Received: 3 January 2021 | Revised: 16 September 2021 | Accepted: 12 October 2021

DOI: 10.1111/1755-0998.13534

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system

Manolo F. Perez[1,2] | Isabel A. S. Bonatelli[1,3] | Monique Romeiro-Brito[1] | Fernando F. Franco[1] | Nigel P. Taylor[4] | Daniela C. Zappi[5] | Evandro M. Moraes[1]

[1]Departamento de Biologia, Universidade Federal de São Carlos, Sorocaba, Brazil

[2]Departamento de Genética e Evolução, Universidade Federal de São Carlos, São Carlos, Brazil

[3]Departamento de Ecologia e Biologia Evolutiva, Universidade Federal de São Paulo, Diadema, Brazil

[4]University of Gibraltar, The Alameda, Gibraltar

[5]Programa de Pós Graduação em Botânica, Instituto de Ciências Biológicas, Universidade de Brasília, Brasília, Brazil

Correspondence
Evandro M. Moraes, Departamento de Biologia, Universidade Federal de São Carlos, Sorocaba, SP, Brazil.
Email: emarsola@ufscar.br

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number: 03940/2019-0 and 305301/2018-7; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Grant/Award Number: 001; Fundação de Amparo à Pesquisa do Estado de São Paulo, Grant/Award Number: 2012/22857-8, 2012/22943-1 and 2015/06160-5

Abstract

Delimiting species boundaries is a major goal in evolutionary biology. An increasing volume of literature has focused on the challenges of investigating cryptic diversity within complex evolutionary scenarios of speciation, including gene flow and demographic fluctuations. New methods based on model selection, such as approximate Bayesian computation, approximate likelihoods, and machine learning are promising tools arising in this field. Here, we introduce a framework for species delimitation using the multispecies coalescent model coupled with a deep learning algorithm based on convolutional neural networks (CNNs). We compared this strategy with a similar ABC approach. We applied both methods to test species boundary hypotheses based on current and previous taxonomic delimitations as well as genetic data (sequences from 41 loci) in *Pilosocereus aurisetus*, a cactus species complex with a sky-island distribution and taxonomic uncertainty. To validate our method, we also applied the same strategy on data from widely accepted species from the genus *Drosophila*. The results show that our CNN approach has a high capacity to distinguish among the simulated species delimitation scenarios, with higher accuracy than ABC. For the cactus data set, a splitter hypothesis without gene flow showed the highest probability in both CNN and ABC approaches, a result agreeing with previous taxonomic classifications and in line with the sky-island distribution and low dispersal of *P. aurisetus*. Our results highlight the cryptic diversity within the *P. aurisetus* complex and show that CNNs are a promising approach for distinguishing complex evolutionary histories, even outperforming the accuracy of other model-based approaches such as ABC.

KEYWORDS
approximate Bayesian computation, convolutional neural networks, deep learning, fragmented systems, recent diversification, species delimitation

1 | 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

Manolo F. Perez and Isabel A. S. Bonatelli contributed equally to this article.

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, cytogenetics, 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, 2011). Highly fragmented systems impose critical caveats on such methods, potentially resulting in the oversplit of existing diversity, and thus, highly subdivided entities (Sukumaran & Knowles, 2017). Jackson et al. (2017) proposed the incorporation of the genealogical divergence index (*gdi*) as an attempt to reduce the delimitation of population structure, instead of species limits. Furthermore, the developers of the Bayesian delimitation method implemented in Bayesian phylogenetics and phylogeography (BPP); Yang and Rannala (2014) described an approach to integrate *gdi* on the outputs of their software and showed that such a strategy was efficient in mitigating the effect of over splitting (Leaché et al., 2018). However, Rannala and Yang (2020) raised a number of concerns on the use of *gdi*, such as the large interval of *gdi* values (0.2–0.7) where delimitation is ambiguous, and misleading results when there are huge differences among the population sizes of the putative species. Model selection methods using simulated data sets under competing delimitation hypotheses are a promising tool for taxa that have potentially experienced a complex evolutionary history, including recurrent gene flow and demographic fluctuations. These methods also allow us to easily test for different assignments and topologies of the analysed samples into populations/species, a procedure that is not straightforward with other approaches (but see Leaché et al., 2014). Examples of such approaches applied for species delimitation include 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 (Leaché 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 occurence 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

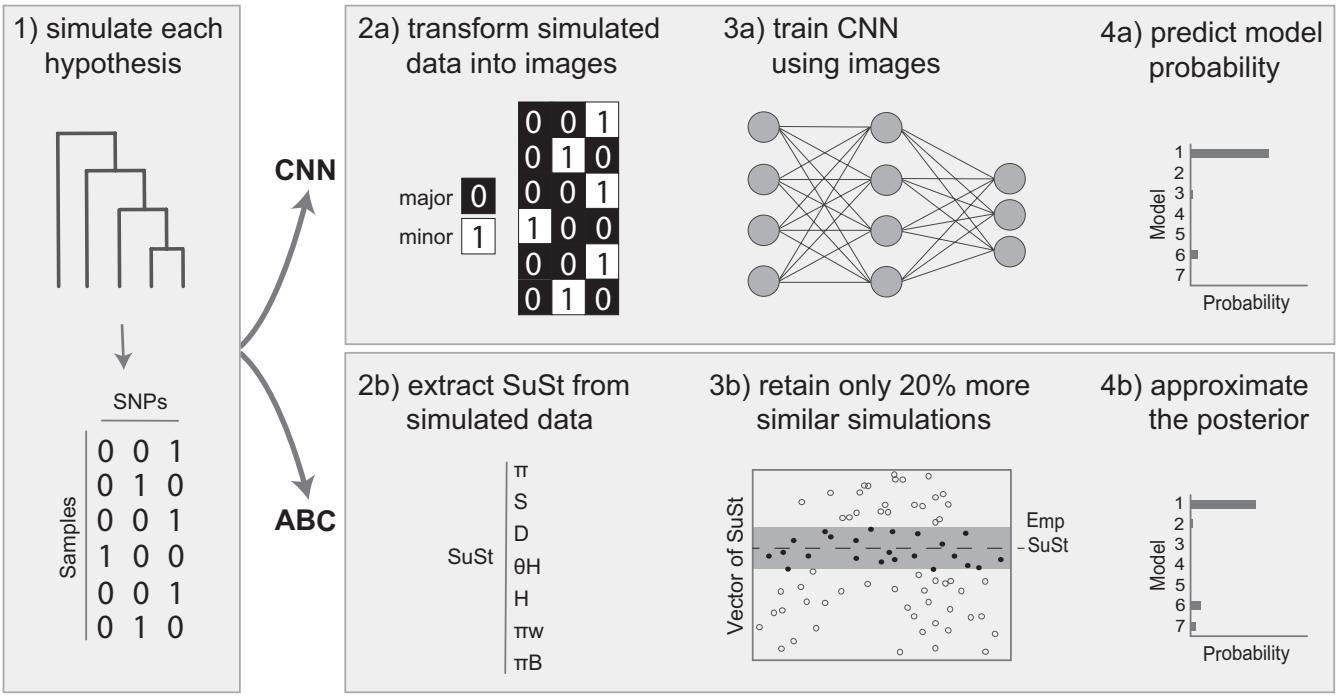

FIGURE 1 A general protocol for species delimitation in the *P. aurisetus* species complex, comparing the adopted CNN and ABC approaches. The first step for both CNN and ABC is to (1) simulate genetic data for each delimitation hypothesis (10,000 simulations per model in CNN and 100,000 in ABC). Then, for the CNN procedure, we (2a) convert the simulated data to images, with black pixels representing the major and white pixels representing the minor frequency alleles for each segregating site; (3a) use the simulated data images to train a neural network that recognize simulations generated from each model; and (4a) predict the softmax probability of each model using the trained CNN on our empirical data. For the ABC approach, we (2b) extract summary statistics (SuSt) averaged over loci for each simulation and for the empirical data; (3b) perform the rejection step, retaining only the 20% most similar simulations; and (4b) obtain the posterior probability of each model by approximating the posterior

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,

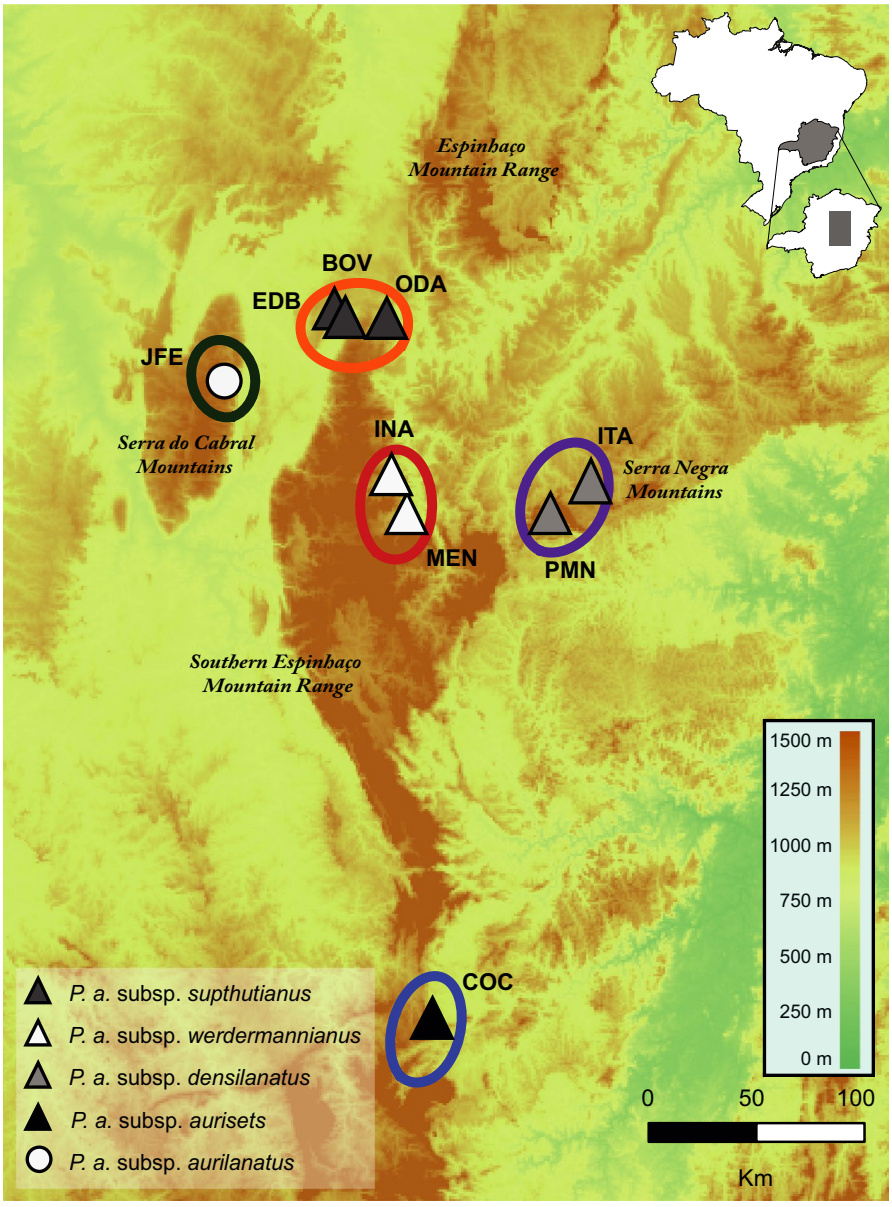

FIGURE 2 (a) Geographic distribution of the sampled localities and elevational range in eastern Brazil (codes according to Table 1). Symbols (triangles and circle) represent the currently recognized *P. aurisetus* subspecies, and grey shades represent heterotypic synonyms of *P. aurisetus* subsp. *aurisetus* (in parenthesis). The delimited species are shown with lines colored according to the delimited entities in Figure 3

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 polymorphism 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 (Lanfear 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 5×10^8 MCMC iterations each, sampling trees every 10^4 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 (Flouri 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 (rjMCMC) 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). θ and τ priors followed an inverse gamma (IG) distribution with $\alpha = 3$ and a β of 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), (6) 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 loci 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 assignment 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

					Delimitation hypothesis			
Taxon/locality	Code	N	Longitude	Latitude	"Splitter"	"Lumper"	"BPP_noGDI"	"BPP_GDI"
P. a. subsp. *aurisetus*								
Cocais-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	−43,755833	−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 Felício-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*, *nhx1*, *psbA-psbB*, *psbC*). No recombination was detected at any of the remaining loci, but kdetrees 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.

densilanatus, 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 *gdi* index, all estimates were below 0.7 and only three (*P. aurisetus* subsp. *aurilanatus*—JFE; *P. a.* subsp. *aurisetus*—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

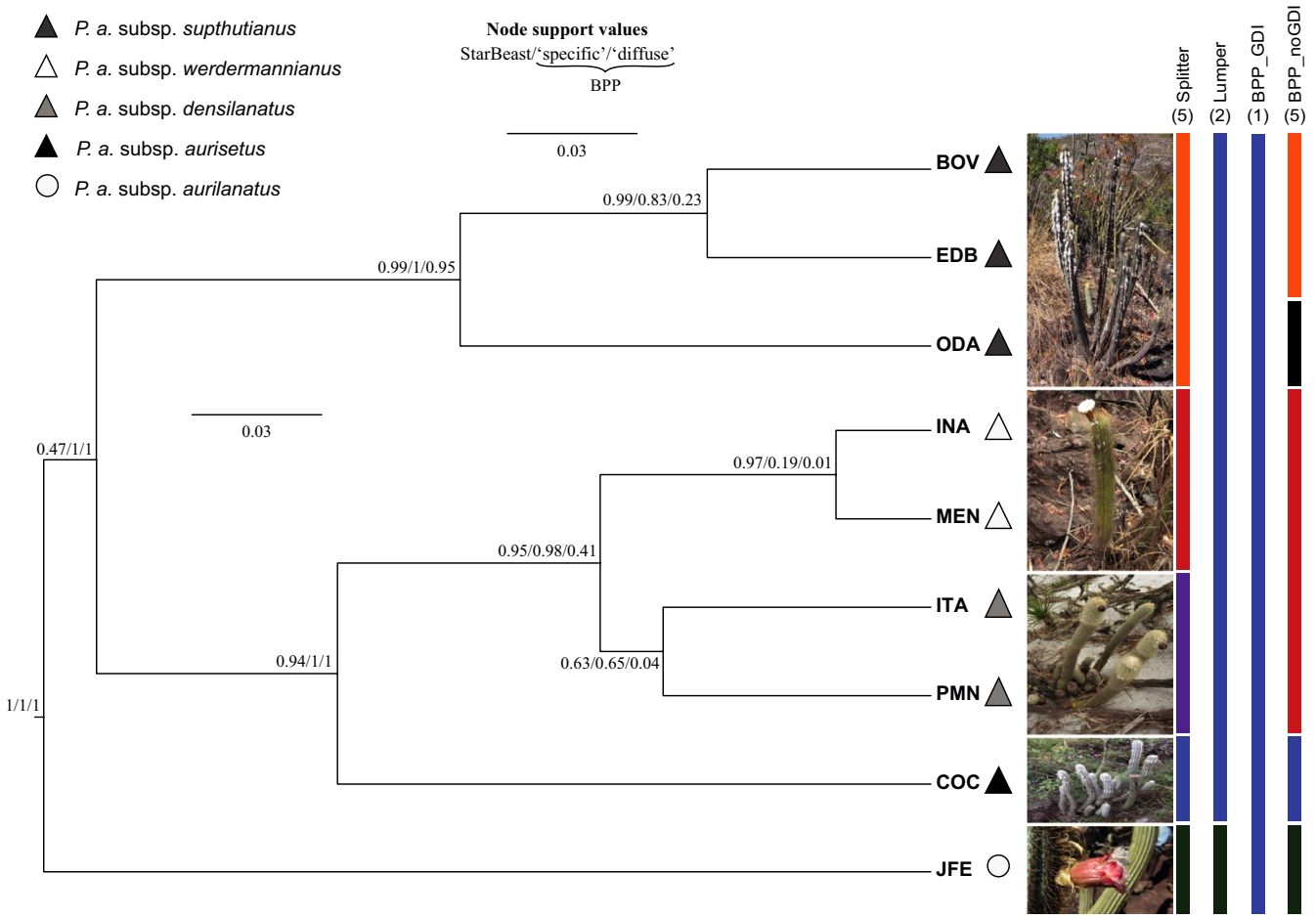

FIGURE 3 Species delimitation results are represented as bars at the tips of the topology obtained in BPP. Values in parentheses above each bar present the number of species for each hypothesis. The numbers at the nodes represent StarBeast PP/BPP "specific"/and "diffuse" prior values. Symbols (triangles and circles) represent both currently recognized and synonymized (in parenthesis) *P. aurisetus* taxa. A representative of each inferred species is shown, next to the limits according to the "splitter" hypothesis

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; Ponti 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

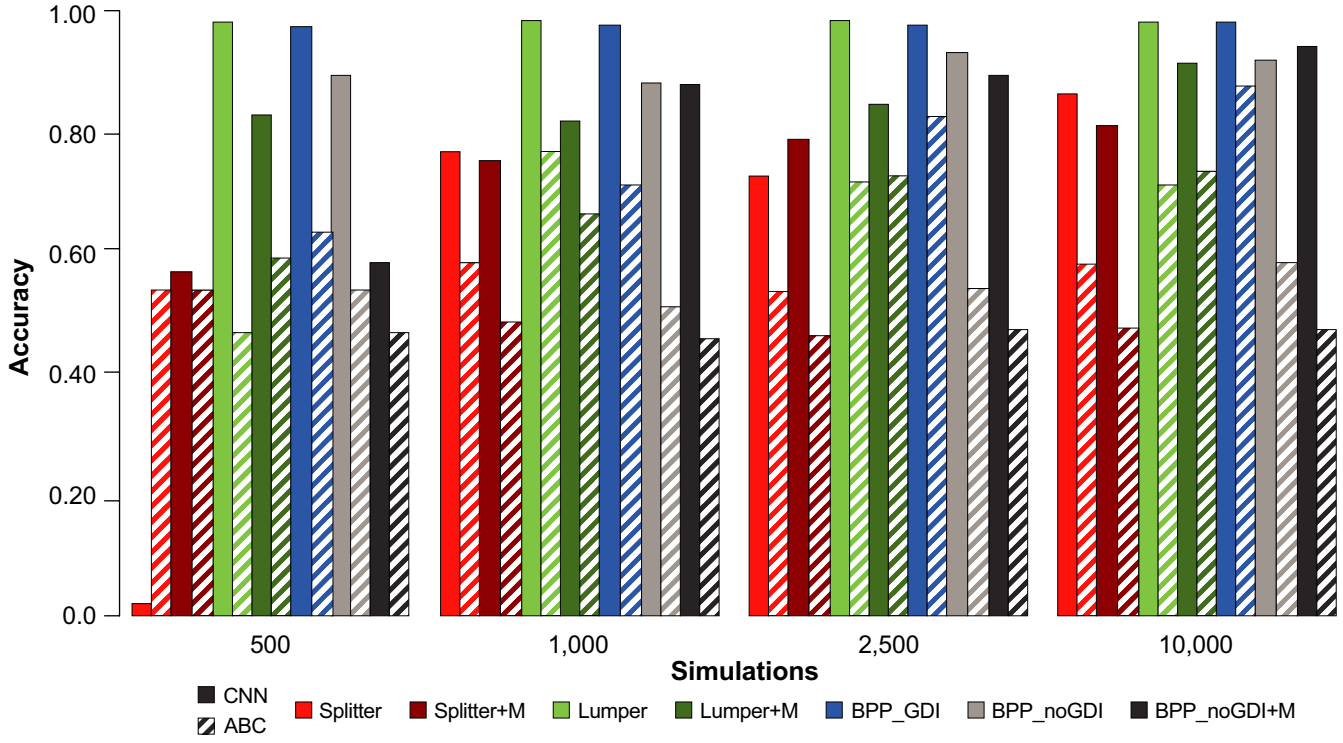

FIGURE 4 Cross-validation test to compare the accuracy of different model comparison methods (CNN—filled bars and ABC—dashed bars) and numbers of simulations. Each bar represents the probability of choosing the "correct" model (i.e., the model from which the data were simulated) over incorrect models. Each colour represents a different scenario, including (+M) or not migration between the OTUs

TABLE 2 Model selection results for the *P. aurisetus* species complex and the two *Drosophila* species pairs using empirical data. CNN values are softmax probabilities for each scenario, while posterior probabilities are shown for ABC. The scenario with the highest probability for each method is shown in bold

	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**

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 overspliting 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

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

BPP	CNN			ABC		
SpTree/Del	10K Sims/model	Training	CV/modSel	100K Sims/model	CV	ModSel
2.89E+03	6.83E+03	1.35E+03	<1	7.77E+04	1.52E+06	1.19E+03

(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. *werdermannianus*, 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-wooly 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, incorporated 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

Fernando F. Franco. Manolo F. Perez, Isabel A. S. Bonatelli and Monique Romeiro-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 MZ509667—MZ510910. 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. Huastecorum*. *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*, *104*, e1003537. https://doi.org/10.1371/journal.pcbi.1003537

Braun, P. J., & Esteves, E. P. (1995). Nieuwe combinaties en namen voor cactussen uit Brazilië, Bolivia en Paraguay. *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 darwinii* complex (Squamata: Liolaemidae). *Evolution*, *66*, 2834–2849. https://doi.org/10.1111/j.1558-5646.2012.01640.x

Campillo, L. C., Barley, A. J., & Thomson, R. C. (2020). Model-based species delimitation: Are coalescent species reproductively 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., McMahong, M. M., Whiteman, N. K., Winga, R. A., Wojciechowskii, 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/10.1080/10635150701701083

Denham, S. S., Brignonea, 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–Gamocarpha" clade (Calyceraceae). *Molecular Phylogenetics and Evolution*, *130*, 211–226. https://doi.org/10.1016/j.ympev.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., & Beheregaray, 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.ympev.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). SpedeSTEM: 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., & Schrider, 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 with BPP using genomic sequences and the multispecies coalescent. *Molecular Biology and Evolution*, *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., & Brumfield, 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.

Jackson, N. D., Carstens, B. C., Morales, A. E., & O'Meara, B. C. (2017). Species delimitation with gene flow. *Systematic Biology*, *66*, 799–812.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*, 772–780. https://doi.org/10.1093/molbev/mst010

Khan, G., Godoy, M. O., Franco, F. F., Perez, M. F., Taylor, N. P., Zappi, D. Z., Machado, M. C., & Moraes, E. M. (2018). Extreme population subdivision or cryptic speciation in the cactus *Pilosocereus jauruensis:* A taxonomic challenge posed by a naturally fragmented system. *Systematics and Biodiversity*, *16*, 188–199.

Kousathanas, A., Leuenberger, C., Helfer, J., Quinodoz, M., Foll, M., & Wegmann, D. (2016). Likelihood-free inference in high-dimensional models. *Genetics*, *203*, 893–904. https://doi.org/10.1534/genetics.116.187567

Lanfear, R., Calcott, B., Ho, S. Y., & Guindon, S. (2012). PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, *29*(6), 1695–1701. https://doi.org/10.1093/molbev/mss020

Lavor, P., Calvente, A., Versieux, L. M., & Sanmartin, I. (2019). Bayesian spatio-temporal reconstruction reveals rapid diversification and Pleistocene range expansion in the widespread columnar cactus *Pilosocereus*. *Journal of Biogeography*, *46*, 238–250.

Leaché, A. D., Fujita, M. K., Minin, V. N., & Bouckaert, R. R. (2014). Species delimitation using genome-wide SNP data. *Systematic Biology*, *63*(4), 534–542.

Leaché, A. D., Zhu, T., Rannala, B., & Yang, Z. (2018). The spectre of too many species. *Systematic Biology*, *68*, 168–181. https://doi.org/10.1093/sysbio/syy051

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 1303.3997.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Math, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Luo, A., Ling, C., Ho, S. Y. W., & Zhu, C. (2018). Comparison of methods for molecular species delimitation across a range of speciation scenarios. *Systematic Biology*, *67*, 830–846. https://doi.org/10.1093/sysbio/syy011

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*, 10–12. https://doi.org/10.14806/ej.17.1.200

McGuire, G., & Wright, F. (2000). TOPAL 2.0: Improved detection of mosaic sequences within multiple alignments. *Bioinformatics*, *16*, 130–134.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110

Milne, I., Lindner, D., Bayer, M., Husmeier, D., McGuire, G., Marshall, D. F., & Wright, F. (2008). TOPALi v2: A rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. *Bioinformatics*, *25*, 126–127. https://doi.org/10.1093/bioinformatics/btn575

Miola, B. P. D., Ramos, V. D. R., & Silveira, F. A. O. (2021). A brief history of research in campo rupestre: Identifying research priorities and revisiting the geographical distribution of an ancient, widespread Neotropical biome. *Biological Journal of the Linnean Society*, *133*, 464–480.

Mondal, M., Bertranpetit, J., & Lao, O. (2019). Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nature Communications*, *10*, 246. https://doi.org/10.1038/s41467-018-08089-7

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

Nguyen, A., Yosinski, J., & Clune, J. (2015). *Deep neural networks are easily fooled: High confidence predictions for unrecognizable images*. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 427–436.

Oliveira, E. A. D., Perez, M. F., Bertollo, L. A. C., Gestich, C. C., Ráb, P., Ezaz, T., Souza, F. H. S., Viana, P., Feldberg, E., Oliveira, E. H. C. D., & Cioffi, M. B. (2020). Historical demography and climate driven distributional changes in a widespread Neotropical freshwater species with high economic importance. *Ecography*, *43*, 1291–1304. https://doi.org/10.1111/ecog.04874

Pei, J., Chong, C., Xin, L., Bin, L., & Yufeng, W. (2018). CLADES: A classification-based machine learning method for species delimitation from population genetic data. *Molecular Ecology Resources*, *18*, 1144–1156. https://doi.org/10.1111/1755-0998.12887

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

Perez, M. F., Bonatelli, I. A. S., Moraes, E. M., & Carstens, B. C. (2016). Model-based analysis supports interglacial refugia over long-dispersal events in the diversification of two South American cactus species. *Heredity*, *116*, 550–557. https://doi.org/10.1038/hdy.2016.17

Perez, M. F., Carstens, B. C., Rodrigues, G. L., & Moraes, E. M. (2016a). Anonymous nuclear markers reveal taxonomic incongruence and long-term disjunction in a cactus species complex with continental-island distribution in South America. *Molecular Phylogenetics and Evolution*, *95*, 11–19. https://doi.org/10.1016/j.ympev.2015.11.005.

Perez, M. F., Carstens, B. C., Rodrigues, G. L., & Moraes, E. M. (2016b). Anonymous nuclear markers data supporting species tree phylogeny and divergence time estimates in a cactus species complex in South America. *Data Brief*, *6*, 456–460. https://doi.org/10.1016/j.dib.2015.12.002.

Pinheiro, F., Dantas-Queiroz, M. V., & Palma-Silva, C. (2018). Plant species complexes as models to understand speciation and evolution: A review of South American Studies. *Critical Reviews in Plant Sciences*, *37*, 54–80. https://doi.org/10.1080/07352689.2018.1471565

Ponti, M., Ribeiro, L., Nazare, T., Bui, T., & Collomosse, J. (2017). *Everything you wanted to know about Deep Learning for Computer Vision but were afraid to ask*. In: 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T), 17–41. IEEE.

Prangle, D. (2015). Summary statistics in approximate Bayesian computation. *arXiv* preprint arXiv:1512.05633.

Rambaut, A., Suchard, M. A., Xie, D., & Drummond, A. J. (2014). *Tracer v1.6*. http://beast.bio.ed.ac.uk/Tracer

Rannala, B., & Yang, Z. (2020). Species delimitation. In C. Scornavacca, F. Delsuc, & N. Galtier (Eds.), *Phylogenetics in the genomic era*. No commercial publisher | Authors open access book, 5.5:1–18. ffhal-02536468.

Robert, C. P., Cornuet, J. M., Marin, J. M., & Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 15112–15117. https://doi.org/10.1073/pnas.1102900108

Rocha, E. A., Domingos-Melo, A., Zappi, D. C., & Machado, I. C. (2019). Reproductive biology of columnar cacti: Are bats the only protagonists in the pollination of *Pilosocereus*, a typical chiropterophilous genus? *Folia Geobotanica*, *54*, 239–256.

Salicini, I., Ibáñez, C., & Juste, J. (2011). Multilocus phylogeny and species delimitation within the Natterer's bat species complex in the Western Palearctic. *Molecular Phylogenetics and Evolution*, *61*, 888–898. https://doi.org/10.1016/j.ympev.2011.08.010

Sanchez, T., Cury, J., Charpiat, G., & Jay, F. (2020). Deep learning for population size history inference: Design, comparison and

combination with approximate Bayesian computation. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13224

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., Gestich, 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