Phylogeographic model selection using convolutional neural networks

Emanuel Masiero da Fonseca[1*] (https://orcid.org/0000-0002-2952-8816)

Guarino R. Colli[2] (https://orcid.org/0000-0002-2628-5652)

Fernanda P. Werneck[3] (https://orcid.org/0000-0002-8779-2607)

Bryan C. Carstens[1] (https://orcid.org/0000-0002-1552-227X)

[1] Department of Evolution, Ecology and Organismal Biology, The Ohio State University, 318 W. 12th Ave, Columbus, OH 43210

[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

[*]Corresponding author: emanuelmfonseca@gmail.com; Department of Evolution, Ecology and Organismal Biology, The Ohio State University, 318 W. 12th Ave, Columbus, OH 43210

Running Title: Phylogeography and deep learning

Abstract

The field of phylogeography has evolved rapidly in terms of the analytical toolkit to analyze the ever-increasing amounts of genomic data. Despite substantial advances, researchers have not fully explored all potential analytical tools to tackle the challenge posed by the huge size of genomic datasets. For example, deep learning techniques, such as convolutional neural networks (CNNs), widely employed in image and video classification, are largely unexplored for phylogeographic model selection. In non-model organisms, the lack of information about their ecology, natural history, and evolution can lead to uncertainty about which set of demographic models should be considered. Here we investigate the utility of CNNs for assessing a large number of competing phylogeographic models using South American lizards as an example, and approximate Bayesian computation (ABC) to contrast the performance of CNNs. First, we evaluated three demographic scenarios (constant, expansion, and bottleneck) for each of four recovered lineages and found that the overall model accuracy was higher than 98% for all lineages. Next, we evaluated a set of 26 models that accounted for evolutionary relationships, gene flow, and changes in effective population size among these lineages and recovered an overall accuracy of 87%. In contrast, ABC was unable to single out a best fit model among 26 competing models. Finally, we used the CNN model to investigate the evolutionary history of two South American lizards. Our results indicate the presence of hidden genetic diversity, gene flow between non-sister populations, and changes in effective population sizes through time, likely in response to Pleistocene climatic oscillations. Our results demonstrate that CNNs can be easily and usefully incorporated into the phylogeographer's toolkit.

Keywords: CNNs, deep learning, machine learning, *Norops* spp., phylogeography

Introduction

One key research goal of phylogeographic research has been to investigate how historical processes have shaped genetic variation across geographic space. In the early years of phylogeography, interpretations were highly qualitative and largely based on gene genealogies and the geographic distribution of the haplotypes. Because of their descriptive nature, such phylogeographic 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), where researchers often interpreted new results in a manner that supported previous findings (Carstens et al., 2009). As the field matured, researchers recognized the importance of statistical approaches that explicitly incorporate uncertainty to draw meaningful conclusions about species' evolutionary history. Therefore, the identification of statistical models relevant for data analysis is a crucial step of any model-based phylogeographical investigation.

Phylogeographers 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). In the first approach, researchers use a particular evolutionary model to estimate a set of parameters of interest based on their expertise about the organism and its environment. Although this approach has enabled the evaluation of complex evolutionary processes, it can lead to unreliable estimates of the parameters of interest due to model misspecification (Koopman & Carstens, 2010). Biological intuition often drives the choice of the analytical framework(s) used to analyze the data. For example, researchers may choose to analyze 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 inferences can be misleading (Beerli & Palczewski, 2010; Hey et al., 2015). Unfortunately, the estimation of many evolutionary processes eventually becomes intractable in a likelihood framework (Beaumont, 2010; Beaumont et al., 2002), Therefore, 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. 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 used by researchers to conduct phylogeographic model selection. Software such as *ms* (Hudson, 2002), *msprime* (Kelleher et al., 2016), and *fastsimcoal2* (Excoffier et al., 2013) can be used to conduct coalescent simulations under customized demographic models that can approximate the details of almost any empirical system. After the simulation procedure, empirical and simulated datasets can be statistically evaluated using a variety of methods, including hypothesis testing (e.g., Knowles et al., 2007), Approximate Bayesian Computation (ABC; e.g., Fagundes et al., 2007), information theory (e.g., Carstens et al., 2009), 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 isn't 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). Such approaches are effective at considering large numbers of models, but at the expense of parameter estimation. 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 of these approaches varies, 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

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., pre-classified data) are available for all the samples. SML involves (i) training a predictive model using a subset of a labeled dataset, (ii) evaluating the model using the remaining portion of the labeled dataset, and (iii) using the now-trained model to predict new, unlabeled 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 dataset. 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. 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 labeled images and train a model to predict the content of new images, one potential advantage of this approach is that data do not need to be summarized using standard genetic summary statistics or a SFS. Rather, prediction can be made directly from the alignment containing the genetic variation from sampled individuals (Flagel et al., 2019). CNNs have been used to address a range of biological questions, from detecting selective sweeps (Flagel et al., 2019) to 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 labeled examples (i.e., DNA alignments), converted to a black and white image by labeling the major allele as the ancestral state and the minor allele as a derived state. After training the model using 80% of the labeled data and evaluating its performance using the remaining 20% of the data, we compare the performance of CNNs and ABC to inquire about the evolutionary history of two species of lizards, from contrasting environments in South America.

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 labeled images (input data) to build a model to differentiate among the various labels (e.g., a model able to differentiate between images of cats and dogs). First, the input images (Figure 1a) are transformed into arrays (Figure 1b), and then a convolution operation is performed by multiplying each value in the array by a learnable weight within a kernel (Figure 1c). After the convolution operation, the arrays are converted into a feature map (Figure 1d) where each value is passed through a non-linear 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 1e). 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).

Material and Methods

South American lizards as a case study

We used lizards as a case study to assess the usefulness of CNNs for phylogeographic model selection. Lizards are a diverse group of vertebrates, recognized as model organisms for evolutionary studies due to low thermal tolerance, relatively short generation times, and low dispersal rates (Camargo et al., 2010). For this study, we selected the sister species *Norops brasiliensis* and *N. planiceps* as targets for objective model selection. Little is known about 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 in the Cerrado and enclaves of Cerrado in Amazonia (Figure 2; Avila-Pires, 1995; Ribeiro, 2015) (Figure 2). While *N. planiceps* is also terrestrial and diurnal, this species is endemic to northern Amazonia, inhabiting mainly “terra firme” forests, which are not periodically flooded (Figure 2; Avila-Pires, 1995; Ribeiro, 2015).

Amazonia and Cerrado are the largest Brazilian biomes, which together originally covered about 73% of the Brazilian territory. Amazonia is a region predominantly covered by tropical rainforests, whereas the Cerrado is a world hotspot priority for conservation (Myers et al., 2000), 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.

Sampling and data collection

We obtained 61 tissue samples; 52 from *N. brasiliensis* (nine localities) and 9 from *N. planiceps* (five localities; Figure 1). Samples were obtained 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).

We extracted DNA from liver or muscle tissues using E.Z.N.A. Tissue DNA Kit and 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: (1) initial denaturation at 72 °C for 5 min; (2) 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; (3) final extension at 72 °C for 5 min. Then, we quantified PCR products using the BR DNA Qubit Quantification Kit. To select DNA fragments of 200–500 bp, we used the Blue Pippin Prep and carried out sequencing at the Ohio State University Comprehensive Cancer Center.

Data processing

We processed (sorted, demultiplexed, clustered, and formatted) raw data from Illumina outputs with ipyrad v 0.9.52 (Eaton & Overcast, 2020), using the resources provided by the Ohio Supercomputer Center. We processed five different datasets: (1) all samples; (2) *N. brasiliensis* (population 1); (3) *N. brasiliensis* (population 2); (4) *N. brasiliensis* (population 3); (5) *N. planiceps*. Datasets 2-5 represent distinct populations recovered in the population assignment analyses (see population assignment section). First, we demultiplexed raw data using individual

barcode adapters. Next, we filtered for adapters using the stricter option. We set the maximum low-quality base calls in the read 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 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.

Population assignments

We used STRUCTURE v2.3.4 (Pritchard et al., 2000) to partition samples into discrete populations before building demographic models. We ran three independent replicates using 100,000 steps of burn-in, followed by 500,000 generations. We performed all runsunder an admixture model for population ancestry and allele frequencies correlated among populations. We evaluated *K*-values ranging from 2 to 6, with ten 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.

Testing diversification history using convolutional neural networks

In phylogeographic model selection, there are countless ways of parameterizing a model. 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 2,000 possible models when incorporating 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 divided the analysis in two parts. First, we independently tested each population for demographic change in population size through time (12 models). Second, we applied this model of population size change while testing models that consider all possible topologies for four tips and also various migration scenarios (26 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).

Testing population trajectory through time

In the first part of model selection, we used a CNN to identify the population trajectory that best described the demographic history of each population. We defined three possible scenarios (Figure 2): (a) *constant population size through time,* (b) *population expansion since the last glacial maximum (LGM), and* (c) *population bottleneck since the LGM*. We used the software fastsimcoal2 to simulate 10,000 data examples for each demographic scenario and population. We simulated short DNA sequences (5 bp) for 100,000 independent loci to ensure that the simulator only generated 1 SNP per locus and kept the same number of SNPs as observed in the empirical datasets. We parameterized the ancestral effective population size, current effective population size, and time of population size changing. All priors are presented in 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,

248 respectively. We labeled the major allele as the ancestral state (0) and the minor allele as the
249 derived state (1), such that the matrix could be converted to a black and white image with each
250 entry corresponding to a pixel in the image.

251 We implemented a two-dimensional CNN architecture as follows: a two-dimensional
252 convolution layer (kernel = 3 x 1), a two-dimensional maximum pooling layer (kernel = 3 x 1), a
253 two-dimensional convolution layer (kernel = 3 x 1), and a two-dimensional maximum pooling
254 layer (kernel = 3 x 1). We then flattened the output layer from the last pooling. Next, we created a
255 fully connected layer with 100 neurons, followed by one with 25 neurons, and a final layer with
256 three neurons, which correspond to our three demographic models (i.e., constant, expansion, and
257 bottleneck; Figure 3). For all layers, we used rectified linear unit activation functions (ReLU),
258 except for the last one where we used a softmax function. This function is a generalization of the
259 logistic function and used for multiclass prediction. We compiled the CNN using the Adam
260 optimization procedure (Kingma & Ba, 2015), a categorical cross-entropy loss function, and a
261 mini-batch size of 100. We ran the CNN for 10 epochs, although without any improvement after
262 three epochs. We did not include a dropout layer because of the lack of evidence of overfitting.
263 We trained the CNN using 80% of the simulated datasets and used the remaining 20% to evaluate
264 model accuracy. Lastly, we used the trained model to predict the model that likely generated the
265 empirical dataset. We built all CNNs with the Keras python library (https://keras.io).
266
267 *Testing evolutionary relationships and gene flow*
268 In the second part, we implemented a CNN architecture to assess the relationships among
269 populations and gene flow between populations that showed evidence of admixture in
270 STRUCTURE. We specified 26 demographic models, which comprise the combination of all 15

possible topologies along with scenarios of isolation after divergence or secondary contact that reflect our identification of individuals that are potentially admixed. For example, because we recovered substantial admixture between populations 2 and 3, we included models with potential secondary contact between these populations (see Figure 4). We did not include models with secondary contact when populations 2 and 3 were sister in the phylogenetic tree, because it was impractical to distinguish between isolation and secondary contact models in our preliminary runs. We used fastsimcoal2 to generate 10,000 data examples per model. As in the first part, we generated short DNA sequences of 5 bp for 100,000 independent loci in a way to simulate 1 SNP per locus. However, we only output the number of SNPs observed in the empirical dataset. 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 S1. We converted alignments nto 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 x 1), a two-dimensional maximum pooling layer (kernel = 3 x 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 (Figure 5), 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 5 epochs; but the model did not improve after the second epoch. Then we split simulations in training (80%) and test datasets (20%). Finally, we used the trained model to predict the empirical dataset. We used the python library Keras throughout to build the CNN.

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 prep.; 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 datasets 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 CNNs models and are presented in Table S1. After simulations, we used the ABC approach to estimate model support using the "postpr" function implemented in "abc" R-package. We set the tolerance value to 1% and used the rejection method to compare models. We evaluate whether simulations produced summary statistics similar to the empirical dataset using PCAs.

Results

Genomic data processing

After genomic data processing, we obtained 4174 unlinked SNPs when all samples were combined, or 6860, 10931, 9396, and 12048 unlinked SNPs for the three *N. brasiliensis*

317 populations and *N. planiceps*, respectively. Because our CNN approach does not accommodate
318 missing data, loci were required to be present in 100% of the samples.
319
320 **Population assignment**
321 The STRUCTURE analysis recovered four geographically structured populations that
322 correspond to *N. planiceps* and three populations within *N. brasiliensis* (hereafter population 1,
323 population 2, and population 3; Figure 2). While *N. planiceps* is distributed in northern
324 Amazonia, population 1 is found in an enclave of Seasonally Dry Tropical Forests within
325 Cerrado. Population 2 is more widespread in Cerrado and population 3 is found in lowlands
326 within Cerrado. In addition, population assignment analysis revealed a region of high admixture
327 between population 2 and 3 (locality #9).
328
329 **Demographic model selection**
330 We recovered population expansion as the best demographic scenario for *N. planiceps*,
331 population 2, and population 3 with a probability of 0.99, 0.59, and 1.0, respectively (Table 2).
332 For population 2, the lower probability value is likely related to the unaccounted gene flow with
333 populations 3, which introduced a genetic variation that was not captured by the model.
334 Conversely, for population 1, we found evidence of constant population size over time
335 (probability = 0.985; Table 2). For all models within each population, the CNN model had a high
336 accuracy when predicting the test set labels, reaching an accuracy higher than 99% for all models
337 (Figure 4).
338 For the second part of model comparison, CNN recovered a single model (#22) as the best
339 evolutionary scenario with a probability of 0.79 (Table 2). 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). Conversely, the posterior probabilities of ABC models were considerably lower. Models 16 and 17 were the best models supported by this analysis, with a posterior probability of 15% each (Table 2). PCAs showed that most models produced summary statistics coincident with empirical datasets, indicating that the choice of priors was plausible (Figure S1).

Discussion

Our simulation testing implies that a deep learning approach for phylogeographic model selection can be very 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*, likely caused by the small number of samples for this species. Model accuracy, therefore, seems to rely on the number of individuals and the number of SNPs. 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%. Undoubtedly, these models are more complex than those dealing only with changes in population size, given that all populations were compared

363 simultaneously, and we also included the relationships among them, gene flow between
364 populations 2 and 3, and divergence times. Still, our approach reached an accuracy similar to
365 other approaches. Conversely, ABC was unable to accurately find the best fit model, given the
366 low posterior probability of all models (see Table 2).

367 CNN and ABC share many similarities, including the use of a simulation-based approach
368 to generate new examples, given a demographic scenario and a set of priors. However, one of the
369 main differences between these approaches is how they summarize the simulated datasets and,
370 most importantly, how empirical and simulated datasets are compared. Therefore, a key feature
371 of any of these methods is to be able to summarize the information in the data in a meaningful
372 way. For ABC, a large number of summary statistics is usually calculated from the simulated
373 datasets, e.g., Tajima's D, nucleotide diversity, F_{ST}, and Fu and Li's D and F statistics. Each
374 summary statistic has been used in phylogeographic investigations. For example, Tajima's D is a
375 summary statistic that detects departures from constant population sizes over time, including
376 population expansion and bottleneck. In addition, fixation indexes have measured the degree of
377 differentiation among populations. The choice of summary statistics is largely subjective, with
378 most studies choosing not to identify a subset of summary statistics that maximize model
379 probability. As stated by Beaumont et al. (2002), "*a crucial limitation of the rejection-sampling*
380 *method is that only a small number of summary statistics can usually be handled*". Our results
381 mirror those from previous research suggesting that ABC does not perform as well with large
382 numbers of models (Pelletier & Carstens, 2014; Smith et al., 2017).

383 Although it is beyond the scope of this study to compare different methods of
384 phylogeographic model selection, a broadly comparison of the accuracy of these approaches can
385 be made based on our approach. For example, PHRAPL summarizes data using gene trees and

because of that, incomplete lineage sorting (ILS) is one of the main sources of model selection inaccuracy (Jackson et al., 2017). At shallower divergence times, a more pronounced discordance in gene trees can be observed and, consequently, it is more difficult to identify the evolutionary scenario that gives birth to the data. Similarly, for CNNs, as the divergence times decreases among lineages the model accuracy decreases, which likely results from ILS (Blischak et al., 2020). As noted above, conventional ABC approaches can attain high accuracy with a high number of models, but this potential liability can be alleviated. 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 compare the Random Forest classifier with ABC. Their overall errors using Random Forest were 7.67% (range: 0–42%) and ~30% for ABC. The overall error of our CNN was 13%, but 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 the tested models, we subset our models and trained a CNN only with isolation models (models 1 to 15). The overall error was 1.5% (0.75–3%; Figure S2) and the best model had a probability of 87%. In a more recent study, Smith & 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 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 there exists a set of statistics that is likely 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 require the data to be summarized as we have done here. 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. Their approach was found to be more accurate than approaches that were based on introgression-specific summary statistics.

Our approach was computationally more demanding than the one proposed by Blischak et al. (2020). It requires an average of two seconds to run the simulation in *fastsimcoal2* and eight seconds to process the image (~ 10 seconds from simulation to generate an image). Since we simulated 10,000 examples per model, it would take about 27 hours to simulate the images that correspond to one scenario. It required 10 hours 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 images storage in the Supercomputer. Our analysis used a total of 380,000 images.

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 & Williams (1970) almost simultaneously proposed the PRH to explain patterns of species diversity and distribution in the Amazon rainforest, where climate oscillations 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 respond for differences in biodiversity between western and eastern Amazonia. In particular, eastern Amazon, which is more connected to the historical and current climate in the Cerrado, held desynchronized interleaved periods of wet and dry climates during the last 250 thousand years (kyr) with western Amazonia. These climatic patterns resulted in habitat fragmentation that isolated species that were previously broadly distributed and led to decreased gene flow and

455 increased genetic differentiation. Some community-level analyses suggest that this model is
456 broadly applicable (Gehara et al., 2017; Silva et al., 2019). In contrast, climate was more stable in
457 western Amazon, which is hypothesized to have generated the observed higher levels of
458 biodiversity across multiple taxonomic groups and likely population stability through time.
459 Our phylogeographic model selection results support the quasi-dipolar scenario of Cheng
460 et al. (2013). We found support for population expansion in *N. planiceps* and populations 2 and 3.
461 While our results showed that population 1 was constant in size through time, this population is
462 located in an enclave of Caatinga within Cerrado (Paranã valley). Caatinga is the largest nucleus
463 of Seasonally Dry Tropical Forests (SDTF) and characterized by xeric vegetation, high
464 seasonality, and unpredictable droughts. It is hypothesized that the climatic oscillations during
465 the Pleistocene led the expansion and connection of now disjunct SDTFs (the Pleistocenic Arc
466 Hypothesis - PAH; Prado & Gibbs, 1993; Pennington et al., 2000). This hypothesis is supported
467 by the disjunct distribution of plants and animals as well as molecular data (Lanna et al., 2018;
468 Pennington et al., 2000; Werneck & Colli, 2006). However, the exact time of the PAH is
469 uncertain and the SDTFs could have expanded earlier, during the transition between Pliocene and
470 Pleistocene, and have fragmented before the Last Glacial Maximum (Werneck et al., 2011),
471 which could explain the stable population sizes we recovered in the longer term.
472 In addition to climatic oscillations, the pattern of diversification found by our study mirrors the
473 current taxonomic status of both species, though we found a hidden genetic diversity within *N.*
474 *brasiliensis*. The pattern of divergence among lineages within *N. brasiliensis* follows a southeast-
475 northwest pattern of differentiation, which is shared with other squamates in Cerrado (Guarnizo
476 et al., 2016; Prado et al., 2012; Santos et al., 2014). This pattern of differentiation was likely
477 driven by landscape features and climatic conditions.

Conclusion

Deep learning techniques have been successfully used in fields like medical sciences and agriculture, but their usage in evolutionary biology has just begun (but see Blischak et al., 2020; Flagel et al., 2019; Schrider & Kern, 2018). 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 non-sister 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.

Acknowledgements

We thank members of the Carstens Lab for comments and suggestions on the manuscript and Lisa N. Barrow and Megan L. Smith for lab assistance. We thank Ohio Supercomputer Center for computing resources. We thank curators and managers of the INPA-H and INPA-HT (C. Ribas, M. Freitas, A. Silva) for granting and processing samples under their care. EMF thanks the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for his doctoral fellowship (process #88881.170016/2018). FPW thanks Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for her productivity fellowship (process #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

EMF and BCC conceived the ideas and designed methodology; EMF conducted the lab work and conducted the analyses. All authors interpreted the results and participated in the writing of the manuscript and gave final approval for submission.

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

Camargo, A., Sinervo, B., & Sites, J. W. (2010). Lizards as model organisms for linking phylogeographic and speciation studies. *Molecular Ecology*, *19*(16), 3250–3270. https://doi.org/10.1111/j.1365-294X.2010.04722.x

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

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. de 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*. *165*(3889), 131–137.

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

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

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

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

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 BT - Classification in BioApps: Automation of Decision Making. *Springer*, 323–350. 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. In *Zootaxa* (Vol. 3983, Issue 1). https://doi.org/10.11646/zootaxa.3983.1.1

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., Townsend Peterson, A., 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). https://doi.org/10.1126/sciadv.aat5752

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

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

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

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

691 **FIGURE LEGENDS**

692

693 **Figure 1.** Schematic representation of a 2D convolutional neural network (CNN) architecture. (a)
694 input image; (b) array derived from the input image; (c) convolutional kernel (yellow); (d) feature
695 map; (e) pooling kernel (orange). ANN = artificial neural network.
696
697 **Figure 2.** Map showing the geographic distribution of sampled localities. Purple circle = *Norops*
698 *brasilinesis* (population 1); blue circles = *N. brasilinesis* (population 2); red circles = *N.*
699 *brasilinesis* (population 3); green circles = *Norops planiceps*. Bar represents the genetic structure
700 of *Norops* ssp. across the area of study according to STRUCTURE analysis.
701
702 **Figure 3.** Representation of the models tested using convolutional neural networks. (a) set of
703 three models used to test population trajectory through time; (b) set of 26 models used to test the
704 evolutionary relationships and secondary contact of *Norops* ssp. Numbers and colors represent
705 populations recovered in STRUCTURE analysis. Purple circle = *Norops brasilinesis* (population
706 1); blue circles = *N. brasilinesis* (population 2); red circles = *N. brasilinesis* (population 3); green
707 circles = *Norops planiceps*. Gene between populations 2 and 3 is represented by arrows. The
708 best-supported model for CNN in the second part of comparison is marked by a red box.
709
710 **Figure 4.** Confusion matrices measuring the accuracy of the trained CNNs model on the test
711 dataset to detect demographic changes through time. Numbers represent percentages, which were
712 calculated based on 2,000 images for each model. (a) *Norops brasilinesis* (population 1); (b) *N.*
713 *brasilinesis* (population 2); *N. brasilinesis* (population 3); *N. planiceps*.
714
715 **Figure 5.** Confusion matrices measuring the accuracy of the trained CNNs model on the test
716 dataset of 26 phylogeographic models. Numbers represent percentages, which were calculated
717 based on 2,000 images for each model.

718 **Table 1.** A glossary of terms used in this study.

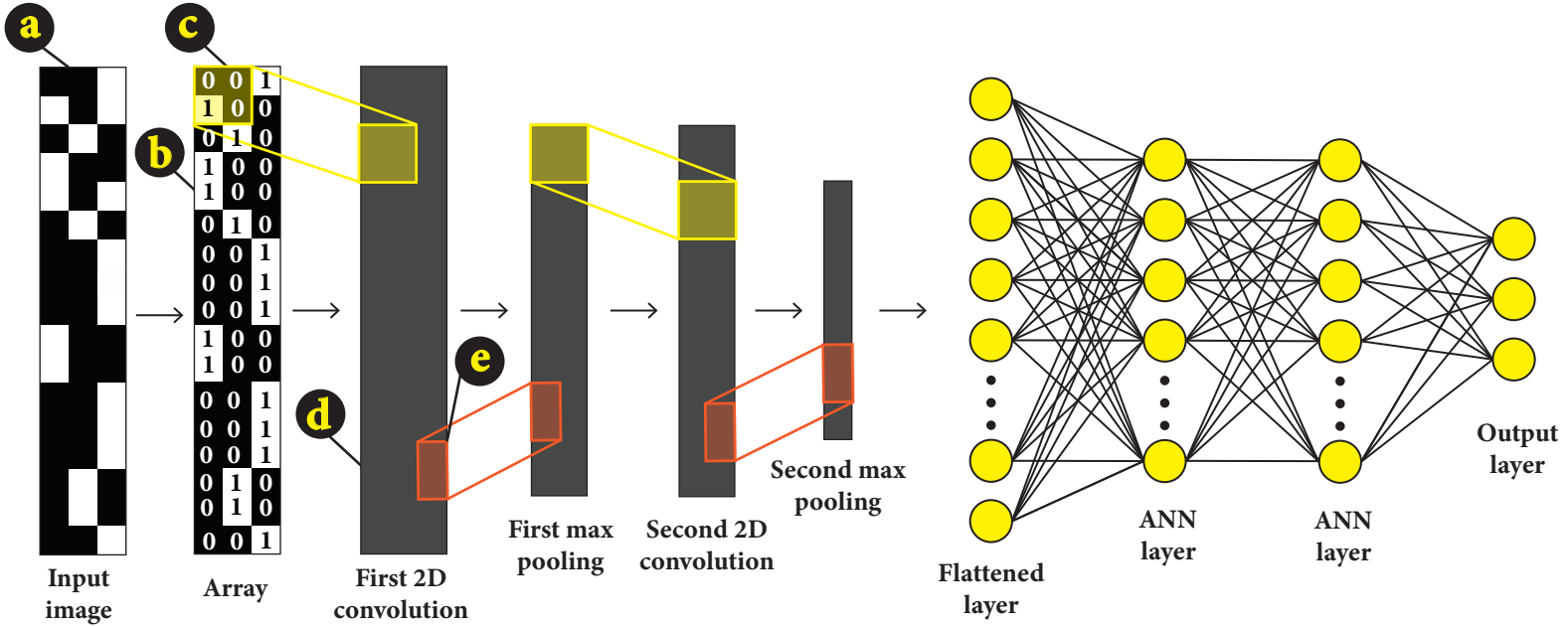
Term	Definition
Artificial neural network-ANN	a model of connected layers that attempt to mimic the way that the brain analyzes and processes information
Convolutional neural network-CNN	a type of artificial neural network used for image classification and recognition
CNN architecture	the general structure 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 method to evaluate how well the model describe the dataset
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

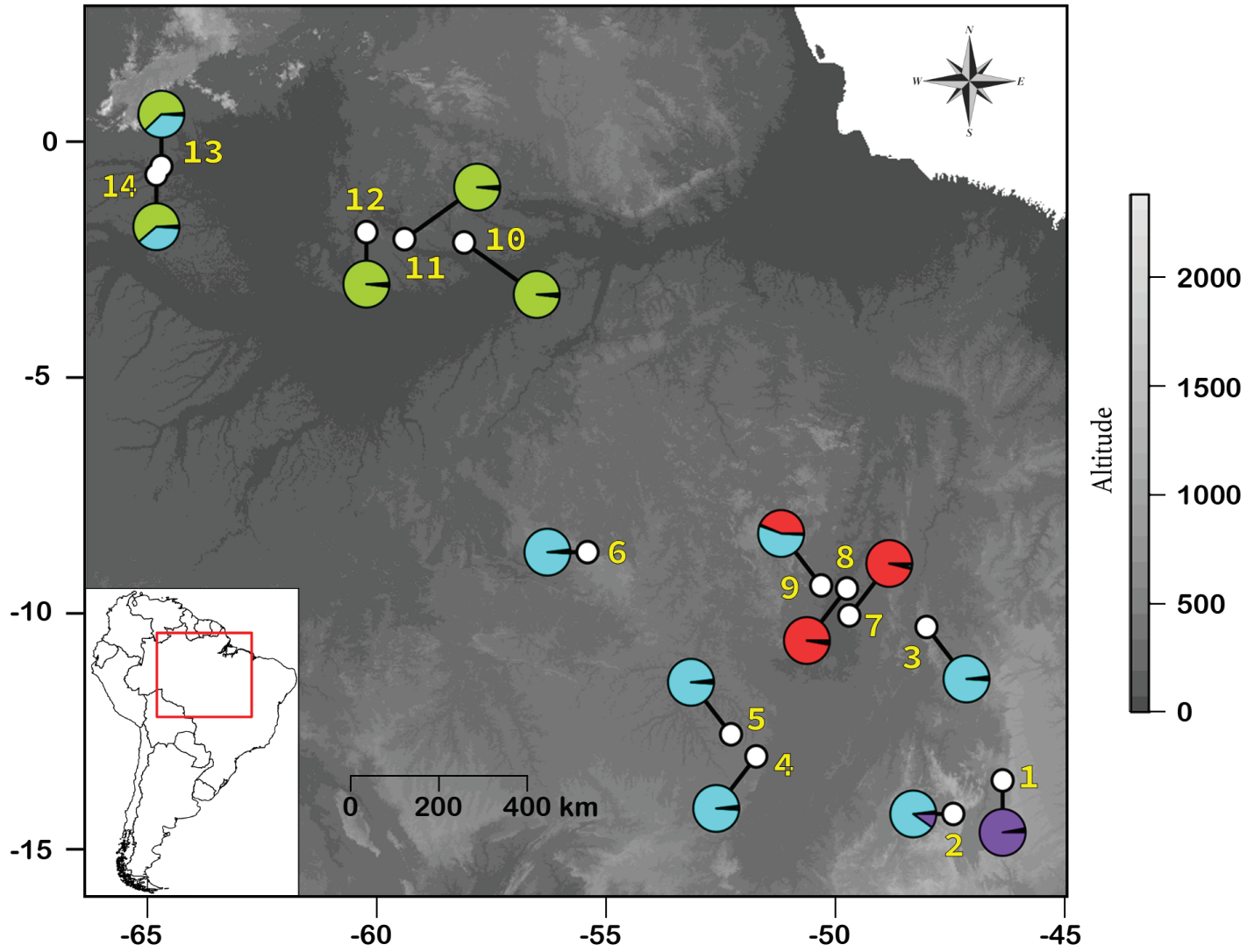
719

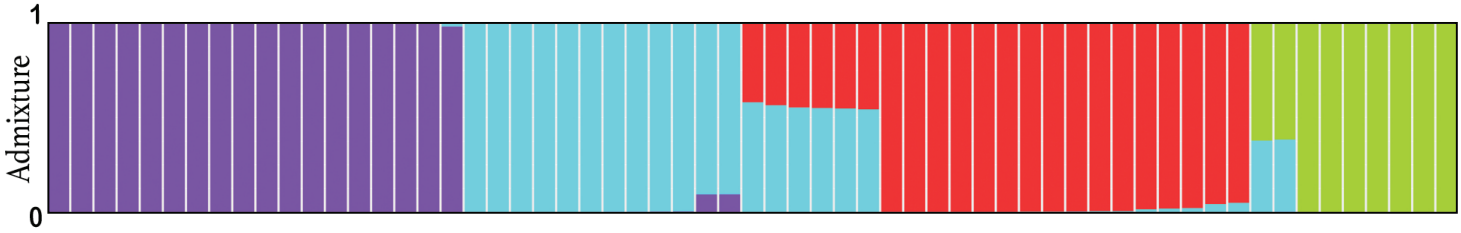
720 **Table 2.** The probability of each model tested using convolutional neural networks and
721 approximate Bayesian computation. Comparisons were first performed within part 1 only using
722 CNNs, and subsequently, models in part 2 were constructed based on demographic scenario
723 inferred in part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

Part 1		**Part 2**		
Model	**CNNs Probability**	**Model**	**CNNs probability**	**ABC posterior probability**
Population 1 - Constant	**0.985**	Model 1	0	0.033
Population 1 - Expansion	0.015	Model 2	0	0.035
Population 1 - Bottleneck	0	Model 3	0	0.025
		Model 4	0	0.11
		Model 5	0	0.11
Population 2 - Constant	0.41	Model 6	0	0
Population 2 - Expansion	**0.59**	Model 7	0	0.1
Population 2 - Bottleneck	0	Model 8	0	0.008
		Model 9	0	0.007
		Model 10	0.01	0
Population 3 - Constant	0	Model 11	0	0
Population 3 - Expansion	**1.0**	Model 12	0	0.007
Population 3 - Bottleneck	0	Model 13	0	0
		Model 14	0	0
		Model 15	0	0
N. planiceps - Constant	0.01	Model 16	0	**0.15**
***N. planiceps* - Expansion**	**0.99**	Model 17	0	**0.15**
N. planiceps - Bottleneck	0	Model 18	0	0.037
		Model 19	0	0.038
		Model 20	0	0.0134
		Model 21	0	0.022
		Model 22	**0.79**	0.060
		Model 23	0	0.0046
		Model 24	0	0.012
		Model 25	0	0.013
		Model 26	0.20	0.065

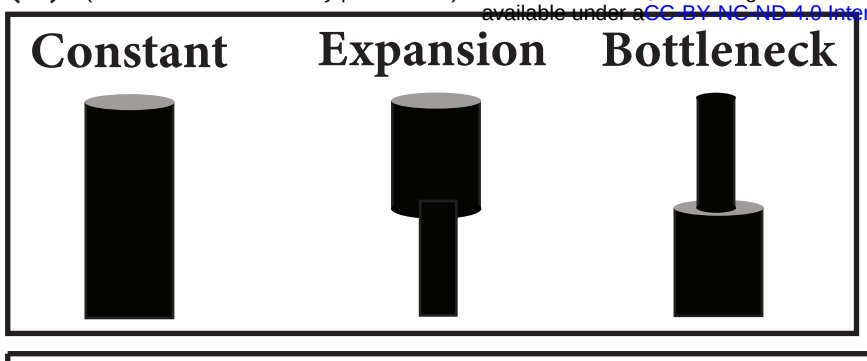
724


a
b
c
d
e
Input image
Array
First 2D convolution
First max pooling
Second 2D convolution
Second max pooling
Flattened layer
ANN layer
ANN layer
Output layer

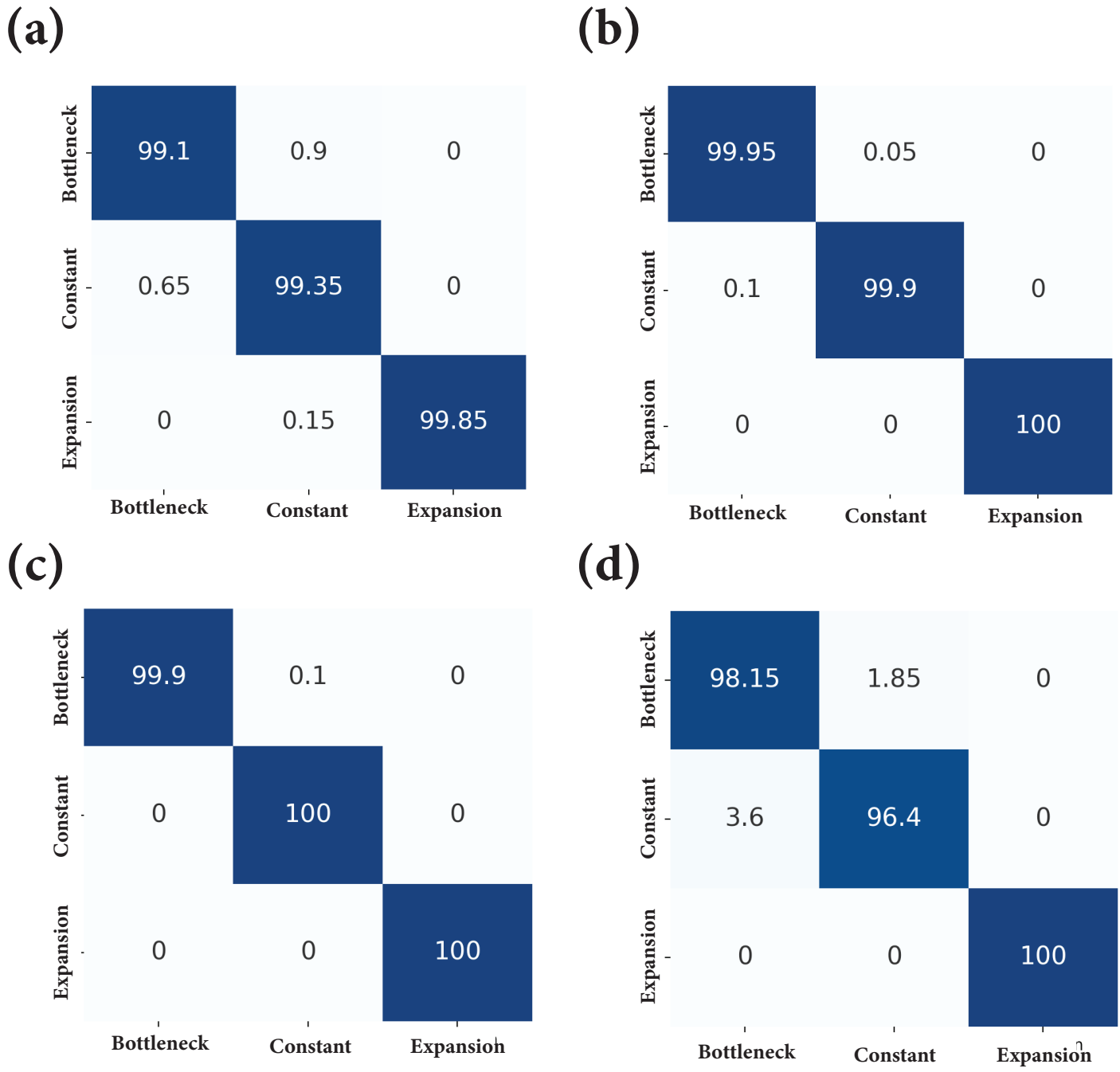


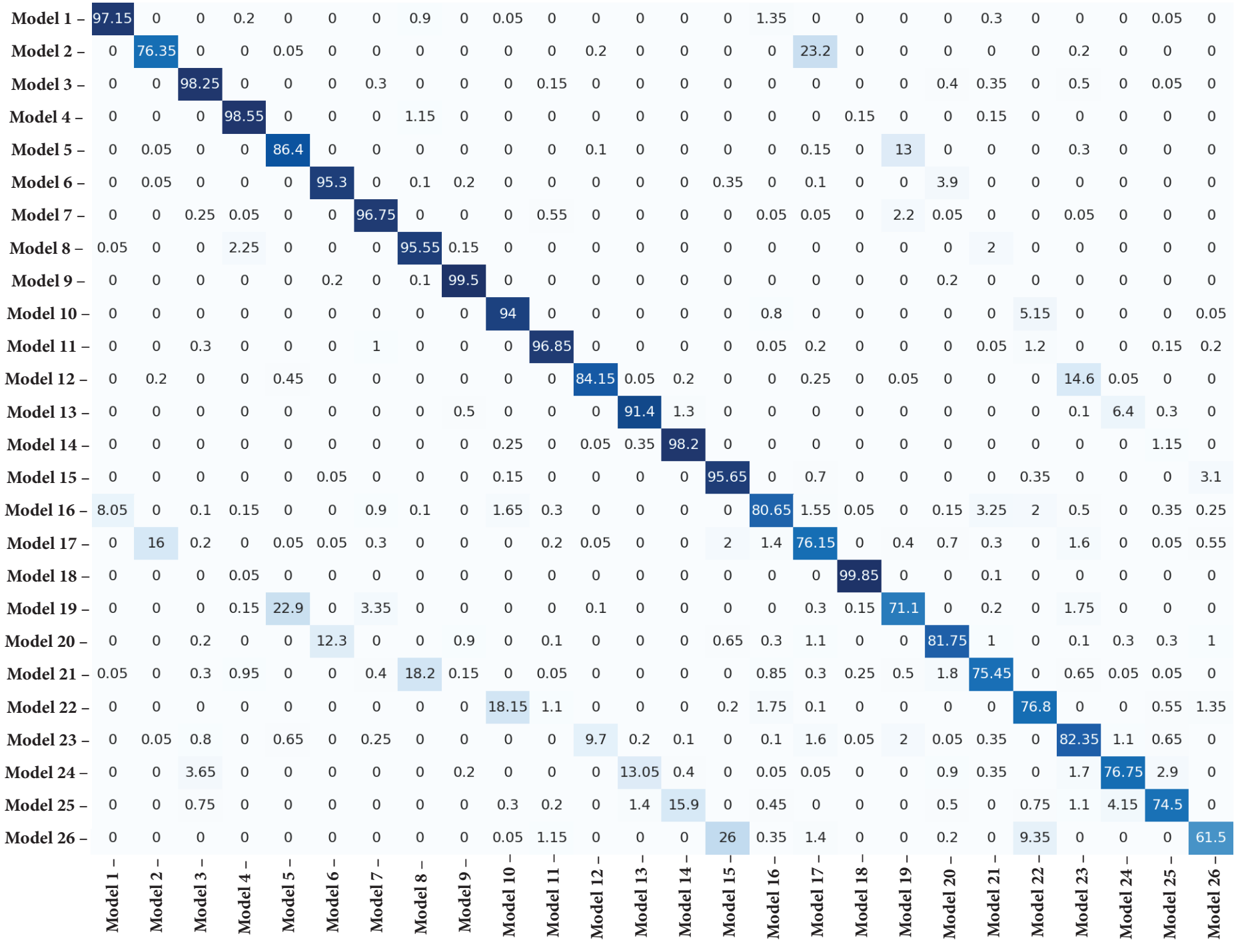


(a)



(b)


(a)
Bottleneck
Constant
Expansion
99.1
0.9
0
0.65
99.35
0
0
0.15
99.85
Bottleneck
Constant
Expansion
(b)
Bottleneck
Constant
Expansion
99.95
0.05
0
0.1
99.9
0
0
0
100
Bottleneck
Constant
Expansion
(c)
Bottleneck
Constant
Expansion
99.9
0.1
0
0
100
0
0
0
100
Bottleneck
Constant
Expansion
(d)
Bottleneck
Constant
Expansion
98.15
1.85
0
3.6
96.4
0
0
0
100
Bottleneck
Constant
Expansion


Model 1 – 97.15 0 0 0.2 0 0 0 0.9 0 0.05 0 0 0 0 0 1.35 0 0 0 0 0.3 0 0 0 0.05 0
Model 2 – 0 76.35 0 0 0.05 0 0 0 0 0 0 0.2 0 0 0 0 23.2 0 0 0 0 0 0.2 0 0 0
Model 3 – 0 0 98.25 0 0 0 0.3 0 0 0 0.15 0 0 0 0 0 0 0 0 0.4 0.35 0 0.5 0 0.05 0
Model 4 – 0 0 0 98.55 0 0 0 1.15 0 0 0 0 0 0 0 0 0 0.15 0 0 0.15 0 0 0 0 0
Model 5 – 0 0.05 0 0 86.4 0 0 0 0 0 0 0.1 0 0 0 0 0.15 0 13 0 0 0 0.3 0 0 0
Model 6 – 0 0.05 0 0 0 95.3 0 0.1 0.2 0 0 0 0 0 0.35 0 0.1 0 0 3.9 0 0 0 0 0 0
Model 7 – 0 0 0.25 0.05 0 0 96.75 0 0 0 0.55 0 0 0 0 0.05 0.05 0 2.2 0.05 0 0 0.05 0 0 0
Model 8 – 0.05 0 0 2.25 0 0 0 95.55 0.15 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0
Model 9 – 0 0 0 0 0 0.2 0 0.1 99.5 0 0 0 0 0 0 0 0 0 0 0.2 0 0 0 0 0 0
Model 10 – 0 0 0 0 0 0 0 0 0 94 0 0 0 0 0 0.8 0 0 0 0 0 5.15 0 0 0 0.05
Model 11 – 0 0 0.3 0 0 0 1 0 0 0 96.85 0 0 0 0 0.05 0.2 0 0 0 0.05 1.2 0 0 0.15 0.2
Model 12 – 0 0.2 0 0 0.45 0 0 0 0 0 0 84.15 0.05 0.2 0 0 0.25 0 0.05 0 0 0 14.6 0.05 0 0
Model 13 – 0 0 0 0 0 0 0 0 0.5 0 0 0 91.4 1.3 0 0 0 0 0 0 0 0 0.1 6.4 0.3 0
Model 14 – 0 0 0 0 0 0 0 0 0 0.25 0 0.05 0.35 98.2 0 0 0 0 0 0 0 0 0 0 1.15 0
Model 15 – 0 0 0 0 0 0.05 0 0 0 0.15 0 0 0 0 95.65 0 0.7 0 0 0 0 0.35 0 0 0 3.1
Model 16 – 8.05 0 0.1 0.15 0 0 0.9 0.1 0 1.65 0.3 0 0 0 0 80.65 1.55 0.05 0 0.15 3.25 2 0.5 0 0.35 0.25
Model 17 – 0 16 0.2 0 0.05 0.05 0.3 0 0 0 0.2 0.05 0 0 2 1.4 76.15 0 0.4 0.7 0.3 0 1.6 0 0.05 0.55
Model 18 – 0 0 0 0.05 0 0 0 0 0 0 0 0 0 0 0 0 0 99.85 0 0 0.1 0 0 0 0 0
Model 19 – 0 0 0 0.15 22.9 0 3.35 0 0 0 0 0.1 0 0 0 0 0.3 0.15 71.1 0 0.2 0 1.75 0 0 0
Model 20 – 0 0 0.2 0 0 12.3 0 0 0.9 0 0.1 0 0 0 0.65 0.3 1.1 0 0 81.75 1 0 0.1 0.3 0.3 1
Model 21 – 0.05 0 0.3 0.95 0 0 0.4 18.2 0.15 0 0.05 0 0 0 0 0.85 0.3 0.25 0.5 1.8 75.45 0 0.65 0.05 0.05 0
Model 22 – 0 0 0 0 0 0 0 0 0 18.15 1.1 0 0 0 0.2 1.75 0.1 0 0 0 0 76.8 0 0 0.55 1.35
Model 23 – 0 0.05 0.8 0 0.65 0 0.25 0 0 0 0 9.7 0.2 0.1 0 0.1 1.6 0.05 2 0.05 0.35 0 82.35 1.1 0.65 0
Model 24 – 0 0 3.65 0 0 0 0 0 0.2 0 0 0 13.05 0.4 0 0.05 0.05 0 0 0.9 0.35 0 1.7 76.75 2.9 0
Model 25 – 0 0 0.75 0 0 0 0 0 0 0.3 0.2 0 1.4 15.9 0 0.45 0 0 0 0.5 0 0.75 1.1 4.15 74.5 0
Model 26 – 0 0 0 0 0 0 0 0 0 0.05 1.15 0 0 0 26 0.35 1.4 0 0 0.2 0 9.35 0 0 0 61.5
Model 1 –
Model 2 –
Model 3 –
Model 4 –
Model 5 –
Model 6 –
Model 7 –
Model 8 –
Model 9 –
Model 10 –
Model 11 –
Model 12 –
Model 13 –
Model 14 –
Model 15 –
Model 16 –
Model 17 –
Model 18 –
Model 19 –
Model 20 –
Model 21 –
Model 22 –
Model 23 –
Model 24 –
Model 25 –
Model 26 –