Received: 18 December 2021 | Accepted: 8 March 2022

DOI: 10.1002/ajb2.1842

RESEARCH ARTICLE

American Journal of Botany

Decoding family-level features for modern and fossil leaves from computer-vision heat maps

Edward J. Spagnuolo[1,2,3] | **Peter Wilf**[1] | **Thomas Serre**[4]

[1]Department of Geosciences and Earth and Environmental Systems Institute, Pennsylvania State University, University Park, Pennsylvania 16802, USA

[2]Millennium Scholars Program, Pennsylvania State University, University Park, Pennsylvania 16802, USA

[3]Schreyer Honors College, Pennsylvania State University, University Park, Pennsylvania 16802, USA

[4]Department of Cognitive, Linguistic and Psychological Sciences, Carney Institute for Brain Science, Brown University, Providence, Rhode Island 02912, USA

Correspondence

Edward J. Spagnuolo, Department of Geosciences and Earth and Environmental Systems Institute, Pennsylvania State University, University Park, Pennsylvania 16802, USA.
Email: spagnuolo@psu.edu

Abstract

Premise: Angiosperm leaves present a classic identification problem due to their morphological complexity. Computer-vision algorithms can identify diagnostic regions in images, and heat map outputs illustrate those regions for identification, providing novel insights through visual feedback. We investigate the potential of analyzing leaf heat maps to reveal novel, human-friendly botanical information with applications for extant- and fossil-leaf identification.

Methods: We developed a manual scoring system for hotspot locations on published computer-vision heat maps of cleared leaves that showed diagnostic regions for family identification. Heat maps of 3114 cleared leaves of 930 genera in 14 angiosperm families were analyzed. The top-5 and top-1 hotspot regions of highest diagnostic value were scored for 21 leaf locations. The resulting data were viewed using box plots and analyzed using cluster and principal component analyses. We manually identified similar features in fossil leaves to informally demonstrate potential fossil applications.

Results: The method successfully mapped machine strategy using standard botanical language, and distinctive patterns emerged for each family. Hotspots were concentrated on secondary veins (Salicaceae, Myrtaceae, Anacardiaceae), tooth apices (Betulaceae, Rosaceae), and on the little-studied margins of untoothed leaves (Rubiaceae, Annonaceae, Ericaceae). Similar features drove the results from multivariate analyses. The results echo many traditional observations, while also showing that most diagnostic leaf features remain undescribed.

Conclusions: Machine-derived heat maps that initially appear to be dominated by noise can be translated into human-interpretable knowledge, highlighting paths forward for botanists and paleobotanists to discover new diagnostic botanical characters.

KEYWORDS

cleared leaves, computer vision, fossil identification, fossil leaves, heat maps, leaf architecture, leaf identification, leaf margin, leaf teeth, leaf venation

Computer vision algorithms categorize complex patterns, often with a capacity far beyond humans (Gouveia et al., 1997; Linsley et al., 2021), and heat maps can be generated from experimental output to visualize diagnostic regions that were not previously noticed (Lapuschkin et al., 2019; Miao et al., 2019; McGrath et al., 2021 [Preprint]). These visualizations are important for interpreting machine-learning results and guiding human users to discover novel information. Leaves contain immense morphological disparity and are widely acknowledged to store unharnessed phylogenetic information (Doyle, 2007; Little et al., 2010; Feild et al., 2011; Seeland et al., 2019); they are the most abundant macroscopic plant organ, both today and in the fossil record (Wilf, 2008). Hickey and Wolfe (1975) surveyed angiosperm leaf architecture variation, but their study preceded the reorganization of the angiosperm phylogeny due to molecular data (Angiosperm Phylogeny Group, 1998, 2016; Doyle, 2007; Leebens-Mack et al., 2019). Despite the significant work that has been done on many groups, most of the more than 400 angiosperm families lack informative leaf architecture characters that could be used for fossil or field identification (Wilf, 2008). Abundant evolutionary dark data is stored in the millions of fossil leaves already in the world's museums that have dubious or

unassigned taxonomy (Dilcher, 1974; Crane et al., 2004; Marshall et al., 2018).

Computer vision has been used extensively for plant identification, although most efforts have focused on the species level; more work on higher taxa would benefit evolutionary interpretations and paleobotanical applications because nearly all fossil species are extinct. Artificial intelligence (AI) can successfully identify species from images of live plants (Kumar et al., 2012; Joly et al., 2016; Tcheng et al., 2016; Rzanny et al., 2019; Champ et al., 2020; Minowa and Nagasaki, 2020; Unger et al., 2020) and herbarium sheets (Belhumeur et al., 2008; Unger et al., 2016; Carranza-Rojas et al., 2017; Little et al., 2020; Romero et al., 2020). Machine learning identification of extant and fossil pollen at the species level has advanced significantly (Punyasena et al., 2012; Tcheng et al., 2016; Romero et al., 2020; White, 2020). Automated species identification of leaf images, in particular, is a well-studied problem in computer vision (Im et al., 1998; Wu et al., 2007; Nam et al., 2008; Park et al., 2008; Caballero and Aranda, 2010; Bama et al., 2011; Hu et al., 2012; Laga et al., 2012; Larese et al., 2012; Mouine et al., 2012; Priya et al., 2012; Charters et al., 2014; Larese et al., 2014a, b; Jamil et al., 2015; Mata-Montero and Carranza-Rojas, 2015, 2016; Zhao et al., 2015; Grinblat et al., 2016; Larese and Granitto, 2016; Carranza-Rojas, Mata-Montero et al., 2018; Wäldchen and Mäder, 2018; Wäldchen et al., 2018; Almeida et al., 2020; Banerjee and Pamula, 2020; Bryson et al., 2020; Pryer et al., 2020; Soltis et al., 2020; Mukherjee et al., 2021; Zhou et al., 2021). However, there have been few efforts to unpack the diagnostic features revealed from AI for the benefit of botanists. Most computer-vision studies on leaves produce black-box results, that is, without visualizations or interpretations of diagnostic regions. Visualizations such as heat maps (Figure 1; Lu et al., 2012; Lee et al., 2015, 2017; Wilf et al., 2016; Champ et al., 2020; Vizcarra et al., 2021) provide botanists with the potential to understand which leaf features are driving identification. Heat maps allow botanists to learn from artificial intelligence, and they provide a novel, but so far apparently unused, pathway to generate potential new taxonomic characters and visual guidance for the identification of extant and fossil leaves.

Field guides and botany courses often emphasize family-level identification as a traditional starting point, and they incorporate leaf-architecture characters to a variable extent. A few guides and systematic works are well known for their use of fine foliar features to recognize plant families (Gentry, 1993; Soepadmo et al., 2000; Keller, 2004; Simpson, 2010; Kubitzki and Bayer, 2013). Flowers and other reproductive organs—the regions that contain the most well-defined taxonomic features (Cronquist, 1981; Rzanny et al., 2019; Seeland et al., 2019)—are ephemeral and often physically inaccessible, which is why vegetative characters are often needed to identify plants out of season (Gentry, 1993). A small but growing number of computer-vision studies have successfully identified extant foliage and other organs at the family level (Wilf et al., 2016; Schuettpelz et al., 2017; Carranza-Rojas, Joly et al., 2018; Seeland et al., 2019). Paleobotany also requires a family-level approach because most fossil angiosperm leaves belong to extinct species and genera from extant families (Wilf, 2008; Wilf et al., 2016).

Wilf et al. (2016) used a machine-learning approach known as sparse coding and trained a support vector machine (SVM) to identify cleared leaves at the family level with 72% overall accuracy (vs. chance accuracy of 5.6%, from 19 families studied using 7597 cleared leaves). The algorithm learned diagnostic features to identify families that have virtually no known leaf-architecture characters with very high accuracy, for example, 90% of Rubiaceae. The algorithm learned entirely from local, small-scale (16×16 pixels, from images rescaled to 1024 pixels in longest dimension) sample crops of the leaf images, providing a wealth of new information about fine leaf features; thus, the method cannot evaluate many of the larger-scale holistic patterns that botanists traditionally use. A heat-mapping algorithm coded the diagnostic significance (classifier weight) for correct computer-vision identification to family of each small image crop directly on the cleared-leaf images. Briefly, the locations and intensities that corresponded to the maximum classifier weights associated with individual features are shown using red saturation (Figure 1). In other words, the redder the heat-map square, the more important the corresponding leaf region was for placing the individual cleared leaf in its correct plant family. Most locations have zero value because only the most representative crops are used by the classifier. The initial study (Wilf et al., 2016) also provided a brief qualitative analysis of the leaf architectural features highlighted in the heat maps.

Here, we present a quantitative analysis of the locations of diagnostic regions for family-level identification, as shown in the heat maps from Wilf et al. (2016). We attempt to decode the machine-learning algorithm's family-level identification of cleared leaves through location-mapping the hottest hotspots. This is, to our knowledge, the first attempt to back-translate and interpret computer-vision heat maps into botanical language. We developed a manual scoring system for the most-diagnostic regions—the most-saturated red hotspot squares (Figure 1)—in the families with large numbers of published heat maps and scored the squares for a set of leaf architectural features (Table 1; following Ellis et al., 2009). This novel method can be used to interpret computer-vision heat maps using ordinary botanical descriptors and to begin the process of converting some computer-vision signals into human-friendly characters. Although the majority of the patterns identified by the machine-learning algorithm probably cannot be extracted and translated into botanical characters, even a handful of new characters obtained from the analysis of heat-map locations could unlock significant new botanical information from angiosperm leaf architecture.

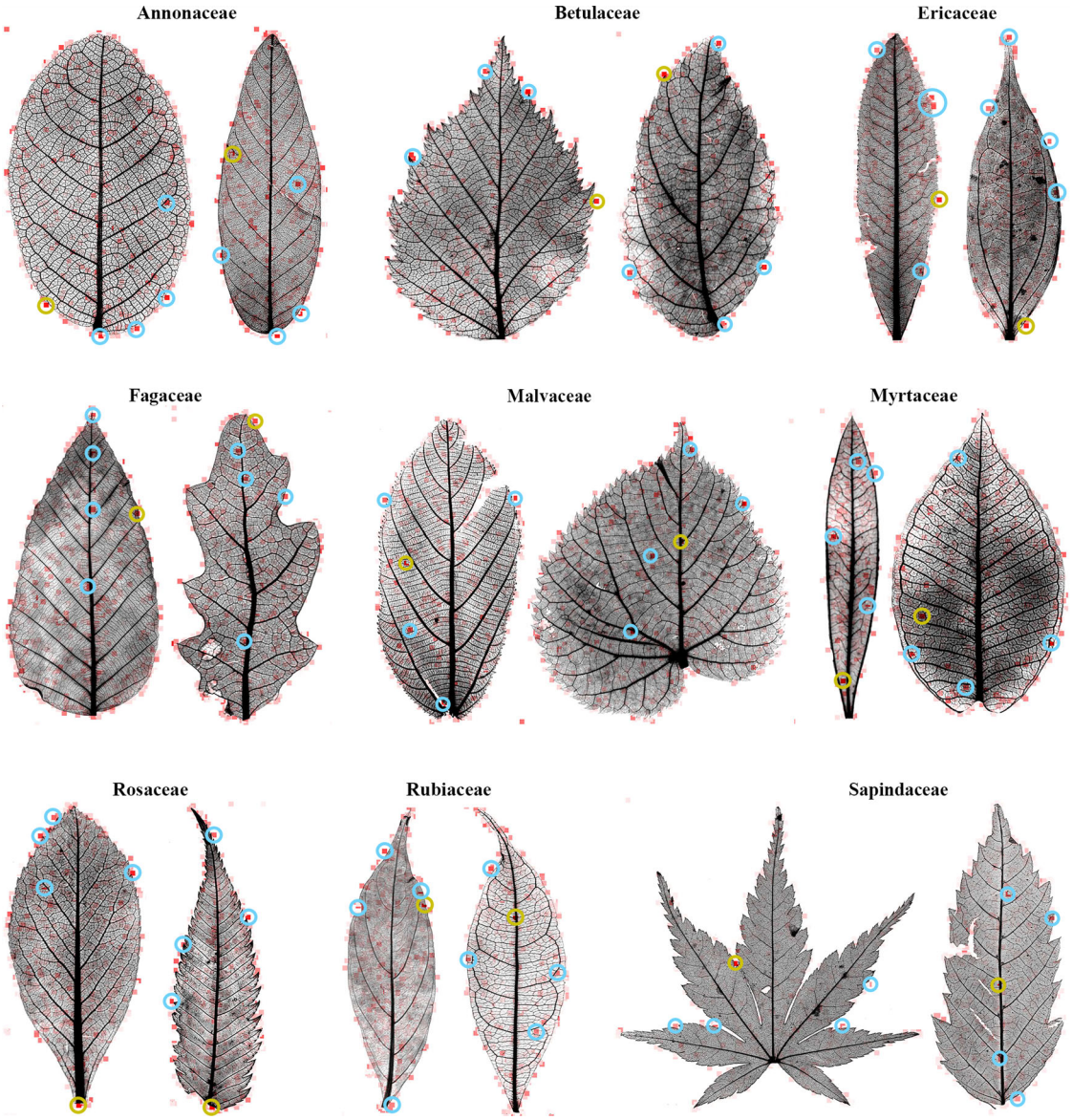

FIGURE 1 Representative heat maps (Wilf et al., 2016) with top-5 squares marked, showing variation in leaf architecture and hotspot locations. Yellow circles, top-1 squares; blue circles, the other four. Top row, left to right: *Fitzalania heteropetala* (NCLC-W catalog no. 14543), *Meiogyne maclurei* (3997), *Betula utilis* (8529), *Alnus trabeculosa* (6718), *Comarostaphylis discolor* (3775), *Psammisia hookeriana* (13044). Middle row, left to right: *Fagus longipetiolata* (1412), *Quercus mohriana* (10721), *Apeiba tibourbou* (1388), *Tilia perneckensis* (16082), *Callistemon citrinus* (12413), *Myrtus lutescens* (10109). Bottom row, left to right: *Crataegus mexicana* (11979), *Sorbaria stellipila* (8806), *Chomelia protracta* (5586), *Faramea anisocalyx* (7375), *Acer sieboldianum* (1220), *Dipteronia sinensis* (1134). For example, using Table 1, the top-1 square in the top-left heat map would receive a score (of 1) both for margin of the basal 25% of the blade and for tertiary veins, with all other features scoring as zero.

MATERIALS AND METHODS

We analyzed the previously published heat maps from the computer-vision experiments of Wilf et al. (2016) (Figure 1). We scored all families with over 50 heat maps available each, totaling 3114 leaves from 14 families and ca. 930 genera. For simplicity, we use "leaves" to refer to both leaves and leaflets. In each heat map, the red intensities of each square represent the diagnostic value of the respective small region of the leaf for correct family placement (Figure 1). All published heat maps used here, placed by Wilf et al. (2016) on Figshare (https://doi.org/10.6084/m9.figshare.1521157.v1), were generated from prepared images of the Jack A. Wolfe contribution to the National Cleared Leaf Collection (NCLC-W), as described by Wilf et al. (2016); NCLC-W is housed in the Division of Paleobotany, Smithsonian National Museum of Natural History, Washington, D.C. Images and voucher data from the collection can be viewed online at the Cleared Leaf Image Database website (www.clearedleavesdb.org; Das et al., 2014; higher-resolution images are available via Wilf et al., 2021).

TABLE 1 Scoring definitions for hotspot squares.[a]

Feature	Definition
In basal 25%	In the first quartile of blade length
Margin of basal 25%	Intersecting basal margin
In midsection 50%	In the second or third quartiles of blade length
Margin of midsection 50%	Intersecting margin of blade midsection
In apical 25%	In the fourth quartile of blade length
Margin of apical 25%	Intersecting blade apex
Margin of lobe	Intersecting margin of leaf lobe
In lobe	In leaf lobe
Primary vein	Intersecting a primary vein; can include tertiaries but not secondaries
Primary–secondary	Intersecting either a primary and a secondary or a primary and an intersecondary vein; veins can be intersecting or separate
Secondary vein	Intersecting any type of secondary vein, including major, minor, intramarginal, and interior secondaries, but not a primary vein
Intersecondary vein	Intersecting an intersecondary vein but not a primary or secondary vein
Tertiary vein	Intersecting tertiary veins but not lower-order veins
Tooth apex	Intersecting the tooth apex
Tooth sinus	Intersecting the tooth sinus
Tooth proximal flank	Intersecting the tooth proximal flank but not the apex
Tooth distal flank	Intersecting the tooth distal flank but not the apex
Mucro	Intersecting a mucronate apex
Petiole insertion	On the petiole insertion point
Damaged area	On a damaged (ripped, torn, folded, contains holes) section of the blade
Off leaf	Not located on the blade

[a]See Materials and Methods section for more details of scoring.

Using Adobe Acrobat Pro DC (continuous release versions; Adobe Inc., San Jose, CA, USA), we manually selected the top-1 and top-5 squares with the highest red intensities for each heat map. We found selection by eye to be more accurate in practice than digital tools such as the Adobe Photoshop eyedropper tool. Although an automated machine ranking and markup could have been generated here from the primary data, we are convinced that our manual markup and the repeated observations involved allowed us to develop a more useful scoring system. The data were scored in two versions: the top-5 squares, manually marked in blue circles, and, of those five, the top-1 square, manually marked in yellow circles (Figure 1).

The 14 families—Anacardiaceae, Annonaceae, Apocynaceae, Betulaceae, Celastraceae, Ericaceae, Fabaceae, Fagaceae, Malvaceae, Myrtaceae, Rosaceae, Rubiaceae, Salicaceae, and Sapindaceae—in nature include ca. 71,000 extant species, or ca. 20% of all angiosperm species, following The Plant List (http://www.theplantlist.org). Wilf et al. (2016) placed the cleared leaf images into their respective, updated families and genera following APG III (Angiosperm Phylogeny Group, 2009) and other standard sources, and a handful of corrections were applied here, namely, the removal of four *Nothofagus* leaves from Fagaceae that had been overlooked. No changes in family placement were warranted here due to subsequent updates to the angiosperm phylogeny (APG IV; Angiosperm Phylogeny Group, 2016). Some of these families have well-studied leaf-fossil records, including Anacardiaceae (e.g., Ramírez et al., 2000; Ramírez and Cevallos-Ferriz, 2002; Sawangchote et al., 2009, 2010), Fagaceae (e.g., Manchester and Crane, 1983; Crepet and Nixon, 1989; Wu et al., 2014; Wilf et al., 2019), Betulaceae (e.g., Crane, 1981; Sun and Stockey, 1992; Pigg et al., 2003; Correa-Narvaez and Manchester, 2021), Malvaceae (e.g., Carvalho et al., 2011; Lebreton Anberrée et al., 2015), Myrtaceae (e.g., MacGinitie, 1969; Manchester et al., 1998; Gandolfo et al., 2011; Tarran et al., 2018), Sapindaceae (e.g., Manchester, 2001;

McClain and Manchester, 2001), Salicaceae (e.g., Manchester et al., 1986, 2006; Boucher et al., 2003), Fabaceae (e.g., Owens et al., 1998; Herendeen and Herrera, 2019), and Rosaceae (e.g., DeVore et al., 2004; DeVore and Pigg, 2007; Kellner et al., 2012). Other families in this study have depauperate leaf-fossil records and, often, poorly understood leaf architecture, including Ericaceae (e.g., Jordan et al., 2010), Apocynaceae (e.g., Del Rio et al., 2020), Annonaceae (e.g., Pirie and Doyle, 2012), Celastraceae (e.g., Bacon et al., 2016), and Rubiaceae (e.g., Roth and Dilcher, 1979; Dilcher and Lott, 2005; Graham, 2009). Many families with poor leaf-fossil records are represented by other organ remains not discussed here (e.g., Friis et al., 2011; Xing et al., 2016).

For each leaf, the top-5 and top-1 square locations were scored using a system we developed (Table 1) based on the definitions of the *Manual of Leaf Architecture* (Ellis et al., 2009). Criteria for the scoring definitions (defined in Table 1) included leaf locations that are unambiguous, likely to be preserved in the fossil record, and rapidly scorable to handle thousands of heat maps in a reasonable amount of time. The 21 scoring definitions, defined in Table 1 and scored as presence–absence, are divided into location categories for the base, apex, or midsection (rest) of the blade; venation features; tooth and other margin features; and noise. The three noise scores report whether the hotspot square is at the petiole insertion, off the leaf, or on a damaged section of the leaf. Due to irregular preservation of petioles in the cleared-leaf collection used, the petioles were previously removed digitally from the cleared leaf images (Wilf et al., 2016), and thus, any signal at the petiole insertion is probably artifactual. Leaf damage includes both natural (insect and fungal damage obliterating parts of leaves) and human (mounting issues, crystallization, and bubbles in mounting medium, breaks) causes. These features do not directly represent leaf architecture and thus were not used in quantitative analyses. We aimed to reduce overlaps in the scores and related ambiguities by increasing the restriction criteria where needed (Table 1). For example, almost any area of most leaves has tertiary veins, sometimes joining lower-order primary or secondary veins within a small selected area and in other cases not. Therefore, we only scored tertiary veins if the hotspot did not also include a primary, secondary, or intersecondary vein. Hotspots with both primary and secondary veins (or primary and intersecondary) were scored as primary–secondary, and hotspots with both secondary and intersecondary veins were scored only as secondary veins. Similarly, hotspots intersecting both a tooth apex and flank were scored for the tooth apex.

For consistency, if a hotspot square was in any way touching the margin of the leaf, its location was scored as on the margin, no matter the percentage of square touching the margin. Lobes and teeth were demarcated with straight lines from sinus to sinus, following the methods of Huff et al. (2003). Basal lobes were demarcated by a perpendicular line across the lobe's primary vein from the lobe's apical sinus, and the lobes of bilobed leaves were demarcated with a line perpendicular to the midvein terminus. The annotated heat maps show marked lobes, when present, and horizontal lines indicate the basal and apical quarters of the leaf. Basal extensions, like those in leaves of many *Bauhinia* spp. (Fabaceae), are not traditionally considered lobes (Ellis et al., 2009) and were not scored as such. We also recorded additional general information, including the percentages of toothed leaves, lobed leaves, and leaves with mucros for each family (Table 2). We note that the red intensity of the hottest heat-map squares varies by family, with some (such as Salicaceae or Betulaceae) having more saturated top-5 squares compared with other families (such as Sapindaceae, Rubiaceae, or Apocynaceae; see Figure 1). However, this pattern seems only to indicate the evenness of the distribution and does not seem to be related to machine-learning accuracy.

The procedure resulted in two presence–absence matrices of scores (i.e., using the terms in Table 1) by specimen for each family, one matrix each for top-1 and top-5 squares, totaling 28 submatrices. We also tabulated the total number of top-5 hotspot squares on teeth, the mean scores for hotspots on toothed vs. untoothed leaves, and the means for lobed vs. unlobed leaves for each family. These matrices and the annotated heat maps can be accessed on Figshare (see Data Availability) at https://10.6084/m9.figshare.17010020. The presence–absence data were analyzed through family-level basic statistics (mean, median) for the top-5 and top-1 matrices (see Data Availability), visualized using box plots, and analyzed using multivariate ordinations and cluster analyses. BoxPlotR software was used to construct the boxplots (http://shiny.chemgrid.org/boxplotr; Spitzer et al., 2014).

For multivariate analyses, principal component (PC), principal coordinate (PCo), and nonmetric multidimensional scaling (NMDS) plots and unweighted pair group method with arithmetic averages (UPGMA) cluster analyses were generated from the median scores for the top-5 matrix and the mean scores for the top-1 matrix for each family (using Euclidean distance measures; other distance measures and linkage strategies gave very similar results). The median values for the top-5 matrix were used to reduce left skewing due to zero values for most scores. To gauge how ordination space varies within families, a separate PCA was conducted for the mean top-5 scores of genera with five or more scored heat maps each. The five heat-map cutoff eliminated many outlier scores of under-sampled genera. Statistical analyses were conducted using Paleontological Statistics Software (PAST; Hammer et al., 2001; available at https://www.nhm.uio.no/english/research/infrastructure/past). Minimal differences were usually observed between PCA, PCoA, and NMDS plots. We present PCA plots here, primarily because the method provides vector biplots through PAST that are easily interpreted. To minimize clutter, we removed generic and leaf-architecture vectors that plotted near the origin from the PCA plots.

For informal demonstrative purposes, we searched manually for possible analogs of the most significant hotspot features in a few fossil leaves of the respective families. No published computer-vision algorithms can

TABLE 2 Summary data by family.

Family	Order	No. of genera[a]	No. of heat maps	%Toothed	%Lobed	%Mucronate	Highest scores
Anacardiaceae	Sapindales	47	101	16.8%	0%	13.9%	Midsection 50%, secondary veins
Annonaceae	Magnoliales	59	164	0%	0%	0%	Margin of basal 25%, margin of midsection 50%, secondary veins, tertiary veins
Apocynaceae	Gentianales	123	206	0%	0%	13.6%	Margin of basal 25%, primary–secondary, secondary veins, intersecondary veins
Betulaceae	Fagales	6	129	100%	0%	15.5%	Margin of apical 25%, secondary veins, tooth apices
Celastraceae	Celastrales	35	121	62%	0%	5%	Midsection 50%, primary–secondary, secondary veins, intersecondary veins
Ericaceae	Ericales	30	161	41%	0%	41.2%	Margin of basal 25%, tooth apices, tertiary veins
Fabaceae	Fabales	260	756	1.5%	2.1%	31.5%	Midsection 50%, margin of apical 25%, margin of basal 25%, secondary veins, tertiary veins
Fagaceae	Fagales	5	135	56.3%	10.4%	0%	Margin of midsection 50%, primary vein, tertiary veins
Malvaceae	Malvales	60	126	56.3%	7.1%	8.7%	Midsection 50%, margin of midsection 50%, secondary veins, tertiary veins, proximal tooth flanks
Myrtaceae	Myrtales	43	77	14.3%	0%	0%	Midsection 50%, in apical 25%, primary–secondary, secondary veins, intersecondary veins
Rosaceae	Rosales	26	187	88.3%	2.1%	9.6%	Margin of apical 25%, secondary veins, tooth apices
Rubiaceae	Gentianales	150	439	0%	0%	0%	Margin of apical 25%, secondary veins
Salicaceae	Malpighiales	24	273	62.6%	0%	0%	Midsection 50%, secondary veins
Sapindaceae	Sapindales	64	239	52.7%	30.1%	2.1%	Margin of midsection 50%, margin of lobe, primary vein, secondary veins, tertiary veins
Totals and among-family means		932	3114	39.4%	3.7%	10.8%	

[a]Genus and species names from Wilf et al. (2016) were retained for compatibility, with very minor adjustments (see Materials and Methods section).

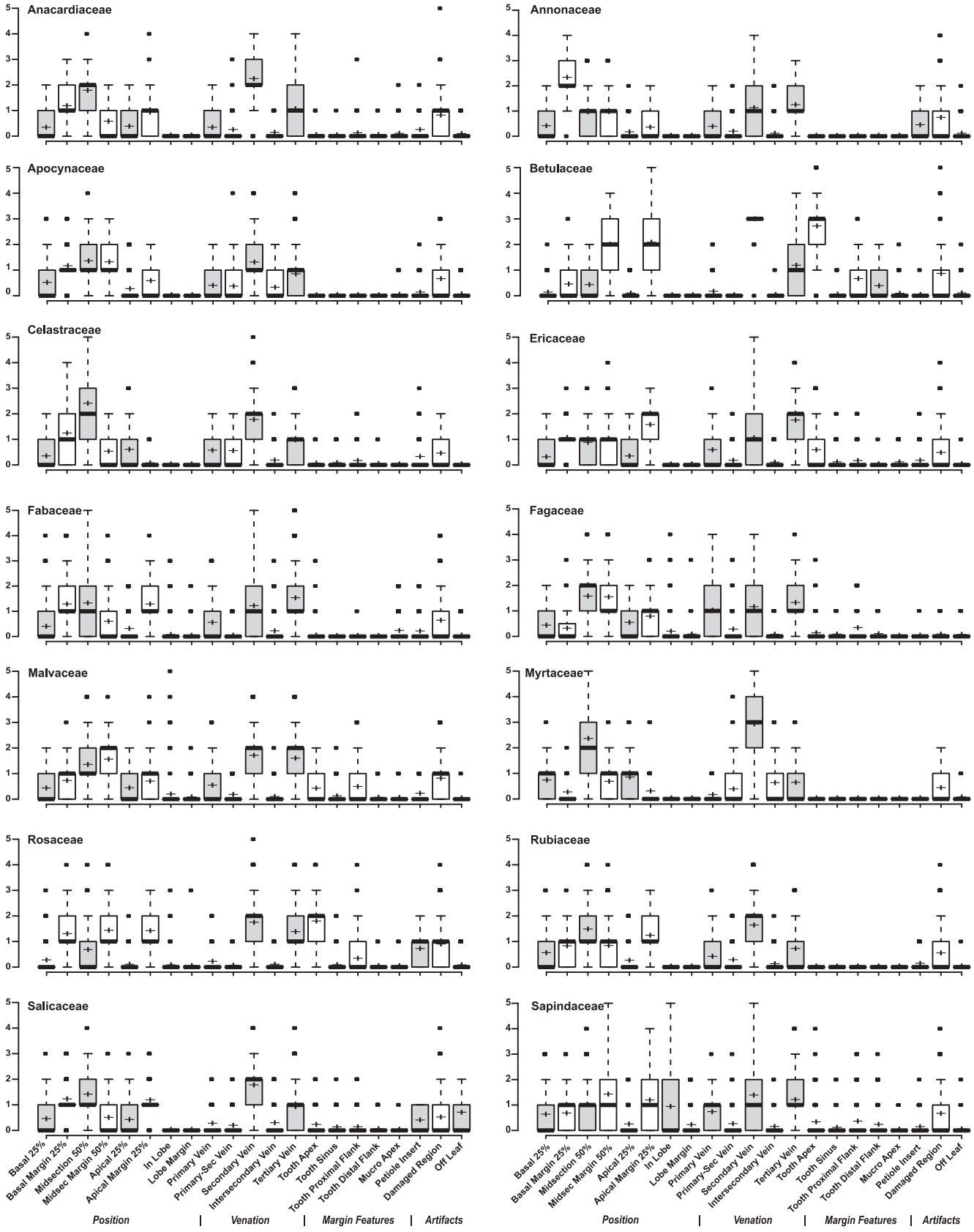

FIGURE 2 Box plots of top-5 scores by family for each of 21 leaf locations (Table 1). Thick bars, medians; box limits, 25th and 75th percentiles; whiskers, 1.5 times the interquartile range; dots, outliers; crosses, means. The sample size per family is five times the number of heat maps (Table 2). Box fills alternate white and gray for visual clarity only; no statistical differences are indicated by the fills. See Data Availability for top-1 box plots.

identify fossil leaves yet, and no computer-vision algorithms were used to find these examples. The examples were found by visually inspecting an open-access image database of vetted fossil leaves identified at the family level (Wilf et al., 2021).

RESULTS

Our analyses found distinctive location signals for leaf hotspots in each family, summarized below by family and illustrated in the box plots (Figure 2; see Data Availability) and selected annotated heat maps (Figure 3; see Data Availability). Univariate and multivariate analyses show similar leaf architecture features as significant; the strongest signals come from locations on apical and basal margins, secondary veins, and tooth apices (Figures 2–5). Comparable locations to those highlighted with the hotspots on the modern leaves can be identified in some fossil representatives from visual observations (Figure 6). Scores are reported below as the within-family means for the top-1 (out of 1.0 possible) or top-5 (out of 5.0 possible) matrices. All summary statistics are archived (see Data Availability).

Anacardiaceae

The highest score for Anacardiaceae was hotspot squares on secondary veins, as seen in both top-5 (mean score of 2.2 out of 5.0; Figures 2 and 3; see Data Availability) and top-1 (mean

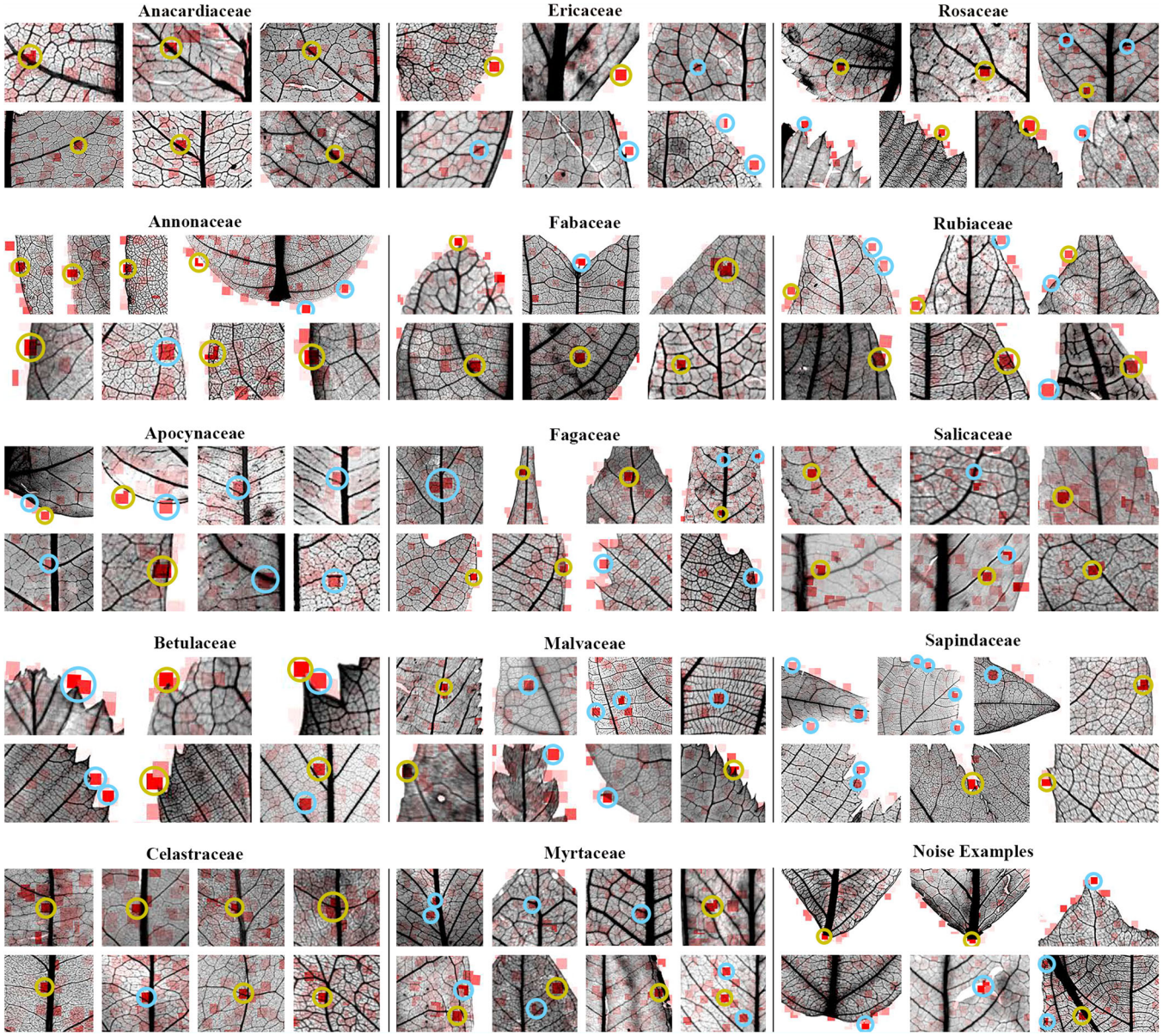

FIGURE 3 (See caption on next page)

score of 0.6 out of 1.0; see Data Availability) squares and exemplified in *Anacardium* and *Buchanania* (see Data Availability). In this family, scores are also high on the blade midsection (i.e., the remaining 50% of the lamina after excluding the basal and apical 25%, see Table 1; top-5 and top-1 squares) and basal 25% margin (top-5 of 1.2). Anacardiaceae are known to have unusual tertiary veins (Wolfe and Wehr, 1987; Martínez-Millán and Cevallos-Ferriz, 2005; Andrés-Hernández and Terrazas, 2009; Mitchell and Daly, 2015); however, the tertiary vein score for Anacardiaceae (top-5 and top-1) is average among the families sampled (see Data Availability). Some tertiary-vein signal is probably present in the hotspot squares that contain both secondary (or primary) and tertiary veins, which in our system are scored for the secondary (primary) vein (Table 1; see Materials and Methods). Only 16.8% of the Anacardiaceae leaves analyzed were toothed (Table 2). All tooth-location scores were low, even when analyzing only toothed leaves (see Data Availability). Across all 14 families, Anacardiaceae has the third-highest scores for both the hotspot squares in the midsection (top-5) and those on the secondary veins (top-5 and top-1), similar to Celastraceae and Myrtaceae for those locations.

Annonaceae

As a completely untoothed and unlobed family, Annonaceae scores are restricted to location and venation (Table 2). All Annonaceae leaves scored seem to have brochidodromous secondary veins. The highest scores within Annonaceae are for the basal margin (top-5 of 2.3; for example, *Cyathostemma*), midsection margin (top-1 of 0.4), and tertiary veins (top-5 of 1.2; top-1 of 0.4; Figure 3). Although below-average in frequency, hotspots on secondary veins are always on secondaries that end in brochidodromous loops or on the loops themselves. Compared with other families, Annonaceae has the third-highest primary vein scores (top-1 of 0.2) and highest tertiary-vein score (top-1 of 0.4).

Apocynaceae

Apocynaceae, another completely untoothed and unlobed family, has its highest scores on the basal margin, secondary veins, and intersecondary veins (Figure 3). The Apocynaceae location scores are for the basal 25% margin (top-5 of 1.1; top-1 of 0.7; see *Baissea*), the midsection margin (top-5 of 1.3), and within the midsection (top-5 of 1.3). The highest score for Apocynaceae venation is for secondary veins (top-5 of 1.3). Compared with other families, Apocynaceae has the third-highest score for primary–secondary intersections (top-5 of 0.3; see *Chilocarpus*) and second-highest score for intersecondary veins (top-5 of 0.3; see *Epigynum*; highest is Myrtaceae; Figures 2 and 3).

Betulaceae

Betulaceae is the only family with 100% toothed and unlobed leaves in the data set; the highest scores for the

FIGURE 3 Selected examples of high-scoring features. **Anacardiaceae**. Secondary veins and secondary–tertiary junctions. Top, left to right: *Anacardium humile* (NCLC-W no. 12854), *Buchanania arborescens* (1758), *Cotinus coggygria* (4306). Bottom, left to right: *Mauria heterophylla* (4219), *Metopium brownei* (4221), *Rhus diversiloba* (12870). **Annonaceae**. Midsection margin, basal margin, brochidodromous secondary loops, tertiary loops: *Malmea depressa* (2885), *Miliusa campanulata* (2453), *Miliusa indica* (7854), *Cyathostemma argenteum* (15483), *Monanthotaxis cauliflora* (5443), *Guatteria ovalifolia* (9517), *Desmopsis microcarpa* (3849), *Monanthotaxis trichocarpa* (5450). **Apocynaceae**. Basal margin, primary–secondary intersection, primary–intersecondary intersection, secondary veins, intersecondary veins: *Heterostemma cuspidatum* (7433), *Baissea axillaris* (5108), *Chilocarpus decipiens* (2034), *Melodinus gracilis* (4824), *Mascarenhasia lisianthiflora* (5118), *Melodinus vitiensis* (6243), *Tabernaemontana hirtula* (10131), *Epigynum miangayi* (8495). **Betulaceae**. Tooth apices, secondary veins: *Alnus oregana* (6710), *Alnus trabeculosa* (6718), *Betula mandshurica* (8521), *Carpinus pubescens* (8497), *Carpinus carpinoides* (8492), *Betula lutea* (11919). **Celastraceae**. Primary vein, primary–secondary intersection, primary–intersecondary intersection, secondary veins: *Celastrus articulatus* (25), *Celastrus articulatus* (13531), *Maytenus tikalensis* (5941), *Pterocelastrus rostratus* (4962), *Cheiloclinium gleasonianum* (8252), *Hippocratea andina* (13608), *Salacia laevigata* (5960), *Schaefferia argentinensis* (6141). **Ericaceae**. Basal margin, teeth, tertiary veins: *Arctostaphylos andersonii* (1454), *Elliottia bracteata* (6888), *Gaultheria miqueliana* (545), *Lyonia lucida* (13034), *Leucothoe axilllaris* (13025), *Vaccinium ciliatum* (13112). **Fabaceae**. Apical margin, mucronate apex, secondary veins, tertiary veins: *Acacia californica* (10636), *Bauhinia divaricata* (30212), *Crudia gabonensis* (13371), *Kunstleria forbesii* (9886), *Kunstleria ridleyi* (9887), *Mimosa glaucescens* (6377). **Fagaceae**. Primary veins, tertiary veins, midsection margin, proximal tooth flanks: *Castanea dentata* (7101), *Castanopsis cuspidata* (190), *Fagus lucida* (8538), *Quercus crassipes* (14728), *Quercus gambelii* (7743), *Quercus hui* (10785), *Quercus libani* (10717), *Quercus donarium* (8549). **Malvaceae**. Secondary veins, minor secondary veins, intercoastal tertiary veins, exterior tertiary veins, tooth apices, tooth proximal flanks: *Corchorus aestuans* (1398), *Pterocymbium tinctorium* (8051), *Microcos paniculata* (11502), *Luehea seemannii* (3609), *Commersonia fraseri* (3662), *Corchorus orinocensis* (3598), *Tilia mongolica* (391), *Tilia noziricola* (8636). **Myrtaceae**. Primary–secondary intersections, primary–intersecondary intersections, secondary veins, intramarginal secondary veins, and tertiary veins: *Eucalyptus sclerophylla* (12430), *Marlierea montana* (3527), *Myrcia affinis* (3521), *Callistemon lanceolatus* (1717), *Calycorectes sellowianus* (3509), *Metrosideros excelsa* (3531), *Myrtus seriocalyx* (3555), *Calyptranthes eugenioides* (3511). **Rosaceae**. Secondary veins, minor secondary veins, tooth apices: *Amelanchier candensis* (1098), *Exochorda racemosa* (1408), *Oemleria cerasiformis* (1008), *Rhodotypos scandens* (12645), *Sorbus japonica* (8671), *Crataegus pubescens* (11981), *Rosa blanda* (12002). **Rubiaceae**. Apical margin and secondary veins in the midsection: *Alibertia nitidula* (10178), *Neobertiera gracilis* (9382), *Tricalysia acocantheroides* (5314), *Chomelia filipes* (5655), *Faramea parvibractea* (13882), *Psychotria longipies* (14056). **Salicaceae**. Secondary vein and midsection: *Abatia stellata* (1702), *Azara dentata* (7953), *Salix acutifolia* (18102), *Salix paradoxa* (18143), *Salix pseudolapponum* (10316), *Samyda yucatanensis* (7030). **Sapindaceae**. Lobes and lobe margin, primary veins, secondary veins, tertiary veins, tooth apex, tooth proximal flank: *Acer* aff. (8604), *Acer caesium* (8580), *Acer barbatum* (480), *Pancovia harmsiana* (4897), *Acer argutum* (8578), *Acer sieboldianum* (1220), *Diploglottis cunninghamii* (7084). **Noise examples**. Petiole insertion, squares off leaf, damaged regions: *Malus toringo* (Rosaceae, 8655), *Prunus americana* (Rosaceae, 7726), *Populus brandegeei* (Salicaceae, 656), *Samyda mexicana* (Salicaceae, 2814), *Albizia saponaria* (Fabaceae, 6366), *Glyphaea grewiodies* (Malvaceae, 4596).

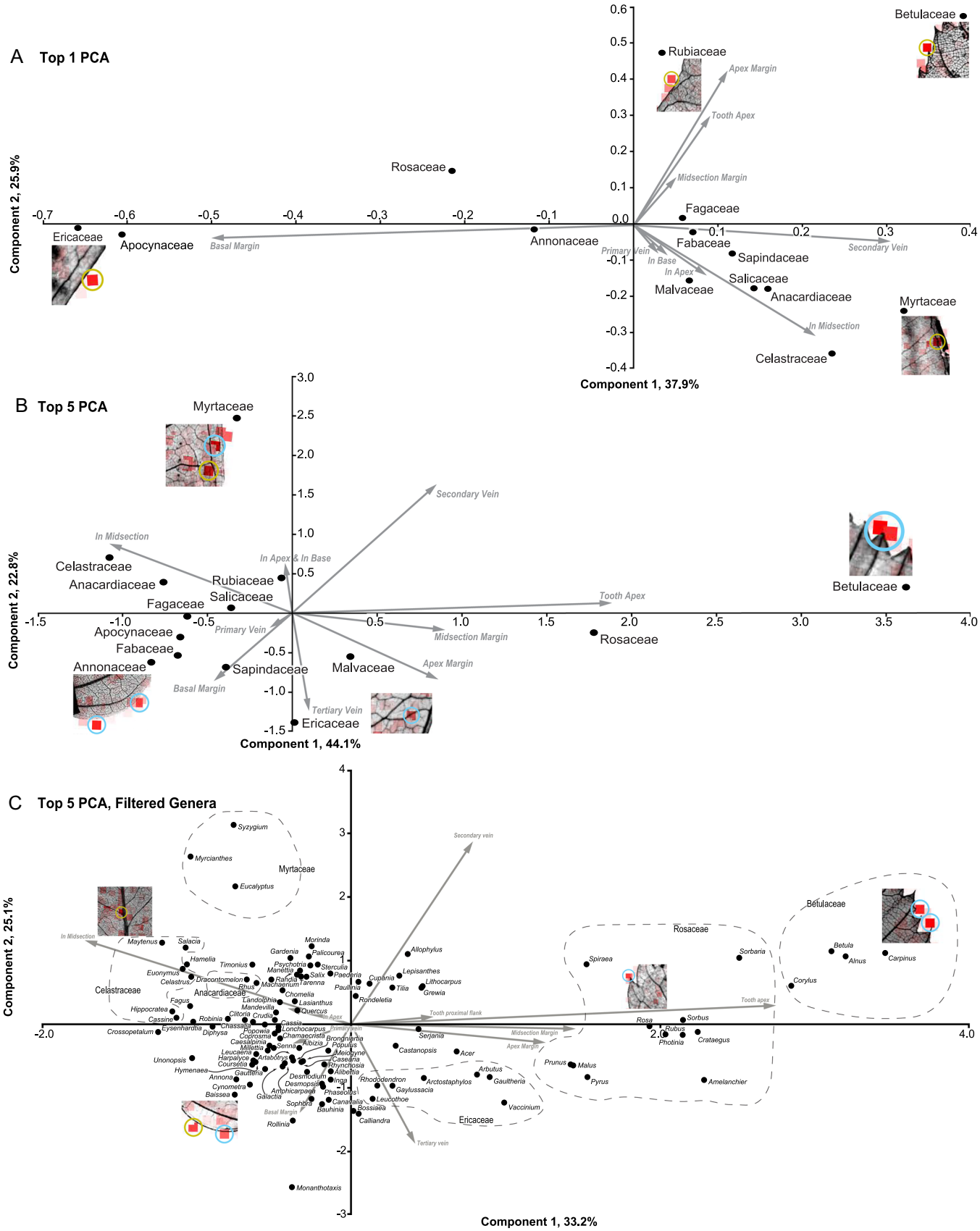

FIGURE 4 (See caption on next page)

family are for leaf margin, secondary veins, and tooth apices (Figure 3). Almost all the hotspot squares are on the leaf margins; the highest location scores for the family are on the apical 25% margin (top-5 of 2.0; top-1 of 0.5), followed by the midsection margin (top-5 of 2.0; top-1 of 0.3). The highest venation scores for Betulaceae are for secondary veins (top-5 of 2.8; top-1 of 0.7; see *Betula* and Figure 3), corresponding to hotspots on both major and minor secondary veins. Betulaceae has very high scores for tooth apices (top-5 of 2.7; top-1 of 0.6; Figures 2 and 3; e.g., *Alnus*), almost always on teeth whose principal veins are secondary or minor secondary veins, rather than tertiary veins (Figure 3). Paleobotanists have used Betulaceae teeth as a distinctive feature when identifying fossil leaves (Hickey and Wolfe, 1975; Wolfe and Wehr, 1987). Betulaceae also has the highest score for all families in the midsection margin (top-5), apical margin 25% (top-5 and top-1), and secondary veins (top-1).

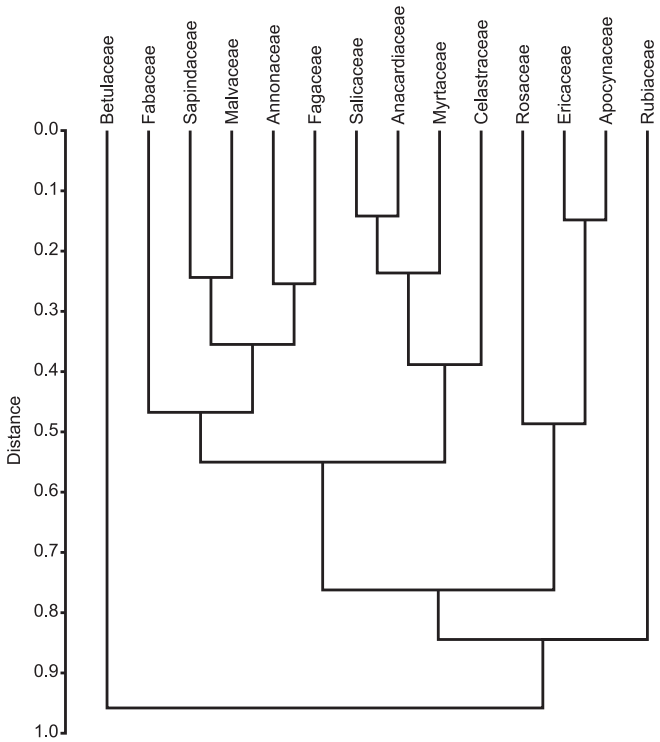

FIGURE 5 UPGMA cluster analysis of the mean top-1 family scores using Euclidean distances (*y*-axis).

Celastraceae

Celastraceae has the highest scores in the midsection, primary–secondary junctions, and secondary veins but has low tooth scores (Figure 3; see Data Availability). The highest location scores within Celastraceae were on the midsection (top-5 of 2.4; top-1 of 0.6) and the basal 25% margin (top-5 of 1.2). Secondary veins generated the highest score for venation (top-5 of 1.7; top-1 of 0.4). Although the Celastraceae image set has one of the highest percentages of toothed leaves (62%; Table 2), all tooth scores are very low, similar to Salicaceae (see Data Availability). Compared with other families, Celastraceae has the highest score for hotspot squares on primary–secondary vein junctions (top-5 of 0.5; top-1 of 0.2; Figure 3). Primary–intersecondary junctions constitute a large portion of the primary–secondary junction score for Celastraceae. However, the intersecondary vein score (top-5 and top-1), representing areas with intersecondaries not at junctions, is low. This result suggests that the junction characteristics (such as angle and gauge; e.g., *Hippocratea*) are more important for identifying Celastraceae compared with the intersecondary or primary veins themselves. Compared with other families, Celastraceae also has the highest hotspot score for the midsection of the blade (top-5 and top-1) and the third-highest score for primary veins (top-5 of 0.5; highest is for Fagaceae and Sapindaceae).

Ericaceae

Ericaceae is a majority untoothed family (41.0% toothed) with teeth small to barely noticeable when present (Table 2). The highest Ericaceae scores are on the basal margin and tooth apices, for toothed leaves (Figure 3). Most Ericaceae hotspot squares were found on the basal 25% margin (top-5 of 1.0; top-1 of 0.8; see *Elliottia*) and apical 25% margin (top-5 of 1.5; Figure 3). The tertiary vein score is high (top-5 of 1.7), along with tooth apices (top-5 of 0.6). Hickey and Wolfe (1975) noted reticulodromous tertiary veins as distinctive in Ericaceae. Top-1 scores have no significant venation or tooth scores. Although most leaves in the family do not have teeth, the toothed leaves contain high frequencies of squares on tooth apices (i.e., *Vaccinium*). Ericaceae has the highest score for the basal 25% margin (top-1) and tertiary veins (top-5) for all families.

FIGURE 4 Principal component analyses (PCA) of top-1 and top-5 results, with vectors shown for influential leaf locations (Table 1) and the percentage of variance represented shown on the respective axis. Selected image patches are included as exemplars. (A) Top-1 analysis for families (means), with vectors longer than 0.08 shown. Exemplars, clockwise from top center: *Tricalysia acocantheroides* (Rubiaceae, NCLC-W no. 5314), *Alnus sieboldiana* (Betulaceae, 980), *Myrtus seriocalyx* (Myrtaceae, 3555), *Elliottia bracteata* (Ericaceae, 6888). (B) Top-5 analysis for families (medians), all vectors retained (some are identical, overlapping, or very short). Exemplars, clockwise from top left: *Calycorectes sellowianus* (Myrtaceae, 3509), *Alnus oregana* (Betulaceae, 6710), *Lyonia lucida* (Ericaceae, 13034), *Cyathostemma argenteum* (Annonaceae, 15483). (C) Top-5 analysis of genera with at least five heat maps each (means), genera less than 0.3 units from origin and vectors shorter than 0.25 units removed. Dashed lines, families with discrete spatial occupation as labeled: Anacardiaceae, Betulaceae, Celastraceae, Ericaceae, Myrtaceae, Rosaceae. Exemplars, left to right: *Maytenus tikalensis* (Celastraceae, 5941), *Baissea axillaris* (Apocynaceae, 5108), *Rosa blanda* (Rosaceae, 12002), *Carpinus carpinoides* (Betulaceae, 8492).

Fabaceae

Fabaceae has high scores on the blade midsection and apical margin along with tertiary veins (Figure 3). Fabaceae is one of the only families with a significant percentage of mucronate apices in the data set, 30.5% (Table 2). The sample only includes a handful of toothed or lobed (mostly bilobed *Bauhinia* spp.) leaves. Hotspot squares are often found on the basal 25% margin (top-5 of 1.3), within the midsection (top-5 of 1.3), and the apical 25% margin (top-5 of 1.2; top-1 of 0.3; Figures 2 and 3; see *Pterocarpus*). For venation, scores for tertiary veins are high (top-5 of 1.5; top-1 of 0.4; Figures 2 and 3). The mucronate apex score is low in this family (top-5 of 0.2) due to the high percentage of leaves lacking mucros, but the feature is probably useful for identifying leaves when it is present. Fabaceae has the third-highest score for the basal 25% margin (top-5) and tertiary veins (top-5).

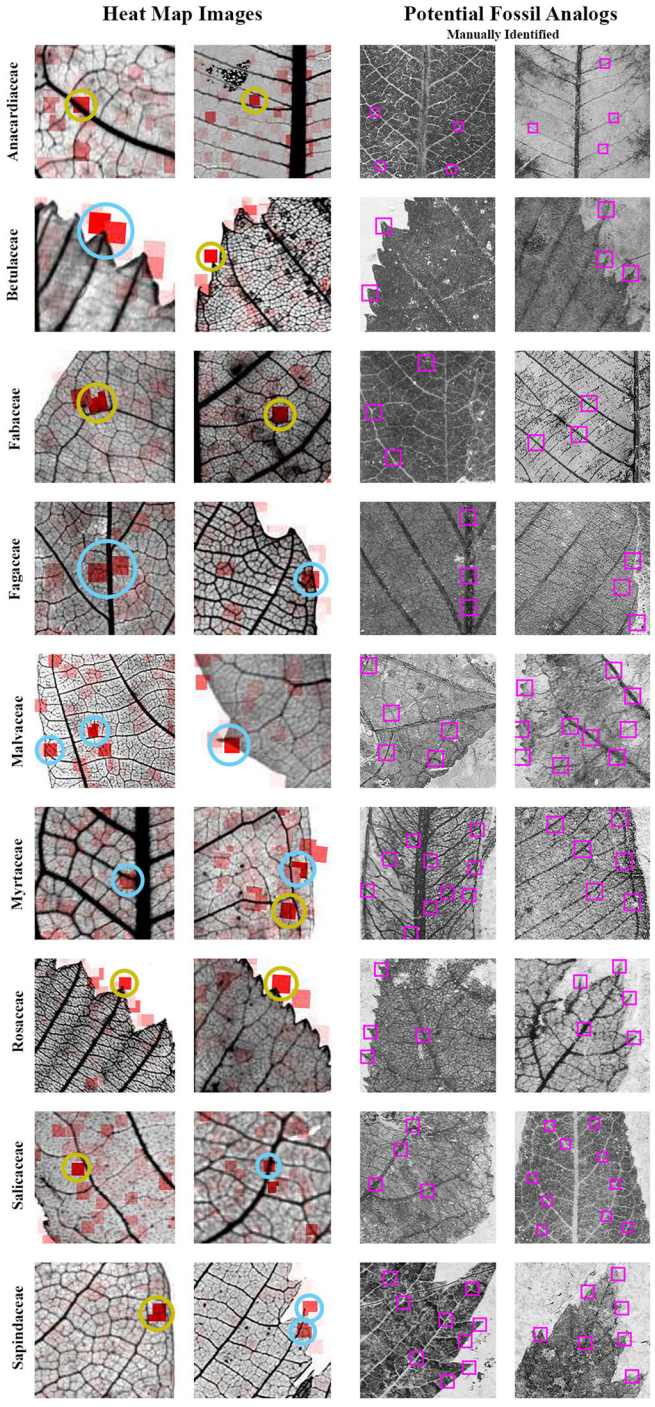

FIGURE 6 (See caption on next page)

Fagaceae

Fagaceae has the highest scores on the midsection margin, primary vein, and tertiary veins (Figure 3). The family has the second-highest percentage of lobed leaves at 10.4%, and 55.6% of the scored cleared leaves are toothed (Table 2). For location, the highest scores for Fagaceae are for hotspot squares in the midsection of the leaf (top-5 of 1.5) and midsection margin (top-5 of 1.5; top-1 of 0.5). For venation, the highest scores are on primary veins (top-5 of 1.0; Figure 3; see *Castanopsis, Fagus*, and *Quercus*) and tertiary veins (top-1 of 0.4; Figure 3). All tooth scores are low because many leaves are untoothed, but the highest tooth score is tooth proximal flanks (top-5 of 0.3; top-5 of 0.5 for only toothed leaves; see Data Availability). Fagaceae has the highest score for the primary veins; however, the primary–secondary junction score is low (Figure 2).

Malvaceae

A family with well-described leaf architecture (Hickey and Wolfe, 1975; Hickey, 1997; Carvalho et al., 2011), approximately half the Malvaceae heat maps are of toothed leaves (Table 2). The highest Malvaceae scores are for squares on the midsection margin (top-5 of 1.5), in the midsection (top-1 of 0.3), on secondary veins (top-5 of 1.7; top-1 of 0.4), on tertiary veins (top-5 of 1.6; top-1 of 0.3), and on proximal tooth flanks (top-5 of 0.5; Figures 2 and 3; e.g., *Tilia*). Hotspot squares are on both secondary and agrophic minor secondary veins with high frequency (Figure 3). Tertiary veins have strong signals on exterior (often tooth principal veins) and intercostal tertiary veins (those veins have a relatively consistent angle and gauge). Although the highest tooth score in Malvaceae is for the tooth proximal flanks (Figure 2), hotspot squares are also on the tooth apex, and the overall tooth score is high in Malvaceae (mean score of 1.0 in top-5 squares and mean score of 1.7 for top-5 squares only on toothed leaves; see Data Availability). Scores are evenly distributed on teeth with secondary and tertiary principal veins. Across families, Malvaceae has the highest score for proximal tooth flanks (top-5 of 0.5; top-5 of 0.8 for only Malvaceae toothed leaves; see Data Availability), and the second-highest score for tertiary veins (top-5 of 1.6; highest is Ericaceae).

Myrtaceae

Myrtaceae leaves are completely untoothed and unlobed (Table 2). High scores in the family are for hotspot squares within the midsection, primary–secondary junctions, secondary veins, and intersecondary veins (Figure 3). The highest Myrtaceae location scores are in the midsection of the blade (top-5 of 2.3; top-1 of 0.4). Although low compared with the midsection scores, the second-highest score is for the apical 25% of the leaf (top-5 of 0.8; top-1 of 0.2). For venation, Myrtaceae has high scores on secondary veins (top-5 of 2.9; top-1 of 0.7; Figure 3), intersecondary veins (top-5 of 0.6; top-1 of 0.1; see *Calyptranthes* and Figure 3), and primary–secondary junctions (top-5 of 0.4; Figure 3). Similar to Celastraceae, many of these are primary–intersecondary junctions (Figure 3). Hotspots are often on thin-gauged secondary and intersecondary veins that join a well-defined intramarginal vein or on the intramarginal vein itself (intramarginal veins are scored as secondary veins; Table 1, Figure 3). The presence of a well-expressed intramarginal vein in many Myrtaceae is well known and has long been used by paleobotanists to help identify fossil myrtaceous leaves (MacGinitie, 1969; Manchester et al., 1998; Gandolfo et al., 2011; Tarran et al., 2018). Compared with other families studied, Myrtaceae has the highest scores for the apical 25% of the blade (top-5), secondary veins (top-5), and intersecondary veins (top-5 and top-1); the second-highest score for the blade midsection (top-5 and top-1; highest is Celastraceae); and the third-highest for primary–secondary intersections (top-5).

FIGURE 6 Potential fossil analogs of selected heat map features. Fossils at right were manually marked, based on visual inspection, with unfilled squares to represent potential regions of similarity to computer-vision hotspot locations on cleared leaves from the same family shown at left. All fossil images are from the open-access image collection of Wilf et al. (2021). **Anacardiaceae**. Hotspots on secondary veins and secondary–tertiary junctions. Left to right: *Astronium graveolens* (NCLC-W no. 8535), *Ozoroa obovata* (10067), Anacardiaceae sp. TY203 (Laguna del Hunco, Chubut, Argentina, Eocene, LH13-0303b, MPEF-Pb), *Rhus malloryi* (Republic Flora, Washington State, Eocene, DMNH 25283). **Betulaceae**. Tooth apices: *Alnus oregana* (6710), *Alnus sieboldiana* (980), *Betula leopoldae* (Republic Flora, DMNH [Stonerose] E155), *Paracarpinus fraterna* (Florissant Fossil Beds, Colorado, Eocene, UCMP 3614). **Fabaceae**. Secondary veins and tertiary veins: *Crudia gabonensis* (13371), *Kunstleria ridleyi* (9887), Fabaceae sp. (Laguna del Hunco, LH13-1173, MPEF-Pb), Fabaceae sp. CJ1 (Cerrejón Coal Mine, Guajíra, Colombia, Paleocene, SGC-ICP-10129). **Fagaceae**. Primary veins, tertiary veins, and midsection margin: *Castanea dentata* (7101), *Quercus donarium* (8549), *Castaneophyllum patagonicum* (Laguna del Hunco, MPEF-Pb 8274), *Fagopsis longifolia* (Florissant Fossil Beds, USNM 332356). **Malvaceae**. Secondary veins, minor secondary veins, intercoastal tertiary veins, exterior tertiary veins, tooth apices, and proximal tooth flanks: *Microcos paniculata* (11502), *Tilia mongolica* (391), *Malvaciphyllum macondicus* (Cerrejón Coal Mine, SGC-ICP 1075), *Tilia johnsoni* (Republic, DMNH 18384). **Myrtaceae**. Primary–secondary intersections, primary–intersecondary intersections, secondary veins, intramarginal secondary veins, intersecondary veins: *Myrcia affinis* (3521), *Calycorectes sellowianus* (3509), *Eucalyptus frenguelliana* (Laguna del Hunco, MPEF-Pb 2329), Myrtaceae sp. TY041 (Laguna del Hunco, MPEF-Pb 976a). **Rosaceae**. Tooth apices, secondary veins, tertiary veins: *Sorbus japonica* (8671), *Crataegus pubescens* (11981), *Prunus gracilis* (Florissant Fossil Beds, UCMP 3644), *Crataegus* sp. (Florissant, FLFO 006827A). **Salicaceae**. Secondary veins and secondary–tertiary junctions: *Abatia stellata* (7021), *Azara dentata* (7953), *Populus wilmattae* (Bonanza site, Green River Formation, Utah, Eocene, DMNH 9763), *Populus crassa* (Florissant Fossil Beds, FLFO 003329A). **Sapindaceae**. Secondary veins, tertiary veins, teeth: *Pancovia harmsiana* (4897), *Acer argutum* (8578), *Koelreuteria allenii* (Florissant Fossil Beds, FLFO 006223B), *Acer florissantii* (Florissant Fossil Beds, UCMP 3831). Repository abbreviations: DMNH, Denver Museum of Nature & Science; FLFO, Florissant Fossil Beds National Monument, Florissant (Colorado); MPEF-Pb, Museo Paleontológico Egidio Feruglio, Trelew (Argentina); SGC-ICP, Colombian Geological Survey and Colombian Petroleum Institute, Bogotá; UCMP, University of California Museum of Paleontology, Berkeley; USNM, National Museum of Natural History, Smithsonian Institution, Washington D.C.

Rosaceae

Rosaceae scores are highest on secondary veins and tooth apices (Figure 3). Rosaceae has the second-highest percentage (highest is Betulaceae) of toothed leaves, 88.3%, and the samples are largely unlobed (Table 2). The hotspots are most often on the margin of the leaf, throughout the margin of the basal 25% (top-5 of 1.3; top-1 of 0.4), the margin of the midsection (top-5 of 1.4), and the margin of the apical 25% (top-5 of 1.4). It is likely that the basal margin 25% score for Rosaceae results from the high score for petiole insertion (top-5 of 0.7; top-1 of 0.2), which is a noise character (see Methods; Table 1; Figure 3). For venation, Rosaceae has high scores for secondary (top-5 of 1.7; top-1 of 0.3) and tertiary veins (top-5 of 1.4; Figure 3), as seen in *Prunus*. Similar to Betulaceae and toothed Ericaceae, Rosaceae also has very high scores for tooth apices (top-5 of 1.8; top-1 of 0.3; Figure 3), notably so in *Crataegus*. However, the machine-learning method was able to discriminate between those families with high accuracy (Wilf et al., 2016), suggesting as-yet-undescribed differences at the family level in tooth-apex morphology. Paleobotanists have used the rosid tooth type as a feature to identify fossil rosaceous leaves (Hickey and Wolfe, 1975; Wolfe and Wehr, 1987; DeVore et al., 2004; Kellner et al., 2012). The majority of the hotspots on tooth apices have secondary or minor secondary principal veins, but there are still some on teeth with tertiary principal veins. Rosaceae scores differ from Betulaceae in the high score of the basal margin and (artifactual) petiole insertion and a higher frequency of hotspots within the leaf interior on secondary and tertiary veins. Compared with other families, Rosaceae has the second-highest scores for the basal 25% margin (top-5; highest is Annonaceae) and tooth apices (top-5 and top-1; highest is Betulaceae) and the third-highest score for the apical 25% margin (top-5).

Rubiaceae

Rubiaceae, a completely untoothed and unlobed family, has high scores for hotspot squares on the apical margin and secondary veins (Figure 3). Rubiaceae species have diagnostic

interpetiolar stipules that have long been used for field identification (Croat, 1978; Gentry, 1993; Simpson, 2010). Unfortunately, the stipules are not preserved in most fossils (but see Roth and Dilcher, 1979), leaving Rubiaceae with a depauperate macrofossil record. The stipules also are not present in the cleared-leaf images used here (or in most or all source slides). The highest hotspot scores in the family are on secondary veins (top-5 of 1.6), within the midsection (top-5 of 1.5), and apical 25% margin (top-5 of 1.2; top-1 of 0.7; see *Tricalysia* and Figures 2 and 3). Compared with other families, Rubiaceae has the highest score for the apical 25% margin (top-1).

Salicaceae

Salicaceae has unexpectedly low scores for tooth characters, despite over 60% of the heat maps being of toothed leaves, no preservation problems observed with the teeth in the images, and the well-known association of the family with the distinctive salicoid tooth type (Figure 2, Table 2; Hickey and Wolfe, 1975; Manchester et al., 1986, 2006; Boucher et al., 2003). The highest location score for Salicaceae is within the blade midsection (top-5 of 1.4; top-1 of 0.4), followed by the basal margin 25% (top-5 of 1.2) and apical margin 25% (top-5 of 1.2). Secondary veins have the highest venation scores for Salicaceae (top-5 of 1.7; top-1 of 0.5; see *Salix* and Figure 3). Frequently, the hotspot squares partially touch the secondary veins or secondary–tertiary junctions (scored as secondary veins; see Materials and Methods, Table 1, Figure 3). Across all families, Salicaceae has the second-highest score for the blade midsection (top-1; highest is Celastraceae).

Sapindaceae

Acer heat maps, comprising more than a third of the Sapindaceae sample, display a different pattern from other Sapindaceae leaves, mostly emphasizing the much higher proportion of lobed leaves in *Acer* compared with other Sapindaceae as well as *Acer* tooth features (Table 3 and Data Availability). In *Acer*, the highest location scores are for hotspot squares on the lobe margin and midsection margin. For non-*Acer* Sapindaceae, the highest leaf location scores are for the midsection, the midsection margin (like *Acer*), and the margin of the apical 25%. For *Acer* venation, primary, secondary, and tertiary vein scores are high, and these are often lobe-forming veins (Figure 3). Only the secondary vein score is high for non-*Acer* Sapindaceae venation. The *Acer* score for tooth proximal flanks is high, and the overall tooth score is more than double that of non-*Acer* taxa; however, the scores are approximately equal for *Acer* and non-*Acer* heat maps for tooth apices. Overall, Sapindaceae (incl. *Acer*) has the highest score on the lobe margin and the second-highest score for the primary veins (Table 3; highest is Fagaceae).

Noise characters

The noise features (hotspots on the digitally clipped petiole, off the leaf, or on a damaged region) did not seem to have a significant impact on the results, attesting to low noise in the system overall as found in the earlier experiments (Wilf et al., 2016). Rosaceae is the only family that has a high score for the petiole insertion (top-5 of 0.7; top-1 of 0.2; Figure 3), and Salicaceae is the only family with a high score for hotspot squares off the leaf (top-5 of 0.7; Figure 3). The score for hotspots on damaged regions of the leaf was low for all families, ranging from 0.05 (Fagaceae) to 0.9 (Rosaceae) for top-5 squares.

Multivariate analyses

The multivariate analyses (Figures 4 and 5) show robust signals from secondary and tertiary veins, several margin features, and tooth apices, generally coinciding with the univariate results just described. Although only a few families sampled here belong in the same order, we note that there is minimal ordinal grouping of those families in the PCA or clusters. However, Anacardiaceae and Sapindaceae (Sapindales) cluster together in the top-1 PCA (not the cluster analysis), most likely due to the high scores for secondary veins in both families. Wilf et al. (2016) found strong identification signals at the ordinal level for cleared leaves, but that work used more families per order than we could examine here.

Top-1 PCA

For the top-1 PCA (Figure 4A), families scoring high on dimension 1 have high scores for secondary veins, as seen in Myrtaceae and Celastraceae, and the secondary vein vector has significant magnitude and almost parallels dimension 1. Families scoring in the negative region of dimension 1 have high scores for the basal margin of the leaf, seen in Ericaceae and Apocynaceae. Families scoring high on dimension 2 have high scores for the apical margin, with positive scores for Betulaceae, Rubiaceae, and Rosaceae. Families scoring in the negative region of dimension 2 have high scores for the blade midsection, as seen in Celastraceae and Myrtaceae. Ericaceae and Apocynaceae plot closely together due to the high frequency of hotspot squares on the basal 25% margin. Most families plot together in the bottom right corner of Figure 4A, that is, with high PC1 and low PC2 scores, due to high scores for secondary veins and the blade midsections. Annonaceae and Rosaceae plot as intermediaries, having high scores for basal margin, midsection margin, and secondary veins. Rubiaceae and Betulaceae are outliers due to their high scores on the apical margin and, for Betulaceae only, tooth apices.

TABLE 3 Selected top-5 means comparisons for *Acer* and non-*Acer* Sapindaceae.[a]

Feature	*Acer*	Non-*Acer*	All Sapindaceae
Midsection 50%	0.7	1.2	1.0
Margin of midsection 50%	2.0	0.9	1.4
Margin of apical 50%	0.8	1.4	1.2
Margin of lobe	2.0	0.02	0.9
Primary vein	0.8	0.6	0.7
Secondary vein	1.2	1.5	1.4
Tertiary vein	1.4	1.0	1.2
Tooth apex	0.4	0.2	0.3
Tooth proximal flank	0.6	0.1	0.3
Total tooth score	1.3	0.6	0.9

[a]*Acer* $N = 107$, non-*Acer* $N = 132$, all Sapindaceae $N = 239$.

Top-5 PCA

For the top-5 PCA (Figure 4B), families scoring high on dimension 1 all have high scores for tooth apices and the midsection and apical 25% margin, such as Betulaceae, Malvaceae, and Rosaceae. The vectors for those features indicate that they are influential on dimension 1. Families scoring low on dimension 1 have high scores for squares within the midsection of the blade, including Myrtaceae and Celastraceae. Families scoring high on dimension 2, such as Myrtaceae, Celastraceae, and Anacardiaceae, have high scores for secondary veins and the blade midsection. Families plotting in the negative region of dimension 2 have high scores for the basal 25% margin and tertiary veins, such as Ericaceae, Fabaceae, and Annonaceae. Most families plot close to the origin, including Fagaceae, Salicaceae, Apocynaceae, and Sapindaceae. Families with very high scores for secondary or tertiary veins are outliers, such as Rosaceae, Betulaceae, Myrtaceae, and Celastraceae. Although the top-1 PCA (Figure 4A) is driven strongly by margin and location vectors, the top-5 PCA (Figure 4B) is driven by margin, tooth, and venation vectors (specifically secondary and tertiary veins, margin, midsection, and tooth apex). In both the top-1 and top-5 PCA, Myrtaceae, Rosaceae, Betulaceae, Ericaceae, and Celastraceae plot near the extremes, but Apocynaceae and Rubiaceae are also extremes in top-1 PCA.

Top-5 PCA for genera

The PCA of top-5 averages at the genus level (Figure 4C) has a similar structure to the corresponding family-level PCA (Figure 4B), and the vectors conserve nearly identical directions. The genera of six of the 14 families—Anacardiaceae, Betulaceae, Celastraceae, Ericaceae, Myrtaceae, and Rosaceae—respectively plot closely together in easily-defined spaces (Figure 4C, dashed outlines). Fabaceae, Annonaceae, Apocynaceae, and Rubiaceae have overlapping and similar ordination space that cannot be easily defined. Other families plot throughout the ordination with no clear pattern, such as Salicaceae, Fagaceae, and Malvaceae.

Cluster analysis

The top-1 cluster dendrogram (Figure 5) follows a similar pattern to the top-1 PCA (Figure 4A), in that Rubiaceae and Betulaceae are outliers and Ericaceae, Apocynaceae, and Rosaceae cluster together. Ericaceae, Apocynaceae, and Rosaceae all have high scores for the basal 25% margin, whereas Betulaceae and Rubiaceae have high apical 25% margin scores. All other families form a pair of clusters. One contains families with high scores for secondary veins (Myrtaceae, Celastraceae, Anacardiaceae, Salicaceae), and the other has high scores for tertiary veins (Fagaceae, Annonaceae, Sapindaceae, Malvaceae).

DISCUSSION

Our results show new possibilities for quantitatively interpreting computer vision signals into human-friendly botanical language by mapping, tabulating, and analyzing the regions of the highest diagnostic value. Although we took a manual approach to develop this pilot study, part of the work involved can be automated, such as selecting regions with the most saturated colors. Our results demonstrate that computer-vision heat maps that may, at first, appear to be all noise in fact provide a new pathway to uncover diagnostic features that were previously unnoticed in the complexity of angiosperm leaf architecture. Although we do not attempt here to define new botanical characters, our work presents new leads for families with few to no established leaf-architectural features and enhances visual learning of leaf architecture (Figure 3). The heat map analyses highlight diagnostic information in several leaf structures, including teeth (Rosaceae, Betulaceae, Ericaceae), marginal features of untoothed leaves (Rubiaceae, Annonaceae, Apocynaceae), and secondary venation (Myrtaceae, Anacardiaceae, Celastraceae, Salicaceae). Some of the highlighted regions appear to correspond to characters used by botanists and paleobotanists or to qualitative observations from the original publication of the heat maps (Wilf et al., 2016). Many others appear to be new observations for the families (such as the apical margin in Rubiaceae). Conversely, other traditional leaf architecture characters (such as the salicoid teeth of Salicaceae; Hickey and Wolfe, 1975) did not correspond to significant signals in our analyses.

For families with few established leaf-architecture characters, such as Celastraceae (Bacon et al., 2016), Rubiaceae (Graham, 2009), Apocynaceae (Del Rio et al., 2020), Annonaceae (Pirie and Doyle, 2012), and Ericaceae (Jordan et al., 2010), the highlighted features (Table 2) provide new leads for identifying their isolated fossil-leaf

representatives. In Celastraceae, features of interest include the primary–secondary and primary–intersecondary junctions, including the relative gauge and angle of junctions. Heat-map signals in Annonaceae include the angle, gauge, and distance from the margin of the secondary and tertiary vein loops. We have also extracted new information from families with well-understood leaf architecture, such as Malvaceae, Salicaceae, and Fagaceae. Malvaceae signals include intercostal tertiary vein gauge and angles, agrophic secondary vein patterns, and tooth proximal flanks. In Salicaceae, features of interest include secondary and tertiary vein gauge and secondary–tertiary junctions and ramifications. There are also robust signals in the Fagaceae primary vein, Fabaceae higher order venation, and tooth apices in Betulaceae, Rosaceae, and Ericaceae.

Distinctive signals are present for leaf margins in most families, in both toothed and untoothed leaves. In many highly toothed families, tooth frequency increases toward the apex of the blade. This observation probably explains the higher frequency of hotspot squares on the apical margin relative to the basal margin in Ericaceae, Betulaceae, and Rosaceae. In Sapindaceae, and to a lesser extent in Malvaceae, hotspots on teeth are not focused on a specific region (such as tooth apices in Rosaceae), producing low mean values across the various tooth scores (Table 3; see Data Availability). The overall combined score for hotspots on teeth, however, is not low for Sapindaceae and Malvaceae, indicating that the whole tooth structure is important for family-level identification (see Data Availability), thus resonating with traditional analyses (e.g., Hickey and Wolfe, 1975). For the untoothed families, we suspect that as-yet not understood marginal microcurvatures of untoothed families in Rubiaceae, Apocynaceae, Ericaceae, and Annonaceae are driving the high frequencies of hotspots on the margins of the blade. The strong signals for leaf margins in untoothed leaves emphasize their poorly understood but clearly significant diagnostic value, which has been generally overlooked compared with the better-understood margins of toothed leaves.

Some of the features identified in this study correspond to qualitative observations noted in the original publication of the heat maps (Wilf et al., 2016; Table 2). The importance of Fagaceae primary veins, Ericaceae teeth, Rosaceae tooth apices, Rubiaceae and Fabaceae apical margins, Annonaceae medial margin, secondary and intersecondary veins in Apocynaceae, and secondary veins in Betulaceae were all noted from holistic examination in the original study (Wilf et al., 2016), and our quantitative scoring affirms those observations. High frequencies of hotspot squares on Salicaceae and Fagaceae tooth flanks, intersecondary veins in Betulaceae, and tertiary veins in Anacardiaceae were also noted qualitatively by Wilf et al. (2016) but did not score highly here. However, the qualitative observations by Wilf et al. (2016) were based on visual inspection of the complete heat maps involving hundreds of sample regions, not through standardized scoring of the filtered hottest spots as done here.

Some leaf-architecture characters that have been used by botanists to identify fossil leaves for decades seem to be echoed in the heat maps, when those features are of similar scale to the small sample squares. The systematic value of tooth and tooth-apex fine architecture is long known (Hickey and Wolfe, 1975). Among families studied here, Betulaceae, Rosaceae, and Malvaceae teeth (Hickey and Wolfe, 1975; Wolfe and Wehr, 1987; DeVore and Pigg, 2007; Carvalho et al., 2011), along with the Myrtaceae intramarginal vein (MacGinitie, 1969; Gandolfo et al., 2011), all have well-known characters. For example, Carvalho et al. (2011) discussed the malvoid tooth type (Hickey and Wolfe, 1975), secondary and tertiary principal veins, and agrophic-vein branching patterns that are diagnostic for Malvaceae, all of which are echoed in the heat maps. On the other hand, the heat mapping cannot respond to some of the holistic leaf architecture characters used to identify fossil Malvaceae leaves, such as actinodromous primary venation (Hickey and Wolfe, 1975; Hickey, 1997; Carvalho et al., 2011) because those features are much larger than the sampling points used in the computer-vision algorithm. In Salicaceae, our results indicate that previously unknown features may have higher diagnostic value than the salicoid tooth type (Hickey and Wolfe, 1975; Boucher et al., 2003), although that tooth feature clearly remains useful for identification. Additionally, families with well-defined ordination space for their genera (Figure 4C)—such as Anacardiaceae, Betulaceae, Rosaceae, Ericaceae, Celastraceae, and Myrtaceae—could be ripe targets for further leaf architecture and computer vision studies. Deep learning algorithms (LeCun et al., 2015; Yosinski et al., 2015; Serre, 2019; Goh et al., 2021) will presumably be responsive to diagnostic regions that are larger than the small sample areas used here, including traditional whole-leaf features. Computer vision interpretability is a new and burgeoning field (Olah et al., 2018; Lapuschkin et al., 2019; Linsley et al., 2021; McGrath et al., 2021 [Preprint]; Voss et al., 2021) that, coupled with the mass digitization of herbaria and fossil plant collections, seems certain to further assist botanists and paleobotanists in the identification of both fossil and extant leaves (Beaman and Cellinese, 2012; Page et al., 2015; Hedrick et al., 2020).

Many hotspot regions that had high scores in our system are similar to those seen in fossil leaves from the respective families, showing the potential for direct applications to the fossil records of the respective families. As seen in Figure 6, most of the features have a high likelihood of preservation in the fossil record. Taken together, our results show that coupling traditional leaf-architecture knowledge with artificial intelligence will lead to improved identification and systematic understanding of modern and fossil leaves.

CONCLUSIONS

Computer vision provides a novel approach for improving understanding of diagnostic features in plant morphology. Here, we show that the interpretation and quantitative analysis of computer-vision heat maps can detect previously unknown leaf-architecture signals that could contribute to the development of new taxonomic characters. This contribution is the first to quantitively back-translate heat-map visualizations to understand and uncover novel leaf architecture signals for family-level leaf identification and, to our knowledge, one of the first to do so for any type of computer heat maps or similar visualizations. Our scoring system yielded distinctive score combinations for each family. Diagnostic regions occurred on, as examples, secondary veins in most families; tooth apices in Rosaceae, Ericaceae, and Betulaceae; tooth flanks and intercostal tertiary veins in Malvaceae; primary–secondary junctions in Celastraceae, Myrtaceae, and Apocynaceae; intersecondary veins in Apocynaceae; and marginal features of untoothed leaves in Rubiaceae, Annonaceae, Fabaceae, Apocynaceae and Ericaceae (Table 2, Figures 2 and 3).

Some of the highlighted features are novel, whereas others, such as the Myrtaceae intramarginal vein and Rosaceae teeth, echo characters that have been used by botanists and paleobotanists for decades. Many, but not all, of the findings quantitatively confirm the initial qualitative observations in the original publication of the heat maps (Wilf et al., 2016) (Table 2). The robust signals from marginal microcurvature in untoothed leaves are a new and promising discovery. Multivariate analyses show high family distinctiveness in diagnostic character combinations. Computer-vision signals from extant leaves have the potential to assist in the identification of millions of unidentified fossil leaves, pending the development of dedicated fossil-leaf applications. Machine-learning visualizations can be combined with traditional leaf architecture to provide the opportunity for botanists to learn from computer vision algorithms, increasing visual integrative learning and uncovering novel botanical characters that have been hiding in plain sight.

AUTHOR CONTRIBUTIONS

E.J.S. and P.W. designed research and the scoring system. E.J.S. scored all cleared-leaf heat-maps, conducted all statistical analyses, and wrote the manuscript with input and revisions from P.W. and T.S.

ACKNOWLEDGMENTS

We thank E. Stiles and M. Patzkowsky for their advice on the statistical analyses, C. N. Nuñez Sanchez and M. D. Feineman for manuscript comments, and Associate Editor A.-L. Decombeix and two anonymous reviewers for constructive feedback. For the preceding study (Wilf et al., 2016) as used here, J. Kissell and A. Young prepared the cleared-leaf images and vetted their taxonomy, and S. Zhang generated the heat maps. We are grateful to the Pennsylvania State University Paleobiology Seminar, Multivariate Analysis in Geosciences, and Geoscholarship courses for fruitful discussions and support from the Pennsylvania State University Millennium Scholars Program, Schreyer Honors College, and Presidential Leadership Academy to E.J.S. We acknowledge funding from the Pennsylvania State University Erickson Discovery Grant (to E.J.S.) and NSF Grants NSF DEB-1556666 and EAR-1925755 (to P.W.) and EAR-1925481 (to T.S.). This research partially fulfilled requirements for a 2022 B.S. in Geobiology with Honors from the Pennsylvania State University for E.J.S.

DATA AVAILABILITY STATEMENT

All marked-up heat maps and data matrices generated for the present paper, along with additional figures and tables, are available open access on Figshare at https://doi.org/10.6084/m9.figshare.17010020. The original heat maps that we scored for this article (Wilf et al., 2016) were previously published on Figshare at (https://doi.org/10.6084/m9.figshare.1521157.v1).

ORCID

Edward J. Spagnuolo http://orcid.org/0000-0001-6720-900X
Peter Wilf http://orcid.org/0000-0001-6813-1937
Thomas Serre http://orcid.org/0000-0003-0846-0039

REFERENCES

Almeida, B. K., M. Garg, M. Kubat, and M. E. Afkhami. 2020. Not that kind of tree: Assessing the potential for decision tree-based plant identification using trait databases. *Applications in Plant Sciences* 8: e11379.

Andrés-Hernández, A. R., and T. Terrazas. 2009. Leaf architecture of *Rhus* s.str. (Anacardiaceae). *Feddes Repertorium* 120: 293–306.

Angiosperm Phylogeny Group. 1998. An ordinal classification for the families of flowering plants. *Annals of the Missouri Botanical Garden* 85: 531–553.

Angiosperm Phylogeny Group. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Botanical Journal of the Linnean Society* 161: 105–121.

Angiosperm Phylogeny Group. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* 181: 1–20.

Bacon, C. D., M. P. Simmons, R. H. Archer, L.-C. Zhao, and J. Andriantiana. 2016. Biogeography of the Malagasy Celastraceae: multiple independent origins followed by widespread dispersal of genera from Madagascar. *Molecular Phylogenetics and Evolution* 94: 365–382.

Bama, B. S., S. M. Valli, S. Raju, and V. A. Kumar. 2011. Context based leaf image retrieval (CBLIR) using shape, color, and texture features. *Indian Journal of Computer Science and Engineering* 2: 202–211.

Banerjee, S., and R. Pamula. 2020. Random Forest boosted CNN: An empirical technique for plant classification. *In* J. K. Mandal and S. Mukhopadhyay [eds.], Proceedings of the Global AI Congress 2019, Advances in intelligent systems and computing, vol. 1112, 251–261. Springer, Singapore. Website: https://doi.org/10.1007/978-981-15-2188-1_20

Beaman, R. S., and N. Cellinese. 2012. Mass digitization of scientific collections: new opportunities to transform the use of biological specimens and underwrite biodiversity science. *ZooKeys* 209: 7–17.

Belhumeur, P. N., D. Chen, S. Feiner, D. W. Jacobs, W. J. Kress, H. Ling, I. Lopez, et al. 2008. Searching the world's herbaria: a system for visual identification of plant species. *In* D. Forsyth, P. Torr, and

A. Zisserman [eds.], Computer Vision—ECCV 2008, Lecture notes in computer science, 116–129. Springer, Berlin, Germany. Website: https://doi.org/10.1007/978-3-540-88693-8_9

Boucher, L. D., S. R. Manchester, and W. S. Judd. 2003. An extinct genus of Salicaceae based on twigs with attached flowers, fruits, and foliage from the Eocene Green River Formation of Utah and Colorado, USA. *American Journal of Botany* 90: 1389–1399.

Bryson, A. E., M. W. Brown, J. Mullins, W. Dong, K. Bahmani, N. Bornowski, C. Chiu, et al. 2020. Composite modeling of leaf shape across shoots discriminates *Vitis* species better than individual leaves. *Applications in Plant Sciences* 8: e11404.

Caballero, C., and M. C. Aranda. 2010. Plant species identification using leaf image retrieval. Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR '10, Xi'an, Shaanxi, China, 327–334. Association for Computing Machinery, NY, NY, USA. Website: https://doi.org/10.1145/1816041.1816089

Carranza-Rojas, J., H. Goëau, P. Bonnet, E. Mata-Montero, and A. Joly. 2017. Going deeper in the automated identification of herbarium specimens. *BMC Evolutionary Biology* 17: 181.

Carranza-Rojas, J., A. Joly, H. Goëau, E. Mata-Montero, and P. Bonnet. 2018. Automated identification of herbarium specimens at different taxonomic levels. *In* A. Joly, S. Vrochidis, K. Karatzas, A. Karppinen, and P. Bonnet [eds.], Multimedia tools and applications for environmental & biodiversity informatics, multimedia systems and applications, 151–167. Springer, Cham, Switzerland. Website: https://doi.org/10.1007/978-3-319-76445-0_9

Carranza-Rojas, J., E. Mata-Montero, and H. Goëau. 2018. Hidden biases in automated image-based plant identification. 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), 1–9. Website: https://doi.org/10.1109/IWOBI.2018.8464187

Carvalho, M. R., F. A. Herrera, C. A. Jaramillo, S. L. Wing, and R. Callejas. 2011. Paleocene Malvaceae from northern South America and their biogeographical implications. *American Journal of Botany* 98: 1337–1355.

Champ, J., A. Mora-Fallas, H. Goëau, E. Mata-Montero, P. Bonnet, and A. Joly. 2020. Instance segmentation for the fine detection of crop and weed plants by precision agricultural robots. *Applications in Plant Sciences* 8: e11373.

Charters, J., Z. Wang, Z. Chi, Ah Chung Tsoi, and D. D. Feng. 2014. EAGLE: A novel descriptor for identifying plant species using leaf lamina vascular features. 2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 1–6. Website: https://doi.org/10.1109/ICMEW.2014.6890557

Correa-Narvaez, J. E., and S. R. Manchester. 2021. Distribution and morphological diversity of *Palaeocarpinus* (Betulaceae) from the Paleogene of the northern hemisphere. Botanical Review. Website: https://doi.org/10.1007/s12229-021-09258-y

Crane, P. R. 1981. Betulaceous leaves and fruits from the British Upper Palaeocene. *Botanical Journal of the Linnean Society* 83: 103–136.

Crane, P. R., P. Herendeen, and E. M. Friis. 2004. Fossils and plant phylogeny. *American Journal of Botany* 91: 1683–1699.

Crepet, W. L., and K. C. Nixon. 1989. Earliest megafossil evidence of Fagaceae: Phylogenetic and biogeographic implications. *American Journal of Botany* 76: 842–855.

Croat, T. B. 1978. Flora of Barro Colorado Island. Stanford University Press, Stanford, CA, USA.

Cronquist, A. 1981. An integrated system of classification of flowering plants. Columbia University Press, NY, NY, USA.

Das, A., A. Bucksch, C. A. Price, and J. S. Weitz. 2014. ClearedLeavesDB: an online database of cleared plant leaf images. *Plant Methods* 10: 8.

Del Rio, C., T.-X. Wang, J. Liu, S.-Q. Liang, R. A. Spicer, F.-X. Wu, Z.-K. Zhou, and T. Su. 2020. *Asclepiadospermum* gen. nov., the earliest fossil record of Asclepiadoideae (Apocynaceae) from the early Eocene of central Qinghai-Tibetan Plateau, and its biogeographic implications. *American Journal of Botany* 107: 126–138.

DeVore, M. L., S. M. Moore, K. B. Pigg, and W. C. Wehr. 2004. Fossil *Neviusia* leaves (Rosaceae: Kerrieae) from the lower-middle Eocene of southern British Columbia. *Rhodora* 106: 197–209.

DeVore, M. L., and K. B. Pigg. 2007. A brief review of the fossil history of the family Rosaceae with a focus on the Eocene Okanogan Highlands of eastern Washington State, USA, and British Columbia, Canada. *Plant Systematics and Evolution* 266: 45–57.

Dilcher, D. L. 1974. Approaches to the identification of angiosperm leaf remains. *Botanical Review* 40: 1–157.

Dilcher, D. L., and T. A. Lott. 2005. A middle Eocene fossil plant assemblage (Powers Clay Pit) from western Tennessee. *Bulletin of the Florida Museum of Natural History* 45: 1–43.

Doyle, J. 2007. Systematic value and evolution of leaf architecture across the angiosperms in light of molecular phylogenetic analyses. *Courier Forschungsinstitut Senckenberg* 258: 21–37.

Ellis, B., D. C. Daly, L. J. Hickey, K. R. Johnson, J. D. Mitchell, P. Wilf, and S. L. Wing. 2009. Manual of leaf architecture. Cornell University Press, Ithaca, NY, USA.

Feild, T. S., T. J. Brodribb, A. Iglesias, D. S. Chatelet, A. Baresch, G. R. Upchurch, Jr., B. Gomez, et al. 2011. Fossil evidence for Cretaceous escalation in angiosperm leaf vein evolution. *Proceedings of the National Academy of Sciences, USA* 108: 8363–8366.

Friis, E. M., P. R. Crane, and K. R. Pedersen. 2011. Early flowers and angiosperm evolution. Cambridge University Press, Cambridge, UK.

Gandolfo, M. A., E. J. Hermsen, M. C. Zamaloa, K. C. Nixon, C. C. González, P. Wilf, N. R. Cúneo, and K. R. Johnson. 2011. Oldest known *Eucalyptus* macrofossils are from South America. *PLoS One* 6: e21084.

Gentry, A. H. 1993. A field guide to the families and genera of woody plants of northwest South America (Colombia, Ecuador, Peru). Conservation International, University of Chicago Press, Chicago, IL, USA.

Goh, G., N. Cammarata, C. Voss, S. Carter, M. Petrov, L. Schubert, A. Radford, and C. Olah. 2021. Multimodal neurons in artificial neural networks. *Distill* 6: e30.

Gouveia, F., V. Filipe, M. Reis, C. Couto, and J. Bulas-Cruz. 1997. Biometry: the characterisation of chestnut-tree leaves using computer vision. ISIE '97 Proceedings of the IEEE International Symposium on Industrial Electronics, 757–760. Website: https://doi.org/10.1109/ISIE.1997.648634

Graham, A. 2009. Fossil record of the Rubiaceae. *Annals of the Missouri Botanical Garden* 96: 90–108.

Grinblat, G. L., L. C. Uzal, M. G. Larese, and P. M. Granitto. 2016. Deep learning for plant identification using vein morphological patterns. *Computers and Electronics in Agriculture* 127: 418–424.

Hammer, Ø., D. A. T. Harper, and P. D. Ryan. 2001. PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* 4: 4.

Hedrick, B. P., J. M. Heberling, E. K. Meineke, K. G. Turner, C. J. Grassa, D. S. Park, J. Kennedy, et al. 2020. Digitization and the future of natural history collections. *BioScience* 70: 243–251.

Herendeen, P. S., and F. Herrera. 2019. Eocene fossil legume leaves referable to the extant genus *Arcoa* (Caesalpinioideae, Leguminosae). *International Journal of Plant Sciences* 180: 220–231.

Hickey, L. J. 1997. Stratigraphy and paleobotany of the Golden Valley Formation (Early Tertiary) of western North Dakota. *Geological Society of America Memoir* 150: 1–183.

Hickey, L. J., and J. A. Wolfe. 1975. The bases of angiosperm phylogeny: vegetative morphology. *Annals of the Missouri Botanical Garden* 62: 538–589.

Hu, R., W. Jia, H. Ling, and D. Huang. 2012. Multiscale distance matrix for fast plant leaf recognition. *IEEE Transactions on Image Processing* 21: 4667–4672.

Huff, P. M., P. Wilf, and E. J. Azumah. 2003. Digital future for paleoclimate estimation from fossil leaves? Preliminary results. *Palaios* 18: 266–274.

Im, C., H. Nishida, and T. L. Kunii. 1998. Recognizing plant species by leaf shapes—a case study of the *Acer* family. Proceedings, Fourteenth International Conference on Pattern Recognition, vol. 2, 1171–1173. Website: https://doi.org/10.1109/ICPR.1998.711904

Jamil, N., N. A. C. Hussin, S. Nordin, and K. Awang. 2015. Automatic plant identification: Is shape the key feature? *Procedia Computer Science* 76: 436–442.

Joly, A., P. Bonnet, H. Goëau, J. Barbe, S. Selmi, J. Champ, S. Dufour-Kowalski, et al. 2016. A look inside the Pl@ntNet experience. *Multimedia Systems* 22: 751–766.

Jordan, G. J., J. M. Bannister, D. C. Mildenhall, R. Zetter, and D. E. Lee. 2010. Fossil Ericaceae from New Zealand: deconstructing the use of fossil evidence in historical biogeography. *American Journal of Botany* 97: 59–70.

Keller, R. 2004. Identification of tropical woody plants in the absence of flowers: a field guide. Birkhäuser Verlag, Basel, Switzerland.

Kellner, A., M. Benner, H. Walther, L. Kunzmann, V. Wissemann, and C. M. Ritz. 2012. Leaf architecture of extant species of *Rosa* L. and the Paleogene species *Rosa lignitum* Heer (Rosaceae). *International Journal of Plant Sciences* 173: 239–250.

Kubitzki, K., and C. Bayer. 2013. Flowering plants. Dicotyledons: Malvales, Capparales and non-betalain Caryophyllales. Springer, Berlin, Germany.

Kumar, N., P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J. V. B. Soares. 2012. Leafsnap: A computer vision system for automatic plant species identification. *In* A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid [eds.], Computer Vision —ECCV 2012, Lecture Notes in Computer Science, vol. 7573, 502–516. Springer, Berlin, Germany. Website: https://doi.org/10.1007/978-3-642-33709-3_36

Laga, H., S. Kurtek, A. Srivastava, M. Golzarian, and S. J. Miklavcic. 2012. A Riemannian elastic metric for shape-based plant leaf classification. 2012 International Conference on Digital Image Computing Techniques and Applications (DICTA), 1–7. Website: https://doi.org/10.1109/DICTA.2012.6411702

Lapuschkin, S., S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications* 10: 1096.

Larese, M. G., A. E. Bayá, R. M. Craviotto, M. R. Arango, C. Gallo, and P. M. Granitto. 2014a. Multiscale recognition of legume varieties based on leaf venation images. *Expert Systems with Applications* 41: 4638–4647.

Larese, M. G., R. M. Craviotto, M. R. Arango, C. Gallo, and P. M. Granitto. 2012. Legume identification by leaf vein images classification. *In* L. Alvarez, M. Mejail, L. Gomez, and J. Jacobo [eds.], Progress in pattern recognition, image analysis, computer vision, and applications. CIARP 2012, Lecture Notes in Computer Science, vol. 7441, 447–454. Springer, Berlin, Germany. Website: https://doi.org/10.1007/978-3-642-33275-3_55

Larese, M. G., and P. M. Granitto. 2016. Finding local leaf vein patterns for legume characterization and classification. *Machine Vision and Applications* 27: 709–720.

Larese, M. G., R. Namías, R. M. Craviotto, M. R. Arango, C. Gallo, and P. M. Granitto. 2014b. Automatic classification of legumes using leaf vein image features. *Pattern Recognition* 47: 158–168.

Lebreton Anberrée, J., S. R. Manchester, J. Huang, S. Li, Y. Wang, and Z.-K. Zhou. 2015. First fossil fruits and leaves of *Burretiodendron* s.l. (Malvaceae s.l.) in Southeast Asia: implications for taxonomy, biogeography, and paleoclimate. *International Journal of Plant Sciences* 176: 682–696.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521: 436–444.

Lee, S. H., C. S. Chan, S. J. Mayo, and P. Remagnino. 2017. How deep learning extracts and learns leaf features for plant classification. *Pattern Recognition* 71: 1–13.

Lee, S. H., C. S. Chan, P. Wilkin, and P. Remagnino. 2015. Deep-plant: plant identification with convolutional neural networks. 2015 IEEE International Conference on Image Processing (ICIP), 452–456. Website: https://doi.org/10.1109/ICIP.2015.7350839

Leebens-Mack, J. H., M. S. Barker, E. J. Carpenter, M. K. Deyholos, M. A. Gitzendanner, S. W. Graham, I. Grosse, et al. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574: 679–685.

Linsley, J. W., D. A. Linsley, J. Lamstein, G. Ryan, K. Shah, N. A. Castello, V. Oza, et al. 2021. Superhuman cell death detection with biomarker-optimized neural networks. Science Advances 7: eabf8142.

Little, D. P., M. Tulig, K. C. Tan, Y. Liu, S. Belongie, C. Kaeser-Chen, F. A. Michelangeli, et al. 2020. An algorithm competition for automatic species identification from herbarium specimens. *Applications in Plant Sciences* 8: e11365.

Little, S. A., S. W. Kembel, and P. Wilf. 2010. Paleotemperature proxies from leaf fossils reinterpreted in light of evolutionary history. *PLoS One* 5: e15161.

Lu, H., W. Jiang, M. Ghiassi, S. Lee, and M. Nitin. 2012. Classification of *Camellia* (Theaceae) species using leaf architecture variations and pattern recognition techniques. *PLoS One* 7: e29704.

MacGinitie, H. D. 1969. The Eocene Green River flora of northwestern Colorado and northeastern Utah. University of California Publications in Geological Sciences, vol. 83, 1–140. University of California Press, Berkeley, CA, USA.

Manchester, S. R. 2001. Leaves and fruits of *Aesculus* (Sapindales) from the Paleocene of North America. *International Journal of Plant Sciences* 162: 985–998.

Manchester, S. R., and P. R. Crane. 1983. 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: 1147–1164.

Manchester, S. R., D. L. Dilcher, and W. D. Tidwell. 1986. Interconnected reproductive and vegetative remains of *Populus* (Salicaceae) from the Middle Eocene Green River Formation, northeastern Utah. *American Journal of Botany* 73: 156–160.

Manchester, S. R., D. L. Dilcher, and S. L. Wing. 1998. Attached leaves and fruits of myrtaceous affinity from the Middle Eocene of Colorado. *Review of Palaeobotany and Palynology* 102: 153–163.

Manchester, S. R., W. S. Judd, and B. Handley. 2006. Foliage and fruits of early poplars (Salicaceae: *Populus*) from the Eocene of Utah, Colorado, and Wyoming. *International Journal of Plant Sciences* 167: 897–908.

Marshall, C. R., S. Finnegan, E. C. Clites, P. A. Holroyd, N. Bonuso, C. Cortez, E. Davis, et al. 2018. Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution. *Biology Letters* 14: 20180431.

Martínez-Millán, M., and S. R. S. Cevallos-Ferriz. 2005. Arquitectura foliar de Anacardiaceae. *Revista Mexicana de Biodiversidad* 76: 137–190.

Mata-Montero, E., and J. Carranza-Rojas. 2015. A texture and curvature bimodal leaf recognition model for identification of Costa Rican plant species. 2015 Latin American Computing Conference (CLEI), 1–12. Website: https://doi.org/10.1109/CLEI.2015.7360026

Mata-Montero, E., and J. Carranza-Rojas. 2016. Automated plant species identification: Challenges and opportunities. *In* F. J. Mata and A. Pont [eds.], ICT for Promoting Human Development and Protecting the Environment. WITFOR 2016, IFIP Advances in Information and Communication Technology, vol. 481, 26–36. Springer, Cham, Switzerland. Website: https://doi.org/10.1007/978-3-319-44447-5_3

McClain, A. M., and S. R. Manchester. 2001. *Dipteronia* (Sapindaceae) from the Tertiary of North America and implications for the phytogeographic history of the Aceroideae. *American Journal of Botany* 88: 1316–1325.

McGrath, T., A. Kapishnikov, N. Tomašev, A. Pearce, D. Hassabis, B. Kim, U. Paquet, and V. Kramnik. 2021. Acquisition of chess knowledge in AlphaZero. *arXiv* 2111.09259. [Preprint].

Miao, Z., K. M. Gaynor, J. Wang, Z. Liu, O. Muellerklein, M. S. Norouzzadeh, A. McInturff, et al. 2019. Insights and approaches using deep learning to classify wildlife. *Scientific Reports* 9: 8137.

Minowa, Y., and Y. Nagasaki. 2020. Convolutional neural network applied to tree species identification based on leaf images. *Journal of Forest Planning* 26: 1–11.

Mitchell, J. D., and D. C. Daly. 2015. A revision of *Spondias* L. (Anacardiaceae) in the Neotropics. *PhytoKeys* 55: 1–92.

Mouine, S., I. Yahiaoui, and A. Verroust-Blondet. 2012. Advanced shape context for plant species identification using leaf image retrieval. Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, ICMR '12, article 49, 1–8. Association for Computing Machinery, NY, NY, USA. Website: https://doi.org/10.1145/2324796.2324853

Mukherjee, G., B. Tudu, and A. Chatterjee. 2021. A convolutional neural network-driven computer vision system toward identification of species and maturity stage of medicinal leaves: case studies with neem, tulsi and kalmegh leaves. *Soft Computing* 25: 14119–14138.

Nam, Y., E. Hwang, and D. Kim. 2008. A similarity-based leaf image retrieval scheme: joining shape and venation features. *Computer Vision and Image Understanding* 110: 245–259.

Olah, C., A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev. 2018. The building blocks of interpretability. *Distill* 3: e10.

Owens, S. A., P. F. Fields, and F. W. Ewers. 1998. Degradation of the upper pulvinus in modern and fossil leaves of *Cercis* (Fabaceae). *American Journal of Botany* 85: 273–284.

Page, L. M., B. J. MacFadden, J. A. Fortes, P. S. Soltis, and G. Riccardi. 2015. Digitization of biodiversity collections reveals biggest data on biodiversity. *BioScience* 65: 841–842.

Park, J., E. Hwang, and Y. Nam. 2008. Utilizing venation features for efficient leaf image retrieval. *Journal of Systems and Software* 81: 71–82.

Pigg, K. B., S. R. Manchester, and W. C. Wehr. 2003. *Corylus*, *Carpinus*, and *Palaeocarpinus* (Betulaceae) from the middle Eocene Klondike Mountain and Allenby Formations of northwestern North America. *International Journal of Plant Sciences* 164: 807–822.

Pirie, M. D., and J. A. Doyle. 2012. Dating clades with fossils and molecules: the case of Annonaceae. *Botanical Journal of the Linnean Society* 169: 84–116.

Priya, C. A., T. Balasaravanan, and A. S. Thanamani. 2012. An efficient leaf recognition algorithm for plant classification using support vector machine. International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012), 428–432. Website: https://doi.org/10.1109/ICPRIME.2012.6208384

Pryer, K. M., C. Tomasi, X. Wang, E. K. Meineke, and M. D. Windham. 2020. Using computer vision on herbarium specimen images to discriminate among closely related horsetails (*Equisetum*). *Applications in Plant Sciences* 8: e11372.

Punyasena, S. W., D. K. Tcheng, C. Wesseln, and P. G. Mueller. 2012. Classifying black and white spruce pollen using layered machine learning. *New Phytologist* 196: 937–944.

Ramírez, J. L., and S. R. S. Cevallos-Ferriz. 2002. A diverse assemblage of Anacardiaceae from Oligocene sediments, Tepexi de Rodriguez, Puebla, Mexico. *American Journal of Botany* 89: 535–545.

Ramírez, J. L., S. R. S. Cevallos-Ferriz, and A. Silva-Pineda. 2000. Reconstruction of the leaves of two new species of *Pseudosmodingium* (Anacardiaceae) from Oligocene strata of Puebla, Mexico. *International Journal of Plant Sciences* 161: 509–519.

Romero, I. C., S. Kong, C. C. Fowlkes, C. Jaramillo, M. A. Urban, F. Oboh-Ikuenobe, C. D'Apolito, and S. W. Punyasena. 2020. Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy. *Proceedings of the National Academy of Sciences, USA* 117: 28496–28505.

Roth, J. L., Jr., and D. L. Dilcher. 1979. Investigations of angiosperms from the Eocene of North America: stipulate leaves of the Rubiaceae including a probable polyploid population. *American Journal of Botany* 66: 1194–1207.

Rzanny, M., P. Mäder, A. Deggelmann, M. Chen, and J. Wäldchen. 2019. Flowers, leaves or both? How to obtain suitable images for automated plant identification. *Plant Methods* 15: 77.

Sawangchote, P., P. J. Grote, and D. L. Dilcher. 2009. Tertiary leaf fossils of *Mangifera* (Anacardiaceae) from Li Basin, Thailand as examples of the utility of leaf marginal venation characters. *American Journal of Botany* 96: 2048–2061.

Sawangchote, P., P. J. Grote, and D. L. Dilcher. 2010. Tertiary leaf fossils of *Semecarpus* (Anacardiaceae) from Li Basin, northern Thailand. *Thai Forest Bulletin (Botany)* 38: 8–22.

Schuettpelz, E., P. B. Frandsen, R. Dikow, A. Brown, S. Orli, M. Peters, A. Metallo, et al. 2017. Applications of deep convolutional neural networks to digitized natural history collections. *Biodiversity Data Journal* 5: e21139.

Seeland, M., M. Rzanny, D. Boho, J. Wäldchen, and P. Mäder. 2019. Image-based classification of plant genus and family for trained and untrained plant species. *BMC Bioinformatics* 20: 4.

Serre, T. 2019. Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science* 5: 399–426.

Simpson, M. G. 2010. Diversity and classification of flowering plants: Eudicots. *In* M. G. Simpson [ed.], Plant systematics, 275–448. Academic Press, San Diego, CA, USA.

Soepadmo, E., S. Julia, and R. Go. 2000. Fagaceae. *In* E. Soepadmo and L. G. Saw [eds.], Tree flora of Sabah and Sarawak, vol. 3, 1–117. Sabah Forestry Department, Forestry Research Institute Malaysia, Kuala Lumpur, Malaysia.

Soltis, P. S., G. Nelson, A. Zare, and E. K. Meineke. 2020. Plants meet machines: prospects in machine learning for plant biology. *Applications in Plant Sciences* 8: e11371.

Spitzer, M., J. Wildenhain, J. Rappsilber, and M. Tyers. 2014. BoxPlotR: a web tool for generation of box plots. *Nature Methods* 11: 121–122.

Sun, F., and R. A. Stockey. 1992. A new species of *Palaeocarpinus* (Betulaceae) based on infructescences, fruits, and associated staminate inflorescences and leaves from the Paleocene of Alberta, Canada. *International Journal of Plant Sciences* 153: 136–146.

Tarran, M., P. G. Wilson, R. Paull, E. Biffin, and R. S. Hill. 2018. Identifying fossil Myrtaceae leaves: the first described fossils of *Syzygium* from Australia. *American Journal of Botany* 105: 1748–1759.

Tcheng, D. K., A. K. Nayak, C. C. Fowlkes, and S. W. Punyasena. 2016. Visual recognition software for binary classification and its application to spruce pollen identification. *PLoS One* 11: e0148879.

Unger, J., D. Merhof, and S. Renner. 2016. Computer vision applied to herbarium specimens of German trees: testing the future utility of the millions of herbarium specimen images for automated identification. *BMC Evolutionary Biology* 16: 248.

Unger, S., M. Rollins, A. Tietz, and H. Dumais. 2020. iNaturalist as an engaging tool for identifying organisms in outdoor activities. *Journal of Biological Education* 55: 537–547.

Vizcarra, G., D. Bermejo, A. Mauricio, R. Z. Gomez, and E. Dianderas. 2021. The Peruvian Amazon forestry dataset: a leaf image classification corpus. *Ecological Informatics* 62: 101268.

Voss, C., N. Cammarata, G. Goh, M. Petrov, L. Schubert, B. Egan, S. K. Lim, and C. Olah. 2021. Visualizing weights. *Distill* 6: e00024.007.

Wäldchen, J., and P. Mäder. 2018. Plant species identification using computer vision techniques: a systematic literature review. *Archives of Computational Methods in Engineering* 25: 507–543.

Wäldchen, J., M. Rzanny, M. Seeland, and P. Mäder. 2018. Automated plant species identification—Trends and future directions. *PloS Computational Biology* 14: e1005993.

White, A. E. 2020. Deep learning in deep time. *Proceedings of the National Academy of Sciences, USA* 117: 29268–29270.

Wilf, P. 2008. Fossil angiosperm leaves: paleobotany's difficult children prove themselves. *Paleontological Society Papers* 14: 319–333.

Wilf, P., K. C. Nixon, M. A. Gandolfo, and N. R. Cúneo. 2019. Eocene Fagaceae from Patagonia and Gondwanan legacy in Asian rainforests. *Science* 364: eaaw5139.

Wilf, P., S. L. Wing, H. W. Meyer, J. Rose, R. Saha, T. Serre, N. R. Cúneo, et al. 2021. An image dataset of cleared, x-rayed, and fossil leaves vetted to plant family for human and machine learning. *PhytoKeys* 187: 93–128.

Wilf, P., S. Zhang, S. Chikkerur, S. A. Little, S. L. Wing, and T. Serre. 2016. Computer vision cracks the leaf code. *Proceedings of the National Academy of Sciences, USA* 113: 3305–3310.

Wolfe, J. A., and W. Wehr. 1987. Middle Eocene dicotyledonous plants from Republic, northeastern Washington. *U.S. Geological Survey Bulletin* 1597: 1–25.

Wu, J.-Y., S.-T. Ding, Q.-J. Li, Z.-R. Zhao, C. Dong, and B.-N. Sun. 2014. A new species of *Castanopsis* (Fagaceae) from the upper Pliocene of west Yunnan, China and its biogeographical implications. *Palaeoworld* 23: 370–382.

Wu, S. G., F. S. Bao, E. Y. Xu, Y. Wang, Y. Chang, and Q. Xiang. 2007. A leaf recognition algorithm for plant classification using probabilistic neural network. 2007 IEEE International Symposium on Signal Processing and Information Technology, 11–16. IEEE. Website: https://doi.org/10.1109/ISSPIT.2007.4458016

Xing, Y., M. A. Gandolfo, R. E. Onstein, D. J. Cantrill, B. F. Jacobs, G. J. Jordan, D. E. Lee, et al. 2016. Testing the biases in the rich Cenozoic angiosperm macrofossil record. *International Journal of Plant Sciences* 177: 371–388.

Yosinski, J., J. Clune, A. Nguyen, T. Fuchs, and H. Lipson. 2015. Understanding neural networks through deep visualization. *In* Deep Learning Workshop, 31st ICML Workshop on Deep Learning, Lille, France, 1–12.

Zhao, C., S. S. F. Chan, W.-K. Cham, and L. M. Chu. 2015. Plant identification using leaf shapes—A pattern counting approach. *Pattern Recognition* 48: 3203–3215.

Zhou, C.-L., L.-M. Ge, Y.-B. Guo, D.-M. Zhou, and Y.-P. Cun. 2021. A comprehensive comparison on current deep learning approaches for plant image classification. *Journal of Physics Conference Series* 1873: 012002.

How to cite this article: Spagnuolo, E. J., P. Wilf, and T. Serre. 2022. Decoding family-level features for modern and fossil leaves from computer-vision heat maps. *American Journal of Botany* 109(5): 768–788. https://doi.org/10.1002/ajb2.1842