Predicting photosynthetic pathway from anatomy using machine learning

Ian S. Gilman[1,2,3], Karolina Heyduk[4], Carlos Maya-Lastra[1,5], Lillian P. Hancock[1] and Erika J. Edwards[1]

[1]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520, USA; [2]Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA; [3]Plant Resilience Institute, Michigan State University, East Lansing, MI 48824, USA; [4]Department of Ecology and Evolutionary Biology, The University of Connecticut, Storrs, CT 06269, USA; [5]Department of Biology, Angelo State University, San Angelo, TX 76909, USA

Author for correspondence:
Ian S. Gilman
Email: gilmania@msu.edu

Received: *31 August 2023*
Accepted: *29 November 2023*

New Phytologist (2024) **242:** 1029–1042
doi: 10.1111/nph.19488

Key words: Asparagaceae, Crassulacean acid metabolism, machine learning, photosynthesis, Portullugo.

Summary

- Plants with Crassulacean acid metabolism (CAM) have long been associated with a specialized anatomy, including succulence and thick photosynthetic tissues. Firm, quantitative boundaries between non-CAM and CAM plants have yet to be established – if they indeed exist.
- Using novel computer vision software to measure anatomy, we combined new measurements with published data across flowering plants. We then used machine learning and phylogenetic comparative methods to investigate relationships between CAM and anatomy.
- We found significant differences in photosynthetic tissue anatomy between plants with differing CAM phenotypes. Machine learning-based classification was over 95% accurate in differentiating CAM from non-CAM anatomy, and had over 70% recall of distinct CAM phenotypes. Phylogenetic least squares regression and threshold analyses revealed that CAM evolution was significantly correlated with increased mesophyll cell size, thicker leaves, and decreased intercellular airspace.
- Our findings suggest that machine learning may be used to aid the discovery of new CAM species and that the evolutionary trajectory from non-CAM to strong, obligate CAM requires continual anatomical specialization.

Introduction

Carbon-concentrating mechanisms increase the efficiency of photosynthesis by raising the concentration of CO_2 inside photosynthetic tissues relative to the ambient environment. The most common carbon-concentrating mechanism, Crassulacean acid metabolism (CAM), was first discovered because of marked physiological differences between succulent and nonsucculent plants (de Saussure, 1804). Generally, CAM species conduct gas exchange at night to reduce transpirational water loss; the nocturnally fixed carbon is stored as malic acid overnight and released the next day behind closed stomata, thereby saturating photosynthetic tissues with CO_2 (Osmond, 1978). Although the co-occurrence of CAM and succulent anatomy is so consistent that botanists have used it as a guide to find new CAM plants (Coutinho, 1969), quantitative relationships between anatomy and CAM remain elusive.

Crassulacean acid metabolism and succulence may be correlated because they are co-selected as adaptations to water limitation. CAM species can be up to eightfold as water use efficient as C_3 species (Winter *et al*., 2005), and the water stored in succulent plants is essential for drought avoidance (Males, 2017). Although such a correlation does not necessarily imply that derived anatomy is a prerequisite of, or is caused by, CAM evolution, there are at least two hypothesized direct functional links between CAM and succulent anatomy. First, storage of nocturnally fixed CO_2 as malic acid in mesophyll vacuoles may require large vacuoles in photosynthetic cells and therefore larger, succulent mesophyll cells (Zambrano *et al*., 2014; Töpfer *et al*., 2020). Second, increased succulence in mesophyll cells may lower intercellular air space (IAS) and therefore mesophyll CO_2 conductance (g_m; Nelson & Sage, 2008; Cousins *et al*., 2020). In a study of *Kalanchoë daigremontiana*, Maxwell *et al*. (1997) found the CO_2 partial pressures of leaf IAS and at Rubisco carboxylation sites to be 205 and 109 μbar, respectively, demonstrating that CO_2 diffusion is strongly limited in this succulent CAM species. Thus, increased succulence may increase selection for CAM by lowering the efficiency of C_3 photosynthesis, particularly in high-temperature environments that exacerbate photorespiration (Nelson & Sage, 2008; Edwards, 2019). It is also possible that the evolution of CAM does not entail selection on succulence *per se*, but that the use of CAM reduces constraints on succulence evolution by removing g_m limitations due to carbon concentration (Leverett *et al*., 2023).

Quantitative studies of CAM and anatomy have mostly been restricted to relatively few taxa at the extremes of the CAM phenotypic spectrum, but have generally found positive correlations between CAM and succulence. Individual studies have reported that CAM species tend to have greater leaf thickness (LT) and larger mesophyll cell area (MA), although mixed trends have been observed for IAS (Nelson *et al.*, 2005; Nelson & Sage, 2008; Zambrano *et al.*, 2014; Earles *et al.*, 2018; Luján *et al.*, 2022); however, a recent meta-analysis of these relationships found inconsistent trends across clades (Herrera, 2020). Recently, hybrids between species with different photosynthetic types have been used to study the relationships between CAM activity and anatomical traits. In both *Yucca* (Agavoideae, Asparagaceae; Heyduk *et al.*, 2020) and *Cymbidium* (Epidendroideae, Orchidaceae; Yamaga-Hatakeyama *et al.*, 2022), hybrids of C_3 × CAM crosses possessed intermediate anatomical phenotypes and CAM activity. Within *Yucca* hybrid genotypes, however, the correlations between CAM activity and anatomy decreased in magnitude or disappeared entirely (Heyduk *et al.*, 2020).

The mosaic of past research provides limited insight into the evolution of CAM and photosynthetic tissue anatomy because it has focused on the extremes of CAM phenotypes (i.e. non-CAM species and species that use CAM as their primary metabolism). However, there are many recognized CAM phenotypes that differ in pattern and magnitude of CAM activity. CAM activity varies along multiple axes, including the degree to which is it facultative or constitutive, the extent of nocturnal stomatal conductance, and the amount of CO_2 sequestered as malic acid (Winter, 2019; Gilman *et al.*, 2023). Despite the diversity in CAM activity, most CAM-capable species either use CAM as their primary method of carbon fixation or use CAM for a small minority of carbon fixation, as evidenced by carbon isotope ratios (Messerschmid *et al.*, 2021). Here, we use term 'CAM' to refer to all species capable of CAM, regardless of strength or pattern of expression, and 'minority CAM' (mCAM) and 'primary CAM' (pCAM) to refer to species that fix the minority and majority of CO_2 with CAM, respectively. Primary CAM is consistent with past definitions of 'CAM plant' (Winter, 2019) and 'strong CAM' (Edwards, 2019, 2023), while mCAM encompasses species that can facultatively use CAM or constitutively use CAM at low levels, but primarily use C_3 or C_4 photosynthesis for CO_2 assimilation (mCAM = 'C_3 + CAM' of Edwards, 2019, but with the inclusion of C_4 + CAM species). We use the term 'succulence' to refer to tissues with enlarged living cells that store large amounts of withdrawable water, and typically exhibit larger cells, thicker leaves, and reduced IAS. This broad definition masks vast diversity in succulent anatomy at the cellular- and whole plant levels (Males, 2017). It is generally assumed that the evolution of pCAM requires transitioning through mCAM (Hancock & Edwards, 2014; Yang *et al.*, 2015; Edwards, 2019), but the relative timings of anatomical shifts during the evolution of mCAM and pCAM – and whether or not mCAM species possess a specialized anatomy – remain open questions.

Here, we combined anatomical measurements from thousands of angiosperm species from over 200 families to draw anatomical boundaries between non-CAM, mCAM, and pCAM phenotypes. Using supervised machine learning, we were able to classify CAM phenotypes from anatomical measurements with moderate to high accuracy. Finally, in a detailed study of the Portullugo clade (Fig. 1), we reconstructed the evolution of CAM and used phylogenetic comparative methods to establish significant relationships between anatomy and CAM evolution. Our findings support the hypothesis that CAM evolution entails anatomical evolution and reveal nuances about the earliest stages of CAM evolution.

Materials and Methods

Public anatomical data sets, taxon sampling, and specimen imaging

Publicly available data were gathered from the TRY (Fraser, 2020) and BROT2 (Tavşanoğlu & Pausas, 2018) plant trait databases and individual studies of CAM anatomy in Orchidaceae (Silvera *et al.*, 2005), Bromeliaceae (Males, 2018), Asparagaceae (Heyduk *et al.*, 2016), Caryophyllales (Ogburn & Edwards, 2012, 2013), Papua New Guinean epiphytes (Earnshaw *et al.*, 1987), Clusiaceae (Luján *et al.*, 2022), and across angiosperms (Nelson & Sage, 2008; Supporting Information Table S1). These data contained observations of MA, LT, mesophyll IAS, leaf dry matter content (LDMC), and specific leaf area (SLA) per unit dry mass. We generated two new datasets of MA, IAS, and LT for members of the Asparagaceae (subfamilies Agavoideae and Nolinoideae) and Portullugo (the clade inclusive of families Anacampserotaceae, Basellaceae, Cactaceae, Didiereaceae, Montiaceae, Molluginaceae, Portulacaceae, and Talinaceae; Table S2). In 2017, leaf cross sections were taken from 15 Portullugo species grown at Brown University, Providence, RI. Tissue sections were immediately placed in 10% neutral buffered formalin and sent to the Veterinary Diagnostic Laboratories of the College of Veterinary Medicine at the University of Georgia (Athens, GA) for fixation, embedding, and sectioning and staining with toluidine blue. In the spring of 2019, we collected leaf or stem cross sections of 41 species of Asparagaceae and 38 species of Portullugo growing at the Desert Botanical Garden (Phoenix, AZ); fixed specimens were created as above and imaged on an Olympus BX51 microscope (Evident Corp., Toyko, Japan) with an Infinity3-3UR camera (Teledyne Lumenera, Ottawa, Canada). To supplement our sampling, we were provided high-resolution images of leaf cross sections of 13 *Portulaca* (Portulacaceae) species used in Ocampo *et al.* (2013) by the authors.

The multiple data sets had some taxonomic overlap and some included multiple measurements from multiple accessions of the same species. To reduce our data set to one observation per species, we took the mean of each feature where multiple accessions were measured; these mean species values were used as the basis for analysis throughout. We binned each taxon into three CAM phenotypes based on Gilman *et al.* (2023) and references therein: C_3, C_3–C_4, and C_4 taxa were coded as 'non-CAM'; taxa that fix the minority of their daily CO_2 with CAM (C_3 + CAM, C_3–C_4 + CAM, and C_4 + CAM) were coded as mCAM; and taxa that primarily use CAM to fix CO_2 (i.e. over 50%, resulting in $\delta^{13}C$ ratios $\geq -18.7‰$; Winter & Holtum, 2002) as pCAM.

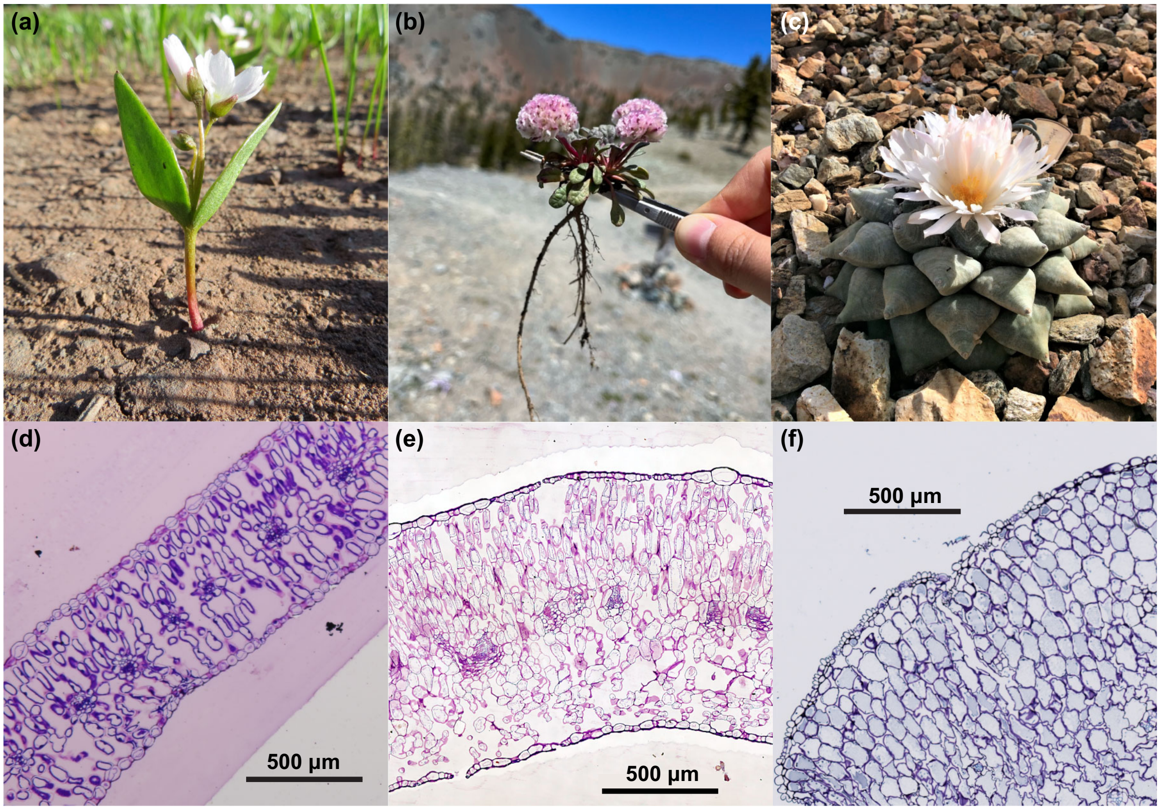

Fig. 1 Gross morphology (a–c) and photosynthetic tissue anatomy (d–f) of Portullugo species with varying Crassulacean acid metabolism (CAM) phenotypes sampled for this study: (a, d) non-CAM *Claytonia lanceolata* Pursh (Montiaceae) (b, e) minority CAM *Calyptridium umbellatum* (Torr.) Greene (Montiaceae), (c, f) primary CAM *Ariocarpus retusus* Scheidw. (Cactaceae). Nonauthor photograph credits: (a) Dr. Thomas Stoughton, (b) Anri Chomentowska, and (c) Desert Botanical Garden, Phoenix, AZ, USA.

The final data set contained observations from 5316 non-CAM, 207 mCAM, and 222 pCAM taxa (Dataset S1).

Measuring plant anatomy

Automated analyses of plant tissues can be difficult because many or most cells are in direct contact with other cells around much of their perimeters, rather than being separated by clear boundaries. We developed a lightweight image segmentation tool built in Python 3 with OpenCV v.4.5.2 (Bradski, 2000) called MiniContourFinder to facilitate measurement of histology slides. Segmentation in MiniContourFinder is accomplished through a combination of thresholding, gradient, and morphological operations (Methods S1; Fig. S1). MiniContourFinder was designed to allow users with minimal experience on the command line or image processing to quickly generate accurate and reproducible contours, particularly from plant histology images. MiniContourFinder can be run through the command line or a graphical user interface to tune contours in real time. We used MiniContourFinder to measure MA in our new Asparagaceae and Portullugo data sets. We used ImageJ v.1.53 (Schneider *et al*., 2012) to calculate LT (for leafy species) and IAS (in roughly 300 μm × 300 μm areas of mesophyll). All measurements were taken within chlorenchymous tissues and therefore excluded hydrenchyma, if present.

Statistical analysis

We investigated group differences in anatomical measurements by assessing normality and homoscedasticity, comparing raw and transformed data, testing for group differences, and finally using *post hoc* tests to identify group differences. We first assessed assumptions of normality using D'Angostino and Pearson's test (D'Agostino & Pearson, 1973) and homoscedasticity using Bartlett's test (Bartlett, 1937) of raw and $\log_{10}$-transformed data. None of the features were normal when raw or transformed, but $\log_{10}$-transformation substantially decreased heteroscedasticity: All transformed features were homoscedastic except SLA, which was much less heteroscedastic (Fig. S2; Table S3). We therefore continued with Kruskal–Wallis (KW) tests for group differences (Kruskal & Wallis, 1952) with the transformed data, and Dunn's *post hoc* tests (Dunn, 1964) where KW tests revealed significant group differences. We tested for correlations between transformed features using Pearson's *r* (Pearson, 1895). All statistical analyses were performed using Python v.3.7.12, SciPy v.1.5.3 (Virtanen *et al*., 2020), and Scikit-posthocs v.0.6.4 (Terpilowski, 2019).

Supervised classification

We attempted to classify species' CAM phenotypes based on anatomy using the supervised learning method gradient boosting, implemented with XGBoost via the Python package XGBOOST v.1.5.0 (Chen & Guestrin, 2016). XGBoost implements gradient tree boosting algorithms (Friedman *et al*., 2000; Friedman, 2001) that use greedy learning over an ensemble of regression trees to train classification models. XGBoost is rare in that it can accept observations with missing values without the need for data imputation. We conducted multiclass classification of non-CAM, mCAM, and pCAM taxa and a simpler, binary classification of non-CAM and CAM taxa, where mCAM and pCAM taxa were combined. We explored a variety of alternative parameterizations: changing the default booster (gbtree) to DART (Rashmi & Gilad-Bachrach, 2015), which can reduce overfitting by randomly dropping decision trees; changing the objective function (softmax or softprob for multiclass classification; logistic probability, logistic raw score, or hinge loss for binary classification); and changing the evaluation metric (multiclass logloss, AUC, or multiclass error rate for multiclass classification; error rate for binary classification; the AUC evaluation metric required a softprob objective function). In all cases, we randomly divided our data set between training (80%) and testing (20%).

We also tried several strategies to reduce the effects of highly imbalanced classes and sparsity. We attempted to reduce class imbalance by adjusting the parameter 'max_delta_step' (MDS), by random over- or under-sampling our training data, and by merging mCAM and pCAM into a binary classification model. Increasing MDS above its default (0) creates an additional penalty that reduces splitting within trees, or the addition of trees entirely, in highly imbalanced data sets. Random oversampling (ROS) resamples minority classes until all class labels are equal (augmenting training data), while random under-sampling (RUS) subsamples classes until all class labels are equal (reducing training data). Our data were also quite sparse (67% missing data) because we merged data from largely nonoverlapping studies. We evaluated three data imputation strategies: median (missing features were imputed with the median), iterative (missing features were imputed by regression of present features), and *K*-nearest neighbors (Knn; missing features were imputed using the nearest neighbors in a Knn embedding).

Phylogenetic tree inference

The Portullugo, the clade inclusive of the Portulacineae (families Anacampserotaceae, Basellaceae, Cactaceae, Didiereaceae, Montiaceae, Portulacaceae, and Talinaceae) and its sister clade (Molluginaceae) is well-suited for large, comparative phylogenetic studies because of recent sequence data, its diversity of CAM phenotypes, and the overlap between anatomical data and extant phylogenies. We constructed a new phylogeny of the Portullugo by merging two previously published sequence matrices that were obtained using different techniques. The first dataset consisted of 841 loci from transcriptomic data used to study the evolution of Portulacineae and its adaptation to harsh environments (Wang *et al*., 2019). The second dataset was a targeted enrichment of 83 gene families, primarily with roles in plant respiration and photosynthesis (Goolsby *et al*., 2018; Hancock *et al*., 2018; Moore *et al*., 2018). To find common loci between the datasets, we independently called consensus sequences for each locus and mapped them against the sugar beet (*Beta vulgaris* ssp. *vulgaris* L.) genome assembly v.EL10_1.0 (McGrath *et al*., 2022) using BLAST v.2.13.0 (Camacho *et al*., 2009). Mapping consensus sequences for each locus proved more accurate than using random representative sequences for a given locus due to high sequence variation. If consensus loci hit multiple reference scaffolds, we retained the reference locus with the highest bitscore. We used the resulting mapping coordinates to search for potential overlapping loci between datasets and aligned them using MAFFT v.7.508 (Katoh & Standley, 2013). Loci showing considerable dataset overlap were concatenated to create an initial matrix of loci represented by both datasets, and then flanked with seven randomly selected nonoverlapping loci from each dataset to increase the number of taxa included and overall matrix size.

We concatenated all loci and constructed a maximum likelihood-based tree using IQ-TREE v.2.2.0.3 (Minh *et al*., 2020). Within IQ-TREE, a model of sequence evolution was selected using the automated model finder (Kalyaanamoorthy *et al*., 2017) constrained to the GTR family of models; node support was assessed using ultrafast bootstrap approximation (Hoang *et al*., 2017). We time-calibrated the tree using the fast least squares dating method (To *et al*., 2016) included in IQ-TREE using the entire concatenated sequence matrix and 13 secondary node calibrations used by Wang *et al*. (2019) from Arakaki *et al*. (2011; Table S4); the topology was constrained to that produced during our initial run of IQ-TREE. Confidence intervals were inferred from 100 resamplings of branch lengths by drawing new clock rates (log-normal distribution with mean 1 and standard deviation 0.2), tip dates were set to 0, and a GTR + F substitution model was selected with the automated model finder.

Phylogenetic trait analyses

We reconstructed the evolutionary history of CAM in the Portullugo using stochastic character mapping (Nielsen, 2002; Huelsenbeck *et al*., 2003) implemented with the 'make.simmap' function of the R package 'PHYTOOLS' v.1.2-0 (Revell, 2012). We modeled CAM evolutionary history assuming (1) an all rates different (ARD) model, and (2) a constrained ARD model without reversions from pCAM to mCAM; both models assumed a root state of non-CAM. The constraint of the latter model was informed by the lack of evidence for reversions from pCAM throughout vascular plants. In all analyses, we pruned our tree to one sample per species and node reconstructions were visualized as pie charts summarizing the state frequencies over 10 000 stochastic maps.

To assess the relationships between CAM phenotypes and anatomical traits in the Portullugo, we used a threshold model of trait evolution (Wright, 1934; Felsenstein, 2005), implemented with the 'threshBayes' function (Revell, 2014) of 'PHYTOOLS' v.1.2-0, and phylogenetic least squares (PGLS) regression

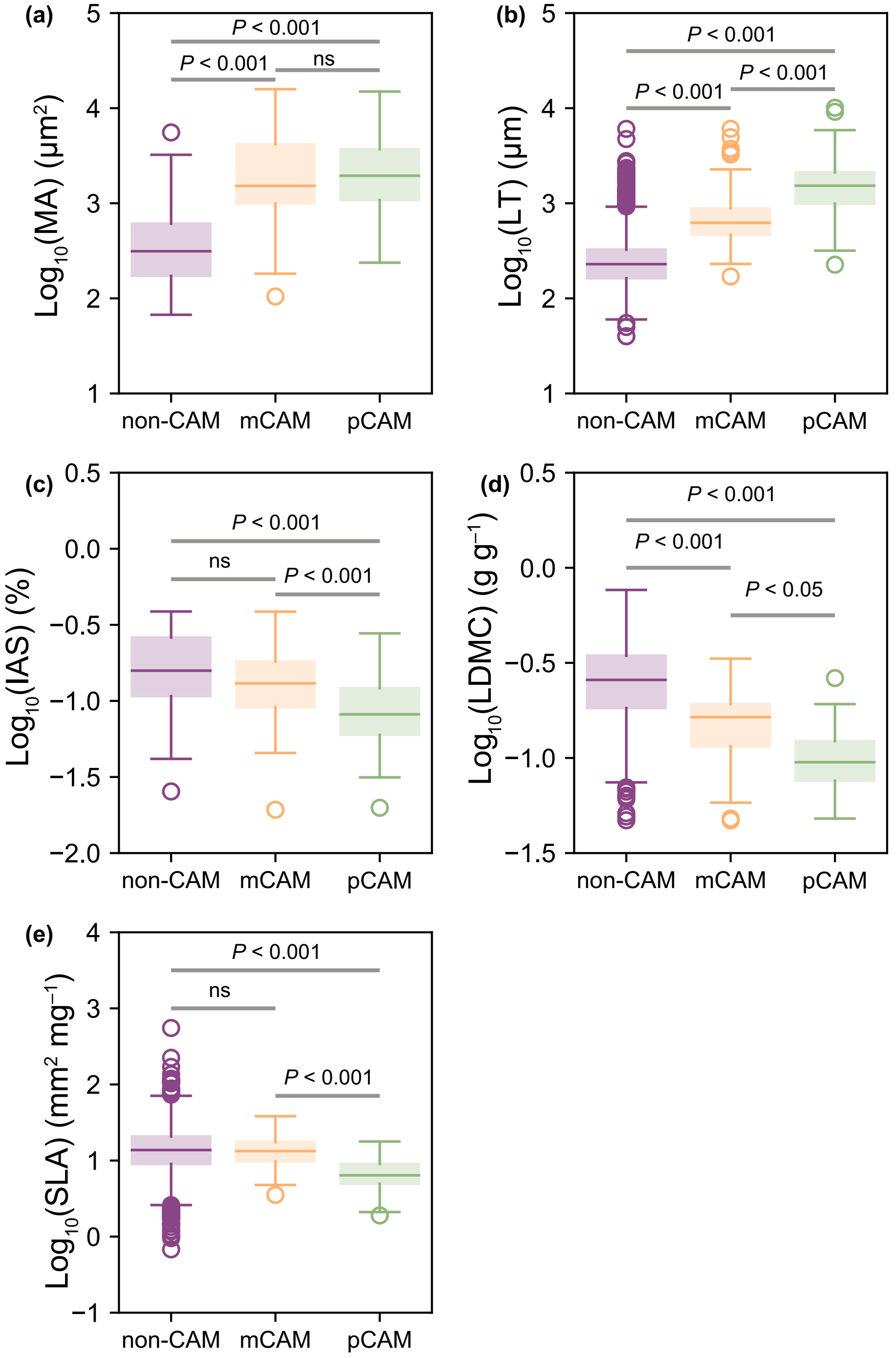

Fig. 2 Results of Dunn's *post hoc* tests for group differences between $\log_{10}$-transformed features: (a) mesophyll cell area (MA); (b) leaf thickness (LT); (c) intercellular airspace (IAS); (d) leaf dry matter content (LDMC); (e) specific leaf area (SLA). Purple, yellow, and green box-and-whisker plots show non-Crassulacean acid metabolism (non-CAM), minority CAM (mCAM), and primary CAM (pCAM) trait distributions; boxes represent the interquartile range (IQR) with a line representing the median, whiskers showing 1.5× the IQR, and points outside considered outliers; ns, nonsignificant.

(Grafen, 1989), implemented with the R package 'NLME' v.3.1-162 (Pinheiro *et al*., 2023). We used PGLS regression to assess relationships between continuous anatomical traits and between anatomical traits and discrete CAM phenotypes (as a predictor variable). We used threshold models to measure the correlations between anatomical traits and CAM phenotype. In all analyses, our tree was pruned to match the taxa with anatomical data and reduced to one sample per taxon where necessary. Each MCMC sampler for threshold analyses was run for 5000 000 steps and posterior distributions were constructed after discarding the initial 20% of samples as burn-in.

Results

Nonphylogenetic analyses of anatomy across angiosperms demonstrated significant group differences for all five anatomical features investigated (Table S5). Dunn's *post hoc* tests identified significant ($P < 0.05$), and generally consistent, differences between CAM phenotypes for most features: the largest differences were observed between non-CAM and pCAM phenotypes, with mCAM intermediate but not always significantly different from both non-CAM and pCAM (Fig. 2). Where sufficient data were available, these trends were supported within individual families (Fig. S3). We found significant negative correlations between MA and LDMC, between LT and SLA, LDMC, and IAS, and between LDMC and SLA; a significant positive correlation was found between LT and MA (Fig. S4).

Multiclass classification with XGBoost yielded similar results regardless of evaluation metric or objective function, with booster choice being the only source of variation (Fig. S5). Because of the similarity of those results, we only continued using models with

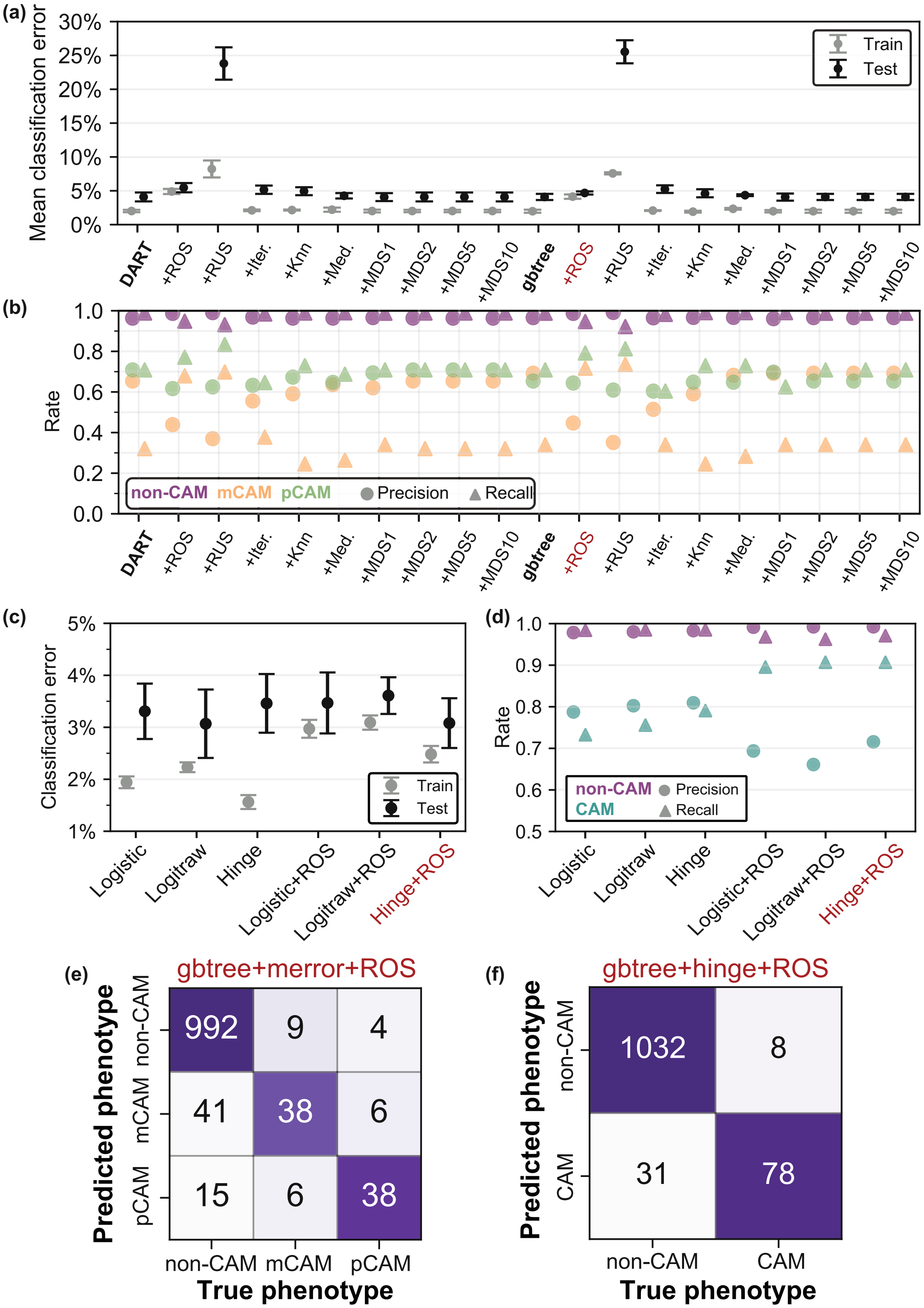

Fig. 3 Machine learning model accuracies. Mean classification error (a, c), precision and recall rates (b, d), and best-performing model confusion matrices (e, f) for multiclass and binary classifiers. Multiclass models (a, b) varied in booster (DART or gbtree), sampling strategy (random oversampling (ROS) or random under-sampling (RUS)), imputation method (iterative, Knn, or median), and MDS (1, 2, 5, or 10); binary models (c, d) varied in objective function (logistic, logitraw, or hinge) and sampling strategy (with or without ROS). Error bars in (a, c) represent standard error of the mean. The columns of each confusion matrix (e, f) show the number of true Crassulacean acid metabolism (CAM) phenotypes in the test data set and the rows show the model predictions. The diagonal in each matrix represents correct model predictions and off-diagonal elements show incorrect predictions; for example, a true pCAM species predicted to be non-CAM would be shown in the first row, the third column of (c). Knn, *K*-nearest neighbors; MDS, max_delta_step; mCAM, minority CAM; pCAM, primary CAM. Base models are in bolded text and the best-performing models are highlighted in red.

the softprob objective function and multiclass error rate (merror) evaluation metric (hereafter, DART and gbtree 'base models'). The two base models had similar cross-validation test accuracies (96.0 ± 1.1%; Fig. 3a), precision and recall of non-CAM, mCAM, and pCAM (Fig. 3b), and feature importance rankings (LT > MA > IAS ≥ LDMC > SLA; Fig. S6; Table S6). No imbalance-reduction sampling, imputation method, or alternative parameterization increased overall accuracy (Fig. 3a); however, ROS and RUS increased recall for mCAM and pCAM taxa (Fig. 3b). Between models of similar accuracy, we prioritized improving mCAM recall (also known as sensitivity in binary classification; true positives/true positives + false negatives) because true negative rates of mCAM are not well known in most CAM-evolving clades. While decreased non-CAM classification accuracy slightly decreased overall model accuracy, ROS raised recall rates of mCAM and pCAM classification to 70% and > 75%, respectively. Although RUS further increased mCAM and pCAM recall (Fig. 3b), the substantial difference between training and testing accuracy (Fig. 3a) suggested that these models were overfit. To further address class imbalance, we combined mCAM and pCAM into a single 'CAM' category and attempted binary classification. Binary classification models had similar test accuracies (Fig. 3c), but the hinge objective function yielded slightly higher CAM precision and recall. As in multiclass classification, ROS greatly increased CAM recall, but the F1-score (2 × precision × recall/precision + recall) remained unchanged because of an equal magnitude drop in precision (Fig. 3d).

Our preferred multiclass and binary classifiers both used gbtree boosters and ROS, and the hinge objective function for binary classification (Fig. 3e,f). Mean cross-validation accuracies were 95.7 ± 0.7% and 96.1 ± 0.6% for multiclass and binary models, respectively (Fig. 3a,c). Most non-CAM taxa incorrectly classified by multiclass models belonged to clades with diverse CAM phenotypes (e.g. Bromeliaceae and Orchidaceae subfamily Epidendroideae), and mCAM taxa were roughly equally classified as non-CAM or pCAM (Fig. 3e; Table S7). Similarly, most incorrect predictions by the binary model were non-CAM species from CAM-evolving lineages classified as CAM (Fig. 3e; Table S8); generally, these taxa have not been thoroughly assessed for mCAM, and so it is possible that they may actually have a facultative or very weak CAM cycle.

Our time-calibrated species tree was mostly congruent with those from which the data were compiled. The 77 transcriptome-based samples (representing 77 unique species) and the 175 target enrichment-based samples (144 unique species) had only 16 species in common, and thus, their combination greatly expanded sampling throughout the Portullugo. Of those species sampled in both datasets, 13 were monophyletic in the final tree (Fig. S7); with samples of *Alluaudia dumosa*, *A. procera*, and *Calyptridium umbellatum* recovered as more closely related to other samples within their respective datasets. Support was generally high, although multiple nodes along the backbone were unresolved and left as polytomies in downstream analyses (Figs 4, S7). Stochastic character map reconstructions of CAM evolution suggested that mCAM evolved at the base of the Portulacineae, and that multiple transitions to pCAM occurred in the Cactaceae and Didiereaceae, while multiple reversions to non-CAM occurred in the Montiaceae (Fig. 4). Though similar, we preferred a constrained ARD model of CAM evolution (Figs 4, S8) to an unconstrained model (Fig. S9) because there is no strong empirical evidence of reversions from pCAM in any vascular plant lineage.

Significant phylogenetic signal was detected in all three traits measured across the Portullugo (Table S9). Phylogenetic least squares regression revealed multiple significant ($P < 0.05$) relationships among anatomical traits and between anatomical traits and CAM phenotype (Fig. 5; Table S10). However, AIC-based model selection favored a model between MA and IAS with a non-significant slope, contrary to our expectation that greater mesophyll cell size would lead to lower IAS (Fig. 5a). Greater MA was a significant ($P < 0.0001$) predictor of greater LT (Fig. 5b), and we found no relationship between IAS and LT (Fig. 5c). CAM phenotype was a significant predictor of MA, LT, and IAS (Fig. 5d–f). We next used phylogenetic threshold analyses to estimate the correlations between CAM phenotype and anatomical traits under the hypothesis that there may be anatomical boundaries between CAM phenotypes. Threshold analyses mostly supported PGLS results, and recovered significant positive correlations between CAM phenotype and both MA and LT (Fig. 6a,b). However, the posterior distribution of correlation coefficients between CAM phenotype and IAS narrowly included 0 (Fig. 6c).

Discussion

From the beaks of Galapagos finches (Darwin, 1839) to unique inflorescence architectures (Waal *et al.*, 2012), the links between form and function have always inspired biologists. Fixed in place, with passive mechanisms for carbon and water acquisition, plants rely on anatomical innovations to adapt to different environments. Succulence has long been understood as a drought avoidance adaptation, but its relationship with CAM has not been resolved as causal or merely coincidental. Through our broad survey of angiosperms and detailed study of the Portullugo, we found support for previous hypotheses of CAM and photosynthetic tissue anatomy co-evolution. Furthermore, we demonstrate that the presence or absence of CAM may be predicted using only a handful of anatomical measurements.

Anatomical measurements from over 200 angiosperm families revealed significant differences in photosynthetic tissue anatomy

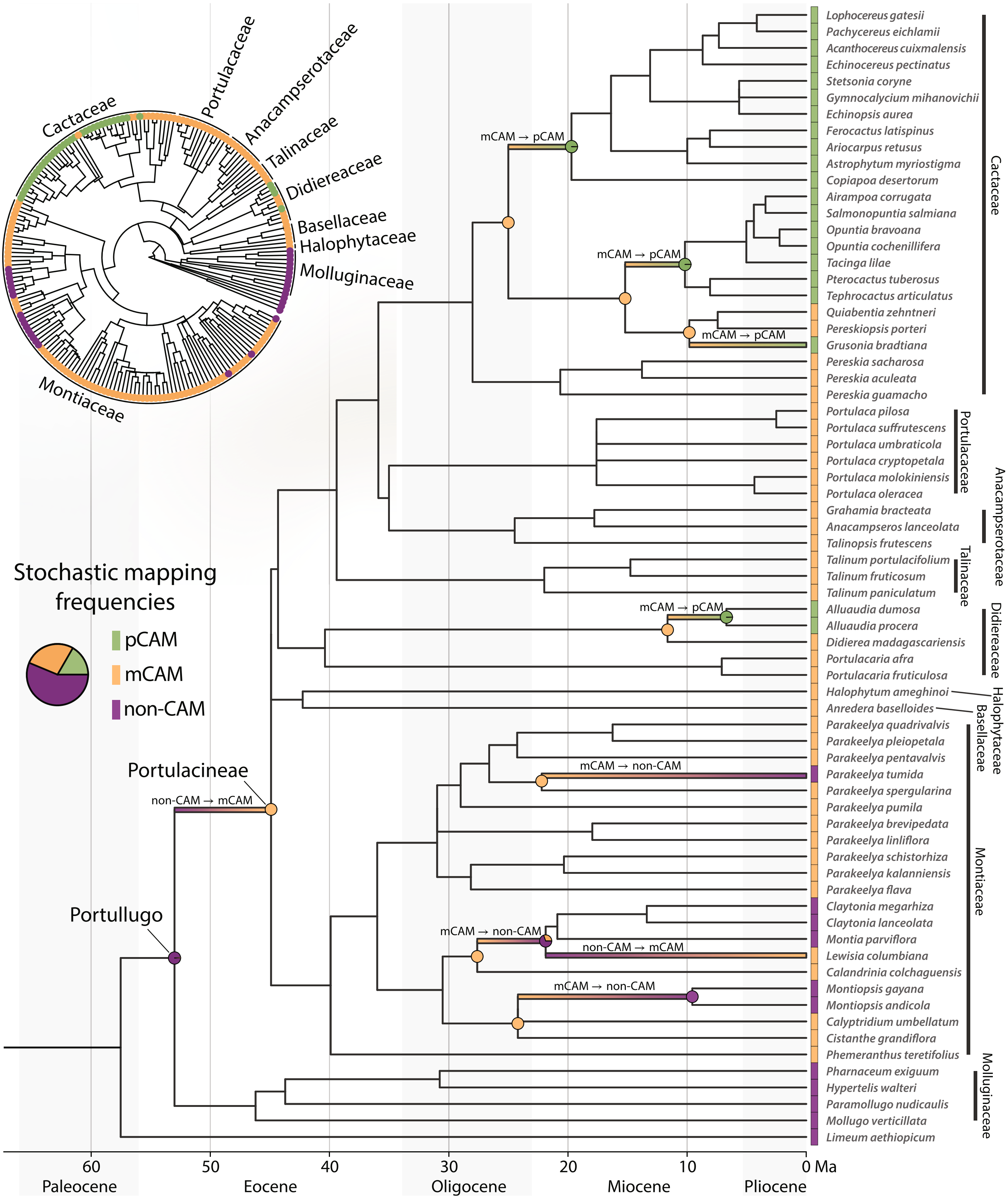

Fig. 4 Time-calibrated phylogeny of the Portullugo with inferred transitions between Crassulacean acid metabolism (CAM) phenotypes. The Portullugo and Portulacineae nodes are highlighted, and color gradients indicate transitions between non-CAM (purple), mCAM (yellow), and pCAM (green) based on the results of our biologically informed ancestral state reconstruction. Pie charts at nodes bracketing inferred transitions show the fractions of stochastic maps supporting each ancestral state. This tree has been pruned to show only those taxa with morphological data used in this study and therefore not all transitions are shown; the full tree is shown in the inset and multiple ancestral state reconstructions are available in the Supporting Information.

of non-CAM, mCAM, and pCAM species. The larger mesophyll cell size of both mCAM and pCAM species suggests some anatomical specialization is required to perform CAM in any capacity, and the reduction in intercellular airspace of pCAM species indicates that further specialization is required to use CAM for primary carbon metabolism. We also showed significant increases in LT and decreases in LDMC from non-CAM to mCAM to pCAM, as well as significantly lower SLA in pCAM species, which support past findings that thicker and more succulent leaves are positively associated with strong CAM activity within individual clades (Teeri *et al*., 1981; Winter *et al*., 1983; Nelson *et al*., 2005; Nelson & Sage, 2008; Zambrano *et al*., 2014; Luján *et al*., 2022).

Because lineage-specific organismal detail will surely influence physiology–anatomy relationships, analyses of anatomy and CAM evolution are best evaluated using phylogenetic comparative methods. PGLS regression and phylogenetic threshold analysis supported the correlated evolution of larger mesophyll cells and thicker leaves. Although PGLS regression further showed a continuous decrease in IAS from non-CAM to pCAM species, we found no significant relationship between IAS and MA. That IAS and MA may evolve independently of one another provides an important nuance to the co-evolution of succulence and CAM. Decreased IAS in CAM species has often been discussed as an adaptation to reduce CO_2 efflux during malate decarboxylation (Nelson & Sage, 2008) or as a consequence of increased succulence restricting g_m, which would limit CO_2 fixation by Rubisco during the day (Maxwell *et al*., 1997; Zambrano *et al*., 2014; Earles *et al*., 2018; Edwards, 2019). More recently, reduced IAS has been hypothesized to be an indirect consequence of increased mesophyll cell volume used for malic acid storage (Leverett *et al*., 2023). While we found that succulence generally increased with CAM evolution, the decoupling of the underlying traits may allow the evolution of intermediate photosynthetic and anatomical phenotypes that efficiently utilize both CAM and C_3 or C_4 photosynthesis. These conclusions are consistent with photosynthetic models that found increased vacuolar volume (and therefore MA) necessary for CAM (Töpfer *et al*., 2020) and empirical findings that the high IAS in mCAM species may allow for C_3 (or C_4) photosynthesis when plants are not engaging CAM (Nelson & Sage, 2008; Zambrano *et al*., 2014). Furthermore, the lowest IAS values in the Portullugo were observed in pCAM species, which reinforces the hypothesis that extremely low IAS may reduce g_m and thus C_3 or C_4 efficiency (Maxwell *et al*., 1997).

In addition to providing support for a positive relationship between CAM and succulence, our findings point toward interactions between life history, CAM, and succulence for those taxa that do not neatly fall along regression lines. Phylogenetic analyses of the Portullugo showed general increases in succulence and a tightening of the distributions of underlying traits for pCAM species. By contrast, mCAM taxa had both the single largest MA and greatest IAS observations, with values that mostly spanned the non-CAM to pCAM range. The eight largest observed MA values in the Portullugo were from mCAM species; most are annual species, with the exceptions of *Parakeelya flava* (a perennial geophyte with aboveground tissues that regrow annually) and *Grahamia bracteata* and *Talinopsis frutescens* (which have nonsucculent woody stems and drought-deciduous leaves). This suggests that the evolution of pCAM requires a shift to a (functional) perennial life history with long-lived photosynthetic tissues (Hancock *et al*., 2019); indeed, we are unaware of any annual pCAM species. The halophyte *Halophytum ameghinoi* had the second-largest observed MA in the Portullugo. While saline soils may select for increased succulence to maintain cytosolic ion balance (Naidoo & Rughunanan, 1990; Ogburn & Edwards, 2010), high salt concentrations inhibit the central CAM enzymes phosphoenolpyruvate carboxylase (PEPC) and malic enzyme (ME; Kluge & Ting, 1978), and may therefore represent an ecological constraint on the evolution of pCAM.

Our ancestral state reconstruction of CAM in the Portullugo was the first to model CAM as an ordered multistate trait, and supported an early- to mid-Eocene origin of mCAM – a time when the Earth's atmosphere had relatively high levels of CO_2 (Rae *et al*., 2021). The reconstruction of mCAM at the crown of the Portulacineae agrees with transcriptomic data that suggest a single recruitment event of a PEPC ortholog for use in CAM (Christin *et al*., 2014; Goolsby *et al*., 2018). All transitions to pCAM were found be within the past 30 Ma (Sage *et al*., 2023), congruent with shifts across angiosperms (including within Caryophyllales) to C_4 photosynthesis, as atmospheric CO_2 fell below 500 ppm (Christin *et al*., 2011). Despite declining CO_2 in the Oligocene and Miocene, multiple lineages within the Montiaceae have lost the ability perform CAM. Although we expect more Montiaceae lineages to exhibit CAM upon experimentation, multiple independent losses of CAM have been experimentally validated in *Parakeelya* (Hancock *et al*., 2019), a clade endemic to hot, dry areas of Australia. While life history may constrain the evolution of pCAM, it remains unclear why some members of the Portulacineae transitioned to C_3 photosynthesis, while others simultaneously transitioned to pCAM, as CO_2 continued to decline. We suspect that these losses of CAM may be linked to shifts in phenology; for example, C_3 *Parakeelya* grow in either monsoonal areas with abundant moisture, or in cooler regions with lower growing season temperatures.

Most clades with CAM lineages show highly bimodal distributions of carbon isotope ratios (Messerschmid *et al*., 2021) that have been used for decades to identify pCAM species, but are

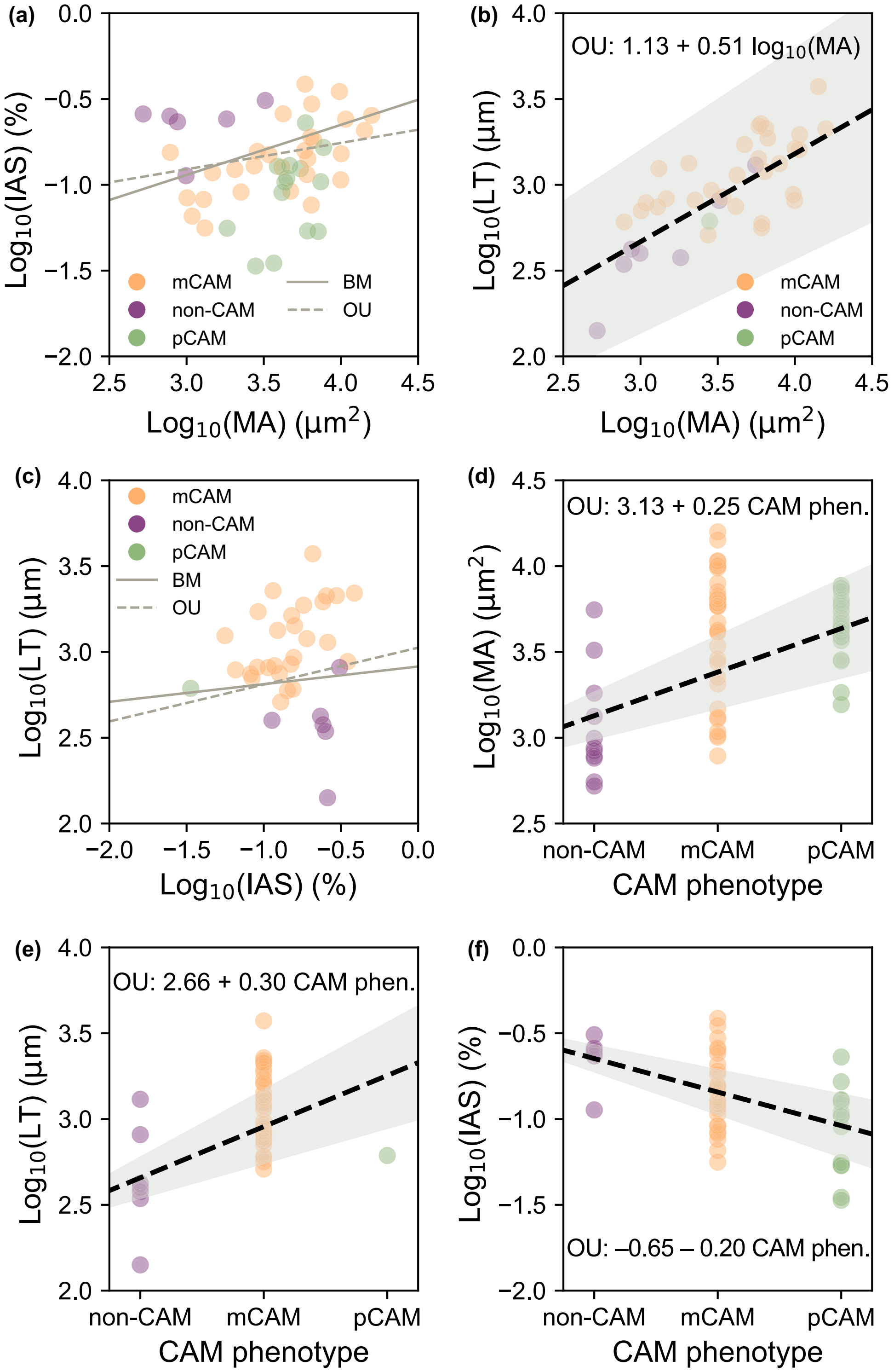

Fig. 5 Results of phylogenetic least squares regressions. Fitted regression results between anatomical features (a–c) and between anatomical features and CAM phenotype (d–f). Predictor and response variables are shown on the horizontal and vertical axes, respectively. Points show trait values for non-Crassulacean acid metabolism (non-CAM; purple), minority CAM (mCAM; yellow), and primary CAM (pCAM; green) species. Solid and dashed grey lines show the fitted regression lines using Brownian motion (BM) and Ornstein–Uhlenbeck (OU) models of trait evolution, respectively. Significant best-fit relationships are shown with bold black lines, associated model coefficients, and grey shading to show standard error. IAS, intercellular air-space; LT, leaf thickness; MA, mesophyll cell area.

generally unable to distinguish mCAM from non-CAM. Laborious controlled experiments (e.g. of gas exchange or malic acid content) with live plants have been the only ways to identify mCAM, but such experiments are not feasible for many long-lived, rare, or difficult-to-cultivate species. We found that differences in photosynthetic anatomy across angiosperms translated into moderate to high accuracy in predicting CAM phenotype. After assessing a variety of machine-learning models, we found that ROS increased the recall of mCAM and pCAM species while not overfitting to training data. To our knowledge, machine learning has not yet been applied to predict the presence or absence of physiological traits from anatomical measurements, such as CAM phenotypes. We believe that the accuracy we obtained represents a lower bound on the true accuracy of our

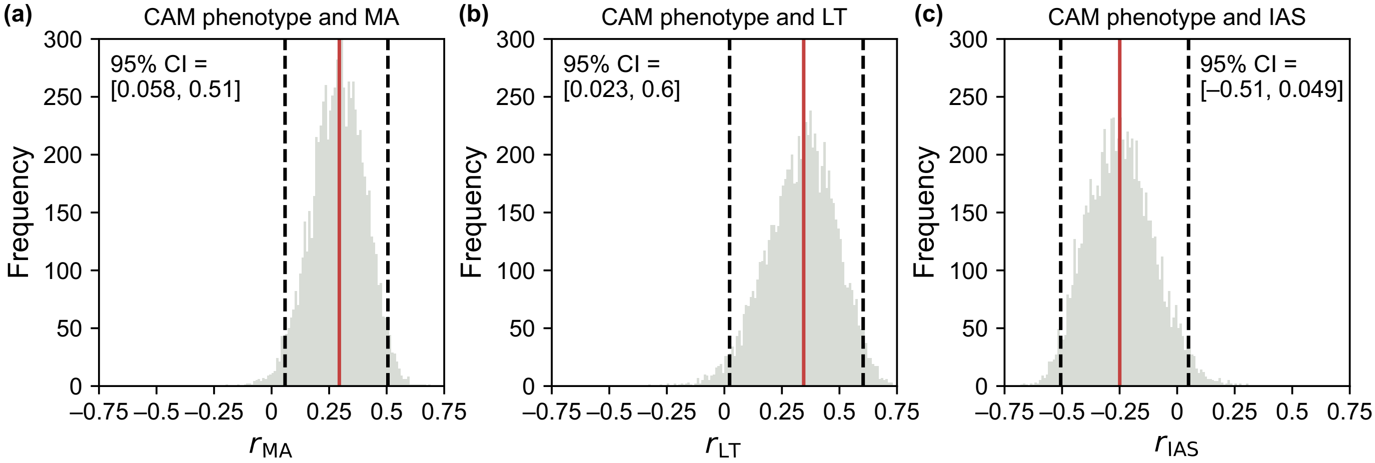

Fig. 6 Phylogenetic threshold model correlations. The distribution of correlation coefficients (*r*) between Crassulacean acid metabolism (CAM) phenotype and (a) $\log_{10}$-transformed mesophyll cell area (MA), (b) intercellular airspace (IAS), and (c) leaf thickness (LT). The grey histograms show the frequency of *r*-values visited by the MCMC sampler following a 20% burn-in period, red lines show the median *r*-values, and dashed black lines show the 95% credible interval (CI).

models because some of our potentially misclassified non-CAM species have not been thoroughly investigated for mCAM. For example, multiple orchid and bromeliad species labeled as non-CAM were predicted to be mCAM, but have not been subjected to drought experiments that might induce CAM activity. We predict that some misclassified species, such as *Nolina bigelovii* (Asparagaceae) will exhibit CAM upon experimentation, resulting in more true positive predictions. Experimentation should continue to be the gold standard for determining CAM phenotypes, but machine learning models, such as those developed here, could play a valuable role in prioritizing study species and would only require small tissue sections for initial fixation and measurement. More broadly, the expansion of online databases, such as the Royal Botanic Gardens' Microscope Slide Collection (https://www.kew.org/science/collections-and-resources/collections/microscope-slide-collection), in combination with machine learning tools for analyzing herbarium specimens (e.g. Wilde *et al.*, 2023), offer new opportunities for testing evolutionary hypotheses about form and function from the cellular- to whole plant-levels.

Applications of machine learning in plant physiology and evolution are only just beginning. Machine learning has been successful in predicting real-time photosynthetic status; for example, deep learning using hyperspectral reflectance in wheat has been used to predict electron transport rate, CO_2 assimilate rate, stomatal conductance, and more (Furbank *et al.*, 2021). Our machine learning models were limited in several ways; perhaps most by the degree of missing data and class imbalance. Our greatest model improvements came when using ROS, suggesting that measuring new mCAM and pCAM species to reduce class imbalance will increase model accuracy. If missing data could be sufficiently reduced, imputation strategies may facilitate the use of models beyond XGBoost, which allows missing data. In addition to our machine learning models, we hope that the tools and methodology developed here for measuring anatomy and merging sequence matrices will facilitate future studies of anatomical evolution. Although software exists for taking measurements from images (e.g. IMAGEJ; Schneider *et al.*, 2012, which we used for portions of this study), making dozens or hundreds of measurements needed for phylogenetic studies remains time consuming and the results are not easily reproducible. Our image segmentation software, MINICONTOURFINDER, can be automated from the command line, quickly segment and measure image features, and record associated metadata so exact measurements can be reproduced. Finally, our strategy for combining sequencing data types into a single phylogenetic analysis is flexible and in theory adaptable to any sequencing strategy. Most clades have reference, or near-reference, quality genomes within *c.* 75 Ma of their focal taxa (as in this study; Cheng *et al.*, 2018) that can serve as common maps to identify overlapping genomic regions, and high-quality transcriptomes (Matasci *et al.*, 2014; Leebens-Mack *et al.*, 2019) or targeted sequencing data (Johnson *et al.*, 2019) for constructing backbones in larger phylogenies.

In conclusion, with a broad sampling of anatomical traits from thousands of angiosperms and a detailed phylogenetic study of the Portullugo clade, we provided support for hypotheses of CAM anatomical evolution. Our findings suggest that even weakly expressed CAM is correlated with larger mesophyll cells, and that decreased intercellular airspace in photosynthetic tissue is associated with a transition to using CAM as the primary carbon fixation pathway. Furthermore, our findings point toward possible evolutionary constraints on pCAM evolution, such as annual life history. We were able predict CAM phenotypes from a handful of anatomical features, which represents a successful first application of machine learning to this problem, but also highlights the paucity of anatomical data for species capable of weak or facultative CAM. As data accumulate, we hope that these correlations will be continuously evaluated across vascular plants with tools that may allow causal evolutionary inference, such as phylogenetic path analysis (von Hardenberg & Gonzalez-Voyer, 2013). We expect that efforts to quantify key anatomical parameters for a diversity of CAM phenotypes will more sharply delineate the anatomical requirements of even a weak CAM cycle and demonstrate the anatomical and biochemical interplay during the evolutionary transition to a pCAM physiology.

Acknowledgements

We would like to thank Dr Jonathan Koss for advice on implementing computer vision algorithms, Dr Elena Voznesenskaya for providing images of *Portulaca* species, Dr Eric Lazo-Wasem, and Lourdes Rojas for help imaging specimens, and Joni Ward, Rual Puente-Martinez, and the rest of the staff at the Desert Botanical Garden (Phoenix, AZ) for assistance with the living collections that contributed to this manuscript. This research was funded by the National Science Foundation (IOS-1754662 to EJE).

Competing interests

None declared.

Author contributions

ISG, KH and EJE designed the research plan. ISG, KH and LPH collected, fixed and imaged specimens. ISG developed image segmentation software, curated anatomical data, time-calibrated the phylogeny, conducted statistical analyses, and wrote the first draft of the manuscript. CM-L designed the sequence merging strategy, generated sequence matrices, and constructed the initial phylogeny. All authors contributed to editing and revising the manuscript.

ORCID

Erika J. Edwards https://orcid.org/0000-0003-0515-2778
Ian S. Gilman https://orcid.org/0000-0002-0390-9370
Lillian P. Hancock https://orcid.org/0000-0002-1394-4970
Karolina Heyduk https://orcid.org/0000-0002-1429-6397
Carlos Maya-Lastra https://orcid.org/0000-0002-0550-3331

Data availability

Additional Supporting Information may be found online in the supporting information section at the end of the article. All statistical analyses of this study can be found at https://github.com/isgilman/Predicting-CAM and all installation and documentation for MiniContourFinder can be found at https://minicontourfinder.readthedocs.io/en/latest/. Raw images generated for this manuscript are available upon request.

References

Arakaki M, Christin P-A, Nyffeler R, Lendel A, Eggli U, Ogburn RM, Spriggs E, Moore MJ, Edwards EJ. 2011. Contemporaneous and recent radiations of the world's major succulent plant lineages. *Proceedings of the National Academy of Sciences USA* **108**: 8379–8384.

Bartlett MS. 1937. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences* **160**: 268–282.

Bradski G. 2000. *The OpenCV library.* Dr. Dobb's J. Software Tools.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. Blast+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Chen T, Guestrin C. 2016. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York, NY, USA: Association for Computing Machinery, 785–794.

Cheng S, Melkonian M, Smith SA, Brockington S, Archibald JM, Delaux P-M, Li F-W, Melkonian B, Mavrodiev EV, Sun W *et al.* 2018. 10KP: a phylodiverse genome sequencing plan. *GigaScience* **7**: 1–9.

Christin P-A, Arakaki M, Osborne CP, Bräutigam A, Sage RF, Hibberd JM, Kelly S, Covshoff S, Wong GK-S, Hancock L *et al.* 2014. Shared origins of a key enzyme during the evolution of C_4 and CAM metabolism. *Journal of Experimental Botany* **65**: 3609–3621.

Christin P-A, Osborne CP, Sage RF, Arakaki M, Edwards EJ. 2011. C_4 eudicots are not younger than C_4 monocots. *Journal of Experimental Botany* **62**: 3171–3181.

Cousins AB, Mullendore DL, Sonawane BV. 2020. Recent developments in mesophyll conductance in C_3, C_4, and Crassulacean acid metabolism plants. *The Plant Journal* **101**: 816–830.

Coutinho LM. 1969. Novas observações sobre a ocorrência do "efeito de de Sassure" e suas relações com a suculência, a temperatura folhear e os movimentos estomáticos. *Boletim Da Faculdade De Filosofia Ciências E Letras Universidade De São Paulo Botânica* **24**: 77–102.

D'Agostino R, Pearson ES. 1973. Tests for departure from normality. Empirical results for the distributions of b2 and $\sqrt{b}1$. *Biometrika* **60**: 613–622.

Darwin C. 1839. *The voyage of the beagle.* New York, NY, USA: Appleton & Co.

Dunn OJ. 1964. Multiple comparisons using rank sums. *Technometrics* **6**: 241–252.

Earles JM, Theroux-Rancourt G, Roddy AB, Gilbert ME, McElrone AJ, Brodersen CR. 2018. Beyond porosity: 3D leaf intercellular airspace traits that impact mesophyll conductance. *Plant Physiology* **178**: 148–162.

Earnshaw MJ, Winter K, Ziegler H, Stichler W, Cruttwell NEG, Kerenga K, Cribb PJ, Wood J, Croft JR, Carver KA *et al.* 1987. Altitudinal changes in the incidence of Crassulacean acid metabolism in vascular epiphytes and related life forms in Papua New Guinea. *Oecologia* **73**: 566–572.

Edwards EJ. 2019. Evolutionary trajectories, accessibility and other metaphors: the case of C_4 and CAM photosynthesis. *New Phytologist* **223**: 1742–1755.

Edwards EJ. 2023. Reconciling continuous and discrete models of C_4 and CAM evolution. *Annals of Botany* mcad125. doi: 10.1093/aob/mcad125.

Felsenstein J. 2005. Using the quantitative genetic threshold model for inferences between and within species. *Philosophical Transactions of the Royal Society, B: Biological Sciences* **360**: 1427–1434.

Fraser LH. 2020. TRY–a plant trait database of databases. *Global Change Biology* **26**: 189–190.

Friedman J, Hastie T, Tibshirani R. 2000. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics* **28**: 337–407.

Friedman JH. 2001. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics* **29**: 1189–1232.

Furbank RT, Silva-Perez V, Evans JR, Condon AG, Estavillo GM, He W, Newman S, Poiré R, Hall A, He Z. 2021. Wheat physiology predictor: predicting physiological traits in wheat from hyperspectral reflectance measurements using deep learning. *Plant Methods* **17**: 108.

Gilman IS, Smith JAC, Holtum JAM, Sage RF, Silvera K, Winter K, Edwards EJ. 2023. The CAM lineages of plant Earth. *Annals of Botany* mcad135. doi: 10.1093/aob/mcad135.

Goolsby EW, Moore AJ, Hancock LP, De VJM, Edwards EJ. 2018. Molecular evolution of key metabolic genes during transitions to C_4 and CAM photosynthesis. *American Journal of Botany* **105**: 602–613.

Grafen A. 1989. The phylogenetic regression. *Philosophical Transactions of the Royal Society, B: Biological Sciences* **326**: 119–157.

Hancock L, Edwards EJ. 2014. Phylogeny and the inference of evolutionary trajectories. *Journal of Experimental Botany* **65**: 3491–3498.

Hancock LP, Holtum JAM, Edwards EJ. 2019. The evolution of CAM photosynthesis in Australian *Calandrinia* reveals lability in C_3+ CAM phenotypes and a possible constraint to the evolution of strong CAM. *Integrative and Comparative Biology* **59**: 517–534.

Hancock LP, Obbens F, Moore AJ, Thiele K, De VJM, West J, Holtum JAM, Edwards EJ. 2018. Phylogeny, evolution, and biogeographic history of *Calandrinia* (Montiaceae). *American Journal of Botany* **105**: 1021–1034.

von Hardenberg A, Gonzalez-Voyer A. 2013. Disentangling evolutionary cause-effect relationships with phylogenetic confirmatory path analysis. *Evolution* **67**: 378–387.

Herrera A. 2020. Are thick leaves, large mesophyll cells and small intercellular air spaces requisites for CAM? *Annals of Botany* **125**: 859–868.

Heyduk K, McKain MR, Lalani F, Leebens-Mack J. 2016. Evolution of a CAM anatomy predates the origins of Crassulacean acid metabolism in the Agavoideae (Asparagaceae). *Molecular Phylogenetics and Evolution* **105**: 102–113.

Heyduk K, Ray JN, Leebens-Mack J. 2020. Leaf anatomy is not correlated to CAM function in a C_3 + CAM hybrid species, *Yucca gloriosa*. *Annals of Botany* **127**: 437–449.

Hoang DT, Chernomor O, Von HA, Minh BQ, Vinh LS. 2017. UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* **35**: 518–522.

Huelsenbeck JP, Nielsen R, Bollback JP. 2003. Stochastic mapping of morphological characters. *Systematic Biology* **52**: 131–158.

Johnson MG, Pokorny L, Dodsworth S, Botigue LR, Cowan RS, Devault A, Eiserhardt WL, Epitawalage N, Forest F, Kim JT *et al.* 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology* **68**: 594–606.

Kalyaanamoorthy S, Minh BQ, Wong TKF, Von HA, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**: 587–589.

Katoh K, Standley DM. 2013. Mafft multiple sequence alignment software v.7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

Kluge M, Ting IP. 1978. *Crassulacean acid metabolism, analysis of an ecological adaptation.* Berlin, Germany: Springer-Verlag.

Kruskal WH, Wallis WA. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* **47**: 583–621.

Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, Grosse I, Li Z, Melkonian M, Mirarab S *et al.* 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**: 679–685.

Leverett A, Borland AM, Inge EJ, Hartzell S. 2023. Low internal air space in plants with crassulacean acid metabolism may be an anatomical spandrel. *Annals of Botany* mcad109. doi: 10.1093/aob/mcad109.

Luján M, Oleas NH, Winter K. 2022. Evolutionary history of CAM photosynthesis in Neotropical *Clusia*: insights from genomics, anatomy, physiology and climate. *Botanical Journal of the Linnean Society* **199**: 538–556.

Males J. 2017. Secrets of succulence. *Journal of Experimental Botany* **68**: 2121–2134.

Males J. 2018. Concerted anatomical change associated with Crassulacean acid metabolism in the Bromeliaceae. *Functional Plant Biology* **45**: 681–695.

Matasci N, Hung L-H, Yan Z, Carpenter EJ, Wickett NJ, Mirarab S, Nguyen N, Warnow T, Ayyampalayam S, Barker M *et al.* 2014. Data access for the 1000 Plants (1KP) project. *GigaScience* **3**: 17.

Maxwell K, Von CS, Evans JR. 1997. Is a low internal conductance to CO_2 diffusion a consequence of succulence in plants with Crassulacean acid metabolism? *Australian Journal of Plant Physiology* **24**: 777–786.

McGrath JM, Funk A, Galewski P, Ou S, Townsend B, Davenport K, Daligault H, Johnson S, Lee J, Hastie A *et al.* 2022. A contiguous *de novo* genome assembly of sugar beet EL10 (*Beta vulgaris* L.). *DNA Research* **30**: dsac033.

Messerschmid TFE, Wehling J, Bobon N, Kahmen A, Klak C, Los JA, Nelson DB, Santos P, Vos JM, Kadereit G. 2021. Carbon isotope composition of plant photosynthetic tissues reflects a Crassulacean acid metabolism (CAM) continuum in the majority of CAM lineages. *Perspectives in Plant Ecology, Evolution and Systematics* **51**: 125619.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von HA, Lanfear R. 2020. IQ-Tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* **37**: 1530–1534.

Moore AJ, Vos JMD, Hancock LP, Goolsby E, Edwards EJ. 2018. Targeted enrichment of large gene families for phylogenetic inference: phylogeny and molecular evolution of photosynthesis genes in the Portullugo Clade (Caryophyllales). *Systematic Biology* **67**: 367–383.

Naidoo G, Rughunanan R. 1990. Salt tolerance in the succulent, coastal halophyte, *Sarcocornia natalensis. Journal of Experimental Botany* **41**: 497–502.

Nelson EA, Sage RF. 2008. Functional constraints of CAM leaf anatomy: tight cell packing is associated with increased CAM function across a gradient of CAM expression. *Journal of Experimental Botany* **59**: 1841–1850.

Nelson EA, Sage TL, Sage RF. 2005. Functional leaf anatomy of plants with Crassulacean acid metabolism. *Functional Plant Biology* **32**: 409–419.

Nielsen R. 2002. Mapping mutations on phylogenies. *Systematic Biology* **51**: 729–739.

Ocampo G, Koteyeva NK, Voznesenskaya EV, Edwards GE, Sage TL, Sage RF, Columbus JT. 2013. Evolution of leaf anatomy and photosynthetic pathways in Portulacaceae. *American Journal of Botany* **100**: 2388–2402.

Ogburn RM, Edwards EJ. 2010. *Advances in botanical research, vol. 55.* Burlington, ON, Canada: Academic Press.

Ogburn RM, Edwards EJ. 2012. Quantifying succulence: a rapid, physiologically meaningful metric of plant water storage. *Plant, Cell & Environment* **35**: 1533–1542.

Ogburn RM, Edwards EJ. 2013. Repeated origin of three-dimensional leaf venation releases constraints on the evolution of succulence in plants. *Current Biology* **23**: 722–726.

Osmond CB. 1978. Crassulacean acid metabolism: a curiosity in context. *Annual Review of Plant Physiology* **29**: 379–414.

Pearson K. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* **58**: 240–242.

Pinheiro JC, Bates D, R Core Team. 2023. *nlme: linear and nonlinear mixed effects models.* R package v.3.1-163. [WWW document] URL https://CRAN.R-project.org/package=nlme [accessed 16 August 2023].

Rae JWB, Zhang YG, Liu X, Foster GL, Stoll HM, Whiteford RDM. 2021. Atmospheric CO_2 over the past 66 million years from marine archives. *Annual Review of Earth and Planetary Sciences* **49**: 609–641.

Rashmi KV, Gilad-Bachrach R. 2015. DART: dropouts meet multiple additive regression trees. In: Lebanon G, Vishwanathan SVN, eds. *Proceedings of the Eighteenth International Conference on artificial intelligence and statistics.* San Diego, CA, USA: Proceedings of Machine Learning Research.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**: 217–223.

Revell LJ. 2014. Ancestral character estimation under the threshold model from quantitative genetics. *Evolution* **68**: 743–759.

Sage RF, Gilman IS, Smith JAC, Silvera K, Edwards EJ. 2023. Atmospheric CO_2 decline and the timing of CAM plant evolution. *Annals of Botany* mcad122. doi: 10.1093/aob/mcad122.

de Saussure T. 1804. Recherches chimiques sur la végétation. Paris, France: Chez la V.^{e} Nyon.

Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9**: 671–675.

Silvera K, Santiago LS, Winter K. 2005. Distribution of Crassulacean acid metabolism in orchids of Panama: evidence of selection for weak and strong modes. *Functional Plant Biology* **32**: 397–411.

Tavşanoğlu C, Pausas JG. 2018. A functional trait database for Mediterranean Basin plants. *Scientific Data* **5**: 180135.

Teeri JA, Tonsor SJ, Turner M. 1981. Leaf thickness and carbon isotope composition in the Crassulaceae. *Oecologia* **50**: 367–369.

Terpilowski M. 2019. Scikit-posthocs: pairwise multiple comparison tests in Python. *The Journal of Open Source Software* **4**: 1169.

To T-H, Jung M, Lycett S, Gascuel O. 2016. Fast dating using least-squares criteria and algorithms. *Systematic Biology* **65**: 82–97.

Töpfer N, Braam T, Shameer S, Ratcliffe RG, Sweetlove LJ. 2020. Alternative CAM modes provide environment-specific water-saving benefits in a leaf metabolic model. *Plant Cell* **32**: 3689–3705.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J *et al.* 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**: 261–272.

Waal C, Barrett SCH, Anderson B. 2012. The effect of mammalian herbivory on inflorescence architecture in ornithophilous *Babiana* (Iridaceae): implications for the evolution of a bird perch. *American Journal of Botany* **99**: 1096–1103.

Wang N, Yang Y, Moore MJ, Brockington SF, Walker JF, Brown JW, Liang B, Feng T, Edwards C, Mikenas J *et al.* 2019. Evolution of Portulacineae marked by gene tree conflict and gene family expansion associated with adaptation to harsh environments. *Molecular Biology and Evolution* **36**: 112–126.

Wilde BC, Bragg JG, Cornwell W. 2023. Analyzing trait-climate relationships within and among taxa using machine learning and herbarium specimens. *American Journal of Botany* **110**: e16167.

Winter K. 2019. Ecophysiology of constitutive and facultative CAM photosynthesis. *Journal of Experimental Botany* **2**: 16178.

Winter K, Aranda J, Holtum JAM. 2005. Carbon isotope composition and water-use efficiency in plants with Crassulacean acid metabolism. *Functional Plant Biology* **32**: 381–388.

Winter K, Holtum JAM. 2002. How closely do the $\delta^{13}C$ values of Crassulacean acid metabolism plants reflect the proportion of CO_2 fixed during day and night? *Plant Physiology* **129**: 1843–1851.

Winter K, Wallace BJ, Stocker GC, Roksandic Z. 1983. Crassulacean acid metabolism in Australian vascular epiphytes and some related species. *Oecologia* **57**: 129–141.

Wright S. 1934. An analysis of variability in number of digits in an inbred strain of Guinea pigs. *Genetics* **19**: 506–536.

Yamaga-Hatakeyama Y, Okutani M, Hatakeyama Y, Yabiku T, Yukawa T, Ueno O. 2022. Photosynthesis and leaf structure of F_1 hybrids between *Cymbidium ensifolium* (C_3) and *C. bicolor* subsp. *pubescens* (CAM). *Annals of Botany* mcac157. doi: 10.1093/aob/mcac157.

Yang X, Cushman JC, Borland AM, Edwards EJ, Wullschleger SD, Tuskan GA, Owen NA, Griffiths H, Smith JAC, Paoli HCD *et al.* 2015. A roadmap for research on Crassulacean acid metabolism (CAM) to enhance sustainable food and bioenergy production in a hotter, drier world. *New Phytologist* **207**: 491–504.

Zambrano VAB, Lawson T, Olmos E, Fernández-García N, Borland AM. 2014. Leaf anatomical traits which accommodate the facultative engagement of Crassulacean acid metabolism in tropical trees of the genus *Clusia*. *Journal of Experimental Botany* **65**: 3513–3523.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Dataset S1 Species' anatomical data mean values.

Fig. S1 Image processing by MiniContourFinder.

Fig. S2 Comparisons of raw and $\log_{10}$-transformed data.

Fig. S3 Results of Dunn's *post hoc* tests for group differences between $\log_{10}$-transformed features for select families.

Fig. S4 Correlations between $\log_{10}$-transformed features.

Fig. S5 Accuracies of base models.

Fig. S6 Relative feature importance scores.

Fig. S7 Time-calibrated phylogeny of the Portullugo.

Fig. S8 Portullugo Crassulacean acid metabolism constrained all rates different reconstruction.

Fig. S9 Portullugo Crassulacean acid metabolism all rates different reconstruction.

Methods S1 Summary of MiniContourFinder image segmentation algorithm.

Table S1 Final anatomical data set information.

Table S2 List of accessions sampled from for this study.

Table S3 Results of D'Angostino and Pearson's test for normality and Bartlett's test for homoscedasticity of raw and $\log_{10}$-transformed data.

Table S4 Node calibrations used from Arakaki *et al.* (2011).

Table S5 Results of Kruskal–Wallis tests for group difference between Crassulacean acid metabolism phenotypes.

Table S6 Relative feature importance of multiclass models.

Table S7 Incorrect predictions of the best-performing multiclass model.

Table S8 Incorrect predictions of the best-performing binary model.

Table S9 Phylogenetic signal in anatomical features of the Portullugo.

Table S10 Results of phylogenetic least squares regressions.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.