Reconstructing the diversity dynamics of paleo-grasslands using deep learning on superresolution images of fossil Poaceae pollen

Marc-Élie Adaimé[1], Shu Kong[2], Michael A. Urban[1,3], F. Alayne Street-Perrott[4], Dirk Verschuren[5], Surangi W. Punyasena[1]

[1] Department of Plant Biology, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA
[2] Faculty of Science and Technology, University of Macau, Macau 999078, China
[3] Department of Biology, University of New Brunswick, Fredericton, NB, E3B 5A3, Canada
[4] Faculty of Science and Engineering, Swansea University, Singleton Park, Swansea, SA2 8PP, UK
[5] Limnology Unit, Department of Biology, Ghent University, B-9000 Gent, Belgium

Corresponding author: spunya1@illinois.edu

Significance

The pollen of most grass species are morphologically indistinguishable using traditional optical microscopy, but we show that they can be differentiated through deep learning analyses of superresolution images. Abstracted morphological features derived from convolutional neural networks can be used to quantify the biological and physiological diversity of grass pollen assemblages, without *a priori* knowledge of the species present, and used to reconstruct past changes in the taxonomic diversity and relative abundance of C_4 grasses in ancient grasslands. This approach unlocks ecological information that had been previously unattainable from the fossil pollen record and demonstrates that deep learning can solve some of the most intractable identification problems in the reconstruction of past vegetation dynamics.

Abstract

Despite its abundance in the fossil record, grass pollen is largely overlooked as a source of ecological and evolutionary data because most Poaceae species cannot be differentiated using traditional optical microscopy. However, deep learning techniques can quantify the small variations in grass pollen morphology visible under superresolution microscopy. We use the abstracted morphological features output by deep learning to estimate the taxonomic diversity and physiology of fossil grass pollen assemblages. Using a semi-supervised learning strategy, we trained convolutional neural networks (CNNs) on pollen images of 60 widely distributed grass species and unlabeled fossil Poaceae. Semi-supervised learning improved the CNN models' capability to generalize feature recognition in fossil pollen specimens. Our models successfully captured both the taxonomic diversity of an assemblage and morphological differences between C_3 and C_4 species. We applied our trained models to fossil grass pollen assemblages from a 25,000-year lake-sediment record from eastern equatorial Africa and correlated past shifts in grass diversity with atmospheric CO_2 concentration and proxy records of local temperature, precipitation, and fire occurrence. We quantified grass diversity for each time window using morphological variability, calculating both Shannon entropy and morphotype counts from the specimens' CNN features. Reconstructed C_3:C_4 ratios suggest a gradual increase in C_4 grasses with rising temperature and fire activity across the late-glacial to Holocene transition. Our results demonstrate that quantitative machine-learned features of pollen morphology can significantly advance palynological analysis, enabling robust estimation of grass diversity and C_3:C_4 ratio in ancient grassland ecosystems.

Keywords

C_4 photosynthesis, deep learning, diversity estimation, Poaceae, superresolution imaging

Introduction

The evolutionary history of grasses and the ecological history of grasslands are recorded in the fossil pollen record. However, because the pollen of Poaceae species are visually indistinguishable using traditional optical microscopy, this vast paleontological record has been largely limited to estimating changes in past abundance (1, 2). As a result, grass pollen data has contributed little insight into the evolution of the ~11,800 species of this diverse plant lineage (3) or the establishment and paleoecological dynamics of grasslands.

The fossil history of grasses is foundational to understanding the development of modern grasslands and their ecological importance. Grasslands are a geologically recent ecosystem, covering ~40% of Earth's terrestrial surface and contributing to ecosystem services such as soil fertility, carbon storage, and atmospheric cooling (4, 5). Grasslands are characterized by unique ecological properties, including high water-use efficiency and high albedo, and play a critical role in regulating global geochemical cycles, including those of carbon, nitrogen, and silica (4, 6). The productivity and ecological resilience of grasslands is linked to its diversity (7), with high species richness increasing resilience to environmental stress (7, 8).

Grass species employ one of two photosynthetic pathways, C_3 and C_4. C_4 photosynthesis increases water use efficiency and tolerance of low atmospheric CO_2 (9). It evolved independently in multiple grass lineages and allowed these species to eventually dominate the world's grasslands by the late Miocene (10, 11). The global expansion of C_4-dominated grasslands occurred alongside a downward trend in atmospheric CO_2 concentrations during the late Cenozoic and is associated with major environmental changes linked to the growth of ice sheets, such as the expansion of seasonally dry tropical and temperate climates (12, 13).

During the Quaternary (the last 2.5 million years), grassland ecosystems expanded throughout the tropics due to the generally cool and dry climatic conditions that characterized glacial periods in low-latitude regions (14). Pollen records from southern and southeastern Brazil indicate that now forested regions were dominated by grasslands during glacial times (15). Similarly, in Africa, grasslands expanded during the Last Glacial Maximum (LGM) (16). Sensitivity to climate, combined with their

widespread geographic distribution and extensive fossil record, make grasses an important biological proxy in the reconstruction of past climates and environments.

However, interpreting the fossil record of grasses has been a long-standing challenge. Poaceae macrofossils can often be difficult to distinguish from other monocotyledonous plants (12), and grass phytoliths do not directly capture taxonomic diversity with multiple morphotypes shared between C_3 and C_4 species (11, 17). Pollen across Poaceae species is morphologically similar, but small variations in the exine patterning of the pollen grain surface are taxonomically meaningful and can be quantitatively characterized (2, 18, 19). However, these previous categorizations of grass morphotypes have been too broad to capture changes in the taxonomic structure of grassland communities on millennial timescales.

New approaches are therefore needed to effectively reconstruct the diversity and composition of the grass pollen record. Machine learning has become an important part of the palynological toolkit over the last decade (e.g., 25–27) but with few exceptions (23–25), it has been largely used to automate, and not improve upon, pollen identifications. In this study, we developed deep learning pipelines that use convolutional neural networks (CNNs) trained on superresolution images (0.04 μm resolution) of modern and fossil Poaceae pollen (Fig. 1A) with the goal of improving on expert classifications of grass pollen. The abstracted morphological features derived from these CNNs capture taxonomic and physiological diversity in grass pollen assemblages. Our first computational pipeline quantifies taxonomic diversity by assessing morphological variability within fossil pollen assemblages using Shannon entropy (Fig. 1B). Our second pipeline uses a random forest classifier (26) to identify the abstracted morphological features associated with each photosynthetic pathway and reconstruct changes in the abundance ratio of C_3 and C_4 grass pollen (Fig. 1C).

We applied our trained CNN models to superresolution images of fossil grass pollen extracted from the 25,000-year sediment record of Lake Rutundu (0.0416° S, 37.4638° E, 3,078 m asl), a subalpine crater lake on Mt. Kenya in eastern equatorial Africa (27) (Fig. 2). We reconstructed the taxonomic diversity and physiological composition of the surrounding high-mountain grasslands relative to local environmental factors such as fire history (28), temperature (29), and precipitation (30), as well as

known atmospheric CO_2 levels (31). The wealth of paleoecological and paleoclimate data available from previous research on the Lake Rutundu record (27, 28, 32, 33) allowed direct comparison of our results with prior work on the history of Mt. Kenya's grasslands. Our findings demonstrate that superresolution imaging and deep learning enables -- for the first time -- the estimation of grass diversity and C_3:C_4 ratios from fossil pollen morphology. With these tools, we can interpret the diversity and composition of grass assemblages even in the absence of species identifications.

Results

CNNs can discriminate among grass species

We trained two separate K-way classification CNNs using superresolution structured illumination (SR-SIM) (34) images of 1,241 modern and 1,503 fossil pollen grains (Fig. 1A). Modern pollen was isolated from herbarium specimens of 60 extant Poaceae species (3 - 37 specimens per species, SI Appendix, Tables S1 and S2). Fossil material was isolated from a Lake Rutundu sediment core (27) and recently deposited sediments from five additional Mt Kenya lakes (SI Appendix, Tables S3 and S4). The first CNN (H-CNN) was trained on maximum intensity projections (MIPs) of the whole grain. The second CNN (P-CNN) used small cropped segments of the MIP for training. CNNs are a series of convolutional and pooling layers with a final classification layer that outputs probabilities of class labels (35). Notably, we employed CNNs to extract abstracted morphological characters for downstream analyses and not for classification. Our analysis ignores the final classification layer and instead uses the penultimate layer containing the highest-level representation of a specimen's morphology, or "features" (25). However, we assessed the robustness of these CNN features by calculating the average classification accuracy of our trained models (Fig. 3 and SI Appendix, Fig. S1). The H-CNN and P-CNN achieved an average accuracy of 61.40% and 72.51%, respectively (SI Appendix, Table S5). Fusion of the two CNNs achieved an average accuracy of 73.95% (SI Appendix, Table S5). Lowest accuracies were for species with the lowest number of training images (3-4 specimens), but all accuracies were significantly greater than chance (1.67%).

Morphological variability reflects taxonomic diversity

We used the Shannon entropy (36) of the CNN features of a grass pollen assemblage as a measure of its taxonomic diversity and tested this assumption by measuring the morphological entropy of simulated grass communities. We randomly split our modern Poaceae dataset into 30 species used to train the CNNs and 30 unseen species from which we drew random selections of specimens to form 1000 artificial assemblages of varying species richness and species abundances. For each artificial assemblage, we calculated Shannon's index (36) based on the number of species selected and their relative abundances. We then forward-passed the specimens through our trained CNN models and calculated the assemblages' Shannon entropy from the distribution of their modified CNN features (see Methods). Linear regression analysis quantified the strong positive correlation between Shannon entropy and the Shannon index in our data ($r = 0.972$, $p < 0.0001$) (SI Appendix, Fig. S2). A unit increase in Shannon entropy corresponded to an equivalent increase in Shannon diversity, meaning morphological variability was proportional to taxonomic (species) diversity.

Shifts in grass diversity on Mt Kenya

The Shannon entropy of fossil grass pollen assemblages from the 25,000-year Lake Rutundu record showed a positive correlation with the evolution of atmospheric CO_2 (Pearson's $r = 0.548$, $p = 0.006$) and the temperature history of Mt Kenya ($r = 0.551$, $p= 0.008$) over the same period (Fig. 4A and SI Appendix, Fig. S3 and Table S6), indicating that higher CO_2 concentration and higher local temperatures are associated with greater morphological variability and taxonomic diversity in these subalpine tropical grasslands.

Using k-means clustering to translate the entropy values into an estimated number of distinct morphotypes, we then calculated the Shannon diversity index for each time window based on the number of morphotypes and their relative abundances (see Methods). These diversity index values exhibited strong positive correlation with

CO_2 ($r = 0.742$, $p < 0.0001$) and temperature ($r = 0.783$, $p < 0.0001$) (Fig. 4A and SI Appendix, Fig. S3 and Table S6). There was no significant relationship, however, between our time series of either Shannon entropy or Shannon diversity values and reconstructed changes in local precipitation (Fig. 4A and SI Appendix, Fig. S3 and Table S6).

Finally, we used the CNN features of our modern pollen dataset and the Rutundu pollen assemblages as a proxy for changes in taxonomic composition, under the assumption that differences in the distribution of CNN features reflected differences in the underlying species composition. We used PCA to reduce the dimensionality of the dataset and visualize the distribution of CNN features across different time windows. We found that the morphological diversity of the fossil dataset was a subset of the diversity present in the modern dataset (SI Appendix, Fig. S4), but the distribution of features were significantly different, particularly for the first principal component (PC1, Welch's t-test, $p < 0.0001$). The morphological space occupied by the fossil assemblages varied through time, with the space occupied by the four LGM samples (21,260 - 18,435 cal yr. BP) distinct from that of other time periods along PC1 and occupying a smaller area in the morphospace defined by PC1 and PC2 (SI Appendix, Fig. S4).

Pollen morphology captures grass photosynthetic pathway

The phylogenetic PCA (pPCA) of the CNN feature matrix of the 60 modern pollen samples yielded 59 phylogenetic principal components (pPCs), of which 12 showed a statistically significant difference between C_3 and C_4 grass taxa (Welch's t-test, $p < 0.05$) (SI Appendix, Fig. S5). We identified an optimal subset of five pPCs that discriminated between the pollen of C_3 and C_4 grasses with an accuracy of 91.25% through recursive feature elimination (RFE) using a random forest classifier (37, 38) (see Methods). We then used these five pPCs to train a second random forest classifier to label fossil pollen grains as C_3 or C_4 and used bootstrapping to calculate confidence intervals for the estimated C_3:C_4 ratios in fossil samples.

Late Quaternary trends in C_4 grass abundance

We projected the standardized CNN features of the fossil specimens onto the pPCA space defined by the CNN features of the modern grass pollen (see Methods). This generated a set of five pPCA scores for each fossil pollen grain that was then forward-passed through the trained random forest classifier for C_3/C_4 classification. The percentage of pollen grains within each fossil assemblage that were assigned to C_4 grasses ranged between 21.03% and 51.61%, with a slight increase through time over the past 25,000 years (Fig. 4B). This temporal trend in C_4 grass abundance is positively correlated with past changes in atmospheric CO_2 (Pearson's $r = 0.437$, $p = 0.033$) and local biomass burning (Pearson's $r = 0.550$, $p = 0.005$ for microcharcoal flux; Pearson's $r = 0.442$, $p = 0.031$ for macrocharcoal flux), suggesting that post-glacial expansion of C_4 grasses in the subalpine zone on Mt Kenya was mediated by increased fire prevalence.

Our results contrast with a previous reconstruction of C_4 grass abundance on Mt Kenya, inferred from the carbon isotope ratio ($\delta^{13}C$) of fossil grass pollen in the Lake Rutundu record, that shows a gradual decrease in C_4 grass abundance over time (Fig. 4B) (28). However, C_4 abundance estimates in this study were highly variable (ranging between 3% and 100%), with wide confidence intervals indicating a large margin of error (Fig. 4B). Our results are more consistent with C_4 grass estimates based on fossil grass cuticle preserved in the Lake Rutundu record, which range between 16.32% and 72.67%, and show no consistent trend over the past 25,000 years (Fig. 4B) (27).

Discussion

Deep learning allows us to quantify the sub-micrometer differences in pollen morphology visible under superresolution microscopy and statistically discriminate among different morphotypes of grass pollen, opening new opportunities in paleobotanical research. With these computational tools, we are able to more consistently discriminate among known species (Fig. 3), capture trends in morphological diversity that approximate species diversity (Fig. 4A and SI Appendix, Fig. S3), and

capture changes in the relative abundance of C_3 and C_4 grasses through time (Fig. 4B and SI Appendix, Fig. S3).

Grass diversity increases in East African grasslands

Shannon entropy estimates species diversity without species identifications. Estimates are most accurate for high-diversity assemblages (SI Appendix, Fig S2). When applied to the fossil pollen record from Lake Rutundu, we found that grass diversity in subalpine grasslands on Mt Kenya remained relatively low from 25,000 years ago until the Last Glacial Maximum (LGM) (~20,000 cal. yr BP), after which there was a gradual increase in taxonomic diversity, as inferred from the variability of CNN features (Fig. 4A). Changes in grass diversity were significantly correlated with changes in CO_2 and temperature (SI Appendix, Fig. S3 and Table S6). Converting entropy measurements into morphotype counts and Shannon diversity index values produced a more pronounced increase in estimated diversity that significantly correlated with CO_2 levels and temperature (SI Appendix, Fig. S3 and Table S6). These findings suggest that elevated CO_2 levels, coupled with warmer conditions, favored diversification of subalpine Poaceae communities around Lake Rutundu following the LGM. This is consistent with multiple studies that have documented increased biomass productivity and biodiversity, particularly in grasslands, under conditions of elevated post-glacial CO_2 levels, as well as in warmer, more fire-disturbed environments (39–43).

Time series of the organic geochemical composition of Lake Rutundu sediments show that the carbon-nitrogen (C/N) ratio, total organic carbon (TOC), and terrestrial biomarker fluxes reached some of their lowest recorded levels during the LGM, before rapidly increasing at ~16,000 cal. yr BP. This suggests terrestrial primary productivity in the subalpine zone of Mt Kenya was lower during the LGM, when grasses were more abundant relative to other herbaceous plants such as heather (Ericaceae) (27, 28, 32). However, our results show that during this period, grass diversity was also at its lowest point in the last 25,000 years. Notably, as the dominance of Poaceae in the subalpine zone declined under the warmer conditions of the Holocene, grass diversity increased.

C_4 grasses have thinner pollen walls and simpler surface texture

Integrating feature selection with a random forest model allowed us to select the optimal set of phylogenetic principal components that best predicted photosynthetic pathway. pPCA allowed us to reduce the influence of relatedness and identify convergent morphological features. We cannot directly link the 12 principal components that captured significant differences between C_3 and C_4 grass pollen (SI Appendix, Fig. S5) to specific morphological characters, but we found that Shannon entropy and pixel intensity values were, on average, significantly greater in C_3 grass species than C_4 grass species (Welch's t-test, $p = 0.029$ and $p = 0.041$ respectively) (Fig. 5). Higher entropy suggests more complex surface texture and increased pixel brightness suggests a thicker exine wall. We tested the hypothesis that the pollen grains of C_3 grasses have thicker exine walls using direct measurements of exine thickness from (19). In this alternate set of Poaceae species, exine thickness was also significantly greater in C_3 taxa (Welch's t-test, $p = 0.001$) (Fig. 5).

Morphological differences between pollen from C_3 and C_4 grasses may be directly linked to their physiological characteristics. C_4 grasses tend to achieve faster growth by producing less dense leaves (higher leaf area to mass ratio), enabling the production of more leaves at the same carbon cost (44). C_4 grasses also invest more in their root systems (44). The systematic difference in resource allocation between C_3 and C_4 grasses may explain why physiological differences are also encoded in pollen morphology, independent of phylogenetic relatedness. Grass pollen morphology may reflect the evolutionary trade-off between investment in roots and shoots. We hypothesize that C_4 grasses may invest less in pollen development and structure, producing grains with thinner exines and less complex surface patterning.

Changing prevalence of C_3 and C_4 grasses in East African grasslands

Our analysis of fossil grass pollen from Lake Rutundu showed that the proportion of C_4 grasses changed through time, with a slight increase from ~18,000 to 5,000 cal. yr BP (Fig. 4B). The increase positively correlated with both local fire prevalence (inferred from micro- and macrocharcoal fluxes) and atmospheric CO_2 (SI Appendix, Fig. S3 and

Table S6) and is consistent with research that identified increased fire activity as a key factor in the expansion of C_4-dominated grasslands in Africa during the Late Miocene (10, 45, 46). Likewise, while full-glacial atmospheric CO_2 concentrations (~180-200 ppm) are limiting for both C_3 and C_4 grasses, the increase in CO_2 concentrations from glacial to postglacial levels provided favorable conditions for an increase in biomass production, and may have been more advantageous for C_4 species under certain environmental conditions, such as water stress (47). Our results reflect this trend, though C_3 grasses remained dominant overall (Fig. 4B). The increased proportion of C_4 grasses we reconstructed across the glacial-to-Holocene transition may also be linked to warmer conditions in the Holocene. However, the modern-day occurrence of cold-tolerant panicoid C_4 grasses in alpine grasslands (48, 49) complicates this interpretation. Our results did not show significant correlation between the proportion of C_4 grasses and precipitation (Fig. 4A and SI Appendix Fig. S3 and Table S6). However, increased biomass burning may have been the result of higher precipitation in eastern equatorial Africa during the Late-Glacial period and early Holocene (~14,300-12,300 and 11,400-9,600 cal. yr BP), as evidenced by more depleted δD values leaf-wax alkanes (52) (Fig. 4C), as well as depleted $\delta^{13}C$ values of organic matter and water-level changes in East African lakes and swamps (50).

Our estimates of C_4 grass abundance are consistent with previous estimates from Lake Rutundu fossil cuticle (27) (Fig. 4B). From ~25,000 to 18,000 cal. yr BP, our results align with the most conservative estimates of C_4 abundance (based on cuticle fragments identified with certainty as C_4 grass species) (27). From ~15,000 cal. yr BP to present, our data align with the median trend, which includes both certain and likely C_4 taxa (27). Our results do not match estimates of C_4 grass abundance based on pollen carbon isotope measurements before ~10,000 cal. yr BP, but align with estimates after ~10,000 cal. yr BP (28). Our underestimation of C_4 grass abundance compared to these other proxies, particularly before the Holocene, may stem in part from differences between the species composition of fossil assemblages from Mt. Kenya and our modern training dataset. If we use the diversity of CNN features as stand-in for taxonomic composition, we see that while there is substantial overlap in the morphological space occupied by our modern and fossil datasets, this range of morphological diversity is

significantly different from that of modern grass pollen (SI Appendix, Fig. S4A). This is especially true for the range of morphospace where the LGM pollen assemblages fall (SI Appendix, Fig. S4B). This may have introduced uncertainty in our random forest estimates and explain some of the mismatch between our time series and those of other proxies for grass abundance.

Deep-time implications

Deep learning can be a powerful tool for revealing shifts in grass diversity and composition. It allows us to quantify and use the morphological data captured in fossil pollen in new and unexpected ways. By analyzing the long-term morphological variability within pollen assemblages, we can test hypotheses about the evolution and physiological adaptations of grasses over thousands to millions of years and provide a deeper understanding of how grasses adapted to changing climates and environments throughout the Cenozoic.

Grasslands provide crucial habitat for many plant and animal species, driving key evolutionary adaptations (51). The first putative Poaceae appeared in the Early Cretaceous (113-101 Mya) (46–48), and grasslands were present on every continent except Antarctica by the middle Eocene (~38 Mya) (12). Although our study focused on Late Quaternary grasslands, this same approach can be applied to fossil grass pollen assemblages further back in the geologic record, providing us deeper insight into the evolution of this important global biome.

Methods

Pollen samples and paleoclimate data

Pollen samples of 60 extant grass species were isolated from herbarium material from the Missouri Botanical Gardens (USA), University of Illinois (USA), and Harvard (USA) or purchased from Sigma-Aldrich (St. Louis, MO, USA) (SI Appendix, Table S1). These are predominately, but not exclusively, widely distributed species that occur in East Africa, including the important African cereals sorghum (*Sorghum halepense*), finger millet (*Eleusine coracana*), and maize (*Zea mays*) (SI Appendix, Table S2).

Thirty-six of these 60 species are C_3 grasses and 24 are C_4. Twenty-four samples of fossil pollen (SI Appendix, Table S4) were isolated from a 7.55-m sediment sequence collected from Lake Rutundu in 1996 (27). Rutundu is a 40-ha oligotrophic lake, 3078 m above sea level (a.s.l.) on the northeast side of Mt. Kenya, just above the current treeline (Fig. 2) (27). Recently deposited pollen was extracted from the surface sediments of five additional small lakes on Mt. Kenya distributed along an altitudinal transect between 1820 and 4585 m a.s.l. (Fig. 2 and SI Appendix, Table S3) (55). All pollen samples were prepared following (56, 57). (Details provided in the SI Appendix.) The ages of the fossil samples are derived from the age model for the Rutundu sediment sequence based on AMS ^{14}C dating of bulk organic carbon (27, 28).

Paleoclimate data for the subalpine zone of Mt. Kenya were taken from published studies on the Rutundu sediment record. The temperature reconstruction is based on abundance ratios of preserved glycerol dialkyl glycerol tetraethers (GDGTs) (29) and the precipitation reconstruction is based on the hydrogen-isotope signature of alkanes in fossil leaf waxes from terrestrial vegetation (30). For fire history, we used data on microscopic and macroscopic charcoal influx to Lake Rutundu as a proxy for past variation in local biomass burning and fire prevalence (28). We used the composite record of atmospheric CO_2 concentration based on air-bubble measurements in several Antarctic ice cores (31).

SR-SIM imaging

Pollen specimens were imaged using a Zeiss Elyra S1 superresolution structured illumination (SR-SIM) microscope. SR-SIM is an optical microscopy method capable of capturing morphological features <150 nm in length, below the diffraction limit of light (34). We used a 63x Plan Apochromat (1.4 NA) oil objective and an excitation wavelength of 561 nm. Image resolution was 0.0397 μm/pixel. Each axial plane of the Z-stack was constructed with five grid rotations, spaced 0.18 μm apart, with 28-108 planes per grain. Images were cropped to the perimeter of each pollen grain.

Deep learning architectures and dataset setup

We developed two separate K-way classification CNNs based on the ResNeXt-101 architecture (58): a holistic image CNN (H-CNN) trained on maximum intensity projections (MIP) of entire image stacks, and a patch CNN (P-CNN) trained on square crops of the MIP, each covering ~10% of the image (25) (Fig. 1A). Both CNNs were trained using images of modern reference pollen (labeled data) and images of fossil pollen (unlabeled data) in a semi-supervised learning (SSL) framework (59). SSL assigns pseudo-labels to unlabeled images and treats them as labeled in training classification models. The labeled reference data were divided into training (70%) and validation (30%) sets for model training and model selection, respectively. Unlabeled images were assigned pseudo-labels by the model being trained, and images that had pseudo-label confidence scores greater ≥95% were retained for training. The validation set of reference images was used for hyperparameter tuning and model selection. SSL improved the model's ability to generalize across both modern and fossil data. Training details are included in the SI Appendix.

The H-CNN generated a single *K*-dimensional logit vector for each pollen specimen, while the P-CNN produced multiple logit vectors corresponding to multiple input patches per specimen (Fig. 1A). The mean of these vectors by P-CNN was calculated and normalized using softmax (25). For each specimen, we fused the outputs of the two CNNs by multiplying the two *K*-dimensional classification probability vectors and normalizing it to unit length (25).

Feature extraction

Our analysis relied on the features characterized by the two CNNs. We concatenated the 2048-dimensional feature vectors derived from the penultimate (global pooling) layer of the H-CNN and P-CNN to produce a 4096-dimensional feature vector for each specimen (Fig. 1A). This combined feature vector represented a high-level abstraction of the pollen image that served as input for our morphological analyses of diversity and physiology. To visualize and compare the morphological features of the modern grass pollen dataset and the 24 fossil pollen assemblages, we stacked their

classification features into a single standardized dataset and performed a PCA on the combined feature matrix (Fig S4).

Morphological measures of diversity

We used simulated communities to test our hypothesis that the morphological diversity within a grass pollen assemblage can serve as a proxy for taxonomic (species) diversity. From our modern pollen dataset of 60 grass species, we randomly selected 30 species for training our CNNs and used the remaining 30 species (572 specimens) as an independent test set. From the test set, we randomly selected a random number of species (between 2 and 30) and a random number of specimens per species (between 1 and the maximum number of available specimens per species). We quantified the diversity of each simulated community using Shannon's index:

$$H' = -\sum_{i=1}^{S} p_i \log(p_i)$$

where S is the total number of species and p_i is the relative frequency of the i^{th}species.

We next measured morphological diversity of each community using its CNN-derived features. We standardized the feature matrix and applied PCA. We then computed Shannon entropy (36) over the probability distribution of the first principal component (PC1). The distribution was estimated using kernel density estimation (KDE) (60). The estimated probability density for a given PC1 score, z_i, is calculated as:

$$p(z_i) = \hat{f}(z_i) = \frac{1}{n \times h} \sum_{j=1}^{n} K\left(\frac{z_i - z_j}{h}\right)$$

Where $\hat{f}$ is the estimated density calculated using KDE, n is the total number of specimens per sample, h is the bandwidth, and K is the kernel function. We computed entropy H from the density values:

$$H = -\sum_{i=1}^{n} p(z_i) \log(p(z_i))$$

440 We calculated Shannon entropy and the Shannon diversity index for 1000 simulations
441 and measured their correlation using linear regression, reporting the Pearson
442 correlation coefficient (r) and p-value.

443

444 ***Fossil grass diversity estimation and paleoenvironmental correlates***

445 We calculated the Shannon entropy values for the CNN-derived morphological
446 features of the 24 fossil Rutundu samples. We then used k-means clustering to
447 translate the entropy values into the number of distinct morphotypes in each time
448 window. We set the number of morphotypes by selecting the Shannon diversity index
449 values that best correlated with Shannon entropy values using Pearson's correlation
450 coefficient (Fig. 1B).

451 We report both the Shannon entropy and Shannon index values and their
452 correlation to five paleoenvironmental variables: atmospheric CO_2 level, local
453 reconstructed temperature and precipitation, and micro- and macrocharcoal influx as
454 proxy for fire prevalence. Our results include Pearson's correlation coefficient (r) and its
455 corresponding p-value for each individual regression (SI Appendix, Fig. S3 and Table
456 S6).

457

458 ***Grass pollen morphology and photosynthetic pathway***

459 We conducted a phylogenetic PCA (pPCA) of the CNN-derived feature matrix
460 using the phyl.pca function of the package phytools (61, 62) (Fig. 1C). We used a
461 Bayesian time-calibrated grass phylogeny based on chloroplast DNA sequences
462 covering approximately 90% of existing grass genera (54) as our input tree. To identify
463 the pPCs most predictive of the photosynthetic pathway, we implemented a feature
464 selection process using recursive feature elimination (RFE) with a random forest
465 classifier, using the functions rfe and randomForest of the 'caret' and 'randomForest'
466 packages in R (26, 38, 63). Photosynthetic pathway was treated as a categorical
467 variable. We established a control for RFE using repeated cross-validation with five
468 folds, repeated 10 times.

Using the selected pPCs, we then trained a final random forest model to classify the photosynthetic pathway based on the modern data's pPCA scores, using 1000 decision trees to reduce potential overfitting (26). To quantify the uncertainty in our estimates, we employed a bootstrapping procedure with 1000 resamples. For each bootstrap sample, we calculated the mean proportion (percent abundance) of C_4 grasses and computed the 95% confidence intervals around the mean estimates for each time window in the Rutundu record.

Estimates of past C_4 abundance and paleoenvironmental correlates

We standardized the fossil dataset using the mean and standard deviation of the modern data and projected the standardized fossil data onto the pPCA space computed using the modern data. We then extracted the optimal pPCs determined using RFE (Fig. 1C). We used the final random forest model, trained on modern data, to determine the photosynthetic pathway (C_3 or C_4) for each fossil specimen based on the selected pPCA components.

We report the correlation between our time series of estimated C_4 abundance and those of the five paleoenvironmental variables: atmospheric CO_2 concentration, local temperature, local precipitation, and micro- and macro-charcoal influx. We calculated Pearson's correlation coefficient (r) and its corresponding p-value for each individual regression (SI Appendix, Fig. S3 and Table S6).

Quantifying morphological differences between C_3 and C_4 pollen

We used entropy to estimate the complexity of the grass pollen surface texture, with higher values indicating more random and complex textural patterning. We calculated Shannon entropy by segmenting the pollen grain image and quantifying the isolated grain's information content (complexity) based on the distribution of its pixel intensities (64). Because pollen exine is autofluorescent, pixel intensity allows us to infer variations in thickness of the pollen wall, with brighter images suggesting a thicker wall (34). We used direct exine thickness measurements (19) for 35 C_3 and 23 C_4 grass species to confirm this relationship. We calculated the median of entropy, pixel intensity, and exine thickness measurements across all specimens of each species in our modern

dataset and used Welch's t-test to assess the statistical significance of observed differences.

Acknowledgments

Imaging costs and postdoctoral support for M.A.U. were provided by NSF-DBI – Advances in Bioinformatics (NSF-DBI-1262561) to S.W.P. S.K. was supported in part by the University of Macau (SRG2023-00044-FST). M.-E.A. was supported in part by the University of Illinois Tom L. Phillips Memorial Fund for Paleobotany. We thank Tim Gallaher for referring us to his recently published grass phylogeny.

Author Contributions

M.-E.A., S.K., and S.W.P. designed the research. M.-E.A. conducted the computational analyses and experiments. M.A.U. conducted the imaging, sample preparation, and lab analyses. F.A.S.-P. and D.V. provided samples and collected data. S.W.P. and S.K. supervised the research. M.-E.A., S.W.P., S.K., and M.A.U. wrote the manuscript, with feedback from all authors.

Data Availability

Superresolution structured illumination images used in this study were submitted to the Illinois Databank (DOI: 10.13012/B2IDB-1391373_V1) (65). The code base is publicly available on Github (DOI: 10.5281/zenodo.13756262) (66), and can be accessed using the following link: https://github.com/paleopollen/Pollen_Diversity_Dynamics

References

1. M. B. Bush, On the interpretation of fossil Poaceae pollen in the lowland humid neotropics. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **177**, 5–17 (2002).

2. L. Mander, S. W. Punyasena, Grass pollen surface ornamentation: A review of morphotypes and taxonomic utility. *J. Micropalaeontology* **35**, 121–124 (2016).

3. Grass Phylogeny Working Group III, *et al.*, Nuclear phylogenomics of grasses (Poaceae) supports current classification and reveals repeated reticulation. [Preprint] (2024). Available at: http://biorxiv.org/lookup/doi/10.1101/2024.05.28.596153 [Accessed 16 August 2024].

4. J. S. Petermann, O. Y. Buzhdygan, Grassland biodiversity. *Curr. Biol.* **31**, R1195–R1201 (2021).

5. G. J. Retallack, Cenozoic expansion of grasslands and climatic cooling. *J. Geol.* **109**, 407–426 (2001).

6. S. E. Melzer, *et al.*, Fire and grazing impacts on silica production and storage in grass dominated ecosystems. *Biogeochemistry* **97**, 263–278 (2010).

7. D. Tilman, J. A. Downing, Biodiversity and stability in grasslands. **367** (1994).

8. A. Lüscher, *et al.*, Using plant diversity to reduce vulnerability and increase drought resilience of permanent and sown productive grasslands. *Grass Forage Sci.* **77**, 235–246 (2022).

9. R. F. Sage, The evolution of C_4 photosynthesis. *New Phytol.* **161**, 341–370 (2004).

10. S. Hoetzel, L. Dupont, E. Schefuß, F. Rommerskirchen, G. Wefer, The role of fire in Miocene to Pliocene C_4 grassland and ecosystem evolution. *Nat. Geosci.* **6**, 1027–1030 (2013).

11. C. A. E. Strömberg, R. E. Dunn, C. Crifò, E. B. Harris, “Phytoliths in paleoecology: Analytical considerations, current use, and future directions” in *Methods in Paleoecology: Reconstructing Cenozoic Terrestrial Environments and Ecological Communities*, D. A. Croft, D. F. Su, S. W. Simpson, Eds. (Springer International Publishing, 2018), pp. 235–287.

12. C. A. E. Strömberg, Evolution of grasses and grassland ecosystems. *Annu. Rev. Earth Planet. Sci.* **39**, 517–544 (2011).

13. R. Sage, T. Sage, F. Kocacinar, Photorespiration and the evolution of C_4 photosynthesis. *Annu. Rev. Plant Biol.* **63**, 19–47 (2012).

14. L. ’t Mannetje, Climate change and grasslands through the ages: An overview. *Grass Forage Sci.* **62**, 113–117 (2007).

15. H. Behling, South and southeast Brazilian grasslands during Late Quaternary times: A synthesis. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **177**, 19–27 (2002).

16. L. Scott, Grassland development under glacial and interglacial conditions in southern Africa: Review of pollen, phytolith and isotope evidence. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **177**, 47–57 (2002).

17. P. C. Twiss, "Predicted world distribution of C_3 and C_4 grass phytoliths" in *Phytolith Systematics: Emerging Issues*, G. Rapp, S. C. Mulholland, Eds. (Springer US, 1992), pp. 113–128.

18. L. Mander, M. Li, W. Mio, C. C. Fowlkes, S. W. Punyasena, Classification of grass pollen through the quantitative analysis of surface ornamentation and texture. *Proc. Biol. Sci.* **280**, 20131905 (2013).

19. C. Wei, P. E. Jardine, W. D. Gosling, C. Hoorn, Is Poaceae pollen size a useful proxy in palaeoecological studies? New insights from a Poaceae pollen morphological study in the Amazon. *Rev. Palaeobot. Palynol.* **308**, 104790 (2023).

20. E. C. Stillman, J. R. Flenley, The needs and prospects for automation in palynology. *Quat. Sci. Rev.* **15**, 1–5 (1996).

21. K. Holt, G. Allen, R. Hodgson, S. Marsland, J. Flenley, Progress towards an automated trainable pollen location and classifier system for use in the palynology laboratory. *Rev. Palaeobot. Palynol.* **167**, 175–183 (2011).

22. C. M. Barnes, *et al.*, Deductive automated pollen classification in environmental samples via exploratory deep learning and imaging flow cytometry. *New Phytol.* **240**, 1305–1326 (2023).

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

24. I. C. Romero, *et al.*, Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy. *Proc. Natl. Acad. Sci.* **117**, 28496–28505 (2020).

25. M.-É. Adaimé, S. Kong, S. W. Punyasena, Deep learning approaches to the phylogenetic placement of extinct pollen morphotypes. *PNAS Nexus* **3**, pgad419 (2024).

26. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).

27. K. J. Ficken, M. J. Wooller, D. L. Swain, F. A. Street-Perrott, G. Eglinton, Reconstruction of a subalpine grass-dominated ecosystem, Lake Rutundu, Mount Kenya: A novel multi-proxy approach. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **177**, 137–149 (2002).

28. M. A. Urban, D. M. Nelson, F. A. Street-Perrott, D. Verschuren, F. S. Hu, A Late-Quaternary perspective on atmospheric pCO_2, climate, and fire as drivers of C_4-grass abundance. *Ecology* **96**, 642–653 (2015).

29. S. E. Loomis, *et al.*, The tropical lapse rate steepened during the Last Glacial Maximum. *Sci. Adv.* **3**, e1600815 (2017).

30. S. Garelick, J. M. Russell, S. Dee, D. Verschuren, D. O. Olago, Atmospheric controls on

precipitation isotopes and hydroclimate in high-elevation regions in Eastern Africa since the Last Glacial Maximum. *Earth Planet. Sci. Lett.* **567**, 116984 (2021).

31. B. Bereiter, *et al.*, Revision of the EPICA Dome C CO_2 record from 800 to 600 kyr before present. *Geophys. Res. Lett.* **42**, 542–549 (2015).

32. M. Wooller, *et al.*, Late Quaternary vegetation changes around Lake Rutundu, Mount Kenya, East Africa: Evidence from grass cuticles, pollen and stable carbon isotopes. *J. Quat. Sci.* **18**, 3–15 (2003).

33. F. A. Street-Perrott, *et al.*, Towards an understanding of late Quaternary variations in the continental biogeochemical cycle of silicon: Multi-isotope and sediment-flux data for Lake Rutundu, Mt Kenya, East Africa, since 38 ka BP. *J. Quat. Sci.* **23**, 375–387 (2008).

34. M. Sivaguru, *et al.*, Comparative performance of airyscan and structured illumination superresolution microscopy in the study of the surface texture and 3D shape of pollen. *Microsc. Res. Tech.* **81**, 101–114 (2018).

35. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).

36. C. E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).

37. I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).

38. M. Kuhn, Building predictive models in R using the caret package. *J. Stat. Softw.* **28**, 1–26 (2008).

39. R. F. Sage, Was low atmospheric CO_2 during the Pleistocene a limiting factor for the origin of agriculture? *Glob. Change Biol.* **1**, 93–106 (1995).

40. W. Jiang, Y. Cheng, X. Yang, S. Yang, Chinese Loess Plateau vegetation since the Last Glacial Maximum and its implications for vegetation restoration. *J. Appl. Ecol.* **50**, 440–448 (2013).

41. J. Cunniff, M. Charles, G. Jones, C. P. Osborne, Reduced plant water status under sub-ambient pCO_2 limits plant productivity in the wild progenitors of C_3 and C_4 cereals. *Ann. Bot.* **118**, 1163–1173 (2016).

42. Y.-F. Yao, *et al.*, Pollen-based reconstruction of vegetational and climatic change over the past ~30 ka at Shudu Lake in the Hengduan Mountains of Yunnan, southwestern China. *PLOS ONE* **12**, e0171967 (2017).

43. T. He, B. B. Lamont, J. G. Pausas, Fire as a key driver of Earth's biodiversity. *Biol. Rev.* **94**, 1983–2010 (2019).

44. R. R. L. Atkinson, *et al.*, C_4 photosynthesis boosts growth by altering physiology, allocation and size. *Nat. Plants* **2**, 16038 (2016).

45. J. E. Keeley, P. W. Rundel, Fire and the Miocene expansion of C_4 grasslands. *Ecol. Lett.* **8**, 683–690 (2005).

46. C. P. Osborne, D. J. Beerling, Nature's green revolution: The remarkable evolutionary rise of C_4 plants. *Philos. Trans. R. Soc. B Biol. Sci.* **361**, 173–194 (2006).

47. J. Cunniff, M. Charles, G. Jones, C. P. Osborne, Was low atmospheric CO_2 a limiting factor in the origin of agriculture? *Environ. Archaeol.* **15**, 113–123 (2010).

48. S. P. Long, C_4 photosynthesis at low temperatures. *Plant Cell Environ.* **6**, 345–363 (1983).

49. M. J. Wooller, D. L. Swain, F. A. Street-Perrott, S. Mathai, A. D. Q. Agnew, An altitudinal and stable carbon isotope survey of C_3 and C_4 graminoids on Mount Kenya. *J. East Afr. Nat. Hist.* **90**, 69–85 (2001).

50. F. A. Street-Perrott, *et al.*, Late Quaternary changes in ecosystems and carbon cycling on Mt. Kenya, East Africa: A landscape-ecological perspective based on multi-proxy lake-sediment influxes. *Quat. Sci. Rev.* **26**, 1838–1860 (2007).

51. O. Toljagić, K. L. Voje, M. Matschiner, L. H. Liow, T. F. Hansen, Millions of years behind: Slow adaptation of ruminants to grasslands. *Syst. Biol.* **67**, 145–157 (2018).

52. E. A. Kellogg, "Fossil record and dates of diversification" in *Flowering Plants. Monocots: Poaceae*, E. A. Kellogg, Ed. (Springer International Publishing, 2015), pp. 103–107.

53. Y. Wu, H.-L. You, X.-Q. Li, Dinosaur-associated Poaceae epidermis and phytoliths from the Early Cretaceous of China. *Natl. Sci. Rev.* **5**, 721–727 (2018).

54. T. J. Gallaher, *et al.*, Grasses through space and time: An overview of the biogeographical and macroevolutionary history of Poaceae. *J. Syst. Evol.* **60**, 522–569 (2022).

55. F. A. Street-Perrott, K. J. Ficken, Y. Huang, G. Eglinton, Late Quaternary changes in carbon cycling on Mt. Kenya, East Africa: An overview of the $\delta^{13}C$ record in lacustrine organic matter. *Quat. Sci. Rev.* **23**, 861–879 (2004).

56. K. Fægri, J. Iversen, P. E. Kaland, K. Krzywinski, *Textbook of Pollen Analysis*, ed. 4th (John Wiley & Sons, 1989).

57. M. A. Urban, I. C. Romero, M. Sivaguru, S. W. Punyasena, Nested cell strainers: An alternative method of preparing palynomorphs and charcoal. *Rev. Palaeobot. Palynol.* **253**, 101–109 (2018).

58. S. Xie, R. Girshick, P. Dollar, Z. Tu, K. He, Aggregated residual transformations for deep neural networks in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (IEEE, 2017), pp. 5987–5995.

59. K. Sohn, *et al.*, FixMatch: Simplifying semi-supervised learning with consistency and confidence in *Advances in Neural Information Processing Systems*, (Curran Associates, Inc., 2020), pp. 596–608.

60. B. W. Silverman, Density Estimation for Statistics and Data Analysis Chapter 1 and 2.

61. L. J. Revell, Size-correction and principal components for interspecific comparative studies. *Evolution* **63**, 3258–3268 (2009).

62. L. J. Revell, phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).

63. A. Liaw, M. Wiener, Classification and regression by randomForest. **2** (2002).

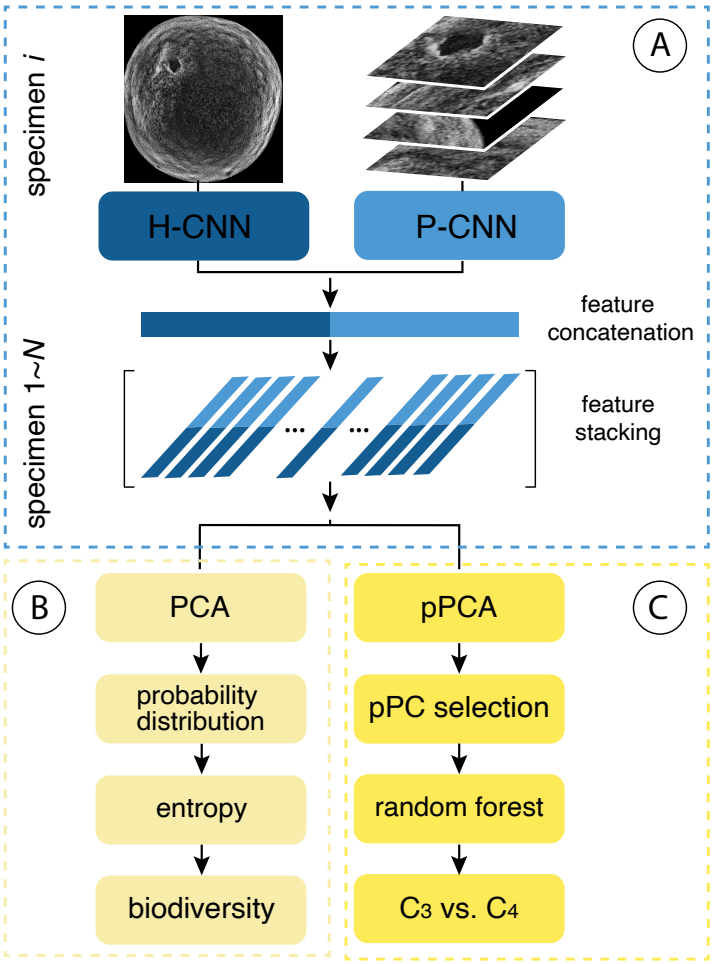
64. R. M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **SMC-3**, 610–621 (1973).

65. S. W. Punyasena, M. A. Urban, M.-É. Adaimé, I. C. Romero, SR-SIM superresolution structured illumination microscopy images of fossil and modern pollen of Poaceae. Champaign, IL, USA: University of Illinois Urbana-Champaign. doi: 10.13012/B2IDB-1391373_V1.

66. M.-É. Adaimé, paleopollen/Pollen_Diversity_Dynamics: Poaceae Pollen Diversity Dynamics. (2024). doi: 10.5281/zenodo.13756262. Deposited 13 September 2024.

680 **Figures**

681



682

683

684 Fig. 1. Flowchart illustrating the pipeline of computational analyses performed in this
685 study. Two representations of superresolution micrographs of a pollen specimen (the
686 maximum intensity projection of the image stack and crops of the projection) are passed
687 to convolutional neural networks (CNNs) trained to recognize the holistic morphology
688 (H-CNN) and surface texture of patches (P-CNN) of grass pollen. The feature values
689 extracted by both CNNs are concatenated to form a comprehensive feature vector
690 representing each specimen. These vectors are then stacked across all *N* specimens in
691 the sample, forming a feature matrix of all the concatenated feature vectors. In our
692 assessments of grass diversity, the PCA scores of the feature matrix are used to

estimate probability distributions for pollen assemblages using kernel density estimation. Shannon entropy is calculated over these probability distributions to provide a quantitative diversity estimate. In our assessments of photosynthetic pathway, we apply a phylogenetic PCA (pPCA) on the feature matrix and pass the most informative phylogenetic principal components (selected based on the modern data) to a random forest classifier trained to discriminate among C_3 and C_4 grasses.

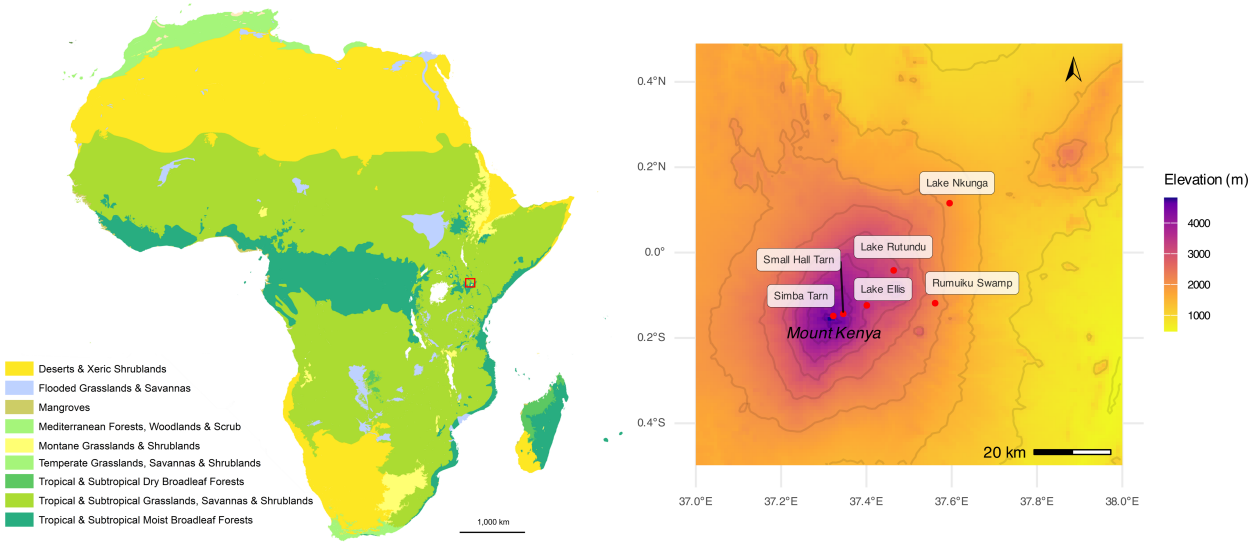


Fig. 2. Location of the study area. The left panel illustrates the distribution of major vegetation biomes across Africa, with Mt. Kenya highlighted by a red square. The right panel shows the elevation gradient in the Mt. Kenya region with the locations of our six sources of fossil grass pollen: the 25,000-year sediment record of Lake Rutundu, and the five additional lakes from which we extracted pollen from recently deposited surface sediments (Lake Nkunga, Rumuiku Swamp, Lake Ellis, Small Hall Tarn, and Simba Tarn).

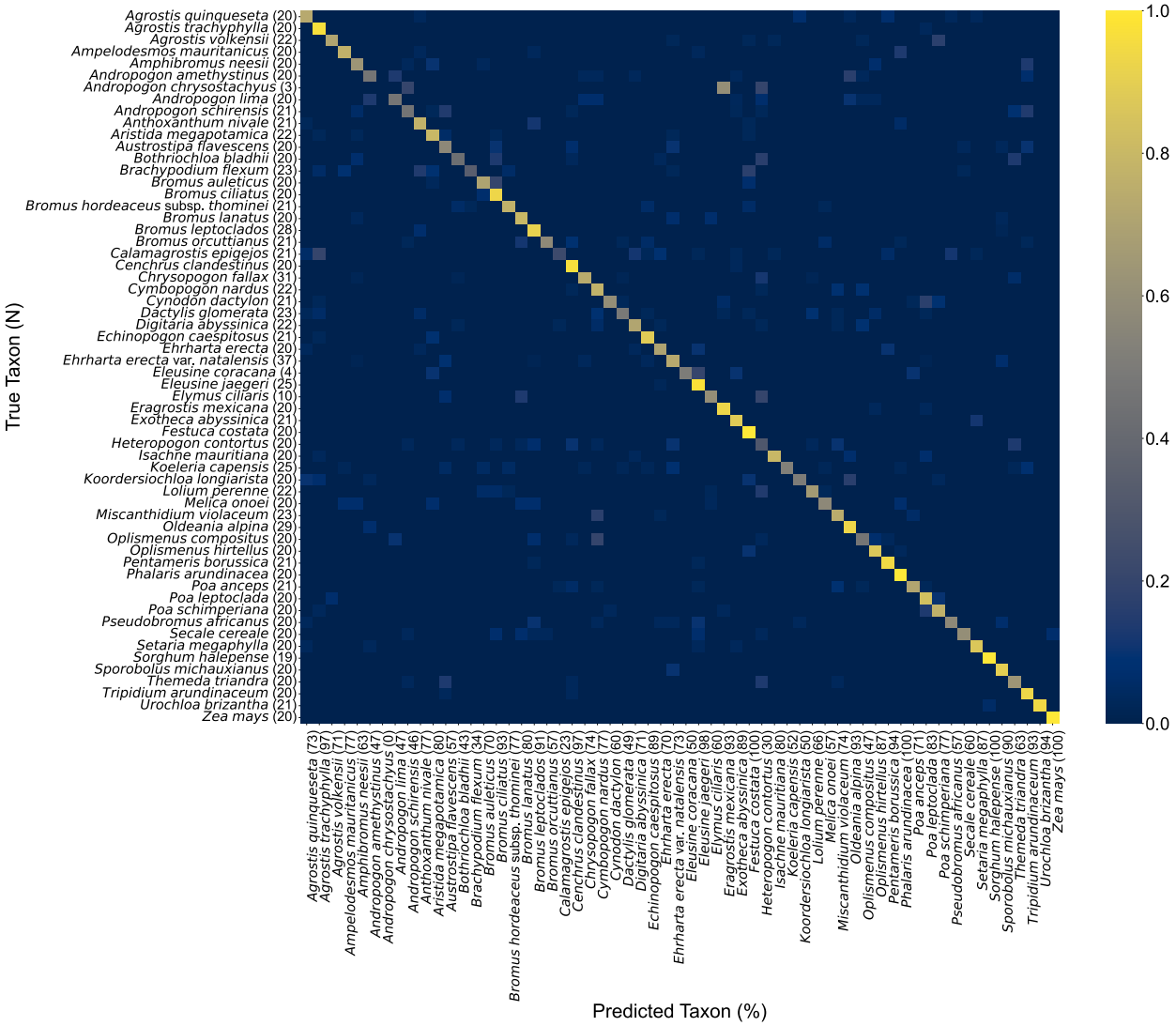


Fig. 3. Confusion matrix depicting the average accuracy of the fused CNN classification model for modern pollen grains from 60 Poaceae species. Rows represent the true taxa, with the number of specimens analyzed per species in parentheses. Columns represent the model's predictions, with the classification accuracy for each species in parentheses. Individual splits are shown in SI Appendix, Fig. S1.

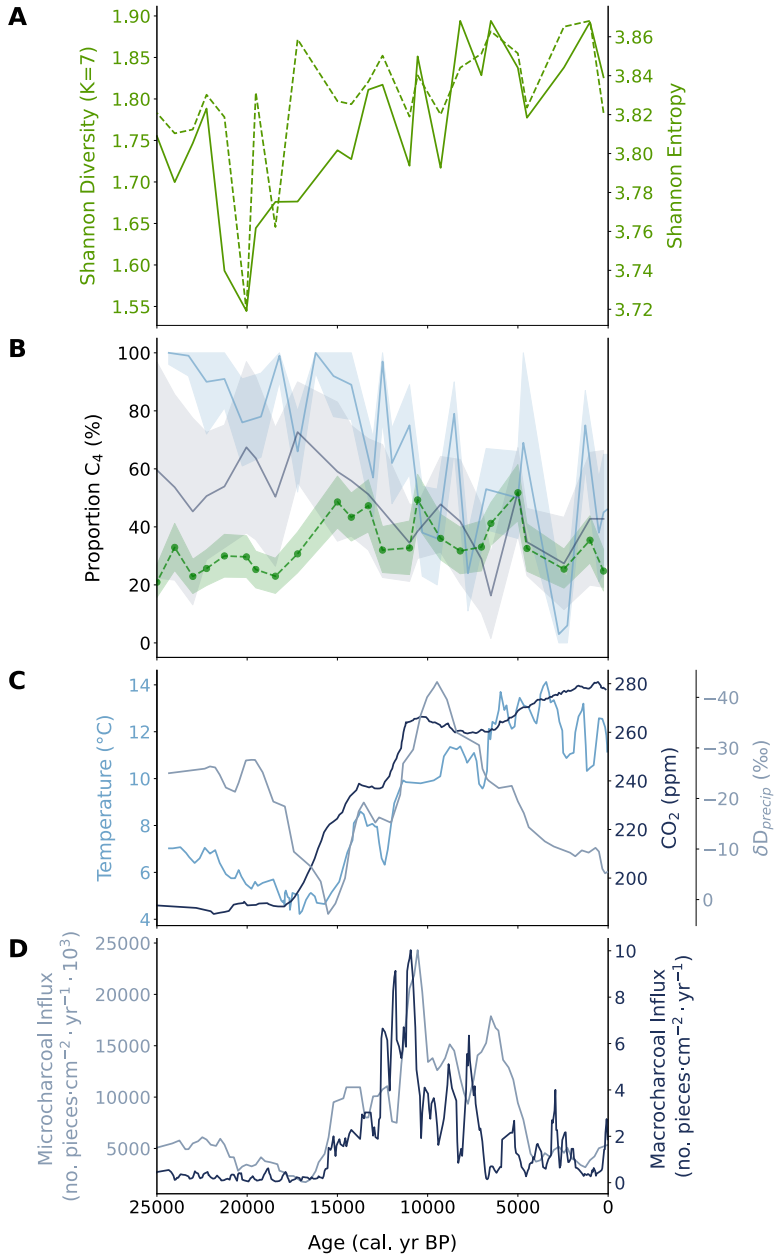


Fig. 4. Inferred changes in grass diversity and proportion of C_4 grasses in subalpine grasslands of Mt Kenya over the past 25,000 years, plotted alongside time series of paleoenvironmental proxies interpolated using moving average with a window size of 4. (A) Grass diversity as measured by Shannon entropy (dashed green line) and a derived Shannon index. (B) Proportion of C_4 grasses determined by this study (dashed green line) compared to results based on the carbon-isotopic signature ($\delta^{13}C$) of individual grass pollen grains (solid blue line, from (28)) and grass cuticle identification (solid purple line, from (27)). (C) The ice-core record of atmospheric CO_2 concentration (31)

plotted with reconstructions of temperature and precipitation on Mt. Kenya derived from, respectively, bacterial membrane lipids and the hydrogen-isotopic signature (δD_{precip}) of plant leaf-wax alkanes extracted from Lake Rutundu sediments (29, 30). More depleted δD_{precip} values indicate wetter conditions. (D) History of fire prevalence on Mt. Kenya reconstructed from estimates of biomass burning based on sedimentary fluxes of micro- and macrocharcoal (28).

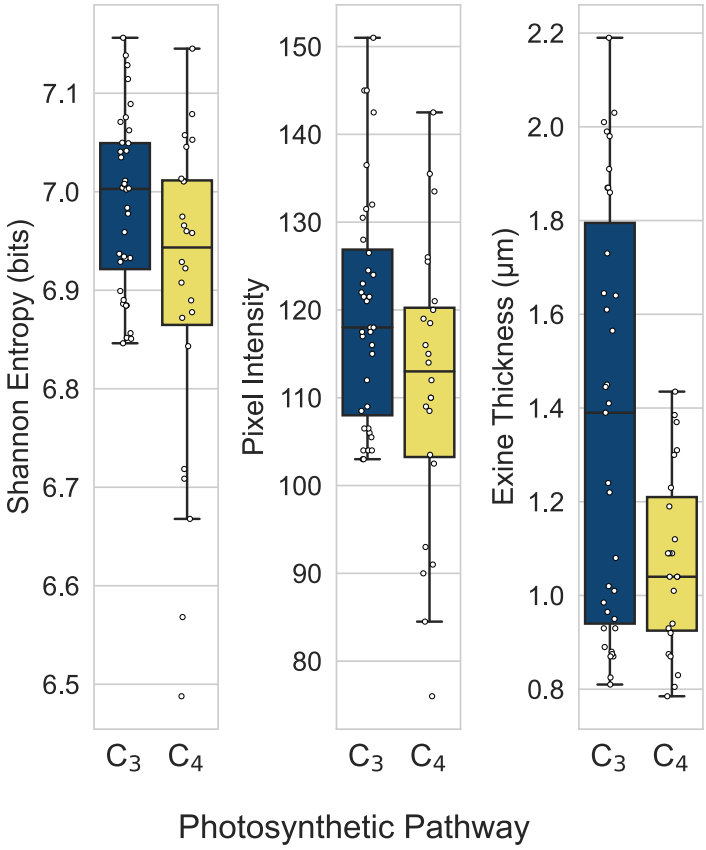


Fig. 5. Box plots showing the median difference in Shannon entropy and pixel intensity (brightness) between pollen from 35 C_3 and 25 C_4 grass species in our modern reference dataset and in exine thickness (µm) of 35 C_3 and 23 C_4 grass species (from (19)). On average, C_3 grass pollen has a significantly thicker exine wall and more complex surface patterning than C_4 grass pollen.