca, R.R. et al. 2016. Next-generation biology: sequencing and data analysis approaches for non-model organisms. Marine genomics 30, 3–13.

Fonseca, E. M., Colli, G. R., Werneck, F. P., & Carstens, B. C. 2021. Phylogeographic model selection using convolutional neural networks. Molecular Ecology Resources 21, 2661–2675. https://doi.org/10.1111/1755-0998.13427.

Fonseca, E. M., & Carstens, B. C. (2024). Artificial intelligence enables unified analysis of historical and landscape influences on genetic diversity. *Molecular Phylogenetics and Evolution*, 108116.

Fountain-Jones, N.M., Smith, M.L., & Austerlitz, F. 2021. Machine learning in molecular ecology. Molecular Ecology Resources 21, 2589–2597. https://doi.org/10.1111/1755-0998.13532.

Fujita, M.K., Leaché, A.D., Burbrink, F.T., McGuire, J.A., & Moritz, C. 2012. Coalescent-based species delimitation in an integrative taxonomy. Trends in Ecology & Evolution 27, 480–488.

Ganin, Y., & Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In International conference on machine learning, 1180–1189.

Gehara, M., Mazzochinni, G.G., & Burbrink, F. 2020. PipeMaster: inferring population divergence and demographic history with approximate Bayesian computation and supervised machine-learning in R. BioRxiv, 2020-12. https://doi.org/10.1101/2020.12.04.410670

Ghifary, M, Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W. 2016. Deep Reconstruction Classification Networks for Unsupervised Domain Adaptation. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision ECCV 2016. Lecture Notes in Computer Science. Cham: Springer International Publishing. p. 597

Ghirotto, S., Vizzari, M.T., Tassi, F., Barbujani, G. & Benazzo, A. 2021. Distinguishing among complex evolutionary models using unphased whole-genome data through random forest approximate Bayesian computation. Molecular Ecology Resources 21, 2614–2628. https://doi.org/10.1111/1755-0998.13263.

Gompert, Z., Mandeville, E. G., & Buerkle, C. A. 2017. Analysis of population genomic data from hybrid zones. Annual Review of Ecology, Evol., and Syst., 48, 207–229.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio Y. 2014. Generative adversarial nets. Advances in Neural Information Processing Systems, 2672–2680.

Greener, J.G., Kandathil, S.M., Moffat, L., & Jones, D.T. 2021. A guide to machine learning for biologists. Molecular Cell Biology 23, 40–55. https://doi.org/10.1038/s41580-021-00407-0.

Hastie, T., Tibshirani, R., & Friedman, J. 2009. Unsupervised learning. In The elements of statistical learning (pp. 485-585). Springer, New York, NY.

Hausdorf, B. 2011. Progress toward a general species concept. Evolution 65, 923–931.

Heil, B.J., Hoffman, M. M., Markowetz, F., Lee, S.I., Greene, C.S. & Hicks, S.C. 2021. Reproducibility standards for machine learning in the life sciences. Nature Methods 18, 1132–1135.

Helmy, M., Awad, M., & Mosa, K.A. 2016. Limited resources of genome sequencing in developing countries: challenges and solutions. Applied & translational genomics 9, 15–19.

*Hodel, R.G., Winslow, S.K., Liu, B.B., Johnson, G., Trizna, M., White, A.E., ... & Wen, J. 2023. A phylogenomic approach, combined with morphological characters gleaned via machine learning, uncovers the hybrid origin and biogeographic diversification of the plum genus. bioRxiv, 2023-09. https://doi.org/10.1101/2023.09.13.557598

Hüllermeier, E., Fober, T. & Mernberger, M. 2013. Inductive bias. Encyclopedia of systems biology, 1018–1019.

Huang, J. P. 2020. Is population subdivision different from speciation? From phylogeography to species delimitation. Ecology and Evolution 10, 6890–6896.

Huelsenbeck, J.P., Andolfatto, P., Huelsenbeck, E.T. 2011. Structurama: Bayesian inference of population structure. Evolutionary Bioinformatics 7, 55–59.

Jackson, N.D., Carstens, B.C., Morales, A.E. & O'Meara B.C. 2017. Species delimitation with gene flow. Systematic Biology 66, 799–812.

Jackson, N.D., Morales, A.E., Carstens, B.C. & O'Meara B.C. 2017. PHRAPL: phylogeographic inference using approximate likelihoods. Systematic Biology 66, 1045–1053.

Jacobs, S. J., Kristofferson, C., Uribe-Convers, S., Latvis, M., & Tank, D. C. (2018). Incongruence in molecular species delimitation schemes: What to do when adding more data is difficult. Molecular Ecology 27, 2397–2413.

*Jamdade, R., Al-Shaer, K., Al-Sallani, M., Al-Harthi, E., Mahmoud, T., Gairola, S., & Shabana, H.A. 2022. Multilocus marker-based delimitation of *Salicornia persica* and its population discrimination assisted by supervised machine learning approach. PLoS ONE 17. https://doi.org/10.1371/journal.pone.0270463.

Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., ... & Zhang, G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346, 1320–1331.

Jörger, K.M., & Schrödl, M. 2013. How to describe a cryptic species? Practical challenges of molecular taxonomy. Frontiers in Zoology 10, 1–27.

Jorna, J. et al. 2021. Species boundaries in the messy middle—A genome-scale validation of species delimitation in a recently diverged lineage of coastal fog desert lichen fungi. Ecology and Evolution 11, 18615-18632.

Karbstein, K., Kösters, L., Hodač, L., Hofmann, M., Hörandl, E., Tomasello, S., ... & Wäldchen, J. (2023). Species delimitation 4.0: integrative taxonomy meets artificial intelligence. *Trends in Ecology & Evolution*.

*Khalighifar, A., Brown, R.M., Goyes Vallejos, J., & Peterson, A.T. 2021. Deep learning improves acoustic biodiversity monitoring and new candidate forest frog species identification (genus *Platymantis*) in the Philippines. Biodiversity and Conservation, 30, 643-657.

Knowles, L. L., & Carstens, B. C. 2007. Delimiting species without monophyletic gene trees. Syst. Biol., 56, 887–895.

Korfmann, K., Gaggiotti, O.E. & Fumagalli, M. 2023. Deep learning in population genetics. Genome Biology and Evolution. https://doi.org/10.1093/gbe/evad008.

Kuzenkov, O., Morozov, A., & Kuzenkova, G. 2020. Exploring evolutionary fitness in biological systems using machine learning methods. Entropy 23, 1–35.

Leaché, A.D., Harris, R.B., Rannala, B. & Yang, Z. 2014. The influence of gene flow on species tree estimation: a simulation study. Systematic Biology 63, 17–30.

Leaché, A.D., Zhu, T., Rannala, B., & Yang, Z. 2019. The spectre of too many species. Systematic Biology 68, 168–181.

Libbrecht, M.W. & Noble, W.S. 2015. Machine learning applications in genetics and genomics. Nature Reviews Genetics 16, 32–332.

*Lima, A.P. et al. 2020a. Not as widespread as thought: Integrative taxonomy reveals cryptic diversity in the Amazonian nurse frog *Allobates tinae* Melo-Sampaio, Oliveira and Prates, 2018 and description of a new species. Journal of Zoological Systematics and Evolutionary Research, 58(4), 1173–1194.

*Lima, L.R. et al. 2020b. Below the waterline: cryptic diversity of aquatic pipid frogs (*Pipa carvalhoi*) unveiled through an integrative taxonomy approach. Systematics and Biodiversity, 18(8), 771–783.

Liu, B. 2019. BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. Briefings in bioinformatics 20, 1280–1294.

Liu, M.Y. & Tuzel, O. 2016. Coupled Generative Adversarial Networks. In: Advances in Neural Information Processing Systems 29. Curran Associates, Inc. https://papers.nips.cc/paper/2016/hash/502e4a16930e414107ee22b6198c578f-Abstract.html.

Lukhtanov, V.A. 2019. Species Delimitation and Analysis of Cryptic Species Diversity in the XXI Century. Entmol. Rev. 99, 463–472. https://doi.org/10.1134/S0013873819040055.

Lürig, M.D., Donoughe, S., Svensson, E.I., Porto, A. & Tsuboi, M. 2021. Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology. Frontiers in Ecology and Evolution 9.

*Magalhães, F.D.M., Lyra, M.L., De Carvalho, T.R., Baldo, D., Brusquetti, F., Burella, P., ... & Garda, A.A. 2020. Taxonomic review of South American Butter Frogs: Phylogeny, geographic patterns, and species delimitation in the *Leptodactylus latrans* species group (Anura: Leptodactylidae). Herpetological Monographs, 34(1), 131–177.

Mallet, J., Besansky, N., & Hahn, M. W. 2016. How reticulated are species? BioEssays, 38, 140–149.

Mangul, S. et al. 2019a. Systematic benchmarking of omics computational tools. Nature communications 10.

2019b. How bioinformatics and open data can boost basic science in countries and universities with limited resources. Nature biotechnology 37, 324–326.

Martin, B.T., Chafin, T.K., Douglas, M.R., Placyk, Jr J.S., Birkhead, R.D., Phillips, C.A., & Douglas, M.E. 2021. The choices we make and the impacts they have: Machine learning and species delimitation in North American box turtles (*Terrapene* spp.). Molecular Ecology Resources 21, 2801–2817.

Mayr, E. M. 1969. The biological meaning of species. Biological Journal of the Linnean society, 1, 311–320.

1996. What is a species, and what is not? Philosophy of science, 63, 262–277.

2000. The biological species concept. Species concepts and phylogenetic theory: a debate, 17–29.

McClure, E.C., Sievers, M., Brown, C.J. Buelow, C.A., Ditria, E.M., Hayes, M.A., ... & Connolly R.M. 2020. Artificial intelligence meets citizen science to supercharge ecological monitoring. Patterns 1.

Messer, P. W. 2013. SLiM: simulating evolution with selection and link-age. Genetics 194, 1037–1039. doi:10.1534/genetics.113. 152181.

Mitchell, T.M. 1997. Machine Learning. McGraw-Hill, New York.

Mo, Z., & Siepel, A. 2023. Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data. PLOS Genetics, 19.

Mo, Y. K., Hahn, M. W., & Smith, M. L. 2024. Applications of machine learning in phylogenetics. *Molecular Phylogenetics and Evolution*, *196*, 108066.

Morimoto, J., Ponchon, A., Sofronov, G., & Travis, J. 2021. Editorial: Applications of Machine Learning to Evolutionary Ecology Data. Frontiers in Ecology and Evolution.

Nesterenko, L., Boussau, B., & Jacob, L. 2022. Phyloformer: towards fast and accurate phylogeny estimation with self-attention networks. bioRxiv, 2022-06. https://doi.org/10.1101/2022.06.24.496975

*Newton, L.G., Starrett, J., Hendrixson, B.E., Derkarabetian, S., & Bond, J.E. (2020). Integrative species delimitation reveals cryptic diversity in the southern Appalachian *Antrodiaetus unicolor* (Araneae: Antrodiaetidae) species complex. Molecular Ecology 29, 2269–2287.

O'Meara B. C. 2010. New heuristic methods for joint species delimitation and species tree inference. Systematic Biology 59, 59–73.

2012. Evolutionary inferences from phylogenies: a review of methods. Annual Review of Ecology, Evolution, and Systematics 43, 267–285.

Padial, J. M., Miralles, A., De la Riva, I., & Vences, M. 2010. The integrative future of taxonomy. Frontiers in zoology, 7, 1–14.

Pan, S. J. & Yang, Q. 2010. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22, 1345–1359.

Pante, E., Puillandre, N., Viricel, A., Arnaud-Haond, S., Aurelle, D., Castelin M., ... & Samadi, S. 2015. Species are hypotheses: avoid connectivity assessments based on pillars of sand. Molecular Ecology 24, 525–544.

Pei, J., Chu, C., Li, X., Lu, B., & Wu, Y. 2018. CLADES: A classification-based machine learning method for species delimitation from population genetic data. Molecular Ecology Resources 18, 1144–1156. https://doi.org/10.1111/1755-0998.12887.

Perez, M.F., Bonatelli, I.A.S., Romeiro-Brito, M., Franco, F.F., Taylor, N.P., Zappi, D.C. et al. 2021. Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system. Molecular Ecology Resources.

Pichler, M., Boreux, V., Klein, A. M., Schleuning, M. & Hartig F. 2020. Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. Methods in Ecology and Evolution 11, 281–293.

Pigliucci, M. 2003. Species as family resemblance concepts: the (dis-)solution of the species problem? BioEssays, 25, 596–602.

Pons, J., Barraclough, T.G., Gomez-Zurita, J. et al. 2006. Sequence-based species delimitation for the DNA taxonomy of unde-scribed insects. Systematic Biology 55, 595–609.

Price, T.D., Qvarnström, A., & Irwin, D.E. 2003. The role of phenotypic plasticity in driving genetic evolution. Proceedings of the Royal Society of London. Series B: Biological Sciences 270, 1433–1440.

*Pritchard, J.K., Stephens, M., Donnelly, P. 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Pudlo, P., Marin, J.M., Estoup, A., Cornuet, J.M., Gautier, M., & Robert, C.P. 2016. Reliable ABC model choice via random forests. Bioinformatics 32, 859–866. https://doi.org/10.1093/bioinformatics/btv684.

Pyron, R.A. 2023. Unsupervised Machine Learning for Species Delimitation, Integrative Taxonomy, and Biodiversity Conservation. Molecular Phylogenetics and Evolution, 189.

Pyron, R.A., O'Connell, K.A., Duncan, S.C., Burbrink, F.T., & Beamer, D.A. 2023. Speciation hypotheses from phylogeographic delimitation yield an integrative taxonomy for Seal Salamanders (*Desmognathus monticola*). Systematic Biology, 72, 179–197.

Pyron, R. A., Kakkera, A., Beamer, D. A., & O'Connell, K. A. 2024. Discerning structure versus speciation in phylogeographic analysis of Seepage Salamanders (*Desmognathus aeneus*) using demography, environment, geography, and phenotype. Molecular Ecology, 33, e17219.

de Queiroz, K. 1998. The general lineage concept of species, species criteria, and the process of speciation. Endless forms: species and speciation.

1999. The General Lineage Concept of Species and the Defining Properties of the Species Category. In book: Species: New Interdisciplinary Essays, Chapter: 3, Publisher: MIT Press, Editors: Robert A. Wilson.

2005a. Ernst Mayr and the modern concept of species. Proceedings of the National Academy of Sciences, 102, 6600–6607.

2005b. Different species problems and their resolution. BioEssays 27, 1263–1269.

2007. Species concepts and species delimitation. Syst. Biol. 56, 879–886.

2011. Branches in the lines of descent: Charles Darwin and the evolution of the species concept. Biol. J. Linn. Soc. 103, 19–35.

2020. An updated concept of subspecies resolves a dispute about the taxonomy of incompletely separated lineages. Herpetological Review.

Qu, K., Guo, F., Liu, X., Lin, Y., & Zou, Q. 2019. Application of machine learning in microbiology. Frontiers in Microbiology 10.

Rannala, B. 2015. The art and science of species delimitation. Current Zoology 61, 846–853.

Rannala, B., & Yang, Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164, 1645–1656.

2010. Bayesian species delimitation using multilocus sequence data. Proceedings of the National Academy of Sciences 107, 9264–9269.

2020. Species Delimitation. In: Phylogenetics in the genomic era.

Rannala, B., Edwards, S.V., Leaché, A., & Yang, Z. 2020. The Multi-species Coalescent Model and Species Tree Inference. Scornavacca, Celine; Delsuc, Frédéric; Galtier, Nicolas. Phylogenetics in the Genomic Era, No commercial publisher | Authors open access book.

Raynal, L., Marin, J.M., Pudlo, P., Ribatet, M., Robert, C.P., & Estoup, A. 2019. ABC random forests for Bayesian parameter inference. Bioinformatics 35, 1720–1728.

Rozantsev, A., Salzmann, M. & Fua, P. 2018. Beyond sharing weights for deep domain adaptation. IEEE transactions on pattern analysis and machine intelligence 41, 801–814.

*Saryan, P., Gupta, S. & Gowda, V. 2020. Species complex delimitations in the genus *Hedychium*: A machine learning approach for cluster discovery. Applications in Plant Sciences 8. https://doi.org/10.1002/aps3.11377.

Sanchez, T., Cury, J., Charpiat, G. & Jay, F. 2020. Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. Molecular Ecology Resources 21, 2645–2660.

Scalon, M.C., Domingos, F. M.C.B., Cruz, W.J.A., Marimon-Júnior, B. H., Marimon, B.S., & Oliveras, I. 2020. Diversity of functional trade-offs enhances survival after fire in Neotropical savanna species. Journal of Vegetation Science, 31, 139-150.

Schrider, D.R. & Kern, A.D. 2016. Discoal: flexible coalescent simulations with selection. Bioinformatics 32, 3839–3841. doi:10.1093/ bioinformatics/btw556.

2018. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends in Genetics 34, 301–312. https://doi.org/10.1016/j.tig.2017.12.005

Searls, D.B. 2010. The Roots of Bioinformatics. PLoS Comput Biol 6. https://doi.org/10.1371/journal.pcbi.1000809.

Sheehan, S., & Song, Y.S. 2016. Deep learning for population genetic inference. PLoS computational biology 12.

Shurtliff, Q. R. 2013. Mammalian hybrid zones: a review. Mammal Review, 43, 1–21.

Sidey-Gibbons, J.A., & Sidey-Gibbons, C.J. 2019. Machine learning in medicine: a practical introduction. BMC medical research methodology 19, 1–18.

Silva, D.C., Oliveira, H.F.M., & Domingos, F.M.C.B. 2024. Cerrado bat community assembly is determined by both present-day and historical factors. Journal of Biogeography.

Simonsen, K.L., Churchill, G.A., Aquadro, C.F. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 1411, 413–429.

Sites, Jr J.W. & Marshall, J.C. 2004. Operational criteria for delimiting species. Annual Review of Ecology, Evolution, and Systematics, 199-227.

Slatko, B.E., Gardner, A.F. & Ausubel, F.M. 2018. Overview of next-generation sequencing technologies. Current protocols in molecular biology 122.

Smith, M.L., Ruffley, M., Espíndola, A., Tank, D.C., Sullivan, J. & Carstens, B.C. 2017. Demographic Model Selection using Random Forests and the Site Frequency Spectrum. Molecular Ecology.

Smith, M.L. & Carstens B.C. 2020. Process-based species delimitation leads to identification of more biologically relevant species. Evolution 74, 216–229. https://doi.org/10.1111/evo.13878.

Smith, M.L., & Hahn, M.W. 2023. Phylogenetic inference using generative adversarial networks. Bioinformatics, 39.

Solis-Lemus, C., Yang, S., & Zepeda-Nunez, L. 2022. Accurate phylogenetic inference with a symmetry-preserving neural network model. arXiv preprint arXiv:2201.04663.

Sukumaran, J. & Knowles, L.L. 2017. Multispecies coalescent delimits structure, not species. Proceedings of the National Academy of Sciences 114, 1607–1612.

Sukumaran, J., Holder, M.T., & Knowles, L.L. 2021. Incorporating the speciation process into species delimitation. PLoS Computational Biology 17.

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

Tagu, D., Colbourne, J.K. & Nègre, N. 2014. Genomic data integration for ecological and evolutionary traits in non-model organisms. BMC genomics 15, 1–16.

Tautz, D., Arctander, P., Minelli, A., Thomas, R.H., Vogler, A.P. 2003. A plea for DNA taxonomy. Trends Ecol. Evol. 18, 70–74.

Truong, H.L., Pham T.V., Thoai N., & Dustdar S. 2012. Cloud computing for education and research in developing countries. In Cloud computing for teaching and learning: strategies for design and implementation, pp. 64–80. IGI Global.

Valletta, J.J., Torney, C., Kings, M., Thornton, A. & Madden J. 2017. Applications of machine learning in animal behaviour studies. Animal Behaviour 124, 203–220.

Veretnik, S., Fink, J.L. & Bourne, P.E. 2008. Computational biology resources lack persistence and usability. PLoS computational biology 4.

Vink, C.J., Paquin, P., & Cruickshank, R.H. 2012. Taxonomy and irreproducible biological science. BioScience 62, 451–452.

Vogler, A.P., Monaghan, M.T. 2007. Recent advances in DNA taxonomy. J. Zool. Syst. Evol. Res. 45, 1–10.

Wake, D.B., Wake, M.H. & Specht C.D. 2011. Homoplasy: from detecting pattern to determining process and mechanism of evolution. Science 331, 1032–1035.

Wäldchen, J. & Mäder, P. 2018. Machine learning for image-based species identification. Methods in Ecology and Evolution 9, 2216–2225.

Wang, G. 2019. Machine learning for inferring animal behavior from location and movement data. Ecological informatics 49, 69–76.

Wang, Z., Wang, J., Kourakos, M., Hoang, N., Lee, H.H., Mathieson, I., & Mathieson, S. 2021. Automatic inference of demographic parameters using generative adversarial networks. Molecular ecology resources 21, 2689–2705.

Wiens, J. J., & Penkrot, T. A. 2002. Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (*Sceloporus*). Syst. Biol., 51, 69–91.

Wiens, J. J. 2007. Species delimitation: new approaches for discovering diversity. Syst. Biol. 56, 875–8.

Wilkins, J. S., Zachos, F. E., & Pavlinov, I. Y. (Eds.). 2022. Species Problems and Beyond: Contemporary Issues in Philosophy and Practice. CRC Press.

Yang, B., Zhang, Z., Yang, C.Q., Wang, Y., Orr, M.C., Wang, H., & Zhang, A.B. 2022. Identification of species by combining molecular and morphological data using convolutional neural networks. Systematic Biology, 71, 690–705.

Yelmen, B. & Jay, F. 2023. An Overview of Deep Generative Models in Functional and Evolutionary Genomics. Annual Reviews of Biomedical Data Science. https://doi.org/10.1146/annurev-biodatasci-020722.

Zachos, F. E. 2016. Species concepts in biology (Vol. 801). Cham: Springer.

2018. (New) Species concepts, species delimitation and the inherent limitations of taxonomy. Journal of genetics, 97, 811–815.

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, 74–89. https://doi.org/10.1089/cmb.2021.0383.