Decoding Fossil Floras with Artificial Intelligence

Ivan Felipe Rodriguez[1†], Thomas Fel[1,2†], Gaurav Gaonkar[1], Mohit Vaishnav[1], Herbert Meyer[3], Peter Wilf[4], Thomas Serre[1*]

[1*]Department of Cognitive and Psychological Sciences, Brown University, Providence, Rhode Island, USA.
[2]Kempner Institute, Harvard University, Boston, Massachusets, USA.
[3]Florissant Fossil Beds National Monument, National Park Service, Florissant, Colorado, USA.
[4]Department of Geosciences, Pennsylvania State University, University Park 16802, Pennsylvania, USA.

*Corresponding author(s). E-mail(s): thomas_serre@brown.edu;
†These authors contributed equally to this work.

Abstract

Accurately identifying fossil angiosperm leaves remains one of paleobotany's most persistent challenges. Although the morphological complexity of leaves has historically hindered manual classification, artificial intelligence (AI) is well-suited to extracting subtle diagnostic patterns that elude human perception. However, applying standard AI approaches to fossil material faces a fundamental limitation: the extreme scarcity of taxonomically vetted fossil specimens precludes conventional supervised training at scale. While modern leaf specimens are abundant, fossilization processes—compression, mineralization, fragmentation—create a challenging domain shift between living and fossil forms. Here, we introduce a deep learning framework that overcomes this challenge by augmenting sparse fossil data with synthetic examples and aligning extant and fossil leaf domains through representational learning. Our approach synthesizes high-fidelity fossil analogs from modern leaf images (cleared and X-rayed) and trains a deep neural network using multi-scale representations that capture features from venation patterns to overall leaf morphology, with triplet-loss embedding to align the two domains. We evaluate the system on the late Eocene Florissant flora of Colorado, achieving a top-5 F1-score of 91.8% (chance: 3.5% across 142 families) for family-level classification of fossil leaves. Remarkably, the system maintains a 73.4% F1-score even when fossil families are entirely excluded from training—demonstrating robust out-of-distribution generalization. To interpret the model's decision process, we developed concept-based dictionary learning

methods that identify and localize visual features driving classification decisions. We demonstrate practical utility by applying our system to 1,723 previously unidentified leaf fossils from the Florissant Formation, providing predictions with visual explanations to guide expert review.
By overcoming data scarcity through generative AI and representational learning, this work offers a pathway to unlocking large-scale paleobotanical "dark data"—the vast collections of unidentified specimens in museum drawers worldwide.

Introduction

Isolated leaves dominate the angiosperm fossil record yet remain notoriously difficult to identify accurately, representing paleobotany's largely untapped source of "dark data" [1]. Historical literature is riddled with botanically incorrect identifications due to the inherent complexity of leaf morphology, insufficient vetted reference samples, and considerable variation within and across fossil sites. Consequently, most well-identified fossil leaves represent only a handful of morphologically distinctive families that are well-represented in the literature, leaving the vast majority of angiosperm diversity in the fossil record unrecognized [2, 3]. Accurate fossil leaf classification is crucial because leaf fossils provide essential data for interpreting evolutionary radiations and extinctions, biome evolution, plant-animal interactions, biogeography, and biotic responses to climate change [4–8]. At the family level—the traditional first step for most extant and fossil plant identifications—classification provides a stable taxonomic anchor, since nearly all fossil plants represent extinct species and many belong to extinct genera. Our previous work demonstrated that computer vision algorithms can generalize morphological features of leaves at this level [9].

Artificial intelligence (AI) shows promise for automating plant and palynomorph identification [10–12], but faces the same fundamental limitations as human experts: small labeled datasets and strong site-to-site taphonomic variation. Overcoming these challenges requires both improved modeling of leaf morphology and strategies to augment scarce fossil data. To control for taphonomic variability, we focus here on a single, exceptionally well-documented fossil locality—the late Eocene Florissant Fossil Beds National Monument in Colorado. Preserved under relatively consistent lacustrine conditions [13], the Florissant flora represents one of the best-understood Cenozoic plant assemblages, first described in MacGinitie's seminal 1953 monograph [14] and subsequently expanded and revised through decades of systematic research [15–18].

Herb Meyer's digital archive [19], recently recompiled as part of a large open-access dataset of fossil, cleared, and X-rayed leaves totaling more than 34,000 images [20], provides a uniquely rich and accessible foundation for AI applications in paleobotany. From this resource, we curated 3,200 taxonomically vetted Florissant fossil leaves spanning 23 plant families (16 families with $\geq$5 specimens each). An additional 1,723 Florissant specimens lacking confident family assignments under modern standards were reserved to test the predictive validity of our models.

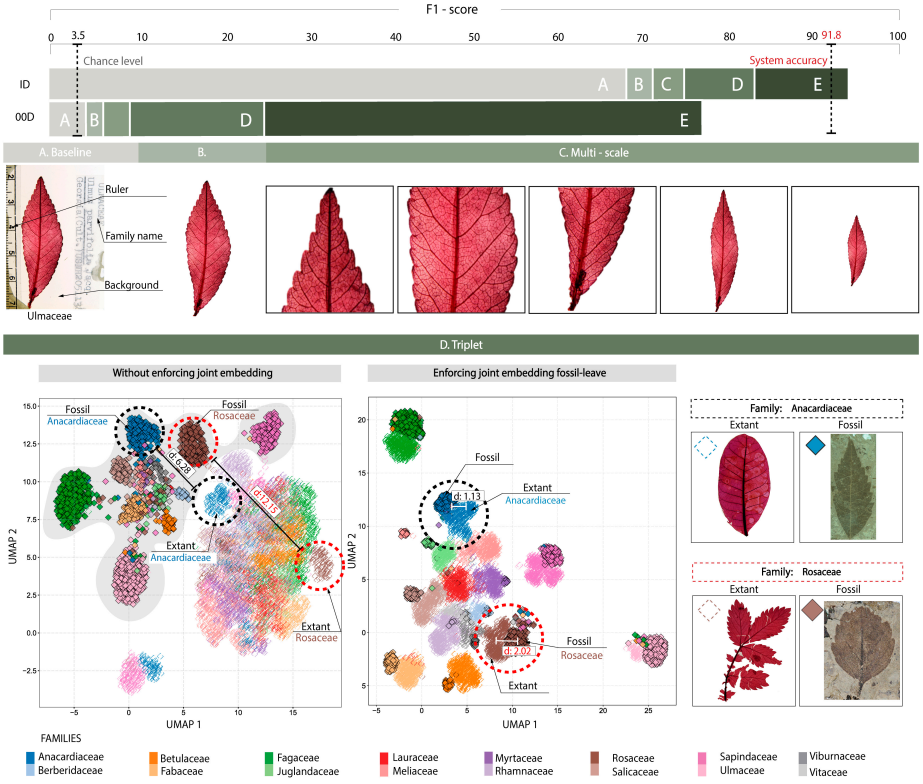
F1 - score
0
3.5
10
20
30
40
50
60
70
80
90
91.8
100
Chance level
System accuracy
ID
00D
A
B
C
D
E
A. Baseline
B.
C. Multi - scale
Ruler
Family name
Background
Ulmaceae
D. Triplet
Without enforcing joint embedding
Enforcing joint embedding fossil-leave
Fossil
Anacardiaceae
Fossil
Rosaceae
d: 6.26
d: 12.13
Extant
Anacardiaceae
Extant
Rosaceae
UMAP 1
UMAP 2
Fossil
Extant
Anacardiaceae
d: 1.13
Fossil
Rosaceae
d: 2.02
Extant
Family: Anacardiaceae
Extant
Fossil
Family: Rosaceae
Extant
Fossil
FAMILIES
Anacardiaceae
Berberidaceae
Betulaceae
Fabaceae
Fagaceae
Juglandaceae
Lauraceae
Meliaceae
Myrtaceae
Rhamnaceae
Rosaceae
Salicaceae
Sapindaceae
Ulmaceae
Viburnaceae
Vitaceae
E. Synthetic fossils
Anacardiceae
Betulaceae
Fagaceae
Rosaceae
Menispermaceae
Malvaceae
Synthetic fossil
Real fossil
No real fossils available only extant

Fig. 1: Overview of system improvements for fossil leaf family identification and their cumulative impact on classification accuracy. **Top:** Top-5 F_1-scores for two generalization scenarios — **in-distribution (ID)**, where real fossils were included during training, and **out-of-distribution (OOD)**, where only synthetic fossils were used. The dashed line marks the 3.5% chance level for 142 families. The final model achieves 91.8% accuracy across extant and fossil leaves. **Panels A–E:** (**A**) Baseline — raw extant leaf image. (**B**) Segmentation — background and artifacts (e.g., ruler, text) removed to ensure the model focuses on leaf morphology. (**C**) Multi-scale learning — each leaf is analyzed at five resolutions to capture both overall shape and fine venation patterns. (**D**) Triplet loss embedding — enforces a shared representation between fossil and extant leaves of the same family: (left) without alignment, clusters separate by domain; (right) with triplet loss, clusters group by family. (**E**) Synthetic fossils — photorealistic fossil images generated to augment families with few or no real fossils. Examples show synthetic fossils (top), real fossils (bottom), and the corresponding extant leaves used as references when real fossils are unavailable.

A central obstacle for fossil classification is that many families have few or no vetted fossil training exemplars. To address this, we employ a generative domain-translation approach to synthesize fossil-like images from extant cleared and X-rayed leaves. Specifically, a cycle-consistent, ControlNet-based model [21] is trained to translate images from 142 modern families (each with $\geq$25 specimens) into fossil-like versions, while using vetted Florissant fossils from 16 families to anchor the mapping between extant and fossil domains. This approach effectively expands the number of families represented in the fossil domain and substantially increases the diversity of training material. The classifier is trained jointly on (i) extant cleared and X-rayed leaves, (ii) real Florissant fossils where available, and (iii) synthetic fossil images generated by the domain translator. Incorporating these synthetic fossils markedly improves family-level accuracy and F1-score for both families with and without real fossil training samples, demonstrating that generative augmentation can overcome long-standing data scarcity in macrofossil paleobotany (Fig. 1).

Results

We developed our deep learning architecture through multiple iterations and cumulative improvements using a ResNet-101 backbone [22], a widely used 101-layer convolutional neural network with strong performance across computer vision tasks. Very similar results were obtained with a transformer architecture (See Supplementary Table S1) Throughout this work, we report the top-5 F1-score of our deep learning model, i.e., the proportion of test images for which the correct family label appears among the five highest-ranked predictions. We chose this metric because it is more informative than simple F1-score, in a multilabel setting with 142 angiosperm families, where top 5 can be informative and exact match F1-score can underestimate performance. The chance-level performance for this classification task is 3.5%. Importantly, we also evaluated model performance in terms of error rate, and found the results to be qualitatively similar, supporting the robustness of our conclusions. All experiments

used (80%, 10%, 10%) train, validation, and test splits, respectively. We systematically tried different model settings (grid search) and selected the best using validation data. All F1-score measures reflect the model F1-score on the test set averaged across independent stratified random splits, $n = 10$.

Evaluating fossil-specific performance is challenging because vetted fossil specimens are available for only 16 of the 142 angiosperm families in our dataset. To quantify classification F1-score specifically for fossil leaves, we restricted our test set to these 16 families while maintaining classification across all 142 families, as a baseline we are only considering the dataset with real samples and without using the synthetic fossils. We evaluated two scenarios to establish performance bounds. In the **in-distribution (ID)** scenario, real fossil specimens from the test family were included during training, with evaluation performed on unseen real specimens of the same family using standard train/validation/test splits. This yielded an average test F1-score of 67.3%, which constitutes the baseline F1 score for our architecture.

In the more challenging **out-of-distribution (OOD)** scenario, we adopted a leave-one-family-out protocol where all real fossils from one family were withheld during training, and the model was tested on the complete set of real fossil leaves from that excluded family. As a baseline, we trained and test only on real examples. This baseline OOD evaluation yielded an overall F1-score of 3.82% on test fossils, reflecting the considerably greater difficulty of generalizing from extant cleared and X-rayed images to fossil leaf images (Fig. 1).

A common challenge in deep learning is the tendency for neural networks to rely on spurious correlations—so-called shortcuts—between image backgrounds and text annotations and class labels. This issue is particularly problematic in scientific applications, where hidden data acquisition biases, annotation artifacts such as handwriting patterns or label sizes, or preparation procedures can lead models to overfit to unintended cues rather than meaningful biological features [23–25].

To assess the robustness of our model against such shortcuts, we computed attribution maps using saliency methods from our custom Xplique toolbox [26]. Visual inspection revealed that the model sometimes based its decisions partly on background elements, potentially due to biases in how family labels were distributed across collections and variations in slide preparation. The model also leveraged annotations and text on slides as classification cues (see Fig. S1 for more details.) Although these shortcuts might improve model ID F1-score, they most likely harm OOD generalization to novel collections and fossil specimens with different backgrounds.

To address this issue, we fine-tuned the Segment Anything Model (SAM) [27] using 576 manually annotated, cleared leaf images, enabling effective leaf segmentation from their background. After applying this model to mask backgrounds across the entire dataset (Fig. 1), we confirmed that classification F1-score remained high (90.5% for the whole system with small, but robust improvements in our OOD (4.6%) and ID (70.1%) evaluations.) Crucially, attribution maps showed that this new model predominantly relies on leaf morphology rather than background artifacts or manual annotations (Fig. S1),

To further improve model performance, we hypothesized, according to standard understanding of leaf architecture [2] , that leaves encode diagnostic features across multiple

spatial scales, from macroscopic traits such as overall shape and margin to mesoscopic patterns such as venation. To capture this morphological information, we extended our architecture to process images at multiple scales. Each input image was decomposed into five sub-images: the original masked image, a tighter crop defined by the smallest bounding box enclosing the leaf mask, and three additional crops corresponding to the basal, middle, and distal thirds of the leaf (Fig. 1). This multiscale approach improved classification F1-score to ID (75.1%) and OOD (8.72%) F1-score .

Fossil leaves often differ markedly from extant specimens due to damage, partial preservation, compression artifacts, and the surrounding rock matrix, which introduces variable background textures. We hypothesized that these differences might lead the model to treat extant and fossil leaves as distinct domains. To test this, we measured the average Euclidean distance between fossil and extant samples within each family in the model's embedding space, normalized by the average inter-family distance (see Fig. 1, Panel D). This analysis showed that fossil and extant samples from the same family were substantially more distant in embedding space (*mean normalized distance* = 11.45, SD = 3.1) than when the model was trained with a triplet loss to enforce cross-domain similarity (*mean normalized distance* = 1.72, SE = 0.2). Because these embedding distances are dimensionless, they directly reflect representational dissimilarity within the neural network. A UMAP visualization of the penultimate-layer embeddings confirms this trend (Fig. 1, Panel D, left), illustrating why general-purpose plant identification systems such as *PlantNet* struggle to recognize fossil leaves.

To address the domain shift, we trained the model with an additional constraint that encourages it to bring leaves from the same family closer together in its internal representation, even when one is a fossil and the other is modern, while pushing apart leaves from different families. This approach (implemented with a triplet-loss; see Methods) explicitly shapes the embedding space so that fossil and extant leaves of the same family are aligned (Fig. 1, Panel D, right). Incorporating this constraint substantially improved performance, reaching F1-score of 82.3% (ID) and 24.6% (OOD), with the largest gains in the OOD scenario where vetted fossil examples were not available during training.

A remaining challenge is the significant imbalance exhibited by our dataset, comprising 34,328 extant and 3,200 fossil leaves—approximately a 10:1 ratio. This problem is compounded by the complete absence of fossil specimens for most families in our dataset (119 of 142 families), with seven additional families having fewer than five samples. To address this limitation, we leveraged generative AI by adapting a stable diffusion model based on the ControlNet architecture [21]. Originally pre-trained on the LAION 5-billion image dataset, this conditional diffusion model synthesizes images guided by text prompts. We fine-tuned the network to generate realistic cleared leaf and fossil images, training it jointly with our ResNet-101 classifier using a combined objective that includes classification F1-score, triplet-loss alignment of extant and fossil representations within families, and the ControlNet loss (Methods). This generative data augmentation dramatically improved OOD generalization, raising OOD F1-score from near-chance levels (3.82%) for our base model to 77.3%, and from 67.4 to 93.2% (ID), substantially narrowing the performance gap between the two evaluation scenarios.

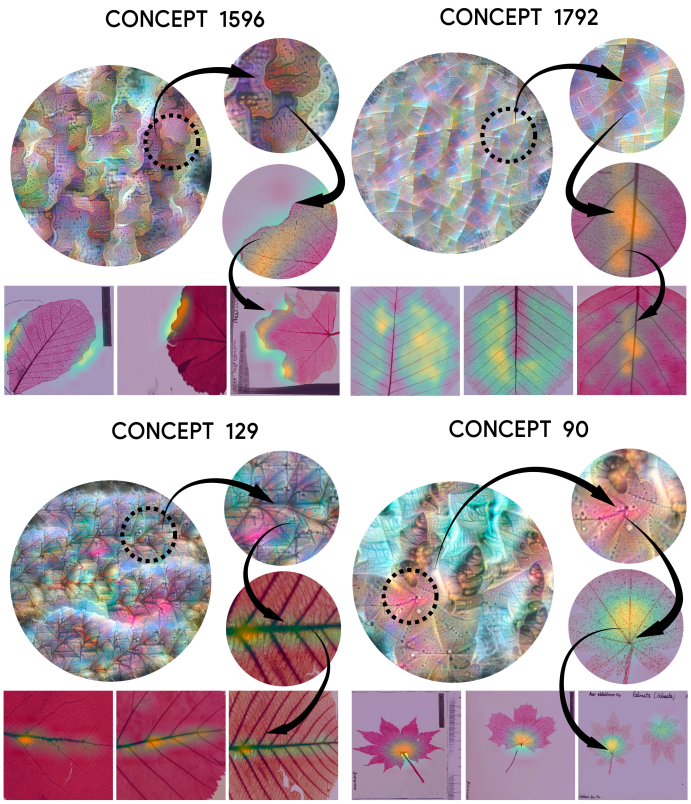

Fig. 2: Examples of visual "concepts" used by the model for classification. Concept 1596: active on the leaf margin. Concept 1792: active in intercostal areas. Concept 129: activates at junctions between primary and secondary veins. Concept 90: activates as leaf base. Additional examples shown on Fig. S3 and Fig. S4. Full set available at https://fel-thomas.github.io/Leaf-Lens/ .

High-performing AI models often function as black boxes, making the rationale behind their predictions difficult to interpret, even though, in the case of leaves, AI they contain abundant, novel taxonomic information [**?** **?**]. To address this, we developed an explainability framework [**?**] to identify internal visual concepts utilized by our fossil-leaf classification model, potentially uncovering diagnostic patterns difficult for human observers to discern [28]. Specifically, our interpretability framework comprises three key stages: (***i***) concept extraction, (***ii***) importance estimation, and (***iii***) concept visualization.

First, we extracted a structured dictionary of visual "concepts" from the model's internal representations using a Top-k Sparse Autoencoder [**?**]. This step is critical because deep neural networks often suffer from the problem of polysemanticity [**?**], with individual model units encoding multiple "semantic" dimensions simultaneously. To reveal this latent structure, we employ sparse dictionary learning [29–31] on the

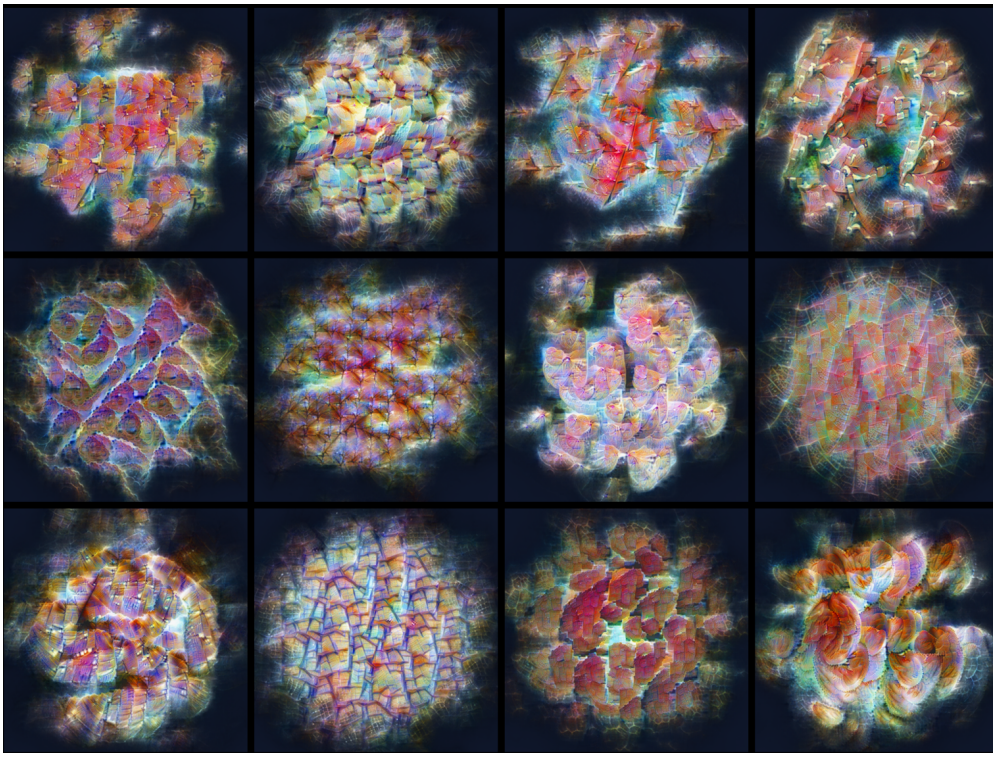

Fig. 3: Sample of high-level features used by the model to perform classification.

penultimate layer activations [**?**] to derive a compact, disentangled set of visual concepts. Representative examples of these visual concepts are shown in Fig. 2 and Fig. 3. See Fig. S3 and Fig. S4 for additional examples and https://serre-lab.github.io/Lens/ for the full set.

Second, we quantified the importance of each extracted concept for the model predictions. Rather than treating all concepts equally, we reparameterized the classifier in concept space, directly attributing model outputs to underlying concepts via linear mapping [**?**]. Each concept, hence, receives a class-specific importance score reflecting its influence on the decision boundary.

Third, we visualized these concepts using feature visualization methods [**? ?**] in two ways to enhance interpretability: heat maps [**? ? ? ?**], which indicate where individual concepts are activated within each image, and maximally exciting images, which reveal the visual patterns that most strongly activate each concept. Together, these stages provide a transparent decomposition of the model's internal reasoning, enhancing interpretability and trust in model predictions (See Fig. 2 also explore more concepts in https://serre-lab.github.io/Lens/). Also observe high level features in 3.

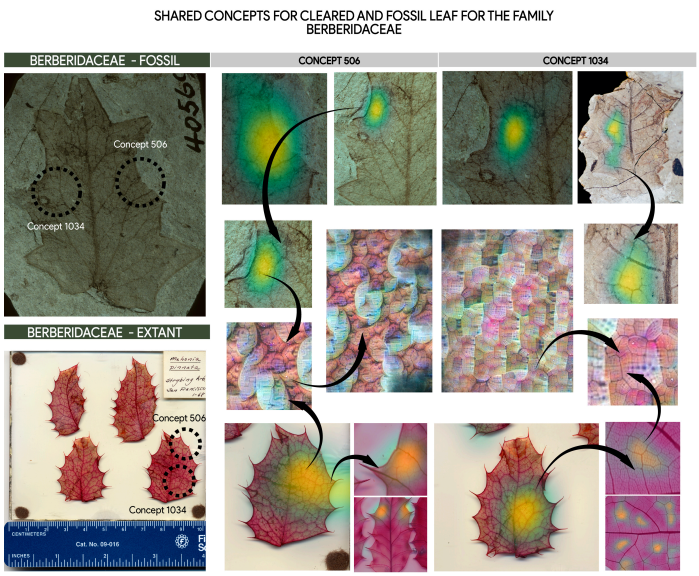

Fig. 4: Shared concepts used by the model to classify extant and fossil examples from the family Berberidaceae.

Demo Application

To facilitate broader accessibility and practical use of our system, we developed an interactive web application available at https://huggingface.co/spaces/Serrelab/fossil_app. This demo enables users to upload fossil leaf images and obtain top family-level predictions generated by our model. In addition to classification results, the application provides saliency maps that visually highlight the regions of the leaf image most influential to the model's decision, offering an interpretable and user-friendly interface for exploring fossil leaf identification.

Finally, we present the first machine-assisted classification of 1,723 fossil samples from the Florissant collection, encompassing specimens with previously unknown family labels or historical identifications no longer considered robust (See Figure 5. These specimens came from the same sources as those in the vetted set [20] but did not pass the vetting criteria. These corresponds only to dicot angiosperm leaves (i.e., no monocot angiosperms due to very limited sample size). The model's predictions and accompanying interpretability results were evaluated by the two team paleobotanists (P.W. and H.M.; Fig. 5). Their assessment revealed that 485 specimens (28.2%) were either severely degraded, belonged to fossil categories other than dicot leaves, or lacked sufficient morphological detail for reliable human identification; these samples were therefore excluded from further evaluation. Among the remaining 1,238 specimens, the paleobotanists identified 585 classifications as intriguing and an additional 505 classifications as plausible, collectively representing promising candidates for detailed follow-up studies and offering abundant new opportunities for paleobotanical research

on the Florissant flora. The experts found the model's predictions to be implausible for only 143 specimens (approximately 12%), highlighting the system's robustness and potential utility for assisting paleobotanical classification tasks. An extensive list of specimens and computer annotations is available on our website [1].

Conclusion

We present a robust framework that advance one of paleobotany's central challenges—accurate identification of fossil angiosperm leaves—and, importantly, ongoing cross-domain training and refinement promise broader applicability across diverse fossil sites, substantially extending its scientific impact. By harnessing modern generative AI to synthesize realistic fossil images from extant leaf data, our system attains high identification F1-scores even for families lacking fossil training examples. Using state-of-the-art interpretability methods, it also surfaces botanically meaningful cues by visually summarizing subtle morphological features that define families across fossil and extant specimens, suggesting new diagnostic characters.
Although our current model substantially improves the classification of previously unknown Florissant fossils, the same cross-domain strategy is readily generalizable to other deposits, positioning this approach for broad use. In the near term, it offers paleobotanists a practical, interpretable tool to reduce longstanding uncertainties in Florissant flora identifications and, more broadly, to advance our understanding of the evolution and ecological dynamics of ancient terrestrial ecosystems.

Methods

Families selected for this study

The dataset used in this study comprises cleared and x-rayed images of specimens from the following dicot (non-monocot angiosperm) families: Acanthaceae, Achariaceae, Actinidiaceae, Altingiaceae, Amaranthaceae, Anacardiaceae, Annonaceae, Apiaceae, Apocynaceae, Aquifoliaceae, Araliaceae, Aristolochiaceae, Asteraceae, Atherospermataceae, Berberidaceae, Betulaceae, Bignoniaceae, Boraginaceae, Burseraceae, Buxaceae, Calophyllaceae, Calycanthaceae, Campanulaceae, Canellaceae, Cannabaceae, Capparaceae, Caprifoliaceae, Cardiopteridaceae, Celastraceae, Cercidiphyllaceae, Chloranthaceae, Chrysobalanaceae, Clusiaceae, Combretaceae, Connaraceae, Coriariaceae, Cornaceae, Crassulaceae, Cucurbitaceae, Cunoniaceae, Dilleniaceae, Dipterocarpaceae, Ebenaceae, Elaeagnaceae, Elaeocarpaceae, Ericaceae, Euphorbiaceae, Eupteleaceae, Fabaceae, Fagaceae, Flacourtiaceae, Garryaceae, Grossulariaceae, Hamamelidaceae, Icacinaceae, Iteaceae, Juglandaceae, Lauraceae, Lardizabalaceae, Lythraceae, Magnoliaceae, Malpighiaceae, Malvaceae, Melastomataceae, Meliaceae, Menispermaceae, Monimiaceae, Moraceae, Myristicaceae, Myrtaceae, Nyssaceae, Oleaceae, Papaveraceae, Passifloraceae, Phyllanthaceae, Pittosporaceae, Platanaceae, Polygonaceae, Proteaceae, Rhamnaceae, Rhizophoraceae, Rosaceae, Rubiaceae, Rutaceae, Salicaceae, Sapindaceae, Sapotaceae, Saxifragaceae,

[1] https://serre-lab.github.io/prj_fossil_unknown/

Schisandraceae, Simaroubaceae, Smilacaceae, Solanaceae, Staphyleaceae, Stemonuraceae, Symplocaceae, Tapisciaceae, Theaceae, Thymelaeaceae, Trochodendraceae, Ulmaceae, Urticaceae, Vitaceae, and Winteraceae.
The families for which there were also fossil samples available are: Anacardiaceae, Berberidaceae, Betulaceae, Fabaceae, Fagaceae, Juglandaceae, Lauraceae, Meliaceae, Myrtaceae, Rhamnaceae, Rosaceae, Salicaceae, Sapindaceae, Ulmaceae, Viburnaceae, and Vitaceae.

Performance measures

We evaluate our method using two distinct scenarios to establish in-distribution (ID) and out-of-distribution (OOD) as upper and lower bounds on performance:

1. **ID scenario:** A standard 80/10/10 training/test/cross-validation where real fossils from each family are included in both training and testing sets.
2. **OOD scenario:** A leave-one-family-out cross-validation, in which the training set excludes all real fossils from the test family (only fossils from the other families are included), and evaluation is performed solely on the withheld family.

These scenarios provide robust estimates of ID (best-case) and OOD (worst-case) classification performance. Specifically, we report the F1-score, defined as:

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

where Precision $= \frac{TP}{TP+FP}$ and Recall $= \frac{TP}{TP+FN}$ with TP, FP, and FN denoting true positives, false positives, and false negatives, respectively.
Additionally, we generated embedding visualizations with and without the triplet-loss regularizer, and performed qualitative analyses of attribution maps using Grad-CAM [**?**] and RISE [32] methods. These analyses further validated that our approach significantly enhances model interpretability and F1-score. Preliminary feedback from expert paleobotanists suggests that our method effectively highlights clear, family-specific visual features, thereby supporting the identification of previously unknown fossil specimens.

Automated training dataset cleanup

To ensure that our models rely exclusively on intrinsic leaf morphology rather than external artifacts, we fine-tuned the Segment Anything Model (SAM; [27]) on 576 manually annotated cleared-leaf images using an 80/20 training-validation split. The fine-tuned model achieved an Intersection over Union (IoU) of 95%, demonstrating highly accurate segmentation performance. This segmentation model was subsequently applied to the entire dataset. Visual inspections revealed that the approach effectively removes background artifacts and annotations, thus encouraging the family classifier to focus solely on relevant morphological features.

Representation learning and triplet loss

Building on [33], we enhanced our classification architecture by adding a dedicated embedding head optimized with triplet-loss regularization alongside the standard cross-entropy classification head. For each training batch, we sample an anchor example with a positive (same-family) and a negative (different-family) using a hard sampling approach, which means that we picked the furthest positive and the closest negative (see S2) from both extant and fossil domains. The triplet loss encourages embeddings such that anchor-positive pairs are closer together than anchor-negative pairs by at least a margin m. Formally, the triplet loss is defined as:

$$\mathcal{L}_{triplet} = \frac{1}{b}\sum_{i=1}^{b}\left[D(a_{i,x}, p_{i,x}) - D(a_{i,x}, n_{i,x}) + m\right], \tag{1}$$

where b denotes batch size, m represents the margin, x being the domain. $D(\cdot)$ is a suitable distance metric, and a, p, and n represent anchor, positive, and negative samples, respectively. The combined objective function used during training is:

$$\mathcal{L} = \mathcal{L}_{crossentropy} + \lambda\mathcal{L}_{triplet}, \tag{2}$$

where λ controls the relative contribution of the triplet loss. To further enhance robustness, synthetic fossil images generated from extant leaves were introduced to augment families with limited or absent fossil data. This strategy reinforces both local (within-family) and global (cross-family) structure within the embedding space, significantly improving model generalization, especially for families lacking representative fossil samples.

To quantitatively assess representational differences between the fossil and extant leaf domains and the impact of the triplet loss, we analyzed the learned embedding space by measuring how closely samples from the same family cluster together, regardless of whether they are extant or fossil. Specifically, for each family, we computed the average Euclidean distance between embeddings of extant and fossil leaves within that family (i.e., the "intra-family, inter-domain" distance). We then compared these "intra-family, inter-domain" distances to the average distance between embeddings from different families ("inter-family" distances). To ensure comparability across families, we normalized the "intra-family, inter-domain" distances by the mean "inter-family" distance for each family. This normalization allows us to interpret the results as a relative measure: values less than one indicate that, on average, extant and fossil leaves from the same family are closer to each other in the embedding space than to leaves from other families.

Our results show that, after applying the triplet loss, the normalized inter-domain, intra-family distances are substantially reduced, reducing from an average of 11.45 with a standard deviation of 3.1 to an average of inter-family distance of just 1.72 with 0.3 standard deviation. This demonstrates that the model learns to align fossil and extant leaves from the same family in the embedding space, effectively bridging the domain gap. In other words, the triplet loss encourages the network to focus on family-level morphological features that are consistent across domains, rather than

domain-specific artifacts. This effect is also visually apparent in the UMAP projection of the embeddings (Fig. 1), where fossil and extant leaves from the same family form tight, overlapping clusters.

Synthetic fossil generation

Let X denote the fossil domain and Y the extant (cleared leaves) domain. While these domains share morphological characteristics, their visual appearance often differs substantially due to preservation conditions, artifacts, and inherent variability in fossilization. Furthermore, many families represented in domain Y lack corresponding fossil examples in domain X.

To bridge this domain gap, we adopted a generative image-to-image translation approach inspired by CycleGAN [34], implemented using the ControlNet architecture [21]. Specifically, we defined two mappings, $G : X \to Y$ (fossil-to-extant) and $F : Y \to X$ (extant-to-fossil), each guided by domain-specific text prompts:

- "A cleared leaf of the family: <Family Name>"
- "A fossilized leaf of the family: <Family Name>"

These text-based prompts guide ControlNet to generate synthetic images that preserve family-specific morphological features. Importantly, the generative model (ControlNet) and the classifier are trained together in an end-to-end fashion. The overall training loss is a weighted sum of the classification loss, the triplet loss, and the generative (ControlNet) loss, with all components optimized jointly during end-to-end training. This joint optimization ensures that the embeddings of both real and generated images are aligned within the same family across domains, and that the synthetic images produced are directly beneficial for the classification objective. At each training step, both the generator and classifier parameters are updated simultaneously, ensuring that synthetic fossil generation and classification performance are tightly coupled.

Specifically,

$$\mathcal{L}_{total} = \mathcal{L}_{classif} + \lambda_{trip}\mathcal{L}_{trip} + \lambda_{gen}\mathcal{L}_{gen} \tag{3}$$

where $\mathcal{L}_{classif}$ is the standard classification loss, $\mathcal{L}_{triplet}$ is the triplet loss as defined above, $\mathcal{L}_{gen}$ is the generative loss for ControlNet, and $\lambda_{triplet}$ and λ_{gen} are weighting coefficients for the triplet and generative losses, respectively. Additionally, we applied the fine-tuned SAM model during synthetic image generation, ensuring that leaf shapes in synthetic images remain faithful to their input counterparts, thereby improving both visual realism and model generalization. Please find details for training in (SI: Details for training) .

Concept Extraction

We consider a dataset of input images $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{n}$, where $\boldsymbol{x}_i \in \mathbb{R}^{H\times W\times 3}$. Each image is processed by a visual encoder $\boldsymbol{f}$, yielding a latent feature tensor $\boldsymbol{A}_i = \boldsymbol{f}(\boldsymbol{x}_i) \in \mathbb{R}^{h\times w\times d}$, with spatial dimensions (h, w) and feature dimension d. We aggregate activations across the dataset by flattening the spatial dimensions of each $\boldsymbol{A}_i$ and concatenating them into a global activation matrix $\boldsymbol{A} \in \mathbb{R}^{nhw\times d}$.

It is tempting to assume that meaningful visual concepts align with the d coordinate axes (neurons) of the activation space, suggesting a hard upper bound of d independently representable features. Yet this view underestimates the complexity of neural representations. Empirical findings indicate that, despite residing in a d-dimensional space, neural networks routinely encode many more than d distinct features [**? ? ?**]—a phenomenon referred to as superposition [35], which bears resemblance to a kind of distributed, compressed coding.

To put it simply, rather than relying on strictly orthogonal directions, as in conventional linear representations, the network appears to adopt a looser geometry, one in which directions are merely sufficiently decorrelated to permit reliable disambiguation. This strategy enables a form of representational overloading: distinct features are encoded along vectors that, while not strictly orthogonal, remain distinguishable within the tolerances of the task.

Given this understanding of the representation geometry of neural networks, overcomplete sparse dictionary learning [29, 36**?** –40] has emerged as the natural tool to reverse-engineer this geometry [**? ? ? ? ? ? ? ?**]. Rather than assuming that concepts align with the original d coordinate axes, we seek to discover a larger set of $c \gg d$ directions in the activation space, each corresponding to a semantically meaningful visual concept. Formally, we decompose the activation matrix $\boldsymbol{A}$ using a learned overcomplete dictionary $\boldsymbol{D} \in \mathbb{R}^{c \times d}$ and sparse coefficient matrix $\boldsymbol{Z} \in \mathbb{R}^{nhw \times c}$ by solving:

$$\min_{\boldsymbol{Z},\boldsymbol{D}} \|\boldsymbol{A} - \boldsymbol{Z}\boldsymbol{D}\|_F^2 \quad \text{subject to} \quad \|\boldsymbol{Z}_i\|_0 \leq k, \quad \boldsymbol{Z}_i \geq 0, \quad \|\boldsymbol{D}_j\|_2 = 1 \quad \forall i, j \in [c] \tag{4}$$

The sparsity constraint is crucial: it reflects the empirical observation that while the model may have access to thousands of concepts, any given input activates only a small subset, consistent with the superposition hypothesis. In practice, we implement this decomposition using a Top-k Sparse Autoencoder [**? ?**]. The encoder applies a learned linear projection, followed by ReLU activation and a Top-k sparsification operator $\boldsymbol{\Pi}_k(\cdot)$ that retains only the k highest activations: $\boldsymbol{Z} = \boldsymbol{\Pi}_k(\mathrm{ReLU}(\boldsymbol{A}\boldsymbol{W}_{\mathtt{enc}}))$. The decoder reconstructs activations as $\hat{\boldsymbol{A}} = \boldsymbol{Z}\boldsymbol{D}$, where each dictionary atom $\boldsymbol{D}_j$ represents a distinct visual "concept" whose semantic meaning emerges through the optimization process.

Concept Quantification

Having decomposed the neural representation into interpretable concepts (See Fig 3), we address the question of quantifying their individual contributions to model predictions. The sparse dictionary learning framework yields a linear reparameterization that renders explicit the relationship between concept activations and classification decisions. The original classifier computes prediction logits as $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{W}$ for weight matrix $\boldsymbol{W} \in \mathbb{R}^{d \times C}$, where C denotes the number of output classes. Our decomposition $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{D}$ permits the reformulation $\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{W}$, which exposes the decision process as a two-stage computation: sparse concept activation followed by linear aggregation according to learned weights. Specifically, we define the concept-to-class mapping

matrix:

$$\boldsymbol{\Gamma} = \boldsymbol{D}\boldsymbol{W} \in \mathbb{R}^{c \times C}, \quad \text{such that} \quad \boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\Gamma} \tag{5}$$

Each entry Γ_{ij} quantifies the influence of concept i on class j logits. The contribution s_i of concept i to target class t follows directly:

$$s_i = \Gamma_{it} Z_i \tag{6}$$

This formulation is in fact the gradient-input and exhibits a notable property: in the linear regime, it converges with other established attribution methods such as Integrated gradients [**?**], Occlusion [**?**] and RISE [**?**]. Interestingly, prior work has demonstrated that gradient-input (still in the linear regime) achieves optimality with respect to standard fidelity metrics [**? ?**]. This convergence provides empirical support for using gradient-input as a principled measure of concept importance in our framework.

Concept Visualization

The preceding analysis establishes a mathematically coherent framework for decomposing neural representations into interpretable concepts and quantifying their contributions to classification decisions. To ground these abstract concept directions in visual semantics, we adopt a dual approach that leverages both natural image contexts and synthetic optimization to characterize the visual pattern encoded by each dictionary atom. We first examine concept activations within their natural spatial context. For a given input image yielding activation tensor $\boldsymbol{A}_i = \boldsymbol{f}(\boldsymbol{x}_i) \in \mathbb{R}^{h \times w \times d}$, we apply the learned Top-k encoder at each spatial location $\boldsymbol{Z}_{\text{spatial}}(u, v) = \boldsymbol{\Pi}_k(\boldsymbol{A}_i(u, v)\boldsymbol{W}_{\texttt{enc}})$, where (u, v) indexes spatial coordinates. This procedure generates concept-specific activation maps that reveal the precise image regions where each learned direction $\boldsymbol{D}_j$ exhibits significant projection, thereby localizing the visual pattern that drive concept activations. Complementing this natural localization, we synthesize prototypical inputs through direct optimization in the input space using Feature Visualization [**? ?**]. For concept j, we solve:

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} Z_j\big(\boldsymbol{\Pi}_k(\boldsymbol{f}(\boldsymbol{x})\boldsymbol{W}_{\texttt{enc}})\big) - \lambda\Omega(\boldsymbol{x}) \tag{7}$$

beginning from random initialization and employing gradient ascent with respect to pixel values. To ensure the resulting visualizations remain within the manifold of natural images, we incorporate regularization strategies $\Omega(\cdot)$ from the MACO framework [41], which constrain the optimization to produce coherent visual patterns. Combined, spatial activation maps and synthetic prototypes provide complementary interpretations: spatial activation maps demonstrate how concepts "fire" with in natural contexts while the Feature Visualization approach allows visualizing the learned dictionary $\boldsymbol{D}$ under the form of synthetic abstract image that capture the minimal visual pattern that drive each concepts.

References

[1] Dilcher, D.L.: Revision of the eocene flora of southeastern north america. Journal of Palaeosciences **20**(1-3), 7–18 (1971) https://doi.org/10.54991/jop.1971.882

[2] Hickey, L.J., Wolfe, J.A.: The bases of angiosperm phylogeny: Vegetative morphology. Annals of the Missouri Botanical Garden **62**(3), 538–589 (1975) https://doi.org/10.2307/2395267

[3] Wilf, P.: Fossil angiosperm leaves: Paleobotany's difficult children prove themselves. Paleontological Society Papers **14**, 319–333 (2008) https://doi.org/10.1017/S1089332600001741

[4] Giraldo, L.A., Wilf, P., Donovan, M.P., Kooyman, R.M., Gandolfo, M.A.: Fossil insect-feeding traces indicate unrecognized evolutionary history and biodiversity on australia's iconic Eucalyptus. New Phytologist **245**, 1762–1773 (2025) https://doi.org/10.1111/nph.20316

[5] Hickey, L.J., Doyle, J.A.: Early cretaceous fossil evidence for angiosperm evolution. The Botanical Review **43**, 2–104 (1977)

[6] Johnson, K.R., Nichols, D.J., Attrep, D.J.J., Orth, C.J.: High-resolution leaf-fossil record spanning the cretaceous–tertiary boundary. Nature **340**, 708–711 (1989) https://doi.org/10.1038/340708a0

[7] Kooyman, R.M., Wilf, P., Barreda, V.D., Carpenter, R.J., Jordan, G.J., Sniderman, J.M.K., Allen, A., Brodribb, T.J., Crayn, D., Feild, T.S., Laffan, S.W., Lusk, C.H., Rossetto, M., Weston, P.H.: Paleo-antarctic rainforest into the modern old world tropics: the rich past and threatened future of the "southern wet forest survivors". American Journal of Botany **101**, 2121–2135 (2014) https://doi.org/10.3732/ajb.1400340

[8] Wing, S.L., Harrington, G.J., Smith, F.A., Bloch, J.I., Boyer, D.M., Freeman, K.H.: Transient floral change and rapid global warming at the paleocene–eocene boundary. Science **310**, 993–996 (2005) https://doi.org/10.1126/science.1116913

[9] Wilf, P., Zhang, S., Chikkerur, S., Little, S.A., Wing, S.L., Serre, T.: Computer vision cracks the leaf code. Proceedings of the National Academy of Sciences **113**(12), 3305–3310 (2016) https://doi.org/10.1073/pnas.1524473113 https://www.pnas.org/doi/pdf/10.1073/pnas.1524473113

[10] Gunjal, M., Thorat, A., Londhe, G., Kamble, A., Gholap, C., Patil, P.: Plant species classification using deep learning. In: 2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), pp. 1–6 (2024). https://doi.org/10.1109/ICONSTEM60960.2024.10568579

[11] Wäldchen, J., Mäder, P.: Machine learning for image based

species identification. Methods in Ecology and Evolution **9**(11), 2216–2225 (2018) https://doi.org/10.1111/2041-210X.13075 https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13075

[12] Adaïmé, M.-É., Kong, S., Punyasena, S.W.: Deep learning approaches to the phylogenetic placement of extinct pollen morphotypes. PNAS Nexus **3**(1), 419 (2024) https://doi.org/10.1093/pnasnexus/pgad419

[13] Harding, I.C., Chant, L.S.: Self-sedimented diatom mats as agents of exceptional fossil preservation in the oligocene florissant lake beds, colorado, united states. Geology **28**(3), 195–198 (2000) https://doi.org/10.1130/0091-7613(2000)28<195:SDMAAO>2.0.CO;2

[14] MacGinitie, H.D.: Fossil Plants of the Florissant Beds, Colorado. Carnegie Institution of Washington Publication, vol. 599, pp. 1–198. Carnegie Institution of Washington, Washington, DC (1953). https://books.google.com/books?id=GeEzQEACAAJ

[15] Manchester, S.R., Crane, P.R.: Attached leaves, inflorescences, and fruits of Fagopsis, an extinct genus of fagaceous affinity from the oligocene florissant flora of colorado, U.S.A. American Journal of Botany **70**(8), 1147–1164 (1983) https://doi.org/10.2307/2443285

[16] Manchester, S.R.: Update on the megafossil flora of florissant, colorado. In: Evanoff, E., Gregory-Wodzicki, K.M., Johnson, K.R. (eds.) Fossil Flora and Stratigraphy of the Florissant Formation, Colorado. Proceedings of the Denver Museum of Nature & Science, Series 4, pp. 137–161. Denver Museum of Nature & Science, Denver, CO (2001). No DOI available. https://npshistory.com/publications/flfo/dmnsp-v4n1-2001.pdf

[17] Jia, H., Manchester, S.R.: Fossil leaves and fruits of Cercis l. (leguminosae) from the eocene of western north america. International Journal of Plant Sciences **175**(5), 601–612 (2014) https://doi.org/10.1086/675693

[18] Herendeen, P.S., Herrera, F.: Eocene fossil legume leaves referable to the extant genus Arcoa (caesalpinioideae, leguminosae). International Journal of Plant Sciences **180**(3), 220–231 (2019) https://doi.org/10.1086/701468

[19] Meyer, H.W., Wasson, M.S., Frakes, B.J.: Development of an integrated paleontological database and web site of florissant collections, taxonomy, and publications. In: Meyer, H.W., Smith, D.M. (eds.) Paleontology of the Upper Eocene Florissant Formation, Colorado. Geological Society of America Special Paper, vol. 435, pp. 159–177. Geological Society of America, Boulder, Colorado (2008). https://doi.org/10.1130/2008.2435(11) . https://doi.org/10.1130/2008.2435(11)

[20] Wilf, P., Wing, S.L., Meyer, H.W., Rose, J.A., Saha, R., Serre, T., Cúneo, N.R., Donovan, M.P., Erwin, D.M., Gandolfo, M.A., González-Akre, E., Herrera, F.,

Hu, S., Iglesias, A., Johnson, K.R., Karim, T.S., Zou, X.: Image collection and supporting data for: An image dataset of cleared, x-rayed, and fossil leaves vetted to plant family for human and machine learning. Figshare+ (2024). https://doi.org/10.25452/figshare.plus.14980698.v2 . https://doi.org/10.25452/figshare.plus.14980698.v2

[21] Zhang, L., Rao, A., Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models (2023). https://arxiv.org/abs/2302.05543

[22] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015) 1512.03385

[23] Brown, A., Tomašev, N., Freyberg, J., Liu, Y., Karthikesalingam, A., Schrouff, J.: Detecting shortcut learning for fair medical AI using shortcut testing. Nature Communications **14**, 4314 (2023) https://doi.org/10.1038/s41467-023-39902-7

[24] Nauta, M., Walsh, R., Dubowski, A., Seifert, C.: Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. Diagnostics **12**(1), 40 (2022) https://doi.org/10.3390/diagnostics12010040

[25] Hill, B.G., Koback, F.L., Schilling, P.L.: The risk of shortcutting in deep learning algorithms for medical imaging research. Scientific Reports **14**, 29224 (2024) https://doi.org/10.1038/s41598-024-79838-6

[26] Fel, T., Hervier, L., Vigouroux, D., Poche, A., Plakoo, J., Cadène, R., Chalvidal, M., Colin, J., Boissin, T., Béthune, L., Picard, A., Nicodeme, C., Gardes, L., Flandin, G., Serre, T.: Xplique: A deep learning explainability toolbox. (2022). CVPR Workshop on Explainable Artificial Intelligence for Computer Vision

[27] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R.: Segment Anything (2023). https://arxiv.org/abs/2304.02643

[28] Spagnuolo, E., Wilf, P., Serre, T.: Decoding family-level features for modern and fossil leaves from computer-vision heat maps. American Journal of Botany **109** (2022) https://doi.org/10.1002/ajb2.1842

[29] Mairal, J., Bach, F., Ponce, J.: Sparse modeling for image and vision processing. Foundations and Trends in Computer Graphics and Vision (2014)

[30] Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature (1996)

[31] Tošić, I., Frossard, P.: Dictionary learning. IEEE Signal Processing Magazine (2011)

[32] Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for

Explanation of Black-box Models (2018). https://arxiv.org/abs/1806.07421

[33] Taha, A., Chen, Y.-T., Misu, T., Shrivastava, A., Davis, L.: Boosting standard classification architectures through a ranking regularizer, pp. 747–755 (2020). https://doi.org/10.1109/WACV45572.2020.9093279 . Publisher Copyright: © 2020 IEEE

[34] Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. (2017)

[35] Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., Olah, C.: Toy Models of Superposition (2022). https://arxiv.org/abs/2209.10652

[36] Tošić, I., Frossard, P.: Dictionary learning. IEEE Signal Processing Magazine (2011)

[37] Rubinstein, R., Bruckstein, A.M., Elad, M.: Dictionaries for sparse representation modeling. Proceedings of the IEEE (2010)

[38] Elad, M.: Sparse and redundant representations: from theory to applications in signal and image processing (2010)

[39] Dumitrescu, B., Irofti, P.: Dictionary learning algorithms and applications (2018)

[40] Hurley, N., Rickard, S.: Comparing measures of sparsity. IEEE Transactions on Information Theory (2009)

[41] Fel, T., Boissin, T., Boutin, V., Picard, A., Novello, P., Colin, J., Linsley, D., Rousseau, T., Cadène, R., Gardes, L., Serre, T.: Unlocking feature visualization for deeper networks with MAgnitude constrained optimization. (2023)

[42] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization **128**(2), 336–359 (2019) https://doi.org/10.1007/s11263-019-01228-7

[43] Bao, H., Dong, L., Piao, S., Wei, F.: BEiT: BERT pre-training of image transformers. (2022). https://openreview.net/forum?id=p-BhZSz59o4

Funding sources and technical assistance

Yuxian Wang developed the gradio demo app. Paula Vargas provided assistance with the final figure editing, and Jacob Rose with the initial curation of the image datasets. We thank Edward Spagnuolo, Teng-Xiang Wang, L. Alejandro Giraldo, and Steven Manchester for helpful discussions on the paleobotanical aspects of this work.

This work was funded by an NSF FRES grant (EAR-1925481 to T.S. and EAR-1925755 to P.W.). Computing support was provided by the Center for Computation and Visualization (CCV) (via NIH Office of the Director grant S10OD025181). We also acknowledge the Cloud TPU hardware resources that Google graciously makes available via the TensorFlow Research Cloud (TFRC) program.

Author Contributions

P.W. and T.S. conceptualized and supervised the research. I.F.R., M.V., T.F., and T.S. developed the artificial intelligence system. T.F. and T.S. developed the interpretability and explainability tools. G.G. conducted analyses, developed software tools, and created website resources. P.W. and H.M. provided vetted fossil samples and expert assessments of the AI-based fossil identifications. I.F.R., T.F., P.W., and T.S. drafted the manuscript. All authors reviewed and edited the manuscript and approved the final version.

Competing Interests

None

Supplementary Information

Shortcut Removal via SAM

We conducted a qualitative evaluation of the attribution maps generated by our Baseline and Segmented models using RISE (Randomized Input Sampling for Explanation) [32] and GradCAM (Gradient-weighted Class Activation Mapping) [42]. These methods were implemented using the Xplique Toolbox [26], developed in our lab. Our findings indicate that the attribution maps of the Segmented model, trained with segmentation, predominantly focus on the leaf regions within the image, demonstrating alignment with relevant features. In contrast, the Baseline model exhibited attribution patterns that relied on spurious shortcuts, such as text and rulers present in the images, as illustrated in Figure S1.

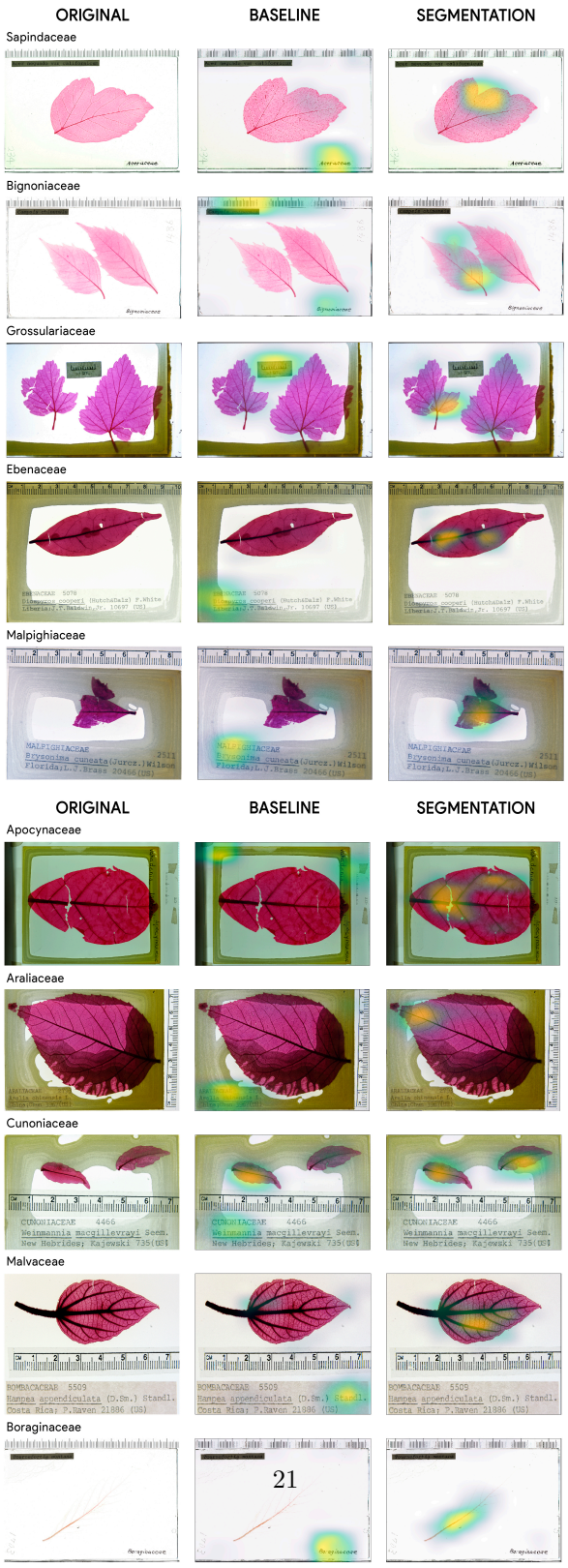

Fig. S1: Attribution map generated using Grad-Cam showing before and after segmentation. For each sample , you will find three columnns : Original image on the left, attributions when no segmentation is performed in the middle and in the third column you will attribution maps from the model trained on segmented leaves reducing the shortcuts used by the model.

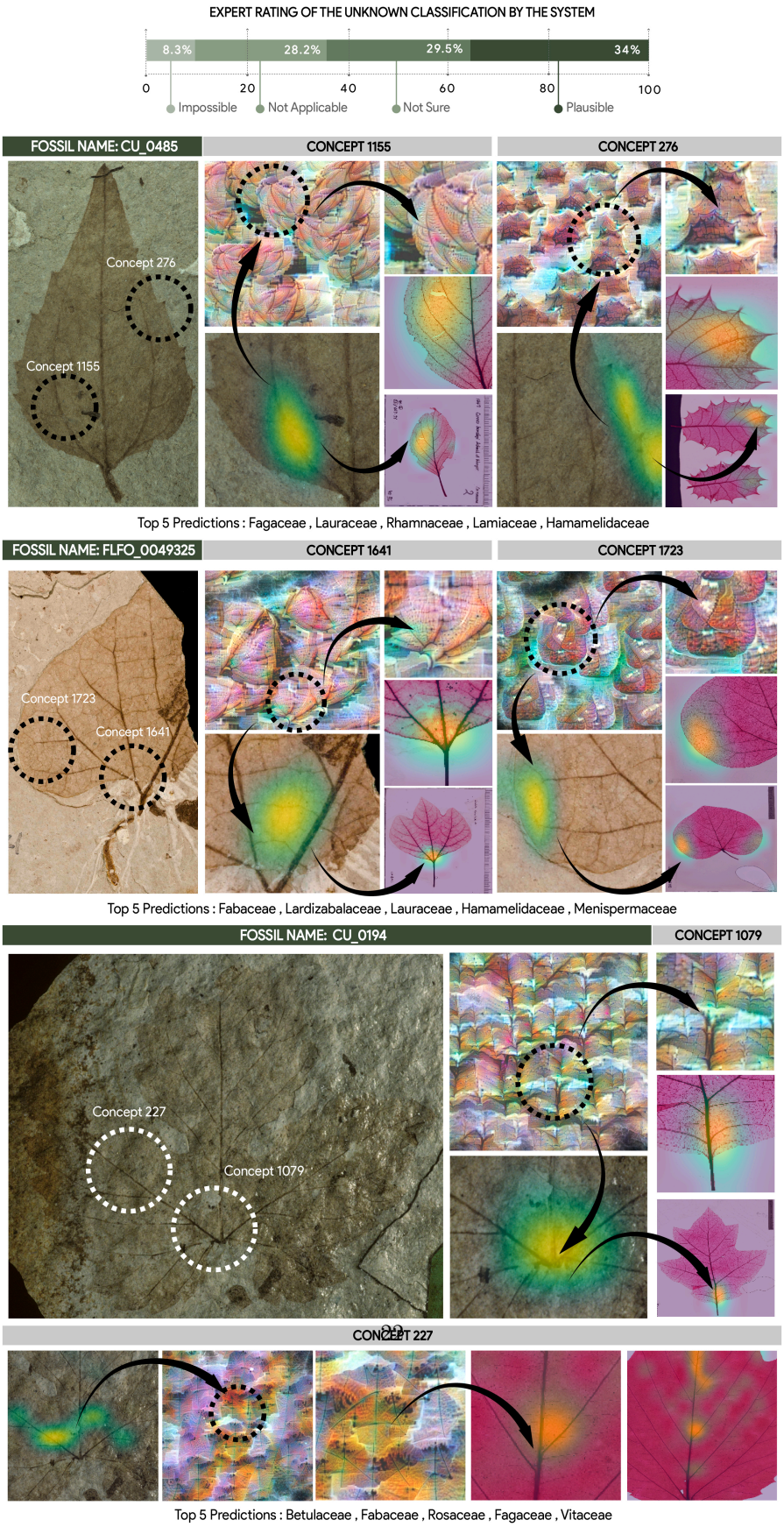

Fig. 5: **Top:** Expert evaluation of model predictions for previously unidentified Florissant fossils (categories: confident, plausible, uncertain, and not applicable). **Bottom:** Example unidentified fossils with the model's Top-5 family predictions and the most influential learned concepts (e.g., secondary-vein arrangement, marginal dentition, honeycomb-like areolation, and the junction of primary and secondary veins). The full set of predictions, visual explanations, and downloadable data are available at https://serre-lab.github.io/prj_fossil_unknown/.

System

Illustration of the system modules.

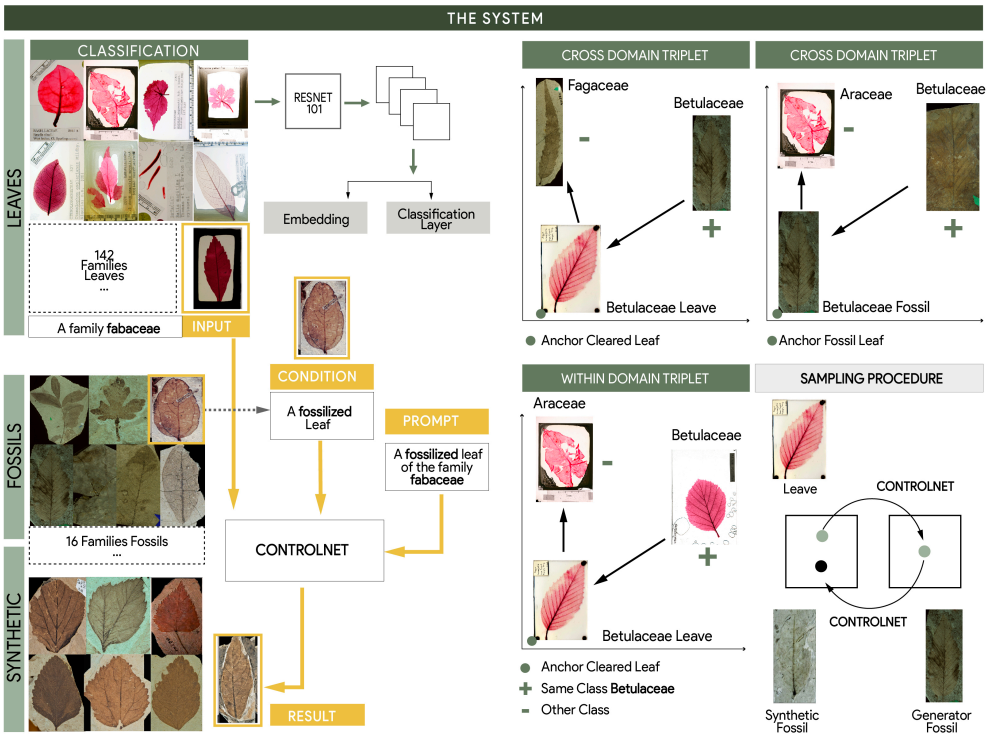

Fig. S2: Illustration of the System. Top left: Representation of the cleared leaf dataset. Images sampled from 142 families are used as input to the Controlnet to generate synthetic fossils. Bottom left: Representation of the fossil leaf dataset. While the dataset only contains 16 vetted families, we extend the dataset by generating missing fossil families. Middle: Overview of the classification architecture and the "ControlNet" prompting approach. Right: Sampling procedure for the triplet loss calculation, within- and cross-domains, taking the furthest positive to bring closer to the anchor and the closest negative to push from the anchor. Bottom right: Illustration of the cycle-consistent approach, where we start from a sample cleared leaf image from the training set and use ControlNet to generate a synthetic fossil. From this synthetic fossil, we then generate a synthetic leaf that is as close as possible to the original sample.

Results of BeIT

We also investigated a transformer-based architecture, BEiT (BERT pre-training of Image Transformers) [43], which belongs to a family of vision transformers that have recently achieved state-of-the-art performance across multiple image recognition benchmarks. Unlike convolutional neural networks (CNNs), transformers model long-range dependencies in images via self-attention, enabling them to capture global contextual relationships that can be critical for fine-grained classification tasks. BEiT, in particular, is considered a leading transformer approach, pretrained with a masked image modeling objective inspired by natural language processing. We tested BEiT to evaluate whether transformer-based models could outperform CNNs for fossil leaf classification. However, in our experiments, BEiT showed slightly lower performance than the ResNet-101 baseline presented in the main text (Table SS1 summarizes the Top-5 F1-score under the different conditions).

Condition	Top 5 Lower Bound	Top 5 Upper Bound	#families
Unsegmented	4.21%	64.3%	142
Segmented	6.21%	65.1%	142
Triplet + Segmented	21.4%	72%	142
Triplet + Segmented + Synthetic Fossils	**75.1%**	**86.4%**	142

Table S1: BEiT classification results under the different conditions studied.

Details for Training

We use a Large Memory IBM Node with 4 V100 GPUS and 1tB of RAM with fast interconnect between RAM and GPU. One GPU was exclusively hosting the SAM model trained for image silhouette detection, while the other three hosted the Control Net running on Stable Diffusion 2.1 and the classification model The Batch size was 64 with a total resolution of 512x512.

Additional Concepts

Find Examples of more concepts that our model finds below, and please Please visit our website https://fel-thomas.github.io/Leaf-Lens/ for an exhaustive list of concepts.

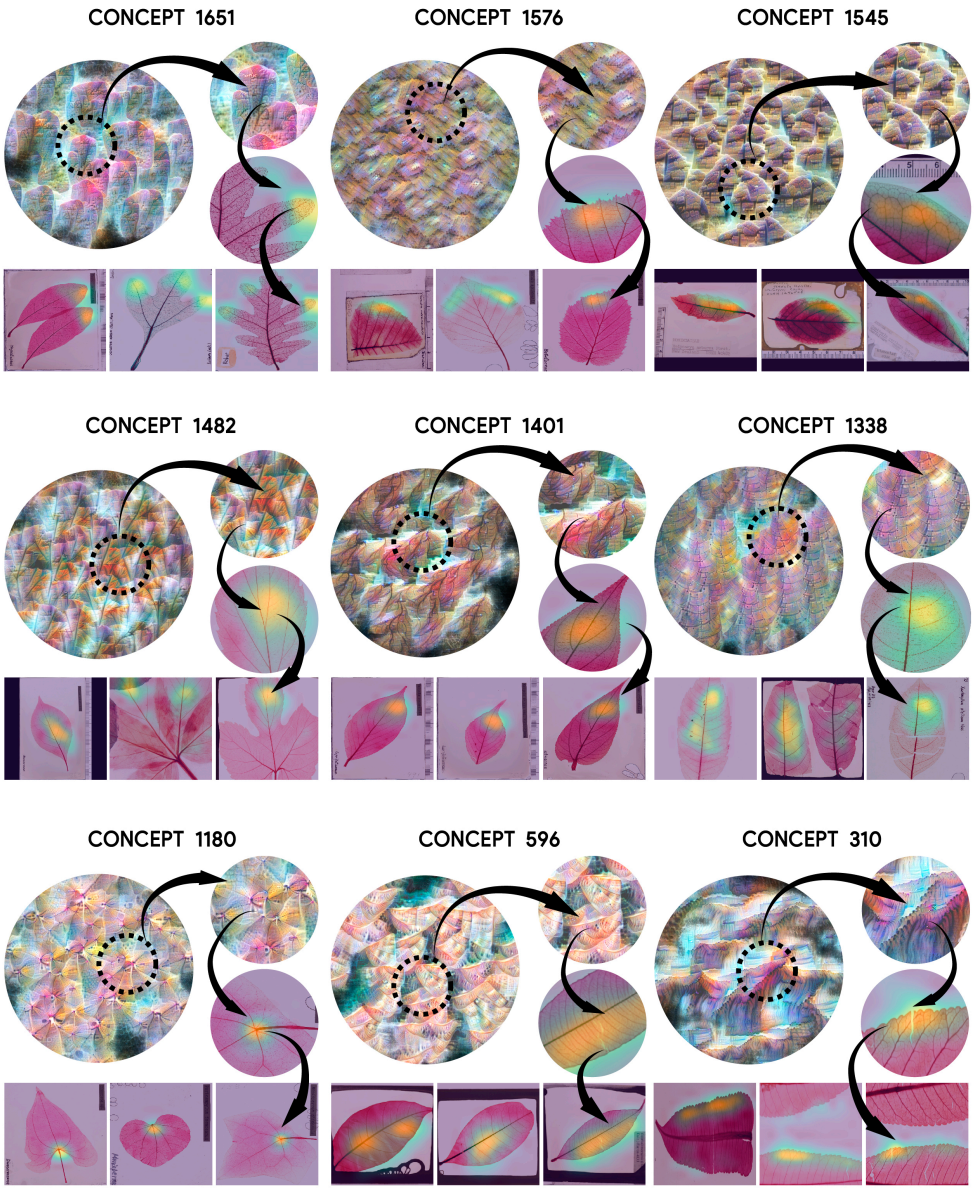

Fig. S3: Examples of more concepts used by the model.

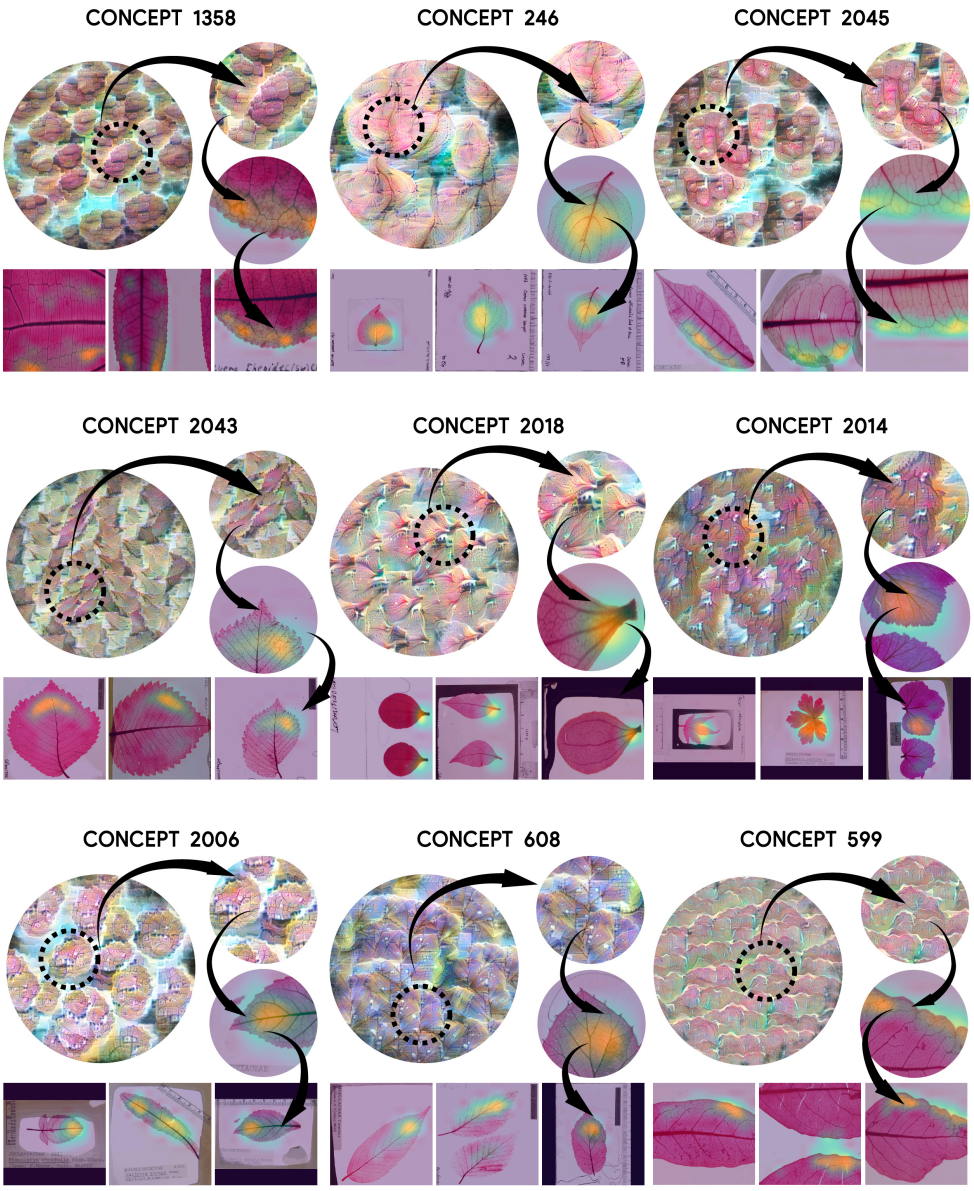

Fig. S4: Additional examples of concepts used by the model.