THE PENNSYLVANIA STATE UNIVERSITY
MILLENNIUM SCHOLARS PROGRAM

DEPARTMENT OF GEOSCIENCES

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

EDWARD J. SPAGNUOLO

SPRING 2022

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Geobiology
with honors in Geobiology

Reviewed and approved* by the following:

Peter Wilf
Professor of Geosciences
Thesis Supervisor and Honors Advisor

Mark Patzkowsky
Professor of Geosciences
Faculty Reader

* Electronic approvals are on file.

ABSTRACT

Angiosperm leaves present a classic identification problem due to their morphological complexity. Computer-vision heat maps illustrate diagnostic regions for identification, providing novel insights through visual feedback. I investigate the potential of analyzing leaf heat maps to reveal novel, human-friendly botanical information with applications for extant- and fossil-leaf identification. I 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 analyzed using cluster and principal component analyses and visualized using box plots. I manually identified similar features in fossil leaves to informally demonstrate potential fossil applications. The method successfully mapped machine feedback using standard botanical language, and distinctive patterns emerged for each family. Hotspots were concentrated on secondary veins (Salicaceae, Myrtaceae, Anacardiaceae, Rubiaceae, Celastraceae), tooth apices (Betulaceae, Rosaceae), and on the little-studied leaf margins of untoothed leaves (Rubiaceae, Annonaceae, Ericaceae, Apocynaceae, Fabaceae). Results from multivariate analyses were driven by similar leaf features. The results echo many traditional observations, while also showing that most diagnostic leaf features remain undescribed. Heat maps that initially appear to be noise can be translated, and the knowledge obtained can be used offline, highlighting paths forward for botanists and paleobotanists to discover new, family-diagnostic botanical characters.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

This thesis corresponds to a manuscript that is now published in *American Journal of Botany*. The manuscript includes two coauthors: Drs. Peter Wilf (Pennsylvania State University) and Thomas Serre (Brown University). Peter Wilf assisted in designing the scoring system and other methods, funded the project, and provided extensive feedback on drafts. Thomas Serre also provided manuscript draft comments. Both Peter Wilf and Thomas Serre led the initial publication of the heat maps (Wilf et al., 2016).

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

ACKNOWLEDGEMENTS

I am incredibly grateful for all of the support from my friends and family, most notably my parents (Elizabeth Gonzalez-Spagnuolo and Edward Spagnuolo) and grandparents (Eunice G. Spagnuolo and Joseph A. Spagnuolo) and Cassandra N. Nuñez Sanchez, Deja N. Workman, Michael A. Zaidel, Maddison A. Williams and V. Abigail Boehm. I could not be more thankful for the unending support from my advisor, Dr. Peter Wilf, who has supported me since my first year at the Pennsylvania State University, provided me the skills and funding to engage in three independent research projects, and always furthered my paleontological and botanical curiosity. I have had unending support and encouragement from all current and past members of the Penn State Paleobotany Lab including Elena Stiles, Gabriella Rossetto-Harris, Xiaoyu "Elliott" Zou, L. Alejandro Giraldo, and Tengxiang Wang as well as Drs. Mark E. Patzkowsky and Sarah Ivory and all past and current Penn State paleobiology graduate students. I want to thank Dr. Maureen Feineman for being my advisor and her commitment to my success before I was even a Penn State Student, and Drs. Gene Hunt and Camilla Souto for their mentorship and continuing support. I would not be here without the Pennsylvania State University Millennium Scholars Program and its amazing advisors and program directors including Georjanne P. Rosa, Katelyn Jones, R. Adidi Etim-Hunting, Amy Freeman, and Charles E. Fisher, who have financially, academically, and personally supported me since Summer Bridge 2018, as well as all other members of Cohort 6. I am grateful for financial and academic support from the Millennium Scholars Program, College of Earth and Mineral Science, Schreyer Honors College, Presidential Leadership Academy, PSU Office of the Provost, and PSU Geoscience Department.

On a more technical note, discussions with Elena Stiles and Mark E. Patzkowsky on statistical analyses and fruitful discussions in Pennsylvania State University Paleobiology Seminar and Geoscholarship course were invaluable. Thomas Serre, Cassandra N. Nuñez Sanchez, Maureen Feineman, Anne-Laure Decombeix, and two anonymous reviewers provided helpful comments on the thesis or the corresponding submitted manuscript. For the preceding study (Wilf et al., 2016) as used here, Jennifer Kissell and Alysa Young prepared the cleared-leaf images and vetted their taxonomy, and Shengping Zhang generated the heat maps. This project was funded by the Pennsylvania State University Erickson Discovery Grant as well as NSF Grants NSF DEB-1556666 and EAR-1925755 (to Peter Wilf) and EAR-1925481 (to Thomas Serre).

Chapter 1

Introduction

Many species of plants today are struggling due to anthropogenic impacts including climate change, deforestation, overexploitation, invasive species, and many other issues (Brummitt et al., 2021). Approximately one third of all tree species are currently threatened with extinction. While all species have intrinsic value and should be conserved, many of these species are also ecologically, medicinally, economically, or ethnobotanically important (Cámara-Leret et al., 2019; Cámara-Leret & Bascompte, 2021). Conservation biologists are rapidly designing conservation strategies to protect and preserve these species but for many, their ecologies and environmental requirements are poorly understood. Most plant species today live only in a subset of the habitat that they could survive in, bracketed by their evolutionary histories, biotic interactions, and human impacts. This theoretical habitat range is known as a fundamental niche and the region where the species is currently present is known as the realized niche (Kearney & Porter, 2004). Mapping a fundamental niche can be very difficult if the organism's current distribution is heavily restricted by humans, however, fossils can be used to remediate this issue.

Fossils can track the environmental conditions at which a species can survive; especially for plant fossils, where species can persist for tens of millions of years and some genera over 100 million years. Fossils are able to record how plants respond to changes in ecosystems due to climate and other environmental perturbations (Ivory et al., 2016), which can be applied to modern pressures and used to make informed conservation decisions. Unfortunately, this method is mostly restricted to the Cenozoic for most plant species, but the

applications of the plant deep time fossil record, while more abstract at times, are no less important. The fossil record can be used like as a natural evolutionary laboratory (Dietl & Flessa, 2011), similar to the Galapagos Islands or Evolution Canyon in Israel, but with the added fourth dimension of time. Many paleobotanists study how plants respond to extinction by studying past mass extinction events (Harnik et al., 2012) and their recovery period. Following the End-Permian mass extinction event, Looy et al. (1999) discovered a crash of global forests and a domination of herbaceous taxa. This echoes pioneer and disaster taxa today that will thrive in regions following severe disturbance (e.g., Marler and del Moral, 2011). Similar results have been found following the Cretaceous-Paleogene mass extinction in North America (Barclay et al., 2003; Barclay & Johnson, 2004) and New Zealand (Vajda et al., 2001) but without the extinction, modern rainforests would likely not exist as the event led to a massive restructuring of plant communities in the neotropics (Carvalho et al., 2021).

Leaves are the most abundant macroscopic plant organ produced today, and they are the most common plant macrofossil of the last 350 million years. Fossil leaves are oftentimes very difficult to identify. Most leaves lack diagnostic characters for taxonomic identification (Wilf, 2008), which can make linking these fossil applications to the modern almost impossible. Leaf architecture displays immense morphological disparity and complexity (Doyle, 2007; Feild et al., 2011; Hickey & Wolfe, 1975), and are widely acknowledged to contain unharnessed phylogenetic signals (Doyle, 2007; Little et al., 2010; Seeland et al., 2019). 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). The mass digitization of natural history collections and herbaria (Bakker et al., 2020; Beaman & Cellinese, 2012; Bebber et al.,

2010; Belhumeur et al., 2008; Hedrick et al., 2020; Marshall et al., 2018; Page et al., 2015; Soltis et al., 2020) provides botanists access to abundant sources of data for leaf architecture analyses and computer vison applications. Despite the significant work that has been done on many groups, most of the more than 400 angiosperm families lack known leaf architecture characters that could be used for fossil identification (Wilf, 2008).

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 are well-known for their use of fine foliar features to recognize plant families (Gentry, 1993; Keller, 2004; Kubitzki & Bayer, 2013; Simpson, 2010; Soepadmo & Wong, 1995). Flowers and other reproductive organs — the regions that contain the most well-defined taxonomic features (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). 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). Millions of fossil leaves are currently housed in museum collections worldwide with incorrect or no known identification (Dilcher, 1974; Marshall et al., 2018), unlocking this evolutionary dark data can help reconstruct the fossil records for thousands of plant species worldwide. Recent advancements in computer vision technology suggest that it might be possible to learn new taxonomic features stored in leaf architecture that could possibly be used to identify modern and fossil leaves.

Here, I present a quantitative analysis of the locations of diagnostic regions for family-level identification that were found using computer vision in Wilf et al. (2016). I attempt to decode the SIFT algorithm's family-level identification of cleared leaves through location-

mapping the hottest hotspots on the Wilf et al. (2016) heat maps. This is, to my knowledge, the first attempt to back-translate and interpret computer-vision heat-maps into botanical terms. By selecting the most diagnostic regions in the families with large numbers of published heat maps and scoring the squares for strictly defined leaf architecture features (following Ellis et al., 2009), I developed a novel method to interpret any computer vision heat map in ordinary botanical terms and to begin the process of converting some computer vision signals into human-friendly characters. Although the majority of the patterns identified by the SIFT 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 dark data stored within angiosperm leaf architecture.

Chapter 2

Literature Review

Computer vision algorithms categorize complex patterns, often with a capacity far beyond humans (Gouveia et al., 1997), and heat maps can be generated to visualize diagnostic regions that were not previously noticed. These visualizations are important for interpreting computer vision results and guiding human users to discover novel information. 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. There have been few efforts to unpack the diagnostic features revealed from AI for the benefit of botanists.

Computer vision studies have successfully identified species using live plants (Champ et al., 2020; Joly et al., 2016; Kumar et al., 2012; Minowa & Nagasaki, 2020; Rzanny et al., 2019; Tcheng et al., 2016; S. Unger et al., 2020) and herbarium sheets (Belhumeur et al., 2008; Carranza-Rojas et al., 2017; Little et al., 2020; Romero et al., 2020; J. Unger et al., 2016). Machine learning identification of fossil pollen at the species level has advanced significantly (Punyasena et al., 2012; Romero et al., 2020; White, 2020). Automated species identification of leaf images, in particular, is a well-studied problem in computer vision (Almeida et al., 2020; Bama et al., 2011; Banerjee & Pamula, 2020; Bryson et al., 2020; Caballero & Aranda, 2010; Carranza-Rojas, Mata-Montero, et al., 2018; Charters et al., 2014; Grinblat et al., 2016; Hu et al., 2012; Im et al., 1998; Jamil et al., 2015; Laga et al., 2012; Larese, Namías, et al., 2014; Larese, Bayá, et al., 2014; Larese et al., 2012; Larese & Granitto, 2016; Mata-Montero & Carranza-Rojas, 2015; Mouine et al., 2012; Nam et al., 2008; Park et al., 2008; Priya et al., 2012; Pryer et

al., 2020; Soltis et al., 2020; Wäldchen et al., 2018; Wäldchen & Mäder, 2018; S. G. Wu et al., 2007; Zhao et al., 2015). A small but growing number of computer-vision studies have successfully achieved identification of extant leaves at the family level (Carranza-Rojas, Joly, et al., 2018; Schuettpelz et al., 2017; Seeland et al., 2019; Wilf et al., 2016). Most computer-vision studies on leaves produce black box results, i.e., without visualizations or interpretations of diagnostic regions. However, visualizations such as heat maps (Fig. 1; Lu et al., 2012; Lee et al., 2015; Wilf et al., 2016; Lee et al., 2017; Champ et al., 2020; Vizcarra et al., 2021) provide botanists with the potential to understand what leaf features are driving identification. Heat maps allow botanists to learn from artificial intelligence and provide a novel, but so far apparently unused, pathway to generate potential new taxonomic characters and "gestalt" visual guidance for the identification of extant and fossil leaves.

Wilf et al. (2016) and Seeland et al. (2019) reported two computer-vision studies that identified plants at the family level and produced heat-map outputs. I build here on the Wilf et al. (2016) study that learned leaf features using 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 7,597 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 (16x16 pixel, 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 (Fig. 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.

Chapter 3

Experimental Details

Data source

I analyzed the previously published heat maps from the set produced by Wilf et al. (2016; Fig. 1). I scored all families with over 50 heat maps available, totaling 3114 leaves from 14 families and ca. 930 genera. For simplicity, I 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 (Wilf et al., 2016). All published heat maps used here, available 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-Wolfe), as described in Wilf et al. (2016); NCLC-Wolfe is housed in the Division of Paleobotany, Smithsonian National Museum of Natural History, Washington, D.C. Images and metadata 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). The National Cleared Leaf Collection is the largest and most phylogenetically diverse compilation of cleared leaves in the world, totaling ca. 25,000 leaves of mounted and stained cleared-leaf slides. Cleared leaves remove the mesophyll of leaves and expose leaf venation, making them the best extant counterparts to fossil leaves. The two cleared leaf contributions (NCLC-Hickey and NCLC-Wolfe) were moved, curated, and combined from 1992-2013 at the NMNH but were created separately by Drs. Hickey and Wolfe beginning in the 1960s to study the leaf architecture of angiosperm leaf fossils (Wilf et al., 2021) and were used extensively in their inspection of angiosperm leaf architecture (Hickey and Wolfe, 1975).

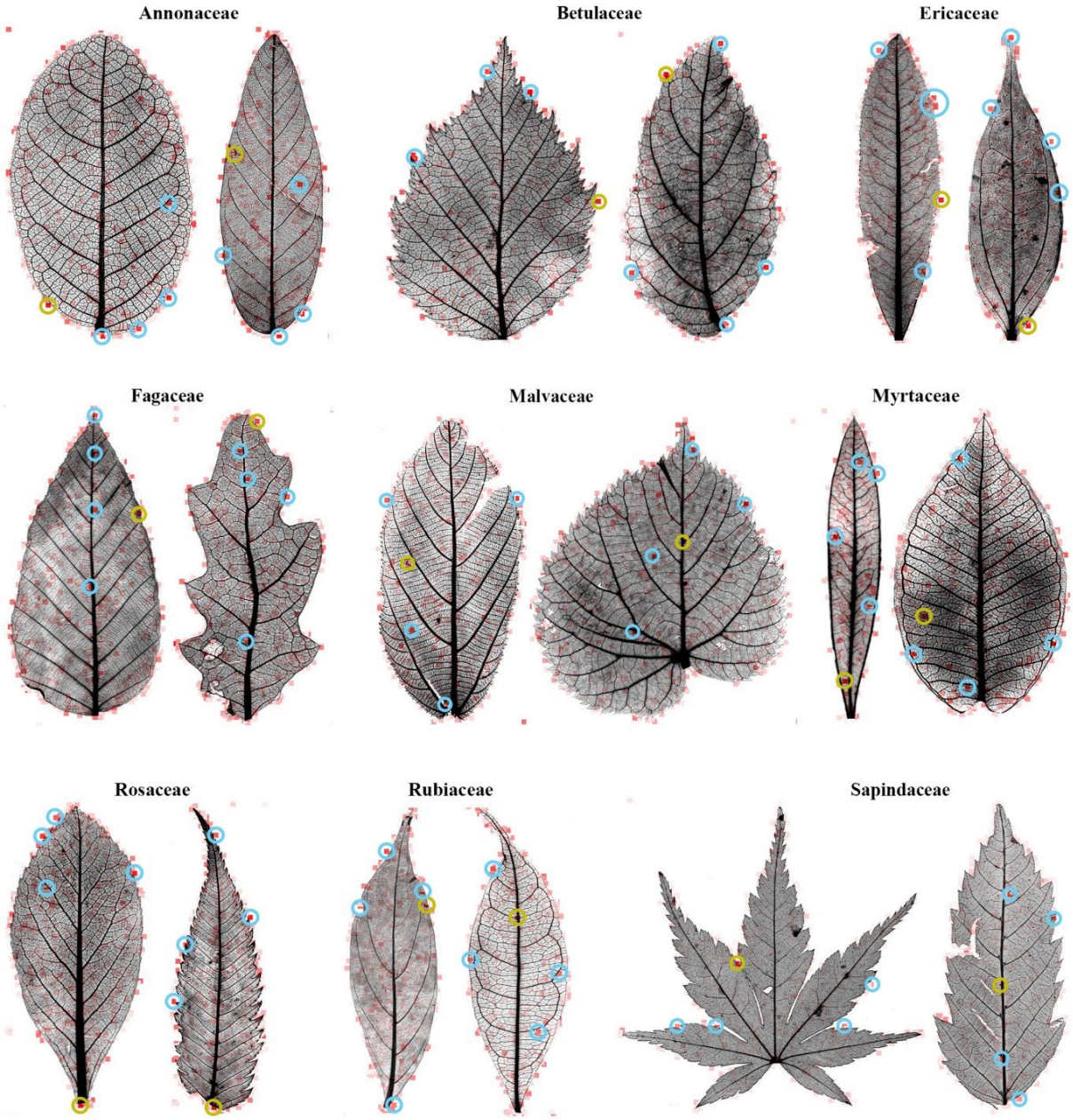

Figure 1. Representative heat maps (Wilf et al. 2016) with top-5 squares marked, showing variation in leaf architecture and hotspot locations. Yellow circles, top-1 squares; blue circles, the other four. Top row, left to right: *Fitzalania heteropetala* (NCLC-W catalog no. 14543), *Meiogyne maclurei* (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.

Scoring system

Using Adobe Acrobat Pro DC (continuous release versions; Adobe Inc., San Jose, California, United States), I manually selected the five squares with the highest red intensities for each cleared-leaf heat map. I found selection by eye to be more accurate in practice than digital tools such as the Adobe eyedropper tool. Although an automated machine ranking and markup could have been generated here from the primary data, the manual markup and the repeated observations involved allowed me 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 (Fig. 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 leaves 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. Some of these families have well-studied leaf-fossil records, including Anacardiaceae (e.g., Ramírez et al., 2000; Ramírez & 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., Herendeen & Herrera, 2019; Lyson et al., 2019; Owens et al., 1998), 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., Taylor et al., 2009; Friis et al., 2011; Xing et al., 2016).

For each leaf, the top-5 and top-1 square locations were scored using a system I developed 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, 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 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 likely 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. I 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, I 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.

Table 1. Scoring definitions for hotspot squares.

Feature[1]	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.

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

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. I also recorded additional, general information, including the percentages of toothed leaves, lobed leaves, and leaves with mucros for each family (Table 2). 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 Fig. 1). However, this pattern seems only to indicate the evenness of the distribution and does not seem to be related to SIFT 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, thus totaling 28 submatrices. The presence-absence data were analyzed through family-level basic statistics (mean, median) for the top-5 and top-1 matrices, visualized using box plots, and analyzed using multivariate ordinations and cluster analyses. BoxPlotR software was used to construct the box plots (http://shiny.chemgrid.org/boxplotr/; Spitzer et al., 2014).

Table 2. Summary data by family.

Family	Order	# Heat maps	%Toothed	%Lobed	%Mucronate	Highest scores
Anacardiaceae	Sapindales	101	16.8%	0%	13.9%	Midsection 50%, secondary veins
Annonaceae	Magnoliales	164	0%	0%	0%	Margin of basal 25%, margin of midsection 50%, secondary veins, tertiary veins
Apocynaceae	Gentianales	206	0%	0%	13.6%	Margin of basal 25%, primary-secondary,

						secondary veins, intersecondary veins
Betulaceae	Fagales	129	100%	0%	15.5%	Margin of apical 25%, secondary veins, tooth apices
Celastraceae	Celastrales	121	62%	0%	5%	Midsection 50%, primary-secondary, secondary veins, intersecondary veins
Ericaceae	Ericales	161	41%	0%	41.2%	Margin of basal 25%, tooth apices, tertiary veins
Fabaceae	Fabales	756	1.5%	2.1%	31.5%	Midsection 50%, margin of apical 25%, margin of basal 25%, secondary veins, tertiary veins
Fagaceae	Fagales	135	56.3%	10.4%	0%	Margin of midsection 50%, primary vein, tertiary veins
Malvaceae	Malvales	126	56.3%	7.1%	8.7%	Midsection 50%, margin of midsection 50%, secondary veins, tertiary veins, proximal tooth flanks
Myrtaceae	Myrtales	77	14.3%	0%	0%	Midsection 50%, in apical 25%, primary-secondary, secondary veins, intersecondary veins
Rosaceae	Rosales	187	88.3%	2.1%	9.6%	Margin of apical 25%, secondary veins, tooth apices
Rubiaceae	Gentianales	439	0%	0%	0%	Margin of apical 25%, secondary veins
Salicaceae	Malpighiales	273	62.6%	0%	0%	Midsection 50%, secondary veins
Sapindaceae	Sapindales	239	52.7%	30.1%	2.1%	Margin of midsection 50%, margin of lobe, primary vein, secondary veins, tertiary veins
Total, among-family means		3114	39.4%	3.7%	10.8%	

Multivariate analyses

Multivariate analyses have been used by ecologists and paleobiologists for decades to understand complex, multivariate data (Foote, 1994, 1995; Kooyman et al., 2014, 2019; Krug & Patzkowsky, 2007; McCune & Grace, 2002; Roy & Foote, 1997; Stiles et al., 2020). Multivariate exploratory analyses are useful when searching for structure in a matrix, and for validating qualitative results. Cluster analysis, while most simplistic, results in a dendrogram that shows the

similarity between samples, in this case families, with more similar samples forming clusters close together compared to less similar samples. A distance measurement is applied to the data matrix, producing a sample-by-sample matrix with larger numbers representing higher dissimilarity between samples. Common distance measurements in paleobiological analyses are either Pythagorean, measuring the hypotenuse between two samples (e.g., Euclidean), or city-block, measuring the legs of the triangle (e.g., Bray-Curtis; McCune and Grace, 2002). Samples that are most similar to each other are then combined (linkage strategies also vary; McCune and Grace, 2002), which will eventually produce a cluster dendrogram. Ordination analyses, specifically principal component and principal coordinate (PCA and PCoA, respectively) analyses produce two dimensional plots, unlike the one-dimensional cluster analyses. These analyses use eigenanalysis matrix algebra, producing as many components, or coordinates, as there are samples, that together explain all the variance and relationships in a matrix. Usually, the first two components are selected as x-y axes because they will contain the largest fraction of the variance. Eigenvector loadings are often plotted as vectors in the ordination as well to show the weightings and importance of a variable to an axis and its relationship with the samples (McCune & Grace, 2002).

Principal component (PCA), principal coordinate (PCoA), and nonmetric multidimensional scaling (NMDS) plots, as well as unweighted pair group method with arithmetic averages (UPGMA) and 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. A separate PCA was conducted for the mean top-5 scores of genera with five or more

scored specimens each, to examine variation within families at the genus level. 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. I present PCA plots here, primarily because the method provides vector biplots through PAST that are easily interpreted. To minimize clutter, genera and leaf-architecture vectors that plotted near the origin were removed from PCA plots.

Fossil applications

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

Chapter 4

Results

My analyses found distinctive location signals for leaf hotspots in each family, summarized below by family and illustrated in the box plots (Fig. 2, see Data Availability) and selected annotated heat maps (Fig. 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 (Figs. 2-5). Comparable locations to those highlighted with the hotspots on the modern leaves can be identified in some fossil representatives from visual observations (Fig. 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 and top-1 box plots are archived (see Data Availability).

Anacardiaceae

The highest score for Anacardiaceae is hotspot squares on secondary veins, as seen in both top-5 (mean score of 2.2 out of 5.0; Figs. 2, 3; see Data Availability) and top-1 (mean 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 (Andrés-Hernández & Terrazas, 2009; Martínez-Millán & Cevallos-Ferriz, 2005; Mitchell & Daly, 2015; Wolfe & Wehr, 1987); however, the tertiary vein score for Anacardiaceae (top-5 and top-1) is average among the families sampled (Fig. 2, see Data

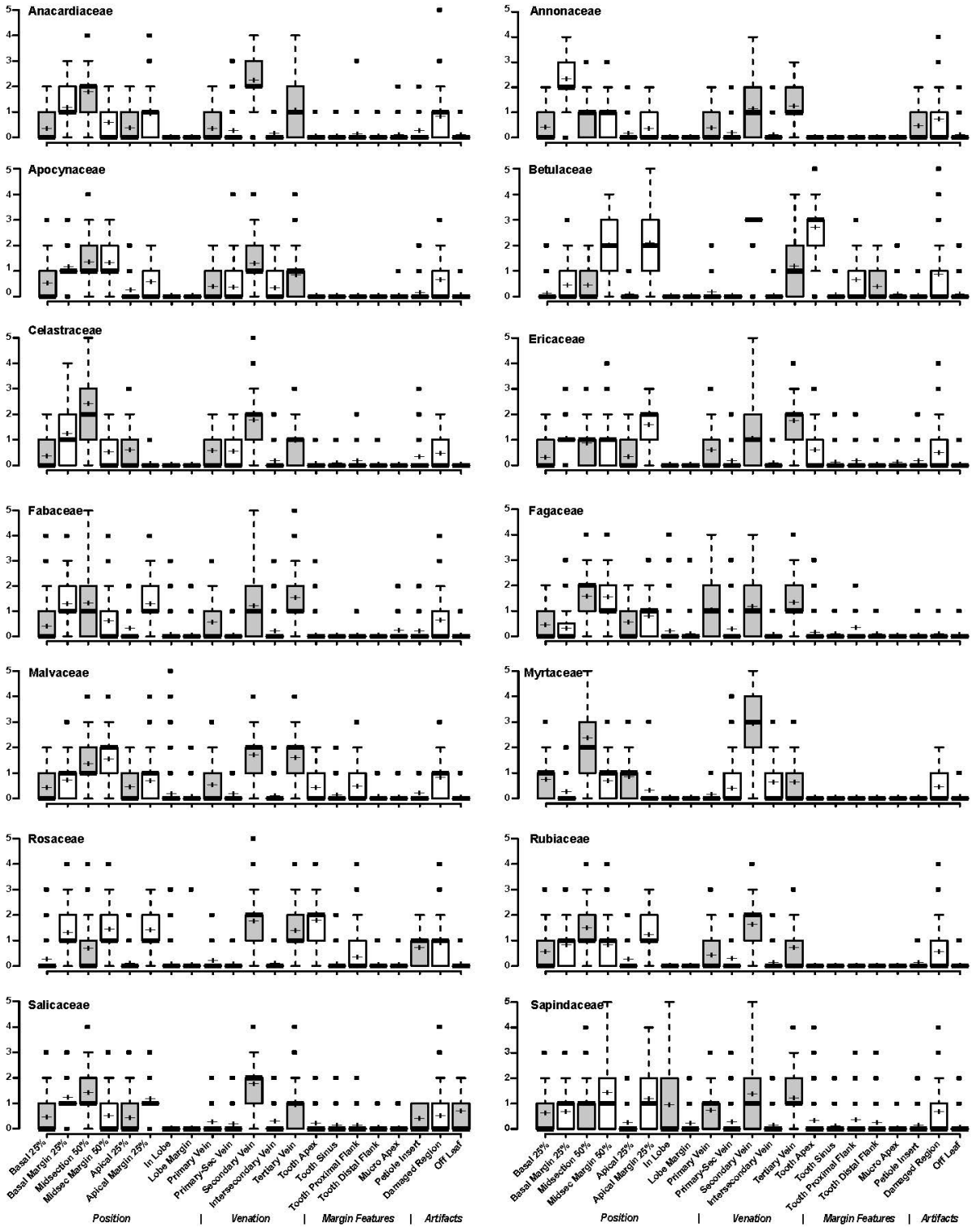

Figure 2. Box plots of top-5 scores by family for each of 21 leaf locations (Table 1). Thick bars, medians; box limits, 25th and 75th percentiles; whiskers, 1.5 times the interquartile range; dots, outliers; crosses, means. The sample size per family is five times the number of heat maps (Table 2). Box 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.

Availability). Some tertiary-vein signal is probably present in the hotspot squares that contain both secondary (or primary) and tertiary veins, which in the system are scored for the secondary (primary) vein (Table 1; see 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; Fig. 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 (Fig. 3). The Apocynaceae location scores are for the basal 25% margin (top-5 of 1.1; top-1 of 0.7; see *Baissea*),

themidsection 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; Figs. 2, 3).

Betulaceae

Betulaceae is the only family with 100% toothed and unlobed leaves in the dataset; the highest scores for the family are for leaf margin, secondary veins, and tooth apices (Fig. 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 Fig. 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; Figs. 2, 3; e.g., *Alnus*; see Data Availability), almost always on teeth whose principal veins are secondary or minor secondary veins, rather than tertiary veins (Fig. 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).

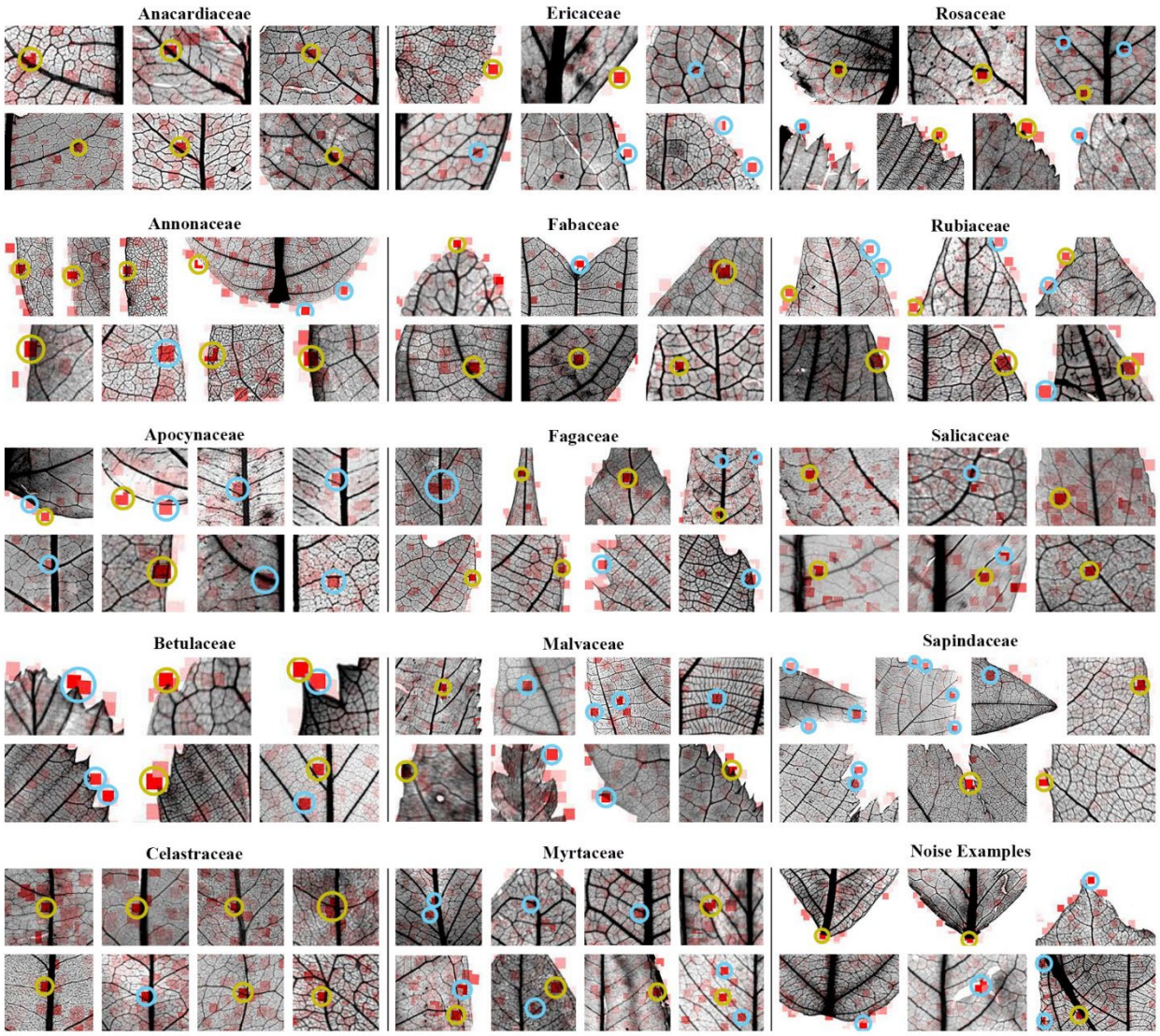

Figure 3. Selected examples of high-scoring features. **Anacardiaceae**. Secondary veins and secondary-tertiary junctions. Top, left to right: *Anacardium humile* (NCLC-W no. 12854), *Buchanania arborescens* (1758), *Cotinus coggygria* (4306). Bottom, left to right: *Mauria heterophylla* (4219), *Metopium 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 gracilus* (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).

Celastraceae

Celastraceae has the highest scores in the midsection, primary-secondary junctions, and secondary veins but has low tooth scores (Fig. 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; Fig. 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 could mean 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 (Fig. 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; Fig. 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.

Fabaceae

Fabaceae has high scores on the blade midsection and apical margin along with tertiary veins (Fig. 3). Fabaceae is one of the only families with a significant percentage of mucronate apices in the dataset, 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; Figs. 2, 3; see *Pterocarpus*; see Data Availability). For venation, scores for tertiary veins are high (top-5 of 1.5; top-1 of 0.4; Figs. 2, 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).

Fagaceae

Fagaceae has the highest scores on the midsection margin, primary vein, and tertiary veins (Fig. 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; Fig. 3; see *Castanopsis, Fagus*, and *Quercus*) and tertiary veins (top-1 of 0.4; Fig. 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 (Fig. 2; see Data Availability).

Malvaceae

A family with well-described leaf architecture (Carvalho et al., 2011; Hickey, 1997; Hickey & Wolfe, 1975), 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; Figs. 2, 3; e.g., *Tilia*). Hotspot squares are on both secondary and agrophic minor secondary veins with high frequency (Fig. 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 (see Data Availability), 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 (Fig. 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; Fig. 3), intersecondary veins (top-5 of 0.6; top-1 of 0.1; see *Calyptranthes* and Fig. 3), and primary-secondary junctions (top-5 of 0.4; Fig. 3). Similar to Celastraceae, many of these are primary-intersecondary junctions (Fig. 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; Fig. 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 (Gandolfo et al., 2011; MacGinitie, 1969; Manchester et al., 1998; 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).

Rosaceae

Rosaceae scores are highest on secondary veins and tooth apices (Fig. 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; Fig. 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; Fig. 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; Fig. 3), and notably so in *Crataegus.* However, the SIFT 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 (DeVore et al., 2004; Hickey & Wolfe, 1975; Kellner et al., 2012; Wolfe & Wehr, 1987). 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 (Fig. 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 Figs. 2, 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 (Table 2; see Data Availability), 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 (Boucher et al., 2003; Hickey & Wolfe, 1975; Manchester et al., 1986, 2006). 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; Fig. 3; see *Salix*). Frequently, the hotspot squares partially touch the secondary veins or secondary-tertiary junctions (scored as secondary veins; see Methods and Table 1; Fig. 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 (Fig. 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).

Table 3. Selected top-5 means comparisons for *Acer* and non-*Acer* Sapindaceae.

Feature	*Acer*[1]	Non-*Acer*[1]	All Sapindaceae[1]
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

[1] *Acer* (n=107), non-*Acer* (n=132), all Sapindaceae (n=239)

Noise features

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; Fig. 3), and Salicaceae is the only family with a high score for hotspot squares off the leaf (top-5 of 0.7; Fig. 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 (Figs. 4, 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, I note that there is minimal grouping of families due to ordinal membership 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 I could examine here. Studies of pteridophytes using

traditional leaf architecture characters have shown phylogenetic signals at the ordinal level or higher in PCA and cluster analyses (Tan & Buot, 2019).

Top-1 PCA.

For the top-1 PCA (Fig. 4A), families scoring high on axis 1 have high scores for secondary veins, as seen in Myrtaceae and Celastraceae, and the secondary vein vector has significant magnitude and almost parallels axis 1. Families scoring in the negative region of axis 1 have high scores for the basal margin of the leaf, seen in Ericaceae and Apocynaceae. Families scoring high on axis 2 have high scores for the apical margin, with positive scores for Betulaceae, Rubiaceae, and Rosaceae. Families scoring in the negative region of axis 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, i.e., 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.

Top-5 PCA.

For the top-5 PCA (Fig. 4B), families scoring high on axis 1 all have high scores for tooth apices and hotspot squares on the midsection and apical 25% margin, such as Betulaceae, Malvaceae, and Rosaceae. The vectors for these leaf architecture features indicate that they are influential on

axis 1. Families scoring low on axis 1 have high scores for squares within the midsection of the blade, including Myrtaceae and Celastraceae. Families scoring high on axis 2, such as Myrtaceae, Celastraceae, and Anacardiaceae, have high scores for secondary veins and the blade midsection. Families plotting in the negative region of axis 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 (Fig. 4A) is driven strongly by margin and location vectors, the top-5 PCA (Fig. 4B) is driven by margin, tooth, and venation vectors (specifically secondary and tertiary veins, midsection, basal, apex margin, in 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 within-genera averages (Fig. 4C) has a similar structure to the corresponding family PCA (Fig. 4B), and the vectors conserve a near-identical direction to their families. The genera of six of the fourteen families — Anacardiaceae, Betulaceae, Celastraceae, Ericaceae, Myrtaceae, and Rosaceae — respectively plot closely together in easily-defined ordination spaces, outlined in dashed lines (Fig. 4C). Fabaceae, Annonaceae, Apocynaceae, and Rubiaceae have overlapping and similar morphospaces that cannot be easily defined. Other families plot throughout the morphospace with no clear pattern, such as Salicaceae, Fagaceae, and Malvaceae.

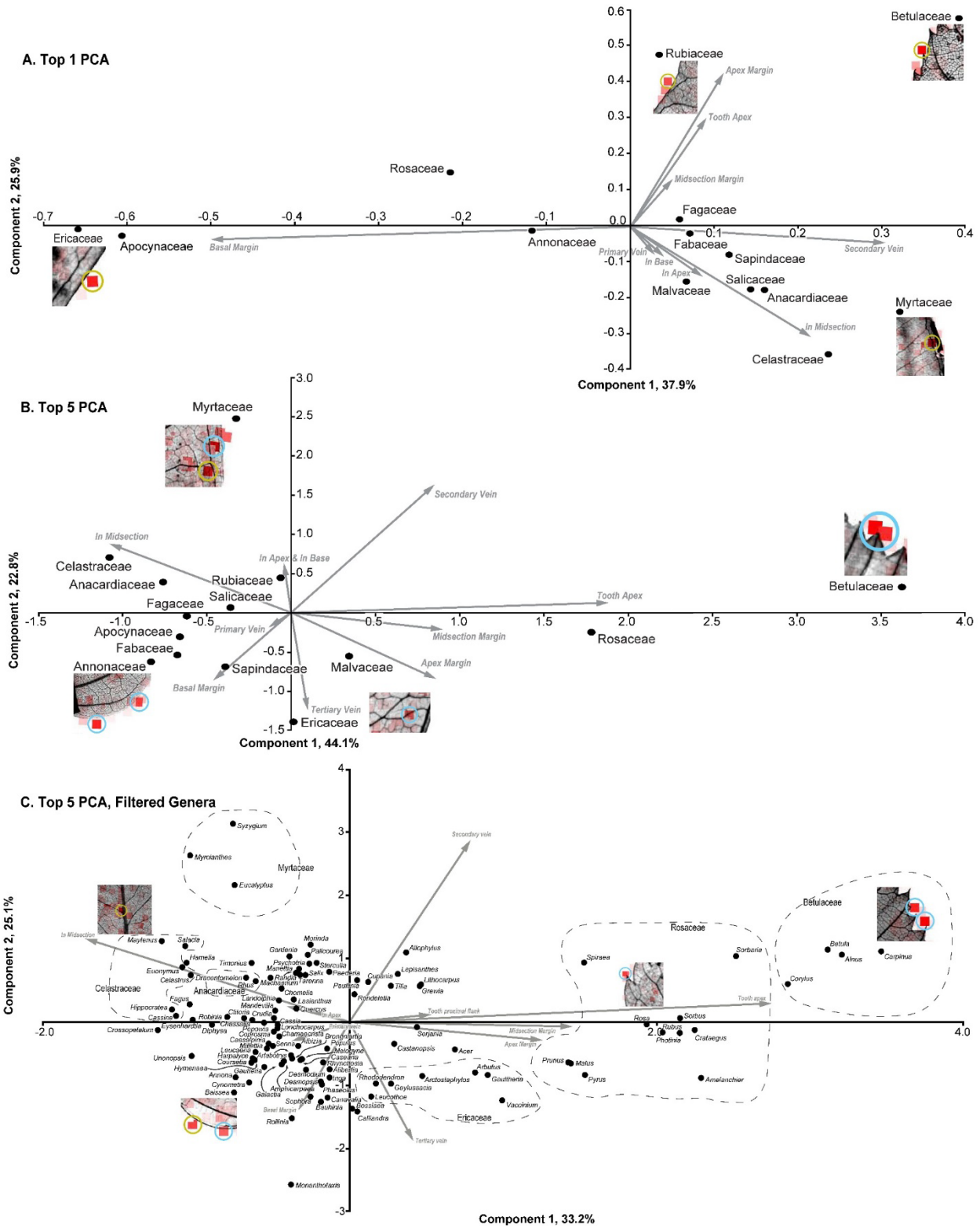

Figure 4. Principal component analyses (PCA) of top-1 and top-5 results, with vectors shown for influential leaf locations (Table 1) and the percentage of variance represented shown on the respective axis. Selected image patches are included as exemplars. **A.** Top-1 analysis for families

(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).

Cluster analysis.

The top-1 cluster dendrogram (Fig. 5) follows a similar pattern to the top-1 PCA (Fig. 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 paired 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).

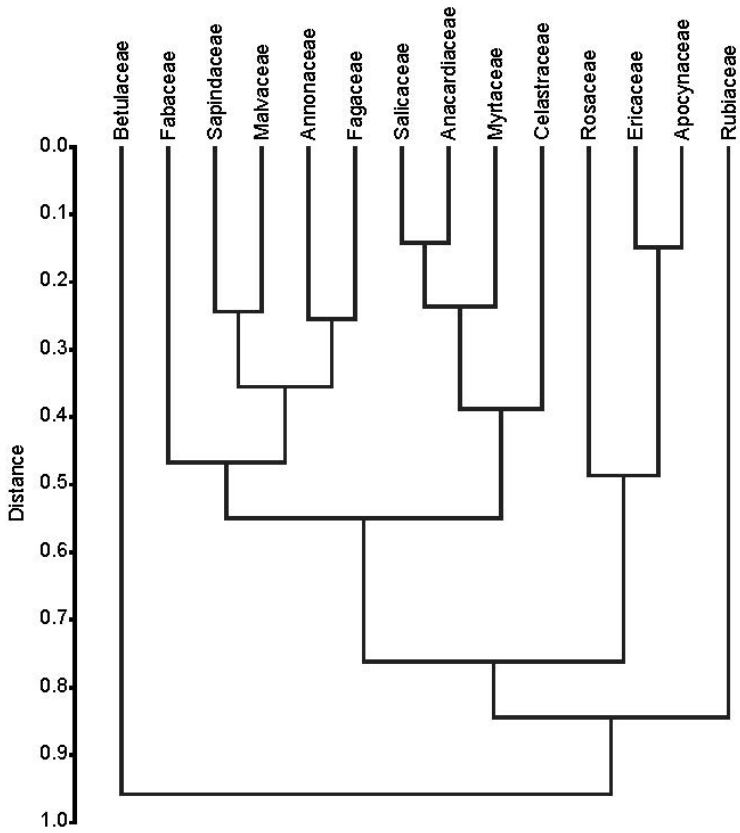

Figure 5. UPGMA cluster analysis of the mean top-1 family scores using Euclidean distances (y-axis).

Chapter 5

Discussion

These results show new possibilities for quantitatively interpreting computer vision signals into human-friendly botanical language by mapping, tabulating, and analyzing the regions of highest diagnostic value. Although I undertook 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. My results demonstrate that computer-vision heat maps that may, at first, appear to be noise, in fact provide a new pathway to uncover diagnostic features that were previously unnoticed in the complexity of angiosperm leaf-architecture. Although I do not attempt here to define new botanical characters, the work presents new leads for families with few to no established leaf-architectural features and enhances visual "gestalt" learning of leaf architecture (Fig. 3; see Data Availability). 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 the 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 & Doyle, 2012), and Ericaceae (Jordan et al., 2010), the highlighted features (Table 2) can be viewed as new leads in 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. I have also extracted new information in 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, 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 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 and 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, I 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). 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 my 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.

More broadly, 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 (Carvalho et al., 2011; DeVore & Pigg, 2007; Hickey & Wolfe, 1975; Wolfe & Wehr, 1987), along with the Myrtaceae intramarginal vein (Gandolfo et al., 2011; MacGinitie, 1969), all have well-known characters. For example, Carvalho et al. (2011)

discussed the malvoid tooth type (Hickey & Wolfe, 1975), secondary and tertiary principal veins, and agrophic-vein branching patterns that are diagnostic to Malvaceae, all of which are echoed in the heat maps. The heat maps cannot respond to some of the holistic leaf architecture characters used to identify fossil Malvaceae leaves such as actinodromous primary venation (Carvalho et al., 2011; Hickey, 1997; Hickey & Wolfe, 1975) because those features are much larger than the sampling points used in the SIFT 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 (Fig. 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 (Goh et al., 2021; LeCun et al., 2015; Serre, 2019; Voss et al., 2021; Yosinski et al., 2015) 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 (Linsley et al., 2021; Olah et al., 2018; 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 leaves, both fossil and extant (Bakker et al., 2020; Beaman & Cellinese, 2012; Bebber et al., 2010; Belhumeur et al., 2008; Hedrick et al., 2020; Mata-Montero & Carranza-Rojas, 2016; Page et al., 2015; Soltis et al., 2020).

Many hotspot regions that had high scores in the system are similar to those seen in fossil leaves from the respective families (Fig. 6), 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, my results show that coupling

traditional leaf-architecture knowledge with artificial intelligence will lead to improved identification and systematic understanding of modern and fossil leaves.

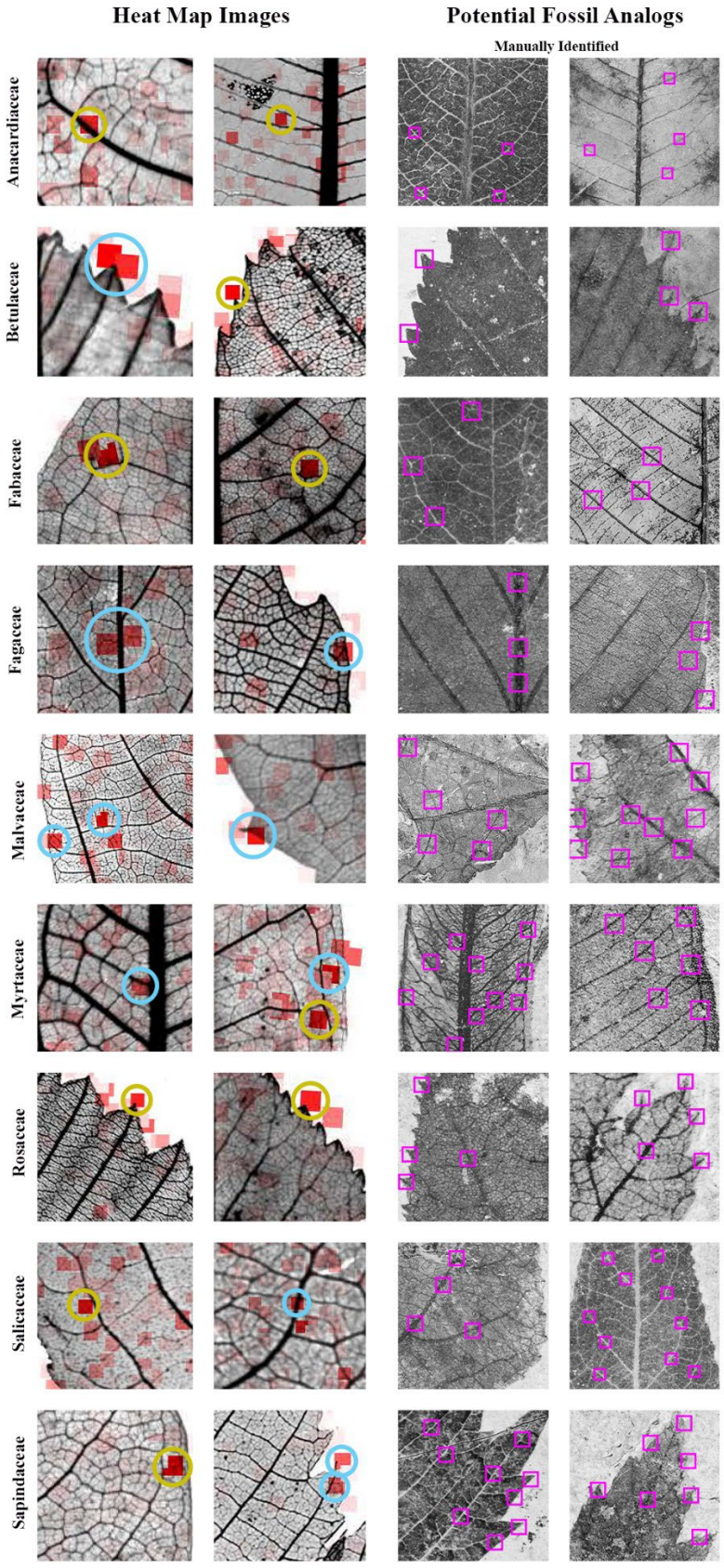
Heat Map Images
Potential Fossil Analogs
Manually Identified
Anacardiaceae
Betulaceae
Fabaceae
Fagaceae
Malvaceae
Myrtaceae
Rosaceae
Salicaceae
Sapindaceae

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 an open-access image collection organized by 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.

Chapter 6

Conclusions

Visual outputs from machine-learning experiments provide a novel approach for improving understanding of diagnostic features in plant morphology. Here, I 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 one of the first to do so for any type of computer-vision heat maps. The 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.

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). 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. Searching for computer-vision signals from extant leaves in fossil leaves has

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 “gestalt” learning and uncovering novel botanical characters that have been hiding in plain sight.

Chapter 7

Reflections

I always knew I wanted to be a paleontologist but my time at Penn State and my senior thesis work changed my trajectory and paleontological career interests. I was always a dinosaur kid growing up and had a steadfast resolve to work in a museum studying dinosaurs and vertebrate paleontology but when I arrived at PSU, there were no vertebrate paleontologists to work with. I joined the PSU Paleobotany lab with Dr. Peter Wilf my first year thinking that whatever skills I learn will be transferable to vertebrate paleontology. Dr. Wilf saw my passion for paleontology and helped me start my independent project, which has led to this thesis. Through this project, and additional paleobotanical projects, I fully immersed myself in paleobotany and the paleobotanical community. After almost four years in the lab, I now plan to pursue a Ph.D. studying paleobotany and can easily see myself researching plant evolution and their fossil record for a career.

Even without the giant asterisk that is COVID-19, my undergraduate experience has been quite non-traditional. I came to Penn State in June 2018, before I even graduated high school, for the Millennium Scholars Program (MSP) summer bridge. Coming into fall 2018, I already knew my way around campus, knew over 40 students at PSU, my advisors, my course trajectory, and how to join a lab. It was because of MSP that I joined a research lab my first year and started my thesis work as early as possible. I would not be at Penn State without MSP and I couldn't be more grateful for their unending academic, personal, and financial support over the last four years. I am also a member of the Schreyer Honors College and Presidential Leadership Academy but, in all honesty, receive significantly less support from these programs. In terms of coursework, I will be graduating with over 220 credits, five minors, and two certificates. The geobiology degree allowed me to specialize and take courses that interested me without the courses that I'd never use required for a biology or geology degree (ex. biochemistry or geophysics). My five minors (wildlife and fisheries science, marine science, biology,

global and international studies, astrobiology), allowed me to expand my knowledge base into other disciplines oftentimes related to paleontology.

While COVID-19 did fundamentally change my education, despite what the administration might argue, it also allowed me to expand my research through additional virtual projects. My second year, I started another project with Dr. Wilf and Dr. Robert Kooyman (Macquarie University, Australia) compiling modern rainforest plot data from Southeast Asia to study the ecological importance and distribution of Gondwanan plant lineages and assessing paleo-heritage in the region. This project strengthened my data management skills and also my mentoring skills, as I oversaw several students on this project. Additionally, I participated in an REU with the Smithsonian National Museum of Natural History with Drs. Camilla Souto and Gene Hunt in which I studied cassiduloid evolution. This project allowed me to try my hand at invertebrate paleontology and paleobiology and while I enjoyed my project, I did not feel the same spark that I felt for my paleobotany projects. While COVID allowed me to expand my research projects, it also stymied my field skills and hands-on work with fossils. To remedy this, I applied for, and received, an Erickson Grant and worked on plant macrofossils from the Tanjung Formation, Borneo, Indonesia. While the collection is small, it has allowed me to learn the skills necessary to describe a fossil flora and how to describe fossil specimens.

My thesis itself was a labor of love that had many stalls along the way. Starting this project early allowed me to conduct a more thorough and comprehensive thesis than most other senior theses. I studied all 14 families, learned and conducted all analyses, designed all figures, and wrote the manuscript at a level to submit for publication. Creating the scoring system itself took over six months of trial and error and another year to score all of the heat maps and run analyses. I spent another year and a half writing the manuscript and creating the figures using Adobe Photoshop and Illustrator (which has a very large learning curve). Along the way, I fine-tuned the project through presentations at conferences including the Midcontinent Paleobotanical Colloquium (MPC) 2020, Botany 2020 and 2021, and the Geobiology Symposium 2021. This project also allowed me to explore the process of scientific publication and peer

review. As mentioned in the Preface, my honors thesis has been published in *American Journal of Botany*. I am the first undergraduate researcher in the PSU paleobotany lab to publish a paper as first author and corresponding author before graduating. As first and corresponding author, I also wrote the cover letter, created the Figshare file, submitted the manuscript for publication, and responded to peer reviews. This experience put into perspective what every peer-reviewed paper goes through before it can be read by scientists. After almost four years working with Dr. Peter Wilf and my extensive paleobotanical research, I cannot see myself leaving paleobotany. I do not yet know where I will spend the next five years, but I know it will be in a Ph.D. program studying plant fossils.

Appendix A

Data Availability

All marked-up heat maps and data matrices generated from this work are available open access on Figshare, https://doi.org/10.6084/m9.figshare.17010020. This includes separate pdf heat map files for each family, a csv file with raw scores for all heat maps, and summary files for mean and median scores for each family. I also included within-family means and medians separated by toothed vs. untoothed and lobed vs. unlobed leaves, in separate csv files. Finally, box plots of top-1 scores by family for each of 21 leaf locations as pdf file are also available. See Fig. 2 of this thesis for top-5 box plots and more information.

The original heat maps that I scored for this article (Wilf et al., 2016) were previously published on Figshare (https://doi.org/10.6084/m9.figshare.1521157.v1).

BIBLIOGRAPHY

Almeida, B. K., Garg, M., Kubat, M., & Afkhami, M. E. (2020). Not that kind of tree: Assessing the potential for decision tree–based plant identification using trait databases. *Applications in Plant Sciences*, *8*(7), e11379. https://doi.org/10.1002/aps3.11379

Andrés-Hernández, A. R., & Terrazas, T. (2009). Leaf architecture of *Rhus* s.str. (Anacardiaceae). *Feddes Repertorium*, *120*(5–6), 293–306. https://doi.org/10.1002/fedr.200911109

Angiosperm Phylogeny Group. (1998). An ordinal classification for the families of flowering plants. *Annals of the Missouri Botanical Garden*, *85*(4), 531–553. https://doi.org/10.2307/2992015

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*(2), 105–121. https://doi.org/10.1111/j.1095-8339.2009.00996.x

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), 1–20. https://doi.org/10.1111/boj.12385

Bacon, C. D., Simmons, M. P., Archer, R. H., Zhao, L.-C., & Andriantiana, J. (2016). Biogeography of the Malagasy Celastraceae: Multiple independent origins followed by widespread dispersal of genera from Madagascar. *Molecular Phylogenetics and Evolution*, *94*(A), 365–382. https://doi.org/10.1016/j.ympev.2015.09.013

Bakker, F. T., Antonelli, A., Clarke, J. A., Cook, J. A., Edwards, S. V., Ericson, P. G. P., Faurby, S., Ferrand, N., Gelang, M., Gillespie, R. G., Irestedt, M., Lundin, K., Larsson, E., Matos-Maraví, P., Müller, J., Proschwitz, T. von, Roderick, G. K., Schliep, A., Wahlberg, N., … Källersjö, M. (2020). The Global Museum: Natural history collections and the future of evolutionary science and public education. *PeerJ*, *8*, e8225. https://doi.org/10.7717/peerj.8225

Bama, B. S., Valli, S. M., Raju, S., & Kumar, V. A. (2011). Context based leaf image retrieval (CBLIR) using shape, color, and texture features. *Indian Journal of Computer Science and Engineering*, *2*(2), 202–211. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.300.4263&rep=rep1&type=pdf

Banerjee, S., & Pamula, R. (2020). Random Forest boosted CNN: An empirical technique for plant classification. In J. K. Mandal & S. Mukhopadhyay (Eds.), *Proceedings of the Global AI Congress 2019* (Vol. 1112, pp. 251–261). Springer. https://doi.org/10.1007/978-981-15-2188-1_20

Barclay, R. S., & Johnson, K. R. (2004). West Bijou Site Cretaceous-Tertiary boundary, Denver Basin, Colorado. In *Nelsen, E.P. and Erslev, E.A., eds., Field Trips in the Southern Rocky Mountains, USA,* (pp. 59–68). Geological Society of America Field Guide 5. https://doi.org/10.1130/0-8137-0005-1.59

Barclay, R. S., Johnson, K. R., Betterton, W. J., & Dilcher, D. L. (2003). Stratigraphy and megaflora of a K-T boundary section in the eastern Denver Basin, Colorado. *Rocky Mountain Geology*, *38*(1), 45–71. https://doi.org/10.2113/gsrocky.38.1.45

Beaman, R. S., & Cellinese, N. (2012). Mass digitization of scientific collections: New opportunities to transform the use of biological specimens and underwrite biodiversity science. *ZooKeys*, *209*, 7–17. https://doi.org/10.3897/zookeys.209.3313

Bebber, D. P., Carine, M. A., Wood, J. R. I., Wortley, A. H., Harris, D. J., Prance, G. T., Davidse, G., Paige, J., Pennington, T. D., Robson, N. K. B., & Scotland, R. W. (2010). Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences*, *107*(51), 22169–22171. https://doi.org/10.1073/pnas.1011841108

Belhumeur, P. N., Chen, D., Feiner, S., Jacobs, D. W., Kress, W. J., Ling, H., Lopez, I., Ramamoorthi, R., Sheorey, S., White, S., & Zhang, L. (2008). Searching the world's herbaria: A system for visual identification of plant pecies. In D. Forsyth, P. Torr, & A. Zisserman (Eds.), *Computer Vision – ECCV 2008* (Vol. 5305, pp. 116–129). Springer. https://doi.org/10.1007/978-3-540-88693-8_9

Boucher, L. D., Manchester, S. R., & Judd, W. S. (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*(9), 1389–1399. https://doi.org/10.3732/ajb.90.9.1389

Brummitt, N., Araújo, A. C., & Harris, T. (2021). Areas of plant diversity—What do we know? *Plants, People, Planet*, *3*(1), 33–44. https://doi.org/10.1002/ppp3.10110

Bryson, A. E., Brown, M. W., Mullins, J., Dong, W., Bahmani, K., Bornowski, N., Chiu, C., Engelgau, P., Gettings, B., Gomezcano, F., Gregory, L. M., Haber, A. C., Hoh, D., Jennings, E. E., Ji, Z., Kaur, P., Raju, S. K. K., Long, Y., Lotreck, S. G., … Chitwood, D. H. (2020). Composite modeling of leaf shape across shoots discriminates *Vitis* species

better than individual leaves. *Applications in Plant Sciences*, *8*(12), e11404. https://doi.org/doi.org/10.1002/aps3.11404

Caballero, C., & Aranda, M. C. (2010). Plant species identification using leaf image retrieval. *Proceedings of the ACM International Conference on Image and Video Retrieval*, 327–334. https://doi.org/10.1145/1816041.1816089

Cámara-Leret, R., & Bascompte, J. (2021). Language extinction triggers the loss of unique medicinal knowledge. *Proceedings of the National Academy of Sciences*, *118*(24), e2103683118. https://doi.org/10.1073/pnas.2103683118

Cámara-Leret, R., Raes, N., Roehrdanz, P., De Fretes, Y., Heatubun, C. D., Roeble, L., Schuiteman, A., van Welzen, P. C., & Hannah, L. (2019). Climate change threatens New Guinea's biocultural heritage. *Science Advances*, *5*(11), eaaz1455. https://doi.org/10.1126/sciadv.aaz1455

Carranza-Rojas, J., Goëau, H., Bonnet, P., Mata-Montero, E., & Joly, A. (2017). Going deeper in the automated identification of herbarium specimens. *BMC Evolutionary Biology*, *17*(181), 1–14. https://doi.org/10.1186/s12862-017-1014-z

Carranza-Rojas, J., Joly, A., Goëau, H., Mata-Montero, E., & Bonnet, P. (2018). Automated identification of herbarium specimens at different taxonomic levels. In A. Joly, S. Vrochidis, K. Karatzas, A. Karppinen, & P. Bonnet (Eds.), *Multimedia Tools and Applications for Environmental & Biodiversity Informatics* (pp. 151–167). Springer International Publishing. https://doi.org/10.1007/978-3-319-76445-0_9

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

Carvalho, M. R., Herrera, F. A., Jaramillo, C. A., Wing, S. L., & Callejas, R. (2011). Paleocene Malvaceae from northern South America and their biogeographical implications. *American Journal of Botany*, *98*(8), 1337–1355. https://doi.org/10.3732/ajb.1000539

Carvalho, M. R., Jaramillo, C., Parra, F. de la, Caballero-Rodríguez, D., Herrera, F., Wing, S., Turner, B. L., D'Apolito, C., Romero-Báez, M., Narváez, P., Martínez, C., Gutierrez, M., Labandeira, C., Bayona, G., Rueda, M., Paez-Reyes, M., Cárdenas, D., Duque, Á., Crowley, J. L., … Silvestro, D. (2021). Extinction at the end-Cretaceous and the origin of modern Neotropical rainforests. *Science*, *372*(6537), 63–68. https://doi.org/10.1126/science.abf1969

Champ, J., Mora-Fallas, A., Goëau, H., Mata-Montero, E., Bonnet, P., & Joly, A. (2020). Instance segmentation for the fine detection of crop and weed plants by precision agricultural robots. *Applications in Plant Sciences*, *8*(7), e11373. https://doi.org/10.1002/aps3.11373

Charters, J., Wang, Z., Chi, Z., Ah Chung Tsoi, & Feng, D. D. (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. https://doi.org/10.1109/ICMEW.2014.6890557

Correa-Narvaez, J. E., & Manchester, S. R. (2021). Distribution and morphological diversity of *Palaeocarpinus* (Betulaceae) from the Paleogene of the Northern Hemisphere. *The Botanical Review*. 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*(2), 103–136. https://doi.org/10.1111/j.1095-8339.1981.tb01224.x

Crepet, W. L., & Nixon, K. C. (1989). Earliest megafossil evidence of Fagaceae: Phylogenetic and biogeographic implications. *American Journal of Botany*, *76*(6), 842–855. JSTOR. https://doi.org/10.2307/2444540

Croat, T. B. (1978). *Flora of Barro Colorado Island*. Stanford University Press.

Das, A., Bucksch, A., Price, C. A., & Weitz, J. S. (2014). ClearedLeavesDB: An online database of cleared plant leaf images. *Plant Methods*, *10*, 8. https://doi.org/10.1186/1746-4811-10-8

Del Rio, C., Wang, T.-X., Liu, J., Liang, S.-Q., Spicer, R. A., Wu, F.-X., Zhou, Z.-K., & Su, T. (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*(1), 126–138. https://doi.org/10.1002/ajb2.1418

DeVore, M. L., Moore, S. M., Pigg, K. B., & Wehr, W. C. (2004). Fossil *Neviusia* leaves (Rosaceae: Kerrieae) from the lower-middle Eocene of southern British Columbia. *Rhodora*, *106*(927), 197–209. https://www.jstor.org/stable/23314752

DeVore, M. L., & Pigg, K. B. (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*(1), 45–57. https://doi.org/10.1007/s00606-007-0540-3

Dietl, G. P., & Flessa, K. W. (2011). Conservation paleobiology: Putting the dead to work. *Trends in Ecology & Evolution*, *26*(1), 30–37. https://doi.org/10.1016/j.tree.2010.09.010

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

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

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

Ellis, B., Daly, D. C., Hickey, L. J., Johnson, K. R., Mitchell, J. D., Wilf, P., & Wing, S. L. (2009). *Manual of Leaf Architecture* (2nd ed.). Cornell University Press.

Feild, T. S., Brodribb, T. J., Iglesias, A., Chatelet, D. S., Baresch, A., Upchurch Jr., G. R., Gomez, B., Mohr, B. A. R., Coiffard, C., Kvacek, J., & Jaramillo, C. (2011). Fossil evidence for Cretaceous escalation in angiosperm leaf vein evolution. *Proceedings of the National Academy of Sciences*, *108*(20), 8363–8366. https://doi.org/10.1073/pnas.1014456108

Foote, M. (1994). Morphological disparity in Ordovician-Devonian crinoids and the early saturation of morphological space. *Paleobiology*, *20*(3), 320–344. JSTOR. https://www.jstor.org/stable/2401006

Foote, M. (1995). Morphological diversification of Paleozoic crinoids. *Paleobiology*, *21*(3), 273–299. https://doi.org/10.1017/S0094837300013300

Friis, E. M., Crane, P. R., & Pedersen, K. R. (2011). *Early Flowers and Angiosperm Evolution*. Cambridge University Press.

Gandolfo, M. A., Hermsen, E. J., Zamaloa, M. C., Nixon, K. C., González, C. C., Wilf, P., Cúneo, N. R., & Johnson, K. R. (2011). Oldest known *Eucalyptus* macrofossils are from South America. *PLOS ONE*, *6*(6), e21084. https://doi.org/10.1371/journal.pone.0021084

Gentry, A. H. (1993). *A Field Guide to the Families and Genera of Woody Plants of Northwest South America (Colombia, Ecuador, Peru). Conservation International* (1st ed.). University of Chicago Press.

Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., & Olah, C. (2021). Multimodal neurons in artificial neural networks. *Distill*, *6*(3), e30. https://doi.org/10.23915/distill.00030

Gouveia, F., Filipe, V., Reis, M., Couto, C., & Bulas-Cruz, J. (1997). Biometry: The characterisation of chestnut-tree leaves using computer vision. *ISIE '97 Proceeding of the IEEE International Symposium on Industrial Electronics*, *3*, 757–760. https://doi.org/10.1109/ISIE.1997.648634

Graham, A. (2009). Fossil record of the Rubiaceae. *Annals of the Missouri Botanical Garden*, *96*(1), 90–108. https://doi.org/10.3417/2006165

Grinblat, G. L., Uzal, L. C., Larese, M. G., & Granitto, P. M. (2016). Deep learning for plant identification using vein morphological patterns. *Computers and Electronics in Agriculture*, *127*, 418–424. https://doi.org/10.1016/j.compag.2016.07.003

Harnik, P. G., Lotze, H. K., Anderson, S. C., Finkel, Z. V., Finnegan, S., Lindberg, D. R., Liow, L. H., Lockwood, R., McClain, C. R., McGuire, J. L., O'Dea, A., Pandolfi, J. M., Simpson, C., & Tittensor, D. P. (2012). Extinctions in ancient and modern seas. *Trends in Ecology & Evolution*, *27*(11), 608–617. https://doi.org/10.1016/j.tree.2012.07.010

Hedrick, B. P., Heberling, J. M., Meineke, E. K., Turner, K. G., Grassa, C. J., Park, D. S., Kennedy, J., Clarke, J. A., Cook, J. A., Blackburn, D. C., Edwards, S. V., & Davis, C. C. (2020). Digitization and the future of natural history collections. *BioScience*, *70*(3), 243–251. https://doi.org/10.1093/biosci/biz163

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

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. https://doi.org/10.1130/MEM150

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

Hu, R., Jia, W., Ling, H., & Huang, D. (2012). Multiscale distance matrix for fast plant leaf recognition. *IEEE Transactions on Image Processing*, *21*(11), 4667–4672. https://doi.org/10.1109/TIP.2012.2207391

Huff, P. M., Wilf, P., & Azumah, E. J. (2003). Digital future for paleoclimate estimation from fossil leaves? Preliminary results. *Palaios*, *18*(3), 266–274. https://doi.org/10.1669/0883-1351(2003)018<0266:DFFPEF>2.0.CO;2

Im, C., Nishida, H., & Kunii, T. L. (1998). Recognizing plant species by leaf shapes-a case study of the *Acer* family. *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No.98EX170)*, *2*, 1171–1173. https://doi.org/10.1109/ICPR.1998.711904

Ivory, S. J., Early, R., Sax, D. F., & Russell, J. (2016). Niche expansion and temperature sensitivity of tropical African montane forests. *Global Ecology and Biogeography*, *25*(6), 693–703. https://doi.org/10.1111/geb.12446

Jamil, N., Hussin, N. A. C., Nordin, S., & Awang, K. (2015). Automatic plant identification: Is shape the key feature? *Procedia Computer Science*, *76*, 436–442. https://doi.org/10.1016/j.procs.2015.12.287

Joly, A., Bonnet, P., Goëau, H., Barbe, J., Selmi, S., Champ, J., Dufour-Kowalski, S., Affouard, A., Carré, J., Molino, J.-F., Boujemaa, N., & Barthélémy, D. (2016). A look inside the Pl@ntNet experience. *Multimedia Systems*, *22*(6), 751–766. https://doi.org/10.1007/s00530-015-0462-9

Jordan, G. J., Bannister, J. M., Mildenhall, D. C., Zetter, R., & Lee, D. E. (2010). Fossil Ericaceae from New Zealand: Deconstructing the use of fossil evidence in historical biogeography. *American Journal of Botany*, *97*(1), 59–70. https://doi.org/10.3732/ajb.0900109

Kearney, M., & Porter, W. P. (2004). Mapping the fundamental niche: Physiology, climate, and the distribution of a nocturnal lizard. *Ecology*, *85*(11), 3119–3131. https://doi.org/10.1890/03-0820

Keller, R. (2004). *Identification of Tropical Woody Plants in the Absence of Flowers: A Field Guide.* (2nd ed., Issue Ed. 2). Birkhäuser.

Kellner, A., Benner, M., Walther, H., Kunzmann, L., Wissemann, V., & Ritz, C. M. (2012). Leaf architecture of extant species of *Rosa* L. and the Paleogene species *Rosa lignitum* Heer (Rosaceae). *International Journal of Plant Sciences*, *173*(3), 239–250. https://doi.org/10.1086/663965

Kooyman, R. M., Morley, R. J., Crayn, D. M., Joyce, E. M., Rossetto, M., Slik, J. W. F., Strijk, J. S., Su, T., Yap, J.-Y. S., & Wilf, P. (2019). Origins and assembly of Malesian

rainforests. *Annual Review of Ecology, Evolution, and Systematics*, *50*(1), 119–143. https://doi.org/10.1146/annurev-ecolsys-110218-024737

Kooyman, R. M., Wilf, P., Barreda, V. D., Carpenter, R. J., Jordan, G. J., Sniderman, J. M. K., Allen, A., Brodribb, T. J., Crayn, D., Feild, T. S., Laffan, S. W., Lusk, C. H., Rossetto, M., & Weston, P. H. (2014). Paleo-Antarctic rainforest into the modern Old World tropics: The rich past and threatened future of the "southern wet forest survivors." *American Journal of Botany*, *101*(12), 2121–2135. https://doi.org/10.3732/ajb.1400340

Krug, A. Z., & Patzkowsky, M. E. (2007). Geographic variation in turnover and recovery from the Late Ordovician mass extinction. *Paleobiology*, *33*(3), 435–454. https://doi.org/10.1666/06039.1

Kubitzki, K., & Bayer, C. (2013). *Flowering plants. Dicotyledons: Malvales, Capparales and Non-betalain Caryophyllales* (Vol. 5). Springer Science & Business Media.

Kumar, N., Belhumeur, P. N., Biswas, A., Jacobs, D. W., Kress, W. J., Lopez, I. C., & Soares, J. V. B. (2012). Leafsnap: A computer vision system for automatic plant species identification. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), *Computer Vision – ECCV 2012* (pp. 502–516). Springer.

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

Larese, M. G., Bayá, A. E., Craviotto, R. M., Arango, M. R., Gallo, C., & Granitto, P. M. (2014). Multiscale recognition of legume varieties based on leaf venation images. *Expert Systems with Applications*, *41*(10), 4638–4647. https://doi.org/10.1016/j.eswa.2014.01.029

Larese, M. G., Craviotto, R. M., Arango, M. R., Gallo, C., & Granitto, P. M. (2012). Legume identification by leaf vein images classification. In L. Alvarez, M. Mejail, L. Gomez, & J. Jacobo (Eds.), *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2012* (Vol. 7441, pp. 447–454). Springer. https://doi.org/10.1007/978-3-642-33275-3_55

Larese, M. G., & Granitto, P. M. (2016). Finding local leaf vein patterns for legume characterization and classification. *Machine Vision and Applications*, *27*(5), 709–720. https://doi.org/10.1007/s00138-015-0732-8

Larese, M. G., Namías, R., Craviotto, R. M., Arango, M. R., Gallo, C., & Granitto, P. M. (2014). Automatic classification of legumes using leaf vein image features. *Pattern Recognition*, *47*(1), 158–168. https://doi.org/10.1016/j.patcog.2013.06.012

Lebreton Anberrée, J., Manchester, S. R., Huang, J., Li, S., Wang, Y., & Zhou, Z.-K. (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*(7), 682–696. https://doi.org/10.1086/682166

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, S. H., Chan, C. S., Mayo, S. J., & Remagnino, P. (2017). How deep learning extracts and learns leaf features for plant classification. *Pattern Recognition*, *71*, 1–13. https://doi.org/10.1016/j.patcog.2017.05.015

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

Leebens-Mack, J. H., Barker, M. S., Carpenter, E. J., Deyholos, M. K., Gitzendanner, M. A., Graham, S. W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M., Rensing, S. A., Soltis, D. E., Soltis, P. S., Stevenson, D. W., Ullrich, K. K., Wickett, N. J., DeGironimo, L., … One Thousand Plant Transcriptomes Initiative. (2019). One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, *574*(7780), 679–685. https://doi.org/10.1038/s41586-019-1693-2

Linsley, J. W., Linsley, D. A., Lamstein, J., Ryan, G., Shah, K., Castello, N. A., Oza, V., Kalra, J., Wang, S., Tokuno, Z., Javaherian, A., Serre, T., & Finkbeiner, S. (2021). Super-human cell death detection with biomarker-optimized neural networks. *Sciences Advances* 7: eabf8142.

Little, D. P., Tulig, M., Tan, K. C., Liu, Y., Belongie, S., Kaeser-Chen, C., Michelangeli, F. A., Panesar, K., Guha, R. V., & Ambrose, B. A. (2020). An algorithm competition for automatic species identification from herbarium specimens. *Applications in Plant Sciences*, *8*(6), e11365. https://doi.org/10.1002/aps3.11365

Little, S. A., Kembel, S. W., & Wilf, P. (2010). Paleotemperature proxies from leaf fossils reinterpreted in light of evolutionary history. *PLOS ONE*, *5*(12), e15161. https://doi.org/10.1371/journal.pone.0015161

Looy, C. V., Brugman, W. A., Dilcher, D. L., & Visscher, H. (1999). The delayed resurgence of equatorial forests after the Permian–Triassic ecologic crisis. *Proceedings of the National Academy of Sciences*, *96*(24), 13857–13862. https://doi.org/10.1073/pnas.96.24.13857

Lu, H., Jiang, W., Ghiassi, M., Lee, S., & Nitin, M. (2012). Classification of *Camellia* (Theaceae) species using leaf architecture variations and pattern recognition techniques. *PLOS ONE*, *7*(1), e29704. https://doi.org/10.1371/journal.pone.0029704

Lyson, T. R., Miller, I. M., Bercovici, A. D., Weissenburger, K., Fuentes, A. J., Clyde, W. C., Hagadorn, J. W., Butrim, M. J., Johnson, K. R., Fleming, R. F., Barclay, R. S., Maccracken, S. A., Lloyd, B., Wilson, G. P., Krause, D. W., & Chester, S. G. B. (2019). Exceptional continental record of biotic recovery after the Cretaceous–Paleogene mass extinction. *Science*, *366*(6468), 977–983. https://doi.org/10.1126/science.aay2268

MacGinitie, H. D. (1969). *The Eocene Green River flora of northwestern Colorado and northeastern Utah*. University of California Publications in Geological Sciences. https://books.google.com/books?id=Q2LpM5cB_X8C&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Manchester, S. R. (2001). Leaves and fruits of *Aesculus* (Sapindales) from the Paleocene of North America. *International Journal of Plant Sciences*, *162*(4), 985–998. https://doi.org/10.1086/320783

Manchester, S. R., & Crane, P. R. (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*(8), 1147–1164. https://doi.org/10.2307/2443285

Manchester, S. R., Dilcher, D. L., & Tidwell, W. D. (1986). Interconnected reproductive and vegetative remains of *Populus* (Salicaceae) from the Middle Eocene Green River Formation, Northeastern Utah. *American Journal of Botany*, *73*(1), 156–160. https://doi.org/10.1002/j.1537-2197.1986.tb09691.x

Manchester, S. R., Dilcher, D. L., & Wing, S. L. (1998). Attached leaves and fruits of myrtaceous affinity from the Middle Eocene of Colorado. *Review of Palaeobotany and Palynology*, *102*(3), 153–163. https://doi.org/10.1016/S0034-6667(98)80002-X

Manchester, S. R., Judd, W. S., & Handley, B. (2006). Foliage and fruits of early poplars (Salicaceae: *Populus*) from the Eocene of Utah, Colorado, and Wyoming. *International Journal of Plant Sciences*, *167*(4), 897–908. https://doi.org/10.1086/503918

Marler, T. E., & del Moral, R. (2011). Primary succession along an elevation gradient 15 years after the eruption of Mount Pinatubo, Luzon, Philippines. *Pacific Science*, *65*(2), 157–173. https://doi.org/10.2984/65.2.157

Marshall, C. R., Finnegan, S., Clites, E. C., Holroyd, P. A., Bonuso, N., Cortez, C., Davis, E., Dietl, G. P., Druckenmiller, P. S., Eng, R. C., Garcia, C., Estes-Smargiassi, K., Hendy, A., Hollis, K. A., Little, H., Nesbitt, E. A., Roopnarine, P., Skibinski, L., Vendetti, J., & White, L. D. (2018). Quantifying the dark data in museum fossil collections as palaeontology undergoes a second digital revolution. *Biology Letters*, *14*(9), 20180431. https://doi.org/10.1098/rsbl.2018.0431

Martínez-Millán, M., & Cevallos-Ferriz, S. R. S. (2005). Arquitectura foliar de Anacardiaceae. *Revista Mexicana de Diodiversidad*, *76*(2), 137–190. http://www.scielo.org.mx/scielo.php?script=sci_abstract&pid=S1870-34532005000200003&lng=es&nrm=iso&tlng=es

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

Mata-Montero, E., & Carranza-Rojas, J. (2016). Automated plant species identification: Challenges and opportunities. In F. J. Mata & A. Pont (Eds.), *ICT for Promoting Human Development and Protecting the Environment. WITFOR 2016* (Vol. 481, pp. 26–36). Springer International Publishing. https://doi.org/10.1007/978-3-319-44447-5_3

McClain, A. M., & Manchester, S. R. (2001). *Dipteronia* (Sapindaceae) from the Tertiary of North America and implications for the phytogeographic history of the Aceroideae. *American Journal of Botany*, *88*(7), 1316–1325. https://doi.org/10.2307/3558343

McCune, B., & Grace, J. B. (2002). *Analysis of ecological communities*. MjM software design Gleneden Beach, OR.

Minowa, Y., & Nagasaki, Y. (2020). Convolutional neural network applied to tree species identification based on leaf images. *Journal of Forest Planning*, *26*, 1–11. https://doi.org/10.20659/jfp.2020.001

Mitchell, J. D., & Daly, D. C. (2015). A revision of *Spondias* L. (Anacardiaceae) in the Neotropics. *PhytoKeys*, *55*, 1–92. https://doi.org/10.3897/phytokeys.55.8489

Mouine, S., Yahiaoui, I., & Verroust-Blondet, A. (2012). Advanced shape context for plant species identification using leaf image retrieval. *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, 1–8. https://doi.org/10.1145/2324796.2324853

Nam, Y., Hwang, E., & Kim, D. (2008). A similarity-based leaf image retrieval scheme: Joining shape and venation features. *Computer Vision and Image Understanding*, *110*(2), 245–259. https://doi.org/10.1016/j.cviu.2007.08.002

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. *Distill*, *3*(3), e10. https://doi.org/10.23915/distill.00010

Owens, S. A., Fields, P. F., & Ewers, F. W. (1998). Degradation of the upper pulvinus in modern and fossil leaves of *Cercis* (Fabaceae). *American Journal of Botany*, *85*(2), 273–284. https://doi.org/10.2307/2446316

Page, L. M., MacFadden, B. J., Fortes, J. A., Soltis, P. S., & Riccardi, G. (2015). Digitization of biodiversity collections reveals biggest data on biodiversity. *BioScience*, *65*(9), 841–842. https://doi.org/10.1093/biosci/biv104

Park, J., Hwang, E., & Nam, Y. (2008). Utilizing venation features for efficient leaf image retrieval. *Journal of Systems and Software*, *81*(1), 71–82. https://doi.org/10.1016/j.jss.2007.05.001

Pigg, K. B., Manchester, S. R., & Wehr, W. C. (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*(5), 807–822. https://doi.org/10.1086/376816

Pirie, M. D., & Doyle, J. A. (2012). Dating clades with fossils and molecules: The case of Annonaceae. *Botanical Journal of the Linnean Society*, *169*(1), 84–116. https://doi.org/10.1111/j.1095-8339.2012.01234.x

Priya, C. A., Balasaravanan, T., & Thanamani, A. S. (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. https://doi.org/10.1109/ICPRIME.2012.6208384

Pryer, K. M., Tomasi, C., Wang, X., Meineke, E. K., & Windham, M. D. (2020). Using computer vision on herbarium specimen images to discriminate among closely related horsetails (*Equisetum*). *Applications in Plant Sciences*, *8*(6), e11372. https://doi.org/10.1002/aps3.11372

Punyasena, S. W., Tcheng, D. K., Wesseln, C., & Mueller, P. G. (2012). Classifying black and white spruce pollen using layered machine learning. *New Phytologist*, *196*(3), 937–944. https://doi.org/10.1111/j.1469-8137.2012.04291.x

Ramírez, J. L., & Cevallos-Ferriz, S. R. S. (2002). A diverse assemblage of Anacardiaceae from Oligocene sediments, Tepexi de Rodriguez, Puebla, Mexico. *American Journal of Botany*, *89*(3), 535–545. https://doi.org/10.3732/ajb.89.3.535

Ramírez, J. L., Cevallos-Ferriz, S. R. S., & Silva-Pineda, A. (2000). Reconstruction of the leaves of two new species of *Pseudosmodingium* (Anacardiaceae) from Oligocene Strata of Puebla, Mexico. *International Journal of Plant Sciences*, *161*(3), 509–519. https://doi.org/10.1086/314261

Romero, I. C., Kong, S., Fowlkes, C. C., Jaramillo, C., Urban, M. A., Oboh-Ikuenobe, F., D'Apolito, C., & Punyasena, S. W. (2020). Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy. *Proceedings of the National Academy of Sciences*, *117*(45), 28496–28505. https://doi.org/10.1073/pnas.2007324117

Roth Jr., J. L., & Dilcher, D. L. (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*(10), 1194–1207. JSTOR. https://doi.org/10.2307/2442218

Roy, K., & Foote, M. (1997). Morphological approaches to measuring biodiversity. *Trends in Ecology & Evolution*, *12*(7), 277–281. https://doi.org/10.1016/S0169-5347(97)81026-9

Rzanny, M., Mäder, P., Deggelmann, A., Chen, M., & Wäldchen, J. (2019). Flowers, leaves or both? How to obtain suitable images for automated plant identification. *Plant Methods*, *15*(77). https://doi.org/10.1186/s13007-019-0462-4

Sawangchote, P., Grote, P. J., & Dilcher, D. L. (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*(11), 2048–2061. https://doi.org/10.3732/ajb.0900086

Sawangchote, P., Grote, P. J., & Dilcher, D. L. (2010). Tertiary leaf fossils of *Semecarpus* (Anacardiaceae) from Li Basin, Northern Thailand. *Thai Forest Bulletin (Botany)*, *38*, 8–22. https://li01.tci-thaijo.org/index.php/ThaiForestBulletin/article/view/24335

Schuettpelz, E., Frandsen, P. B., Dikow, R., Brown, A., Orli, S., Peters, M., Metallo, A., Funk, V. A., & Dorr, L. (2017). Applications of deep convolutional neural networks to digitized natural history collections. *Biodiversity Data Journal*, *5*, e21139. https://doi.org/10.3897/BDJ.5.e21139

Seeland, M., Rzanny, M., Boho, D., Wäldchen, J., & Mäder, P. (2019). Image-based classification of plant genus and family for trained and untrained plant species. *BMC Bioinformatics*, *20*(1), 4. https://doi.org/10.1186/s12859-018-2474-x

Serre, T. (2019). Deep learning: The good, the bad, and the ugly. *Annual Review of Vision Science*, *5*(1), 399–426. https://doi.org/10.1146/annurev-vision-091718-014951

Simpson, M. G. (2010). 8 - Diversity and Classification of Flowering Plants: Eudicots. In M. G. Simpson (Ed.), *Plant Systematics* (2nd ed., pp. 275–448). Academic Press. https://doi.org/10.1016/B978-0-12-374380-0.50008-7

Soepadmo, E., & Wong, K. M. (1995). *Tree Flora of Sabah and Sarawak* (Vol. 1). Forest Research Institute Malaysia.

Soltis, P. S., Nelson, G., Zare, A., & Meineke, E. K. (2020). Plants meet machines: Prospects in machine learning for plant biology. *Applications in Plant Sciences*, *8*(6), e11371. https://doi.org/10.1002/aps3.11371

Spitzer, M., Wildenhain, J., Rappsilber, J., & Tyers, M. (2014). BoxPlotR: A web tool for generation of box plots. *Nature Methods*, *11*(2), 121–122. https://doi.org/10.1038/nmeth.2811

Stiles, E., Wilf, P., Iglesias, A., Gandolfo, M. A., & Cúneo, N. R. (2020). Cretaceous–Paleogene plant extinction and recovery in Patagonia. *Paleobiology*, *46*(4), 445–469. https://doi.org/10.1017/pab.2020.45

Sun, F., & Stockey, R. A. (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*(1), 136–146. https://doi.org/10.1086/297015

Tan, J. M. P., & Buot, I. E. (2019). Cluster and ordination analyses of leaf architectural characters in classifying Polypodiales *sensu* PPG. *Thailand Natural History Museum Journal*, *13*(1), 27–42.

Tarran, M., Wilson, P. G., Paull, R., Biffin, E., & Hill, R. S. (2018). Identifying fossil Myrtaceae leaves: The first described fossils of *Syzygium* from Australia. *American Journal of Botany*, *105*(10), 1748–1759. https://doi.org/10.1002/ajb2.1163

Taylor, E. L., Taylor, T. N., & Krings, M. (2009). *Paleobotany: The biology and evolution of fossil plants* (2nd ed.). Academic Press.

Tcheng, D. K., Nayak, A. K., Fowlkes, C. C., & Punyasena, S. W. (2016). Visual recognition software for binary classification and its application to spruce pollen identification. *PLOS ONE*, *11*(2), e0148879. https://doi.org/10.1371/journal.pone.0148879

Unger, J., Merhof, D., & Renner, S. (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*(1), 248. https://doi.org/10.1186/s12862-016-0827-5

Unger, S., Rollins, M., Tietz, A., & Dumais, H. (2020). INaturalist as an engaging tool for identifying organisms in outdoor activities. *Journal of Biological Education*. https://doi.org/10.1080/00219266.2020.1739114

Vajda, V., Raine, J. I., & Hollis, C. J. (2001). Indication of global deforestation at the Cretaceous-Tertiary Boundary by New Zealand fern spike. *Science*, *294*(5547), 1700–1702. https://doi.org/10.1126/science.1064706

Vizcarra, G., Bermejo, D., Mauricio, A., Gomez, R. Z., & Dianderas, E. (2021). The Peruvian Amazon forestry dataset: A leaf image classification corpus. *Ecological Informatics*, *62*, 101268. https://doi.org/10.1016/j.ecoinf.2021.101268

Voss, C., Cammarata, N., Goh, G., Petrov, M., Schubert, L., Egan, B., Lim, S. K., & Olah, C. (2021). Visualizing Weights. *Distill*, *6*(2), e00024.007. https://doi.org/10.23915/distill.00024.007

Wäldchen, J., & Mäder, P. (2018). Plant species identification using computer vision techniques: A systematic literature review. *Archives of Computational Methods in Engineering*, *25*(2), 507–543. https://doi.org/10.1007/s11831-016-9206-z

Wäldchen, J., Rzanny, M., Seeland, M., & Mäder, P. (2018). Automated plant species identification—Trends and future directions. *PLOS Computational Biology*, *14*(4), e1005993. https://doi.org/10.1371/journal.pcbi.1005993

White, A. E. (2020). Deep learning in deep time. *Proceedings of the National Academy of Sciences*, *117*(47), 29268–29270. https://doi.org/10.1073/pnas.2020870117

Wilf, P., S.L. Wing, H.W. Meyer, J. Rose, R. Saha, T. Serre, N.R. Cúneo, M.P. Donovan, D.M. Erwin, M.A. Gandolfo, E. González-Akre, F. Herrera, S. Hu, A. Iglesias, K.R. Johnson, T.S. Karim, X. Zou. 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. (2008). Fossil angiosperm leaves: Paleobotany's difficult children prove themselves. *Paleontological Society Papers*, *14*, 319–333. https://doi.org/10.1017/S1089332600001741

Wilf, P., Nixon, K. C., Gandolfo, M. A., & Cúneo, N. R. (2019). Eocene Fagaceae from Patagonia and Gondwanan legacy in Asian rainforests. *Science*, *364*(6444), eaaw5139. https://doi.org/10.1126/science.aaw5139

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

Wolfe, J. A., & Wehr, W. (1987). Middle Eocene dicotyledonous plants from Republic, northeastern Washington. In *Bulletin* (U. S. Geological Survey Bulletin No. 1597; Bulletin, p. 67). U.S. Government Printing Office; USGS Publications Warehouse. https://doi.org/10.3133/b1597

Wu, J.-Y., Ding, S.-T., Li, Q.-J., Zhao, Z.-R., Dong, C., & Sun, B.-N. (2014). A new species of *Castanopsis* (Fagaceae) from the upper Pliocene of West Yunnan, China and its biogeographical implications. *Palaeoworld*, *23*(3–4), 370–382. https://doi.org/10.1016/j.palwor.2014.10.005

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

Xing, Y., Gandolfo, M. A., Onstein, R. E., Cantrill, D. J., Jacobs, B. F., Jordan, G. J., Lee, D. E., Popova, S., Srivastava, R., Su, T., Vikulin, S. V., Yabe, A., & Linder, H. P. (2016). Testing the biases in the rich Cenozoic angiosperm macrofossil record. *International Journal of Plant Sciences*, *177*(4), 371–388. https://doi.org/10.1086/685388

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *In ICML Workshop on Deep Learning*, 1–12.

Zhao, C., Chan, S. S. F., Cham, W.-K., & Chu, L. M. (2015). Plant identification using leaf shapes—A pattern counting approach. *Pattern Recognition*, *48*(10), 3203–3215. https://doi.org/10.1016/j.patcog.2015.04.00

Edward J. Spagnuolo

Education

The Pennsylvania State University — *University Park, PA*
GPA: 3.930 — Expected Graduation: May 2022
Major: Geobiology Bachelor of Science with Honors
Minors: Biology, Wildlife and Fisheries Science, Astrobiology, Marine Science, Global and International Studies
Certificate: International Science
Dean's List — August 2018—Present

Research Experience

The Pennsylvania State University — University Park, PA
Paleobotany Research — 01/2019—Present

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

- Interpreted 3115 published computer vision heat map outputs on cleared leaves in 14 angiosperm families to isolate new features that can be used to identify extant and fossil leaves at the family level
- Used PAST and R to analyze data through principal component analyses, clusters, and box plots
- Found features used for computer vision analyses in all 14 families, some of which echoed in paleobotanical literature and others are new discoveries
- Presentation given at the Midcontinent Paleobotanical Colloquium 2020, Botany 2020 and 2021, Geobiology Symposium 2021, and Yale University Buds to Biomes virtual workshop 2020
- Manuscript accepted in press with American Journal of Botany

An open vegetation-plot database for Southeast Asia: tool for ecology, conservation, and paleo-conservation

- Generating database of over 450 published rainforest plots from Southeast Asia to study the penetration and ecological abundance of Gondwanan floras into modern SE Asia and its paleo-heritage
- Using Tabula to extract data and R to taxonomically update all species names
- Training and overseeing other students on project to increase dataset size
- Poster presented at Midcontinent Paleobotanical Colloquium 2021 and Botany 2021

First modern macrofloral reconnaissance of Paleogene Malay Archipelago

- Morphotyping only plant macrofossils collected from Cenozoic Indonesia within last century using Adobe Bridge and *Manual of Leaf Architecture*
- 45 fossil specimens collected in Tanjung Formation, South Kalimantan, Indonesian Borneo by Wilf and colleagues in 2014, 42 of which are fossil leaves
- Two large seed fossils are tentatively assigned to *Castanospermum* but will be CT scanned for internal structures and lead to systematic and phylogenetic analyses
- Some leaves express features characteristic of Melastomataceae, Myrtaceae, and Fabaceae
- Recipient of Penn State Erickson Discovery Grant to fund project

The National Museum of Natural History *Washington, D.C.*
Smithsonian NHRE REU 06/2021—Present

Exploring Morphological Disparity in the Cassiduloida (Echinodermata, Echinoidea) using geometric morphometrics

- Collected an image library of cassiduloid echinoids using public and private image sources from fossil and extant specimens from Jurassic to modern distributed in six continents
- Digitized 2D landmarks and semilandmarks on test outline and petals to gauge burrowing efficiency and respiration in 7 cassiduloid clades using tpsDig and generated principal component analyses in RStudio
- Discovered a selective extinction event of epifunal cassiduloids at the end-Cretaceous and a delayed partial experimentation into the morphospace in modern species
- Oversaw and mentored high school researcher who collected images and digitized specimens
- Won best undergraduate poster presentation by Paleontological Society at the Geological Society of America meeting in Portland, Oregon 2021

International and Field Experiences

The SSV Corwith Crammer *Caribbean Sea*
Biological Oceanography 12/2019—04/2020

- Studied aboard a sailing research vessel for nine days in the Northern Caribbean
- Conducted experiments on sediment characters, nutrient levels with depth, plankton density and diversity
- Learned how to deploy and use hydrocasts, neuston nets, phytoplankton nets, and spectrophotometers
- Capstone project on dissolved oxygen, chlorophyll-a, and phosphate levels throughout the water column

Soltis Center, Corcovado National Park *Campanario*, *Costa Rica*
Tropical Field Ecology 12/2018—01/2019

- Studied the ecological processes of several Costa Rican ecosystems
- Used statistical methods to analyze lizard, paca, and bat populations
- Designed and conducted individual project on the frequency of ectoparasites on bat species and their pregnancy status

Natural Hazards in Thailand *Bangkok, Sukhothai, Phuket, Kanchanaburi, Thailand*
May 2019

- Studied how developing nations differ in response to natural disasters compared to the United States
- Focused on the effects of earthquakes, tsunamis, and saltwater encroachment
- Developed intercultural communication skills through collaboration with professors at Kasetsart University

National Institute of Technology, Fukushima College
Fukushima, Japan Renewable Energy, Disaster Mitigation & Nuclear to Renewable Transitions
March 2019

- Competitively selected among 11 other global student leaders to join The GREEN Program: an international career focused experience in sustainable development

- Gained exclusive access to top-tier renewable energy facilities, networked with industry professionals, and engaged in coursework focused on energy, economics, and policy taught by industry experts
- Led an interdisciplinary team to develop an entrepreneurial business plan addressing a modern sustainability challenge connected to the Sustainable Development Goals (SDGs)

Volunteer Work and Campus Involvement

East End Hospice *Southampton, NY*
Camp Good Grief Volunteer 08/2014 — 04/2020

- Bereavement volunteer at Camp Good Grief, a camp dedicated to helping children properly grieve after the loss of a loved one
- Volunteer speaker to 30-40 volunteers each year
- Assisted in teaching proper grieving practices to groups of approximately 15 campers each year
- Recipient of Helping Makes You Happy Award from 4 years at camp

The Pennsylvania State University *University Park, PA*
Millennium Scholars Tutor and Mentor 09/2019—Present

- Tutoring students in calculus, biology, Earth materials, and geochemistry
- Tutor up to 7 hours a week through the Millennium Scholars Program
- Selected to serve as STEM mentor to minority underclass Millennium Scholars

The Pennsylvania State University *University Park, PA*
Millennium Society and Millennium THON 09/2019—Present
Vice President, Interim Treasurer, Fundraising Co-Chair 02/2021—Present

- Millennium Society is a service-based club dedicated to STEM education and engagement
- THON is the largest student run philanthropic event in the nation dedicated to generating funding for pediatric cancer victims and their families through the Four Diamonds Foundation
- Projects include Nittany Greyhounds, and educating elementary students on the benefits of STEM
- Organized and ran a plant sale fundraiser that earned over $350 for pediatric cancer and a book drive and book sale for the Midstate Literacy Center

The Pennsylvania State University *University Park, PA*
Minorities in the College of Earth and Mineral Sciences (MEMS) 04/2020—Present
Secretary and Founding Member

- Founded by underrepresented students in the College of EMS to provide a safe and uplifting space for all students, simplify the pathway to research, graduation, and graduate education, and increase networking and professional development opportunities for underrepresented students
- Helped create a seminar series to hear from underrepresented faculty members, their research experiences and experiences as a minority researcher in academia
- Assist in organizing bi-weekly meetings, study sessions, and social events.

Awards, Scholarships, and Accomplishments

The Pennsylvania State University *University Park, PA*
06/2018—Present

Millennium Scholars Program

- The Millennium Scholars Program is a merit-based scholarship program designed to prepare students for the pursuit of doctoral degrees in science, technology, engineering, and mathematics (STEM) disciplines. The program fosters not only academic excellence but values and fosters community service and engagement.

- Scholarships: Ira Lubert Scholarship, Millennium Scholars Scholarship, EMS Internal Merit Grant, Millennium Scholars Travel Grant (Japan, Caribbean, Geological Society of America Conference)

Schreyer Honors College
- Penn State honors program with the goal of academic excellence, global perspective, and leadership
- Scholarships: Schreyer Academic Excellence Scholarship, Matthew J. Wilson Honors Scholarship

The Presidential Leadership Academy
- Leadership program within Schreyer Honors College led by dean of Schreyer and president of Penn State Eric Barron focused on leadership in complex situations and gray areas
- PLA Academic Excellence Scholarship, PLA Travel Scholarship (Geoscience Field Camp)

College of Earth and Mineral Sciences Academy of Global Experiences
- EMS undergraduate award to recognize students exceling and fostering experiential learning, global experiences, service and integrity

Erickson Discovery Grant
- Penn State research grant for undergraduate independent project. Used to study the Tanjung fossil flora

Student Engagement Network Grant
- Penn State Scholarship funding international experiences. Funded for travel to Caribbean

Hedberg Geoscience Academic Excellence Scholarship and Edwin Drake Geoscience Scholarship

Penn State Provost Scholarship

Paleontological Society Best Undergraduate Poster Award 09/2021
- Awarded at Geological Society of America meeting at Portland, Oregon for work at Smithsonian

Professional Society Memberships:
- Botanical Society of America 05/2020—Present
- Paleontological Society 08/2021—Present
- Geological Society of America 08/2021—Present