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, for instance regarding complex evolutionary scenarios and large datasets. 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. Future ML enterprises to delimit species should consider the

constraints related to the use of simulated data, as in other model-based methods relying on simulations. Conversely, the flexibility of ML algorithms offers a significant advantage by enabling the analysis of diverse data types (e.g., genetic and phenotypic) and handling large datasets effectively. 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.

An important concept in this context is the **General Lineage Concept** (**GLC**, terms in bold are defined in the Glossary, available in Appendix A), which unifies diverse contemporary views on the nature of species, 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 and Penkrot, 2002; Dayrat, 2005; Padial et al., 2010; Fujita et al., 2012; Karbstein et al., 2024).

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

101 has emerged as a crucial tool for inferring species limits, offering a more objective
102 approach for species delimitation (Fujita et al., 2012), while complementing
103 traditional morphological methods (Jörger and Schrödl, 2013).

104 Modern species delimitation methods (SDMs) aiming at identifying
105 evolutionary units (Tautz et al., 2003; Vogler and Monaghan, 2007) have grown
106 due to advancements in statistical frameworks for phylogenetic inference
107 (Edwards, 2009; O'Meara, 2012), along with Molecular Biology tools (e.g., next-
108 generation sequencing (NGS); Slatko et al., 2018) and Bioinformatics (Searls,
109 2010). They mostly operate with molecular data under the principles of
110 Coalescent Theory, notably, the multispecies coalescent (MSC; Rannala and
111 Yang, 2003; Degnan and Rosenberg, 2009; Rannala et al., 2020). The MSC
112 analytical framework has many evolutionary assumptions, such as the absence
113 of recombination and hybridization, independence of gene trees and their
114 coalescent processes, random mating within species, among others (a review on
115 the subject can be found in Mirarab et al., 2021). However, these conditions are
116 typically only met in tree-like speciation scenarios involving diploid, sexually
117 reproducing organisms. In any case, MSC methods are capable of managing
118 common problems in phylogenetic inference, such as conflicts among different
119 gene trees due **incomplete lineage sorting** (**ILS**; Knowles and Carstens, 2007;
120 Carstens et al., 2013; Jacobs et al., 2018).

121 Therefore, while they are valuable for inferring evolutionary relationships,
122 coalescence-based SDMs may fail to distinguish population structure from
123 species-level divergence (Sukumaran and Knowles, 2017), and may also be
124 affected by the above-mentioned assumptions of the MSC model (Rannala and
125 Yang, 2003; Degnan and Rosenberg, 2009; Edwards, 2009; Fujita et al., 2012).

Some methods have their functionality and performance compromised in scenarios when there is introgression between putative 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). Besides, 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). Naturally, different coalescence-based SDMs have varying capabilities to address particular evolutionary scenarios, and while such methods may be biased under certain evolutionary and analytical conditions, they are certainly an important part of the evolutionary biologist toolkit. For a more detailed discussion on SDMs based on the MSC for genomic data, see Rannala and Yang (2020).

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; Fountain-Jones et al., 2021; Greener et al., 2021; Morimoto et al., 2021; Borowiec et al., 2022). 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 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 and 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 Fountain-Jones et al., 2021; Morimoto et al., 2021; Borowiec et al., 2022).

Therefore, ML's potential in evolutionary biology, and particularly in species delimitation, is evident (Karbstein et al., 2024). Specific examples can also 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 and Hahn, 2023; Zaharias et al., 2022; Mo et al., 2024), and species delimitation (Pei et al., 2018; Derkarabetian et al., 2019; Smith and 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**. For example, by avoiding the need for explicit likelihood calculations, these methods might be computationally advantageous, particularly when combined with approaches optimized for high-throughput data processing, making them particularly suitable for analyzing large datasets with many taxa.

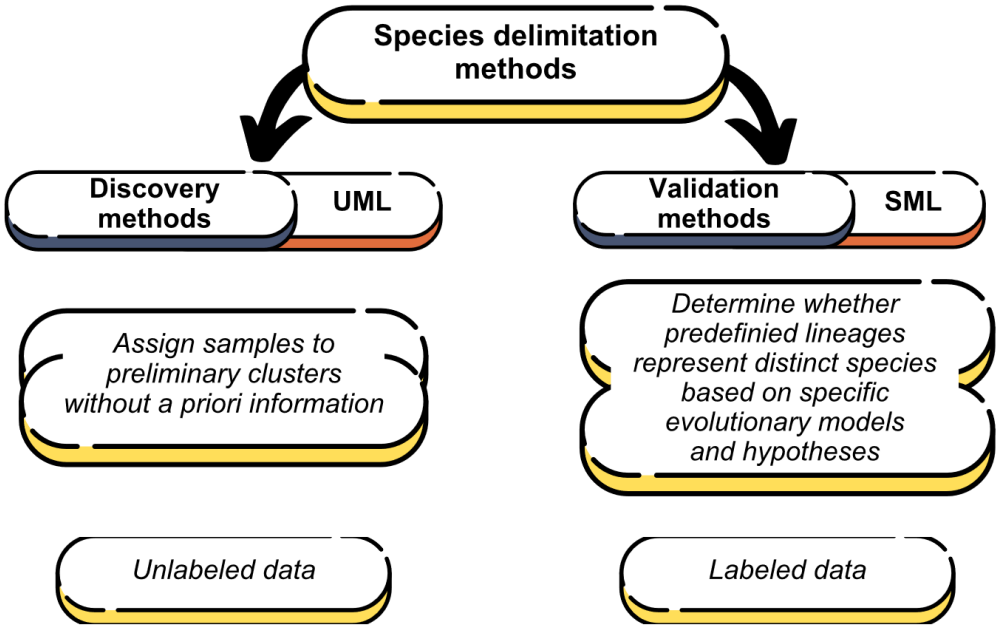

Fig. 1. Comparative diagram categorizing species delimitation methods and machine learning algorithms, along with some of their key characteristics. Species delimitation methods can be broadly categorized as discovery and validation methods, akin to unsupervised and supervised machine learning algorithms,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

222 **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	SML & UML	t-Distributed Stochastic Neighbor Embedding, Random Forest, Variational autoencoders	NA	SNP data matrix	One-hot-encoding of the SNP data matrix (VAE), *axis* from a discriminant analysis of principal components (t-SNE), scaled data from DAPC + cMDS and isoMDS ouput (Random forest)
Process-based species delimitation leads to identification of more biologically relevant species (Smith and 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

223 Online repositories where it is possible to find more information about the currently existing platforms. [1] https://github.com/pjweggy/CLADES;
224 [2]https://www.sciencedirect.com/science/article/abs/pii/S1055790319301721; [3]https://github.com/meganlsmith/delimitR; [4]https://github.com/manolofperez/CNN_spDelimitation_Piloso;
225 [5]https://github.com/kyleaoconnell22/Pyron_et_al_UML_sp_delim/tree/main

2.1 Discovery and unsupervised methods

Unsupervised machine learning (UML) relies 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, a combination of both, 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). UML algorithms are often regarded as methods lacking strong predefined assumptions about the underlying structure of the dataset (such as population parameters, species numbers, or sample categorization, in the case of species delimitation). Nevertheless, it is possible to incorporate heuristic or pragmatic assumptions in an UML framework to facilitate their operation. Either way, UML will be particularly useful in cases where prior hypotheses are limited or unavailable, provided that the assumptions of the chosen method are evaluated.

UML 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. Meanwhile, UML dimensionality reduction techniques compress data to identify a smaller distinct set of variables that retain essential features of the original data, while minimizing information loss. We highlight this as, when it comes to species delimitation, UML approaches often operate through clustering and/or dimensionality reduction algorithms (Fig. 2), extracting and condensing the necessary information to identify limits between biological groups (Derkarabetian et al., 2019; Pyron, 2023; Pyron et al., 2023), while also enabling the simultaneous use of different types of data (e.g., genetical, phenotypical and ecological).

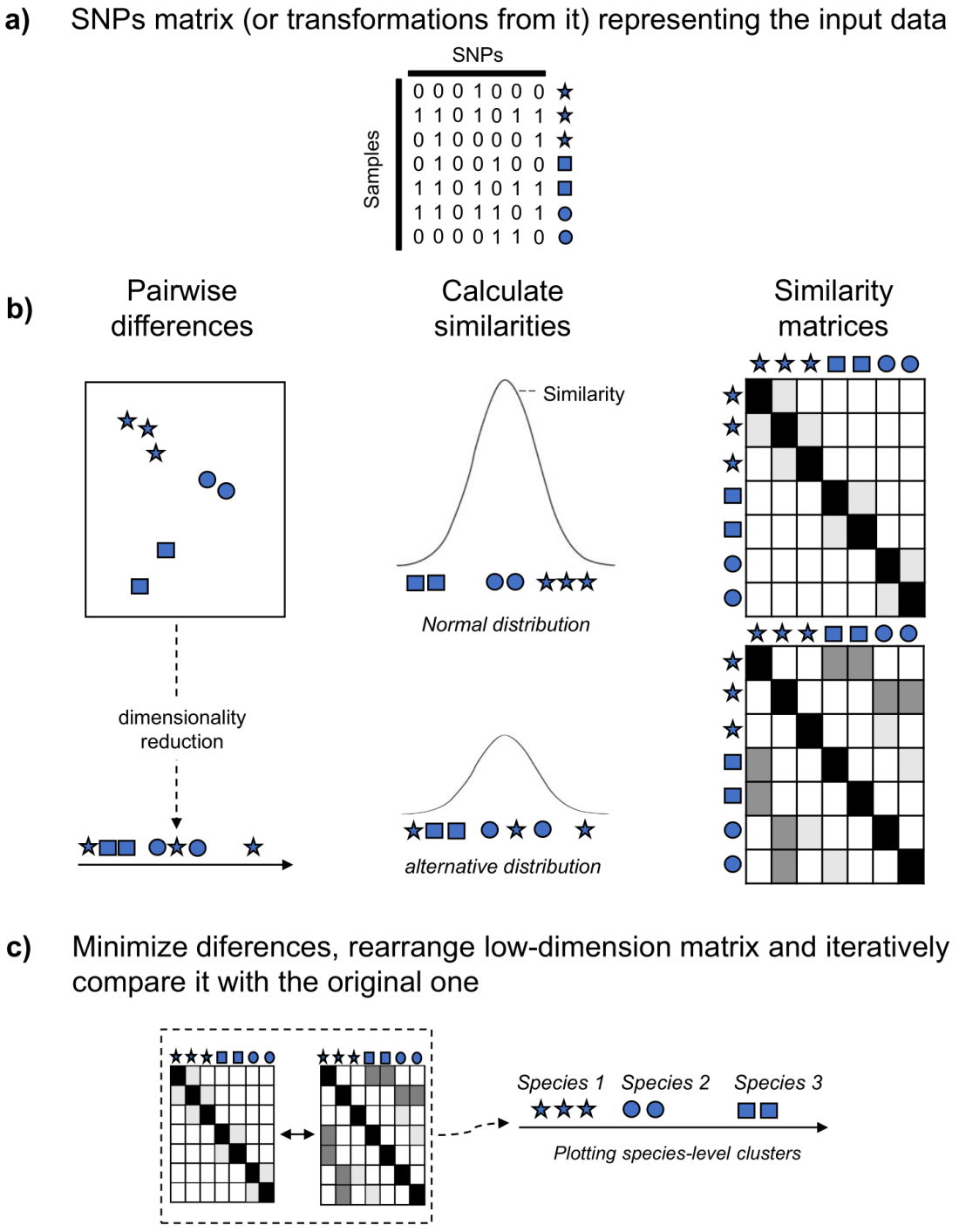

Fig. 2. Diagram outlining a potential UML workflow for species delimitation, utilizing the t-SNE algorithm as an example (inspired by Derkarabetian et al., 2019). a) Data representation is the initial step, and it varies depending on the chosen ML tool, which may work with sequence data, SNP matrices, or population genetics metrics extracted from them; in this case, samples from different populations are represented by distinct symbols. 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) evaluated the performance of different ML methods for species delimitation, including both SML and UML algorithms. Specifically, they evaluated the capacity of three approaches: **Random Forest** (RF, including a supervised and a non-supervised alternative), and two unsupervised models, **variational autoencoders (VAE)** and **t-Distributed Stochastic Neighbor Embedding (t-SNE).** In the t-SNE approach, data derived from a **principal component analysis (PCA)** were used as input variables, followed by clustering techniques using the output from the UML algorithms. In the VAE approach, **single-nucleotide polymorphism (SNP)** matrices were converted via **one-hot encoding,** where nucleotides were transformed into binary variables. In this case, the encoder takes the transformed SNP data and infers the distribution of latent variables, given as a normal distribution with a mean (μ) and standard deviation (σ). The decoder maps the latent distribution to a reconstruction of the one-hot encoded SNP data, offering a two-dimensional depiction. Finally, the RF approaches were performed with scaled data derived from a **Discriminant Analysis of Principal Components (DAPC)**, and the resulting proximity matrix was then used for **classical multidimensional scaling (cMDS)** and **isotonic multidimensional scaling (isoMDS)**. In sum, all approaches yielded, through different clustering strategies (depending on the algorithm being investigated), more readily interpretable outcomes compared to other traditional delimitation methods (or population structure detection methods) assessed by the authors, revealing distinct species groupings (Derkarabetian et

al., 2019). Notably, the identified groups also corresponded to those of an integrative taxonomy approach, suggesting that the limits identified by UML algorithms probably 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 based on **self-organizing maps (SOMs)**. This approach produces discrete outcomes rather than continuous ones, grouping genotypes based on similarity. 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 candidate species, Pyron et al. (2023) recommend performing cluster analyses, such as k-means. The determination of the optimal number of **classes,** or species, is achieved by selecting the value that maximizes the sequential reduction in the weighted sum of squares from k to $k + 1$. 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 genetic, morphological and ecological variables (Pyron, 2023).

2.2. Validation and supervised methods

While UML approaches are powerful and widely applicable, SML offers distinct analytical advantages in certain scenarios, contributing to its widespread

316 use in population genetics and evolutionary biology (Schrider and Kern, 2018). A
317 primary requirement for SML is the availability of **labeled training data**, which is
318 used to teach the algorithm to recognize patterns and make predictions. In the
319 context of population genetic analyses, such labeled datasets are often
320 unavailable or insufficient in size. To overcome this limitation, simulated genetic
321 data based on known evolutionary models are usually generated to represent
322 evolutionary scenarios. This simulated data is then encoded, along with observed
323 genetic data, into feature vectors used to train the algorithm, which is used to
324 recognize specific patterns in new observed data points (Fig. 3).
325

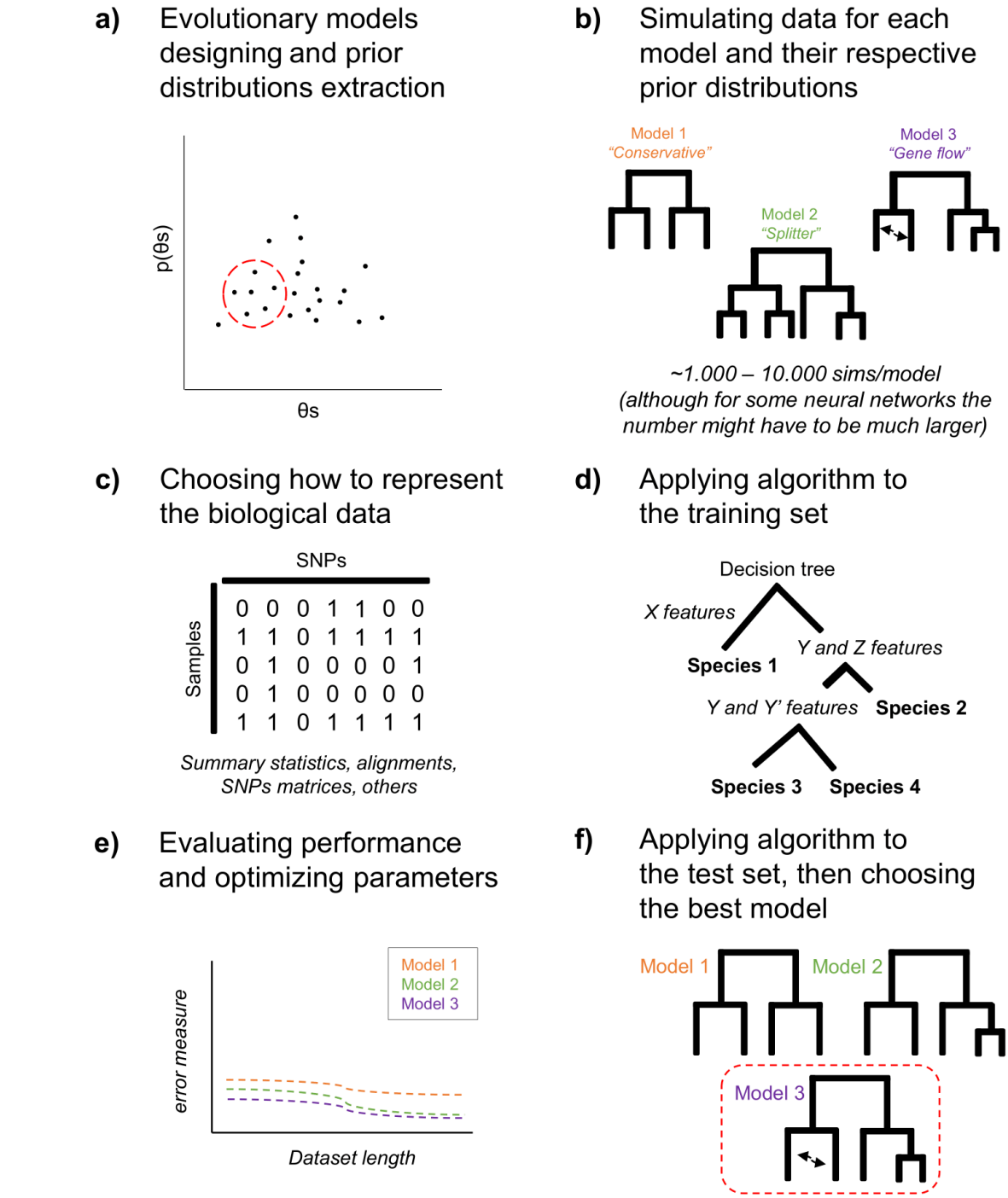

Fig. 3. Diagram illustrating a potential SML workflow for species delimitation, here using a decision-tree based algorithm as an example (inspired by the work of Smith and Carstens, 2020). a) The initial step involves, from a wider set of priors, extracting relevant subsets of priors for the evolutionary models of interest (clusters of dots circled in red). b) Simulated data is generated for each model, typically ranging from 1,000 to 10,000 simulations per model, using relevant simulation software; the specified models may involve variations in topology, such as scenarios with differing numbers of potential species ('*conservative*' versus *'splitter'*), and can also account for the possibility of gene flow. 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) After the algorithm is properly trained and evaluated, it can be used to predict species limits for new datasets (whether they are newly simulated or empirical data), using the model identified as best representing the species limits in the biological system (indicated here by the dashed red line).

The reliability of SML methods rely on the resemblance between the training data (typically simulated) and the actual biological data. Thus, the process of applying ML algorithms in species delimitation is influenced by the assumptions of the underlying evolutionary processes, such as population size, selection strength, and gene flow. 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) is an SML-based approach for species delimitation that employs **classification models** trained and tested on *multilocus* sequence data. Within this framework, **support vector machines (SVM)** are used to classify population pairs as either belonging to the same species or distinct species. Regarding model training, datasets at the population level are simulated, with and without gene flow. Then, each training sample is represented as a list of **summary statistics**, and a SVM **regression** is estimated, through iterative training, to minimize the misclassification cost. Subsequently, the SVM classifier computes the probability of the training samples belonging to each potential grouping.

Notably, the training dataset in this study 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 A and B was allowed at a rate of M = N*m* migrants per generation, with *m* representing the migration rate per generation. Additionally, symmetrical migration between A and B was accounted for prior to their divergence into two distinct populations each (A1 and A2, and B1 and B2). Multilocus sequence data of length L were simulated under diverse parameter combinations using the MCcoal software (Rannala and Yang, 2003). 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. 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. With regard to CLADES' performance, longer loci improved its efficiency, and this approach 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 RF classifier. In essence, this framework aims to discriminate between various divergence models compatible with virtually any species concept, as asserted by the authors. Besides, working with data summarized through the mSFS, delimitR facilitates the evaluation of models that vary in terms of lineage numbers. Either way, given its supervised nature,

392 delimitR demands researchers to define reasonable priors, such as divergence
393 times or migration rates, and decide which models will be assessed. For each
394 model evaluated in their study, Smith and Carstens (2020) simulated 10,000
395 mSFS. A RF classifier was constructed using 1,000 **decision trees** to
396 accommodate the extensive number of models. delimitR's performance improved
397 with larger SNP matrices and increasing divergence times. Compared to ABC
398 methods, delimitR showed lower error rates, even though the detection of
399 migration was challenging in cases of recent divergence between lineages (Smith
400 and Carstens, 2020). The authors acknowledge that further research is needed
401 to elucidate the association between the model space, number of parameters,
402 and delimitation accuracy.
403
404 *2.3. Deep learning*
405 Deep learning is a subset of ML that focuses on training **artificial neural**
406 **networks (ANNs)** with multiple layers (hence "deep") to perform complex tasks
407 (Sheehan and Song, 2016). In terms of data labeling, deep learning algorithms
408 can be both supervised or unsupervised. Deep learning techniques have found
409 success in various fields in the Biological Sciences (Angermueller et al., 2016;
410 Sheehan and Song, 2016; Schrider and Kern, 2018). However, its adoption in
411 Evolutionary Biology is relatively recent (see Angermueller et al., 2016; Sheehan
412 and Song, 2016; Fonseca et al., 2021; Blischak et al., 2021; Yelmen and Jay,
413 2023). The popularity of deep learning can be attributed to their highly flexible
414 data input and output structure, allowing networks trained for one task to be
415 repurposed for another by modifying their final **layers**, for instance, through
416 **transfer learning** approaches. This versatility enables the resolution of intricate

417 tasks that might prove challenging for **shallow learning** algorithms. Conversely,
418 deep learning often demands meticulous and more specific fine-tuning compared
419 to shallow learning methods. For a detailed description of how neural networks
420 work, and their general structure, see Sheehan and Song (2016), Borowiec et al.
421 (2022), and Korfmann et al. (2023).

422 The fundamental stages involved in creating a deep learning framework
423 for species delimitation, especially a supervised one, closely parallel those of a
424 shallow SML workflow, as both typically involve formulating evolutionary models
425 and simulating data. Broadly, these steps include data simulation and
426 representation, **model** training and optimization, and ultimately, predicting
427 relevant categories from empirical data (Fig. 3).

428

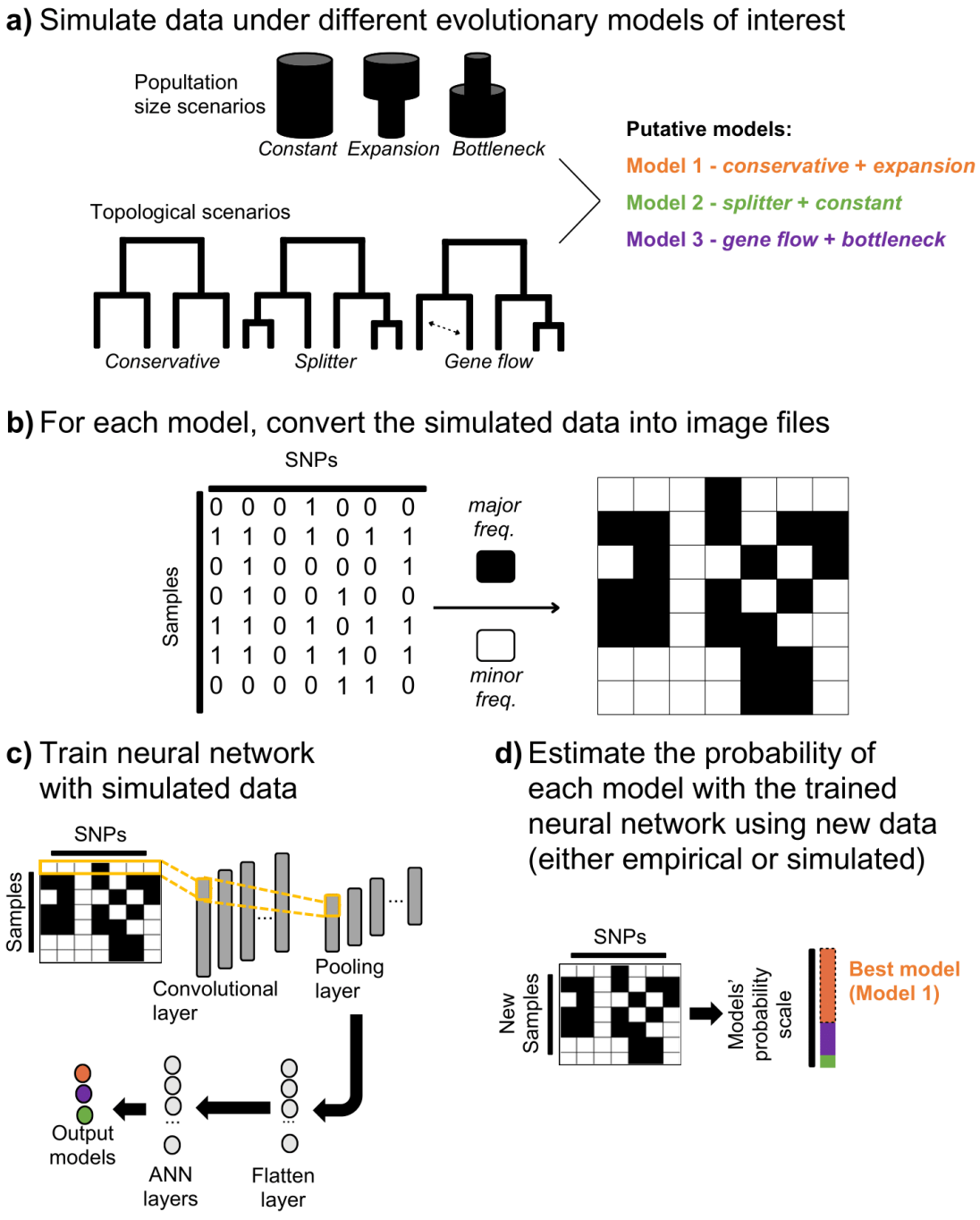

Fig. 4. Diagram illustrating a potential deep learning workflow applied in the context of species delimitation, using CNNs as an example of algorithm that can be used in this context (inspired by Perez et al., 2021). a) The process typically begins with the simulation of biological data under various evolutionary models, considering factors like topology (e.g., scenarios with differing numbers of potential species, namely 'conservative' and 'splitter'), population size (considering potential demographic variations over time, whether of population contraction or expansion), gene flow, and many more, similar to a SML pipeline. 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) proposed a delimitation approach that accommodates the integration of coalescence-based methods with model selection using **convolutional neural networks (CNNs)**. In short, this approach can integrate models from coalescent analyses, such as BPP (Flouri et al., 2018; 2020), to compare different evolutionary scenarios while incorporating information from multiple sources. Specifically, it allows users to combine insights from genetic analyses (e.g., coalescent-based methods) with hypotheses derived from other data types (e.g., phenotypic traits) that reflect different taxonomic arrangements. This flexibility enables the formulation of models informed by multiple lines of evidence. The initial steps in this approach involve simulating genetic data for each delimitation hypothesis, with the study encompassing 10,000 simulations per model. These simulations are then converted into images, which serve as input for training a neural network. It is worth noting that while CNNs used 10,000 simulations per model in this study, ABC required 100,000 simulations per model. Finally, each species hypothesis probability can be predicted through the trained CNNs using a **test set**. Perez et al. (2021) also compared their model selection approach with ABC using empirical data. The CNNs consistently demonstrated superior performance in distinguishing between the simulated evolutionary 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, see Appendix B) have specifically explored ML techniques for species delimitation, in particular 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 demographic inferences, phylogeography and population genetics, 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). Also, while ML algorithms are often used similarly to simulation-based approaches like ABC, additional steps are generally incorporated, such as: i) selecting a more comprehensive subset of summary statistics based on specific criteria (Smith and Carstens, 2020; Martin et al., 2021); ii) handling larger or more complex genetic datasets more efficiently compared to model selection tools such as ABC (Smith and Carstens, 2020; Collin et al., 2021; Ghirotto et al., 2021). These advantages stem from the fact that ML approaches usually require less

497 specificity in summary statistic selection and can manage high-dimensional data
498 with fewer concerns about the **curse of dimensionality**.
499
500 *2.5. What types of species ML methods might be detecting?*

501 A significant part of the studies we analyzed were philosophically based
502 on species concepts grounded on evolutionary or genealogical independence
503 criteria. This might stem from our focus on workflows using molecular data, which
504 generally aims at identifying lineages and genetic clusters characterized by
505 significant levels of genetic divergence and restricted amounts of gene flow. Also,
506 some studies specifically model parameters like migration, which make them in
507 line with concepts focused on reproductive criteria. While evolutionary and
508 genealogical independence evidence (or reproductive criteria) may have their
509 limitations in investigating species limits, results generated by ML methods in this
510 context can still serve as hypotheses for further investigations (e.g., Fujita et al.,
511 2012), aligning with the GLC perspective (de Queiroz, 1998; 1999; 2005b).

512 So far, there are no definitive coalescent-based solutions to differentiate
513 between population structure and species (Sukumaran and Knowles, 2017;
514 Leaché et al., 2019). In this context, it is reasonable to assert that ML-based
515 delimitation methods, just as coalescence-based methods, might not always be
516 identifying species *per se*, but rather: i) incompletely separated (or incipient)
517 species, which may eventually be classified as distinct (Burbrink et al., 2021), or
518 even as 'subspecies' (de Queiroz, 2020); or ii) population or phylogeographic
519 variation (Rosenblum et al., 2012; Sukumaran et al., 2021). Consequently, while
520 genetic structure (either at population or species level) detected through ML can
521 be biologically relevant for species delimitation, additional data and an

evolutionary process-based perspective remain crucial to discern the nature of the inferred biological entities (Smith and Carstens, 2020; Sukumaran et al., 2021). 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 could serve as secondary evidence of divergence and may be also analyzed using ML frameworks. Given ML's versatility in handling diverse data types, future applications in species delimitation should prioritize the explicit incorporation and evaluation of diverse biological properties—such as genomic divergence, ecological adaptation, and phenotypic differentiation—to enhance species hypothesis testing (e.g., Karbstein et al., 2024; Pyron et al., 2024). Several strategies can achieve this, including integrating different feature categories as distinct layers within a deep learning architecture. Besides, investigating how the contribution of various traits impacts species delimitation across different biological systems also represents a key avenue for future research. Only a few detailed ML pipelines have been proposed in this context, aiming to explore the relationship between evolutionary models and divergence scenarios in terms of distinct characteristics, whether genetic, phenotypic, geographic or ecological. Pyron (2023), for instance, 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). Yang et al. (2022) is another great example, which introduced a CNN

method that successfully integrates morphological and molecular data for species identification.

In sum, integrating genetic, ecological, and phenotypic data may be essential for achieving robust and reliable species limits assessments, particularly in systems with complex evolutionary histories—such as cryptic species complexes or hybridizing lineages. In this context, ML-based species delimitation offers a powerful framework to combine domain expertise with quantitative hypothesis testing, optimizing the reconciliation of conflicting evidence. This approach also paves the way for establishing comprehensive frameworks rooted in modern Integrative Taxonomy, potentially enabling the automated synthesis of diverse data to accurately define taxonomic units (Karbstein et al., 2024).

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 based on genetic data offer some advantages over coalescent or traditional simulation-based methods. Despite particular constraints, ML algorithms can be as accurate (in biological terms) as 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). Because of their likelihood-free nature, 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).

A major advantage of deep learning, in particular, is the capacity to automatically extract information from alignments (commonly treated as images), as opposed to relying on summary statistics typically required by other 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), holding promise in future species delimitation studies. Besides, especially in supervised approaches, which often use explicit evolutionary 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. Factors requiring careful consideration in ML-based species delimitation

Certain algorithms, especially supervised ones trained on simulated data, may become 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, as models 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 must carefully design prior distributions to ensure that the data generated under these models closely reflect the actual biological system being studied. This is a challenge for non-model organisms, where data availability may limit the quality of parameter estimates (Tagu et al., 2014; Fonseca et al., 2016; Cerca et al., 2021; Jorna et al., 2021). Importantly, 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). Furthermore, 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), 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)). Also, limited information is available regarding the asymptotic statistical performance of most ML methods applied for species delimitation. Thus, such models may never be comprehensive

622 enough, have limitations in representing real data, and demand substantial
623 computational resources (Arenas, 2012; Mangul et al., 2019; Zaharias et al.,
624 2022). This leads to an inherent challenge in avoiding some degree of
625 misspecification in the training data, even considering the variety of powerful
626 genetic data simulators currently available.

627 Moreover, all model-based methods depend on the chosen models and
628 their parameters, whether they are used for simulations or for direct likelihood
629 estimation. As a result, even methods that do not rely on simulations can still be
630 sensitive to model misspecification. For example, coalescence-based
631 approaches depend on MSC assumptions, which may not always accurately
632 represent specific biological systems. Likelihood-based approaches offer
633 advantages in exploring parameter space within a defined model—due to their
634 optimality and iterative nature—though they can be computationally intensive
635 (e.g., Flouri et al., 2018; Sukumaran et al., 2021). Thus, these methods remain
636 important alternatives especially when there is no clear reference for simulations.
637 ML approaches, on the other hand, have the potential to explore a broader model
638 space. That said, no approach can account for all possible evolutionary
639 processes, leaving both traditional SDMs and ML approaches limited in their
640 ability to comprehensively explore the broad space of evolutionary models. Only
641 UML approaches might be relatively immune to some of these constraints, as
642 they do not rely on predefined models. Either way, regardless of the analytical
643 framework, misrepresenting evolutionary relationships can lead to misleading
644 outcomes. This underscores the need for biologically informed feature processing
645 and modeling.

Another important perspective to consider is related to data representation. While ML can uncover patterns in high-dimensional datasets, its performance heavily relies on the quality and relevance of the input data and how it is representated (Guyon and Elisseeff, 2003; Domingos, 2012; LeCun et al., 2015). In the context of the present study, there are ML pipelines that utilize data derived from SNPs 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), and only a few are extensible to different genetic markers (e.g., Collin et al., 2021). Also, numerous studies using ML frameworks, whether focusing on species delimitation, demography, or population genetics, use genetic summary statistics as the main input data (e.g., Pei et al., 2018; Collin et al., 2021; Ghirotto et al., 2021).

While summary statistics can be valuable for distinguishing between models, some may not be suitable for making inferences about species limits. Besides, 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). Therefore, unless we have a clear understanding of which type of data is truly sufficient to capture the target biological phenomena, relying solely on a specific set of statistics can lead to inevitable information loss (Rannala and Yang, 2020). An alternative to this is to consider the sequence alignment itself as input, as demonstrated in the deep learning approach introduced by Perez et al. (2021), which implicitly enables dimensionality reduction while capturing structures within the input data. Notably, deep learning

techniques are valuable tools in this context, offering the capability to analyze both raw sequence data and summary statistics (Korfmann et al., 2023).

Even considering that data representation is a critical component of any analytical framework, its role in species delimitation demands particular attention, where complex evolutionary processes such as gene flow and incomplete lineage sorting (ILS) can leave distinct signatures in the data. For instance, gene flow may produce localized discordance in allele frequencies and introgressed genomic segments, whereas ILS typically results in widespread gene tree incongruence due to ancestral polymorphisms. Consequently, how data is represented (e.g., via full sequence alignments, SNP matrices, or gene tree reconstructions) can impact the ability to detect and distinguish these phenomena, and consequently the robustness and interpretability of delimitation results. Therefore, priority should be given to representations that retain detailed information for detecting key processes in species delimitation while preserving the inherent variability and structure of the data. Future research should focus on understanding the extent to which the flexibility of ML to handle various input data types provides a true advantage for species delimitation. Moreover, despite the challenges associated with comparing ML approaches due to differing assumptions, employing diverse training data representations—from genomic sequences to summary statistics—could help illuminate the strengths and limitations of each method for detecting species limits.

3.3. Possible avenues and prospects for future studies

Mitigating the effects of misspecification during simulation might involve designing or using a simulator that enforces high 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 Wang et al., 2021; Callier, 2022). GANs operate with two networks, the generator and the discriminator, trained together (Goodfellow et al., 2014). While the generator simulates data, the discriminator distinguishes between real and synthetic data. During 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. When training is complete, the generator network can 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 (Nesterenko et al., 2022 preprint; Booker et al., 2023; Yelmen and Jay, 2023). Smith and Hahn (2023) 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.

The application of GANs is still incipient in Evolutionary Biology. Although the above-mentioned approaches perform well in relatively simple scenarios, methodological challenges arise as the complexity of the evolutionary model space increases. This can result from additional variables in evolutionary models or larger phylogenetic trees and sequence alignments, potentially affecting both accuracy and computational efficiency (Nesterenko et al., 2022 preprint; Smith and Hahn, 2023). Therefore, future advancements in the use of GANs in should focus on enhancing the efficiency of exploring parameter spaces, reducing computational training times, and accommodating more complex evolutionary

models (Smith and Hahn, 2023). To fully harness the potential of this tool in species delimitation, further efforts are required to refine the population genetics parameters estimates (e.g., Wang et al., 2021), and to improve the accuracy of species limits inferences based on these parameters. Future GAN applications in this context should also focus on generating synthetic datasets to model realistic scenarios of species divergence under complex evolutionary processes.

Potential errors in data simulation can be linked to a "domain adaptation" problem as well, 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 the only labeled training data available are 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 solutions involve learning a "domain-invariant" data representation through a feature extractor neural network. This is accomplished by minimizing domain disparities (Rozantsev et al., 2018), using 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 robust deep learning models in evolutionary biology inference (Mo and Siepel, 2023), potentially including species delimitation. In this area, future efforts should focus on mitigating problems introduced by sampling bias or model misspecifications across diverse evolutionary scenarios, particularly in supervised frameworks that rely on simulated data. Employing domain adaptation strategies to facilitate the integration of naturally heterogeneous datasets—such as genomic and morphological, environmental and geographical data—by extracting of domain-invariant features will also enhance the resolution and reliability of delimitation outcomes.

Future ML applications to infer species limits should also focus on the development of new transfer learning structures. For example, a deep learning **architecture** originally trained for inferring historical population sizes can be repurposed for classifying demographic scenarios (Pan and Yang, 2010), even though reusing trained models can be challenging due to differences in data dimensionality (Sanchez et al., 2020). Particularly regarding species delimitation, improving model generalizability could be achieved by transferring learning between well-studied taxonomic groups and those with limited data availability. In order to establish baseline models, a practical workflow could involve using ML algorithms to identify species limits in a broad training dataset, such as population-level genomic or morphological data from different species. Then, these baseline models could be fine-tuned with smaller, taxonomically specific datasets to validate or identify taxa in understudied groups. This iterative

771 approach can also optimize computational efficiency by avoiding the need to train
772 models from scratch. A methodology relatively aligned with this reasoning is
773 exemplified in the study by Derkarabetian et al. (2022). Moreover, ML methods
774 initially designed for other model selection purposes, such as phylogeography
775 (Fonseca et al., 2021), could be reasonably adapted for species delimitation,
776 provided that the simulated data and initial models adequately capture species
777 limits nuances.
778
779 **4. Optimizing the use of ML in the context of species delimitation**
780 *4.1. Enhancing Species Delimitation through accessible and purpose-built ML*
781 The introduction of new ML approaches will increasingly enhance
782 researchers' ability to make biologically precise decisions, especially when these
783 methods are purpose-built, from conception to implementation, for the specific
784 task of delimiting evolutionary lineages. In order to choose the appropriate
785 species delimitation method, researchers must consider the available data and
786 putative evolutionary scenarios, while considering the available statistical
787 evaluation and performance optimization of each method (Greener et al., 2021;
788 Morimoto et al., 2021). However, a comprehensive comparison of the recently
789 proposed ML methods characteristics, advantages and disadvantages, and
790 overall performance compared to other SDMs is still lacking.
791 Such evaluation should consider the inherent properties of the ML
792 algorithms, such as how the workflows manipulate the data attributes, and the
793 different types of data. In nearly all studies using ML methods to infer species
794 limits, at least a minimal approach to estimating error or noise is employed (Pei
795 et al., 2018; Smith and 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. The quantity and quality of data clearly influences the effectiveness of ML applications, as analyses conducted on larger, well-filtered datasets consistently yield better delimitation results (Pei et al., 2018; Smith and 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, as should be the case in any scientific field, evaluating the suitability of an ML tool for species delimitation also involves assessing its accessibility and reproducibility, particularly compared to traditional SDMs. 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. A good example that circumvents these problems can be found on the study by Derkarabetian et al. (2022), where the authors assessed a ML approach capability to delimit cryptic species constructing a "customized" training dataset from a well-studied lineage with biological characteristics akin to their focal taxon, and clearly explained the rationale behind the customizations made to the datasets and pre-trained models. 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 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).

4.2. Integrating 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. In some systems, certain methods should be prioritized based on the processes driving divergence, and using multiple methods with similar biases might not enhance biological interpretability. Therefore, the choice of which species delimitation method to use should be done before or during the hypothesis-formulation process, considering the nature of the available data and, possibly, prior relevant biological information regarding the evolution of the organisms.

To effectively prioritize methods, researchers should consider key factors such as the evolutionary context (e.g., presence of gene flow, divergence times, demographic parameters) and the type and quality of the available data (e.g., genomic vs. phenotypic data). For instance, Smith and Carstens (2020) precisely 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 address more efficiently. Thus, while some ML-based methods may be particularly well-suited for systems where distinguishing between divergence with gene flow and strict isolation is critical,

coalescent-based methods may perform better in detecting fine-scale population structure.

As previously discussed, incorporating multiple delimitation criteria is essential for capturing the diverse mechanisms driving speciation, as different evolutionary processes leave distinct signatures in genetic and non-genetic data. But even considering that inferring species limits from molecular data and integrating phenotypic data can be a solution in some cases, robust species delimitation still requires mechanistic hypotheses about the speciation process itself (Padial and De la Riva, 2021; Pyron et al., 2023; Pyron et al., 2024), as distinguishing between population structure and diverging or collapsing species require explicit hypotheses and quantifiable tests (Sukumaran and Knowles, 2017; Derkarabetian et al., 2019; Huang, 2020; Pyron et al., 2024). In this context, a promising approach that could shape the future of genetic-based species delimitation (and would greatly benefit from integrating different delimitation approaches) is the empirical validation of **speciation-based models**, which offers a more nuanced understanding of the speciation process (see Sukumaran et al., 2021; Hua and Moritz, 2025). For instance, divergence with gene flow can create a pattern of genomic heterogeneity where some loci reflect historical connectivity while others indicate reproductive isolation (see Harrison and Larson, 2016). Such cases may be better captured by models that incorporate allele frequency shifts across loci or explicit tests for introgression (e.g., network-based methods or ML classifiers trained on specific genomic features). Also, the temporal dynamics of divergence, such as recent speciation events with ILS, may be more appropriately addressed by coalescent-based models that account for stochasticity in gene tree variation. Therefore, such a process-based approach

might be particular useful in distinguishing intraspecific genetic structure from interspecific divergence.

In the current state of affairs, while ML methods are still being developed and refined, explicitly considering the evolutionary mechanisms underlying species divergence, and strategically integrating this approach with different delimitation tools can enhance both the accuracy and biological realism of species limits. ML-based methods will probably not replace coalescent or tree-based approaches in the near future, but rather complement them by leveraging their particular strengths. Even so, ML is sure 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.

5. Conclusions

- Relatively few studies have yet applied ML techniques to species delimitation using molecular data. Nonetheless, these approaches have already proved to be computationally efficient, and capable of being readily integrated into diverse analytical frameworks, providing a robust way to explore dataset structure when species-level divergences are hypothesized.
- Existing ML-based genetic species delimitation frameworks use various data representation as input (e.g., sequences treated as images, summary statistics, SNPs). Although this flexibility can be advantageous, the reliance on particular representations may bias the accurate delineation of species limits. Assessing the impact of data transformation on delimitation outcomes remains a challenge, and is a key avenue for future research.

- Overly specialized ML algorithms might perform well within the specific ranges of evolutionary parameters present in their training data, but struggle when applied beyond that parameter space. This is particularly critical given the heavy reliance on simulated data in evolutionary biology, where overspecialization can compromise generalizability and transferability—especially in supervised pipelines. Emerging approaches offer promising solutions, including but not limited to the use of transfer learning approaches to exchange knowledge across datasets, GANs to produce more realistic simulated data, and domain adaptation techniques to address the challenges of working with inherently heterogeneous datasets.
- Given the flexibility of ML workflows in handling different types of data—and the multifactorial nature of divergence processes among evolutionary lineages—future research should focus on quantifying how different biological traits contribute to and influence species delimitation results across distinct biological systems.
- A key priority is the development of robust ML-based species delimitation frameworks within the context of Integrative Taxonomy. This will enable the automated integration of multiple lines of evidence to accurately define taxonomic units, while facilitating the reconstruction of the evolutionary processes underlying species limits patterns.
- Although no universally superior species delimitation method currently exists, ML algorithms present promising prospects for their integration into systematic protocols tailored for species delimitation.

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 providing feedback on the first draft of this manuscript. Finally, we wish to thank Emanuel Masiero da Fonseca for his suggestions, which helped improve the final version of the 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.

946 Beaumont, M.A., Zhang, W., Balding, D.J. 2002. Approximate Bayesian computation in
947 population genetics. Genetics 162, 2025–2035. doi:10.1093/genetics/162.4.2025
948 Bertorelle, G., Benazzo, A., Mona, S. 2010. ABC as a flexible framework to estimate
949 demography over space and time: some cons, many pros. Mol Ecol. 19, 2609–2625.
950 doi:10.1111/j.1365-294X.2010.04690.x
951 Blischak, P.D., Barker, M.S., & Gutenkunst, R. N. 2021. Chromosome-scale inference of hybrid
952 speciation and admixture with convolutional neural networks. Molecular Ecology
953 Resources 21, 2676–2688. https://doi.org/10.1111/1755-0998.13355
954 Booker, W.W., Ray, D.D., & Schrider, D.R. 2023. This population does not exist: learning the
955 distribution of evolutionary histories with generative adversarial
956 networks. Genetics, 224(2), iyad063.
957 Borowiec, M.L., Dikow, R.B., Frandsen, P.B., McKeeken, A., Valentini, G., & White, A.E. 2022.
958 Deep learning as a tool for ecology and evolution. Methods in Ecology and Evolution
959 13, 1640–1660.
960 Bortolus, A. 2008. Error cascades in the biological sciences: the unwanted consequences of
961 using bad taxonomy in ecology. AMBIO: A journal of the human environment 37, 114–
962 118.
963 Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., & Krishnan, D. 2017. Unsupervised pixel-
964 level domain adaptation with generative adversarial networks. Proceedings of the IEEE
965 conference on computer vision and pattern recognition, 3722–3731.
966 Breitman, M.F., Domingos, F.M., Bagley, J.C., Wiederhecker, H.C., Ferrari, T.B., Cavalcante,
967 V.H., ... & Colli, G.R. 2018. A new species of *Enyalius* (Squamata, Leiosauridae)
968 endemic to the Brazilian Cerrado. Herpetologica 74, 355–369.
969 Burbrink, F.T., & Ruane, S. 2021. Contemporary philosophy and methods for studying
970 speciation and delimiting species. Ichthyology & Herpetology 109, 874–894.
971 Callier, V. 2022. Machine learning in evolutionary studies comes of age. Proceedings of the
972 National Academy of Sciences 119.
973 Carstens, B. C., Pelletier, T. A., Reid, N. M., & Satler, J. D. 2013. How to fail at species
974 delimitation. Molecular Ecology 22, 4369–4383.

975 Cerca, J., Maurstad, M.F., Rochette, N. C., Rivera-Colón, A.G., Rayamajhi, N., Catchen, J.M., &
976 Struck, T H. 2021. Removing the bad apples: A simple bioinformatic method to improve
977 loci-recovery in de novo RADseq data for non-model organisms. Methods in Ecology
978 and Evolution 12, 805–817.
979 Chadha, A., Kumar, V., Kashyap, S., & Gupta, M. 2021. Deepfake: An overview. In Proceedings
980 of Second International Conference on Computing, Communications, and Cyber-
981 Security, pp. 557-566. Springer, Singapore.
982 Chicco, D. 2017. Ten quick tips for machine learning in computational biology. BioData Mining
983 10, 1–17. https://doi.org/10.1186/s13040-017-0155-3
984 Christin, S., Hervet, É., & Lecomte, N. 2019. Applications for deep learning in ecology. Methods
985 in Ecology and Evolution 10, 1632–1644.
986 Cochran, K., Srivastava, D., Shrikumar, A., Balsubramani, A., Hardison, R.C., Kundaje, A.,
987 Mahony, S. 2022. Domain adaptive neural networks improve cross-species prediction of
988 transcription factor binding. Genome Res. 32, 512–523.
989 Collin, F.D., Durif, G., Raynal, L., Lombaert, E., Gautier, M., Vitalis, R., Marin, J.M., & Estoup, A.
990 2021. Extending approximate Bayesian computation with supervised machine learning
991 to infer demographic history from genetic polymorphisms using DIYABC Random
992 Forest. Molecular Ecology Resources 21, 2598–2613. https://doi.org/10.1111/1755-
993 0998.13413.
994 Csilléry, K., Blum, M.G., Gaggiotti, O.E., & François, O. 2010. Approximate Bayesian
995 computation (ABC) in practice. Trends in Ecology & Evolution 25, 410–418.
996 Dayrat, B. 2005. Towards integrative taxonomy. Biological journal of the Linnean society, 85,
997 407–417.
998 Degnan, J. H. & Rosenberg, N.A. 2009. Gene tree discordance, phylogenetic inference and the
999 multispecies coalescent. Trends Ecol. Evol. 24, 332–340.
1000 Derkarabetian, S., Castillo, S., Koo, P.K., Ovchinnikov, S., & Hedin M. 2019. A demonstration of
1001 unsupervised machine learning in species delimitation. Molecular Phylogenetics and
1002 Evolution 139. https://doi.org/10.1016/j.ympev.2019.106562

1003 Derkarabetian, S., Starrett, J., & Hedin, M. 2022. Using natural history to guide supervised
1004 machine learning for cryptic species delimitation with genetic data. Frontiers in Zoology
1005 19, 1–15.
1006 Dike, H.U., Zhou, Y., Deveerasetty, K.K., & Wu, Q. 2018. Unsupervised learning based on
1007 artificial neural network: A review. In 2018 IEEE International Conference on Cyborg
1008 and Bionic Systems (CBS), pp. 322-327.
1009 Domingos, P. 2012. A few useful things to know about machine learning. *Communications of*
1010 *the ACM*, *55*, 78-87.
1011 Domingos, F.M., Bosque, R.J., Cassimiro, J., Colli, G.R., Rodrigues, M.T., Santos, M.G., &
1012 Beheregaray, L. B. 2014. Out of the deep: cryptic speciation in a Neotropical gecko
1013 (Squamata, Phyllodactylidae) revealed by species delimitation methods. Molecular
1014 Phylogenetics and Evolution 80, 113–124.
1015 *Duan, L., Han, L.N., Liu, B., Leostrin, A., Harris, A.J., Wang, L., Arslan, E., Ertuğrul, K.,
1016 Knyazev, M., Hantemirova, E., Wen, J., & Chen, H.F. 2023. Species delimitation of the
1017 liquorice tribe (Leguminosae: Glycyrrhizeae) based on phylogenomic and machine
1018 learning analyses. Journal of Systematics and Evolution 61, 22–41.
1019 https://doi.org/10.1111/jse.12902.
1020 Edwards, S. V. 2009. Is a new and general theory of molecular systematics emerging?
1021 Evolution 63, 1–19.
1022 Edwards, S.V., Xi, Z., Janke, A., Faircloth, B.C., McCormack, J.E., Glenn, T.C., ... & Davis, C.C.
1023 2016. Implementing and testing the multispecies coalescent model: a valuable
1024 paradigm for phylogenomics. Molecular Phylogenetics and Evolution 94, 447–462.
1025 Edwards, S. V., Potter, S., Schmitt, C. J., Bragg, J. G., & Moritz, C. 2016. Reticulation,
1026 divergence, and the phylogeography–phylogenetics continuum. Proceedings of the
1027 National Academy of Sciences, 113, 8025–8032.
1028 Ely, C.V., de Loreto Bordignon, S.A., Trevisan, R., & Boldrini, I.I. 2017. Implications of poor
1029 taxonomy in conservation. Journal for Nature Conservation 36, 10–13.
1030 *Fan, X.K., Wu, J., Comes, H.P., Feng, Y., Wang, T., Yang, S.Z., Iwasaki, T., Zhu, H., Jiang, Y.,
1031 Lee, J., & Li, P. 2023. Phylogenomic, morphological, and niche differentiation analyses
1032 unveil species delimitation and evolutionary history of endangered maples in Acer

1033 series Campestria (Sapindaceae). Journal of Systematics and Evolution 61, 284–298.
1034 https://doi.org/10.1111/jse.12919.
1035 Farahani, A., Voghoei, S., Rasheed, K., & Arabnia, H.R. 2021. A brief review of domain
1036 adaptation. Advances in data science and information engineering: proceedings from
1037 ICDATA 2020 and IKE 2020, 877–894.
1038 Flagel, L., Brandvain, Y., & Schrider, D.R. 2019. The unreasonable effectiveness of
1039 convolutional neural networks in population genetic inference. Molecular Biology and
1040 Evolution 36, 220–238.
1041 Flouri, T., Jiao, X., Rannala, B., Yang, Z. 2018. Species Tree Inference with BPP using
1042 Genomic Sequences and the Multispecies Coalescent. Molecular Biology and Evolution
1043 35, 2585–2593. doi:10.1093/molbev/msy147.
1044 2020. A Bayesian implementation of the multispecies coalescent model with
1045 introgression for phylogenomic analysis. Molecular Biology and Evolution 37, 1211–
1046 1223.
1047 Fonseca, R.R. et al. 2016. Next-generation biology: sequencing and data analysis approaches
1048 for non-model organisms. Marine genomics 30, 3–13.
1049 Fonseca, E. M., Colli, G. R., Werneck, F. P., & Carstens, B. C. 2021. Phylogeographic model
1050 selection using convolutional neural networks. Molecular Ecology Resources 21, 2661–
1051 2675. https://doi.org/10.1111/1755-0998.13427.
1052 Fonseca, E. M., & Carstens, B. C. (2024). Artificial intelligence enables unified analysis of
1053 historical and landscape influences on genetic diversity. *Molecular Phylogenetics and*
1054 *Evolution*, 108116.
1055 Fountain-Jones, N.M., Smith, M.L., & Austerlitz, F. 2021. Machine learning in molecular
1056 ecology. Molecular Ecology Resources 21, 2589–2597. https://doi.org/10.1111/1755-
1057 0998.13532.
1058 Fujita, M.K., Leaché, A.D., Burbrink, F.T., McGuire, J.A., & Moritz, C. 2012. Coalescent-based
1059 species delimitation in an integrative taxonomy. Trends in Ecology & Evolution 27, 480–
1060 488.
1061 Ganin, Y., & Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation.
1062 In International conference on machine learning, 1180–1189.

1063 Gehara, M., Mazzochinni, G.G., & Burbrink, F. 2020. PipeMaster: inferring population
1064 divergence and demographic history with approximate Bayesian computation and
1065 supervised machine-learning in R. BioRxiv, 2020-12.
1066 https://doi.org/10.1101/2020.12.04.410670
1067 Ghifary, M, Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W. 2016. Deep Reconstruction
1068 Classification Networks for Unsupervised Domain Adaptation. In: Leibe B, Matas J,
1069 Sebe N, Welling M, editors. Computer Vision ECCV 2016. Lecture Notes in Computer
1070 Science. Cham: Springer International Publishing. p. 597
1071 Ghirotto, S., Vizzari, M.T., Tassi, F., Barbujani, G. & Benazzo, A. 2021. Distinguishing among
1072 complex evolutionary models using unphased whole-genome data through random
1073 forest approximate Bayesian computation. Molecular Ecology Resources 21, 2614–
1074 2628. https://doi.org/10.1111/1755-0998.13263.
1075 Gompert, Z., Mandeville, E. G., & Buerkle, C. A. 2017. Analysis of population genomic data
1076 from hybrid zones. Annual Review of Ecology, Evol., and Syst., 48, 207–229.
1077 Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. &
1078 Bengio Y. 2014. Generative adversarial nets. Advances in Neural Information
1079 Processing Systems, 2672–2680.
1080 Greener, J.G., Kandathil, S.M., Moffat, L., & Jones, D.T. 2021. A guide to machine learning for
1081 biologists. Molecular Cell Biology 23, 40–55. https://doi.org/10.1038/s41580-021-00407-
1082 0.
1083 Guyon, I., & Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of*
1084 *machine learning research*, 3, 1157-1182.
1085 Harrison, R. G., & Larson, E. L. 2016. Heterogeneous genome divergence, differential
1086 introgression, and the origin and structure of hybrid zones. Molecular ecology 25, 2454–
1087 2466.
1088 Hastie, T., Tibshirani, R., & Friedman, J. 2009. Unsupervised learning. In The elements of
1089 statistical learning (pp. 485-585). Springer, New York, NY.
1090 Hausdorf, B. 2011. Progress toward a general species concept. Evolution 65, 923–931.

1091 Heil, B.J., Hoffman, M. M., Markowetz, F., Lee, S.I., Greene, C.S. & Hicks, S.C. 2021.
1092 Reproducibility standards for machine learning in the life sciences. Nature Methods 18,
1093 1132–1135.
1094 *Hodel, R.G., Winslow, S.K., Liu, B.B., Johnson, G., Trizna, M., White, A.E., ... & Wen, J. 2023.
1095 A phylogenomic approach, combined with morphological characters gleaned via
1096 machine learning, uncovers the hybrid origin and biogeographic diversification of the
1097 plum genus. bioRxiv, 2023-09. https://doi.org/10.1101/2023.09.13.557598
1098 Hüllermeier, E., Fober, T. & Mernberger, M. 2013. Inductive bias. Encyclopedia of systems
1099 biology, 1018–1019.
1100 Hua, X., & Moritz, C. 2025. A phylogenetic approach to delimitate species in a probabilistic
1101 way. *Systematic Biology*, syaf004.
1102 Huang, J. P. 2020. Is population subdivision different from speciation? From phylogeography to
1103 species delimitation. Ecology and Evolution 10, 6890–6896.
1104 Huelsenbeck, J.P., Andolfatto, P., Huelsenbeck, E.T. 2011. Structurama: Bayesian inference of
1105 population structure. Evolutionary Bioinformatics 7, 55–59.
1106 Jackson, N.D., Carstens, B.C., Morales, A.E. & O'Meara B.C. 2017. Species delimitation with
1107 gene flow. Systematic Biology 66, 799–812.
1108 Jacobs, S. J., Kristofferson, C., Uribe-Convers, S., Latvis, M., & Tank, D. C. (2018).
1109 Incongruence in molecular species delimitation schemes: What to do when adding more
1110 data is difficult. Molecular Ecology 27, 2397–2413.
1111 *Jamdade, R., Al-Shaer, K., Al-Sallani, M., Al-Harthi, E., Mahmoud, T., Gairola, S., & Shabana,
1112 H.A. 2022. Multilocus marker-based delimitation of *Salicornia persica* and its population
1113 discrimination assisted by supervised machine learning approach. PLoS ONE 17.
1114 https://doi.org/10.1371/journal.pone.0270463.
1115 Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., ... & Zhang, G. 2014. Whole-
1116 genome analyses resolve early branches in the tree of life of modern birds. Science
1117 346, 1320–1331.
1118 Jörger, K.M., & Schrödl, M. 2013. How to describe a cryptic species? Practical challenges of
1119 molecular taxonomy. Frontiers in Zoology 10, 1–27.

1120 Jorna, J. et al. 2021. Species boundaries in the messy middle—A genome-scale validation of
1121 species delimitation in a recently diverged lineage of coastal fog desert lichen fungi.
1122 Ecology and Evolution 11, 18615-18632.
1123 Karbstein, K., Kösters, L., Hodač, L., Hofmann, M., Hörandl, E., Tomasello, S., ... & Wäldchen,
1124 J. (2023). Species delimitation 4.0: integrative taxonomy meets artificial intelligence.
1125 *Trends in Ecology & Evolution*.
1126 *Khalighifar, A., Brown, R.M., Goyes Vallejos, J., & Peterson, A.T. 2021. Deep learning
1127 improves acoustic biodiversity monitoring and new candidate forest frog species
1128 identification (genus *Platymantis*) in the Philippines. Biodiversity and Conservation, 30,
1129 643-657.
1130 Knowles, L. L., & Carstens, B. C. 2007. Delimiting species without monophyletic gene
1131 trees. Syst. Biol., 56, 887–895.
1132 Korfmann, K., Gaggiotti, O.E. & Fumagalli, M. 2023. Deep learning in population genetics.
1133 Genome Biology and Evolution. https://doi.org/10.1093/gbe/evad008.
1134 Kuzenkov, O., Morozov, A., & Kuzenkova, G. 2020. Exploring evolutionary fitness in biological
1135 systems using machine learning methods. Entropy 23, 1–35.
1136 Leaché, A.D., Harris, R.B., Rannala, B. & Yang, Z. 2014. The influence of gene flow on species
1137 tree estimation: a simulation study. Systematic Biology 63, 17–30.
1138 Leaché, A.D., Zhu, T., Rannala, B., & Yang, Z. 2019. The spectre of too many
1139 species. Systematic Biology 68, 168–181.
1140 LeCun, Y., Bengio, Y., & Hinton, G. 2015. Deep learning. *Nature*, *521*, 436-444.
1141 Libbrecht, M.W. & Noble, W.S. 2015. Machine learning applications in genetics and genomics.
1142 Nature Reviews Genetics 16, 32–332.
1143 *Lima, A.P. et al. 2020a. Not as widespread as thought: Integrative taxonomy reveals cryptic
1144 diversity in the Amazonian nurse frog *Allobates tinae* Melo-Sampaio, Oliveira and
1145 Prates, 2018 and description of a new species. Journal of Zoological Systematics and
1146 Evolutionary Research, 58(4), 1173–1194.
1147 *Lima, L.R. et al. 2020b. Below the waterline: cryptic diversity of aquatic pipid frogs (*Pipa
1148 carvalhoi*) unveiled through an integrative taxonomy approach. Systematics and
1149 Biodiversity, 18(8), 771–783.

1150 Liu, B. 2019. BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based
1151 on machine learning approaches. Briefings in bioinformatics 20, 1280–1294.
1152 Liu, M.Y. & Tuzel, O. 2016. Coupled Generative Adversarial Networks. In: Advances in Neural
1153 Information Processing Systems 29. Curran Associates, Inc.
1154 https://papers.nips.cc/paper/2016/hash/502e4a16930e414107ee22b6198c578f-
1155 Abstract.html.
1156 Lukhtanov, V.A. 2019. Species Delimitation and Analysis of Cryptic Species Diversity in the XXI
1157 Century. Entmol. Rev. 99, 463–472. https://doi.org/10.1134/S0013873819040055.
1158 Lürig, M.D., Donoughe, S., Svensson, E.I., Porto, A. & Tsuboi, M. 2021. Computer vision,
1159 machine learning, and the promise of phenomics in ecology and evolutionary
1160 biology. Frontiers in Ecology and Evolution 9.
1161 *Magalhães, F.D.M., Lyra, M.L., De Carvalho, T.R., Baldo, D., Brusquetti, F., Burella, P., ... &
1162 Garda, A.A. 2020. Taxonomic review of South American Butter Frogs: Phylogeny,
1163 geographic patterns, and species delimitation in the *Leptodactylus latrans* species
1164 group (Anura: Leptodactylidae). Herpetological Monographs, 34(1), 131–177.
1165 Mallet, J., Besansky, N., & Hahn, M. W. 2016. How reticulated are species? BioEssays, 38,
1166 140–149.
1167 Mangul, S. et al. 2019. Systematic benchmarking of omics computational tools. Nature
1168 communications 10.
1169 Martin, B.T., Chafin, T.K., Douglas, M.R., Placyk, Jr J.S., Birkhead, R.D., Phillips, C.A., &
1170 Douglas, M.E. 2021. The choices we make and the impacts they have: Machine
1171 learning and species delimitation in North American box turtles (*Terrapene*
1172 spp.). Molecular Ecology Resources 21, 2801–2817.
1173 Mayr, E. M. 1969. The biological meaning of species. Biological Journal of the Linnean
1174 society, 1, 311–320.
1175 1996. What is a species, and what is not? Philosophy of science, 63, 262–277.
1176 2000. The biological species concept. Species concepts and phylogenetic
1177 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.

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

Mirarab, S., Nakhleh, L., & Warnow, T. (2021). Multispecies coalescent: theory and applications in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *52*, 247-268.

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.

1208 Pei, J., Chu, C., Li, X., Lu, B., & Wu, Y. 2018. CLADES: A classification-based machine learning
1209 method for species delimitation from population genetic data. Molecular Ecology
1210 Resources 18, 1144–1156. https://doi.org/10.1111/1755-0998.12887.
1211 Perez, M.F., Bonatelli, I.A.S., Romeiro-Brito, M., Franco, F.F., Taylor, N.P., Zappi, D.C. et al.
1212 2021. Coalescent-based species delimitation meets deep learning: Insights from a
1213 highly fragmented cactus system. Molecular Ecology Resources.
1214 Pichler, M., Boreux, V., Klein, A. M., Schleuning, M. & Hartig F. 2020. Machine learning
1215 algorithms to infer trait-matching and predict species interactions in ecological networks.
1216 Methods in Ecology and Evolution 11, 281–293.
1217 Pigliucci, M. 2003. Species as family resemblance concepts: the (dis-)solution of the species
1218 problem? BioEssays, 25, 596–602.
1219 Pons, J., Barraclough, T.G., Gomez-Zurita, J. et al. 2006. Sequence-based species delimitation
1220 for the DNA taxonomy of unde-scribed insects. Systematic Biology 55, 595–609.
1221 Price, T.D., Qvarnström, A., & Irwin, D.E. 2003. The role of phenotypic plasticity in driving
1222 genetic evolution. Proceedings of the Royal Society of London. Series B: Biological
1223 Sciences 270, 1433–1440.
1224 *Pritchard, J.K., Stephens, M., Donnelly, P. 2000. Inference of population structure using
1225 multilocus genotype data. Genetics 155, 945–959.
1226 Pudlo, P., Marin, J.M., Estoup, A., Cornuet, J.M., Gautier, M., & Robert, C.P. 2016. Reliable
1227 ABC model choice via random forests. Bioinformatics 32, 859–866.
1228 https://doi.org/10.1093/bioinformatics/btv684.
1229 Pyron, R.A. 2023. Unsupervised Machine Learning for Species Delimitation, Integrative
1230 Taxonomy, and Biodiversity Conservation. Molecular Phylogenetics and Evolution, 189.
1231 Pyron, R.A., O'Connell, K.A., Duncan, S.C., Burbrink, F.T., & Beamer, D.A. 2023. Speciation
1232 hypotheses from phylogeographic delimitation yield an integrative taxonomy for Seal
1233 Salamanders (*Desmognathus monticola*). Systematic Biology, 72, 179–197.
1234 Pyron, R. A., Kakkera, A., Beamer, D. A., & O'Connell, K. A. 2024. Discerning structure versus
1235 speciation in phylogeographic analysis of Seepage Salamanders (*Desmognathus*
1236 *aeneus*) using demography, environment, geography, and phenotype. Molecular
1237 Ecology, 33, e17219.

1238 de Queiroz, K. 1998. The general lineage concept of species, species criteria, and the process
1239 of speciation. Endless forms: species and speciation.
1240 1999. The General Lineage Concept of Species and the Defining Properties of
1241 the Species Category. In book: Species: New Interdisciplinary Essays,
1242 Chapter: 3, Publisher: MIT Press, Editors: Robert A. Wilson.
1243 2005a. Ernst Mayr and the modern concept of species. Proceedings of the
1244 National Academy of Sciences, 102, 6600–6607.
1245 2005b. Different species problems and their resolution. BioEssays 27,
1246 1263–1269.
1247 2007. Species concepts and species delimitation. Syst. Biol. 56, 879–886.
1248 2011. Branches in the lines of descent: Charles Darwin and the evolution of the
1249 species concept. Biol. J. Linn. Soc. 103, 19–35.
1250 2020. An updated concept of subspecies resolves a dispute about the
1251 taxonomy of incompletely separated lineages. Herpetological Review.
1252 Qu, K., Guo, F., Liu, X., Lin, Y., & Zou, Q. 2019. Application of machine learning in
1253 microbiology. Frontiers in Microbiology 10.
1254 Rannala, B. 2015. The art and science of species delimitation. Current Zoology 61, 846–853.
1255 Rannala, B., & Yang, Z. 2003. Bayes estimation of species divergence times and ancestral
1256 population sizes using DNA sequences from multiple loci. Genetics 164, 1645–1656.
1257 2010. Bayesian species delimitation using multilocus sequence data.
1258 Proceedings of the National Academy of Sciences 107, 9264–9269.
1259 2020. Species Delimitation. In: Phylogenetics in the genomic era.
1260 Rannala, B., Edwards, S.V., Leaché, A., & Yang, Z. 2020. The Multi-species Coalescent Model
1261 and Species Tree Inference. Scornavacca, Celine; Delsuc, Frédéric; Galtier, Nicolas.
1262 Phylogenetics in the Genomic Era, No commercial publisher | Authors open access
1263 book.
1264 Raynal, L., Marin, J.M., Pudlo, P., Ribatet, M., Robert, C.P., & Estoup, A. 2019. ABC random
1265 forests for Bayesian parameter inference. Bioinformatics 35, 1720–1728.

1266 Rozantsev, A., Salzmann, M. & Fua, P. 2018. Beyond sharing weights for deep domain
1267 adaptation. IEEE transactions on pattern analysis and machine intelligence 41, 801–
1268 814.
1269 *Saryan, P., Gupta, S. & Gowda, V. 2020. Species complex delimitations in the genus
1270 *Hedychium*: A machine learning approach for cluster discovery. Applications in Plant
1271 Sciences 8. https://doi.org/10.1002/aps3.11377.
1272 Sanchez, T., Cury, J., Charpiat, G. & Jay, F. 2020. Deep learning for population size history
1273 inference: Design, comparison and combination with approximate Bayesian
1274 computation. Molecular Ecology Resources 21, 2645–2660.
1275 Scalon, M.C., Domingos, F. M.C.B., Cruz, W.J.A., Marimon-Júnior, B. H., Marimon, B.S., &
1276 Oliveras, I. 2020. Diversity of functional trade-offs enhances survival after fire in
1277 Neotropical savanna species. Journal of Vegetation Science, 31, 139-150.
1278 Schrider, D.R. & Kern, A.D. 2016. Discoal: flexible coalescent simulations with selection.
1279 Bioinformatics 32, 3839–3841. doi:10.1093/ bioinformatics/btw556.
1280 2018. Supervised Machine Learning for Population Genetics: A New Paradigm.
1281 Trends in Genetics 34, 301–312. https://doi.org/10.1016/j.tig.2017.12.005
1282 Searls, D.B. 2010. The Roots of Bioinformatics. PLoS Comput Biol 6.
1283 https://doi.org/10.1371/journal.pcbi.1000809.
1284 Sheehan, S., & Song, Y.S. 2016. Deep learning for population genetic inference. PLoS
1285 computational biology 12.
1286 Shurtliff, Q. R. 2013. Mammalian hybrid zones: a review. Mammal Review, 43, 1–21.
1287 Sidey-Gibbons, J.A., & Sidey-Gibbons, C.J. 2019. Machine learning in medicine: a practical
1288 introduction. BMC medical research methodology 19, 1–18.
1289 Silva, D.C., Oliveira, H.F.M., & Domingos, F.M.C.B. 2024. Cerrado bat community assembly is
1290 determined by both present-day and historical factors. Journal of Biogeography.
1291 Simonsen, K.L., Churchill, G.A., Aquadro, C.F. 1995. Properties of statistical tests of neutrality
1292 for DNA polymorphism data. Genetics 1411, 413–429.
1293 Sites, Jr J.W. & Marshall, J.C. 2004. Operational criteria for delimiting species. Annual Review
1294 of Ecology, Evolution, and Systematics, 199-227.

1295 Slatko, B.E., Gardner, A.F. & Ausubel, F.M. 2018. Overview of next-generation sequencing
1296 technologies. Current protocols in molecular biology 122.
1297 Smith, M.L., Ruffley, M., Espíndola, A., Tank, D.C., Sullivan, J. & Carstens, B.C. 2017.
1298 Demographic Model Selection using Random Forests and the Site Frequency
1299 Spectrum. Molecular Ecology.
1300 Smith, M.L. & Carstens B.C. 2020. Process-based species delimitation leads to identification of
1301 more biologically relevant species. Evolution 74, 216–229.
1302 https://doi.org/10.1111/evo.13878.
1303 Smith, M.L., & Hahn, M.W. 2023. Phylogenetic inference using generative adversarial networks.
1304 Bioinformatics, 39.
1305 Solis-Lemus, C., Yang, S., & Zepeda-Nunez, L. 2022. Accurate phylogenetic inference with a
1306 symmetry-preserving neural network model. arXiv preprint arXiv:2201.04663.
1307 Sukumaran, J. & Knowles, L.L. 2017. Multispecies coalescent delimits structure, not
1308 species. Proceedings of the National Academy of Sciences 114, 1607–1612.
1309 Sukumaran, J., Holder, M.T., & Knowles, L.L. 2021. Incorporating the speciation process into
1310 species delimitation. PLoS Computational Biology 17.
1311 Suvorov, A., Hochuli, J. & Schrider, D.R. 2020. Accurate inference of tree topologies from
1312 multiple sequence alignments using deep learning. Systematic biology 69, 221–233.
1313 Tagu, D., Colbourne, J.K. & Nègre, N. 2014. Genomic data integration for ecological and
1314 evolutionary traits in non-model organisms. BMC genomics 15, 1–16.
1315 Tautz, D., Arctander, P., Minelli, A., Thomas, R.H., Vogler, A.P. 2003. A plea for DNA
1316 taxonomy. Trends Ecol. Evol. 18, 70–74.
1317 Valletta, J.J., Torney, C., Kings, M., Thornton, A. & Madden J. 2017. Applications of machine
1318 learning in animal behaviour studies. Animal Behaviour 124, 203–220.
1319 Vink, C.J., Paquin, P., & Cruickshank, R.H. 2012. Taxonomy and irreproducible biological
1320 science. BioScience 62, 451–452.
1321 Vogler, A.P., Monaghan, M.T. 2007. Recent advances in DNA taxonomy. J. Zool. Syst. Evol.
1322 Res. 45, 1–10.
1323 Wake, D.B., Wake, M.H. & Specht C.D. 2011. Homoplasy: from detecting patterns to
1324 determining process and mechanism of evolution. Science 331, 1032–1035.

1325 Wäldchen, J. & Mäder, P. 2018. Machine learning for image-based species identification.
1326 Methods in Ecology and Evolution 9, 2216–2225.
1327 Wang, G. 2019. Machine learning for inferring animal behavior from location and movement
1328 data. Ecological informatics 49, 69–76.
1329 Wang, Z., Wang, J., Kourakos, M., Hoang, N., Lee, H.H., Mathieson, I., & Mathieson, S. 2021.
1330 Automatic inference of demographic parameters using generative adversarial
1331 networks. Molecular ecology resources 21, 2689–2705.
1332 Wiens, J. J., & Penkrot, T. A. 2002. Delimiting species using DNA and morphological variation
1333 and discordant species limits in spiny lizards (*Sceloporus*). Syst. Biol., 51, 69–91.
1334 Wiens, J. J. 2007. Species delimitation: new approaches for discovering diversity. Syst. Biol. 56,
1335 875–8.
1336 Wilkins, J. S., Zachos, F. E., & Pavlinov, I. Y. (Eds.). 2022. Species Problems and Beyond:
1337 Contemporary Issues in Philosophy and Practice. CRC Press.
1338 Yang, B., Zhang, Z., Yang, C.Q., Wang, Y., Orr, M.C., Wang, H., & Zhang, A.B. 2022.
1339 Identification of species by combining molecular and morphological data using
1340 convolutional neural networks. Systematic Biology, 71, 690–705.
1341 Yelmen, B. & Jay, F. 2023. An Overview of Deep Generative Models in Functional and
1342 Evolutionary Genomics. Annual Reviews of Biomedical Data Science.
1343 https://doi.org/10.1146/annurev-biodatasci-020722.
1344 Zachos, F. E. 2016. Species concepts in biology (Vol. 801). Cham: Springer.
1345 2018. (New) Species concepts, species delimitation and the inherent limitations
1346 of taxonomy. Journal of genetics, 97, 811–815.
1347 Zaharias, P., Grosshauser, M. & Warnow, T. 2022. Re-evaluating Deep Neural Networks for
1348 Phylogeny Estimation: The Issue of Taxon Sampling. Journal of Computational Biology
1349 29, 74–89. https://doi.org/10.1089/cmb.2021.0383.