Received: 7 September 2020 | Revised: 2 April 2021 | Accepted: 28 April 2021

DOI: 10.1111/1755-0998.13427

SPECIAL ISSUE

MOLECULAR ECOLOGY RESOURCES WILEY

Phylogeographic model selection using convolutional neural networks

Emanuel M. Fonseca[1] | Guarino R. Colli[2] | Fernanda P. Werneck[3] | Bryan C. Carstens[1]

[1]Department of Evolution, Ecology and Organismal Biology, The Ohio State University, Columbus, OH, USA

[2]Departamento de Zoologia, Universidade de Brasília, Brasília, Brazil

[3]Coordenação de Biodiversidade, Programa de Coleções Científicas Biológicas, Instituto Nacional de Pesquisas da Amazônia (INPA), Manaus, Brazil

Correspondence
Emanuel M. Fonseca, Department of Evolution, Ecology and Organismal Biology, The Ohio State University, 318 W. 12th Avenue, Columbus, OH 43210, USA.
Email: emanuelmfonseca@gmail.com

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number: 305535/2017-0; National Science Foundation, Grant/Award Number: DBI 1661029 and DEB 1831319; Ohio Supercomputer Center, Grant/Award Number: PAA0202; USAID's PEER program under cooperative agreement, Grant/Award Number: AID-OAA-A-11-00012; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Grant/Award Number: 88881.170016/2018

Abstract

The discipline of phylogeography has evolved rapidly in terms of the analytical toolkit used to analyse large genomic data sets. Despite substantial advances, analytical tools that could potentially address the challenges posed by increased model complexity have not been fully explored. For example, deep learning techniques are underutilized for phylogeographic model selection. In non-model organisms, the lack of information about their ecology and evolution can lead to uncertainty about which demographic models are appropriate. Here, we assess the utility of convolutional neural networks (CNNs) for assessing demographic models in South American lizards in the genus *Norops*. Three demographic scenarios (constant, expansion, and bottleneck) were considered for each of four inferred population-level lineages, and we found that the overall model accuracy was higher than 98% for all lineages. We then evaluated a set of 26 models that accounted for evolutionary relationships, gene flow, and changes in effective population size among the four lineages, identifying a single model with an estimated overall accuracy of 87% when using CNNs. The inferred demography of the lizard system suggests that gene flow between non-sister populations and changes in effective population sizes through time, probably in response to Pleistocene climatic oscillations, have shaped genetic diversity in this system. Approximate Bayesian computation (ABC) was applied to provide a comparison to the performance of CNNs. ABC was unable to identify a single model among the larger set of 26 models in the subsequent analysis. Our results demonstrate that CNNs can be easily and usefully incorporated into the phylogeographer's toolkit.

KEYWORDS
convolutional neural networks, deep learning, machine learning, *Norops* spp., phylogeography

1 | INTRODUCTION

One key goal of phylogeography has been to investigate how historical processes have shaped genetic variation across geographic space. Early phylogeographic investigations were highly qualitative, with inferences based largely on gene genealogies and the geographic distribution of the haplotypes. Due to their descriptive nature, these investigations were susceptible to overinterpretation (Knowles & Maddison, 2002), where a detailed explanation of the causes of intraspecific diversification usually went beyond the evidence supported by the data, and confirmation bias (Nickerson, 1998), in which researchers often interpreted new results in a manner that supported previous findings (Carstens et al., 2009). As the discipline matured, researchers recognized the importance of statistical approaches that explicitly incorporate uncertainty to draw meaningful conclusions about a species' evolutionary history.

Therefore, the identification of statistical models relevant for data analysis is a crucial step of any model-based phylogeographical investigation.

Researchers have employed three general approaches to identify the models used to describe the data and make inference: (i) Intuitive model identification, (ii) phylogeographic hypothesis testing, and (iii) objective model selection (Carstens et al., 2017). Biological intuition often drives the choice of the analytical framework(s) used to analyse the data. For example, researchers may choose to analyse their data with an isolation with migration model or an *n*-island migration model due to beliefs regarding the processes that have influenced their system. In practice, if the chosen model has a poor fit to the evolutionary history of the organism, the resulting parameter estimates (Koopman & Carstens, 2010) and other inferences can be misleading (Beerli & Palczewski, 2010; Hey et al., 2015). Notably, the estimation of many evolutionary processes eventually becomes intractable in a likelihood framework (Beaumont et al., 2002) such that no single analytical method can incorporate all possible evolutionary processes and use maximum likelihood or Bayesian methods to identify parameter values that maximize the probability of the model given the data (Beaumont, 2010). Similarly, hypothesis testing (e.g., Knowles et al., 2007) is conducted under an assumed model and thus subject to the same potential flaws as intuitive approaches. For these reasons, many researchers now utilize model selection approaches in phylogeographic research.

Simulation-based and likelihood-free approaches, which can accommodate complex demographic scenarios (Pritchard et al., 1999), are often utilized to conduct phylogeographic model selection. Software such as ms (Hudson, 2002), msprime (Kelleher et al., 2016), SliM (Haller & Messer, 2019), and *fastsimcoal2* (Excoffier et al., 2013) can be used to conduct simulations under customized demographic models that can approximate the details of almost any empirical system. After the simulation procedure, empirical and simulated data sets can be statistically evaluated using a variety of methods, including approximate Bayesian Computation (ABC; e.g., Fagundes et al., 2007), information theory (e.g., Carstens et al., 2009; Morales et al., 2017), and machine learning approaches such as random forest (Smith et al., 2017). While these have in common the flexibility to assess multiple demographic models given the observed data, factors such as the type of data collected and details about the empirical system make it likely that there is not a single "best" approach for all questions.

Information theoretic approaches can be conducted either on SNP data, summarized as site frequency spectra (SFS; e.g., Thomé & Carstens, 2016), or gene trees (e.g., Jackson et al., 2017). They appear effective at considering large numbers of models, but potentially at the expense of parameter estimation. In contrast, approximate Bayesian computation (ABC) remains a widely used approach in demographic model selection but can potentially suffer from the "curse of dimensionality" when comparing more than a handful of demographic models (Pelletier & Carstens, 2014; Schrider & Kern, 2018). The computational effort required by these approaches varies by application, but ABC becomes computationally expensive when the data are summarized on a locus-by-locus basis. For this reason, methods that summarize SNP data as SFS and use machine learning to identify the best model are increasingly being applied (e.g., Pudlo et al., 2016; Smith et al., 2017). As genomic data become easier to collect and more common in non-model systems, increased exploration of the usefulness of these (and other) approaches to phylogeographic model selection is warranted.

Supervised machine learning (SML) is a branch of artificial intelligence that gives computers the ability to learn from data without being explicitly programmed and where labels (i.e., preclassified data) are available for a subset of the samples. SML involves (i) training a predictive model using a subset of a labelled data set, (ii) evaluating the model using the remaining portion of the labelled data set, and (iii) using the now-trained model to predict new, unlabelled examples. One example of a SML approach to phylogeographic inferences is implemented in the R package delimitR (Smith & Carstens, 2020), which uses a random forest classifier to create hundreds of individual decision trees (a forest) from SNP data, summarized using SFS, to train the model. Next, the set of decision trees are combined via a consensus tree, and this tree is used to classify a new data set. Results from a simulation study indicate that delimitR is able to compare hundreds of alternative models with high accuracy, even when comparing complex evolutionary scenarios (Smith & Carstens, 2020; Smith et al., 2017). However, results in other fields that apply SML approaches indicate that random forest may not be as efficient as other approaches, such as convolutional neural networks (CNN; Box 1; Razzak et al., 2018). Since CNNs take as input a set of labelled images and train a model to predict the content of new images, one potential advantage of this approach is that predictions can be made directly from the alignment containing the genetic variation from sampled individuals (Blischak et al., 2020; Cheng et al., 2013; Flagel et al., 2019; Sanchez et al., 2020; Torada et al., 2019) rather than from data that are summarized using either summary statistics or a SFS. CNNs have been used to address a range of biological questions, from detecting natural selection (Flagel et al., 2019; Torada et al., 2019), reconstructing phylogenetic history (Suvorov et al., 2020), and predicting cancer outcomes (Mobadersany et al., 2018). In spite of all its benefits, the potential applicability of CNNs to phylogeographic model selection remains largely unexplored.

Here, we explore the usefulness of CNNs for phylogeographic model selection. We use a simulation-based approach to create labelled examples (i.e., DNA alignments), converted to a black and white image by labelling the major allele and the minor allele as 0 and 1, respectively. After training the model using 80% of the labelled data and evaluating its performance using the remaining 20% of the data, we compare the performance of CNNs and ABC to enquire about the evolutionary history of two species of lizards from contrasting environments in South America.

2 | MATERIALS AND METHODS

2.1 | South American lizards as a case study

We used SNP data collected from the lizard sister species *Norops brasiliensis* and *N. planiceps* as a case study to assess the usefulness of CNNs for phylogeographic model selection. Little is known about

BOX 1 Overview of Convolutional Neural Networks (CNNs)

Artificial neural networks (ANNs) were proposed as an attempt to mimic the network of neurons that constitute the animal brain. In human brains, for example, an external stimulus is passed through a chain of neurons that culminate in a response. Likewise, ANNs are fed with data (i.e., stimulus) which are passed through an artificial network of neurons to make a prediction (i.e., response). CNNs (also known as ConvNets) are a class of artificial neural networks that use a set of labelled images (input data) to build a model to differentiate among the various labels. First, a convolution operation is performed by multiplying each value in the input (Figure 1a) by a learnable weight within a kernel (Figure 1b). After the convolution operation, the images are converted into a feature map (Figure 1c) where each value is passed through a nonlinear function (e.g., ReLU, tanh, sigmoid). Next, a pooling method (maximum, average pooling, etc.) is applied to the feature maps within a kernel to reduce the dimensions of the feature maps and maintain conceivably important features from the convolutional kernel (Figure 1d). These steps can be replicated "n" times inside the CNN architecture. For example, in Figure 1, the convolutional and pooling steps were replicated twice. Lastly, the resulting array of all these operations is flattened into a one-dimensional array and fully connected to an ANN. Together, these steps comprise the forward propagation, in which the goal is to pass the data through the CNN (or ANN) and compute a loss function with respect to the weights. Once the loss function is computed, the CNN works backward (back-propagation) to optimize the weights and minimize the total loss function of the model using partial derivatives. In summary, a set of images is forward propagated into a CNN to calculate a loss function, which in turn is back-propagated to optimize the model weight and minimize the loss function. Thus, the training of a CNN consists of an iterative process of forward and backward propagation. Definitions of commonly used terms in this study are presented in Table 1 and a more detailed description of CNNs is available in Lecun et al. (2015) and Flagel et al. (2019).

their ecology, natural history, and evolution, which poses great uncertainty about which set of models are appropriate. *Norops brasiliensis* is a terrestrial and diurnal species that occurs predominantly in open areas and riparian forests (gallery forests) in the Cerrado savanna and enclaves of Cerrado within the Amazonian rainforest (Figure 2; Avila-Pires, 1995; Ribeiro, 2015) (Figure 2). *Norops planiceps* is also terrestrial and diurnal, but endemic to northern Amazonia, where it inhabits "terra firme" forests, which are not periodically flooded (Figure 2; Avila-Pires, 1995; Ribeiro, 2015). Because *Norops* inhabits Amazonia and the Cerrado, the largest Brazilian biomes, understanding its evolutionary history provides data regarding the evolution of these important regions. Amazonia is a region predominantly covered by tropical rainforests, whereas the Cerrado, a world hotspot priority for conservation (Myers et al., 2000), is characterized by sclerophyllous, fire-adapted flora, abundant grasses and short, thick-barked, and twisted trees (savanna-like vegetation). The Cerrado is part of the South American diagonal of "open formations" (also known as "dry diagonal" or "savanna corridor") and shares its north-western boundary with Amazonia.

2.2 | Sampling and data collection

We obtained 61 tissue samples; 52 from *N. brasiliensis* (nine localities) and nine from *N. planiceps* (five localities; Figure 2) from the Herpetological Collection of Brasília University (CHUNB) and the Collections of Amphibians and Reptiles and Genetic Resources from the National Institute of Amazonian Research (INPA-H and INPA-HT). DNA was extracted from liver or muscle tissues using E.Z.N.A. Tissue DNA Kit. Prepared libraries from each species for sequencing using a modified version of the genotyping-by-sequencing (GBS) protocol described in Elshire et al. (2011). For DNA digestion, we used 100 ng of freshly extracted DNA and the restriction enzyme Sbf1. After digestion–ligation reactions, we pooled all samples and purified using Agencourt AMPure beads. We amplified samples with polymerase chain reaction (PCR) as follows: (i) initial denaturation at 72°C for 5 min, (ii) 16 cycles consisting of: 98°C for 10 s for denaturation, 65°C for 30 s for annealing, and 72°C for 30 s for extension and (iii) final extension at 72°C for 5 min. We then quantified PCR products using the BR DNA Qubit Quantification Kit. We used the Blue Poppin Prep to select DNA fragments of 200–500 bp. Sequencing was carried out at the Ohio State University Comprehensive Cancer Center on an Illumina HiSeq 4000 and paired-end reads of 150 bp were generated.

2.3 | Data processing

We processed (i.e., sorted, demultiplexed, clustered, and formatted) raw data from Illumina outputs using ipyrad v 0.9.52 (Eaton & Overcast, 2020) and resources provided by the Ohio Supercomputer Center. We processed five different data sets: (i) all samples, (ii) *N. brasiliensis* (population 1), (iii) *N. brasiliensis* (population 2), (iv) *N. brasiliensis* (population 3), and (v) *N. planiceps*. Data sets ii–v represent distinct populations recovered in the population assignment analyses (see population assignments section). First, we demultiplexed raw data using individual barcode adapters. Next, we filtered for adapters using the stricter option. All reads were trimmed to 75 bp before analysis. We set the maximum low-quality base calls in the read to 5, only allowing reads longer than 35 bp. We clustered reads within each sample if their similarity was greater than 85%, set the maximum cluster depth within samples to 10,000 reads, and used

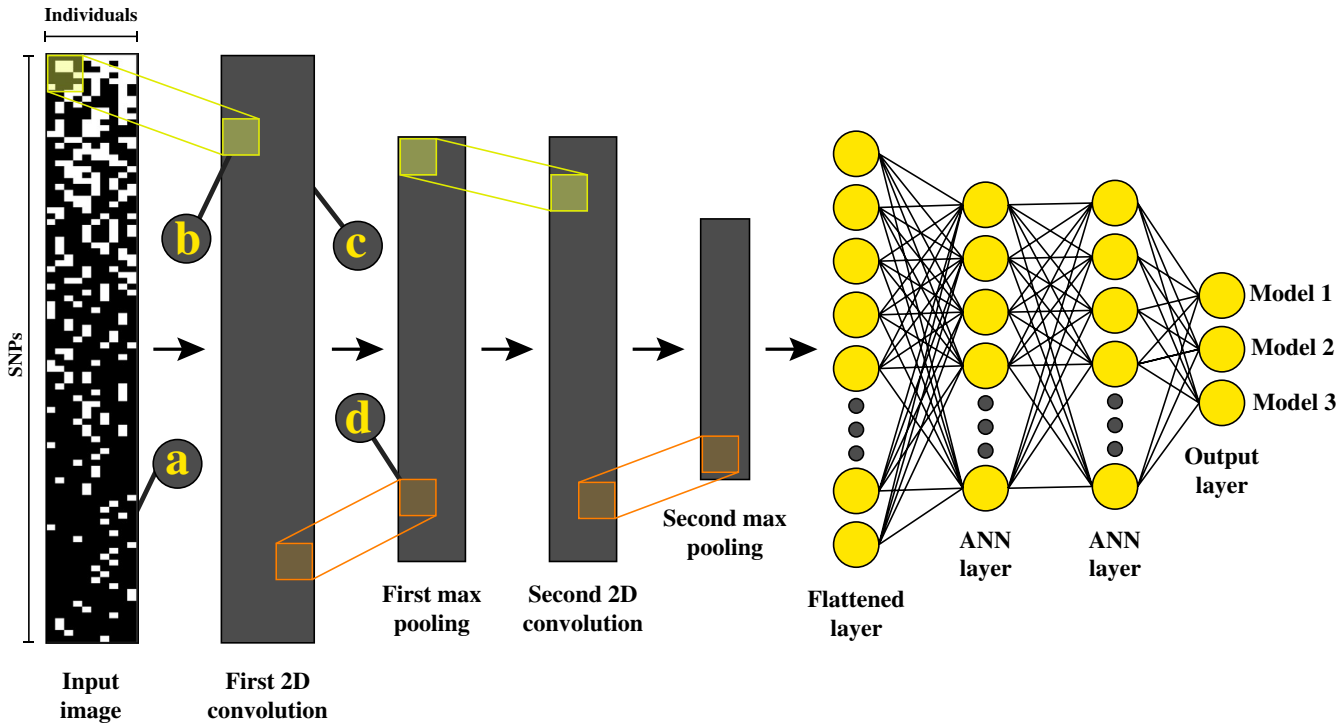

FIGURE 1 A general schematic representation of a two-dimensional convolutional neural network (CNN) architecture. (a) Input image, (b) convolutional kernel (yellow), (c) feature map, and (d) pooling kernel (orange). ANN, artificial neural network

TABLE 1 A glossary of terms used in this study

Term	Definition
Artificial neural network - ANN	A computational network of connected layers that attempt to mimic the way that the brain analyses and processes information (biological neural network)
Convolutional neural network -CNN	A type of artificial neural network used for image classification and recognition
CNN architecture	The general structure of the model that includes the number of convolution and pooling layers, size and numbers kernels, and the number of neurons in each hidden layer
Kernel	Vector of weights used for feature detection
Neuron	A mathematical function that takes a group of input and weights, applies an activation function (e.g., ReLU, tanh, sigmoid) and output a value
Loss function	A variety of methods designed to calculate the distance between actual and predicted outcomes
Epoch	The number of times that all images are fed into the model
Optimizer	A mathematical function used to update the weights of the model to minimize the loss function

a minimum depth for statistical base calling of six reads. Because CNNs do not allow missing data (see CNN section), we removed loci with missing data. While errors in estimates of admixture and summary statistics may accompany low-coverage data (e.g., Korneliussen et al., 2013; Skotte et al., 2013), it is unclear if these factors lead to biases in model selection. We expect that the potential for bias is reduced for CNN in comparison to approaches that summarize the data using summary statistics.

2.4 | Population assignments

STRUCTURE v2.3.4 (Pritchard et al., 2000) was used to partition samples into discrete populations before building demographic models. We ran three independent replicates using 100,000 steps of burnin, followed by 500,000 generations. We performed all runs under an admixture model for population ancestry and allele frequencies correlated among populations. We evaluated *K*-values ranging from 2 to 6, with 10 replications. Using the ad hoc statistic ΔK, we evaluated the optimal value of *K*, calculating the rate of change in the log probability of data between successive *K* values (Evanno et al., 2005), as in STRUCTURE HARVESTER (Earl & vonHoldt, 2012). We combined all replicate analyses under the best value of *K* using the software CLUMPP (Jakobsson & Rosenberg, 2007), and assigned individuals to populations based on their admixture proportion. For example, if an individual was assigned jointly to two populations, we placed that individual in the population with the higher admixture proportion.

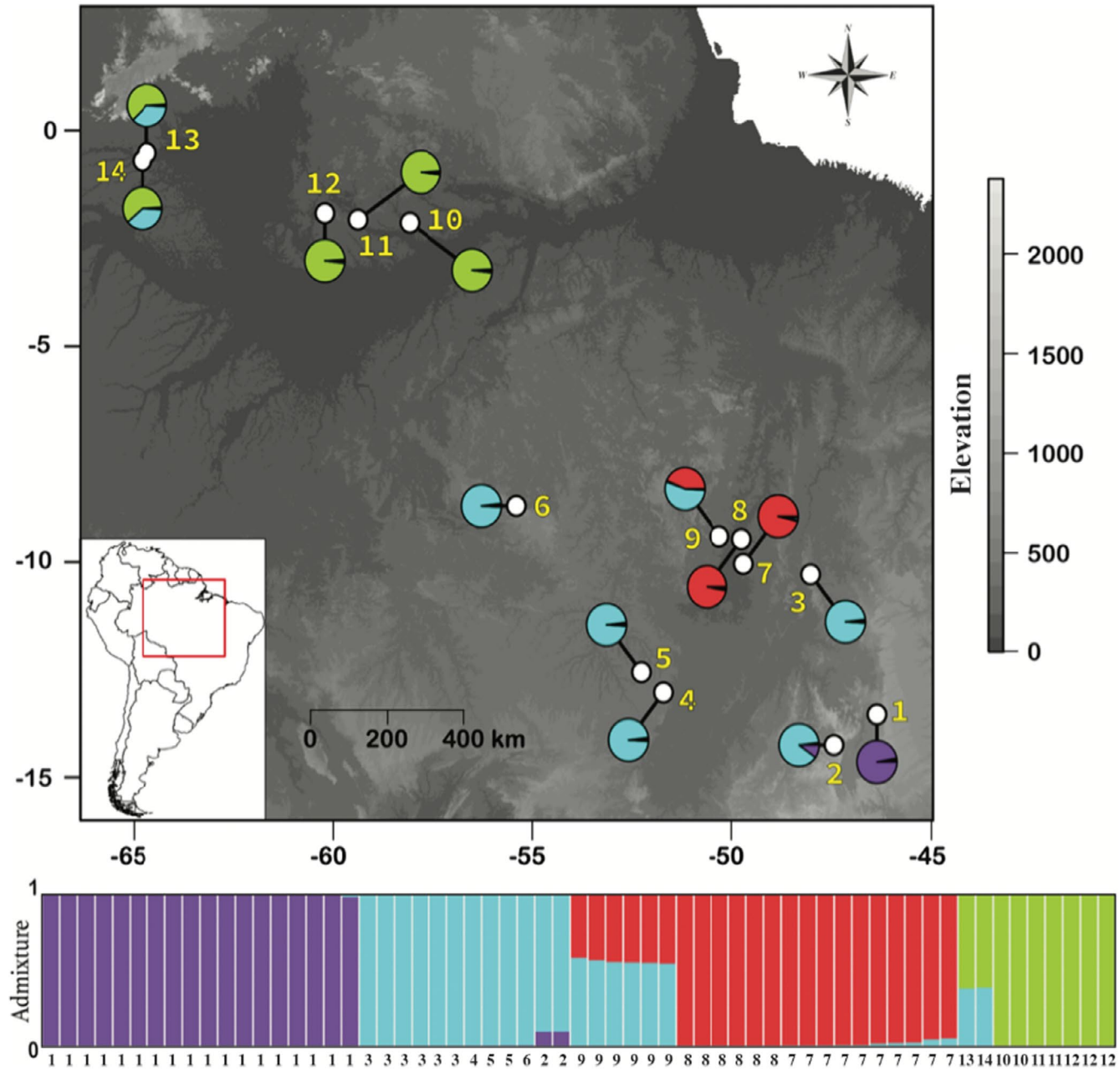

FIGURE 2 Map showing the geographic distribution of sampled localities. Information about samples and localities are available in Table S1. Purple circle, *Norops brasiliensis* (population 1); blue circles, *N. brasiliensis* (population 2); red circles, *N. brasiliensis* (population 3); green circles, *Norops planiceps*. Bar represents the admixture plot of *Norops* ssp. across the area of study according to STRUCTURE analysis. Numbers below the admixture plot represent the individuals sampled in each locality. Population 1, purple; population 2, blue; population 3, red; *Norops planiceps*, green

2.5 | Testing diversification history using convolutional neural networks

In phylogeographic model selection, there are countless ways of parameterizing a given model but, as the number of lineages and possible parameters increase, the number of possible models grows at a greater than exponential rate. For example, for the four populations we inferred based on the STRUCTURE results, there are more than 2000 possible models that could be designed if one incorporates topology (four populations), gene flow (isolation vs. secondary contact), and changes in population size (constant, bottleneck, and expansion). To facilitate comparison of all potential models, we conducted a hierarchical analysis by dividing our model selection into two components. First, we independently tested each population for demographic change in population size through time (12 models). Second, we applied the model of population size change that was individually identified in each population to a broader analysis that considered all possible topologies for four lineages with assorted migration scenarios (26 total models). With this approach, we reduced the model space from more than 2,000 to 38 competing models,

which greatly facilitated the comparison between the CNN and ABC approaches to model selection (below).

2.5.1 | *Testing population trajectory through time*

In the first part of model selection, we used a CNN to identify the population trajectory that best describes the demographic history of each population. Pleistocene climate oscillations are one of the main drivers of genetic variation across the globe (Haffer, 1969; Hewitt, 2000, 2004) and also hypothesized to have impacted the evolutionary and demographic history of *N. brasiliensis* and *N. planiceps* (Vanzolini & Williams, 1970). Because of that, priors were selected to mirror this hypothesis and are presented in Table S1. We defined three possible scenarios (Figure 3a): (i) Constant population size through time, (ii) population expansion since the last glacial maximum (LGM), and (iii) population bottleneck since the LGM. We used the software fastsimcoal2 (Excoffier et al., 2013) to simulate 10,000 data examples for each demographic scenario and population. We simulated short DNA sequences (1 to 5 bp) for 1,000,000 independent loci to ensure that the simulator only generated one SNP per locus and kept the same number of SNPs as observed in the empirical data sets. Also, individuals were randomly organized with respect to each other in the alignment, and SNPs were sorted based on major allele frequency (higher to a lower frequency). We parameterized the ancestral effective population size, current effective population size, and time of population size changing (Table S1). Next, we wrote custom R scripts to convert the alignment of each simulation into a biallelic matrix, with *n* rows and *k* columns, corresponding to the number of samples and SNPs, respectively. We labelled the major allele as (0) and the minor allele as (1), such that the matrix could be converted to a black and white image with each entry corresponding to a pixel in the image. Finally, we sorted SNPs based on their allele frequency (higher to a lower frequency).

We implemented a two-dimensional CNN architecture as follows: a two-dimensional convolution layer (kernel = 3 × 1), a two-dimensional maximum pooling layer (kernel = 3 × 1), a two-dimensional convolution layer (kernel = 3 × 1), and a two-dimensional maximum

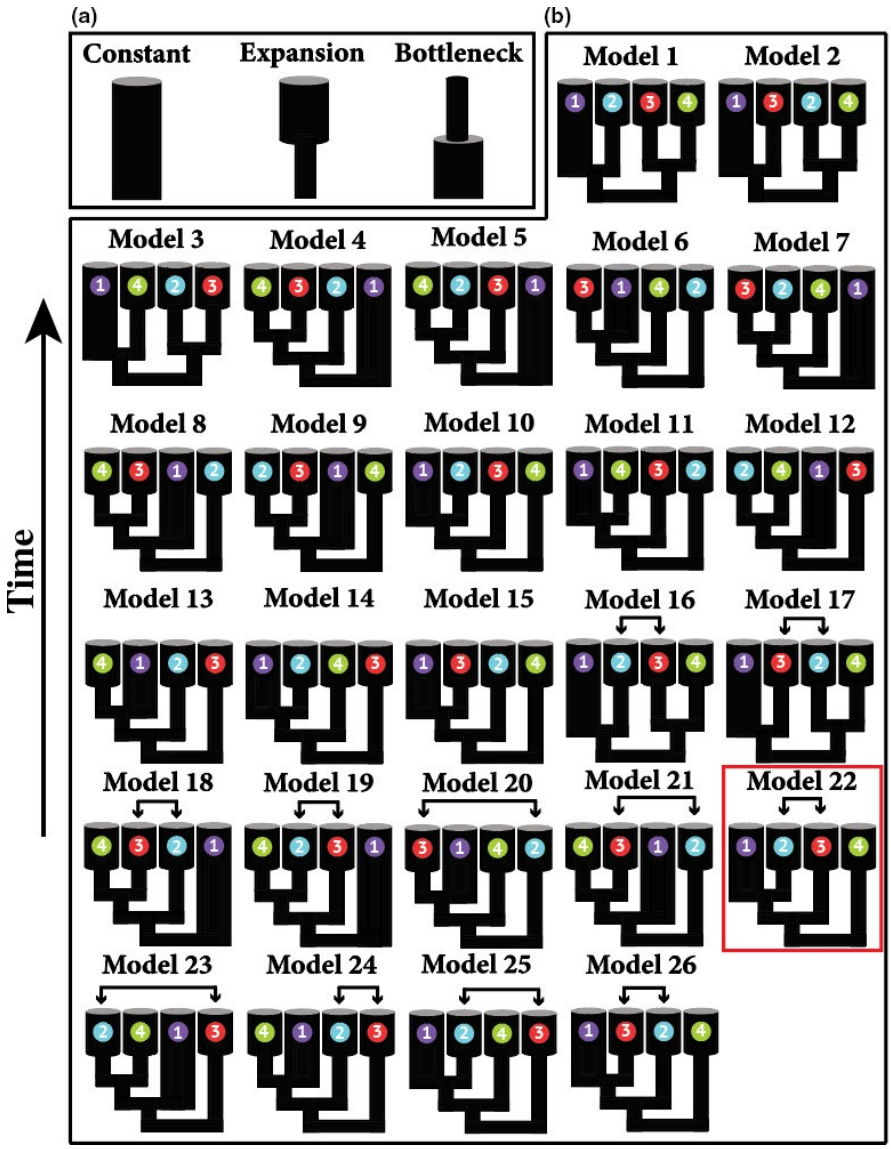

FIGURE 3 Representation of the models tested using convolutional neural networks. (a) Set of three models used to test population trajectory through time (constant, expansion, and bottleneck). (b) Set of 26 models used to test the evolutionary relationships and secondary contact of *Norops* ssp. Numbers and colours represent populations recovered in STRUCTURE analysis. Purple circle, *Norops brasiliensis* (population 1); blue circles, *N. brasiliensis* (population 2); red circles, *N. brasiliensis* (population 3); and green circles, *N. planiceps*. The best-supported model for CNN in the second part of comparison is marked by a red box. Time elapses from bottom to top in all models. Gene flow between populations 2 and 3 is represented by arrows (Models 16 to 26).

pooling layer (kernel = 3 × 1). We then flattened the output layer from the last pooling. Next, we created a fully connected layer with 100 neurons, followed by one with 25 neurons, and a final layer with three neurons, which correspond to our three demographic models (i.e., constant, expansion, and bottleneck; Figure 3). For all layers, we used rectified linear unit activation functions (ReLU), except for the last one where we used a softmax function. This function is a generalization of the logistic function and used for multiclass prediction. We compiled the CNN using the Adam optimization procedure (Kingma & Ba, 2015), a categorical cross-entropy loss function, and a mini-batch size of 100. We ran the CNN for 10 epochs, although without any improvement after three epochs. We did not include a dropout layer because of the lack of evidence of overfitting. Lastly, we trained the CNN using 80% of the simulated data sets and used the remaining 20% to evaluate model accuracy. We used the trained model to predict the model that probably generated the empirical data set. We built all CNNs with the Keras python library (https://keras.io).

We also evaluated the impact of sampling and the number of SNPs on CNN performance. Specifically, we conducted simulations under different sampling schemes (5, 10, and 25 individuals) and number of SNPs (100, 1000, and 5000) using the same CNN architecture and data summarization as described above.

2.5.2 | Testing evolutionary relationships and gene flow

In the second step of the analysis, we implemented a CNN architecture to assess the relationships among populations and gene flow between populations that showed evidence of admixture in STRUCTURE. We specified 26 demographic models consisting of a combination of 15 possible topologies along with scenarios of isolation or secondary contact after divergence that reflect our identification of individuals that are potentially admixed (Figure 3b). For example, because we recovered substantial admixture between populations 2 and 3, we included models with potential secondary contact between these populations (see Figure 2). We did not include models with secondary contact when populations 2 and 3 were sisters in the phylogenetic tree, because it was impractical to distinguish between isolation and secondary contact models in our preliminary runs. In addition, we did not include the possibility of gene flow between *N. planiceps* and population 2 of *N. brasiliensis* because given their disjunct geographic distributions (they are more than 1000 km apart) and knowing that lizards have low dispersal rates, gene flow between these lineages appears to be very unlikely. We used fastsimcoal2 to generate 10,000 data examples per model. As in the first part, we generated short DNA sequences of 1 bp for 500,000 independent loci in a way to simulate one SNP per locus. However, we only output the number of SNPs observed in the empirical data set. Individuals were randomly organized within each population, and SNPs were sorted based on allele frequency (higher to a lower frequency). Parameters in these models include ancestral and current population size, the time of population size changing, divergence time, migration rate, time of migration, and topology. Priors are available in Table S2. We converted alignments into images as described previously. In addition, because the relationship among populations is a key parameter in the models, images always presented populations in the same order: *N. brasiliensis* (population 1), *N. brasiliensis* (population 2), *N. brasiliensis* (population 3), and *N. planiceps*.

We used a simpler CNN architecture for the second part because it achieved a higher accuracy when compared to the CNN architecture used in the first part. We built the CNN using a two-dimensional convolution layer (kernel = 3 × 1; corresponding to three SNPs over one sample), and a two-dimensional maximum pooling layer (kernel =3 × 1). After that, we flattened the output layer from the pooling and generated a fully connected layer with 500 neurons using the hyperbolic tangent function (tanh) for all layers, followed by our final layer with 26 neurons, corresponding to different models, where we used the softmax function. We compiled our model similar to the first part: Adam optimization and categorical cross-entropy loss function, but we used a mini-batch size of 50. We trained the CNN for five epochs; but the model did not improve after the second epoch (i.e., accuracy did not decrease over epochs). Finally, CNN was trained using 80% of the simulated data set as training and the remaining 20% was used to evaluate the model. We used the trained model to predict the empirical data set. We used the python library Keras throughout to build the CNN. CNN architectures, for both parts, were selected after preliminary runs with varying combinations of activation functions (ReLU, tanh, sigmoid functions), numbers of convolutions layers and neurons, and kernel dimensions. For all CNNs, we evaluated the calibration of the softmax function by computing the absolute output probability of each simulation on each model on the test data set and assigned this value into five classes (0%–20%, 20%–40%, 40%–60%, 60%–80%, 80%–100%).

2.6 | Model selection in an approximate Bayesian computation framework

We also evaluated ABC performance for the second part of comparisons (from models 1 to 26). First, we used the R-package "PipeMaster" to perform 100,000 simulations for each model to generate summary statistics (Gehara et al., *in preparation*.; www.github.com/gehara/PipeMaster). PipeMaster is a user-friendly R-package that builds demographic models and then simulates data under the coalescent process using msABC (Pavlidis et al., 2010). Demographic models mirrored empirical data sets with respect to the number of populations, the number of individuals within each population, and the number of loci. Priors used to build the models were the same used to construct CNN models and are presented in Table S1. After simulations, we used the ABC approach to estimate model support using the "postpr" function implemented in the "abc" R package (Csilléry et al., 2012). We set the tolerance value to 1% and used the rejection and mnlogistic method

to compare models. We evaluate whether simulations produced summary statistics similar to the empirical data set using PCAs.

3 | RESULTS

3.1 | Genomic data processing

After genomic data processing, we obtained 4174 unlinked SNPs when all samples were combined, or 6860, 10,931, 9,396, and 12,048 unlinked SNPs for the three *N. brasiliensis* populations and *N. planiceps*, respectively. Because our CNN approach does not accommodate missing data, loci were required to be present in 100% of the samples.

3.2 | Population assignment

The STRUCTURE analysis recovered four geographically structured populations that correspond to *N. planiceps* and three populations within *N. brasiliensis* (hereafter population 1, population 2, and population 3; Figure 2). While *N. planiceps* is distributed in northern Amazonia, population 1 is found in an enclave of Seasonally Dry Tropical Forests within Cerrado. Population 2 is more widespread in Cerrado and population 3 is found in lowlands within Cerrado. In addition, population assignment analysis revealed a region of high admixture between population 2 and 3 (locality no. 9). Population assignments of $K = 3$ and $K = 5$ are shown in the supporting information (Figure S1).

3.3 | Demographic model selection

We inferred population expansion as the best demographic scenario for *N. planiceps*, population 2, and population 3 with a probability of 0.99, 0.59, and 1.0, respectively (Table 2). For population 2, the lower probability value is probably related to the unaccounted gene flow with population 3, which introduced a genetic variation that was not captured by the model. Conversely, for population 1, we found evidence of constant population size over time (probability = 0.985; Table 2). For all models within each population, the CNN model had a high accuracy when predicting the test set labels, reaching an overall accuracy higher than 98% for all models (Figure 4). Precision and recalls values were also higher than 98% and are shown in Table S3.

We found that CNN performance is influenced primarily by the number of sampled individuals and to a less extent by the number of SNPs (Table S4; Figure S2). The CNN model reached an accuracy of about 80% with five individuals (10 sequences for a diploid species) and 100 SNPs. However, CNN performance improved slightly when the number of SNPs increased to 1000 and 5000 (82% and 83%, respectively; Table S4). Conversely, with the increase of the number of individuals (to 10 and 25 individuals), all models reached an overall accuracy of over 92%, with a small improvement with a larger number of SNPs (Table S4). Our results showed that sample size is positively associated with CNN performance.

For the second part of model comparison, CNN recovered a single model (no. 22) as the best evolutionary scenario with a probability of 0.79 (Table 2; Figure 3). As expected, *N. planiceps* was recovered as the sister species of *N. brasiliensis* and population 1 is more closely related to population 2 than to population 3. In addition, we found evidence of secondary contact between populations 2 and 3. The second-best model (model 26; probability = 0.20; Table 2) is similar to the best model but, in this scenario, population 1 is more closely related to population 3. All other scenarios had a probability of less than 1% (Table 2). Even comparing complex evolutionary histories, our CNN showed a high average accuracy: 87%; range: 62%–99%; Figure 5). The posterior probabilities of ABC models were lower on both rejection and mnlogistic methods. The rejection method selected scenario 19 as the best models with a posterior probability of 18% (Table 2). The mnlogistic method performed better than the rejection method, selecting model 8 as the best evolutionary history (pp = 0.55), followed by model 16 (pp = 0.16). In model 8, *N. brasiliensis* was found paraphyletic *with N. planiceps*. Population 3 was recovered as the sister lineage of *N. planiceps* and these lineages formed another clade with population 1 with population 2 in a more external position. PCAs showed that most models produced summary statistics coincident with empirical data sets, indicating that the choice of priors was plausible (Figure S3). Overall, CNN produced results more robust than ABC in terms of accuracy (Figure 5 and Figure S4) and recall and precision (Table S5). Because of that, we discussed the results in the light of CNN findings. Overall, for both demographic parts, our calibration analysis showed that the CNNs are satisfactorily calibrated (Figures S5–S6).

4 | DISCUSSION

Our simulation testing suggests that a deep learning approach for phylogeographic model selection can be accurate for certain types of demographic processes. For example, the best CNN model had an accuracy of over 99% when testing for changes in effective population size through time in population 1 (i.e., constant, expansion, and bottleneck). We also found similar results for populations 2 and 3 (accuracy >99%). Model accuracy was slightly lower for *N. planiceps*, probably caused by the small number of samples for this species. Even though we generated fewer SNPs for population 1, this model achieved higher accuracy than the one for *N. placenips* probably because we had twice the number of samples for population 1. For models 1 to 26, the average accuracy was 87%. These models are more complex than those that deal only with changes in population size since we evaluated the evolutionary relationship of all populations and also included gene flow between populations 2 and 3, and the temporal divergence among populations. Nevertheless, the accuracy of the CNN model selection retained an accuracy similar to that seem in other approaches (below) while ABC was unable to identify a single best model (see Table 2).

TABLE 2 The probability of each model tested using convolutional neural networks (CNNs) and approximate Bayesian computation (ABC). Comparisons were first performed within part 1 only using CNNs, and subsequently, models in part 2 were constructed based on demographic scenario inferred in part 1. The best-fit model selected in each part is highlighted in bold

Part 1		Part 2			
				ABC posterior probability	
Model	CNN Probability	Model	CNN probability	Rejection	Mnlogistic
Population 1 - Constant	**0.98**	Model 1	0	0.01	0
Population 1 - Expansion	0.02	Model 2	0	0.01	0
Population 1 - Bottleneck	0	Model 3	0	0.01	0
		Model 4	0	0.09	0
		Model 5	0	0.10	0
Population 2 - Constant	0.41	Model 6	0	0	0
Population 2 - Expansion	**0.59**	Model 7	0	0.09	0
Population 2 - Bottleneck	0	Model 8	0	0	**0.56**
		Model 9	0	0	0
		Model 10	0.01	0	0
Population 3 - Constant	0	Model 11	0	0	0
Population 3 - Expansion	**1.0**	Model 12	0	0	0
Population 3 - Bottleneck	0	Model 13	0	0	0
		Model 14	0	0	0
		Model 15	0	0	0
N. planiceps - Constant	0.01	Model 16	0	0.16	0.21
***N. planiceps* - Expansion**	**0.99**	Model 17	0	0.15	0
N. planiceps - Bottleneck	0	Model 18	0	0.16	0.01
		Model 19	0	**0.18**	0.01
		Model 20	0	0	0
		Model 21	0	0	0
		Model 22	**0.79**	0.02	0.08
		Model 23	0	0	0
		Model 24	0	0	0
		Model 25	0	0	0
		Model 26	0.20	0	0.11

Our CNN implementation and ABC share many similarities, including the use of a simulations to generate new examples, given a demographic scenario and a set of priors. However, these approaches summarize the simulated data sets in different ways, leaning to different methods for comparison between empirical and simulated data. For ABC, a large number of summary statistics are usually calculated from the simulated data sets, e.g., Tajima's *D*, nucleotide diversity, F_{ST}, and Fu and Li's D. These summary statistics have traditionally been used in phylogeographic investigations, for example, Tajima's *D* has been used to detect deviations from constant population sizes caused by population expansions or bottlenecks and F_{STs} have measured the degree of differentiation among populations. This choice of summary statistics is subjective, with most studies choosing not to identify a subset of summary statistics that maximize model probability. Moreover, as model complexity increases, more summary statistics are required to describe the evolutionary history, for example, it is necessary to calculate at least seven pairwise population divergence metrics (e.g., F_{ST}) to describe the divergence sequence between the populations found in this study. Furthermore, the aforementioned statistics are pairwise metrics, and as such nonindependent, leading to consequences such as the "curse of dimensionality" (Beaumont, 2010; Beaumont et al., 2002) which leads to poor performance as the number of models grows. Our results mirror those from previous studies suggesting that ABC does not perform as well with large numbers of summary statistics and models (Pelletier & Carstens, 2014; Schrider & Kern, 2018; Smith et al., 2017).

Although it is beyond the scope of this study to compare different methods of phylogeographic model selection, the accuracy of the CNN approach used here can be placed into a broader context. It appears to be at least as accurate as other model selection approaches. For example, it appears to perform at least as well as PHRAPL, which summarizes the data using gene trees (Jackson et al.,

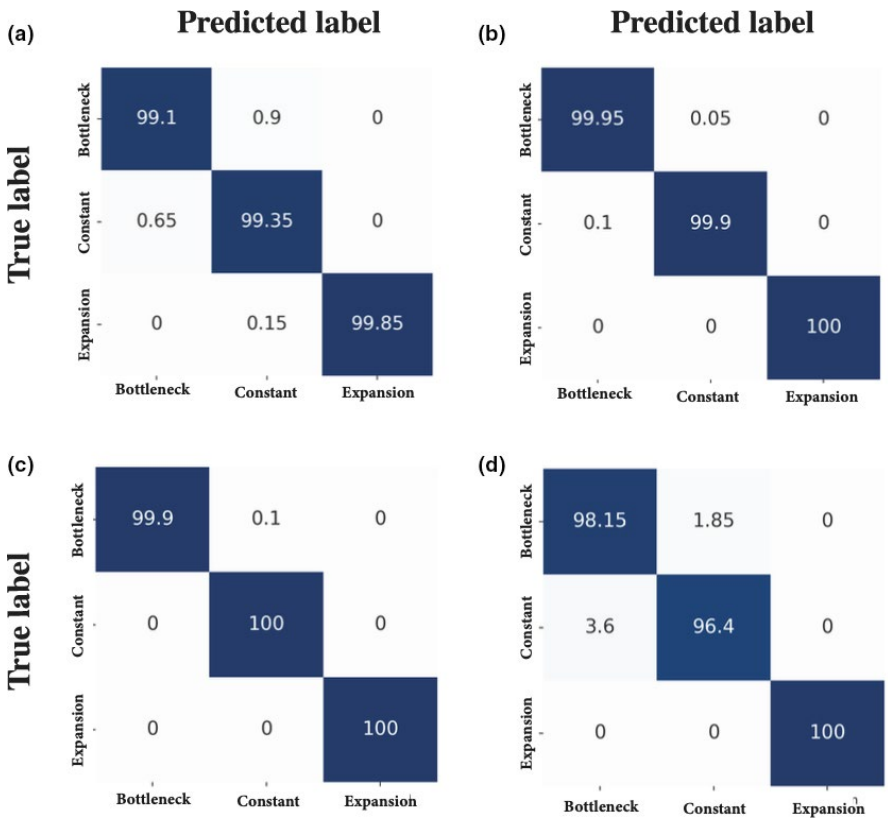

FIGURE 4 Confusion matrix measuring the accuracy of the trained CNN model on the test dataset to detect demographic changes through time. Numbers represent percentages, which were calculated based on 2,000 images for each model. (a) *Norops brasiliensis* (population 1), (b) *N. brasiliensis* (population 2), (c) *N. brasiliensis* (population 3), and (d) *N. planiceps*

2017) and, because there is more gene-tree to species-tree discordance at shallow levels of population divergence, becomes more accurate as population divergence increases. Similarly, for CNNs, the model accuracy decreases as the divergence between populations decrease, a phenomenon which has been attributed to incomplete lineage sorting (ILS; Blischak et al., 2020). The accuracy estimated here is also similar to that of other machine learning approaches to phylogeographic model selection. For example, Smith et al. (2017) proposed a random forest approach to test 15 evolutionary scenarios for a land snail endemic to the Pacific Northwest of North America and also compared the random forest classifier with ABC. Their overall error rates using random forest were 7.67% (range: 0%–42%) and ~30% for ABC. While our overall error for CNNs in step 2 was 13%, we noticed that most misclassification was between models that only differed on the presence or absence of secondary contact. Since Smith et al. (2017) did not include gene flow in their tested models, we subset our models and trained a CNN only with isolation models (models 1 to 15) in order to estimate a comparable error rate of 1.5% (0.75%–3%; Figure S7), which is lower than that described by Smith et al. (2017). In a more recent study, Smith and Carstens (2020) applied random forest to the reticulate taildropper slug (*Prophysaon andersoni*) and found an average error of 5.2% when comparing 208 demographic models. These results show that CNN has an accuracy comparable to the best results reported for other methods (i.e., ABC with random forest). Unfortunately, the comparison between CNN and AIC-based methods (such as PHRAPL) is not as straightforward because they use different frameworks to measure model performance. In particular, AIC-based approaches to model selection lack the built-in approach for assessing model accuracy (i.e., identifiability) that deep learning approaches such as CNN and ABC with random forest include.

One advantage of CNNs is that researchers are absolved of the requirement to summarize their data using summary statistics. Since a set of statistics exists that is probably best used with a particular demographic history, this is particularly challenging for investigations into non-model systems. In our system (*N. planiceps* and *N. brasiliensis*) and others, there is a scarcity of a priori ecological and evolutionary information that limits the ability of researchers to specify a small set of candidate models and choose appropriate summary statistics. In such a scenario, approaches such as CNNs, PHRAPL, and delimitR offer the potential to compare among a large number of competing alternatives models without the need to make choices that are likely to influence the outcome. That is not to say that CNN approaches are flexible, as the image-based nature of the analysis enables the data to be summarized in different ways. For example, Blischak et al. (2020) used CNNs to detect hybridization in simulated and an empirical system from *Heliconius* butterflies. They simulated chromosome-scale data for four species and generated images based on the pairwise Nei's genetic distance among populations and found that this approach was more accurate than those based on introgression-specific summary statistics.

On disadvantage of the approach used here is that it was computationally more demanding than the one proposed by Blischak et al. (2020). It requires an average of 2 s to run the simulation in fastsimcoal2 and 8 s to process the image (~10 s from simulation to generate an image). Since we simulated 10,000 examples per model,

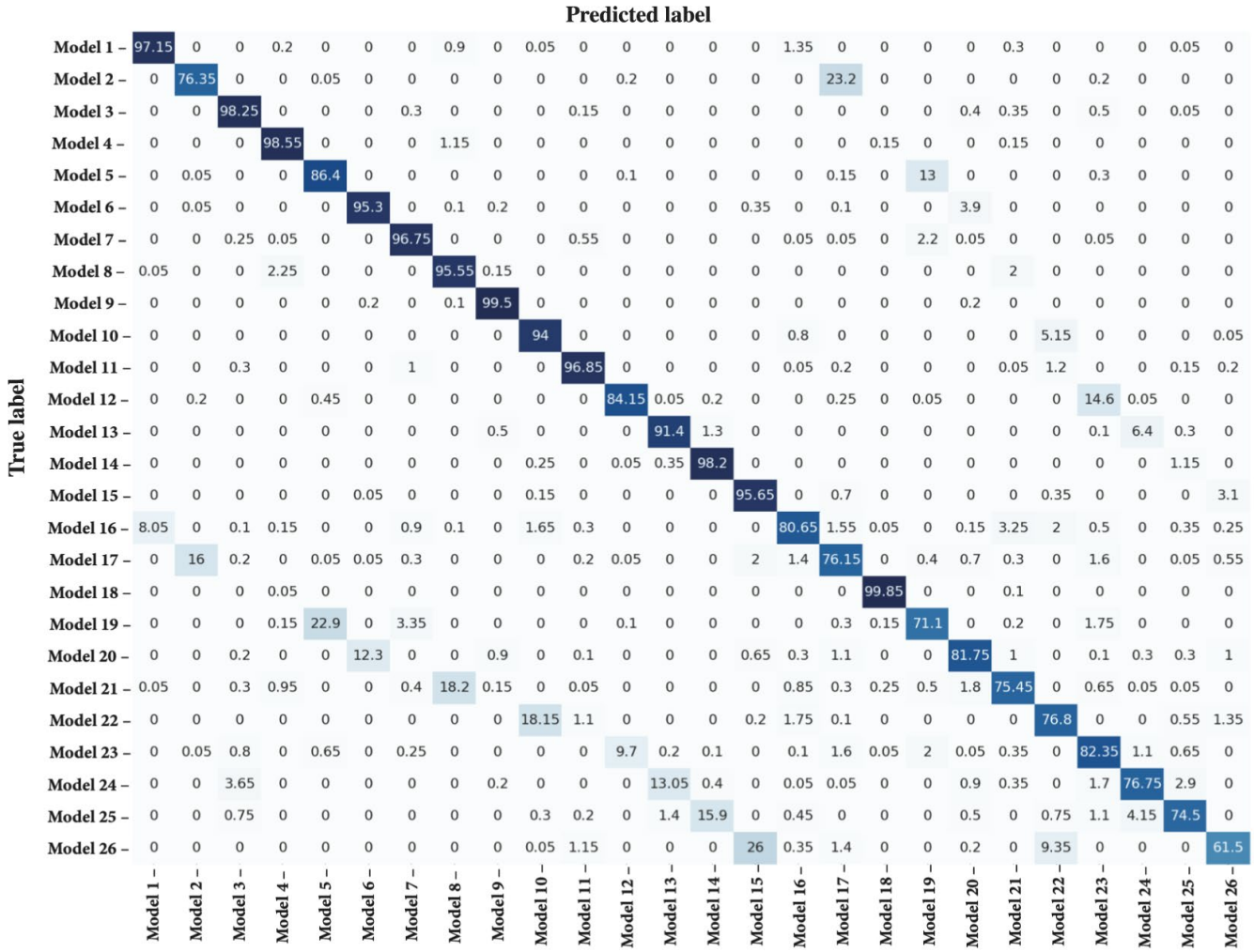

FIGURE 5 Confusion matrices measuring the accuracy of the trained CNN model on the test data set of 26 phylogeographic models. Numbers represent percentages which were calculated based on 2,000 images for each model

approximately 27 h are required to simulate the images that correspond to one scenario. It then requires an additional 10 h to run one epoch in the comparison among 26 models (208,000 training images and 52,000 test images), but this time can be optimized by using Graphical Processing Unit (GPU) instead of Central Processing Unit (CPU). Although the simulation and CNN were performed using the resources provided by the Ohio Supercomputer Center, we used a Mac mini (1.6 GHz Intel Core i5, 8 GB RAM, 2 cores) to generate these reference values to provide context for potential users of this approach who do not have access to supercomputing centers. By far the biggest computational hurdle was the number of stored images in the supercomputer, as our analysis used a total of 380,000 images totalling 7.5 GB.

4.1 | Evolutionary history of South American lizards

Pleistocene climate change has been proposed as one of the main drivers of speciation at higher latitudes (Burbrink et al., 2016; Hewitt, 2000, 2004). The Pleistocene refugia hypothesis (PRH) posits that species had to inhabit favorable refugia to persist and thrive under the new environmental conditions (Vanzolini & Williams, 1970). In South America, Haffer (1969) and Vanzolini and Williams (1970) almost simultaneously proposed the PRH to explain patterns of species diversity and distribution in the Amazon rainforest, where climate oscillations putatively led to a series of contraction events of rainforests and expansions of dry vegetations during glacial periods, which would enable allopatric speciation of the associated biota. While this has been a popular hypothesis, many investigations have dismissed the Pleistocene refugia model based on multiple biological and paleoenvironmental sources of evidence (Bush & Oliveira, 2006; Lessa et al., 1997; Thomé et al., 2010; Wang et al., 2017). Cheng et al. (2013), based on speleothem oxygen isotope records, proposed an alternative speciation model for the Late Pleistocene in South America, in which a quasi-dipolar precipitation pattern during the Pleistocene would impact biodiversity differently in western and eastern Amazonia. In eastern Amazon, which is more connected to the historical and current climate in the Cerrado, this model posits that the interleaved periods

of wet and dry climates during the last 250 thousand years (kyr) were desynchronized with those in western Amazonia, resulting in habitat fragmentation that isolated species previously broadly distributed and led to decreased gene flow and increased genetic differentiation. Community-level analyses have suggested that the model is broadly applicable (e.g Gehara et al., 2017; Silva et al., 2019). In contrast to the more stable climate in western Amazon, which is hypothesized to have generated the observed higher levels of biodiversity across multiple taxonomic groups and probably population stability through time. Our phylogeographic model selection results support the quasi-dipolar scenario of Cheng et al. (2013). For example, the population expansion inferred in *N. planiceps* and populations 2 and 3 of *N. brasiliensis* are consistent in timing and magnitude with the predictions of Cheng et al. (2013). Population 1 of *N. brasiliensis*, which was inferred to be constant in size, is located in an enclave of Caatinga within Cerrado (Paranã valley). The Caatinga is the largest nucleus of Seasonally Dry Tropical Forests (SDTF) and characterized by xeric vegetation, high seasonality, and unpredictable droughts. It is hypothesized that the climatic oscillations during the Pleistocene led the expansion and connection of now disjunct SDTFs (the Pleistocenic Arc Hypothesis - PAH; Prado & Gibbs, 1993; Pennington et al., 2000). This hypothesis is supported by the disjunct distribution of plants and animals as well as molecular data (Lanna et al., 2018; Pennington et al., 2000; Werneck & Colli, 2006). However, the exact time of the PAH is uncertain and the SDTFs could have expanded earlier, during the transition between Pliocene and Pleistocene, and have fragmented before the Last Glacial Maximum (Werneck et al., 2011), which could explain the stable population sizes we recovered in the longer term. In addition to climatic oscillations, the pattern of diversification found by our study mirrors the current taxonomic status of both species, though we found a hidden genetic diversity within *N. brasiliensis*. The pattern of divergence among lineages within *N. brasiliensis* follows a southeast-northwest pattern of differentiation, which is shared with other squamates in Cerrado (Guarnizo et al., 2016; Prado et al., 2012; Santos et al., 2014). Although the causes of this southeast-northwest pattern are unknown, it is hypothesized that this pattern was probably driven by landscape features (e.g., topography, rivers) and climatic conditions that have been acting over time.

4.2 | Conclusion

Deep learning techniques have been successfully used in fields such as medical sciences (Mobadersany et al., 2018) and agriculture (Kamilaris & Prenafeta-Boldú, 2018), but their usage in evolutionary biology has just begun (see Blischak et al., 2020; Flagel et al., 2019; Sanchez et al., 2020; Schrider & Kern, 2018; Torada et al., 2019). Our results showed that CNNs can be an effective and promising approach for phylogeographic model selection. We showed that a DNA alignment can be used as the source of comparison of a large number of models, without the need of genetic summary statistics. Also, our approach revealed a complex evolutionary scenario among lizards distributed in contrasting environments in South America, which involves hidden genetic diversity, gene flow between nonsister populations, and changes in effective population size through time. Finally, we encourage future investigations to compare the relative performance of different approaches for phylogeographic model selection and assess how key demographic parameters (e.g., divergence times, migration rates, changes in population size through time, etc.) affect the accuracy of different approaches.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ACKNOWLEDGEMENTS

We thank members of the Carstens Laboratory, Lex Flagel, Matteo Fumagalli, and one anonymous reviewer for comments and suggestions on the manuscript. We also thank Lisa N. Barrow and Megan L. Smith for laboratory assistance. We thank curators and managers of the INPA-H and INPA-HT (C. Ribas, M. Freitas, and A. A. Silva) for granting and processing samples under their care. We thank Ohio Supercomputer Center for providing computational resources via a grant to BCC (PAA0202). We thank the National Science Foundation for supporting this work via a grant to BCC (DEB-1831319). EMF thanks the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for his doctoral fellowship (process no. 88881.170016/2018). FPW thanks Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for her productivity fellowship (process no. 305535/2017-0). GRC thanks CAPES, CNPq, Fundação de Apoio à Pesquisa do Distrito Federal – FAPDF and the USAID's PEER program under cooperative agreement AID-OAA-A-11-00012 for financial support.

AUTHOR CONTRIBUTIONS

Emanuel M. Fonseca and Bryan C. Carstens conceived the ideas and designed methodology. Emanuel M. Fonseca conducted the laboratory work and conducted the analyses. All authors interpreted the results and participated in the writing of the manuscript and gave final approval for submission.

DATA AVAILABILITY STATEMENT

Scripts used in the convolutional neural network analyses have been made available on GitHub https://github.com/emanuelmfonseca/Model_selection_using_CNN.

ORCID

Emanuel M. Fonseca https://orcid.org/0000-0002-2952-8816
Guarino R. Colli https://orcid.org/0000-0002-2628-5652
Fernanda P. Werneck https://orcid.org/0000-0002-8779-2607
Bryan C. Carstens https://orcid.org/0000-0002-1552-227X

REFERENCES

Avila-Pires, T. C. S. (1995). Lizards of Brazilian Amazonia (Reptilia: Squamata). *Zoologische Verhandelingen*, *299*, 1–706.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, *41*, 379–406. https://doi.org/10.1146/annurev-ecolsys-102209-144621.

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, *162*(4), 2025–2035. https://doi.org/10.1111/j.1937-2817.2010.tb01236.x

Beerli, P., & Palczewski, M. (2010). Unified framework to evaluate panmixia and migration direction among multiple sampling locations. *Genetics*, *185*(1), 313–326. https://doi.org/10.1534/genetics.109.112532

Blischak, P. D., Barker, M. S., & Gutenkunst, R. N. (2020). Chromosome-scale inference of hybrid speciation and admixture with convolutional neural networks. *BioRxiv*, 2020.06.29.159673. https://doi.org/10.1101/2020.06.29.159673

Burbrink, F. T., Chan, Y. L., Myers, E. A., Ruane, S., Smith, B. T., & Hickerson, M. J. (2016). Asynchronous demographic responses to Pleistocene climate change in Eastern Nearctic vertebrates. *Ecology Letters*, *19*(12), 1457–1467. https://doi.org/10.1111/ele.12695

Bush, M. B., & de Oliveira, P. E. (2006). The rise and fall of the Refugial Hypothesis of Amazonian speciation: a paleoecological perspective. *Biota Neotropica*, *6*(1), 1–17. https://doi.org/10.1590/s1676-06032006000100002

Carstens, B. C., Morales, A. E., Jackson, N. D., & O'Meara, B. C. (2017). Objective choice of phylogeographic models. *Molecular Phylogenetics and Evolution*, *116*(April), 136–140. https://doi.org/10.1016/j.ympev.2017.08.018

Carstens, B. C., Stoute, H. N., & Reid, N. M. (2009). An information-theoretical approach to phylogeography. *Molecular Ecology*, *18*(20), 4270–4282. https://doi.org/10.1111/j.1365-294X.2009.04327.x

Cheng, H., Sinha, A., Cruz, F. W., Wang, X., Edwards, R. L., D'Horta, F. M., Ribas, C. C., Vuille, M., Stott, L. D., & Auler, A. S. (2013). Climate change patterns in Amazonia and biodiversity. *Nature Communications*, *4*, 1411. https://doi.org/10.1038/ncomms2415

Csilléry, K., François, O., & Blum, M. G. B. (2012). Abc: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, *3*(3), 475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x

Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*(2), 359–361. https://doi.org/10.1007/s12686-011-9548-7

Eaton, D. A. R., & Overcast, I. (2020). ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics*, *36*(8), 2592–2594. https://doi.org/10.1093/bioinformatics/btz966

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*, *6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, *14*(8), 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., & Foll, M. (2013). Robust Demographic Inference from Genomic and SNP Data. *PLoS Genetics*, *9*(10), https://doi.org/10.1371/journal.pgen.1003905

Fagundes, N. J. R., Ray, N., Beaumont, M., Neuenschwander, S., Salzano, F. M., Bonatto, S. L., & Excoffier, L. (2007). Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(45), 17614–17619. https://doi.org/10.1073/pnas.0708280104

Flagel, L., Brandvain, Y., & Schrider, D. R. (2019). The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Molecular Biology and Evolution*, *36*(2), 220–238. https://doi.org/10.1093/molbev/msy224

Gehara, M., Garda, A. A., Werneck, F. P., Oliveira, E. F., da Fonseca, E. M., Camurugi, F., Magalhães, F. D. M., Lanna, F. M., Sites, J. W., Marques, R., Silveira-Filho, R., São Pedro, V. A., Colli, G. R., Costa, G. C., & Burbrink, F. T. (2017). Estimating synchronous demographic changes across populations using hABC and its application for a herpetological community from northeastern Brazil. *Molecular Ecology*, *26*(18), 4756–4771. https://doi.org/10.1111/mec.14239

Guarnizo, C. E., Werneck, F. P., Giugliano, L. G., Santos, M. G., Fenker, J., Sousa, L., D'Angiolella, A. B., dos Santos, A. R., Strüssmann, C., Rodrigues, M. T., Dorado-Rodrigues, T. F., Gamble, T., & Colli, G. R. (2016). Cryptic lineages and diversification of an endemic anole lizard (Squamata, Dactyloidae) of the Cerrado hotspot. *Molecular Phylogenetics and Evolution*, *94*, 279–289. https://doi.org/10.1016/j.ympev.2015.09.005

Haffer, J. (1969). Speciation in Amazonian Forest Birds. *Science*, *165*(3889), 131–137.

Haller, B. C., & Messer, P. W. (2019). SLiM 3: Forward genetic simulations beyond the Wright-Fisher model. *Molecular Biology and Evolution*, *36*(3), 632–637.

Hewitt, G. (2000). The genetic legacy of the Quaternary ice ages. *Revue Des Maladies Respiratoires*, *405*(4), 907–913.

Hewitt, G. (2004). Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*(1442), 183–195. https://doi.org/10.1098/rstb.2003.1388

Hey, J., Chung, Y., & Sethuraman, A. (2015). On the occurrence of false positives in tests of migration under an isolation-with-migration model. *Molecular Ecology*, *24*(20), 5078–5083. https://doi.org/10.1111/mec.13381

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, *18*(2), 337–338. https://doi.org/10.1093/bioinformatics/18.2.337

Jackson, N. D., Morales, A. E., Carstens, B. C., & O'Meara, B. C. (2017). PHRAPL: Phylogeographic Inference Using Approximate Likelihoods. *Systematic Biology*, *66*(6), 1045–1053. https://doi.org/10.1093/sysbio/syx001

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, *23*(14), 1801–1806. https://doi.org/10.1093/bioinformatics/btm233

Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, *147*, 70–90.

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLoS Computational Biology*, *12*(5), 1–22. https://doi.org/10.1371/journal.pcbi.1004842

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv*, *1412*, 6980.

Knowles, L. L., Carstens, B. C., & Keat, M. L. L. (2007). Coupling Genetic and Ecological-Niche Models to Examine How Past Population Distributions Contribute to Divergence. *Current Biology*, *17*(11), 940–946. https://doi.org/10.1016/j.cub.2007.04.033

Knowles, L. L., & Maddison, W. P. (2002). Statistical Phylogeography. *Molecular Ecology*, *11*(12), 2623–2635.

Koopman, M. M., & Carstens, B. C. (2010). Conservation genetic inferences in the carnivorous pitcher plant Sarracenia alata (Sarraceniaceae). *Conservation Genetics*, *11*(5), 2027–2038. https://doi.org/10.1007/s10592-010-0095-7

Korneliussen, T. S., Moltke, I., Albrechtsen, A., & Nielsen, R. (2013). Calculation of Tajima's *D* and other neutrality test statistics from

low depth next-generation sequencing data. *BMC Bioinformatics*, *14*, 289.

Lanna, F. M., Werneck, F. P., Gehara, M., Fonseca, E. M., Colli, G. R., Sites, J. W., Rodrigues, M. T., & Garda, A. A. (2018). The evolutionary history of Lygodactylus lizards in the South American open diagonal. *Molecular Phylogenetics and Evolution*, *127*(August), 638–645. https://doi.org/10.1016/j.ympev.2018.06.010

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lessa, E. P., Van Valkenburgh, B., & Fariña, R. A. (1997). Testing hypotheses of differential mammalian extinctions subsequent to the Great American Biotic Interchange. *Palaeogeogrdphy, Palaeoclimatology, Palaeoecology*, *135*, 157–162.

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Velázquez Vega, J. E., Brat, D. J., & Cooper, L. A. D. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(13), E2970–E2979. https://doi.org/10.1073/pnas.1717139115

Morales, A. E., Jackson, N. D., Dewey, T. A., O'Meara, B. C., & Carstens, B. C. (2017). Speciation with gene flow in North American *Myotis* bats. *Systematic Biology*, *66*, 440–452. https://doi.org/10.1093/sysbio/syw100

Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, *403*, 853–858.

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Pavlidis, P., Laurent, S., & Stephan, W. (2010). MsABC: A modification of Hudson's ms to facilitate multi-locus ABC analysis. *Molecular Ecology Resources*, *10*(4), 723–727. https://doi.org/10.1111/j.1755-0998.2010.02832.x

Pelletier, T. A., & Carstens, B. C. (2014). Model choice for phylogeographic inference using a large set of models. *Molecular Ecology*, *23*(12), 3028–3043. https://doi.org/10.1111/mec.12722

Pennington, T. R., Prado, D. E., & Pendry, C. A. (2000). Neotropical seasonally dry forests and Quaternary vegetation changes. *Journal of Biogeography*, *27*(2), 261–273. https://doi.org/10.1046/j.1365-2699.2000.00397.x

Prado, C. P. A., Haddad, C. F. B., & Zamudio, K. (2012). Cryptic lineages and Pleistocene population expansion in a Brazilian Cerrado frog. *Molecular Ecology*, *21*(4), 921–941. https://doi.org/10.1111/j.1365-294X.2011.05409.x

Prado, D. E., & Gibbs, P. E. (1993). Patterns of species distributions in the dry seasonal forests of South America. *Annals of the Missouri Botanical Garden*, *80*(4), 902–927.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. (1999). Population growth of human Y chromosomes: A study of y chromosome microsatellites. *Molecular Biology and Evolution*, *16*(12), 1791–1798. https://doi.org/10.1093/oxfordjournals.molbev.a026091

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945–959. https://doi.org/10.1007/s10681-008-9788-0

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*(6), 859–866. https://doi.org/10.1093/bioinformatics/btv684

Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep Learning for Medical Image Processing: Overview, Challenges and the Future. In N. Dey, A. S. Ashour, & S. Borra (Eds.), *Classification in BioApps: Automation of Decision Making, Lecture Notes in Computational Vision and Biomechanics 26*, 323–350. Cham: Springer. https://doi.org/10.1007/978-3-319-65981-7_12

Ribeiro, M. A. (2015). Catalogue of distribution of lizards (Reptilia: Squamata) from the Brazilian Amazonia. I. Dactyloidae, Hoplocercidae, Iguanidae, Leiosauridae, Polychrotidae, Tropiduridae. *Zootaxa*, *3983*(1), 1–110. https://doi.org/10.11646/zootaxa.3983.1.1

Sanchez, T., Cury, J., Charpiat, G., & Jay, F. (2020). Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13224. [Early view]

Santos, M. G., Nogueira, C., Giugliano, L. G., & Colli, G. R. (2014). Landscape evolution and phylogeography of Micrablepharus atticolus (Squamata, Gymnophthalmidae), an endemic lizard of the Brazilian Cerrado. *Journal of Biogeography*, *41*(8), 1506–1519. https://doi.org/10.1111/jbi.12291.

Schrider, D. R., & Kern, A. D. (2018). Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics*, *34*(4), 301–312. https://doi.org/10.1016/j.tig.2017.12.005.

Silva, S. M., Peterson, A. T., Carneiro, L., Burlamaqui, T. C. T., Ribas, C. C., Sousa-Neves, T., Miranda, L. S., Fernandes, A. M., d'Horta, F. M., Araújo-Silva, L. E., Batista, R., Bandeira, C. H. M. M., Dantas, S. M., Ferreira, M., Martins, D. M., Oliveira, J., Rocha, T. C., Sardelli, C. H., Thom, G., … Aleixo, A. (2019). A dynamic continental moisture gradient drove Amazonian bird diversification. *Science*. *Advances*, *5*(7), eaat5752. https://doi.org/10.1126/sciadv.aat5752.

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, *195*, 693–702.

Smith, M. L., & Carstens, B. C. (2020). Process-based species delimitation leads to identification of more biologically relevant species. *Evolution*, *74*(2), 216–229. https://doi.org/10.1111/evo.13878.

Smith, M. L., Ruffley, M., Espíndola, A., Tank, D. C., Sullivan, J., & Carstens, B. C. (2017). Demographic model selection using random forests and the site frequency spectrum. *Molecular Ecology*, *26*(17), 4562–4573. https://doi.org/10.1111/mec.14223.

Suvorov, A., Hochuli, J., & Schrider, D. R. (2020). Accurate inference of tree topologies from multiple sequence alignments using deep learning. *Systematic Biology*, *69*(2), 221–233. https://doi.org/10.1093/sysbio/syz060.

Thomé, M. T. C., & Carstens, B. C. (2016). Phylogeographic model selection leads to insight into the evolutionary history of four-eyed frogs. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(29), 8010–8017. https://doi.org/10.1073/pnas.1601064113.

Thomé, M. T. C., Zamudio, K. R., Giovanelli, J. G. R., Haddad, C. F. B., Baldissera, F. A., & Alexandrino, J. (2010). Phylogeography of endemic toads and post-Pliocene persistence of the Brazilian Atlantic Forest. *Molecular Phylogenetics and Evolution*, *55*(3), 1018–1031. https://doi.org/10.1016/j.ympev.2010.02.003

Torada, L., Lorenzon, L., Beddis, A., Isildak, U., Pattini, L., Mathieson, S., & Fumagalli, M. (2019). ImaGene: a convolutional neural network to quantify natural selection from genomic data. *BMC Bioinformatics*, *20*(S9), 337.

Vanzolini, P., & Williams, E. (1970). South american anoles: the geographic differentiation and evolution of the anolis Chrysolepis species group (Sauria, Iguanidae). *Arquivos De Zoologia*, *19*(3–4), 125–298.

Wang, X., Edwards, R. L., Auler, A. S., Cheng, H., Kong, X., Wang, Y., Cruz, F. W., Dorale, J. A., & Chiang, H. W. (2017). Hydroclimate changes across the Amazon lowlands over the past 45,000 years. *Nature*, *541*(7636), 204–207. https://doi.org/10.1038/nature20787

Werneck, F. P., & Colli, G. R. (2006). The lizard assemblage from seasonally dry tropical forest enclaves in the Cerrado biome, Brazil, and its association with the Pleistocenic Arc. *Journal of Biogeography*, *33*(11), 1983–1992. https://doi.org/10.1111/j.1365-2699.2006.01553.x

Werneck, F. P., Costa, G. C., Colli, G. R., Prado, D. E., & Sites, J. W. (2011). Revisiting the historical distribution of Seasonally Dry Tropical Forests: New insights based on palaeodistribution modelling and palynological evidencegeb. *Global Ecology and Biogeography*, *20*(2), 272–288. https://doi.org/10.1111/j.1466-8238.2010.00596.x

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Fonseca EM, Colli GR, Werneck FP, Carstens BC. Phylogeographic model selection using convolutional neural networks. *Mol Ecol Resour*. 2021;21: 2661–2675. https://doi.org/10.1111/1755-0998.13427