INTRODUCTION

Recognizing species boundaries has long been a major challenge for biologists. The main difficulty is to some degree related to the numerous existing species concepts. The use of specific definitions can lead to alternative strategies for identifying species boundaries in empirical data sets (Carstens et al., 2013; de Queiroz, 2007). However, different species concepts can be considered elements of diverse properties that are associated with the dynamics of the speciation continuum. After the proposal of the unified species concept (Queiroz, 2007), the roles of species concept theory and species delimitation methodologies became apparent. The view of species as independent segments...
of a metapopulation indicated lineage independence as the only criterion necessary for delimiting species boundaries, avoiding any disagreement purely related to the species concept.

Selecting a suitable approach for species delimitation has been difficult, especially in species complexes (Pinheiro et al., 2018). Identifying discontinuities among incipient stages of divergence, which are commonly found in species complexes, demands great effort and a multidisciplinary approach to assess different sources of evidence supporting species limits (Carstens et al., 2013). In this context, estimates based on independent sources of data such as morphology, cytotgenetics, anatomy, ecology, and genetics have been used to achieve a better resolution in species delimitation (Alvarado-Sizzo et al., 2018; Denham et al., 2019; Domingos et al., 2014). In particular, phenotypes and geographic distributions can be a starting point for hypotheses about species circumscriptions (Luo et al., 2018; Solís-Lemus et al., 2015).

Species delimitation methods based on multilocus data and the multispecies coalescent model (MSC) compare the probability of species trees with different numbers of operational taxonomic units (OTUs) to identify optimal partitions for the data (Ence & Carstens, 2018). Species delimitation methods that implement approximate Bayesian computation (ABC; Camargo et al., 2012) and approximate likelihood analysis (Morales et al., 2017). ABC represents a class of flexible likelihood-free algorithms for performing Bayesian inference (Beaumont et al., 2002). The principle behind the method relies on the massive simulation of genetic data using parameter values drawn from a prior distribution, followed by the calculation of summary statistics (SuSt) for each simulation. SuSt are values generated from the raw genetic data, which are expected to capture important information to differentiate among the simulated models. Simulations that produce genetic variation patterns (SuSt) close to the observed data are retained to form an approximate sample from the posterior distribution.

A number of machine learning methods are increasingly being applied to species delimitation. The applied strategies include coupling ABC with random forest (Smith & Carstens, 2020), use a Support Vector Machine on SuSt (Pei et al., 2018) and using different unsupervised machine learning approaches (Derkarabetian et al., 2019). Another promising approach recently developed for demographic model selection, based on deep learning image classification, uses convolutional neural networks (CNNs) to retrieve as much information as possible from genetic data sets by converting them to images without requiring user-specified SuSt (Flagel et al., 2019). Simulations have shown that CNNs maximize the use of available information from the data and increase the capacity to distinguish divergent evolutionary histories, frequently overcoming the limitations of other traditional approaches (Flagel et al., 2019). Despite the recent use of CNNs in population genetics (Flagel et al., 2019; Sanchez et al., 2020) and phylogeography (Fonseca et al., 2021; Oliveira et al., 2020; Souza et al., 2019), this approach has not previously been tested in the context of species delimitation using genetic data.

Here, we introduce an approach for species delimitation that combines the use of MSC-based methods with model selection via CNN. We compared our model selection method with ABC (detailed in Figure 1) to evaluate distinct species delimitation hypotheses in the cactus species complex *Pilosocereus aurisetus*, currently arranged in two subspecies. We chose the *P. aurisetus* complex because (1) it presents a controversial taxonomy, and (2) it has a sky-island distribution, which imposes difficulties on MSC-based methods (Leachê et al., 2018). We also validated our method by following this same procedure using a recently published data set from two *Drosophila* species pairs (Campillo et al., 2020), which have information about pre- and post zygotic reproductive isolation and do not have controversial taxonomy.

## 2 | MATERIALS AND METHODS

### 2.1 | The study system

The taxon *P. aurisetus* is a puzzling system regarding species delimitation. The taxonomy of *P. aurisetus* complex has been unstable over the years, which is a common attribute in the family Cactaceae, probably due to convergent and parallel evolution causing the absence of clear diagnostic characters for some groups (Copetti et al., 2017; Hernández-Hernández et al., 2011). The morphological variation observed within this taxon has led to repeated taxonomic evaluations, with species and subspecies being recurrently synonymized and re-established (Hunt et al., 2006; Taylor & Zappi, 2004; Zappi, 1994). Another complicating issue is the naturally fragmented sky-island distribution of *P. aurisetus*. The taxon occurrence is restricted to the mountaintops of the Espinhaço Mountain Range in eastern Brazil and is a typical element of the campo rupestre landscapes (Figure 2). The campo rupestre is a rock outcrop vegetation harbouring outstanding species richness and endemism in South American
ancient mountaintops (Miola et al., 2021). In line with this sky-island vegetation system, *P. aurisetus* shows high intraspecific genetic structure and restricted or absent gene flow, even among neighbour populations (Bonatelli et al., 2014; Perez, Bonatelli, et al., 2016; Perez et al., 2016a). Although the details of the reproductive biology of *P. aurisetus* is unknown, the flower and fruit characteristics are similar to other congeneric taxa, which are predominantly pollinated by bats (Rocha et al., 2019) and seed-dispersed by birds and bats (Vázquez-Castillo et al., 2019). Currently, two subspecies are recognized based on differences in the size and diameter of stems and the colour and density of hairs on flower-bearing areoles: the more widespread *P. aurisetus* (Werderm.) Byles & Rowley subsp. *aurisetus*, and *P. aurisetus* subsp. *aurilanatus* (Ritter) Zappi, which is limited to only a few populations restricted to a disjunct mountain (Serra do Cabral) west of the main Espinhaço Mountain Range. We also included plants from three heterotypic synonyms of *P. aurisetus* subsp. *aurisetus* (Figure S1). The heterotypic synonym *P. aurisetus* subsp. *densilanatus* (Ritter) Braun & Esteves occurs on the easternmost side of the Serra Negra Mountains and is distinguished by their densely wooly stems reaching up to 4 cm in diameter. *P. aurisetus* subsp. *supthutianus* (Braun) Braun & Esteves is a northern outlier of the species that also presents densely hairy stems, which can reach more than 5 cm in diameter. Finally, *P. aurisetus* subsp. *werdermannianus* (Buining & Brederoo) Braun & Esteves occurs in the central range of the *P. aurisetus* distribution and shows more slender stems, fewer ribs and sometimes green epidermis.

### 2.2 Sampling and DNA extraction

We sampled four individuals in each of nine localities in eastern Brazil, totaling 36 sampled individuals and covering the entire known distribution of the *P. aurisetus* species complex (Figure 2; Table 1). Hereafter, we refer to these sampled localities as populations justified by the clear limits of the *P. aurisetus* habitat patches and the marked genetic structure in this taxon (Bonatelli et al., 2014). Root tissue was stored in silica gel and then transferred to a −80°C freezer until DNA extraction. We maintained a distance of approximately 10 m between sampled individuals to avoid collecting clones. Total DNA was extracted using the Extract All kit (Applied Biosystems) and purified with 95% ethanol and 3 M sodium acetate buffer. The DNA concentration was measured with a Qubit fluorometer (Invitrogen).

### 2.3 Microfluidic PCR and sequencing

Massive parallel target amplification of 41 loci in 36 samples was carried out in microfluidic PCR reactors (Access Array System,
Fluidigm). The selected loci included 26 anonymous nuclear markers developed for *P. aurisetus* (Perez et al., 2016b) and another five nuclear, eight plastid and two mitochondrial genic regions selected from the available data in GenBank (Table S1). The development of these new markers was based on available sequences from Cactaceae or from other related plant family species to detect conserved regions for primer design. All primers were developed with PRIMER3 v4.0.0 (Untergrasser et al., 2012), with parameters suggested by the Fluidigm Assay Design Group as follows: primer size from 18 to 23 bases, annealing temperature between 59 and 61°C, maximum poli- X of 3. The selected markers were synthesized according to Fluidigm specifications and applied to a single 48 × 48 access array chip with reactions performed according to the manufacturer’s protocol. The obtained PCR products were pooled and purified with 0.6X AMPure beads (Agencourt). The quality and amplification range were visualized in a BioAnalyser, and the concentration was estimated via real- time PCR using KAPA qPCR (Kapa Biosystems). The samples were subjected to a MiSeq (Illumina) 300 bp paired-end run along with 156 other samples from other projects.

After sequencing, a first filtering step was performed with cutadapt (Martin, 2011), excluding sequences smaller than 60 bp and removing 3’ bases with a PHRED score <Q22. PHRED quality for all reads was then visualized in FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) and graphically compiled in MultiQC (http://multiqc.info/). Reference sequences for each marker were generated from the alignments used to design the primers. Mapping reads for each marker was performed with BWA-MEM (Li, 2013), which is suited for paired-end long reads with indels, while SAMtools (Li et al., 2009) was used to retain only sequences that were mapped with high fidelity. Single-nucleotide polymorphism (SNP) calling was carried out with GATK (McKenna et al., 2010) by transforming low-quality bases to missing data, identifying possible indels with realignment, applying polymor- phism filters for quality and coverage, and generating FASTA.
files for each sample. To confirm the obtained polymorphisms, each detected SNP was then mapped against reference files for each marker that were built from reads from each analysed sample. The phasing of SNPs in nuclear markers was achieved with a Python script, modified from Harvey et al. (2016). The aligned sequences from each marker were obtained with MAFFT v.7 (Katoh & Standley, 2013) with default parameters. Recombination within each marker was tested with the DSS method (McGuire & Wright, 2000) implemented in TOPALi v2 (Milne et al., 2008) and topological incongruences were tested in KDETrees (Weyenberg et al., 2014). The most likely substitution model for each locus was obtained with PartitionFinder v.1.1.1 (Lanford et al., 2012).

2.4 | Species tree and coalescent delimitation

The phylogenetic relationships among populations were recovered with a species tree in StarBEAST v2.1.3 (Bouckaert et al., 2014), assigning the sequences from each population to a single OTU. For all markers, we used the selected substitution model (Table S1) and a relaxed clock log-normal, with the mean sampled from a broad log normal distribution (according to Perez et al., 2016a). We performed three runs with \(5 \times 10^5\) MCMC iterations each, sampling trees every \(10^5\) generations with 75% burnin. Convergence was observed in TRACER 1.6 (Rambaut et al., 2014) and a maximum clade credibility (MCC) tree with median heights was obtained in TreeAnnotator (Drummond & Rambaut, 2007).

We used the obtained species tree topology and the genomic sequences obtained from microfluidic PCR as input to test the species limits in *P. aurisetus* complex with the method (A10 analysis) implemented in BPP v4.2.9 (Froui et al., 2018; Yang & Rannala, 2014). For this analysis, we concatenated all mitochondrial (cox1 and cox2) and chloroplast (atpB-rbcL, petL-psbE and rpl16) markers. The BPP method uses a reversible jump Bayesian (rJ-MCMC) framework to estimate the posterior probability of a species split at each node of the phylogeny. To ensure convergence, five independent runs were performed to estimate the best delimitation scheme, with parameters selected according to the values suggested by Leaché et al. (2018) combined with previous estimates for the species from Perez, Bonatelli, et al. (2016). \(\theta\) and \(\tau\) priors followed an inverse gamma (IG) distribution with \(a = 3\) and \(b = 0.01\) and 0.002, respectively (hereafter BPP “specific prior values”). We evaluated the potential effect of priors in the delimitation results by also performing the same number of runs with broad prior values \(\theta \sim IG(3, 0.002)\) and \(\tau \sim IG(3, 0.03)\) as suggested in the BPP manual (hereafter “diffuse prior values”). We considered a node as valid only if both analyses support its existence with a posterior probability (PP) of 0.95 or higher. The runs were carried out for \(10^5\) MCMC iterations sampled every five steps after discarding the first \(10^4\) iterations as burnin. The obtained delimitations were mapped to the tips of the species tree topology. To evaluate the potential effect of tree uncertainty in species delimitation, we also performed additional BPP runs using the same strategies described above, but with a joint species delimitation and species tree inference (A11 analysis). Also, as a more conservative alternative, we evaluated our BPP results using the gdi index, as suggested by Leaché et al. (2018). We considered the threshold values suggested by Jackson et al. (2017) and Leaché et al. (2018), in which pairwise gdi values smaller than 0.2 indicate a single species and above 0.7, two species.

2.5 | Comparing delimitation hypotheses with deep learning and ABC

We adopted an integrative approach to establish putative species limits by considering information on the *P. aurisetus* complex from different sources and analyses, as suggested by Carstens et al. (2013). For that, we compared the delimitation hypotheses generated by the BPP results (with and without the gdi index), as well as two morphological hypotheses based on one lumper and one splitter taxonomic arrangement (Table 1). The genetic hypotheses, based on the BPP results, were composed of one species when gdi was considered and five species when it was not incorporated (see results below); the “splitter” morphological hypothesis, considering both currently accepted and previously synonymized taxa, was composed of five species; and the “lumper” morphological hypothesis, considering only the currently recognized subspecies, was composed of two species (Hunt et al., 2006). To test these competing delimitation hypotheses, we considered scenarios including the number of species and the potential migration between the delimited entities. Therefore, seven scenarios (Table 1; Figure S4) were tested: (1) “splitter”, (2) “splitter” combined with migration, (3) “lumper”, (4) “lumper” with migration, (5) BPP combined with gdi (BPP_GDI), (5) BPP without gdi (BPP_noGDI), and (7) BPP_noGDI with migration. As the BPP_GDI scenario consists of a single species, an additional model including migration was not necessary. We performed model comparison using both a deep learning approach based on CNN and a regular ABC method (Figure 1). For these two strategies, we simulated genetic data sets under each scenario with a modified version of the scripts from Perez, Bonatelli, et al. (2016). To simplify our simulations and avoid using an overly heterogeneous data set, we included only nuclear markers and simulated data mirroring sample sizes with the same number of segregating sites per locus of our empirical data set. For the deep learning approach, 10,000 simulated data sets per model were converted to images that were used to train a CNN to recognize simulations generated by different scenarios (for details on the simulation and training procedure, see Supporting Information text). The comparison of delimitation scenarios using regular ABC analysis was performed following the approach of Perez, Bonatelli et al. (2016) with 100,000 simulations per model conducted under the same conditions used for the CNN (details in Supporting Information text).

To validate our newly proposed deep learning method for species delimitation, we also carried out the same analyses using a data set recently published by Campillo et al. (2020). These authors compared the results of reproductive isolation and coalescent-based
TABLE 1  Sampling localities for *P. aurisetus* species complex and their assignement to operational taxonomic units (OTU) according to four different species delimitation hypotheses (“splitter”, “lumper”, “BPP_noGDI” and “BPP_GDI”). For all hypotheses except BPP_GDI (which has only one OTU), we also included a model incorporating migration (+M) between the OTUs.

| Taxon/locality       | Code | N  | Longitude | Latitude | "Splitter" | "Lumper" | "BPP_noGDI" | "BPP_GDI" |
|----------------------|------|----|-----------|----------|------------|----------|-------------|-----------|
| *P. a. subsp. aurisetus* |      |    |           |          |            |          |             |           |
| Cocaís-MG            | COC  | 4  | -43,450000 | -19,866667| 1          | 1        | 1           | 1         |
| Itamarandiba-MG (a)  | ITA  | 4  | -42,935972 | -18,001944| 2          | 1        | 2           | 1         |
| Pedra Menina-MG (a)  | PMN  | 4  | -43,044500 | -18,136250| 2          | 1        | 2           | 1         |
| Olhos D’Água-MG (b)  | ODA  | 4  | -43,622361 | -17,438833| 3          | 1        | 3           | 1         |
| Engenheiro Dolabela-MG (b) | EDB  | 4  | -43,793667 | -17,451417| 3          | 1        | 4           | 1         |
| Bocaiúva-MG (b)      | BOV  | 4  | -43,758333 | -17,545833| 3          | 1        | 4           | 1         |
| Mendanha-MG (c)      | MEN  | 4  | -43,536583 | -18,118917| 4          | 1        | 2           | 1         |
| Inhaí-MG (c)         | INA  | 4  | -43,605833 | -17,146444| 4          | 1        | 2           | 1         |
| *P. a. subsp. aurilanatus* |      |    |           |          |            |          |             |           |
| Joaquim Felicio-MG   | JFE  | 4  | -44,178333 | -17,757833| 5          | 2        | 5           | 1         |

Note: (a) (b) and (c) indicate the samples from the heterotypic synonyms *P. a. subsp. densilanatus*, *P. a. subsp. supthutianus* and *P. a. subsp. werdermannianus*, respectively.

(BPP) delimitation methods to infer species boundaries in several species pairs of the genus *Drosophila*, a classic model system in speciation research. We focused our approach on data from two species pairs of the Melanogaster group: *D. melanogaster*—*D. sechellia* and *D. melanogaster*—*D. simulans*. We chose these species because they are widely accepted and corroborated by both BPP and reproductive isolation metrics (Campillo et al., 2020). Furthermore, while the *D. melanogaster*—*D. sechellia* pair is allopatric, the *D. melanogaster*—*D. simulans* pair is sympatric, allowing us to explore whether geographic context might affect our ability to identify these species. Therefore, we tested a model with all conspecifics against a model in which the specimens were divided into two species, which was the expected outcome according to the currently supported classification. As with *P. aurisetus*, our simulations were based on empirical sample sizes and the number of segregating sites (details in Supporting Information text), with priors based on the values suggested by Campillo et al. (2020).

3 | RESULTS

3.1 | Microfluidic PCR and sequencing

Four markers showed no amplification in any sample (PaANL_46, petA, rps16, and ycf1). To further minimize missing data and ensure that at least one individual was sequenced from each population, we discarded 16 loci (PaANL_8, PaANL_10, PaANL_50, PaANL_82, PaANL_123, PaANL_134, PaANL_142, PaANL_155, PaANL_160, PaANL_187, PaANL_196, its2, ppc, mh1, psba-psbB, psbC). No recombination was detected at any of the remaining loci, but kdtrees result suggested an incongruent topology in the isi1 marker, which was removed. Therefore, the final data set included 20 markers that were considered in the subsequent analyses for *P. aurisetus* (PaANL_15, PaANL_17, PaANL_28, PaANL_35, PaANL_80, PaANL_87, PaANL_96, PaANL_126, PaANL_140, PaANL_147, PaANL_164, PaANL_165, PaANL_182, PaANL_205, apk1, atpB-rbcL, petL-psbE, rpl16, cox1 and cox2), with a total of 8908 bp.

3.2 | Species tree and coalescent delimitation

The StarBEAST species tree recovered most nodes with high support, especially the ones defining previously and currently valid subspecies (Figure 3). Population JFE of *P. aurisetus* subsp. *aurilanatus* was the most external clade. Other populations were arranged in two subclades. One of them contained populations in the centre (INA and MEN, previously classified as *P. a. subsp. werdermannianus*; ITA and PMN, formerly *P. a. subsp. densilanatus*) and south (COC) of the *P. a. subsp. aurisetus* distribution. The second subclade contained populations from northern distribution (EDB, BOV and ODA, formerly classified as *P. a. subsp. supthutianus*).

All BPP runs (A10 analysis) under each prior set showed the same results, therefore we provided the mean posterior probability of each node (Figure 3). Results with “specific” and “diffuse” priors were similar, though the latter supported a lower number of species. The congruent delimited species were *P. a. subsp. aurisetus* (COC), *P. a. subsp. supthutianus* populations EDB and BOV separated from ODA and the currently recognized subspecies *P. aurisetus* subsp. *aurilanatus* (JFE). The only exception was for the node separating *P. a. subsp. werdermannianus* (INA and MEN) and *P. a. subsp. densilanatus* (ITA and PMN), which was validated (PP = 0.98) only in the "specific" but not (PP = 0.41) in the "diffuse" prior set. To be conservative, we decided to collapse this node, considering them as a single unit. Therefore, BPP resulted in five delimited species that coincided with the "splitter" hypothesis, except for the OTUs composed by populations once recognized as *P. a. subsp. werdermannianus* and *P. a. subsp. aurilanatus*. subsp. *werdermannianus*).
densilanus, which were combined, and for population ODA that was separated from the other P. a. subsp. supthutianus populations (BOV and EDB) (Figure 3). BPP results from the joint species delimitation and species tree (A11 analysis) were similar, with the “specific” prior runs resulting in more delimited units (7—assigning each population to a single species, expect for populations INA, ITA and MEN that were collapsed), while the runs with the “diffuse” priors recovered the same five species described above (BPP_noGDI in Figure 3). When we analysed our BPP results with the gdl index, all estimates were below 0.7 and only three (P. auriisetus subsp. aurilanus—JFE; P. a. subsp. aurilanus—COC; and P. a. subsp. supthutianus—ODA, BOV and EDB) had values between 0.2 and 0.7 (Figure S2). Therefore, to be conservative, we decided to collapse all populations in a single species for this hypothesis.

3.3 | Delimitation hypotheses comparison

After 250 epochs, our CNN showed accuracies of 96.81% and 92.49% for the training and validation sets, respectively. Some degree of overfitting was observed when we plotted the accuracy of the training and validation sets throughout the epochs. However, this effect decreased inversely proportional to the number of evaluated simulations (Figure S3). Our cross-validation procedure, using a test set of simulations not evaluated during training, also showed that increasing the number of simulations resulted in accuracy improvement for all models (Figure 4). It was also possible to observe that CNN presented a higher proportion of simulations that were correctly predicted in relation to their model than the ABC results, with all models being correctly assigned with more than 80% accuracy when 10,000 simulations per model were used (Figure 4 and Figure S4). The percentages of correct predictions with the ABC approach were all higher than 70% when 100,000 and 10,000 simulations per model were used, except for the “splitter” and “BPP_noGDI” models that showed ~50% accuracy (Figure 4 and Figure S4). These two latter models also showed the highest levels of confusion with each other (Figure S4). The “splitter” hypothesis was selected with a softmax probability of 99.9% (Table 2) when the empirical data were submitted to the trained CNN, and this result was consistent even when the order of SNPs was shuffled (Supporting Information text). The ABC approach showed similar results, also selecting the “splitter” hypothesis (PP = 0.962). Further, when we plotted the simulated and empirical genotypes and the ABC posterior with a PCA (Figure S5) the empirical data were located inside the cloud of simulations from the “splitter” model.

The Drosophila data set also pointed to a higher accuracy of CNN over ABC. The cross-validation tests indicate a correct assignment of CNN simulations just above 78% and 81% in the D. melanogaster-D. simulans and D. melanogaster-D. sechellia pairs, respectively (Figure S4). The accuracy of ABC was lower, reaching values slightly higher than 72% and 68% for the same species pairs (Figure S4). When the empirical data were used, both methods supported the hypothesis of two species for the D. melanogaster-D. simulans pair, showing a probability of 99.2% in CNN and 93.6% for ABC. In D. melanogaster-D. sechellia, both methods also pointed to two species with a probability of 94.9% for CNN and 52.6% for ABC (Table 2).

4 | DISCUSSION

4.1 | Deep learning in coalescent-based species delimitation

The advantages of the simulation approaches adopted here (CNN and ABC) over explicit statistical methods are their ability to test models with complex demographic scenarios, such as migration events or size fluctuations. In addition, these approaches enable us to easily test for different assignments of samples into OTUs (but see Leaché et al., 2014) and even compare hypotheses relying on distinct topologies. Here, we compared hypotheses derived from two previous taxonomic arrangements with the results we obtained from BPP. For each hypothesis, we also evaluated whether gene flow among the delimited entities was present, totaling seven evaluated scenarios (Table 1). We decided not to include even more complex scenarios (e.g., by incorporating demographic fluctuations), as evaluating a high number of scenarios can negatively impact the performance of the methods applied, especially in ABC (Pelletier & Carstens, 2014). Indeed, our comparison of the two approximate approaches suggests that deep learning (CNN) showed a higher capacity to distinguish among the simulated demographic scenarios, outperforming ABC (Figure 4 and Figure S4). It is important to highlight that CNN achieved this performance with ten times less simulations and much smaller running times (~22 min to train the network in a Nvidia K80 GPU with an Intel Xeon 2.30 GHz CPU) than ABC (~17 days to perform cross-validation and 19 min for model selection in an Intel Xeon 2.40 GHz CPU; see Table 3 for a full comparison of running times for each step). The results from the Drosophila data set also supported a higher accuracy of CNN compared to ABC (Figure S4). Moreover, when the empirical data was evaluated, CNN recovered two species for both Drosophila species pairs with higher probability, especially for D. melanogaster-D. sechellia (Table 2). These results validated our CNN approach, as they are in agreement with our expectations based on Campillo et al. (2020).

The main difference between CNN and ABC functioning is the ability of the former to take information directly from SNP matrices without using SuSt, which is required by the latter. Moreover, while ABC relies on a rejection step that discards most of the simulations and leaves only a small proportion that are more similar to the empirical data (Csilléry et al., 2012), the CNN uses information from the whole set of simulations to learn how to distinguish the concurrent scenarios (Schrider & Kern, 2018). The inferior performance of the ABC analyses could also be related to the choice of SuSt with overly high dimensionality (Robert et al., 2011; but see Kousathanas et al., 2016) or with little information to distinguish among the tested models. To reduce dimensionality, we adopted the suggestion of Perez, Bonatelli, et al. (2016) applying a PCA transformation
of the raw SuSt and taking only the most explanatory axes necessary to ensure at least 99% of the variance in the data. Other strategies usually applied to reduce dimensionality (reviewed in Prangle, 2015) include partial least squares regression (Wegmann et al., 2009), linear regression (Fearnhead & Prangle, 2012) or boosting (Aeschbacher et al., 2012).

Although our CNN results were very encouraging, we note that care must be taken in interpretation, as deep learning techniques can suffer from overfitting (Nguyen et al., 2015; Porti et al., 2017). To evaluate overfitting, we observed the history plot of the accuracy throughout the epochs (Figure S3) and adopted a cross-validation approach based on a test set, which consisted of simulated data that were not evaluated during training. The results pointed to high accuracy (Figure 4 and Figure S4), which is in agreement with the similarity of our empirical and simulated data (Figure S5). It is important to note that our CNN approach is very flexible and can potentially be combined with other recently developed machine learning applications. For instance, our CNN approach can be used to compare models and estimate parameters by using the CNN predictions (Mondal et al., 2019) or combining them with SuSt (Sanchez et al., 2020) to perform ABC. Some machine learning approaches have also shown promising results for the specific field of species delimitation, and they could also be combined or compared to our CNN method (Derkarabetian et al., 2019; Pei et al., 2018; Smith & Carstens, 2020).

### 4.2 Delimiting species in *P. aurisetus* complex

Both CNN and ABC approaches selected the "splitter" model without migration as the most likely scenario on the *P. aurisetus* data set (Table 2). It might seem counterintuitive that two methods relying on coalescent simulations and sequence data selected a scenario conceived using morphological taxonomy instead of those based on the BPP results. We believe this result is related to different assumptions from each method, which can return distinct results even when comparing methods based on the multispecies coalescent model (for example, see the comparison of BPP and spedeSTEM results in Figure 1 of Carstens et al., 2013), especially in recent diverging lineages. *Pilosocereus aurisetus* populations diverged very recently in
the last 1 million years (Perez, Bonatelli, et al., 2016, Perez et al., 2016b). Delimiting species with recent diversification has been considered challenging, and the main difficulty arises from the short time for species to accumulate diagnostic characteristics that allow taxonomic distinction (Salicini et al., 2011). A similar effect is observed in genetic data and is commonly referred to as incomplete lineage sorting (ILS). ILS appears when ancestral genetic variants persist as polymorphisms across successive diversification events. This phenomenon has been indicated as a major cause of discordance between gene trees and species trees in Cactaceae (Copetti et al., 2017).

Another issue that imposes additional challenges for species delimitation in *P. aurisetus* complex is its sky-island distribution, constraining gene flow and leading to high population divergence (Bonatelli et al., 2014). Topography-driven isolation in species of the Brazilian *campo rupestre* environments, such as *P. aurisetus*, has been considered an important driver of speciation and high micro-endemism in these landscapes (Vasconcelos et al., 2020; Zappi et al., 2017). The major concern in this kind of system is that MSC based methods may fail to discriminate between genetic structure associated with population isolation and species boundaries (Jackson et al., 2017; Leaché et al., 2018; Sukumaran & Knowles, 2017).

Although *gdi* appears as a possible solution for the over-splitting tendency of MSC methods (Jackson et al., 2017), mainly in highly structured systems, it is still not clear if it could lead to an opposite effect—over lumping differentiated species. Incorporating a priori knowledge on the structure associated with population-level processes could probably overcome this oversplitting tendency of MSC methods. An ideal scenario would incorporate data from a plethora of classes, such as morphological, ecological, and genetic data, to recover more robust species boundaries (Pinheiro et al., 2018). Since no single method can discriminate between these processes, the results from species delimitation using multispecies coalescent models

| Scenario            | CNN  | ABC  |
|---------------------|------|------|
| *P. aurisetus*      |      |      |
| Splitter            | 0.999| 0.962|
| Splitter +M         | 0.000| 0.002|
| Lumper              | 0.000| 0.000|
| Lumper +M           | 0.000| 0.000|
| BPP +GDI           | 0.000| 0.000|
| BPP_noGDI          | 0.001| 0.033|
| BPP_noGDI+M        | 0.000| 0.003|

| *D. melanogaster_D.simulans* |      |      |
|-------------------------------|------|------|
| One species                   | 0.008| 0.064|
| Two species                   | 0.992| 0.936|

| *D. melanogaster_D. sechellia* |      |      |
|--------------------------------|------|------|
| One species                    | 0.051| 0.474|
| Two species                    | 0.949| 0.526|

FIGURE 4 Cross-validation test to compare the accuracy of different model comparison methods (CNN—filled bars and ABC—dashed bars) and numbers of simulations. Each bar represents the probability of choosing the “correct” model (i.e., the model from which the data were simulated) over incorrect models. Each colour represents a different scenario, including (+M) or not migration between the OTUs.
(MSC) in sky-island systems should be treated as species hypotheses that need further confirmation from multiple sources (Sukumaran & Knowles, 2017).

4.3 Taxonomic implications for *P. aurisetus* complex

Historically, Cactaceae taxonomy reflected the amateur views, always looking for differences with the objective to describe all possible variation found within and between populations as new species. Zappi (1994), Taylor and Zappi (2004) and Hunt et al. (2006) sought similarities and relationships between species, leading to a lumping approach. The present species delimitation investigation within the *P. aurisetus* complex is discordant from the currently recognized taxonomic diversity of this taxon (Hunt et al., 2006), recovering the “splitter” hypothesis with higher probability than different models. Convergent evolution as well as the level of plasticity commonly found in cactus species (Guerrero et al., 2018) might be related to the disagreement in the number of species delimited by current taxonomy and the detected genetic lineages, as observed within the genus *Pilosocereus* (Bonatelli et al., 2014; Perez et al., 2016a).

Furthermore, the “splitter” hypothesis is in line with the sky-island distribution and the pronounced population geographic isolation in this taxon, which agrees with the high incidence of microendemic plant species in the campo rupestre landscapes (Miola et al., 2021). Therefore, our results show the need for a taxonomic revision of the *P. aurisetus* complex and provide recommendations for a taxonomic treatment that better reflects the diversity of species. In this context, our “splitter” hypothesis may be used as a guide to investigate new diagnostic characters among the proposed species, contributing to fill gaps in the traditional taxonomy of cactus. Previous genetic studies have identified well-differentiated lineages within *P. aurisetus* complex and related species, with monophyletic groups more frequently being associated with the geographic distribution than the proposed taxonomic boundaries (Figure 2; Bonatelli et al., 2014; Calvente et al., 2017; Khan et al., 2018; Lavor et al., 2019; Perez et al., 2016a). Different taxonomic circumscriptions have been considered in the past for some of the recovered lineages within the *P. aurisetus* complex that agree with our results. For instance, the populations occurring in the Serra Negra Mountains (ITA and PMN localities), in the centre (MEN and INA) and in the northern distribution of the species (EDB, BOV and ODA localities) were previously reported as *P. aurisetus* subsp. *densilanatus*, *P. aurisetus* subsp. *wermannianus*, and *P. aurisetus* subsp. *supthutianus* (Braun & Esteves, 1995), respectively. These subspecies propositions were based on some level of morphological variation that initially seemed rather distinctive, such as the densely wooly stems observed in the ITA populations (Figure S1). However, Taylor and Zappi (2004) argued that such differences were not sufficiently important to merit recognition as an additional subspecies. These authors reasoned that if this level of morphological difference was to be recognized, the logical extension would be to give similar status to other populations that show slightly divergent characteristics in terms of size, white-woolly stems, and glaucousness.

5 CONCLUSION

Although the splits within species could be an artefact of the MSC, we might recognize previous taxonomic issues that agree with the inferred species in this investigation. *Pilosocereus aurisetus* complex exhibits many biological characteristics that make species delimitation challenging, such as recent divergence, a sky-island distribution and unstable taxonomic history. Here, we sought to investigate the species limits in this complex system and, for the first time, incorporate a deep learning approach to select the best species hypothesis according to our data and simulations. Finally, we stress that all the species hypotheses reported here should be considered, as any of the taxonomic species in fact are hypotheses prone to validation with other methods and sources of information.

ACKNOWLEDGEMENTS

This work was supported by grants from the São Paulo Research Foundation (FAPESP) (2015/06160-5 to EMM, 2012/22943-1 to MFP, and 2012/22857-8 to IASB); the National Council for Scientific and Technological Development (CNPq) (03940/2019-0 to EMM, 305301/2018-7 to DCZ); and the Coordination for the Improvement of Higher Education Personnel (CAPES) (Finance Code 001 to MRB).

We thank Heidi Utsunomiya and Juliana de Fátima Martinez for laboratory assistance and Gerardus Olsthoorn and Marlon Machado for sampling assistance. We also thank the four anonymous reviewers for their helpful comments.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Manolo F. Perez, Isabel A. S. Bonatelli, Fernando F. Franco, Daniela C. Zappi, Nigel P. Taylor, and Evandro M. Moraes conceived and designed the study. Manolo F. Perez performed the analysis, with critical inputs from Isabel A. S. Bonatelli, Evandro M. Moraes, and

---

**TABLE 3** Computational time in seconds to estimate the species tree and delimit (SpTree/Del) in BPP (average time for each separate run), perform 10,000 (10 K) simulations per model (Sims/Model) for CNN and 100,000 (100K) for ABC, train the neural network (only CNN), and perform cross validation (CV) and model selection (ModSel) for CNN and ABC approaches.

| Model | SpTree/Del | CNN | ABC |
|-------|------------|-----|-----|
|       | 2.89E+03   | 10K | 1.19E+03 |
|       | 6.83E+03   | CV  | 7.77E+04 |
|       | 1.35E+03   | modSel | 1.52E+06 |

---

**PerEZ ETL Al.**
Fernando F. Franco. Manolo F. Perez, Isabel A. S. Bonatelli and Monique Romeo-Brito drafted the manuscript. Daniela C. Zappi and Nigel P. Taylor guided the sampling of the taxonomic complexity of the study group, identified the species, and helped with the fieldwork. All authors participated in discussions and contributed critically to data interpretation and the final version of the text.

DATA AVAILABILITY STATEMENT
All sequences have been deposited in GenBank with accession numbers MZ2509667–MZ2510910. All scripts and data sets used are available in GitHub: https://github.com/manolofperez/CNN_spDelimitation_Piloso.

ORCID
Manolo F. Perez https://orcid.org/0000-0002-4642-7793
Evandro M. Moraes https://orcid.org/0000-0003-4197-0794

REFERENCES
Aeschbacher, S., Beaumont, M. A., & Futschik, A. (2012). A novel approach for choosing summary statistics in approximate Bayesian computation. Genetics, 192, 1027–1047. https://doi.org/10.1534/genetics.112.143164
Alvarado-Sizzo, H., Casas, A., Parra, F., Arreola-Nava, H. J., Terrazas, T., & Sánchez, C. (2018). Species delimitation in the Stenocereus griseus (Cactaceae) species complex reveals a new species, S. Haustecorum. PLoS One, 13, e0190385. https://doi.org/10.1371/journal.pone.0190385
Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. Genetics, 162, 2025–2035. https://doi.org/10.1093/genetics/162.4.2025
Bonatelli, I. A. S., Perez, M. F., Townsend, A. P., Taylor, N. P., Zappi, D. C., Machado, M. C., Koch, I., Pires, A. H. C., & Moraes, E. M. (2014). Interglacial microrefugia and diversification of a cactus species complex: Phylogeography and palaeodistributional reconstructions for Pilosocereus aurisetus and allies. Molecular Ecology, 23, 3044–3063.
Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., Suchard, M. A., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: A software platform for Bayesian evolutionary analysis. PLoS Computational Biology, 10, e1003537. https://doi.org/10.1371/journal.pcbi.1003537
Braun, P. J., & Esteves, E. P. (1995). Nieuwe combinaties en namen voor Succulenta (Netherlands), 74, 134–135.
Calvente, A., Moraes, E. M., Lavor, P., Bonatelli, I. A. S., Nacaguma, P., Versieux, L. M., Taylor, N. P., & Zappi, D. C. (2017). Phylogenetic analyses of Pilosocereus (Cactaceae) inferred from plastid and nuclear sequences. Botanical Journal of the Linnean Society, 183, 25–38.
Camargo, A., Morando, M., Avila, L. J., & Sites, J. W. Jr (2012). Species delimitation with ABC and other coalescent-based methods: A test of accuracy with simulations and an empirical example with lizards of the Liolaemus darwini complex (Squamata: Liolaemidae). Evolution, 66, 2834–2849. https://doi.org/10.1111/j.1558-5646.2012.01640.x
Campillo, L. C., Barlet, A. J., & Thomson, R. C. (2020). Model-based species delimitation: Are coalescent species reproducibly isolated? Systematic Biology, 69(4), 708–721. https://doi.org/10.1093/sysbio/syz072
Carstens, B. C., Pelletier, T. A., Reid, N. M., & Satler, J. D. (2013). How to fail at species delimitation? Molecular Ecology, 22, 4369–4383.
Copetti, D., Búrquez, A., Bustamante, E., Charboneau, J. L. M., Childs, K. L., Eguiarte, L. E., Lee, S., Liu, T. L., McMahon, M. M., Whiteman, N. K., Winga, R. A., Wojciechowski, M. F., & Sanderson, M. J. (2017). Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. Proceedings of the National Academy of Sciences of the United States of America, 114, 12003–12008. https://doi.org/10.1073/pnas.1706367114
Csilléry, K., François, O., & Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC). Methods in Ecology and Evolution, 3, 475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x
de Queiroz, K. (2007). Species concepts and species delimitation. Systematic Biology, 56, 879–886. https://doi.org/10.1080/10653507.150701701083
Denham, S. S., Bignonea, N. F., Johnson, L. A., & Pozner, R. E. (2019). Using integrative taxonomy and multispecies coalescent models for phylogeny reconstruction and species delimitation within the “Nastanthus–Gamocarphe” clade (Calyceraceae). Molecular Phylogenetics and Evolution, 130, 211–226. https://doi.org/10.1016/j.mpev.2018.10.015
Derkarabetian, S., Castillo, S., Koo, P. K., Ovchinnikov, S., & Hedin, M. (2019). A demonstration of unsupervised machine learning in species delimitation. Molecular Phylogenetics and Evolution, 139, 106562. https://doi.org/10.1016/j.ympev.2019.106562
Domingos, F. M., Bosque, R. J., Cassimiro, J., Colli, G. R., Rodrigues, M. T., Santos, M. G., & Beheragaray, L. B. (2014). Out of the deep: Cryptic speciation in a Neotropical gecko (Squamata, Phyllodactylidae) revealed by species delimitation methods. Molecular Phylogenetics and Evolution, 80, 113–124. https://doi.org/10.1016/j.mpev.2014.07.022
Drummond, A. J., & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology, 7, 214. https://doi.org/10.1186/1471-2148-7-214
Ence, D. D., & Carstens, B. C. (2011). SpeeDESTM: A rapid and accurate method for species delimitation. Molecular Ecology Resources, 11, 473–480. https://doi.org/10.1111/j.1755-0998.2010.02947.x
Fearnhead, P., & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic ABC. Journal of the Royal Statistical Society Series B, 74, 419–474.
Flagel, L., Brandvain, Y., & Schröder, D. R. (2019). The unreasonable effectiveness of convolutional neural networks in population genetic inference. Molecular Biology and Evolution, 36, 220–238. https://doi.org/10.1093/molbev/msy224
Flouri, T., Jiao, X., Rannala, B., & Yang, Z. (2018). Species tree inference using integrative taxonomy and multispecies coalescent models. Molecular Ecology, 35(10), 2585–2593.
Fonseca, E. M., Colli, G. R., Werneck, F. P., & Carstens, B. C. (2021). Phylogeographic model selection using convolutional neural networks. Molecular Ecology Resources, https://doi.org/10.1111/1755-0998.13427
Guerrero, P. C., Majure, L. C., Cornejo-Romero, A., & Hernández-Hernández, T. (2018). Phylogenetic relationships and evolutionary trends in the Cactus family. Journal of Heredity, 110, 4–21. https://doi.org/10.1093/jhered/esy064
Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., & Bromfield, R. T. (2016). Sequence capture versus restriction site associated DNA sequencing for shallow systematics. Systematic Biology, 65, 910–924. https://doi.org/10.1093/sysbio/syw036
Hernández-Hernández, T., Hernández, H. M., Arturo, D.-N., Puente, R., Eguiarte, L. E., & Magallón, S. (2011). Phylogenetic relationships and evolution of growth form in Cactaceae (Caryophyllales, Eudicotyledoneae). American Journal of Botany, 98, 44–61. https://doi.org/10.3732/ajb.1000129
Hunt, D., Taylor, N. P., & Charles, G. (2006). The new Cactus Lexicon. Atlas & Text. dh Books.
combination with approximate Bayesian computation. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13224

Schrider, D. R., & Kern, A. D. (2018). Supervised machine learning for population genetics: A new paradigm. *Trends in Genetics*, 34, 301–312. https://doi.org/10.1016/j.tig.2017.12.005

Smith, M. L., & Carstens, B. C. (2020). Disentangling the process of speciation using machine learning. *Evolution*, 74, 216–229.

Solís-Lemus, C., Knowles, L. L., & Ané, C. (2015). Bayesian species delimitation combining multiple genes and traits in a unified framework. *Evolution*, 69, 492–507. https://doi.org/10.1111/evo.12582

Souza, F. H. S., Perez, M. F., Bertollo, L. A. C., Oliveira, E. A., Lavoué, S., Gestaich, C. C., Ráb, P., Ezaz, T., Liehr, T., Viana, P. F., Feldberg, E., & Cioffi, M. B. (2019). Interspecific genetic differences and historical demography in South American Arowanas (Osteoglossiformes, Osteoglossidae, Osteoglossum). *Genes*, 10, 693. https://doi.org/10.3390/genes10090693

Sukumaran, J., & Knowles, L. L. (2017). Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences of the United States of America*, 114, 1607–1612. https://doi.org/10.1073/pnas.1607921114

Taylor, N. P., & Zappi, D. C. (2004). *Cacti of Eastern Brazil*. Royal Botanic Gardens.

Untergrasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3 – New capabilities and interfaces. *Nucleic Acids Research*, 40, e115. https://doi.org/10.1093/nar/gks596

Vasconcelos, T. N. C., Alcantara, S., Andrino, C. O., Forest, F., Reginato, M., Simon, M. F., & Pirani, J. R. (2020). Fast diversification through a mosaic of evolutionary histories characterizes the endemic flora of ancient Neotropical mountains. *Proceedings of the Royal Society B*, 287, 20192933. https://doi.org/10.1098/rspb.2019.2933

Vázquez-Castillo, S., Miranda-Jácome, A., & Inzunza, E. R. (2019). Patterns of frugivory in the columnar cactus *Pilosocereus leucocephalus*. *Ecology and Evolution*, 9, 1268–1277.

Wegmann, D., Leuenberger, C., & Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, 182, 1207–1218. https://doi.org/10.1534/genetics.109.102509

Weyenberg, G., Huggins, P. M., Schardl, C. L., Howe, D. K., & Yoshida, R. (2014). *kdetrees*: Non-parametric estimation of phylogenetic tree distributions. *Bioinformatics*, 30, 2280–2287. https://doi.org/10.1093/bioinformatics/btu258

Yang, Z., & Rannala, B. (2014). Unguided species delimitation using DNA sequence data from multiple loci. *Molecular Biology and Evolution*, 31, 3125–3135. https://doi.org/10.1093/molbev/msu279

Zappi, D. C. (1994). *Pilosocereus (Cactaceae)*. The genus in Brazil. Royal Botanic Gardens.

Zappi, D. C., Moro, M. F., Meagher, T. R., & Lughadha, E. N. (2017). Plant biodiversity drivers in Brazilian campos rupestres: Insights from phylogenetic structure. *Frontiers in Plant Science*, 8, 2141. https://doi.org/10.3389/fpls.2017.02141

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

**How to cite this article:** Perez, M. F., Bonatelli, I. A. S., Romeiro-Brito, M., Franco, F. F., Taylor, N. P., Zappi, D. C., & Moraes, E. M. (2022). Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system. *Molecular Ecology Resources*, 22, 1016–1028. https://doi.org/10.1111/1755-0998.13534