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

Matheus M. A. Salles[1], Fabricius M. C. B. Domingos[1]

E-mails for correspondence: matheus.salles@ufpr.br; fabricius.domingos@ufpr.br

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

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, with most of them being rooted in Coalescent Theory. Their primary goal is to identify independently evolving lineages that should represent separate species. Coalescent models have improved species delimitation by enabling explicit testing of hypotheses regarding evolutionary independence among lineages. However, they have some limitations, especially regarding complex evolutionary scenarios, large datasets, and varying genetic data types. In this context, machine learning (ML) can be considered as a promising analytical tool, and clearly provides an effective way to explore dataset structures when species-level divergences are hypothesised. 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 researchers and students interested in the field. 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 traditional coalescent methods for species delimitation. For instance, there are clear limitations regarding the utilisation of simulated data, especially in supervised

and deep learning approaches, and the type of data representation used by each ML approach. We then discuss the strengths and weaknesses of existing pipelines, propose best practices for the use of ML methods in species delimitation, and offer insights into potential future applications. Generative adversarial networks and domain adaptation techniques, for instance, could be used to partially address the misspecification issue related to simulating genetic data. Besides, integrating ML methods into the hypothesis testing process, alongside available coalescent-based methods, could enable a more comprehensive exploration of evolutionary models and parameters, improving the accuracy and biological interpretability of species delimitation analyses. Additionally, we suggest guidelines for enhancing the accessibility, effectiveness, and objectivity of ML in species delimitation processes, aiming to offer a transformative perspective on this subject.

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

CONTENTS

I. Introduction

II. Machine learning

(1) Supervised Learning

(2) Unsupervised Learning

(3) Deep Learning

III. Current ML applications for species delimitation

IV. Advantages, limitations and future perspectives

V. Optimising the use of ML in the context of species delimitation

VI. Discussion

VII. Conclusions

VIII. Acknowledgements

IX. References

I. INTRODUCTION

Species represent fundamental entities across all biological disciplines. Consequently, the review, categorisation, and characterisation 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, characterising, and defining a species is both 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. Estimating the number of species in a particular biological system is challenging not only due to the great number of yet-undescribed species (Strain, 2011; Locey & Lennon, 2016), but also because species limits often lack clarity (see Rannala, 2015; Larsen *et al*., 2017; Rannala & Yang, 2020). Furthermore, some conceptual issues surrounding the definition of species concepts still attract debates among taxonomists and evolutionary biologists (Pante *et al*., 2015; Zachos, 2018).

Despite considerable empirical and theoretical progress, it is noteworthy that debates concerning species definition criteria remain prevalent today (de Queiroz, 2007; Saikia *et al.*, 2008; Sangster, 2013; Zachos, 2018). A multitude of operational criteria are employed to characterise species, whether they pertain to the particular species concept adopted within each empirical study or the delineation of species themselves (see de Queiroz, 2007). Interestingly, while a clear relationship exists between these components, namely the species concept and species delimitation, scientific endeavours have

historically focused on the former (see Sites Jr & Marshall, 2004; Wiens, 2007; de Queiroz, 2011; Hausdorf, 2011). Only within the past two decades has the field seen an increased emphasis on theoretical considerations related to species delimitation, accompanied by the introduction of new criteria and associated statistical methods (Lukhtanov, 2019; Rannala & Yang, 2020).

In practice, identifying species limits demands methods that precisely determine which individuals or populations should be assigned to existing species names and which entities constitute new species. Traditionally, species assignment and description, whether for recognised species or higher taxonomic categories (e.g., genus, family), have primarily relied on morphological characters (Rannala, 2015; Rannala & Yang, 2020), usually based on specific levels of morphological similarity to delineate species. However, this becomes especially problematic as morphological characters can exhibit significant plasticity and be influenced by environmental factors that do not necessarily reflect genetic or evolutionary relationships among lineages (Price *et al*., 2003; Wake *et al*., 2011; Jarvis *et al*., 2014).

Species delimitation and identification involve some degree of subjectivity, particularly in determining the levels of difference required for systematic and taxonomic classification. This time-consuming process demands high specialisation from researchers, and involves both delimiting evolutionary lineages and subsequently creating a formal diagnosis and nomenclature system (Jörger & Schrödl, 2013). Hence, semi-automated processes, in which experts primarily verify and refine results obtained from genomic data and computer algorithms, present an appealing alternative (Rannala & Yang, 2020). Typically, the delimitation process is initiated with a null hypothesis regarding the recognised species, and the evidence required to refute this hypothesis can sometimes be more substantial than that needed to justify a new species discovery (Sites

& Marshall, 2004; Camargo, 2013; Carstens *et al*., 2013). This emphasis on evidence is crucial as this process must be grounded in sound data, considering the population genetic structure, and integrating information from multiple sources to understand aspects like the species phylogenetic relationships, and the extent of hybridisation with closely-related lineages (Rannala & Yang, 2020).

Modern species delimitation methods (SDMs) aiming at identifying evolutionary units (Tautz *et al*., 2003; Vogler & Monaghan, 2007) are primarily based on the generalized species concept (de Queiroz, 1999; 2007), and mostly operate with molecular data under the principles of Coalescent Theory, notably, the multi-species coalescent (MSC; Rannala & Yang, 2003; Degnan & Rosenberg, 2009). Its use has grown due to advancements in statistical frameworks for phylogenetic inference (Edwards, 2009; O'Meara, 2012), along with Molecular Biology tools (e.g., next-generation sequencing (NGS); Slatko *et al*., 2018) and Bioinformatics (Searls, 2010). Nonetheless, researchers face many challenges when using SDM's in empirical systems, related to the vast amount of data generated by NGS platforms, and to inferential challenges Using SDMs with genetic data may fail to distinguish population structure from species-level divergence (Sukumaran & Knowles, 2017), and be affected by other issues associated with the reliance on the MSC model (Rannala & Yang, 2003; Degnan & Rosenberg, 2009; Edwards, 2009; Fujita *et al*., 2012). These can arise from conflicts among different gene trees, stemming from introgression events, **incomplete lineage sorting** (terms in bold are defined in the Glossary available in the Supplementary Material) and mixing between groups that constitute potential species (Rannala & Yang, 2010; Leaché *et al*., 2014; Jackson *et al*., 2017), and also from potential errors in phylogenetic inference (Carstens *et al*., 2013; Jacobs *et al*., 2018).

Consequently, some methods have their functionality limited to situations in which gene flow ceases immediately after population divergence, corresponding to an allopatric model of speciation (Fujita *et al*., 2012; Smith & Carstens, 2020). Simulations have also shown that ignoring gene flow leads the MSC to overestimate **population sizes** and underestimate divergence times (e.g. Leaché *et al.*, 2014). Hence, despite the clear usefulness of the MSC framework, its usefulness is still limited, to some extent, when additional processes influence divergence during speciation (Smith & Carstens, 2020). Different SDMs have varying capabilities to address each of these potential difficult scenarios (Camargo *et al*., 2012; Giarla *et al*., 2014; Luo *et al*., 2018). While some studies have examined the impact of various parameters (under various models of divergence, gene flow, and speciation) on species delimitation, further research should extend these comparisons to include MSC methods and alternatives based on different analytical frameworks (e.g. Camargo *et al*., 2012; Jackson *et al*., 2017; Luo *et al*., 2018). Besides, considering the intricate nature of the speciation process, it is unrealistic to anticipate that the use of specific models or metrics alone will result in error-free species delimitation (Burbrink & Ruane, 2021).

Machine learning (ML), a branch of artificial intelligence (AI) known for its computational efficiency and predictive accuracy, gained popularity mainly due to its ability to analyse and process large, complex, and high-dimensional datasets (Chicco, 2017; Borowiec *et al*., 2022; Fountain-Jones *et al*., 2021; Greener *et al*., 2021; Morimoto *et al*., 2021). Many ML algorithms are known to be extremely useful in various aspects of biology. This includes photo-based species identification (Wäldchen & 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 & Noble, 2015; Liu, 2019),

population genetics (Sheehan & Song 2016; Schrider & Kern, 2018), ecology (Christin *et al*., 2019; Pichler *et al*., 2020; Lürig *et al*., 2021), medicine (Sidey-Gibbons & Sidey-Gibbons, 2019), microbiology (Qu *et al*., 2019), and more (see Borowiec *et al*., 2022; Fountain-Jones *et al*., 2021; Morimoto *et al*., 2021). Therefore, its potential in evolutionary biology, and particularly in species delimitation, is evident. Specific examples can already be found in studies involving model selection in demography and phylogeography (Pudlo *et al.*, 2016; Fonseca *et al*., 2021), speciation (Blischak *et al*., 2021), phylogenetics (Suvorov *et al.*, 2020; C. Solis-Lemus, S. Yang, L. Zepeda-Nunez unpublished data; Smith & Hahn, 2023; Zaharias *et al*., 2022; Y.K. Mo, M. Hahn, M.L. Smith unpublished data), and species delimitation (Pei *et al*., 2018; Derkarabetian *et al*., 2019; Smith & Carstens, 2020; Pyron *et al.*, 2023), with the last one forming the primary focus of this review. In the following sections, we provide a brief overview of ML applications in the context of species delimitation.

II. MACHINE LEARNING

(1) Supervised learning

Supervised machine learning (SML) algorithms offer valuable solutions for statistical inference in diverse contexts. They enable predictions of new data points using a **training** set containing **labelled data** (often simulated) with known response variable values. This capacity is particularly significant in Evolutionary Biology, where obtaining large empirical datasets with high-confidence labels is challenging. Additionally, certain SML pipelines can effectively handle high-dimensional input data, mitigating issues related to the curse of dimensionality (see Schrider & Kern, 2018), unlike some coalescent or Bayesian approximation methods, which face increasing challenges when estimating

functions as the number of input variables rises. Nevertheless, while SML approaches have already transformed various fields, their application in phylogeographic and population genetic inference is relatively recent (e.g., Schrider & Kern, 2016; Sheehan & Song, 2016; Smith & Carstens, 2020; Fonseca *et al.*, 2021; Smith & Hahn, 2023).

In analytical terms, SML involves using a dataset comprising predictor variables (input) and response variables (output) to establish and predict the relationship between them. Formally, SML methods employ a function, denoted here as *f*, to predict a response variable, *y*, based on a **feature vector**, *x*, containing *n* input variables. This relationship is expressed as $y = f(x)$ within a typical analytical framework. When *y* represents a categorical variable, such as a specific evolutionary scenario, it constitutes a **classification problem**. Conversely, if *y* is continuous, the task becomes a **regression problem**, applicable, for instance, in estimating population genetic parameters. In supervised learning, the objective is to optimize $y = f(x)$ using a labelled training set, where response variable values are known. Besides, the dataset comprises values from a feature vector, which is a multidimensional representation of any point in the initial dataset or features extracted from it.

A workflow for applying any SML method to population genetic data comprises multiple steps, especially in the genomic scale. These include data simulation for various evolutionary scenarios, encoding both simulated and observed genetic data into feature vectors, training the algorithm, applying it to new observed data points, and assessing its predictive performance through error calculations and accuracy estimates (Fig. 1). Crucially, the use of simulated genetic data based on known evolutionary models is essential, given the scarcity of adequately sized datasets with high-confidence labels. However, it introduces concerns, primarily related to potential inaccuracies in **model** specifications (Schrider & Kern, 2018; Callier, 2022). Essentially, the entire process of

training and applying ML algorithms is influenced by the assumptions made about the underlying evolutionary processes, such as population size, selection strength, and gene flow. Consequently, the reliability of results obtained from SML methods hinges on the resemblance between the training data (typically simulated) and the real biological data used for posterior inferences.

Numerous SML algorithms are efficient in classification or regression tasks when it comes to Evolutionary Biology (refer to Schrider & Kern, 2018 and Greener *et al*., 2021 for reviews). Generally, the initial step in SML analysis is designing a training set (Fig. 1A). This training set can consist of data simulated across various scenarios, with parameter values drawn from prior distributions (Fig. 1B). At this point, it is important to consider that such data may not always be readily available, as some scenarios cannot be efficiently simulated, or may lack certain desired characteristics necessary for training and analysing specific evolutionary models. Besides, depending on the research goals, the simulated data can be characterised using a set of summary statistics or represented in another relevant biological format (Fig. 1C). In the context of Evolutionary Biology, this is particularly crucial given the challenges of acquiring high-quality data for testing complex hypotheses. Besides, one should also consider the need of acquiring a fine balance during the training phase of an SML algorithm, between achieving accuracy through the trained model and ensuring the model's ability to generalize its learning when faced with a **test set** or new empirical data (see Korfmann *et al*., 2023).

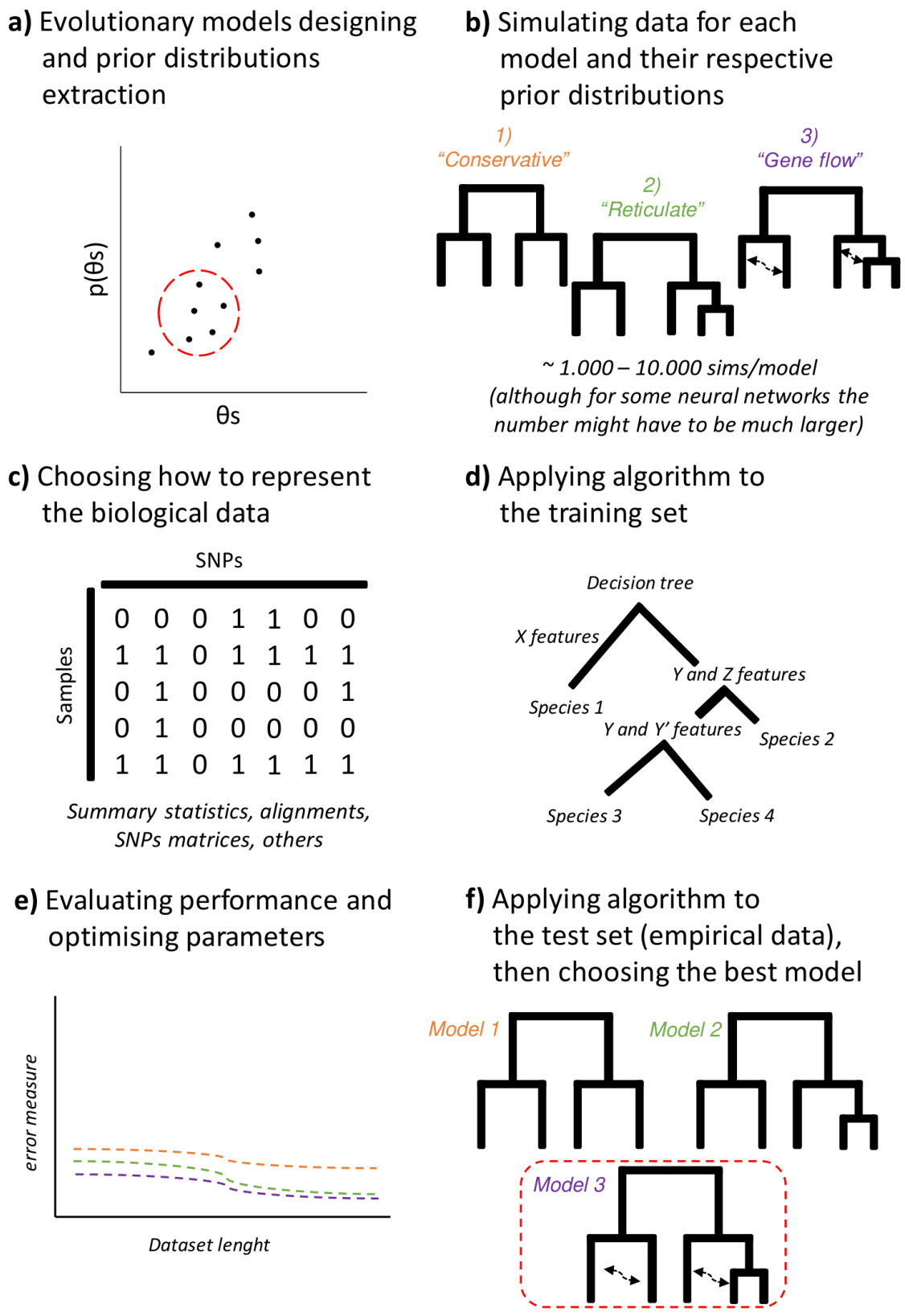

Fig. 1 - Diagram illustrating a potential SML workflow for species delimitation, inspired by the work of Smith & Carstens (2020). a) The initial step involves designing priors for the evolutionary models considered in the study. b) Simulated data is generated for each model, typically ranging from 1,000 to 10,000 simulations per model, using relevant simulation software. c) The data is represented according to

the requirements of the chosen ML tool. d) Following data simulation and representation, ML model training begins, involving various preliminary steps like data pre-processing, dataset division, feature selection, and algorithm choice. e) Model performance (both in terms of biological accuracy and computationally) is assessed using statistical metrics, allowing for retraining and adjustment based on the results. f) Once the model is adequately trained and evaluated, it can be used to predict species categories for new data, which can be either newly simulated data or empirical data consistent with the model's proposal, determining how many species exist in that particular biological system.

Supervised algorithms offer the advantage of extracting maximum information from a set of diverse metrics (e.g., various summary statistics) within the feature vector. This eliminates the need for arbitrary subset selection, a practice often employed in methods like Approximate Bayesian computation (ABC; Collin *et al.*, 2021). Consequently, SML mitigates issues related to inference accuracy, as its performance remains stable even when the number of variables increases, contrasting with traditional methods such as ABC (Raynal *et al*., 2019; Collin *et al*., 2021). This is partially due to the ability of SML's algorithms to utilise all simulations during the training phase, which enables the mapping of an entire dataset regarding different scenarios and parameters (Collin *et al.*, 2021). Additionally, it's essential to note that SML algorithms are highly effective in handling large, intricate datasets, as many of them can create a high-dimensional hyperplane to differentiate between various classes across multiple features. Consequently, adding extra features is unlikely to cause analytical issues.

Moreover, SML algorithms demand a significantly smaller training set compared to other methods, 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). Given the growing dimensionality of genetic data from NGS technologies, SML methods have emerged as a suitable choice for researchers seeking to analyse complex scenarios and large datasets, especially in the context of selecting evolutionary scenarios and

demographic estimates. These characteristics underscore SML's enormous potential to revolutionise genetic data analysis in the near future.

(2) Unsupervised learning

SML algorithms require some initial human intervention for proper sample labelling and perform well in several dataset scenarios, as they do not necessitate a large labelled training set for achieving reliable results (Libbrecht & Noble, 2015; Shen *et al*., 2022). Semi-supervised learning (SEMI-ML) is another approach used when only part of the input data is labelled, and has proved advantageous in situations where labelling data is challenging, either due to the time required for labelling or uncertainties associated with assigning labels. Even so, while SML and SEMI-ML approaches are powerful and widely applicable, there are situations where unsupervised machine learning (UML) becomes a more viable option. Unlike SML and SEMI-ML, UML relies solely on the inherent data structure to group samples. Consequently, UML algorithms operate without predefined assumptions about the data's underlying structure, population parameters, species numbers, or sample categorisation, making them particularly suitable for species delimitation where no prior hypotheses are put forward about these aspects.

UML algorithms generally fall into three problem categories: clustering, association, and dimensionality reduction (Hastie et al., 2009; Libbrecht & Noble, 2015; Dike *et al.*, 2018). Clustering methods group input data into subsets or clusters, where samples with high similarities are placed in the same cluster and exhibit less or no similarity with samples in other clusters. Conversely, association algorithms uncover relationships between variables within the dataset by employing various metrics to assess interdependencies among variables, effectively partitioning them into groups based on meaningful associations. Dimensionality reduction techniques focus on compressing data

to identify a smaller, distinct set of variables that could capture essential features of the original data while minimising information loss. When combined with clustering approaches, UML dimensionality reduction may provide intuitive data visualisation and accommodate various data types (Libbrecht & Noble, 2015). In sum, UML is another promising option for species delimitation, as such algorithms enable the simultaneous use of diverse data types, extracting and condensing the necessary information to try to identify the limits of biological groups.

However, because UML methods do not rely on preconceived notions on the nature of the data, researchers using it for species delimitation must ensure that the analyses are effectively operating at the species level (Derkarabetian *et al.*, 2019). Either way, in species delimitation practices, UML dimensionality reduction algorithms are generally employed (Fig. 2), having demonstrated effectiveness in cases where coalescent methods tend to split potential species too narrowly, particularly when there is species-level divergence but not significant population structure (Derkarabetian *et al.*, 2019). Also, as mentioned before, UML approaches might even be able to accommodate diverse data types commonly found in integrative taxonomic studies, including genetic, morphometric, continuous, and categorical data (Pyron, 2023; Pyron *et al*., 2023).

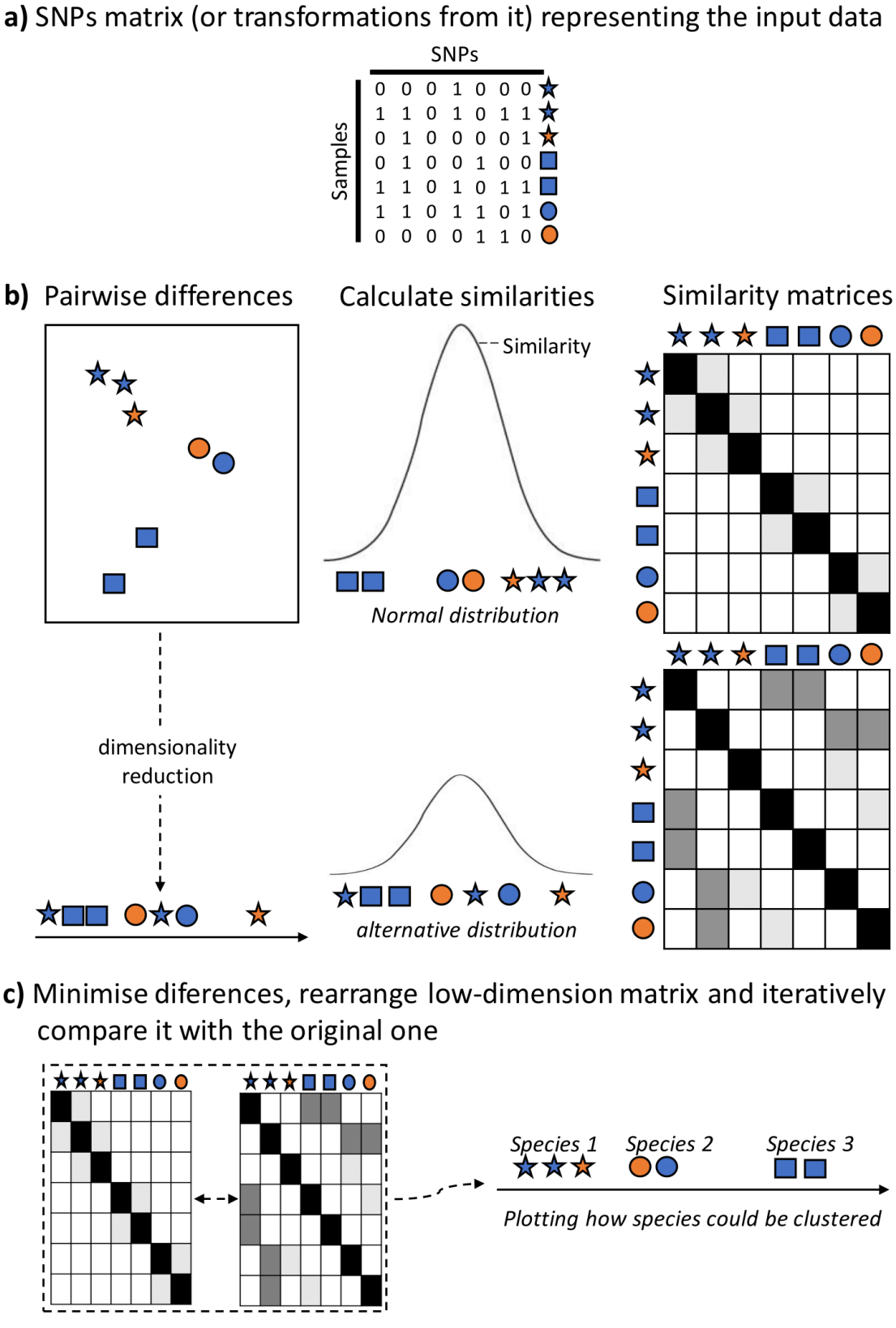

Fig. 2 - Diagram outlining a potential UML workflow for species delimitation, utilizing the t-SNE algorithm (inspired by Derkarabetian *et al*., 2019). a) Data representation is the initial step, and it varies depending on the chosen ML tool, which may work with sequence data, SNP matrices, or population genetics metrics extracted from them. b) t-SNE, as a dimensionality reduction technique, iteratively finds a lower-

dimensional representation of the original data. It identifies local similarity spaces between sample pairs by analysing 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 neighbours 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).

(3) Deep learning

Artificial neural networks (ANNs) are increasingly employed in Evolutionary Biology, often referred to as '**deep learning**' (Sheehan & Song, 2016). Deep learning techniques have found success in various fields in Biological Sciences (Angermueller *et al*., 2016; Sheehan & Song, 2016; Kamilaris & Prenafeta-Boldú, 2018; Mobadersany *et al.*, 2018; Schrider & Kern, 2018). However, its adoption in Evolutionary Biology is relatively recent (see Blischak *et al*., 2021; Yelmen & Jay, 2023). The recent popularity of ANNs can be mostly attributed to their highly flexible data input and output structure, allowing networks trained for one task to be repurposed for another by modifying their final **layers**. For instance, a network originally trained for inferring population size history can theoretically be adapted to identify optimal population genetics parameters within various demographic scenarios. **Transfer learning** approaches, for example, can be useful when limited training data are available from a new domain, with reduced computational expenses compared to training an algorithm from scratch. Additionally, the knowledge acquired during the initial task could improve the new network, reducing errors and enhancing learning efficiency (Sanchez *et al.*, 2021). Also, ANNs possess a unique capability to establish parameterised functions that facilitate non-linear mappings from one parameter space to another. This versatility enables the resolution of intricate tasks that might prove challenging for **shallow learning** algorithms.

Similar to traditional machine learning, neural networks are trained by adjusting their parameters using a training set, typically composed of pairs of known or simulated inputs and desired outputs. Optimisation relies on minimizing the value of a **loss function** that gauges the degree of error in the network's performance based on the current parameters. Parameter adjustments are executed through an optimisation algorithm driven by **gradient descent** and **backpropagation**. This process typically necessitates a substantial volume of training data to ensure effective learning and generalisation, enabling the network to perform well when faced with previously unseen data (Sanchez *et al*., 2021). In essence, a deep learning algorithm aims to project a function, embodied as a neural network, which can be conceptualised as a differentiable computational graph organised into a series of stacked linear and non-linear layers (Angermueller *et al*., 2016; Sanchez *et al*., 2021). These layers are replete with numerous trainable parameters, from thousands to trillions, depending on the case. Within the neural network, each layer receives inputs from the preceding layer(s), causing every node in the layer to execute a linear combination of these inputs. This is succeeded by a non-linear transformation (the **activation function**), culminating in the calculated value being forwarded to the subsequent layer. ANNs can vary in their **architecture**, encompassing the number of layers and nodes, as well as the connections between nodes. In light of this context, the design of the neural network architecture holds paramount importance when employing deep learning techniques. A suboptimal design can result in reduced inferential capabilities, information loss, issues like **underfitting** or **overfitting**, and unnecessary complexities, all of which can negatively influence the training process (Cartwright, 2008; Angermueller *et al*., 2016; Sanchez *et al*., 2021).

Conversely, deep learning methods come with their share of intricacies and often demand meticulous and more specific fine-tuning compared to shallow learning methods.

This includes defining the number of layers in the neural network, configuring network **hyperparameters**, and exploring the control parameters of the loss function. In this regard, it is reasonable to assert that simpler machine learning algorithms remain competitive, especially when detailed parameter adjustment is unfeasible or unwarranted. This is particularly applicable in scenarios where an extensive volume of data or variables is not necessary to study a particular phenomenon, favouring the simplicity of shallow learning over the inherent complexity that neural networks typically entail. Nevertheless, the fundamental stages involved in creating a supervised shallow learning framework for species delimitation can be broadly paralleled with the primary phases found in a deep learning workflow. These encompass data simulation and representation, model training and optimisation, all the way to predicting the relevant categories from empirical data (Fig. 3).

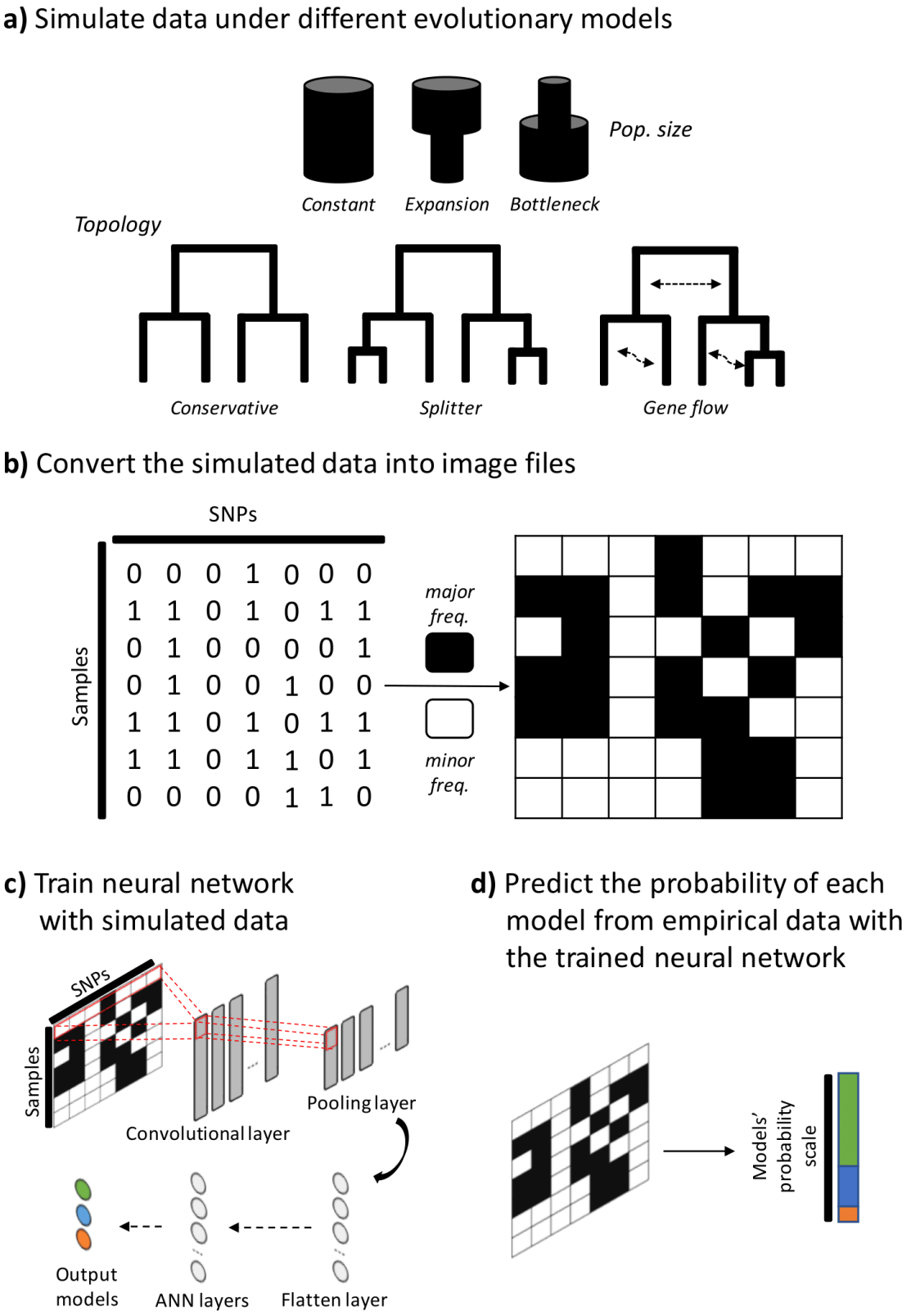

Fig. 3 – Diagram illustrating a potential deep learning workflow applied in the context of species delimitation, using CNNs (inspired by Perez *et al.*, 2021). a) The process typically begins with the simulation of biological data under various evolutionary models, considering factors like topology, population size, gene flow, and more, similar to SML. b) Next, data representation is crucial. For CNNs,

SNP matrices are often converted into arrays or image files, where pixel contrast reflects differences in minor and major frequencies between samples. c) With the simulated and properly represented data, the network training phase can commence. The parameter configuration and network architecture may vary, depending on the specific study's requirements. d) Once each model is trained and its performance is rigorously evaluated, the final stage of the workflow involves predicting categories for new data. This can include using new simulated data with slight parametric modifications, still within the trained model's limits, as well as empirical data whose evolutionary history aligns with the proposed model. In both cases, the goal is to determine which delimitation model best applies to the biological system being investigated.

III. CURRENT ML APPLICATIONS FOR SPECIES DELIMITATION

In the same way that there are two primary categories of ML (excluding deep learning), species delimitation methods can also be broadly categorised 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) (Flouri *et al.*, 2018; Sukumaran *et al.*, 2021). This draws a conceptual parallel between traditional discovery approaches and UML methods, as well as between validation methods and supervised algorithms (Fig. 4). In practice, UML delimitation approaches typically use clustering or dimensionality reduction techniques (e.g. Derkarabetian *et al.*, 2019), while SML approaches often involve using simulated datasets to train a classifier, which is then used to label new datasets accurately.

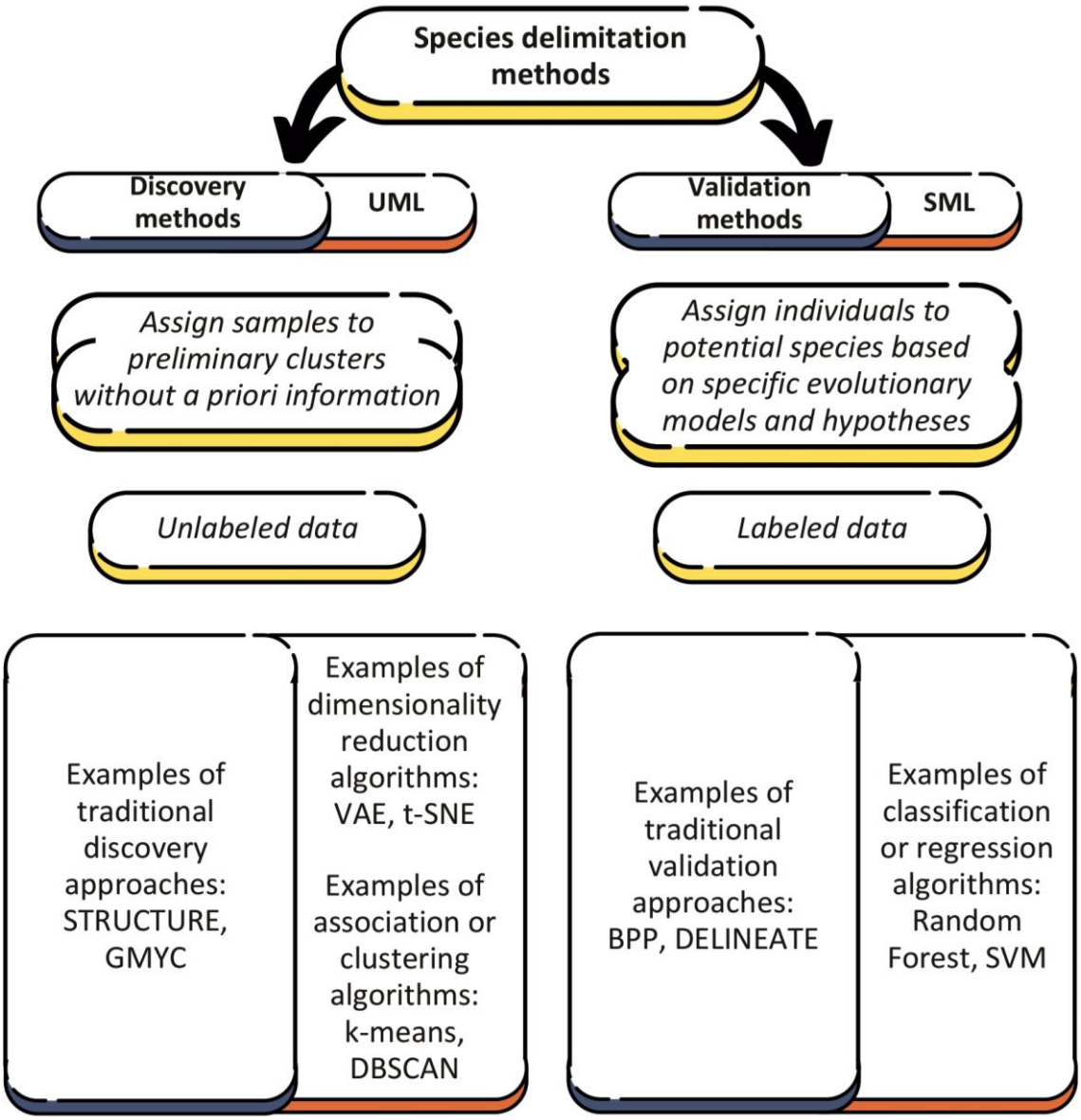

Fig. 4 - Comparative diagram categorising species delimitation methods and machine learning algorithms, along with some of their key characteristics. Species delimitation methods can be broadly categorised as discovery and validation methods, akin to unsupervised and supervised machine learning algorithms, respectively.

Below, we present a comprehensive overview of recently applied ML methods in the domain of species delimitation, emphasising 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 prioritised research projects that either introduced novel methodologies

(see Table 1) or enhanced and tested existing techniques in this context (Supplementary Material). 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 analysing DNA sequence data. The categorised methods include SML, UML, and deep learning. Also, there are some studies utilizing ML techniques and 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 the Supplementary Material.

429 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 (Pei et al., 2018)[1]	python	SML	Support vector machines	MCcoal	Multiple sequence alignment (MSA) or SNP matrix	Population genetics summary statistics
A demonstration of unsupervised machine learning in species delimitation (Derkarabetian et al., 2019)[2]	R/python	UML	Variational autoencoders and t-Distributed Stochastic Neighbour Embedding	NA	SNP data matrix	One-hot-encoding of the SNP data matrix and *axis* from a discriminant analysis of principal components
delimitR (Smith & Carstens, 2020)[3]	python	SML	Random forest	fastsimcoal	SNP data matrix	Folded multi-dimensional SFS
Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system (Perez et al., 2021)[4]	python	*Deep learning*	Convolutional neural networks	ms	SNP data matrix	NumPy 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

430 Online repositories where it is possible to find more information about the currently existing platforms. [1] https://github.com/pjweggy/CLADES;
431 [2] https://www.sciencedirect.com/science/article/abs/pii/S1055790319301721; [3] https://github.com/meganlsmith/delimitR;
432 [4] https://github.com/manolofperez/CNN_spDelimitation_Piloso; [5]https://github.com/kyleaoconnell22/Pyron_et_al_UML_sp_delim/tree/main

Derkarabetian *et al.* (2019) conducted a study to assess the performance of UML and deep learning methods in the context of species delimitation. Their research highlighted the effectiveness of variational autoencoder (VAE) and t-Distributed Stochastic Neighbour Embedding (t-SNE) algorithms in particular scenarios for accurately identifying species clusters and estimating the correct number of species. In the case of VAE, SNP matrices were converted via 'one-hot coding,' where nucleotides were transformed into binary variables (e.g., A = [1, 0, 0, 0]; C = [0, 1, 0, 0], and so on), including ambiguous bases (e.g., M = [0.5, 0.5, 0.0, 0.0]). A custom script was developed to perform this transformation (Derkarabetian et al., 2019). This VAE approach employed multiple layers of encoding to compress high-dimensional input data, followed by the reconstruction of data through successive decoding layers. The latent variables, represented as a normal distribution with mean (μ) and standard deviation (σ), offered a two-dimensional depiction of the SNP matrix, facilitating a clear visualisation that accounted for the uncertainty surrounding groupings due to standard deviations among samples and groups. In the case of t-SNE, data derived from a discriminant analysis of principal components (DAPC) was used as input variables, preceded by clustering tests using SNP matrices. Both approaches yielded more readily interpretable outcomes compared to other methods assessed by the authors, revealing distinct species groupings in a two-dimensional space (Derkarabetian *et al*., 2019). Notably, the identified groupings in this study aligned with the species delimitation results achieved through an integrative taxonomy approach, demonstrating a high degree of concordance between the datasets, suggesting that the limits identified by UML algorithms indeed correspond to species-level divergence rather than population structure (Derkarabetian *et al*., 2019).

Smith & Carstens (2020) introduced delimitR, a SML approach designed to frame species delimitation as a model selection challenge; delimitR employs the

multidimensional **site frequency spectrum** (mSFS) with a **binning** strategy as a predictor variable for a **Random Forest (RF)** classifier. Working with data summarised through the mSFS, delimitR facilitates the evaluation of models that vary in terms of lineage numbers. In essence, this framework aims to discriminate between various divergence models compatible with virtually any species concept, as asserted by the authors. Given its supervised nature, delimitR demands researchers to define reasonable priors, such as divergence times or migration rates, and to make decisions about the inclusion of models within the set (Smith & Carstens, 2020). Moreover, delimitR offers users the flexibility to customize the parameter space by incorporating custom models generated using fastsimcoal (Excoffier *et al*., 2021) and integrating them into the R workflow.

In the context of Smith & Carstens' (2020) study, each model was used to simulate 10,000 mSFS. These sets of simulated mSFS were subsequently summarised and observed by binning into four classes per population. A RF classifier was constructed using 1,000 **decision trees** to accommodate the extensive number of models. delimitR's performance demonstrates an improvement with larger SNP matrices and increasing divergence times. Also, compared to traditional ABC methods, the RF approach implemented in delimitR demonstrates lower error rates, even though the detection of migration becomes more challenging in cases of recent divergence between lineages (Smith & Carstens, 2020). However, the authors acknowledge that further research is needed to elucidate the association between the model space, number of parameters, and delimitation accuracy. Also, one of the fundamental principles underlying delimitR is the encouragement of explicit predictions based on hypotheses. This approach necessitates researchers to articulate the species concept employed in their data analysis, enhancing

transparency and repeatability in species delimitation studies, by connecting biological units with the evolutionary processes that gave rise to them.

CLADES (Pei *et al*., 2018) is another SML approach designed for species delimitation, utilizing classification models trained and evaluated on multilocus sequence data. Notably, this study introduced the application of the **support vector machines (SVM)** algorithm to species delimitation. For model training, genetic datasets at the population level were simulated, with and without gene flow (although only the dataset without gene flow was shared by the authors). To manage computational complexity, five summary statistics were employed instead of raw sequence data. Also, in contrast to other existing methods, CLADES eliminate the need for users to supply guide trees or priors related to divergence time and population size (Pei *et al*., 2018). Within this framework, species delimitation is framed as a classification task, where the goal is to classify pairs of populations as either belonging to the same species or different species, using new observations based on training data derived from simulations conducted by the authors, across various evolutionary scenarios. Each training sample was represented as a list of summary statistics, and a SVM regression is calculated using these statistics. Through iterative training, the classification weights for each statistic were adjusted to minimise the misclassification cost. Being a SVM, it is assumed that the training data fell within a standard range, so all summary statistics were normalised to a range between 0 and 1. Subsequently, the SVM classifier computed the probability of the training samples belonging to each potential grouping.

To create the training dataset, the authors conducted simulations based on a two-species model (A and B) where both species diverged at time τ with identical population size parameters ($\theta A = \theta B = \theta$). Each species further consisted of two populations that recently split at time τp. Migration between species A and B was allowed at a rate of M

= Nm migrants per generation, with m representing the migration rate per generation. The MCcoal software (Rannala & Yang, 2003) was used to simulate multilocus sequence data of length L under various parameter combinations for training. For each possible parameter combination (θ, τ, M), sequences were simulated at 100 loci with a length of L = 100Kbp for all populations. For each locus, 40 sequences were sampled, with 10 sequences per population. Additionally, symmetrical migration between species A and B was assumed before the populations of the species split at time τp. Following this parameter configuration, a classifier was trained on the simulated data for each parameter configuration, with **cross-validation** used to assess accuracy. Subsequently, all training samples were combined to train a global classifier, enabling it to adapt to various values of θ and M and not assume fixed parameters. Regarding its performance, Pei *et al.* (2018) demonstrated that longer loci sequences improved CLADES' efficiency. Moreover, CLADES exhibited robustness to different modelling structures because it can accommodate various demographic events and evolutionary parameters, and achieved reasonable delimitation results even in the presence of gene flow (Pei *et al*., 2018).

Perez *et al.* (2021) propose a species delimitation approach that combines coalescent-based methods with model selection using CNNs. The initial step involves simulating genetic data for each delimitation hypothesis, with the study encompassing 10,000 simulations per model (the alignments of empirical and simulated data are accessible on GitHub). Subsequently, the simulated data is transformed into images, where black pixels represent alleles with the highest frequency, and white pixels represent those with the lowest frequency at each segregating site. These images of simulated data are used to train a neural network capable of recognising simulations generated from each model. The network can predict the associated probability using CNNs when tested with new empirical data. In the same study, the authors conducted a comparison between their

model selection approach and ABC to assess various species delimitation hypotheses within the *Pilosocereus aurisetus* cactus species group. To validate the new method, they also employed a previously published dataset consisting of two pairs of *Drosophila* species. It's worth noting that while CNNs used 10,000 simulations per model, ABC required 100,000 simulations per model. The CNNs consistently demonstrated superior performance in distinguishing between the simulated demographic scenarios, outperforming ABC in all cases, with fewer simulations and faster execution times (Perez *et al*., 2021).

Pyron *et al.* (2023) introduced a novel UML approach designed for delineating species limits from extensive genomic datasets. Their method is primarily grounded in **self-organizing maps (SOMs)**, which aim to arrange multidimensional data into a two-dimensional configuration to maximise similarity between the input data's distance matrix and the output data. Notably, it produces discrete outcomes rather than continuous ones, as it groups genotypes based on shared descent or state. This approach is posited as more advantageous than prior workflows, such as those presented by Derkarabetian *et al*. (2019). 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 genotypes, have been effectively mapped from the original SNP matrix. Subsequently, the method estimates the cumulative distances from each sample to its immediate neighbours. Notably, these distances should show an increase near to **class** limits, which correspond to the demarcation between different candidate species. To effectively separate these candidate species, Pyron *et al.* (2023) recommend performing cluster analyses, such as k-means. The determination of the optimal number of classes or species in the dataset is achieved by selecting the value that maximises the sequential reduction in the weighted sum of squares from k to k + 1. Also,

we highlight that this technique is rooted in the assessment of similarity rather than dissimilarity. Besides, recently an extension of this method has been proposed in the form of a SuperSOM approach, incorporating the possibility of utilising several trait classes simultaneously, such as alleles, morphological and ecological variables (Pyron, 2023).

IV. ADVANTAGES, LIMITATIONS AND FUTURE PERSPECTIVES

In general, it is reasonable to assert that the ML methods applied to infer species limits offer some advantages over coalescent or traditional Bayesian computation methods. Despite some constraints, ML algorithms perform as well as or even outperform model selection methods like ABC and coalescent-based methods (Pei *et al*., 2018; Smith & Carstens, 2020; Perez *et al*., 2021; Derkarabetian *et al*., 2021). Moreover, 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 & Carstens, 2020; Suvorov *et al*., 2020; Martin *et al*., 2021; Perez *et al*., 2021). Some of these algorithms also have proven to be highly efficient in complex evolutionary scenarios, including situations involving gene flow (Pei *et al*., 2018; Perez *et al*., 2021).

It is reasonable to anticipate that the introduction of new ML approaches for species delimitation will increasingly enhance researchers' ability to make biologically precise decisions particularly when these methods are purpose-built, from conception to implementation, for the specific task of delimiting evolutionary lineages. As a consequence, a critical step in any study at the intersection of ML approaches and species delimitation methods involves selecting the methods to be employed. This decision can be quite challenging due to the broad array of coalescent-based and ML methods available in the modern Evolutionary Biology toolkit (Schrider & Kern, 2018; Smith & Carstens, 2020; Greener *et al*., 2021; Yelmen & Jay, 2023). With this multitude of possibilities, the

582 ideal choice should not only consider an appropriate fit with the biological problem under
583 investigation, but also a statistical evaluation and performance optimisation (Greener *et*
584 *al*., 2021; Morimoto *et al*., 2021), under various diversification scenarios, while
585 estimating historical parameters like divergence time, population size, and **migration**
586 rates.

587 In this regard, one primary advantage of ML approaches over some formal
588 Bayesian or maximum likelihood methods is their efficiency in testing complex
589 demographic models, including scenarios with migration events or population size
590 fluctuations (Perez *et al*., 2021). This efficiency does not compromise the ability to
591 distinguish between different models (Smith *et al.*, 2017). Even simple SML methods
592 provide high selection accuracy when comparing multiple models in a single analysis (M.
593 Gehara, G.G. Mazzochinni, F. Burbrink, unpublished data). In sum, different empirical
594 studies using simulated data have demonstrated that ML algorithms can perform at least
595 as effectively as coalescent-based species delimitation methods and, in certain scenarios,
596 they can be more efficient in delineating species limits, especially when lineages continue
597 to exhibit gene flow. Additionally, studies have indicated that deep learning methods,
598 such as **convolutional neural networks (CNNs)**, show promise as effective tools for
599 model selection in evolutionary biology (Fonseca *et al*., 2021), being applicable even in
600 complex evolutionary scenarios involving hidden genetic diversity, gene flow between
601 populations, and changes in effective population size over time. Thus, even when ML
602 methods such as these are not designed as delimitations approaches *per se*, they can
603 function as one depending on its application, for instance in a transfer learning approach.

604 However, it is essential to consider that certain algorithms, especially those in
605 SML or deep learning, can be overly specialised. Modern ML methods are proficient at
606 interpolating within the observed range of values in the training data, even in cases where

specific values haven't been encountered before, being adaptive and not solely reliant on memorising specific training instances. Even so, because such algorithms are typically trained on simulated data with specific values of evolutionary parameters, such as θ and M, their performance might be compromised when applied far outside the training parameter space (Schrider & Kern, 2018; Borowiec *et al*., 2022). Besides, ML algorithms such as those used in the studies described in the previous section do exhibit some degree of **inductive bias**, leading to potential inaccuracies in this context (Hüllermeier *et al*., 2013). Therefore, exploring in further details the association between training capacity and predictive power should be a priority for future studies.

Machine Learning is certainly becoming more prevalent in Evolutionary Biology due to its extensive use of simulated data for training classification and regression models (Yuan *et al*., 2012; Yelmen & Jay, 2022; Korfmann *et al.*, 2023), as modern computer simulators can efficiently generate substantial amounts of labeled data in diverse evolutionary scenarios (Haller & Messer, 2019; Baumdicker *et al.*, 2021). Methods relying on a substantial volume of simulated data across diverse evolutionary scenarios need to consider the careful design of prior distributions to simulate models that closely resemble the real biological system under investigation. However, this model specialisation might yield models that lack generalisability and transferability across different studies or data types, an area warranting further empirical exploration (Schrider & Kern, 2018; Borowiec *et al*., 2021). This challenge becomes more pronounced for non-model organisms, where data availability may severely limit the quality of parameter estimates (Tagu *et al*., 2014; Fonseca *et al*., 2016; Cerca *et al*., 2021; Jorna *et al.*, 2021).

Furthermore, it may be unfeasible to simulate data or train an ML algorithm across an entire parameter space, especially in complex evolutionary models (Rannala & Yang, 2020). Limited information is available regarding the asymptotic statistical performance

of most ML methods applied for species delimitation, and important phenomena may be entirely missing from the simulations (e.g. background selection, Mo & Siepel (2023), or missing data Arnab *et al*. (2023)). This leads to an inherent challenge in avoiding some degree of misspecification in the training data, even considering the variety of powerful genetic data simulators currently available, such as SLiM (Messer, 2013), discoal (Kern & Schrider, 2016), msprime (Baumdicker *et al*., 2021), and fastsimcoal2 (Excoffier *et al*., 2021). In the context of species delimitation, formal statistical methods based on coalescence still offer the means to address such issues. These methods possess optimality and iterability properties that span a reasonable portion of the parameter space, albeit at a considerable computational cost (e.g., Flouri *et al.*, 2018; Sukumaran *et al.*, 2021).

Regarding ML itself, one approach to mitigate the effects of misspecification during simulation involves designing or using a simulator that enforces greater compatibility between simulated and actual data. Generative adversarial networks (GANs), a type of deep learning algorithm commonly used for creating synthetic images and voices (Chadha *et al*., 202), have shown promise in this regard (see Callier, 2022; Wang *et al.*, 2021). GANs operate with two networks, the generator and the discriminator, trained together (Goodfellow *et al*., 2014). While the generator generates simulated data, the discriminator distinguishes between real and fake data. Over the course of training, the generator network becomes more adept at producing realistic **examples**, and the discriminator network becomes more skilled at distinguishing between real and synthetic data. Once training is complete, the generator network can be utilised to generate new examples that are indistinguishable from real data, providing a reliable way to work with labelled data where ground truth is known. Researchers have already assessed the utility of GANs in various fields, including genomics, phylogenetics, and population genetics (Booker *et al.*, 2023; L. Nesterenko, B. Boussau, L. Jacob unpublished data; Yelmen &

Jay, 2023). Smith & Hahn (2023), for instance, introduced phyloGAN, a workflow that takes a concatenated alignment (or a set of alignments) as input and infers a phylogenetic tree, potentially accounting for gene tree heterogeneity.

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

Besides, some researchers argue that issues related to potential errors in data simulation can be likened to a "domain adaptation" problem, where a model trained on one data distribution is applied to a dataset originating from a different distribution (Farahani *et al.*, 2021; Mo & Siepel, 2023). Such problems often arise in scenarios involving extensive and diverse datasets, where generating adequately representative labelled training examples can be challenging. A classic illustration of domain adaptation is found in image classification. Consider a situation in which a recognition model needs to identify different dog breeds from photographs ("target domain"), but there is an abundance of labelled training data available only in cartoon drawings of dogs ("source

682 domain"). In such cases, a ML model must be trained on one dataset with the expectation
683 of performing well on another, even in the presence of systematic differences between
684 the two distributions.

685 Domain adaptation techniques encompass a broad array of methods historically
686 prevalent in fields like computer vision and natural language processing (Li 2012; Xu *et*
687 *al.*, 2019; Farahani *et al*., 2021). Recent approaches typically involve learning a "domain-
688 invariant" data representation through a feature extractor neural network. This is
689 accomplished by minimising domain disparities (Rozantsev *et al*., 2018), utilizing
690 adversarial networks (Ganin & Lempitsky, 2015; Liu & Tuzel, 2016; Bousmalis *et al*.,
691 2017), or employing auxiliary reconstruction tasks (Ghifary *et al*., 2016). It is noteworthy
692 that domain adaptation techniques have found applications in fields such as genomics
693 (Cochran *et al*., 2022) and population genetics (Mo & Siepel, 2023), particularly as an
694 unsupervised domain adaptation problem. In this context, initial simulations generated
695 substantial amounts of meticulously labelled training data in the source domain;
696 subsequently, the trained model was deployed on unlabelled real data in the target domain
697 to explicitly consider the disparities between these domains during model training.
698 Through extensive simulation studies, Mo & Siepel (2023) convincingly demonstrated
699 that their domain-adapted models significantly outperformed standard networks across
700 various simulation misspecification scenarios. This outcome underscores the potential of
701 domain adaptation techniques as a promising avenue for developing more robust deep
702 learning models in the realm of population genetic inference (Mo & Siepel, 2023),
703 potentially including species delimitation.

704 Another crucial perspective to consider is that numerous studies, whether focusing
705 on species delimitation, population demography, or genetics, incorporate ML for
706 inferences based on summary statistics (Pei *et al.*, 2018; Smith & Carstens, 2020; Collin

et al., 2021; Ghirotto *et al*., 2021). Furthermore, there are methodologies tailored for handling data derived from SNP matrices (Derkarabetian *et al*., 2019; Sanchez *et al*., 2020; Smith & Carstens, 2020; Blischak *et al*., 2021; Fonseca *et al*., 2021; Martin *et al*., 2021; Perez *et al*., 2021) or raw sequence data (Pei *et al*., 2018; Ghirotto *et al*., 2021), and only a few pipelines offer extensibility to various genetic markers (e.g., Collin *et al*., 2021). Notably, deep learning techniques are valuable tools in this context, offering the capability to analyse both raw genetic data and summary statistics (Korfmann *et al*., 2023). Either way, it is crucial to recognize that this diversity in data representation is a notable constraint when employing ML for species delimitation, as ML approaches typically handle the delimitation problem differently than traditional coalescent methods like BPP, which base their inferences on parameters directly derived from DNA sequences (Flouri *et al*., 2018; 2020).

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

data, an approach solely based on a particular set of summary statistics can inevitably result in a degree of information loss (Rannala & Yang, 2020).

Thus, the challenge in species delimitation context extends beyond the selection and optimisation of ML algorithms; it encompasses the development of workflows that effectively represent the input data's information, translating evolutionary processes under a given biological signal into testable hypotheses about species limits. Particularly, an alternative to learning from summary statistics is to consider the alignment itself as input, as demonstrated in the CNNs approach introduced by Perez *et al*. (2021). Remarkably, CNNs, along with other deep learning techniques, implicitly enable dimensionality reduction while capturing structures within the input data. This capacity facilitates accurate and efficient classification or regression tasks, as observed in studies by Sanchez *et al*. (2020), Fonseca *et al.* (2021), Perez *et al.* (2021), and Borowiec *et al.* (2022), thus holding promise in future species delimitation studies. Even so, while it might be feasible to compare results across different approaches, it is important to recognise that such comparisons could be somewhat misleading due to the variability in the biological foundations employed in each ML workflow. In other words, it is not always reasonable to strictly compare results produced by different ML approaches, as they are generally trained on specific parameterisations and ways of representing data. Comparisons should be performed considering the statistical properties of the used ML algorithms, such as how the workflows manipulate the data attributes, and the different types of input and output data.

This issue gains further significance when we consider that ML techniques are primarily lauded for their adaptability, especially in transfer learning frameworks. It is reasonable to assume that a neural network initially trained for a specific task can be repurposed for different learning contexts with the simple modification of some of its

layers. As an example, a deep learning architecture originally trained for inferring historical population sizes can be repurposed for classifying demographic scenarios (Pan & Yang, 2010). Moreover, when coupled with its capacity to simultaneously address phenotypic, ecological, and phylogeographic variables, the integration of ML analyses into species delimitation contributes to the construction of more profound and enlightening insights into taxonomical and speciation processes (e.g., Yang *et al.*, 2022). Interestingly, this kind of approach of ML for species delimitation would also align with de Queiroz's generalized species concept (1998; 1999), mainly due to ML's capability to accommodate diverse data types. Within this context, while the primary criterion for recognising a species would still be evolutionary independence, other characteristics may serve as secondary evidence of divergence and could be also analysed using ML approaches.

V. OPTIMISING THE USE OF ML IN THE CONTEXT OF SPECIES DELIMITATION

While it is undeniable that the development of new ML-based methods (or the adaptation of methods from other fields) contributed to the species delimitation literature, it is crucial for researchers in this field to maintain a set of guiding questions. In this regard, we present a basic framework for the selection and assessment of ML workflows in the context of species delimitation (Fig. 5). Our aim is not to comprehensively outline all steps for implementing a broad ML project (for a broader overview, refer to Chicco *et al*., 2017; Fountain-Jones *et al*., 2021; Greener *et al.*, 2021; Lee *et al*., 2022), but rather to propose key considerations for ML applications targeted at species limits inference. To accomplish this, we drew upon and adapted certain questions from Greener *et al.* (2021), which addressed the critical aspects to contemplate when reading or reviewing articles

782 employing ML on biological data. Furthermore, the proposed framework acknowledges
783 the potential use of both ML and coalescent-based methods.

784 Choosing a ML method should not be grounded on popularity but on its suitability
785 for the specific data and research questions at hand (Greener *et al.*, 2021). It is crucial for
786 researchers to thoroughly assess how new proposed methods truly differ from existing
787 ones. While developing a new method for species delimitation is undoubtedly a valuable
788 endeavour, it is equally important to consider the extent to which it contributes to the
789 current literature, given the existing diversity of methods. Many of these, despite their
790 limitations, have historically demonstrated their utility and effectiveness in tackling
791 various biological challenges. In the context of considering ML as an alternative to
792 coalescent methods, it is important to assess whether there are specific evolutionary
793 scenarios where fully coalescent methods exhibit limitations, and whether a new ML
794 workflow might outperform others in terms of performance. Additionally, users and
795 developers should bear in mind that many ML frameworks still rely on coalescent
796 principles, as many genetic simulators used in SML and deep learning approaches,
797 operate within the framework of Coalescent Theory (Hoban *et al.*, 2012; Hoban, 2014;
798 Peng *et al*., 2015).

799 An evaluation should encompass both the algorithm's biological predictions and
800 computational performance. Thus, a comprehensive analysis of its characteristics,
801 advantages, disadvantages, and overall performance compared to existing SDMs,
802 especially coalescent ones, is desired. For instance, Smith & Cartens (2020) argue that
803 traditional methods like BPP can accurately infer the number of species but may overlook
804 significant processes, such as secondary contact, something that ML workflows like
805 delimitR could be more efficient in dealing with. Also, one must consider that ML's
806 ability to efficiently compare a wide range of models using large datasets in less

computational time could provide a significant advantage over traditional model comparison approaches. The primary computational load typically involves simulating the training dataset, which can be alleviated by using multiple processors or graphical processing units.

A thorough description of the ML method, 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. Consequently, one of the initial steps within this process involves evaluating the dataset itself. For instance, is the dataset adequately described in terms of its structure and biological representation for species delimitation purposes? For example, Derkarebetian *et al.* (2022) assessed a ML approach's capability to delimit cryptic species, and constructed a "customised" training dataset from a well-studied lineage with biological characteristics akin to their focal taxon. In cases like these, where a specific ML classifier has been designed and trained with a particular dataset based on a specific evolutionary model's parameters, it is important to ensure both the dataset and the classifier are meticulously described and made accessible to the public. Furthermore, it is always pertinent to question the adequacy of the test set for addressing each biological problem, as it must be comprehensive enough to yield results congruent with the spectrum of examples encompassed in the training set.

Furthermore, especially within deep learning structures, where discerning the actual knowledge acquired by the neural network is challenging, achieving accurate predictions does not equate to learning a causal mechanism, even when the predictions are precise (Lee *et al*., 2022). Deep learning frameworks are intricate statistical models trained on high-dimensional data, and caution should be taken to avoid overinterpretation. Considering that tools employed for testing hypotheses regarding species limits span a

spectrum from interpretability to inferential power, deep learning workflows often find themselves at the extremes – offering high inferential power but limited interpretability. All of these factors accentuate the need for researchers to exercise prudence within this domain, bearing in mind the idiosyncrasies associated with each method and the specific biological or evolutionary models under investigation.

Once these considerations weigh in favour of developing or adapting a new ML method, it is imperative to plan its statistical evaluation and comparison to existing methods, whether coalescent-based or not, primarily focusing on predictive performance. The appropriate statistical metrics for assessing the algorithm's ability to predict species limits should be determined (see Moses, 2017; Ramsundar *et al.*, 2019). For example, it is common for researchers to evaluate the ML model's performance using genetic datasets of varying sizes, such as matrices containing 1,000, 5,000, and 10,000 SNPs, or alignments of different dimensions. Clearly, the quantity and quality of data significantly influence the effectiveness of ML applications. ML analyses conducted on larger, well-filtered datasets consistently yield better results (Pei *et al*., 2018; Smith & Carstens, 2020; Martin *et al*., 2021; Derkarebetian, *et al*., 2022). This effect is particularly pronounced in UML approaches, as they tend to be more susceptible to data-related issues (Martin *et al*., 2021). Additionally, it is essential to devise strategies to prevent overfitting. This becomes particularly significant when we consider that current ML methods are addressing various challenges (such as performance, handling of missing data, prevention of overfitting, and manipulation of evolutionary model parameters) in diverse ways.

While nearly all ML methods incorporate error or noise estimates in classification tasks (Pei *et al*., 2018; Smith & Carstens, 2020; Martin *et al.*, 2021; Derkarabetian *et al*., 2022), there is substantial variation in the metrics and evaluation methods chosen by researchers, something that can further complicate comparisons among studies. There

have been limited comparisons among ML methods used in species delimitation, and the limitations of those already found in the literature are still not completely understood. Besides, it is challenging to compare existing ML methods, as they often operate on different data transformations in terms of their biological representation and are generally trained with specific parameters tailored to the study's aims. Thus, it is prudent for researchers to question the appropriateness of prioritising one method over another, and to consider that an integrative framework encompassing various methods may also offer a sensible approach.

From a practical perspective, evaluating the suitability of an ML tool for species delimitation also involves assessing its accessibility, particularly when compared to established traditional methods. To promote the widespread adoption of ML tools in species delimitation, it is crucial to ensure that analyses are accessible and reproducible. This minimises 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). This reasoning, similar to that applied to the use of deep learning in Population Genetics (Korfmann *et al*., 2023), emphasizes the importance of making ML applications more user-friendly. Several factors can facilitate the integration of ML into a broader range of datasets, whether for species delimitation or other applications. For example, providing well-documented workflows, **pre-trained** models, and clear parameterisation details enables users to tailor model settings to the specific requirements of their biological systems. Likewise, the adoption of open-source software and programs, which is common practice in the field of ML, plays an important role in enhancing accessibility (Chicco, 2017; Heil *et al*., 2021).

Finally, it is essential to consider the diversity of programming environments used by different ML tools, as this can either facilitate or hinder researchers' usage, depending on their familiarity with specific coding structures or computing environments. In the studies we reviewed, Python and R workflows were the most commonly employed programming languages (Table 1; Pei *et al*., 2018; Derkarabetian *et al*., 2019; Smith & Carstens, 2020; Martin *et al*., 2021; Perez *et al*., 2021; Derkarabetian *et al*., 2021). This is not surprising, given the widespread adoption of Python and R in the biological sciences (Ekmekci *et al*., 2016; Perkel, 2021). Also, access to adequate computing resources remains a challenge for many researchers in species delimitation and various scientific disciplines (Veretnik *et al*., 2008; Truong *et al*., 2012; Helmy *et al*., 2016; Mangul *et al*., 2019b). Efforts to provide resources like graphics processing units, cloud storage, and computational clusters are all crucial steps toward making ML more accessible and inclusive for scientists across diverse domains of knowledge. We echo the existing literature's call (Chicco, 2017; Greener *et al*., 2022; Korfmann *et al*., 2023) and emphasise the importance of integrating and broadening ML in terms of equity and inclusion within the field of Evolutionary Biology as a whole, including increased training opportunities and participation in scientific events. As these conditions are increasingly met, ML is poised to become an integral part of the toolkit used by scientists not only in the field of species delimitation, but for various Evolutionary Biology applications worldwide.

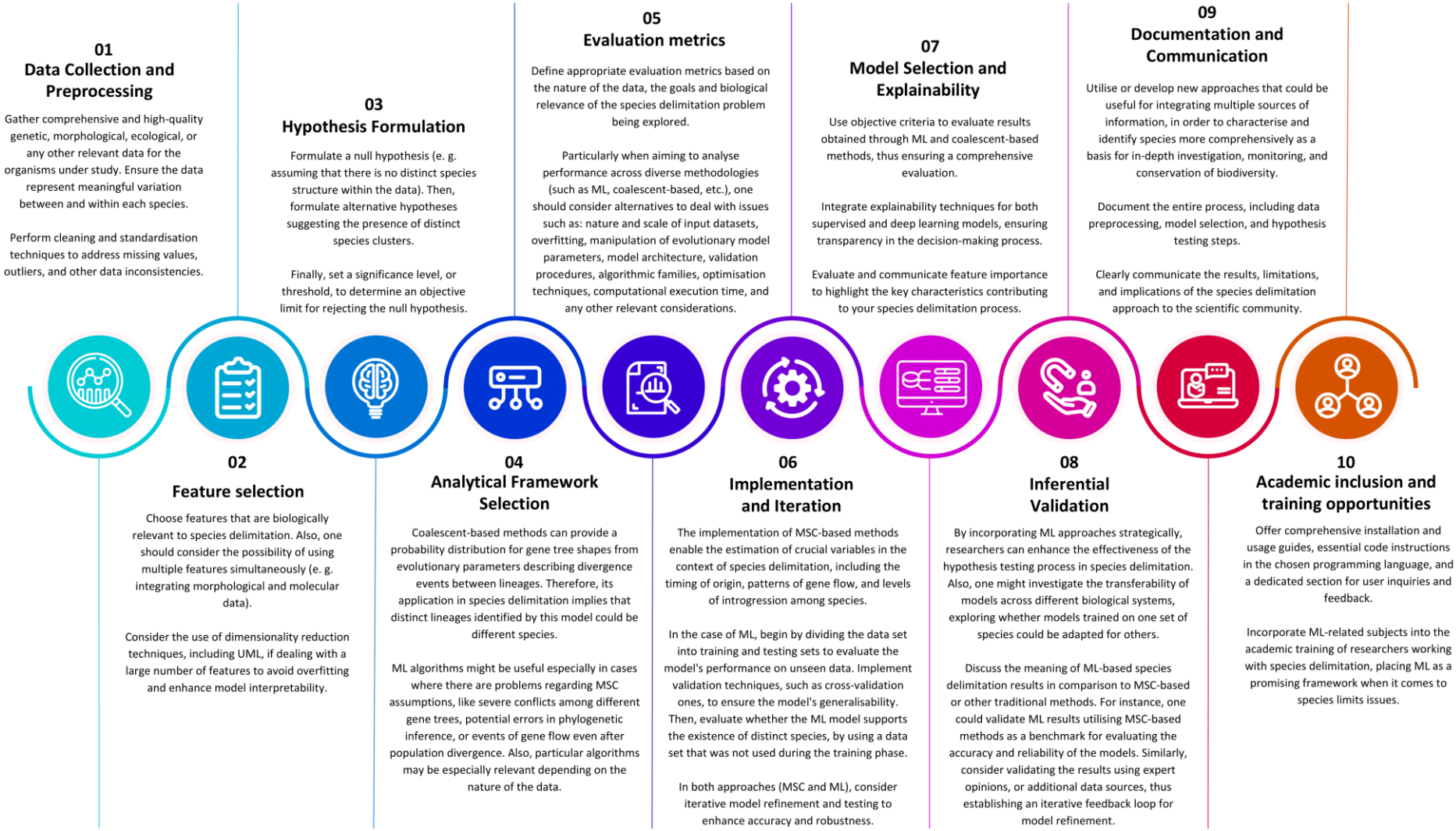

Fig. 5. Important considerations when evaluating, analysing, or creating machine learning (ML) approaches for inferring species limits involve several crucial steps. First, assessing a new ML method's potential contribution in species delimitation is key, especially in comparison to existing efficient methods across different evolutionary scenarios. A detailed understanding of the ML workflow, including data representation, model parameterisation, and training procedures is fundamental. Robust statistical evaluations of the ML method's performance, both computationally and in predicting species limits, are imperative. Additionally, emphasis should be placed on ensuring reproducibility and accessibility by documenting platforms and sharing data and models for broader utilisation. Finally, promoting inclusivity and encouraging broader participation of ML developers and researchers within the field of Evolutionary Biology should be a priority.

VI. DISCUSSION

Coalescence-based methods are robust tools for inferring species limits, but their predictive capacity can be limited, particularly in scenarios where gene flow rapidly halts post-population divergence (Fujita *et al*., 2012; Smith & Carstens, 2020). While certain coalescent approaches can identify populations as distinct species even with moderate gene flow (Jackson *et al*., 2017; Leaché *et al.*, 2019; Flouri *et al*., 2020), these species delimitation approaches should be used with caution when additional demographic processes influence lineage divergence during speciation (Smith & Carstens, 2020). Complexities like these suggest that using a single analytical approach, whether coalescent or otherwise, is unlikely to fully explore the intricate parameter space required for accurate species boundary inference. In this context, ML applications have emerged as a promising alternative. To date, relatively few studies (<20, see Supplementary Material) have specifically explored ML techniques for species delimitation, particularly when focusing on molecular data. While the potential for ML to revolutionise species delimitation, akin to its impact in various areas of Evolutionary Biology, is promising, this transformation will only be feasible with a comprehensive understanding of the diverse methodologies in existence. Among the studies examined here, only five introduced novel ML approaches for species delimitation, providing comprehensive details for researchers to follow—from initial simulations to statistical performance evaluations (Pei *et al*., 2018; Derkarabetian *et al*., 2019; Smith & Carstens, 2020; Perez *et al*., 2021; Pyron *et al.*, 2023).

Such approaches, also including some applied in phylogeography and demographic inferences, are often justified on the following arguments: i) challenges with coalescent method assumptions, as some researchers turn to ML techniques due to limitations associated with the assumptions of coalescent methods (Derkarabetian *et al*.,

2019; Smith & Carstens, 2020; Blischak *et al*., 2021; Martin *et al*., 2021; Derkarabetian *et al*., 2021); ii) computational efficiency and handling complex models, both in terms of computational efficiency and model adaptability (Pei *et al*., 2018; Martin *et al*., 2021; Perez *et al*., 2021; Derkarabetian *et al*., 2021; Pyron *et al*., 2023); and iii) integration with ABC methods, as ML can be combined with or used as an alternative to ABC methods (Sanchez *et al*., 2020; Smith & Carstens, 2020; Martin *et al*., 2021; Perez *et al*., 2021). This integration is often achieved through methods such as: i) adapting traditional summary statistics or selecting a more informative subset based on specific criteria (Smith & Carstens, 2020; Martin *et al*., 2021); and ii) incorporating ML techniques (e.g., RF) into the ABC framework to handle a larger number of summary statistics (Ghirotto *et al*., 2020; Smith & Carstens, 2020; Collin *et al*., 2021).

It is also notable that the criterion of evolutionary independence among metapopulation lineages (de Queiroz, 1998; 1999; 2005) takes precedence over other operational methods for species delimitation when it comes to ML frameworks. This preference may stem from our focus on workflows using molecular data, which aim to define evolutionary lineages and genetic groupings characterised by significant genetic divergence and restricted gene flow. While these criteria may have their limitations in investigating species limits, the results generated by ML methods in this context can serve as strong hypotheses for further investigations (e.g., Fujita *et al*., 2012). In cases where the independent evolutionary lineages or genetic groupings identified through ML methods may not precisely correspond to distinct species, these methods can still be adapted to analyse the same subjects using additional data sources. Consequently, there are scenarios where an integrative approach that builds upon methods with distinct statistical properties while comparing results and implications regarding species limits

will be appropriate. Robust and well-designed methods comparisons will help determine which methods are most suitable for particular biological questions.

Especially in SML or deep learning approaches, which often use explicit speciation models to validate species (e.g., Smith & 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 analysing complex evolutionary models, ML might serve a dual role: providing primary evidence for examining species limits patterns while aiding in the formulation of initial hypotheses, and assisting in the investigation and reconstruction of the processes responsible for these patterns. To empirically evaluate these methods for estimating unknown evolutionary parameters, a practical approach involves simulating data under various evolutionary models. However, data simulation carries significant limitations, particularly in complex evolutionary scenarios: the models may never be comprehensive enough, have limitations in representing real data, and demand substantial computational resources (Arenas, 2012; Mangul *et al*., 2019a; Zaharias *et al.*, 2022).

While these issues are not unique to ML-based workflows (inferential frameworks like ABC also employ simulated data; Beaumont *et al*., 2002; Bertorelle *et al*., 2010), simulations in this context appear to pose additional challenges. To address uncertainties related to the data simulation process, especially in Population Genetics studies, several solutions have been proposed. These include training networks on multiple "mis specified" models (Flagel *et al*., 2019; Torada *et al*., 2019; Adrion *et al.*, 2020), employing GANs (Booker *et al*., 2023; L. Nesterenko, B. Boussau, L. Jacob unpublished data; Smith & Hahn, 2023; Yelmen & Jay, 2023), as well as utilising domain adaptation techniques (Cochran *et al*., 2022; Mo & Siepel, 2023). Furthermore, the increasing

availability of trained models in the literature, particularly those with comprehensive documentation and trained under various parameterisations, is likely to facilitate future implementations. Concerns also arise regarding the true nature of species as identified by ML-based delimitation methods. As most of the approaches we presented rely on SNP data or in particular population genetics metrics, it is valid to question whether these methods genuinely discern species or primarily detect population structure (Sukumaran & Knowles, 2017; Huang, 2020).

Although some ML approaches incorporate tests to deal with such limitations, ML-based delimitation methods, just as some coalescent-based methods, might not always be identifying species *per se*, but rather: i) incompletely separated (or incipient) species, which may eventually be classified as distinct ones (Burbrink *et al*., 2022), or even as "subspecies" (de Queiroz, 2020); ii) ephemeral population or phylogeographic variation (Rosenblum *et al*., 2012; Sukumaran *et al*., 2021). Consequently, while ML methods hold increasing promise for species limits inference, even under the Generalized Lineage Concept of Species (de Queiroz, 1998; 1999; 2007), it is necessary to evaluate the extent to which the ML methods (just as coalescent-based ones) could effectively discern evolutionary independence among metapopulation lineages. Results obtained from these methods may not always provide definitive support for species delimitation hypotheses, but additional evidence for taxonomic decisions. Just as phenotypic, ecological, or other 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.

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. The choice on which species delimitation method to use should be done before hypothesis-testing, considering the nature of the available data, and possibly prior relevant biological information regarding the evolution of organisms, such that the best available model for the specific situation could be used. However, since ML methods for species delimitation are still in their infancy, this would be a difficult task for non-model organisms, and those for which no information on their diversification process is available. Thus, integrating coalescent-based methods into the hypothesis-testing process, alongside available ML methods, could enable a more comprehensive exploration of genetic and evolutionary models and parameters, improving the accuracy and biological interpretability of species delimitation analyses, and pave the way for the future use and applicability of ML methods.

Currently, by leveraging the strengths of both of those powerful analytical approaches, researchers will be able to construct a more reliable and defensible process for hypothesis testing in species delimitation, while accumulating evidence on the particular strengths of the methods. One particular type of approach that would benefit greatly from the combination of coalescence-based methods and machine learning algorithms, and that could shape the future direction of genetic-based species delimitation, involves the empirical validation of speciation-based models, which can provide a nuanced understanding of the speciation process. Different speciation-based delimitation models, whether relying on ML, coalescence, or a combination of both, could be employed to capture different facets of the process of evolutionary divergence, with model formulation serving as a means to articulate expert knowledge to statistical tools for hypothesis testing.

In addition to that, due to its great versatility in handling diverse data types, ML's future applications to infer species limits may also focus on evaluating which of the

different biological properties could be most effectively integrated into the species hypotheses testing process. This approach would also strongly align with de Queiroz's generalized species concept (1998; 1999; 2005), providing a deeper understanding of speciation processes through multiple biological perspectives. ML applications for species delimitation may serve as a robust tool for developing integrative taxonomy approaches by accommodating various types of input data, something that is fundamental in the light of the complex nature and variability observed within species. This becomes particularly appealing as AI-assisted approaches can be employed not only to test delimitation hypotheses, but also to analyse the relationships between evolutionary models and phylogeographic scenarios in terms of distinct characteristics, whether genetic, phenotypic, or ecological. While combining morphological and ecological analysis with molecular approaches can enhance inference quality (Wahlberg *et al*., 2005; Edwards & Knowles, 2014; Derkarabetian *et al*., 2022), relying solely on either method poses challenges. Only a few detailed ML pipelines have been proposed to address this challenge so far. For example, Yang *et al.* (2022) introduced a CNN method that successfully integrates morphological and molecular data for species identification. Pyron (2023), on the other hand, implemented a UML method using SOMs for learning high-dimensional associations between observations (e.g., individual specimens) across a wide set of input features (e.g., genetics, geography, environment, and phenotype). Future methodologies could explore this integration of multiple sources of information, both regarding species delimitation and integrative taxonomy.

Species delimitation is an increasingly challenging enterprise due to the growing availability of large-scale genomic data and the necessity to examine diverse evolutionary scenarios. While currently no universally superior species delimitation method exists, ML algorithms offer promising prospects for their integration into systematic protocols

tailored for species delimitation. Moving forward, it is imperative to conduct research on the performance of ML applications in terms of their adaptability to various parameterisations and the representation of genetic data. We suggest that ML species delimitation methods should follow a thorough evaluation of its strengths and weaknesses concerning the specific biological problem at hand, and preferably in comparison with coalescent-based approaches. Even if a particular ML algorithm is identified as a potential solution for addressing complex evolutionary problems, traditional coalescent methods could at least be used for benchmarking the ML algorithm's performance. As issues like these are solved, ML should progressively become a more practical, objective and robust alternative, paving the way for more concrete advancements when it comes to species delimitation.

VII. CONCLUSIONS

(1) Relatively few studies have explored ML techniques for species delimitation using molecular data so far. They are generally employed due to coalescent-based methods specific assumptions and limitations. Besides, they are computationally efficient, can be easily integrated with Bayesian approximation methods, and clearly provides a concrete and robust way to explore dataset structures when species-level divergences are hypothesised.

(2) ML approaches and coalescent-based methods provide a wide array of choices, necessitating careful selection considering multiple factors. Particularly, ML algorithms offer promising prospects but require thorough evaluation, comparison, and adaptation to specific biological problems, potentially in combination with traditional SDMs. Besides, selecting an appropriate ML method for species delimitation should prioritize suitability for specific data and research questions over popularity. This assessment includes

biological predictions, computational performance, and comparisons to existing methods, even considering that comparing existing methods can be challenging. Either way, there's a need for better comparative studies among ML methods and consideration of an integrative approach encompassing various methods.

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

(4) Just as some coalescent-based methods, ML-based delimitation methods may not always discern species, but might identify incompletely separated species or ephemeral population variations, offering strong hypotheses for further investigations. Therefore, ML should be progressively developed and used alongside coalescent-based methods to enhance objectivity and robustness in species delimitation processes, combining the strengths of both for hypothesis testing. Also, future applications of ML methods in species delimitation may focus on integrating various biological properties into species hypothesis testing, aiding in understanding speciation processes, accommodating different types of input data, and dealing more effectively with problems associated with data simulation. Besides, there is potential in utilizing ML methods in Integrative

Taxonomy approaches, as combining morphological, ecological, and molecular data, is crucial for robust species delimitation and may benefit from the flexibility of these AI-based approaches.

VIII. ACKNOWLEDGEMENTS

We thank André Luiz Gomes de Carvalho, Fernanda de Pinho Werneck and Renato José Pires Machado for their helpful suggestions in earlier versions of the text. We extend our gratitude to Daniel R. Schrider for critically reviewing the manuscript and providing suggestions. Matheus Salles is funded by a PhD scholarship granted by the Brazilian federal institution "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior" (CAPES).

IX. REFERENCES

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

Adrion, J. R. *et al.* (2020). A community-maintained standard library of population genetic models. *eLife* **9**. doi:10.7554/eLife.54967

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**.

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

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

1135 Bertorelle, G., Benazzo, A., Mona, S. (2010). ABC as a flexible framework to estimate demography over
1136 space and time: some cons, many pros. *Mol Ecol.* **19**, 2609–2625. doi:10.1111/j.1365-
1137 294X.2010.04690.x
1138 Blei, D. M. & Smyth, P. (2017). Science and data science. *Proceedings of the National Academy of*
1139 *Sciences* **114**, 8689–8692.
1140 Blischak, P. D., Barker, M. S., & Gutenkunst, R. N. (2021). Chromosome-scale inference of hybrid
1141 speciation and admixture with convolutional neural networks. *Molecular Ecology Resources* **21**,
1142 2676–2688. https://doi.org/10.1111/1755-0998.13355
1143 Booker, W. W., Ray, D. D., & Schrider, D. R. (2023). This population does not exist: learning the
1144 distribution of evolutionary histories with generative adversarial networks. *Genetics*, *224*(2),
1145 iyad063.
1146 Borowiec, M. L., Dikow, R. B., Frandsen, P. B., McKeeken, A., Valentini, G., & White, A. E. (2022).
1147 Deep learning as a tool for ecology and evolution. *Methods in Ecology and Evolution* **13**, 1640–
1148 1660.
1149 Bortolus, A. (2008). Error cascades in the biological sciences: the unwanted consequences of using bad
1150 taxonomy in ecology. *AMBIO: A journal of the human environment* **37**, 114–118.
1151 Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., & Krishnan, D. (2017). Unsupervised pixel-level
1152 domain adaptation with generative adversarial networks. *Proceedings of the IEEE conference on*
1153 *computer vision and pattern recognition*, 3722–3731.
1154 Breitman, M. F., Domingos, F. M., Bagley, J. C., Wiederhecker, H. C., Ferrari, T. B., Cavalcante, V. H.,
1155 ... & Colli, G. R. (2018). A new species of *Enyalius* (Squamata, Leiosauridae) endemic to the
1156 Brazilian Cerrado. *Herpetologica* **74**, 355–369.
1157 Burbrink, F. T., & Ruane, S. (2021). Contemporary philosophy and methods for studying speciation and
1158 delimiting species. *Ichthyology & Herpetology* **109**, 874–894.
1159 Callier, V. (2022). Machine learning in evolutionary studies comes of age. *Proceedings of the National*
1160 *Academy of Sciences* **119**.
1161 Camargo, A. (2013). Species delimitation: a decade after the renaissance. In *The species problem-ongoing*
1162 *issues*. IntechOpen.

1163 Camargo, A., Morando, M., Avila, L. J. & Sites, J. W. (2012). Species delimitation with abc and other
1164 coalescent-based methods: A test of accuracy with simulations and an empirical example with
1165 lizards of the *Liolaemus darwinii* complex (Squamata: Liolaemidae). *Evolution* **66**, 2834–2849.
1166 Carstens, B. C., Pelletier, T. A., Reid, N. M., & Satler, J. D. (2013). How to fail at species
1167 delimitation. *Molecular Ecology* **22**, 4369–4383.
1168 Cartwright, H. M. (2008). Artificial neural networks in biology and chemistry - the evolution of a new
1169 analytical tool. *Artificial neural networks*, 1-13.
1170 Cerca, J., Maurstad, M. F., Rochette, N. C., Rivera-Colón, A. G., Rayamajhi, N., Catchen, J. M., &
1171 Struck, T. H. (2021). Removing the bad apples: A simple bioinformatic method to improve loci-
1172 recovery in de novo RADseq data for non-model organisms. *Methods in Ecology and Evolution*
1173 **12**, 805–817.
1174 Chadha, A., Kumar, V., Kashyap, S., & Gupta, M. (2021). Deepfake: An overview. In *Proceedings of*
1175 *Second International Conference on Computing, Communications, and Cyber-Security*, pp. 557-
1176 566. Springer, Singapore.
1177 Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining* **10**, 1–
1178 17. https://doi.org/10.1186/s13040-017-0155-3
1179 Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in*
1180 *Ecology and Evolution* **10**, 1632–1644.
1181 Cochran, K., Srivastava, D., Shrikumar, A., Balsubramani, A., Hardison, R. C., Kundaje, A., Mahony, S.
1182 (2022). Domain adaptive neural networks improve cross-species prediction of transcription
1183 factor binding. *Genome Res.* **32**, 512–523.
1184 Collin, F. D., Durif, G., Raynal, L., Lombaert, E., Gautier, M., Vitalis, R., Marin, J. M., & Estoup, A.
1185 (2021). Extending approximate Bayesian computation with supervised machine learning to infer
1186 demographic history from genetic polymorphisms using DIYABC Random Forest. *Molecular*
1187 *Ecology Resources* **21**, 2598–2613. https://doi.org/10.1111/1755-0998.13413.
1188 Csilléry, K., Blum, M. G., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian computation
1189 (ABC) in practice. *Trends in Ecology & Evolution* **25**, 410–418.
1190 Degnan, J. H. & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the
1191 multispecies coalescent. *Trends Ecol. Evol.* **24**, 332–340.

1192 Derkarabetian, S., Castillo, S., Koo, P. K., Ovchinnikov, S., & Hedin M. (2019). A demonstration of
1193 unsupervised machine learning in species delimitation. *Molecular Phylogenetics and Evolution*
1194 **139**. https://doi.org/10.1016/j.ympev.2019.106562
1195 Derkarabetian, S., Starrett, J., & Hedin, M. (2022). Using natural history to guide supervised machine
1196 learning for cryptic species delimitation with genetic data. *Frontiers in Zoology* **19**, 1–15.
1197 Dike, H. U., Zhou, Y., Deveerasetty, K. K., & Wu, Q. (2018). Unsupervised learning based on artificial
1198 neural network: A review. In *2018 IEEE International Conference on Cyborg and Bionic*
1199 *Systems (CBS)*, pp. 322-327.
1200 Domingos, F. M., Bosque, R. J., Cassimiro, J., Colli, G. R., Rodrigues, M. T., Santos, M. G., &
1201 Beheregaray, L. B. (2014). Out of the deep: cryptic speciation in a Neotropical gecko (Squamata,
1202 Phyllodactylidae) revealed by species delimitation methods. *Molecular Phylogenetics and*
1203 *Evolution* **80**, 113–124.
1204 *Duan, L., Han, L. N., Liu, B., Leostrin, A., Harris, A. J., Wang, L., Arslan, E., Ertuğrul, K., Knyazev,
1205 M., Hantemirova, E., Wen, J., & Chen, H. F. (2023). Species delimitation of the liquorice tribe
1206 (Leguminosae: Glycyrrhizeae) based on phylogenomic and machine learning analyses. *Journal*
1207 *of Systematics and Evolution* **61**, 22–41. https://doi.org/10.1111/jse.12902.
1208 Edwards, S. V. (2009). Is a new and general theory of molecular systematics emerging? *Evolution* **63**, 1–
1209 19.
1210 Edwards, D. L., & Knowles, L. L. 2014. Species detection and individual assignment in species
1211 delimitation: can integrative data increase efficacy? *Proceedings of the Royal Society B:*
1212 *Biological Sciences* **281**, 20132765.
1213 Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., ... & Davis, C. C.
1214 (2016). Implementing and testing the multispecies coalescent model: a valuable paradigm for
1215 phylogenomics. *Molecular Phylogenetics and Evolution* **94**, 447–462.
1216 Ence, D. D., Carstens, B. C. (2011). SpedeSTEM: a rapid and accurate method for species delimitation.
1217 *Mol. Ecol. Resour.* **11**, 473–480.
1218 Ekmekci, B., McAnany, C. E., & Mura, C. (2016). An introduction to programming for bioscientists: a
1219 Python-based primer. *PLoS computational biology* **12**.
1220 Ely, C. V., de Loreto Bordignon, S. A., Trevisan, R., & Boldrini, I. I. (2017). Implications of poor
1221 taxonomy in conservation. *Journal for Nature Conservation* **36**, 10–13.

1222 Excoffier, L. *et al*. (2021). fastsimcoal2: demographic inference under complex evolutionary scenarios.
1223 *Bioinformatics* **37**, 4882–4885. doi:10.1093/bioinformatics/btab468.
1224 *Fan, X. K., Wu, J., Comes, H. P., Feng, Y., Wang, T., Yang, S. Z., Iwasaki, T., Zhu, H., Jiang, Y., Lee,
1225 J., & Li, P. (2023). Phylogenomic, morphological, and niche differentiation analyses unveil
1226 species delimitation and evolutionary history of endangered maples in Acer series Campestria
1227 (Sapindaceae). *Journal of Systematics and Evolution* **61**, 284–298.
1228 https://doi.org/10.1111/jse.12919.
1229 Farahani, A., Voghoei, S., Rasheed, K., & Arabnia, H. R. (2021). A brief review of domain adaptation.
1230 *Advances in data science and information engineering: proceedings from ICDATA 2020 and
1231 IKE 2020*, 877–894.
1232 Feder, J. L., Egan, S. P., & Nosil, P. (2012). The genomics of speciation-with-gene-flow. *Trends in
1233 Genetics* **28**, 342–350.
1234 Fišer, C., Robinson, C. T., & Malard, F. 2018. Cryptic species as a window into the paradigm shift of the
1235 species concept. *Molecular Ecology* **27**, 613–635.
1236 Flagel, L., Brandvain, Y., & Schrider, D. R. (2019). The unreasonable effectiveness of convolutional
1237 neural networks in population genetic inference. *Molecular Biology and Evolution* **36**, 220–238.
1238 Flouri, T., Jiao, X., Rannala, B., Yang, Z. (2018). Species Tree Inference with BPP using Genomic
1239 Sequences and the Multispecies Coalescent. *Molecular Biology and Evolution* **35,** 2585–2593.
1240 doi:10.1093/molbev/msy147.
1241 (2020). A Bayesian implementation of the multispecies coalescent model with introgression for
1242 phylogenomic analysis. *Molecular Biology and Evolution* **37**, 1211–1223.
1243 Fonseca, R. R. *et al.* (2016). Next-generation biology: sequencing and data analysis approaches for
1244 non-model organisms. *Marine genomics* **30**, 3–13.
1245 Fonseca, E. M., Colli, G. R., Werneck, F. P., & Carstens, B. C. (2021). Phylogeographic model selection
1246 using convolutional neural networks. *Molecular Ecology Resources* **21**, 2661–2675.
1247 https://doi.org/10.1111/1755-0998.13427.
1248 Fountain-Jones, N. M., Smith, M. L., & Austerlitz, F. (2021). Machine learning in molecular ecology.
1249 *Molecular Ecology Resources* **21**, 2589–2597. https://doi.org/10.1111/1755-0998.13532.
1250 Fraïsse, C., Popovic, I., Mazoyer, C., Spataro, B., Delmotte, S., Romiguier, J., Loire, É., Simon, A.,
1251 Galtier, N., Duret, L., Bierne, N., Vekemans, X., & Roux, C. (2021). DILS: Demographic

1252 inferences with linked selection by using ABC. *Molecular Ecology Resources* **21**, 2629–2644.
1253 https://doi.org/10.1111/1755-0998.13323.
1254 Fujita, M. K., Leaché, A. D., Burbrink, F. T., McGuire, J. A., & Moritz, C. (2012). Coalescent-based
1255 species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution* **27**, 480–488.
1256 Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation.
1257 In *International conference on machine learning*, 1180–1189.
1258 Ghifary, M, Kleijn, W. B., Zhang, M., Balduzzi, D., Li, W. (2016). Deep Reconstruction Classification
1259 Networks for Unsupervised Domain Adaptation. In: Leibe B, Matas J, Sebe N, Welling M,
1260 editors. *Computer Vision ECCV 2016. Lecture Notes in Computer Science*. Cham: Springer
1261 International Publishing. p. 597
1262 Ghirotto, S., Vizzari, M. T., Tassi, F., Barbujani, G. & Benazzo, A. (2021). Distinguishing among
1263 complex evolutionary models using unphased whole-genome data through random forest
1264 approximate Bayesian computation. *Molecular Ecology Resources* **21**, 2614–2628.
1265 https://doi.org/10.1111/1755-0998.13263.
1266 Giarla, T. C., Voss, R. S., & Jansa, S. A. (2014). Hidden diversity in the Andes: comparison of species
1267 delimitation methods in montane marsupials. *Molecular Phylogenetics and Evolution* **70**, 137–
1268 151.
1269 Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. &
1270 Bengio Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing*
1271 *Systems*, 2672–2680.
1272 Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2021). A guide to machine learning for
1273 biologists. *Molecular Cell Biology* **23**, 40–55. https://doi.org/10.1038/s41580-021-00407-0.
1274 Haller, B. C. & Messer, P. W. (2019). SLiM 3: forward genetic simulations beyond the Wright–Fisher
1275 model. *Molecular biology and evolution* **36**, 632–637.
1276 Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. In *The elements of statistical*
1277 *learning* (pp. 485-585). Springer, New York, NY.
1278 Haubold, B., & Börsch-Haubold, A. (2017). *Bioinformatics for Evolutionary Biologists.* Springer.
1279 Hausdorf, B. 2011. Progress toward a general species concept. *Evolution* **65**, 923–931.

1280 Heil, B. J., Hoffman, M. M., Markowetz, F., Lee, S. I., Greene, C. S. & Hicks, S. C. (2021).
1281 Reproducibility standards for machine learning in the life sciences. *Nature Methods **18***, 1132–
1282 1135.
1283 Helmy, M., Awad, M., & Mosa, K. A. (2016). Limited resources of genome sequencing in developing
1284 countries: challenges and solutions. *Applied & translational genomics* **9**, 15–19.
1285 Hoban, S., Bertorelle, G. & Gaggiotti, O. E. (2012). Computer simulations: tools for population and
1286 evolutionary genetics. *Nature Reviews Genetics* **13**, 110–122.
1287 Hoban, S. (2014). An overview of the utility of population simulation software in molecular
1288 ecology. *Molecular ecology* **23**, 2383–2401.
1289 Hüllermeier, E., Fober, T. & Mernberger, M. (2013). Inductive bias. *Encyclopedia of systems biology*,
1290 1018–1019.
1291 Huang, J. P. (2020). Is population subdivision different from speciation? From phylogeography to species
1292 delimitation. *Ecology and Evolution* **10**, 6890–6896.
1293 Huelsenbeck, J. P., Andolfatto, P., Huelsenbeck, E. T. (2011). Structurama: Bayesian inference of
1294 population structure. *Evolutionary Bioinformatics* **7,** 55–59.
1295 Jackson, N. D., Carstens, B. C., Morales, A. E. & O'Meara B. C. (2017). Species delimitation with gene
1296 flow. *Systematic Biology* **66**, 799–812.
1297 Jackson, N. D., Morales, A. E., Carstens, B. C. & O'Meara B. C. (2017). PHRAPL: phylogeographic
1298 inference using approximate likelihoods. *Systematic Biology* **66**, 1045–1053.
1299 Jacobs, S. J., Kristofferson, C., Uribe-Convers, S., Latvis, M., & Tank, D. C. (2018). Incongruence in
1300 molecular species delimitation schemes: What to do when adding more data is
1301 difficult. *Molecular Ecology* **27**, 2397–2413.
1302 *Jamdade, R., Al-Shaer, K., Al-Sallani, M., Al-Harthi, E., Mahmoud, T., Gairola, S., & Shabana, H. A.
1303 2022. Multilocus marker-based delimitation of *Salicornia persica* and its population
1304 discrimination assisted by supervised machine learning approach. *PLoS ONE* **17**.
1305 https://doi.org/10.1371/journal.pone.0270463.
1306 Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., ... & Zhang, G. (2014). Whole-genome
1307 analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331.
1308 Jörger, K. M., & Schrödl, M. (2013). How to describe a cryptic species? Practical challenges of molecular
1309 taxonomy. *Frontiers in Zoology* **10**, 1–27.

1310 Jorna, J. *et al.* (2021). Species boundaries in the messy middle—A genome-scale validation of species
1311 delimitation in a recently diverged lineage of coastal fog desert lichen fungi. *Ecology and*
1312 *Evolution* **11**, 18615-18632.
1313 Kapli, P., Yang, Z. & Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nature*
1314 *Reviews Genetics* **21,** 428–444.
1315 Kamilaris, A. & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and*
1316 *electronics in agriculture* **147**, 70–90.
1317 Korfmann, K., Gaggiotti, O. E. & Fumagalli, M. (2023). Deep learning in population genetics. *Genome*
1318 *Biology and Evolution*. https://doi.org/10.1093/gbe/evad008.
1319 Kuzenkov, O., Morozov, A., & Kuzenkova, G. (2020). Exploring evolutionary fitness in biological
1320 systems using machine learning methods. *Entropy* **23**, 1–35.
1321 Larsen, B. B., Miller, E. C., Rhodes, M. K. & Wiens J. J. (2017). Inordinate fondness multiplied and
1322 redistributed: the number of species on earth and the new pie of life. *The Quarterly Review of*
1323 *Biology* **92**, 229–265.
1324 Leaché, A. D., Harris, R. B., Rannala, B. & Yang, Z. (2014). The influence of gene flow on species tree
1325 estimation: a simulation study. *Systematic Biology* **63**, 17–30.
1326 Leaché, A. D., Zhu, T., Rannala, B., & Yang, Z. (2019). The spectre of too many species. *Systematic*
1327 *Biology* **68**, 168–181.
1328 Lee, B. D., Gitter, A., Greene, C. S., Raschka, S., Maguire, F., Titus, A. J., ... & Boca, S. M. (2022). Ten
1329 quick tips for deep learning in biology. *PLoS computational biology* **18**.
1330 Li, Q. (2012*). Literature survey: domain adaptation algorithms for natural language processing.*
1331 Department of Computer Science The Graduate Center, The City University of New York, 8–10.
1332 Libbrecht, M. W. & Noble, W. S. (2015). Machine learning applications in genetics and genomics.
1333 *Nature Reviews Genetics* **16**, 32–-332.
1334 Liu, B. (2019). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on
1335 machine learning approaches. *Briefings in bioinformatics* **20**, 1280–1294.
1336 Liu, M. Y. & Tuzel, O. (2016). Coupled Generative Adversarial Networks. In: *Advances in Neural*
1337 *Information Processing Systems* **29**. Curran Associates, Inc.
1338 https://papers.nips.cc/paper/2016/hash/502e4a16930e414107ee22b6198c578f-Abstract.html.

1339 Locey, K. J., & Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proceedings of the*
1340 *National Academy of Sciences* **113**, 5970–5975.
1341 Lukhtanov, V. A. (2019). Species Delimitation and Analysis of Cryptic Species Diversity in the XXI
1342 Century. *Entmol. Rev*. **99**, 463–472. https://doi.org/10.1134/S0013873819040055.
1343 Luo, A., Ling, C., Ho, S. Y., & Zhu, C. D. (2018). Comparison of methods for molecular species
1344 delimitation across a range of speciation scenarios. *Systematic Biology* **67**, 830–846.
1345 Lürig, M. D., Donoughe, S., Svensson, E. I., Porto, A. & Tsuboi, M. (2021). Computer vision, machine
1346 learning, and the promise of phenomics in ecology and evolutionary biology. *Frontiers in*
1347 *Ecology and Evolution* ***9***.
1348 Mangul, S. *et al.* 2019a. Systematic benchmarking of omics computational tools. *Nature communications*
1349 **10**.
1350 2019b. How bioinformatics and open data can boost basic science in countries
1351 and universities with limited resources. *Nature biotechnology* **37**, 324–326.
1352 Martin, B. T., Chafin, T. K., Douglas, M. R., Placyk, Jr J. S., Birkhead, R. D., Phillips, C. A., & Douglas,
1353 M. E. (2021). The choices we make and the impacts they have: Machine learning and species
1354 delimitation in North American box turtles (*Terrapene* spp.). *Molecular Ecology Resources* **21**,
1355 2801–2817.
1356 McClure, E. C., Sievers, M., Brown, C. J. Buelow, C. A., Ditria, E. M., Hayes, M. A., ... & Connolly R.
1357 M. (2020). Artificial intelligence meets citizen science to supercharge ecological monitoring.
1358 *Patterns* **1**.
1359 Messer, P. W. (2013). SLiM: simulating evolution with selection and link-age. *Genetics* **194**, 1037–1039.
1360 doi:10.1534/genetics.113. 152181.
1361 Mo, Z., & Siepel, A. (2023). Domain-adaptive neural networks improve supervised machine learning
1362 based on simulated population genetic data. *PLOS Genetics*, **19**.
1363 Mobadersany, P. *et al.* (2018). Predicting cancer outcomes from histology and genomics using
1364 convolutional networks. *Proceedings of the National Academy of Sciences* **115**, E2970-E2979.
1365 Morimoto, J., Ponchon, A., Sofronov, G., & Travis, J. (2021). Editorial: Applications of Machine
1366 Learning to Evolutionary Ecology Data. *Frontiers in Ecology and Evolution.*
1367 Moses, A. (2017). *Statistical modeling and machine learning for molecular biology.* CRC Press.
1368 *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.

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

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

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

Peng, B., Chen, H. S., Mechanic, L. E., Racine, B., Clarke, J., Gillanders, E., & Feuer, E. J. (2015). Genetic data simulators and their applications: an overview. *Genetic epidemiology* **39**, 2–10.

Perkel, J. M. (2021). Ten computer codes that transformed science. *Nature* **589**, 344–349.

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

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

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

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

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

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

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

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

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

(2005). Different species problems and their resolution. *BioEssays* **27**, 1263–1269.

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

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

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

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

Ramsundar, B., Eastman, P., Walters, P., & Pande, V. (2019). *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. O'Reilly Media.

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

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

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

(2020). Species Delimitation. In: *Phylogenetics in the genomic era.*

1428 Rannala, B., Edwards, S. V., Leaché, A., & Yang, Z. (2020). The Multi-species Coalescent Model and
1429 Species Tree Inference. Scornavacca, Celine; Delsuc, Frédéric; Galtier, Nicolas. *Phylogenetics*
1430 *in the Genomic Era*, No commercial publisher | Authors open access book.
1431 Raynal, L., Marin, J. M., Pudlo, P., Ribatet, M., Robert, C. P., & Estoup, A. (2019). ABC random forests
1432 for Bayesian parameter inference. *Bioinformatics* **35**, 1720–1728.
1433 Rozantsev, A., Salzmann, M. & Fua, P. (2018). Beyond sharing weights for deep domain adaptation.
1434 *IEEE transactions on pattern analysis and machine intelligence* **41**, 801–814.
1435 *Saryan, P., Gupta, S. & Gowda, V. (2020). Species complex delimitations in the genus *Hedychium*: A
1436 machine learning approach for cluster discovery. *Applications in Plant Sciences* **8**.
1437 https://doi.org/10.1002/aps3.11377.
1438 Saikia, U., Sharma, N. & Das, A. (2008). What is a Species? An endless debate. *Reson.* **13**, 1049–1064.
1439 https://doi.org/10.1007/s12045-008-0125-7.
1440 Sanchez, T., Cury, J., Charpiat, G. & Jay, F. (2020). Deep learning for population size history inference:
1441 Design, comparison and combination with approximate Bayesian computation. *Molecular*
1442 *Ecology Resources* **21**, 2645–2660.
1443 Sangster, G. (2013). The application of species criteria in avian taxonomy and its implications for the
1444 debate over species concepts. *Biological Reviews* **89**, 199–214. doi:10.1111/brv.12051.
1445 Schrider, D. R. & Kern, A. D. (2016). S/HIC: robust identification of soft and hard sweeps using machine
1446 learning. *PLoS Genetics* **12**.
1447 (2016). Discoal: flexible coalescent simulations with selection. *Bioinformatics* **32**,
1448 3839–3841. doi:10.1093/ bioinformatics/btw556.
1449 (2018). Supervised Machine Learning for Population Genetics: A New Paradigm.
1450 *Trends in Genetics* **34**, 301–312. https://doi.org/10.1016/j.tig.2017.12.005
1451 Searls, D. B. (2010). The Roots of Bioinformatics. *PLoS Comput Biol* **6**.
1452 https://doi.org/10.1371/journal.pcbi.1000809.
1453 Sheehan, S., & Song, Y. S. (2016). Deep learning for population genetic inference. *PLoS computational*
1454 *biology* **12**.
1455 Shen, X., Jiang, C., Wen, Y., Li, C., & Lu, Q. (2022). A Brief Review on Deep Learning Applications in
1456 Genomic Studies. *Frontiers in Systems Biology* **10**.

1457 Sidey-Gibbons, J. A., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical
1458 introduction. *BMC medical research methodology* **19**, 1–18.
1459 Simonsen, K. L., Churchill, G. A., Aquadro, C. F. (1995). Properties of statistical tests of neutrality for
1460 DNA polymorphism data. *Genetics* **1411**, 413–429.
1461 Sites, Jr J. W. & Marshall, J. C. (2004). Operational criteria for delimiting species. *Annual Review of*
1462 *Ecology, Evolution, and Systematics*, 199-227.
1463 Slatko, B. E., Gardner, A. F. & Ausubel, F. M. (2018). Overview of next-generation sequencing
1464 technologies. *Current protocols in molecular biology* **122**.
1465 Smith, M. L., Ruffley, M., Espíndola, A., Tank, D. C., Sullivan, J. & Carstens, B. C. (2017).
1466 Demographic Model Selection using Random Forests and the Site Frequency Spectrum.
1467 *Molecular Ecology*.
1468 Smith, M. L. & Carstens B. C. (2020). Process-based species delimitation leads to identification of more
1469 biologically relevant species. *Evolution* **74**, 216–229. https://doi.org/10.1111/evo.13878.
1470 Smith, M. L., & Hahn, M. W. (2023). Phylogenetic inference using generative adversarial networks.
1471 *Bioinformatics*, **39.**
1472 Strain, D. (2011). 8.7 million: A new estimate for all the complex species on Earth. *Science* **333**.
1473 Sukumaran, J. & Knowles, L. L. (2017). Multispecies coalescent delimits structure, not
1474 species. *Proceedings of the National Academy of Sciences* **114**, 1607–1612.
1475 Sukumaran, J., Holder, M. T., & Knowles, L. L. (2021). Incorporating the speciation process into species
1476 delimitation. *PLoS Computational Biology* **17**.
1477 Suvorov, A., Hochuli, J. & Schrider, D. R. (2020). Accurate inference of tree topologies from multiple
1478 sequence alignments using deep learning. *Systematic biology* **69**, 221–233.
1479 Tagu, D., Colbourne, J. K. & Nègre, N. (2014). Genomic data integration for ecological and evolutionary
1480 traits in non-model organisms. *BMC genomics* **15**, 1–16.
1481 Tautz, D., Arctander, P., Minelli, A., Thomas, R.H., Vogler, A. P. (2003). A plea for DNA taxonomy.
1482 *Trends Ecol. Evol.* **18**, 70–74.
1483 Torada, L., *et al.* (2019). Imagene: a convolutional neural network to quantify natural selection from
1484 genomic data. *BMC Bioinform.* **20**. doi:10.1186/s12859-019-2927-x

1485 Truong, H. L., Pham T. V., Thoai N., & Dustdar S. (2012). Cloud computing for education and research
1486 in developing countries. In *Cloud computing for teaching and learning: strategies for design and*
1487 *implementation*, pp. 64–80. IGI Global.
1488 Valletta, J. J., Torney, C., Kings, M., Thornton, A. & Madden J. (2017). Applications of machine learning
1489 in animal behaviour studies. *Animal Behaviour* **124**, 203–220.
1490 Veretnik, S., Fink, J. L. & Bourne, P. E. (2008). Computational biology resources lack persistence and
1491 usability. *PLoS computational biology* **4**.
1492 Vink, C. J., Paquin, P., & Cruickshank, R. H. (2012). Taxonomy and irreproducible biological science.
1493 *BioScience* **62**, 451–452.
1494 Vogler, A. P., Monaghan, M. T. (2007). Recent advances in DNA taxonomy. *J. Zool. Syst. Evol. Res.* **45**,
1495 1–10.
1496 Wahlberg, N., Braby, M. F., Brower, A. V. Z., Jong, R. de, Lee, M. M., Nylin, S., Pierce, N. E., Sperling,
1497 F. A. H., Vila, R., Warren, A. D., Zakharov, E. (2005). Synergistic effects of combining
1498 morphological and molecular data in resolving the phylogeny of butterflies and skippers. *Proc.*
1499 *R. Soc. B-Biol. Sci.* **272**, 1577–1586.
1500 Wake, D. B., Wake, M. H. & Specht C. D. (2011). Homoplasy: from detecting pattern to determining
1501 process and mechanism of evolution. *Science* **331**, 1032–1035.
1502 Wäldchen, J. & Mäder, P. (2018). Machine learning for image-based species identification. *Methods in*
1503 *Ecology and Evolution* **9**, 2216–2225.
1504 Wang, G. (2019). Machine learning for inferring animal behavior from location and movement data.
1505 *Ecological informatics* **49**, 69–76.
1506 Wang, Z., Wang, J., Kourakos, M., Hoang, N., Lee, H. H., Mathieson, I., & Mathieson, S. (2021).
1507 Automatic inference of demographic parameters using generative adversarial
1508 networks. *Molecular ecology resources* **21**, 2689–2705.
1509 Wiens, J. J. (2007). Species delimitation: new approaches for discovering diversity. *Syst. Biol.* **56**, 875–8.
1510 Xu, J., Xiao, L., & López, A. M. (2019). Self-supervised domain adaptation for computer vision tasks.
1511 *IEEE Access 7*, 156694-156706.
1512 Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. *Curr. Zool.* **61**,
1513 854–865.

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

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

Yuan, X., Miller, D. J., Zhang, J., Herrington, D, Wang, Y. (2012). An overview of population genetic data simulation. *J. Comput. Biol.* **19,** 42–54.

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

Zaharias, P., Grosshauser, M. & Warnow, T. (2022). Re-evaluating Deep Neural Networks for Phylogeny Estimation: The Issue of Taxon Sampling. *Journal of Computational Biology* **29**, 74–89. https://doi.org/10.1089/cmb.2021.0383.