Deep Learning Approaches to the Phylogenetic Placement of Extinct Pollen Morphotypes

Marc-Élie Adaïmé[1], Shu Kong[2], Surangi W. Punyasena[1,*]

1 Department of Plant Biology, University of Illinois Urbana–Champaign, Urbana, IL 61801
2 Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843
[*]spunya1@illinois.edu

Significance Statement

Machine learned features from deep neural networks can do more than categorize and classify biological images. We demonstrate that these features can also be used to quantify morphological differences among pollen taxa, discover novel morphotypes, and place fossil specimens on a phylogeny using Bayesian inference. Deep learning can be used to characterize and identify and morphological features with evolutionary significance. These features can then be used to infer phylogenetic distance. This approach fundamentally changes how fossil pollen morphology can be interpreted, allowing greater evolutionary inference of fossil pollen specimens. The analysis framework, however, is not specific to pollen and can be generalized to other taxa and other biological images.

Abstract

The phylogenetic interpretation of pollen morphology is limited by our inability to recognize the evolutionary history embedded in pollen features. Deep learning offers tools for connecting morphology to phylogeny. Using neural networks, we developed an explicitly phylogenetic toolkit for analyzing the overall shape, internal structure, and texture of a pollen grain. Our analysis pipeline classifies pollen specimens and determines whether a testing specimen is from any novel species based on uncertainty estimate. Features of novel specimens are passed to a multi-layer perceptron network trained to transform these features into predicted phylogenetic distances from known taxa. We used these predicted distances to place specimens in a phylogeny using Bayesian inference. We trained and evaluated our models using optical superresolution micrographs of 30 *Podocarpus* species. We then used trained models to place nine fossil *Podocarpidites* specimens within the phylogeny. We demonstrate that the phylogenetic history encoded in pollen morphology can be recognized by neural networks and that deep-learned features can be used in phylogenetic placement. Our approach makes extinction and speciation events that would otherwise be masked by the limited taxonomic resolution of the fossil pollen record visible to palynological analysis.

1 Introduction

Machine learning in the computational biological literature has largely focused on biological classification and categorization, that is, developing neural networks for K-way classification, where K represents the number of taxa in a training set. Neural networks are used to identify features that define individual classes (taxa) and not the features that define the evolutionary relationships among taxa. This poses a challenge when classifying fossil specimens, many of which may be derived from extinct species that are entirely new to a trained network.

New methods are needed to detect novel fossil taxa and to accurately place them within a phylogeny. Deep learning techniques, such as convolutional neural networks (CNNs) are capable of extracting key discriminative features from a given image [23, 22]. Novelty can be identified by measuring the uncertainty within the K-way classification for a given specimen: high uncertainty implies a high likelihood that the specimen is from an unknown taxon [13, 20]. For phylogenetic placement, however, neural networks must explicitly incorporate evolutionary distances. Models must be trained not to identify the traits which define classes, but to recognize phylogenetic synapomorphies - derived features that are shared between sister taxa.

Our insight is that phylogenetic distances can be used to re-learn and re-weight morphological features derived from CNNs. CNN features can be treated as a vector of continuous morphological traits for placing taxa within a

known phylogeny using Bayesian inference. We obtained these weighted features by first extracting the features from classification CNNs and then passing them to a second embedding model trained to transform features according to a phylogenetically-informed distance function. This is a fundamentally new approach in applying machine learning to biological classification.

Fossil pollen is an example of a fossil record where many morphotypes have unknown or uncertain biological affinities [26, 2, 40]. So although fossil pollen comprises one of the most extensive paleontological archives of terrestrial ecosystems, poor taxonomic resolution has limited evolutionary and paleoecological inference [26, 34]. Our approach builds on our previous work demonstrating that well-constructed CNN analyses can transfer learning from modern pollen morphology to fossil [37] and expands the range of questions that can be addressed by pollen data, offering new insights into paleoecology, plant evolution, and phylogenetics. It allows us to recognize extinct morphotypes and reconstruct their phylogenetic relationship to extant taxa. It allows us to reassess and recalibrate existing phylogenetic trees using microfossil data. The development of phylogenetically informed deep learning models is not specific to pollen, but also can be generalized to other taxa with robust phylogenies and standardized morphological imaging.

1.1 General analysis pipeline

1.1.1 Image classification

We began by training three independent convolutional neural networks (CNNs) for K-way classification by exploiting three individual modalities (shape, internal structure, and texture) (Fig. 1A). The first, holistic CNN (H-CNN) took as input maximum intensity projections (MIPs) of the whole grain. The second, cross-sectional CNN (C-CNN) used axial slices taken through the entirety of a pollen grain. The third, patch CNN (P-CNN), was trained on overlapping square crops of the initial MIP image, each covering $\sim$10% of the entire grain. The result of each CNN was a K-dim vector of classification scores that captured probabilities for each known taxonomic classification. We used a probabilistic fusion model (FM) (Fig. 1B) to combine these single-modal probabilities into a fused multimodal probability for better performance.

1.1.2 Novel taxon detection

The total uncertainty of each CNN and FM classification was calculated using entropy computed over the product of their classification probabilities. We used receiver-operating characteristic (ROC) analysis [43] to establish the uncertainty threshold above which a specimen should be considered novel (Fig. 1C). ROC is widely used to evaluate the performance of binary classification tasks [43, 16]. The area under the ROC curve (AUROC) [20] quantified our models' ability to discriminate between known and novel pollen taxa. We determined the threshold that optimized both the precision and recall of novel taxa. If the model's uncertainty score was below this threshold, the specimen's most likely predicted taxon was reported (Fig. 1D). Otherwise, the sample was flagged as novel, and the features extracted from the three classification CNNs (Fig. 1E) were forward passed through an embedding model for phylogenetic placement (Fig. 1F).

1.1.3 Phylogenetic placement

Our phylogenetic embedding model is a trained multi-layer perceptron (MLP) that transforms a specimen's classification features to embedding features based on a learned phylogenetic distance function (Fig. 1F). The resulting embedding features capture the phylogenetic information within classification features. The embedding features can then be used to place a novel specimen on an established phylogeny (Fig. 1G). We train the MLP over specimens from K known taxa by minimizing the differences in the computed Euclidean distances between pairs of specimens' embedding features and ground-truth phylogenetic distances. We then use Bayesian inference to reconstruct the original tree topology from the continuous machine-learned features, applying the Brownian motion (BM) model for character evolution [8, 9, 31].

2 Materials and Methods

2.1 Specimens and molecular phylogenies

We used 30 extant *Podocarpus* (Podocarpaceae) species and nine fossil *Podocarpidites* specimens (Fig. 2) [35] and a fossil-calibrated phylogeny [24] to train and evaluate our neural networks. Time-calibrated phylogenies allowed us to work with absolute ages (in millions of years, Ma) rather than relative phylogenetic distances.

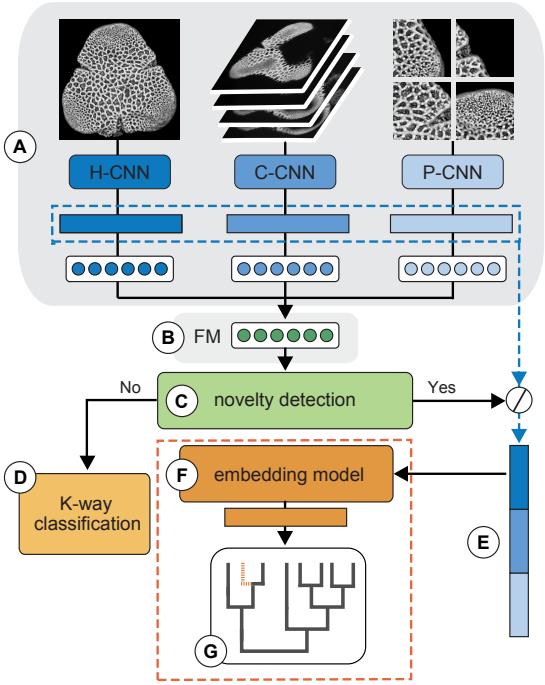

Figure 1. Flowchart illustrating the trained multimodal neural network pipeline. Three representations of superresolution images are passed through CNNs capturing shape, internal structure, and texture (H-CNN, C-CNN, P-CNN; A). The three sets of classification scores are fused (fused model, FM; B) and the analysis determines whether a specimen belongs to a known taxon by assessing the network's uncertainty during image classification (ROC analysis; C). If the specimen is recognized as one of the K taxa with high confidence (low uncertainty), its predicted taxon is reported (D). Otherwise, its features are extracted across all three image modalities and concatenated (E) as input to a multi-layer perceptron (F), which is trained to transform these features to an embedding feature to better compute phylogenetic distances from known taxa. Embedding features are clustered with the features of known taxa for phylogenetic placement, based on Bayesian inference (G).

Specimens were from the collections of Smithsonian Tropical Research Institute and Utrecht University. The pollen image dataset included 309 specimens from Australasia/Indomalaya (16 species), the Neotropics (10 species), Africotropics (2 species), and Asian Palearctic (1 species) (Supplemental Table S4). The dataset also includes nine fossil *Podocarpidites* specimens from the Late Cretaceous and Neogene of Colombia, Peru, and Panama (Supplemental Table S5) [17, 27, 4]. Modern taxon names were reviewed and amended according to the taxonomic nomenclature in Tropicos [42]. Specimens were imaged with Airyscan confocal superresolution microscopy at 630x magnification (63x/1.4NA oil objective, 0.08 µm per pixel resolution) as a series of axial focal planes (Z-stacks) at 0.19 µm increments [35]. We manually cropped and masked grains to remove background debris.

For the phylogenetic model, we relied on a conifer-wide fossil-calibrated phylogeny based on a concatenated dataset of two chloroplast DNA (cpDNA) protein-coding genes, rbcL and matK, and the 18S nuclear ribosomal DNA (rDNA) gene [24]. The reconstruction used a partitioned maximum likelihood analysis and assumed a GTR+Γ nucleotide substitution model. Most clades were well-supported, with high bootstrap support. Of the 45 species in the *Podocarpus* image dataset [35], only 30 were included in the phylogeny and were therefore selected for this study.

2.2 Image modalities

For our first imaging modality, we generated images of the external structure and shape of an entire pollen grain using maximum intensity projections (MIP) of the entire specimen image stack [34]. We applied contrast-limited adaptive histogram equalization (CLAHE) to standardize all images [47].

For the second modality, we used cross-sectional images. For the modern specimens, we generated a series of internal cross-sections by producing MIPs of staggered subsections of the image stack [37]. Each substack included 15 axial cross-sectional planes (2.85 µm depth), offset by 10 planes (1.90 µm). Because our fossil specimens were

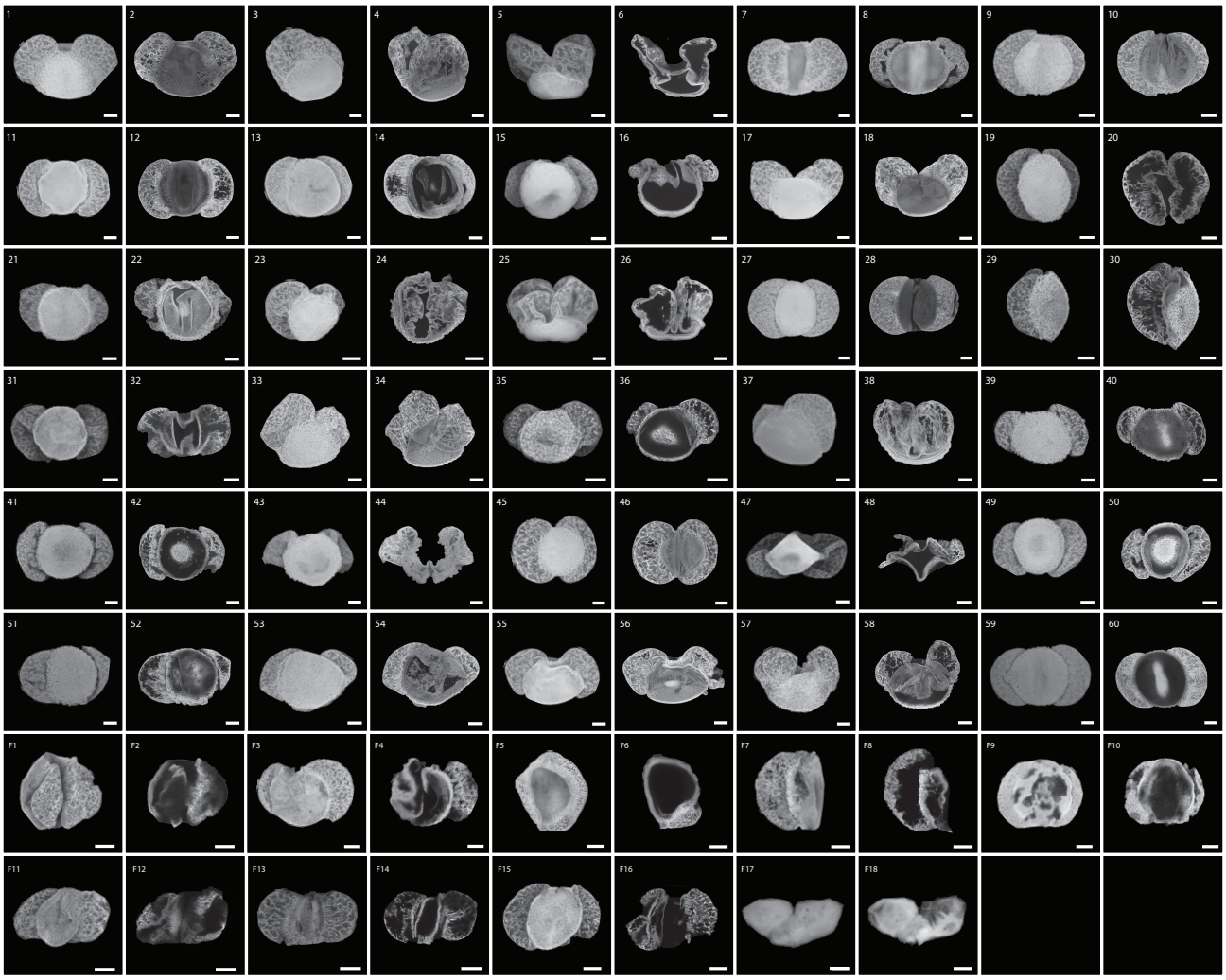

Figure 2. Modern and fossil specimens of *Podocarpus*. Each specimen is represented by a pair of images consisting of a maximum intensity projection (MIP) (first image) and a cross-sectional image (second image). Plate numbers and their corresponding species can be found in Supplemental Tables S4 and S5 (scale bar: 10 µm).

highly compressed, instead of subsections we used each individual plane of the image stack as our cross-sections.

The final modality was image patches that sampled a small portion of the grain. The whole-grain MIP for each specimen was divided into 10-13 smaller square patches, each of which covered ∼10% of the original MIP image and overlapped ∼25% with adjacent patches [19].

2.3 Deep learning models

2.3.1 Network architectures

We trained three separate K-way classification CNNs for each of our three modalities: a holistic image CNN (H-CNN), a cross-sectional CNN (C-CNN), and a patch CNN (P-CNN). All three CNNs incorporated ResNeXt-101 architecture [45] with the exception of the last layer, which was modified to output K-dim logit vectors. We normalized the logit vectors using softmax to produce K-dim probability vectors, which indicated the probability of the input image being classified as each of the K pollen taxon classes.

We fine-tuned the ResNeXt CNNs (pre-trained on ImageNet images) using our modern reference pollen dataset [7, 11]. For a given specimen, the H-CNN produced a K-dim logit vector (i.e., the feature prior to the K-dim softmax scores). The C-CNN and P-CNN used multiple cross-sections or patches and produced multiple logit vectors, respectively. We next computed the mean of the per-patch logit vectors and the mean of the

per-cross-sectional logit vectors. Lastly, we summed the three logit vectors and normalized by softmax to obtain a K-dim classification probability vector.

2.3.2 Training details

Specimens of the K known taxa were randomly split into training (70%) and validation (30%) sets. We augmented the training data by adopting common augmentation techniques including random flip, random rotation (with probability p=0.5 by degrees in [-90°, +90°]), and random translation (within pixel displacement in [-30, +30]). We resized the augmented images to 224×224×3 pixel resolution using bilinear interpolation.

We trained the K-way classification CNNs by minimizing the cross-entropy loss:

$$\mathcal{L}^{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} T_{i,k} \cdot \log(P_{i,k}) \tag{1}$$

where N is the number of specimens in the training set, K is the number of taxa, and $P_{i,k}$ is the softmax probability of specimen-i classified as taxon-k. $T_{i,k}$ is a binary indicator:

$$T_{i,k} = \begin{cases} 1, & \text{if specimen-}i\text{'s label is taxon-}k \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

We used the stochastic gradient descent optimizer with momentum (0.9). We adopted a stage-wise learning rate schedule with an initial learning rate of 0.0009, decreased by half every two epochs. All models were trained for 20 epochs with a batch size of 10.

2.3.3 Fused model (FM)

Given a specimen X and its three modalities (x_1, x_2, and x_3), we fused the predictions of the three modalities by assuming conditional independence given the class label [5], i.e., $p(x_1, x_2, x_3|y = k) = p(x_1|y = k) \cdot p(x_2|y = k) \cdot p(x_3|y = k)$. Further, we assumed a uniform prior, i.e., $p(y = 1) = \cdots = p(y = K)$. Below is the fusion result:

$$\begin{aligned} & p(y \mid x_1, x_2, x_3) && (3)\\ =& \frac{p(x_1, x_2, x_3 \mid y)p(y)}{p(x_1, x_2, x_3)} && (4)\\ =& \frac{p(x_1 \mid y)p(y)p(x_2 \mid y)p(y)p(x_3 \mid y)p(y)}{p(x_1, x_2, x_3)p(y)p(y)} && (5)\\ \propto& \, p(y \mid x_1)p(y \mid x_2)p(y \mid x_3) && (6) \end{aligned}$$

(4) applies the Bayes' theorem, (5) assumes conditional independence, (6) assumes uniform prior w.r.t class labels. To derive the probability after fusion, we normalized (6) to sum-to-one over the K classes, so that:

$$p(y = k \mid x_1, x_2, x_3) = \frac{p(y = k \mid x_1)p(y = k \mid x_2)p(y = k \mid x_3)}{\sum_{j=1}^{K} p(y = j \mid x_1)p(y = j \mid x_2)p(y = j \mid x_3)}$$

2.3.4 Model validation

We assessed our models using five random splits of training (70%) and validation (30%) image sets. We report the mean of per-taxon accuracies. Confusion matrices were presented to provide a detailed report of per-taxon accuracies (Fig. 3).

2.4 Novelty detection

2.4.1 Uncertainty estimates

We used entropy [16] to assess whether a specimen should be considered as novel to the trained model. Given specimen X and its probability of being classified as taxon-k, $p(y = k|X)$ for $k = 1, \ldots, K$, we computed entropy $H(p)$ as:

$$H(p) = -\sum_{k=1}^{K} p(y = k|X) \log p(y = k|X) \tag{7}$$

We compared novelty detection performance in each of the three CNN models and the FM. To measure the models' abilities to determine true novel morphologies, we used the Receiver Operating Characteristic (ROC) curve between True Positive Rate (TPR) and False Positive Rate (FPR). ROC is a graph showing the performance of a binary classification model at different thresholds. We computed the area under the ROC curve (AUROC) as the summary number to benchmark methods. Higher AUROC means better performance for novelty detection [20].

2.4.2 Optimal threshold selection

Although AUROC effectively summarizes model performance with varied thresholds, real-world systems require a single operation threshold. We used the method introduced in [43] to find an optimal threshold c^* for fossil pollen analysis. We analyzed the ROC curve to find c^*, using the TPR and FPR for a binary classification model designed to detect specimens of novel pollen types. We evaluated the performance of the model by using our training set to represent the known specimens and the pseudo-novel set to represent the novel specimens. We optimized the following objective function [43] to obtain c^*, the value at which TPR and (1-FPR) values are closest to AUROC while minimizing their absolute difference:

$$c^* = \arg\min_{c \in R}(|\mathrm{TPR}(c) - \mathrm{AUROC}| + |1 - \mathrm{FPR}(c) - \mathrm{AUROC}|) \tag{8}$$

2.5 Phylogenetic placement

2.5.1 Model and loss function

Specimens identified as novel were placed within an existing phylogeny using a trained embedding model, implemented as a multi-layer perceptron (MLP). MLPs are a special form of CNNs in which all the convolution layers have 1×1 kernels. For a novel specimen, the input to the MLP was a concatenation of the specimen's three classification feature vectors (with 2048 dimensions each), extracted from the three classification CNNs. The MLP outputs a low-dimensional embedding feature vector $\mathbf{v} \in R^{256}$ used to place the specimen within a phylogeny.

Our MLPs have five convolutional layers. The first layer learns three non-negative parameters to weigh the three classification features learned by individual CNNs. Between every two consecutive layers, we inserted a rectified linear activation unit (ReLU) for nonlinearity [29] and dropout layers (with a dropout rate of 0.2 to prevent overfitting [39]. The complete network architecture is described in the supplementary materials (Supplemental Fig. S5 and Table S6).

Our MLP is lightweight relative to a CNN, allowing us to train with all training specimens' features as a single batch. We used the Adam optimizer [18] with a constant learning rate of 0.00001 and a weight decay of 10^{-4}. We trained the MLP for 20,000 iterations (∼10 min running time on a GPU).

We developed a simple loss function to train the MLP to learn phylogenetic patterns from classification features. The loss function minimizes the error between ground-truth and predicted evolutionary distances separating pairs of specimens. Because the MLP outputs an embedding feature v_i for specimen-i, between specimen i and j, we compute their Euclidean distance.

With the known ground-truth phylogenetic distance $D_{i,j}$, we minimize the mean square error (MSE) between $D_{i,j}$ and $\hat{D}_{i,j}$:

$$\mathcal{L}^{MSE} = \sum_{i,j} \left(D_{i,j} - \hat{D}_{i,j} \right)^2 \tag{9}$$

All the ground-truth and predicted phylogenetic distances in D and $\hat{D}$ were re-scaled to the range $[0, 1]$.

2.5.2 Tree construction

We constructed phylogenetic trees from the transformed features using Bayesian inference and assuming the Brownian motion (BM) model for character evolution [8, 9, 31]. Analyses were performed using RevBayes [14]. We performed Markov Chain Monte Carlo (MCMC) analyses consisting of two independent runs with four chains each, running for 1 million generations. Trees were sampled every 1000 generations, and initial pre-stationarity generations were discarded according to the burn-in value determined with Tracer (v.1.7.1). Tree files from both independent runs were combined using LogCombiner (v.1.10.4) to generate the final phylogeny. The maximum clade credibility (MCC) tree, which maximizes the posterior probability of each individual clade, was chosen as the final model using TreeAnnotator (v.2.4.2). The resulting trees were visualized and edited using the functions

plotTree and plotFBDTree in the R package "RevGadgets" (v.1.1.1) [41]. Branches in the tree with a posterior probability ≥ 0.95 were considered significantly supported.

2.6 Pseudo-Novel evaluation experiments

2.6.1 Evaluation of novelty detection

To evaluate our methods for discriminating among known and novel taxa, we adopted a "leave-one-out" protocol [21], where a taxon was excluded from the training dataset and its specimens were treated as novel morphotypes. We ran five separate experiments and selected one of five taxa from biologically meaningful subclades in the reference phylogeny: *P. drouynianus*, *P. elongatus*, *P. neriifolius*, *P. oleifolius*, and *P. totara*.

We forward-passed all specimens of a pseudo-novel taxon and the known validation specimens to their corresponding trained CNNs. We computed entropy over their class probability distributions and ran a ROC analysis. The analysis provided an assessment of the model's performance in novelty detection. In each run of these experiments, we also computed an optimal threshold. We averaged these thresholds to obtain a single threshold for identifying new fossil morphotypes.

2.6.2 Evaluation of phylogenetic placement

We extracted the CNN-learned features of all pseudo-novel specimens and passed them to the trained MLPs. We then concatenated the resulting output features with the average feature results for the known species. This produced a continuous character matrix of pseudo-novel specimens and known species. We used the character matrix to construct Bayesian inference trees under a single rate Brownian motion model. The trees were simulated under a birth-death model, with all parameters defined following [31]. Because our modern dataset sampled 30% of extant *Podocarpus* taxa, the sampling rate was set to 0.3.

2.7 Fossil application

We forward-passed the nine fossil *Podocarpidites* specimens through our fully trained CNN-MLP pipeline, where the model had been trained on all modern specimens. This produced a continuous character matrix of fossil specimens and extant species, which were then used to construct a Bayesian inference tree. We performed tip-dating under the fossilized birth-death (FBD) process, enabling the simultaneous analysis of both modern and fossil morphotypes while calibrating divergence times estimates in a Bayesian framework [12, 46]. Tip ages were included as single-point occurrence times corresponding to the most likely ages of the fossil specimens. The diversification and turnover rates were drawn from a uniform distribution over an interval ranging from 0 to 1. Speciation was defined as the sum of diversification and turnover rates, while extinction was set to be equal to the turnover rate. The root age was drawn from a uniform distribution with a lower bound of 67.2 Ma, corresponding to the age of our oldest fossil, and an upper bound of 82 Ma, based on the maximum root age estimated in [36]. Similarly, the fossil rate was drawn from a uniform distribution ranging from 0 to 1. The sampling fraction of extant taxa was set at 0.3. The resulting time-calibrated MCC tree provided us with the most likely placements of our fossil specimens within the reference phylogeny, accompanied by their corresponding posterior probabilities.

3 Results

3.1 Baseline classification accuracy

We measured taxon classification accuracies for the three imaging modalities (H-CNN, C-CNN, and P-CNN), and the fused model (FM). The FM and C-CNN produced the highest average accuracies, 90.40% and 90.60% average accuracy, respectively. The P-CNN scored lower (79.80%) and the H-CNN lower still (55.00%) (Fig. 3, Supplemental Fig. S1, Supplemental Table S1).

3.2 Novel taxon detection

The C-CNN (μ=0.9425) and FM (μ=0.8987) markedly outperformed the P-CNN (μ=0.8407) and H-CNN (μ=0.7104) in detecting novel pollen morphotypes. The FM had the smallest variability overall (σ=0.0325). The P-CNN and H-CNN had higher variability (σ=0.0725 and 0.0591, respectively) (Supplemental Fig. S2 and Table S2).

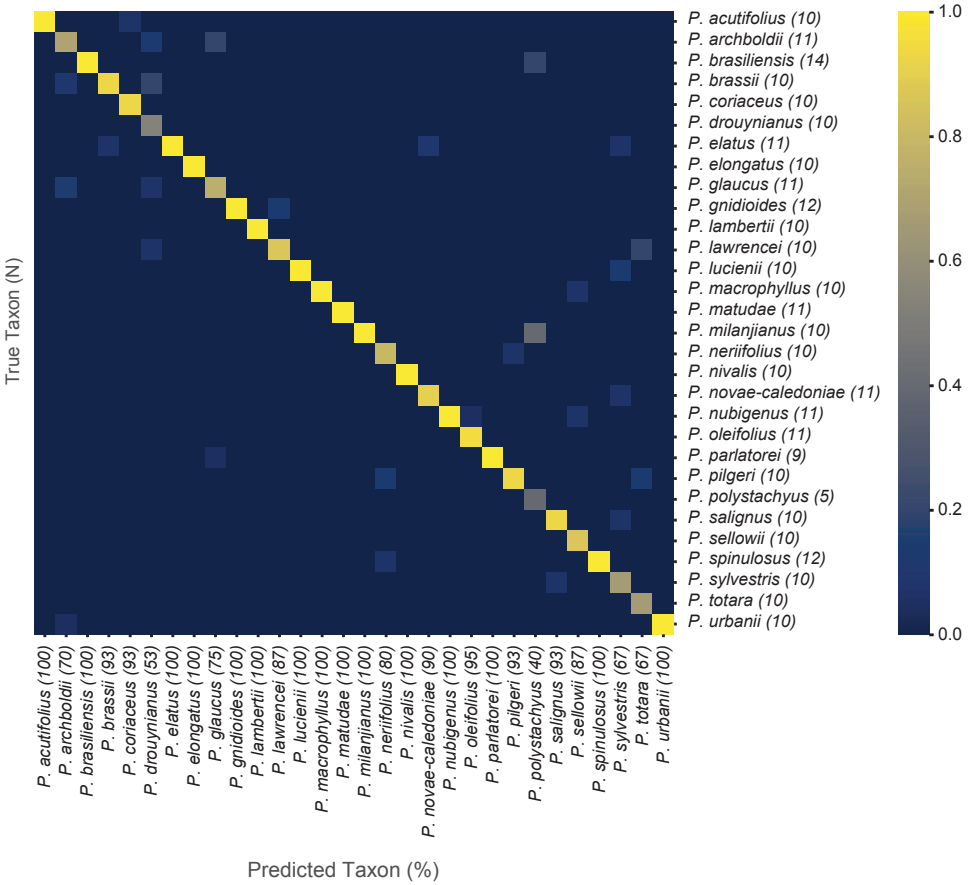

Figure 3. Average confusion matrix depicting the classification accuracy of the fused model (FM) for the modern *Podocarpus* species. Rows correspond to the true taxon while columns represent the model's predictions.

3.3 Phylogenetic placement of pseudo-novel specimens

The Bayesian inference trees constructed from the learned MLP features largely replicated the *Podocarpus* molecular phylogeny and faithfully reconstructed two well-supported subgenera (*Podocarpus* subg. *Podocarpus* and *P.* subg. *Foliolatus*), with strong branch support for most clades across the trees (Supplemental Fig. S3.1-S3.5.). The MLP determined the optimal weights for combining features from the three modalities (H-CNN, C-CNN, and P-CNN). In all five pseudo-novel experiments, the MLPs learned nearly equal weights for each modality (Supplemental Table S3).

For three pseudo-novel taxa *P. drouynianus*, *P. oleifolius*, and *P. totara*), all specimens were placed in their correct respective subclade, with high support values (P=1 in all three experiments) (Supplemental Figs. S3.1, S3.4, and S3.5).

For *P. elongatus*, the accuracy of the placement varied by specimen (Supplemental Fig. S3.2). Six specimens were correctly placed with *P. milanjianus*, with fairly high support (P=0.935). Three were placed elsewhere within subgenus *Podocarpus* and the remaining one was placed within subgenus *Foliolatus*. In the *P. neriifolius* experiments, seven specimens were correctly placed within the subclade comprising *P. brassii*, *P. archboldii*, *P. macrophyllus*, and *P. pilgeri*, with fairly high support (P=0.910) (Supplemental Fig. S3.3). The remaining three specimens were placed as sister to the entire *Foliolatus* subgenus.

3.4 Fossil application

All nine fossil *Podocarpidites* specimens were placed with *Podocarpus* subgenus *Podocarpus* (Fig. 4). Notably, the fossil placements were less certain than the pseudo-novel placements (Fig. 4 and Supplemental Figs. S3.1-S3.5).

The oldest specimen (Colombia, 67.2 Ma) was placed as sister to the entire subgenus, with strong support values (P=1). The three oldest Panamanian specimens (10.15, 4.25, and 3.55 Ma) were placed as sister to the Neotropical clade formed by *P. matudae*, *P. urbanii*, *P. oleifolius*, *P. brasiliensis*, *P. sellowii*, and *P. coriaceus* (P=1).

The youngest fossil and most taphonomically altered fossil (Panama, 2.05 Ma) was placed as sister to the Andean Chilean species *P. salignus*. The four Peruvian specimens (12 Ma) were also placed as sisters to *P. salignus*. The *P. salignus* and *Podocarpidites* clade was poorly supported (P=0.5426).

4 Discussion

CNNs are powerful tools for the analysis of biological morphology. Both our results and previous research, demonstrate their ability to capture the variability of pollen morphology and accurately classify species, even when there are limited morphological differences [34, 37, 3, 30, 38]. Our study takes the application of machine learning in biological classification further, demonstrating that features captured by neural networks can be interpreted and used within a phylogenetic framework. Our results with *Podocarpus* show that machine-learned features can be used to interpret pollen morphology in an evolutionary context and place fossil specimens within a phylogeny.

Molecular phylogenetic studies place the origin of *Podocarpus* in the Late Cretaceous [36] or Paleocene [1, 25, 24], with more uncertain timing for the divergence of the two subgenera, *P. Podocarpus* and *P. Foliolatus*. The underlying uncertainty in published phylogenies introduced noise into our analysis. We expected errors as not all nodes of the original reference tree were well supported or fully resolved [24]. Our reference phylogeny includes more species, but disagreements exist between the phylogeny used to train our models [24] and a second recent phylogeny [36]. Most notably, the African clade within *P. Podocarpus* is more closely related to the tropical Central-South American clade in [24], while in [36], the clade is more closely related to subtropical South American species. Additionally, in [36], the temperate South American species *P. nubigenus* and *P. salignus* are more closely related to the Australasian taxa of *P. Podocarpus* (*P. nubigenus*, *P. lawrencei*, *P. acutifolius*, *P. totara*, *P. gnidioides*, and *P. nivalis*). *P. salignus* is placed as sister to the subgenus *P. Podocarpus* in our reference phylogeny, and *P. nubigenus* is more closely related to the tropical-subtropical South American species of *P. parlatorei* and *P. lambertii* [24]. Differences in the placement of *P. drouynianus* and *P. spinulosus* have less relevance to our results, as none of our fossil specimens were placed within *P. Foliolatus*.

Podocarpus pollen is bisaccate with large sacci that are coarsely endoreticulate. There is little morphological diversity within the genus, and palynologists generally do not attempt to identify individual species [32, 15, 28]. Despite this, we achieved high classification accuracies for our modern species with few per-species examples. Only the H-CNN performed poorly, with an average accuracy of 55.00% (Supplemental Table S1). In contrast, cross-sectional images and patches expanded the number of training examples per species and circumvented the "few-shot" learning problem of training models on a limited number of examples [6, 44].

Many misclassifications were of closely related species (Fig. 3, Supplemental Fig. S1), supporting our prediction of a phylogenetic signal in *Podocarpus* pollen morphology. *P. archboldii*, *P. lawrencei*, *P. neriifolius*, *P. sylvestris*, *P. drouynianus*, *P. pilgeri*, and *P. totara* were confused with their sister species *P. brassii*, *P. gnidiodes*, *P. pilgeri*, *P. lucienii* or species within the same subclade. Several misclassifications, however, were among distantly related species. *P. polystachyus* was misclassified as *P. milanjianus* and *P. brasiliensis*, likely due to its small sample size (n=5). Additionally, *P. glaucus* was confused for *P. archboldii* and *P. parlatorei*, *P. drouynianus* was confused for *P. brassii*, *P. archboldii*, and *P. lawrencei*, and *P. totara* was confused for *P. pilgeri*. Although we did not observe specific morphological features which would explain these misidentifications, we suspect that they may be due to similarities in the endoreticulate patterning of their sacci, resulting from convergence or homoplasy.

Through our pseudo-novel experiments, we were able to successfully establish an uncertainty threshold for identifying taxa that were new or unknown to the model. The C-CNN and FM were the least variable in recognizing pseudo-novel specimens, with consistently high AUROC values. This supports our expectation that the internal structure of the pollen wall retains phylogenetically informative characters for identifying novel morphotypes. The P-CNN produced both the highest (i.e., *P. oleifolius*) and lowest (i.e., *P. elongatus*) AUROC values and was, therefore, the least reliable in detecting new species.

MLP-transformed CNN features closely reproduced the molecular phylogeny used to train our models. We correctly placed 43 out of 51 pseudo-novel specimens in their expected clades (Supplemental Figs. S3.1-S3.5). We could not identify specific morphological differences that explained the inaccurate placement of the four *P. elongatus* specimens that were incorrectly placed. However, the sacci of the three erroneously placed *P. neriifolius* specimens were obscured by the grains' orientation. This may have led to their inaccurate placement in the reference phylogeny.

Our fully trained models were also able to produce plausible placements of fossil specimens. The fossil pollen morphospecies *Podocarpidites* is associated with *Podocarpus*, although other taxa - namely the extant

New Zealand species *Lepidothamnus laxifolius* and extinct Pinaceae species from the Laurasian Paleogene - have similar morphologies [28, 10]. Our *Podocarpidites* fossils are all Neotropical Neogene or younger, with the exception of a Late Cretaceous specimen from Colombia, so we were confident in assuming our *Podocarpidites* specimens were fossil *Podocarpus*. The limited morphological differentiation among *Podocarpus* species makes interpreting the placement of our *Podocarpidites* specimens difficult. However, the spatio-temporal pattern that we observe is consistent with the biogeographic history of the subgenus *P. Podocarpus* as reconstructed by [36]. In our tip-calibrated tree, the 67.2 Ma Colombian *Podocarpidites* specimen was placed as sister to the entire subgenus *P. Podocarpus*, suggesting a lineage now extinct in the Neotropics. The Miocene Peruvian specimens were placed with the early-diverging Chilean species, *P. salignus*. The four Panamanian were the youngest fossils in our analysis. Three specimens (3.55, 4.25 and 10.15 Ma) were placed with the main Neotropical lineage (*P. sellowii*, *P. coriaceus*, *P. brasiliensis*, *P. urbanii*, *P. oleifolius*, and *P. matudae*), with high support.

The youngest (2.05 Ma) Panamanian specimen was highly corroded, with few visible features on the sacci or corpus. Its unlikely placement with *P. salignus* is likely the result of this poor preservation. Its placement, and the placement of the four Miocene (12 Ma) Peruvian specimens, with the Chilean species *P. salignus* had poor support. However, excluding the specimen does not affect the placement of the four Peruvian specimens within the tree (Supplemental Fig. S4). The low node support may reflect the uncertainty in the placement of *P. salignus* within *P. Podocarpus*, as previously discussed [24, 36].

5 Conclusions

Palynologists have long used the morphological conservation of features within clades to classify pollen taxa. However, our results suggest that an even deeper evolutionary history is contained within pollen morphology. The vocabulary of palynology is vast [33], but it does not capture all the gradations and permutations possible in pollen wall structure and ornamentation. Deep learning connects morphology to molecular phylogenies, learning evolutionary distances from existing phylogenies and transforming morphological classification features. Features are transformed from characters that define individual clades to the synapomorphies that link them.

Deep learning is adaptive to the constraints of paleontological data, including incomplete or unresolved phylogenies and fossil morphotypes derived from extinct taxa. Novelty detection recognizes new morphologies and expands our knowledge of a clade's morphospace. We can explicitly quantify our uncertainty. The abundance of fossil pollen in geological samples also means that phylogenetic determinations can be made with populations of specimens, not just individual specimens. Applied at a broad scale, phylogenetically-informed machine learning models will allow palynologists to harness the evolutionary information within the fossil record and identify taxon origins and extinctions that would otherwise go unrecognized.

Acknowledgments

We thank Carlos Jaramillo for sharing fossil and reference pollen samples, Timme Donders for providing additional reference pollen samples, Andrew Leslie for providing the *Podocarpus* phylogeny, and Milton Tan for feedback on the manuscript. We thank Ingrid Romero and Michael Urban for imaging the majority of the material used and cited in this analysis. Support for MEA and SWP was provided by a 2022 National Center for Supercomputing Applications Faculty Fellowship to SWP and the University of Illinois Tom L. Phillips Fund for Paleobotany.

Data availability

The Airyscan images used in this study were submitted to the Illinois Databank (https://doi.org/10.13012/B2IDB-8817604_V1). The codebase is publicly available on GitHub (DOI: 10.5281/zenodo.8128443) and can be accessed using this link: https://github.com/paleopollen/Novel_Pollen_Phylogenetic_Placement.

References

[1] Edward Biffin, John G Conran, and Andrew J Lowe. "Podocarp evolution: a molecular phylogenetic perspective". In: *Ecology of the Podocarpaceae in tropical forests* (2011).

[2] H John B Birks et al. "Does pollen-assemblage richness reflect floristic richness? A review of recent developments and future challenges". In: *Review of Palaeobotany and Palynology* 228 (2016), pp. 1–25.

[3] Mihai Boldeanu et al. "Automatic Pollen Classification Using Convolutional Neural Networks". In: *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE. 2021, pp. 130–133.

[4] Mónica R Carvalho et al. "Extinction at the end-Cretaceous and the origin of modern Neotropical rainforests". In: *Science* 372.6537 (2021), pp. 63–68.

[5] Yi-Ting Chen et al. "Multimodal Object Detection via Probabilistic Ensembling". In: *European Conference on Computer Vision (ECCV)*. 2022.

[6] Wei-Yu Chen et al. "A Closer Look at Few-shot Classification". In: *International Conference on Learning Representations*. 2019. URL: `https://openreview.net/forum?id=HkxLXnAcFQ`.

[7] J. Deng et al. "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR09*. 2009.

[8] Joseph Felsenstein. "Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters". In: *Systematic Biology* 22.3 (1973), pp. 240–249.

[9] Joseph Felsenstein. "Phylogenies and the comparative method". In: *The American Naturalist* 125.1 (1985), pp. 1–15.

[10] David R Greenwood, Christopher R Hill, and John G Conran. "Prumnopitys anglica sp. nov.(Podocarpaceae) from the Eocene of England". In: *Taxon* 62.3 (2013), pp. 565–580.

[11] Kaiming He, Ross Girshick, and Piotr Dollar. "Rethinking ImageNet Pre-Training". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[12] Tracy A Heath, John P Huelsenbeck, and Tanja Stadler. "The fossilized birth–death process for coherent calibration of divergence-time estimates". In: *Proceedings of the National Academy of Sciences* 111.29 (2014), E2957–E2966.

[13] Dan Hendrycks and Kevin Gimpel. "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks". In: *International Conference on Learning Representations*. 2017. URL: `https://openreview.net/forum?id=Hkg4TI9xl`.

[14] Sebastian Höhna et al. "RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language". In: *Systematic biology* 65.4 (2016), pp. 726–736.

[15] Henry Hooghiemstra, Vincent M Wijninga, and Antoine M Cleef. "The paleobotanical record of Colombia: implications for biogeography and Biodiversity1". In: *Annals of the Missouri Botanical Garden* 93.2 (2006), pp. 297–325.

[16] Mohammad Jafarian, Alireza Soroudi, and Andrew Keane. "Resilient identification of distribution network topology". In: *IEEE Transactions on Power Delivery* 36.4 (2020), pp. 2332–2342.

[17] Carlos A Jaramillo et al. "Palynological record of the last 20 million years in Panama". In: *Paleobotany and biogeography: A festschrift for Alan Graham in his 80th year* (2014).

[18] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[19] Shu Kong, Surangi Punyasena, and Charless Fowlkes. "Spatially aware dictionary learning and coding for fossil pollen identification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016, pp. 1–10.

[20] Shu Kong and Deva Ramanan. "Opengan: Open-set recognition via open data generation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 813–822.

[21] Alexey M Kozlov et al. "Phylogeny-aware identification and correction of taxonomically mislabeled sequences". In: *Nucleic acids research* 44.11 (2016), pp. 5022–5033.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems* (2012).

[23] Yann LeCun, Yoshua Bengio, et al. "Convolutional networks for images, speech, and time series". In: *The handbook of brain theory and neural networks* 3361.10 (1995), p. 1995.

[24] Andrew B Leslie et al. "An overview of extant conifer evolution from the perspective of the fossil record". In: *American journal of botany* 105.9 (2018), pp. 1531–1544.

[25] Andrew B Leslie et al. "Hemisphere-scale differences in conifer evolutionary dynamics". In: *Proceedings of the National Academy of Sciences* 109.40 (2012), pp. 16217–16221.

[26] Luke Mander and Surangi W Punyasena. "On the taxonomic resolution of pollen and spore records of Earth's vegetation". In: *International Journal of Plant Sciences* 175.8 (2014), pp. 931–945.

[27] C Martínez et al. "Neogene precipitation, vegetation, and elevation history of the Central Andean Plateau". In: *Science advances* 6.35 (2020), eaaz4724.

[28] Robert J Morley. "Dispersal and paleoecology of tropical podocarps". In: *Smithsonian contributions to botany* (2011).

[29] Vinod Nair and Geoffrey E Hinton. "Rectified linear units improve restricted boltzmann machines". In: *International Conference on Machine Learning (ICML)*. 2010.

[30] Ola Olsson et al. "Efficient, automated and robust pollen analysis using deep learning". In: *Methods in Ecology and Evolution* 12.5 (2021), pp. 850–862.

[31] Caroline Parins-Fukuchi. "Use of continuous traits can improve morphological phylogenetics". In: *Systematic Biology* 67.2 (2018), pp. 328–339.

[32] DT Pocknall. "Pollen morphology of the New Zealand species of Dacrydium selander, Podocarpus L'heritier, and Dacrycarpus endlicher (podocarpaceae)". In: *New Zealand journal of botany* 19.1 (1981), pp. 67–95.

[33] Willem Punt et al. "Glossary of pollen and spore terminology". In: *Review of palaeobotany and palynology* 143.1-2 (2007), pp. 1–81.

[34] Surangi W Punyasena et al. "Classifying black and white spruce pollen using layered machine learning". In: *New Phytologist* 196.3 (2012), pp. 937–944.

[35] Surangi W Punyasena et al. *Pollen of Podocarpus (Podocarpaceae): Airyscan confocal superresolution images. University of Illinois Urbana-Champaign.* 2022. DOI: `https://doi.org/10.13012/B2IDB-8817604_V1`.

[36] Marıa Paula Quiroga et al. "Molecular and fossil evidence disentangle the biogeographical history of Podocarpus, a key genus in plant geography". In: *Journal of Biogeography* 43.2 (2016), pp. 372–383.

[37] Ingrid C Romero et al. "Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy". In: *Proceedings of the National Academy of Sciences* 117.45 (2020), pp. 28496–28505.

[38] Vıctor Sevillano and José L Aznarte. "Improving classification of pollen grain images of the POLEN23E dataset through three different applications of deep learning convolutional neural networks". In: *PLoS One* 13.9 (2018), e0201807.

[39] Nitish Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958.

[40] Alfred Traverse. *Paleopalynology*. Vol. 28. Springer Science & Business Media, 2007.

[41] Carrie M Tribble et al. "RevGadgets: An R package for visualizing Bayesian phylogenetic analyses from RevBayes". In: *Methods in Ecology and Evolution* 13.2 (2022), pp. 314–323.

[42] *Tropicos.org. Missouri Botanical Garden*. `https://www.tropicos.org`. Accessed: 2022-06-20. 2022.

[43] Ilker Unal. "Defining an optimal cut-point value in ROC analysis: an alternative approach". In: *Computational and mathematical methods in medicine* 2017 (2017).

[44] Yaqing Wang et al. "Generalizing from a few examples: A survey on few-shot learning". In: *ACM computing surveys (csur)* 53.3 (2020), pp. 1–34.

[45] Saining Xie et al. "Aggregated residual transformations for deep neural networks". In: *IEEE conference on Computer Vision and Pattern Recognition*. 2017.

[46] Chi Zhang et al. "Total-evidence dating under the fossilized birth–death process". In: *Systematic biology* 65.2 (2016), pp. 228–249.

[47] Karel Zuiderveld. "Contrast limited adaptive histogram equalization". In: *Graphics gems* (1994), pp. 474–485.

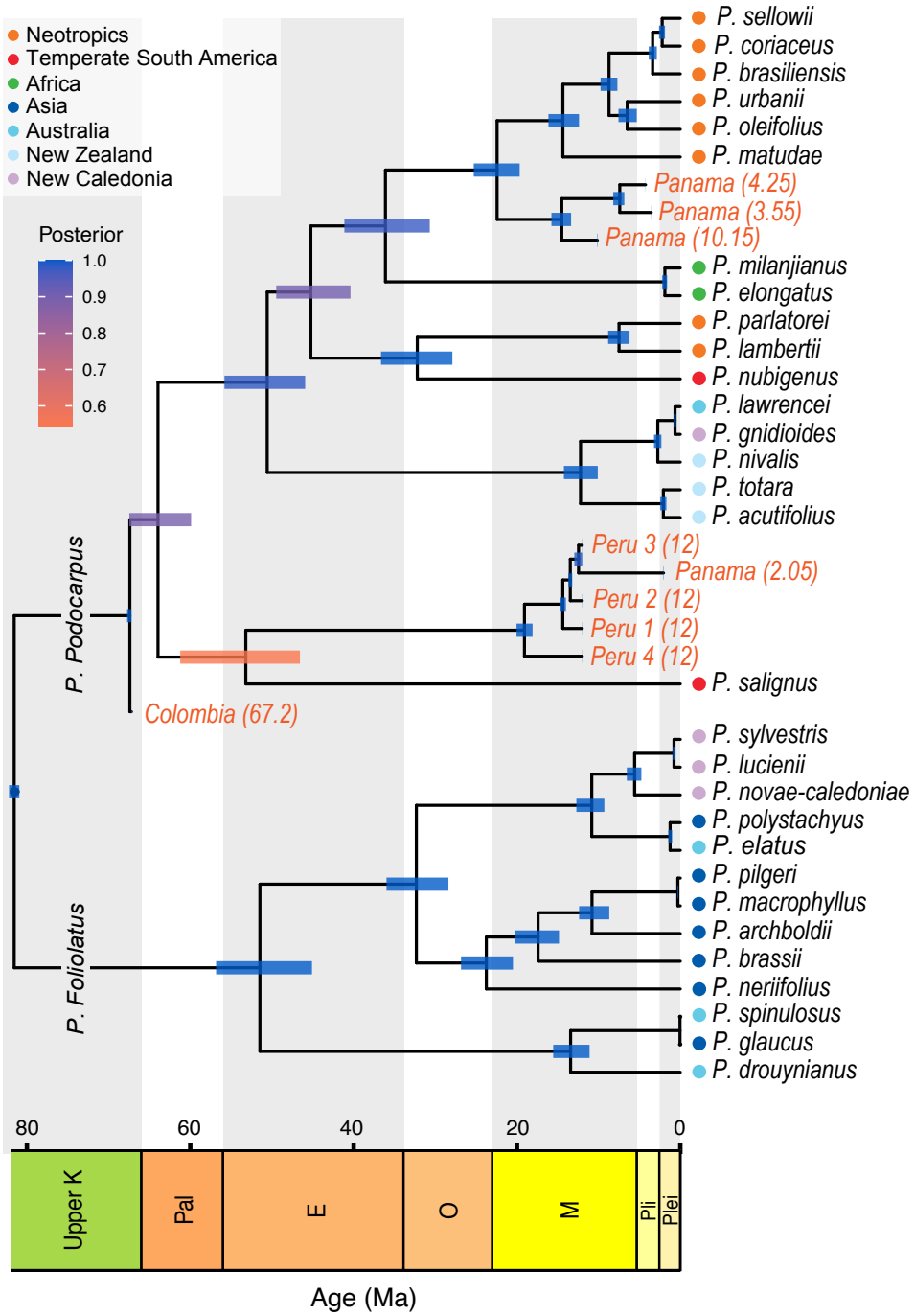

Figure 4. Combined tip-calibrated fossil tree simulated under the fossilized birth-death process (FBDP) using the MLP-transformed features. The tree topology replicates that of the reference phylogeny proposed by [24]. Branch lengths were estimated based on our fossil tip ages. The three oldest Panamanian specimens are placed as sister to the Central-South American tropical clade formed by *P. sellowii*, *P. coriaceus*, *P. brasiliensis*, *P. urbanii*, *P. oleifolius*, and *P. matudae*, while the youngest specimen is placed with the Chilean *P. salignus*. Likewise, the four Peruvian specimens are are placed with *P. salignus*. Finally, the late Cretaceous Colombian specimen is placed as sister to the subgenus *P. Podocarpus*.