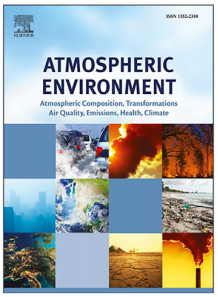
Journal Pre-proof



Vision Transformers for pollen grains classification: A systematic backbone benchmark on UCOPollen

Ricardo Espantaleón-Pérez, Qasim Farooq, Rocío López-Orozco, Mónica González-Alonso, Samer Alashhab, Jose Oteros, Jeroen Buters, Carmen Galán, Manuel J. Marín-Jiménez

PII: S1352-2310(26)00685-0
DOI: https://doi.org/10.1016/j.atmosenv.2026.122455
Reference: AEA 122455

To appear in: *Atmospheric Environment*

Received date : 5 July 2026
Revised date : 4 October 2026
Accepted date : 28 October 2026

Please cite this article as: R. Espantaleón-Pérez, Q. Farooq, R. López-Orozco et al., Vision Transformers for pollen grains classification: A systematic backbone benchmark on UCOPollen. *Atmospheric Environment* (2026), doi: https://doi.org/10.1016/j.atmosenv.2026.122455.

This is a PDF of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability. This version will undergo additional copyediting, typesetting and review before it is published in its final form. As such, this version is no longer the Accepted Manuscript, but it is not yet the definitive Version of Record; we are providing this early version to give early visibility of the article. Please note that Elsevier's sharing policy for the Published Journal Article applies to this version, see: https://www.elsevier.com/about/policies-and-standards/sharing#4-published-journal-article. Please also note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2026 Published by Elsevier Ltd.

Revised manuscript (unmarked, editable source file) Journal Pre-proof Click here to view linked References

Vision Transformers for Pollen Grains Classification: A Systematic Backbone Benchmark on UCOPollen

Ricardo Espantaleón-Pérez[a,*], Qasim Farooq[d,e], Rocío López-Orozco[d,e], Mónica González-Alonso[c], Samer Alashhab[c], Jose Oteros[d,e], Jeroen Buters[c], Carmen Galán[d,e], Manuel J. Marín-Jiménez[a,b]

[a] *Department of Computer Science and Artificial Intelligence, University of Córdoba (UCO), Córdoba, Spain*
[b] *Maimonides Institute for Biomedical Research of Córdoba (IMIBIC), Córdoba, Spain*
[c] *Center of Allergy & Environment (ZAUM), Member of the German Center for Lung Research (DZL), Technische Universität München/Helmholtz Center, Munich, Germany*
[d] *Department of Botany, Ecology and Plant Physiology, Agrifood Campus of International Excellence (ceiA3), University of Córdoba, Rabanales Campus, Celestino Mutis Building, Córdoba, Spain*
[e] *Andalusian Inter-University Institute for Earth System Research (IISTA), University of Córdoba, Córdoba, Spain*

Abstract

Automatic pollen identification is critical for aerobiology, environmental monitoring and allergy forecasting. We introduce UCOPollen, a 48,950-image dataset covering 17 pollen taxa and one negative class, and benchmark eleven backbones spanning convolutional (CNN), hybrid-attention and self-supervised Vision Transformer (ViT) families. Phase I compares frozen representations by linear probing. Phase II fully fine-tunes the strongest model from each family under an equal budget to test whether that ranking persists when optimisation is controlled. Phase II-B adds a hybrid Cross-Entropy–Triplet objective and compact projection to select a model for closed-set classification and gallery-based recognition of unseen taxa without retraining. Phase III, a post hoc analysis on a date-disjoint partition, tests robustness to

*Corresponding author.
Email address: i92esper@uco.es (Ricardo Espantaleón-Pérez)

Journal Pre-proof

Journal Pre-proof

held-out acquisition dates. Frozen probing yields a macro-F1 range (0.668–0.832), led by DINOv2-ViT Base; after equal-budget fine-tuning, the spread collapses to 0.0043 and no pairwise difference is detectable (corrected $p \geq 0.247$). Thus, training regime rather than architecture governs closed-set performance. We introduce PollenFormerUCO, the DINOv2-ViT Base hybrid configuration ($d = 128$, $\lambda = 0.5$), with backbone parameters trainable under differential learning rates. It reaches 98.1% accuracy and 97.5% macro-F1 using cosine 1-nearest-neighbour (1-NN) inference on a 128-dimensional embedding. Gallery extension lets seven unseen taxa be added and queried without retraining; we evaluate out-of-distribution material. Phase III (96.1% accuracy and 88.6% macro-F1 over 17 classes), repeated runs and a cross-entropy control delimit these claims and indicate that retrieval capability derives from the projection head, not uniquely from the triplet term. The dataset, trained weights, code and online demo will be available through the project repository: github.com/AVAuco/pollenformerUCO.

Keywords: Automatic pollen classification, Vision Transformers, Metric learning, Open-set recognition, Deep learning for bioimage analysis, Microscopy images

1. Introduction

Atmospheric pollen monitoring matters for public health and for climatological and ecological research: allergies affect roughly a quarter of the European population (Šaulienė, I. et al., 2019), and real-time taxon-level data enables preventive action (Buters et al., 2024). Traditional Hirst-type traps (Hirst, 1952) require expert microscopic identification and typically take 7 to 10 days to produce results (González-Alonso et al., 2023), which is what motivates automation.

Image-recognition systems such as the BAA500 (*Helmut Hund GmbH*) are among the most successful technologies to date, with operational networks reporting high

Journal Pre-proof

multiclass accuracy and strong correlation with manual counts (Oteros et al., 2015; González-Alonso et al., 2023). Yet a large share of grains goes unclassified, and morphologically similar taxa remain hard to discriminate without adequate training data (Oteros et al., 2015). These difficulties are sharper in Mediterranean environments, with their high botanical diversity, strong seasonality and complex mixtures of pollen with dust, spores and fragmented biological material (Zhang et al., 2016; López-Orozco et al., 2023; Boullayali et al., 2025); in Córdoba, correlations with Hirst measurements vary substantially among taxa (Farooq et al., 2026). While automatic monitoring expands across Europe, models trained and evaluated under Mediterranean conditions remain scarce, and so do the datasets needed to build them.

Hand-crafted features generalised poorly to the natural variability of airborne samples (Konijn et al., 2025), motivating the move to Convolutional Neural Networks (CNNs), on which most pollen-recognition work still relies (Rostami et al., 2023; Shamrat et al., 2024). Systems such as PollenNet provide valuable domain context, but their metrics come from different datasets, taxonomies and protocols and are not directly comparable. No study has benchmarked CNN and Vision Transformer (ViT) architectures for pollen classification under identical training conditions, nor combined them with metric learning to structure the embedding space explicitly.

This work addresses that gap, and its contributions are:

1. UCOPollen, a dataset of 48,950 expert-annotated microscopic images spanning 17 Mediterranean pollen taxa and one negative class, released with its evaluation partitions and acquisition metadata.
2. A controlled three-paradigm benchmark of eleven backbones — classical and modern CNNs, attention-based and hybrid Transformers, and self-supervised Vision

Journal Pre-proof

Transformers — under a common linear probe, followed by full fine-tuning of the strongest of each family under an identical optimisation budget, with Cross-Entropy and with a hybrid Cross-Entropy and Triplet objective. Under Cross-Entropy, the training regime, not the architecture, governs closed-set performance.

3. PollenFormerUCO, a DINOv2-ViT Base model with a learned 128-D projection trained under a hybrid Cross-Entropy and Triplet objective. Its cosine nearest-neighbour readout attains the highest macro-F1 of the study, and a cross-entropy control shows that the learned projection is what makes the embedding queryable.
4. An open-set evaluation in which taxa unseen during training are added to this embedding through gallery extension alone, without retraining, characterising where that capability holds and where it fails.

2. Material and methods

This section describes the dataset (Section 2.1), the architectures evaluated (Section 2.2), the loss functions (Section 2.3), the training protocol (Section 2.4) and the evaluation protocol (Section 2.5). The experiments are organised into four phases, specified in Section 2.4: Phase I compares the eleven backbones as frozen feature extractors under a linear probe; Phase II fine-tunes every parameter of the strongest backbone of each family under an identical budget with Cross-Entropy; Phase II-B repeats Phase II with the hybrid objective of Section 2.3.3; and Phase III, a post hoc robustness analysis, retrains the Phase II-B configurations on a date-disjoint partition.

2.1. Dataset

UCOPollen was provided by the Department of Botany, Ecology and Plant Physiology at the University of Córdoba. It contains 48,950 microscopic images spanning

Journal Pre-proof

17 pollen taxa and one negative class (NoPollen), each manually annotated by domain experts; Figure 1 shows a representative image of every class alongside its composition and acquisition metadata.

Airborne particles were collected with a BAA502 (POMO) automatic bioaerosol analyser, the latest version of the BAA500 system (Helmut Hund GmbH, Wetzlar, Germany). Particles are impacted onto an adhesive plate and scanned by an integrated microscope-camera system; every 3 h the acquired images are processed into synthetic 2D greyscale images, segmented into individual objects, and assigned to broad categories before biological classification (Oteros et al., 2015). Each image corresponds to one particle, so the same grain is never represented twice. Most taxa are ambient samples; three (*Morus*, *Pistacia*, *Chenopodium*) were enriched by introducing reference material during controlled feeding sessions, which underwent the same expert curation.

NoPollen is the largest and most heterogeneous class, grouping images that curation judged unsuitable for assignment to a taxon: mainly fragments, dust, spores and other non-biological particles, together with overlapping grains and mis-segmented images. Its composition was not quantified, because the archived records retain the original segmentation category for only 1.1% of these images. Since a qualitative description cannot establish whether the high accuracy on this class reflects genuine discrimination or an abundance of easy negatives, Section 3.8 evaluates the model on its most pollen-like members specifically.

2.1.1. Composition, partition and acquisition metadata

The dataset was divided by stratified split into 70% training, 15% validation and 15% testing — 34,257 / 7,335 / 7,358 images — with the same proportions in every class (Figure 1a). Support spans a factor of 46, from 160 images for *Castanea* to

Journal Pre-proof

Journal Pre-proof

7,403 for NoPollen.

Hourly acquisition timestamps were recovered for 32,255 images (65.9%). Among them, 191 of the 257 distinct acquisition dates appear in all three subsets and 59.8% of same-hour same-class groups span more than one partition. No image appears in more than one subset, but images sharing an acquisition context do co-occur across splits. Coverage is uneven — 1.1% of NoPollen and about a quarter of *Pistacia* and *Chenopodium* carry a timestamp, against 56–97% for the rest — and dated acquisition is concentrated: *Casuarina* comes entirely from one date and *Populus* draws 54.5% from a single one (Figure 1b–c). Section 2.4.4 builds the strongest date-disjoint partition these metadata support.

2.1.2. Dataset Preprocessing

Images were resized to the square input resolution of each pretrained checkpoint — 256 × 256 pixels for MobileViT-S and 224 × 224 for the other backbones — by bilinear interpolation, before augmentation and without padding or cropping. Most images are close to square (mean aspect ratio 1.03, median 1.01), so the resize changes the typical image negligibly; the extremes (0.15 to 7.0) are rare. Grayscale images are replicated to three channels.

The `timm` models use the normalisation registered with each checkpoint (Wightman, 2019): the standard ImageNet mean and standard deviation (Deng et al., 2009) for most of them, and the $[0, 1]$ range for MobileViT-S. The DINOv2 models consume tensors in $[0, 1]$ ($\mu = 0$, $\sigma = 1$), the convention under which PollenFormerUCO is trained and served, whereas the reference DINOv2 transform applies ImageNet normalisation (Oquab et al., 2024). Every run report records the normalisation used.

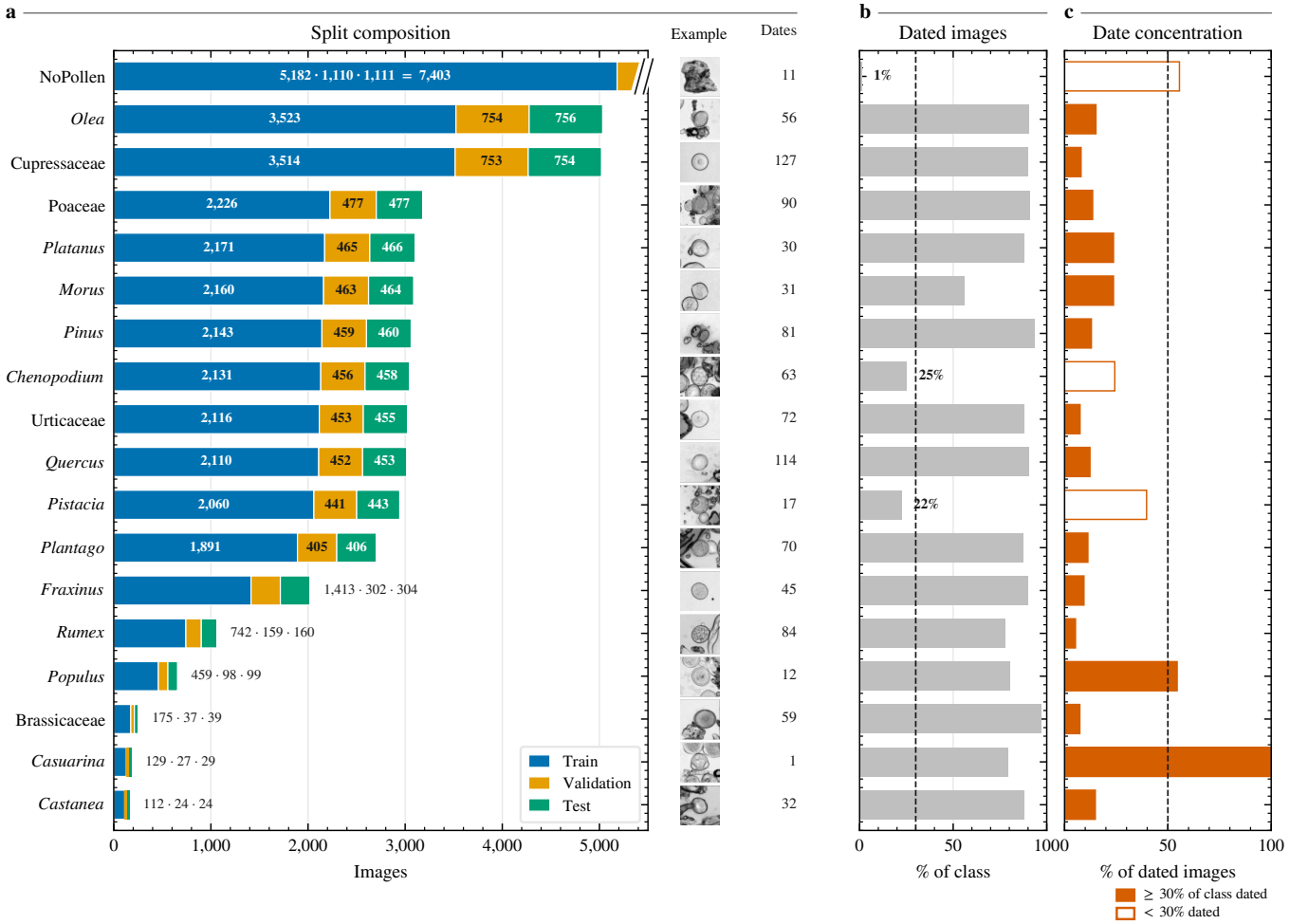


Figure 1: **UCOPollen composition and acquisition concentration.** Rows are classes, ordered by support. (a) Train/validation/test counts; NoPollen is drawn broken at 5,500 and states its total inline. *Example* shows one expert-annotated image per class and *Dates* the number of distinct acquisition dates. (b) Share of each class carrying a recovered timestamp. (c) Share of those dated images contributed by the largest single date; hollow markers flag the three classes below 30% timestamp coverage.

Journal Pre-proof

2.2. Architectures evaluated

Eleven backbones were evaluated: six attention-based or hybrid Transformers and five CNNs. DaViT (Ding et al., 2022) combines spatial and channel attention; DeiT (Touvron et al., 2021) adds distillation from a CNN teacher for data-efficient training; ConViT (d'Ascoli et al., 2022) uses gated positional self-attention to embed convolution-like inductive biases; MobileViT (Mehta and Rastegari, 2022) is a lightweight hybrid for constrained hardware; and DINOv2 (Oquab et al., 2024) is a family of ViTs trained without labels on LVD-142M, a curated collection of 142 million images. On the convolutional side, ResNet18 (He et al., 2016) is the classical residual baseline, ConvNeXt (Liu et al., 2022) modernises a purely convolutional stack, and EfficientNetV2 (Tan and Le, 2021) combines compound scaling with fused inverted-residual blocks. ConvNeXt, EfficientNetV2 and DINOv2 are each evaluated at two capacities, so that capacity and family can be distinguished.

Exact checkpoint identifiers are given in Table A.1. The `timm` checkpoints were trained with supervision on ImageNet-1K, whereas DINOv2 was trained without labels on LVD-142M, so the benchmark compares specific pretrained representations under a common downstream protocol and does not isolate architecture from pretraining data or objective; Section 3.2 addresses that confound.

2.3. Loss Functions

Two training objectives were evaluated: the standard Cross-Entropy loss for closed-set classification, and a hybrid objective that combines Cross-Entropy with a metric-learning term to encourage a more structured embedding space.

Journal Pre-proof

2.3.1. Cross-Entropy Loss

Given logits $\mathbf{z} \in \mathbb{R}^C$ and a ground-truth label $y \in \{1, \ldots, C\}$, the Cross-Entropy (CE) loss is

$$\mathcal{L}_{\mathrm{CE}}(\mathbf{z}, y) = -\log \frac{\exp(z_y)}{\sum_{c=1}^{C} \exp(z_c)}.$$

where z_y is the logit of the true class.

2.3.2. Triplet Loss

Let $f(\mathbf{x}) \in \mathbb{R}^d$ be the embedding produced by the model. The CLS token is projected from 768 dimensions to a compact d-dimensional space by a linear layer (Figure 2). This mitigates the curse of dimensionality (Aggarwal et al., 2001) and reduces the cost of non-parametric retrieval. The projection dimension is selected per model (Section 2.4); PollenFormerUCO uses $d = 128$. Given an anchor $\mathbf{x}_a$, a positive $\mathbf{x}_p$ of the same class and a negative $\mathbf{x}_n$ of another,

$$\mathcal{L}_{\mathrm{Triplet}} = \max\left\{0,\ \|f(\mathbf{x}_a) - f(\mathbf{x}_p)\|_2 - \|f(\mathbf{x}_a) - f(\mathbf{x}_n)\|_2 + m\right\},$$

with margin $m = 1.0$. Triplets already satisfying the constraint contribute zero loss.

Triplets are sampled uniformly within each batch, without semi-hard or hard mining: aggressive selection of the hardest negatives can drive optimisation into poor local minima and collapse the embedding (Schroff et al., 2015), whereas uniform sampling gives an unbiased estimate of the objective and, having no tunable policy, keeps the backbone comparison free of any interaction between mining and architecture. Every sample in the batch serves in turn as anchor; the positive is drawn uniformly from the remaining samples of the anchor's class *within the batch*; the negative class is drawn uniformly from the other classes present, and the negative sample uniformly within it. An anchor is used only when its class contributes at least two samples to the batch, so every triplet rests on a genuine same-class pair. Embeddings are

Journal Pre-proof

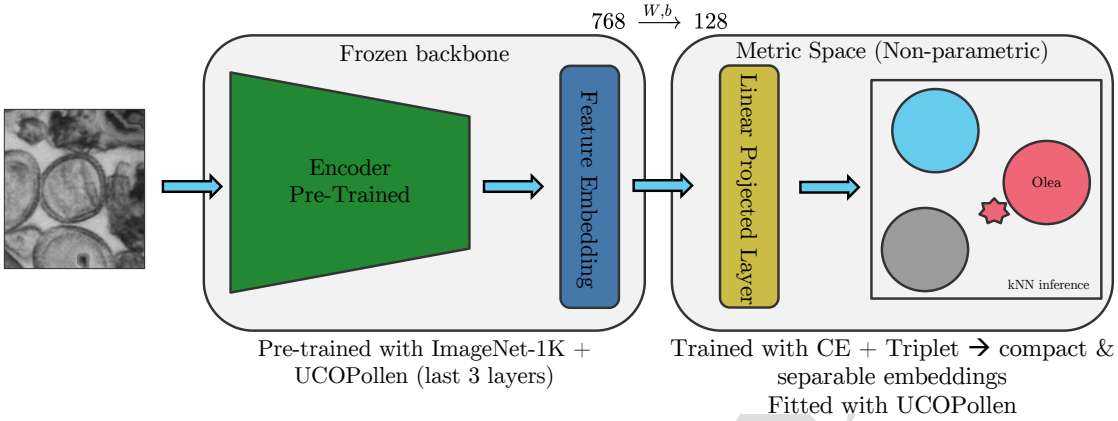


Figure 2: **Hybrid objective architecture.** Workflow of PollenFormerUCO under full fine-tuning. Every parameter of the self-supervised pretrained DINOv2 backbone (left) is trainable (Section 2.4). The resulting 768-D representation is projected to a unit-normalised 128-D metric space and optimised with Cross-Entropy plus Triplet loss ($\lambda = 0.5$). In the metric space (right), coloured circles represent class clusters and the pink star denotes a query assigned by cosine 1-nearest-neighbour (1-NN) classification, enabling gallery extension without retraining, as in *BAA500* (González-Alonso et al., 2023).

L_2-normalised before distances are computed, in training and at nearest-neighbour inference alike.

2.3.3. Hybrid Objective

In Phase II-B, each backbone is trained with a hybrid loss that combines Cross-Entropy with the Triplet term (Figure 2):

$$\mathcal{L}_{\mathrm{H}} = (1 - \lambda)\,\mathcal{L}_{\mathrm{CE}} + \lambda\,\mathcal{L}_{\mathrm{Triplet}}.$$

At $\lambda = 0$ the metric term vanishes and the expression reduces to Cross-Entropy. The weight selected for every Phase II-B model is non-zero (Table 1); for PollenFormerUCO, $\lambda = 0.5$ gives equal weight to the supervised and metric terms.

Journal Pre-proof

2.4. Training protocol

The four phases introduced above have distinct inferential purposes. Phase I evaluates all eleven checkpoints as frozen feature extractors under a common linear probe, with head capacity held constant. Phase II fine-tunes *every* parameter of the strongest member of each family — ConvNeXt-Base, DaViT-Small and DINOv2-ViT Base — under an identical budget, which removes the confound between architecture and optimisation budget and is the basis for comparative statements about architectures. Phase II-B repeats Phase II with the hybrid objective and a learned projection. Phase III is a separate, post hoc robustness analysis: it retrains the Phase II-B configurations, with their hyperparameters unchanged, on a date-disjoint partition in which no acquisition date is shared between training and evaluation (Section 2.4.4).

All phases use AdamW, batch size 64 and seed 42, with up to 50 epochs for final fits and a 10-epoch budget for each grid-search cell. Augmentation is the same biologically admissible geometric set throughout — horizontal and vertical flips at probability 0.5 and rotation within $\pm 30°$, without colour augmentation. Checkpoints are selected by *maximum validation macro-F1* with early stopping (patience 10), never at a fixed epoch, and the test partition is evaluated exactly once, after hyperparameter and epoch selection. Stratified 5-fold cross-validation (CV) on the training partition, with the selected hyperparameters held constant, complements the final test evaluation.

2.4.1. Phase I: frozen-backbone linear probing

One randomly initialised linear head is trained on a frozen backbone, so there is a single rate to set. The grid comprises learning rate (LR) $\in \{10^{-3}, 10^{-4}\}$ and weight decay $\in \{0, 10^{-2}\}$; every checkpoint selected 10^{-3}.

Journal Pre-proof

2.4.2. Phase II: equal-budget full-backbone fine-tuning

Every backbone parameter is trainable together with a new head. Backbone and head form separate AdamW parameter groups whose learning rates are searched independently, so that the pretrained backbone can adapt more conservatively than the randomly initialised head. Both rates are searched over $\{10^{-3}, 10^{-4}, 10^{-5}\}$ and weight decay over $\{0, 10^{-2}\}$. All three models select a head rate of 10^{-3}; the backbone rate is 10^{-4} for ConvNeXt-Base and DaViT-Small and 10^{-5} for DINOv2-ViT Base, and weight decay is 0 for ConvNeXt-Base and 10^{-2} for the other two.

2.4.3. Phase II-B: hybrid objective under full fine-tuning

Every parameter is trainable together with the projector and classifier, which share the head group. The optimiser search of Phase II is reused, and for each model the projection dimension and triplet weight are selected by validation macro-F1 over $d \in \{64, 128, 256\}$ and $\lambda \in \{0.25, 0.5, 0.75\}$ at the 10-epoch budget. The selected rates are those of Phase II — backbone 10^{-4} for ConvNeXt-Base and DaViT-Small, 10^{-5} for DINOv2-ViT Base, and 10^{-3} for every head/projector — with weight decay 0 for the two supervised backbones and 10^{-2} for DINOv2-ViT Base. The selected (d, λ) and stopping epoch of each model are reported in Table 1(b).

Run-time assertions verify that every parameter is trainable in the final Phase II and Phase II-B fits and that the partition sizes are {34,257; 7,335; 7,358}. Phases I to II-B deliberately reuse this image-level partition to isolate the optimisation budget; Section 2.4.4 addresses group-level independence separately.

2.4.4. Phase III: exploratory date-disjoint (group-aware) analysis

Phases I to II-B use the image-level partition, in which images sharing an acquisition date are distributed across training, validation and test (Section 2.1); they measure classification of held-out *images*, not robustness to held-out acquisition dates.

Journal Pre-proof

Phase III is a post hoc sensitivity analysis in which whole acquisition dates are assigned to partitions, so no calendar date is shared between training and evaluation. Only the 65.9% of images carrying a recovered timestamp are eligible. *Casuarina* was captured on a single date and is therefore excluded, leaving 17 classes; NoPollen contributes only 81 timestamped images and is retained to keep the pollen/non-pollen decision in the task, although its per-class scores in this phase are uninformative. The partition holds 21,065 / 5,666 / 5,377 images over 193 / 36 / 27 disjoint dates (Appendix B, Table B.1); the manifests will be published with the code.

Phase III reuses the architectures, objective and selected hyperparameters of Phase II-B unchanged, with no grid search on the grouped partition, so only the partition changes and nothing is tuned on its evaluation dates. Repeated seeds replace cross-validation, which a grouped partition does not support without regrouping every fold.

2.5. Evaluation Protocol

Validation accuracy and macro-F1 are computed from the softmax logits at the end of every epoch, and the validation loss uses $\mathcal{L}_{\mathrm{CE}}$ only, so that model selection is comparable across objectives and with prior work evaluated under Cross-Entropy (Rostami et al., 2023; Shamrat et al., 2024). Four metrics are reported: accuracy, precision, recall and macro-F1. Macro-F1 is the headline metric, because it weights every taxon equally and is not dominated by the abundant classes (Figure 1a); micro-F1, used by related studies, equals accuracy in this single-label setting.

Journal Pre-proof

3. Results

3.1. Phase I: frozen-backbone linear-probe benchmark

Figure 3 reports the linear probe over all eleven backbones, with exact values in Table A.2. DINOv2-ViT Base achieves the highest test macro-F1 (0.8323), and the total spread is wide — 0.6682 to 0.8323, a range of 0.1641. Within the ImageNet-1K-supervised subset the ordering does not separate the architectural families cleanly: DaViT-Small ranks highest, and neither convolutional family improves monotonically with capacity, with ConvNeXt-Small and EfficientNetV2-M placing below several smaller Transformers.

These outcomes do not support a general claim that Transformers outperform CNNs. They show that DINOv2-ViT Base provides the most linearly separable frozen representation of this data, an advantage that reflects its self-supervised pre-training on LVD-142M (Oquab et al., 2024) as much as its architecture. Section 3.2 shows how strongly this ranking depends on the training regime.

3.2. Phase II: equal-budget full-backbone fine-tuning

A linear probe compares frozen representations, so its ranking may reflect how linearly separable each pretrained space happens to be rather than what the architecture can learn. Phase II therefore fine-tunes the strongest member of each family with every parameter trainable under an identical budget (Section 2.4). Two results follow.

First, the regime matters enormously. Every finalist improves by 0.138 to 0.162 macro-F1 over its own linear probe, removing roughly 82–86% of the residual error, and fold dispersion falls from 0.0037–0.0107 to 0.0017–0.0034. Every Phase II and Phase II-B fit stopped early, with no gain in validation macro-F1 over its last ten epochs (Figure C.1), whereas six linear probes selected an epoch after 40 and so

Journal Pre-proof

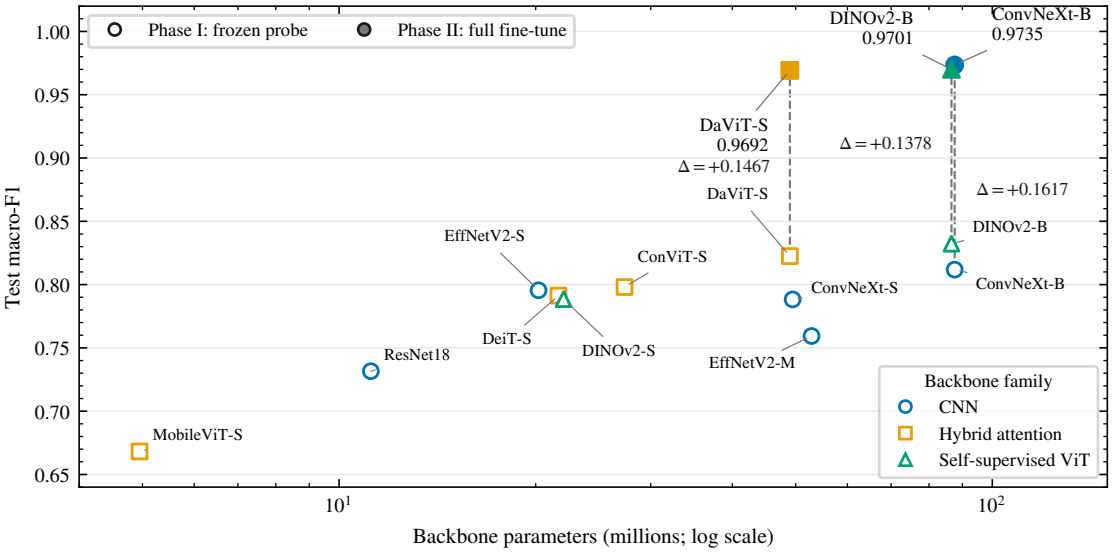


Figure 3: **Comparison of backbones.** Test macro-F1 against parameter count for all eleven frozen-backbone probes (Phase I) and the three matched equal-budget fine-tuned finalists (Phase II). Every Phase I point is labelled; lines connect the finalists and report their descriptive held-out test gains. Marker shape identifies family and fill identifies phase. Exact values are reported in Tables A.2 (Phase I) and 1(a) (Phase II).

reached the 50-epoch ceiling (Table A.2); their frozen baselines may therefore be slightly understated. Second, no architectural difference is detectable: the 0.1641 spread across eleven frozen backbones collapses to 0.0043 across the three finalists (Table 1a), and the ranking inverts. ConvNeXt-Base now leads with 0.9735 macro-F1 and 87.58 M parameters, followed by DINOv2-ViT Base (0.9701, 86.59 M) and DaViT-Small (0.9692), which stays within the cross-validation dispersion of the best model with 48.99 M parameters. This conclusion holds within Phase II's cross-entropy protocol and image-level partition; it does not extend to the hybrid retrieval protocol or to open-set deployment.

Journal Pre-proof

Table 1: **Equal-budget full-backbone fine-tuning.** Every parameter trainable, image-level partition (34,257 / 7,335 / 7,358), identical augmentation, a 50-epoch ceiling with patience 10 and one evaluation on the held-out test partition. (a) Cross-entropy. (b) Hybrid objective, read through the softmax head and through a cosine 1-NN classifier on the projected embedding; silhouette (Sil.) and Davies-Bouldin (DB) are computed on the projected test embeddings. Bold marks the best value per column; no pairwise difference in (a) is statistically detectable (Section 3.5).

(a) Phase II: cross-entropy.

Model	Pretraining	Trainable (M)	Epoch	CV $F1_{macro}$	Test acc.	Test $F1_{macro}$
ConvNeXt-Base	IN-1K sup.	87.58	38	$\mathbf{0.9701 \pm 0.0034}$	**0.9795**	**0.9735**
DINOv2-ViT Base	LVD-142M SSL	86.59	18	0.9684 ± 0.0024	0.9773	0.9701
DaViT-Small	IN-1K sup.	48.99	18	0.9697 ± 0.0017	0.9757	0.9692
Spread across the three finalists				0.0017	0.0038	0.0043

Note: SSL = self-supervised learning; sup. = supervised learning. Learning rates and weight decay are given in Section 2.4. The partition is identical to that used in Phase II-B, so the two phases are directly comparable.

(b) Phase II-B: hybrid objective.

					Softmax head		1-NN (cosine)		Embedding	
Model	d	λ	Epoch	CV $F1_{macro}$	Test acc.	$F1_{macro}$	Test acc.	$F1_{macro}$	Sil. ↑	DB ↓
DINOv2-ViT Base†	128	0.50	39	$\mathbf{0.9668 \pm 0.0025}$	**0.9772**	**0.9707**	**0.9807**	**0.9748**	0.8091	0.6613
ConvNeXt-Base	64	0.25	13	0.9631 ± 0.0027	0.9709	0.9572	0.9719	0.9630	0.8193	0.5601
DaViT-Small	128	0.25	26	0.9592 ± 0.0025	0.9705	0.9601	0.9694	0.9621	**0.8254**	**0.5121**

Note: †Selected configuration (PollenFormerUCO). For every model the 1-NN readout exceeds its own softmax head.

Journal Pre-proof

Journal Pre-proof

Journal Pre-proof

3.3. Phase II-B: hybrid objective under full fine-tuning

The hybrid objective was applied to all three families under full fine-tuning, with hyperparameters and stopping epochs selected by validation macro-F1 (Table 1b). DINOv2-ViT Base attained the highest cross-validation mean (0.9668 ± 0.0025); this validation evidence, together with its queryable 128-D representation, determined the checkpoint carried forward, and the held-out test was used once, for final reporting.

Model selected for the remainder of the paper. The selected DINOv2-ViT Base configuration ($d = 128$, $\lambda = 0.5$; optimiser settings in Section 2.4) is referred to as PollenFormerUCO. On the held-out test, its cosine 1-NN readout reaches 0.9807 accuracy and 0.9748 macro-F1, the highest macro-F1 of the study, and its softmax head is more accurate than both supervised backbones trained under the same objective (Section 3.5). This is selection under the hybrid training regime, not evidence that DINOv2 is universally superior: the Phase II cross-entropy finalists are statistically indistinguishable, and the date-disjoint analysis does not separate this model from the projection-head control. Unless stated otherwise, every subsequent analysis characterises this checkpoint.

3.4. What the hybrid protocol provides: a discriminable, queryable embedding

The accuracy tables compare classifiers; they do not capture the geometry of the space the classifier reads from, which is what allows a gallery of reference images to be queried and extended with new taxa without retraining (Section 3.9). Figure 4 first separates two questions. Hybrid 1-NN exceeds its corresponding softmax macro-F1 by 0.0041 for DINOv2, 0.0058 for ConvNeXt and 0.0020 for DaViT. Against the Phase II CE baseline, however, it is higher only for DINOv2 (+0.0047) and lower for ConvNeXt (−0.0105) and DaViT (−0.0071). The hybrid protocol therefore pro-

Journal Pre-proof

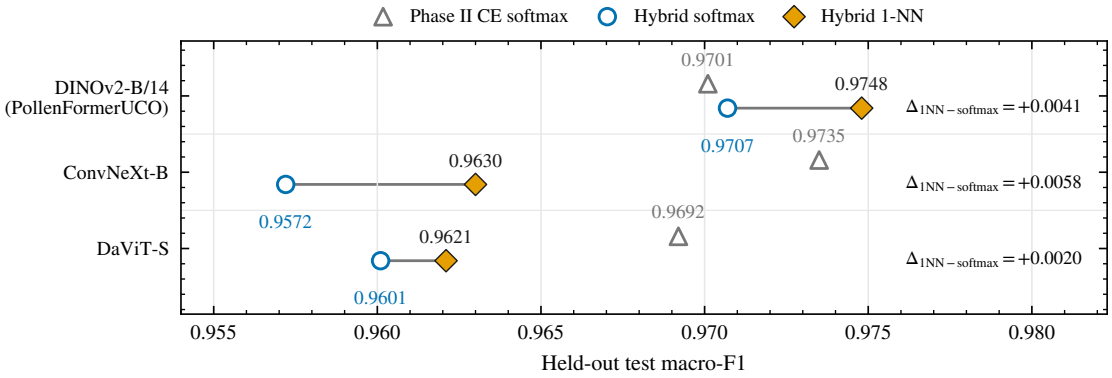


Figure 4: **Closed-set and retrieval readouts across training protocols.** Held-out test macro-F1 for the Phase II CE softmax baseline and both readouts of each fixed Phase II-B checkpoint. Lines join only the softmax and 1-NN readouts of the same hybrid embedding. Within each row the CE baseline (triangle) is drawn slightly above the hybrid readouts so that close values stay legible; the vertical offset carries no information. Although 1-NN improves every hybrid softmax result, the hybrid protocol does not uniformly exceed the CE baseline. These descriptive contrasts were not used for selection; Table 1 reports the configurations.

vides a compact, queryable representation, not a universal closed-set gain; because it changes both the objective and the projection, cross-phase differences are descriptive.

3.4.1. Separability and non-parametric classification

Table 2 reports silhouette (Rousseeuw, 1987), Davies–Bouldin (Davies and Bouldin, 1979) and Calinski–Harabasz (Caliński and Harabasz, 1974) indices on the held-out test partition, computed after $L2$ normalisation for two spaces of the selected PollenFormerUCO checkpoint: the 768-D CLS token immediately before the learned projection, and the 128-D projected embedding. Because both are read from the same forward pass, the comparison isolates the effect of the projection within one trained model. The projection improves every index, raising silhouette from 0.617 to 0.809 and Calinski–Harabasz from 938.6 to 3221.4, while cosine 1-NN macro-F1

Journal Pre-proof

Table 2: **Embedding geometry of PollenFormerUCO.** Separability of the held-out test partition (n = 7,358) before and after the learned 768-to-128-D projection, read from the same selected checkpoint and quantified with Silhouette (Rousseeuw, 1987) (cosine), Davies-Bouldin (DB) (Davies and Bouldin, 1979) and Calinski-Harabasz (CH) (Caliński and Harabasz, 1974) on $L2$-normalised embeddings. Arrows give the favourable direction; bold indicates the best value per column. Classification readouts are reported in Table 1(b).

Space	Silhouette ↑	DB ↓	CH ↑
Pre-projection CLS (768-D)	0.617	1.116	938.6
Projected embedding (128-D)	**0.809**	**0.661**	**3221.4**

is unchanged (0.9752 before and 0.9748 after projection). Compressing the representation six-fold therefore costs nothing in discriminative power while making the space markedly better separated — which is what makes an external reference gallery practical (González-Alonso et al., 2023).

3.4.2. Visual structure

Figure 5 projects both embeddings with Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018), using cosine distance, `n_neighbors` = 15, `min_dist` = 0.1 and seed 42. These settings were fixed after inspecting a prespecified qualitative grid over distance metrics {Euclidean, cosine, Chebyshev}, `n_neighbors` $\in \{5, 15, 50\}$ and `min_dist` $\in \{0.1, 0.5, 0.9\}$: cosine matches the retrieval rule, and 15 neighbours with 0.1 preserve local morphology without fragmenting the dominant classes. They were applied to both spaces and not selected by predictive performance. Because each UMAP is fitted separately, axes and distances are not comparable across panels, and every quantitative conclusion rests on Tables 2 and 1. Both objectives separate the taxa, but the hybrid space shows tighter intra-class groups and wider margins, most visibly for the minority taxa Brassicaceae and

Journal Pre-proof

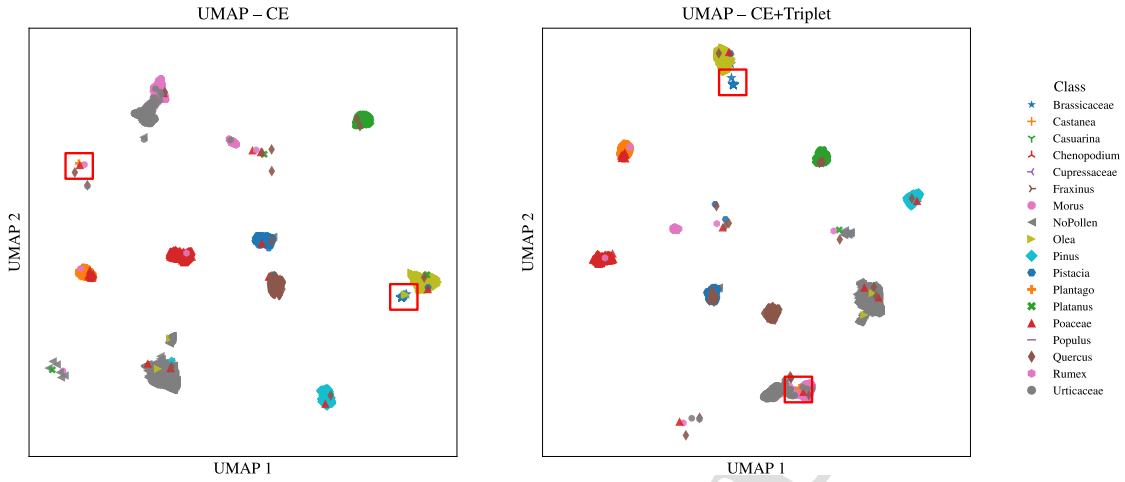


Figure 5: **UMAP projections of the embeddings.** (Left) CE-only. (Right) CE+Triplet ($\lambda = 0.5$). Each is fitted separately with the same prespecified settings, so axes and distances are not comparable across panels; red squares mark regions for qualitative inspection only.

Castanea.

3.5. Statistical validation of the model comparisons

Several comparisons involve differences of a few thousandths of macro-F1. Five-fold cross-validation varies the training/validation partition at a fixed seed; across Phases II and II-B, held-out test macro-F1 differs from the fold mean by -0.0059 to $+0.0039$, in both directions, so the evaluated split is not systematically favourable. This measures partition sensitivity only: optimisation variability is estimated by the Phase III seed repetition on one grouped partition, and CE and CE+Triplet were not repeated under matched seeds.

Paired architecture and objective tests. Over the five folds, the corrected resampled t-test of Nadeau and Bengio (2003) detects no difference among the three Phase II finalists ($p = 0.247$, 0.886 and 0.575). Under the hybrid objective, DINOv2 versus

Journal Pre-proof

ConvNeXt is likewise not detectable ($p = 0.122$), whereas the DINOv2–DaViT fold contrast is $p = 0.007$. These are exact, unadjusted p-values; the Nadeau–Bengio correction accounts for fold dependence, not multiplicity.

On the fixed held-out partition, McNemar's exact test (McNemar, 1947) over 7,358 paired softmax predictions finds fewer DINOv2 errors than ConvNeXt (118 versus 72 discordant outcomes, $p = 1.0 \times 10^{-3}$) and DaViT (118 versus 69, $p = 4.2 \times 10^{-4}$); ConvNeXt and DaViT are indistinguishable (104 versus 101, $p = 0.889$). These small paired-accuracy effects are specific to the hybrid softmax readout on this test split. They do not test the 1-NN macro-F1 used to select PollenFormerUCO and do not establish universal architectural superiority.

Uncertainty when observations are not independent. These intervals treat the 7,358 test images as independent, but images from one session share illumination, focus, preparation and device state. Resampling acquisition dates instead — a cluster bootstrap over the 4,850 dated test images spanning 209 dates — widens accuracy intervals by 1.2 to 1.5× and macro-F1 intervals by 1.9 to 2.8×, because the rare taxa that macro-F1 weights equally are concentrated in few sessions; *Casuarina*'s 29 test images, all from one date, are one independent observation rather than 29. Under date-clustered resampling the three models' macro-F1 intervals overlap across almost their entire range.

What this licenses. On this partition DINOv2-ViT Base is the strongest model under the hybrid objective, and its 1-NN readout of 0.9748 macro-F1 is the highest we report. Against the cross-entropy baselines the reading is parity in accuracy rather than superiority, with PollenFormerUCO additionally providing the queryable embedding used by the gallery and error analyses; Section 3.6 narrows this further.

Journal Pre-proof

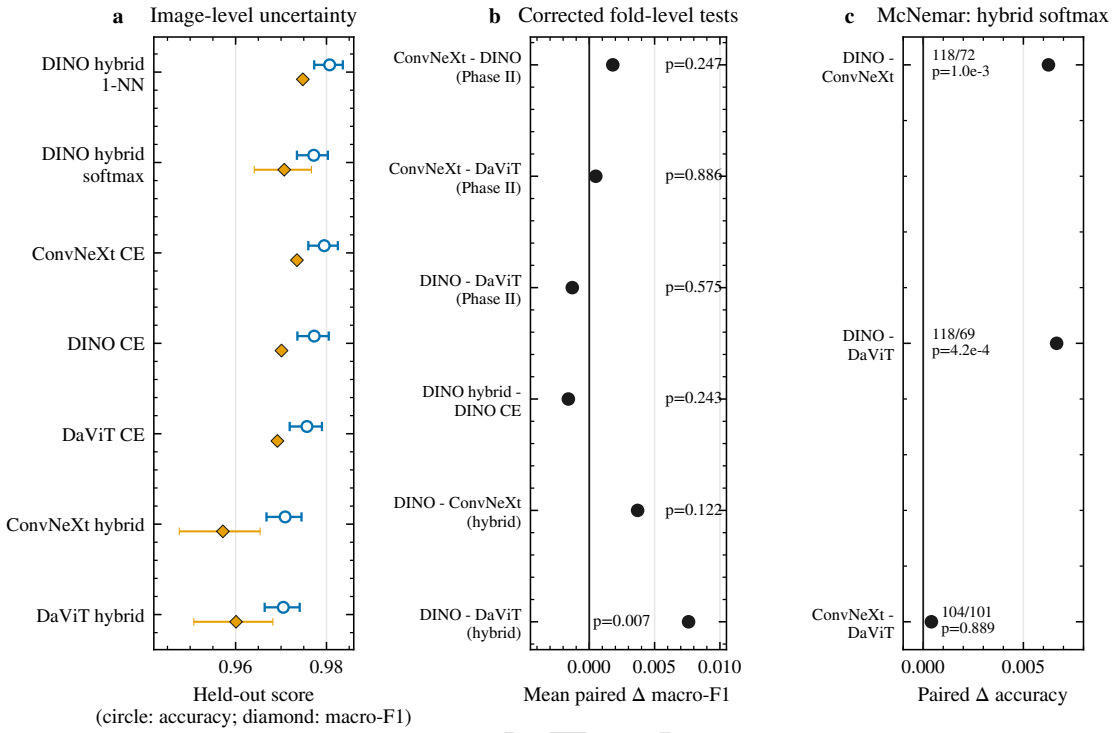


Figure 6: **Uncertainty and paired model comparisons.** (a) Wilson intervals on held-out accuracy and, where test predictions were retained, percentile-bootstrap intervals on macro-F1. (b) Mean paired differences in validation macro-F1 over five folds, with corrected resampled-t p-values. (c) McNemar comparisons of hybrid softmax predictions on the test set: paired accuracy difference, discordant counts and exact unadjusted p-values.

3.6. Phase III: exploratory robustness across acquisition dates

Phase III evaluates the date-disjoint partition of Section 2.4.4 (Table 3). It is a post hoc robustness analysis, not an independent validation: every design choice was made on the image-level split, whose training images include material from dates later assigned to the grouped test subset.

Sensitivity to thin classes. The selected model's softmax head attains 0.9639 accuracy and 0.8854 macro-F1 here, while its 1-NN readout reaches 0.9647 and 0.8946,

Journal Pre-proof

respectively. The corresponding image-level softmax values are 0.9772 and 0.9707. Four of the 17 grouped-test classes fall below the declared 50-image support threshold, and the two smallest — NoPollen (23 images) and *Populus* (21) — score F1 of 0.514 and 0.250 while macro-averaging weights them like *Olea*, measured on 669. Restricting both partitions to the same 13 classes that meet the threshold gives 0.9670 against 0.9754. Even that residual gap cannot be attributed to date separation alone, because the partitions also differ in corpus size, class count and class distribution. The 17-class figure is the headline; the 13-class one is a support diagnostic.

Run-to-run variability matches the differences being compared. Three seeds of the selected configuration estimate optimisation variability on this fixed partition. They give softmax macro-F1 of 0.8854, 0.8781 and 0.8941 — a sample standard deviation of 0.0080 and a range of 0.0160 — against 0.0054 separating the three architectures at a common seed. The ordering reverses under the nearest-neighbour readout, where the seeds span 0.0045 against 0.0199 for the architectures. Both readouts support the same conclusion: differences of this size are comparable to the variability of rerunning one pipeline, so a ranking from single runs is not an architectural result. The paired tests agree — among DINOv2-ViT Base, ConvNeXt-Base and the cross-entropy control no difference is detectable ($p = 0.34$, 0.88, 0.24), with DaViT-Small the exception. As a floor on resolution, the two cross-entropy control runs share seed 42 and differ only in kernel-level nondeterminism, yet their 1-NN macro-F1 differs by 0.0120.

The retrieval capability comes from the head. The hybrid objective changes three things at once: the triplet term, the projection head and the decision rule. A cross-entropy run carrying the same 128-D projection and read with the same 1-NN rule isolates the first. It reaches 0.8901 softmax macro-F1 against the hybrid's 0.8859,

Journal Pre-proof

and 0.8898 under 1-NN against 0.8925; McNemar over the 5,377 paired predictions gives $p = 0.88$. The retrieval capability is therefore a property of the projection head rather than of the triplet term, and two Phase II-B observations prove partition-specific: the 1-NN readout leads in only four of seven runs here, and the hybrid records the lowest silhouette of any configuration tested (0.753, against 0.780 for the control).

3.7. Test-set generalization of PollenFormerUCO

The held-out test partition was not used for training, hyperparameter selection or checkpoint selection, and was evaluated once (Tables 1b and 2). The softmax head achieves 97.72% accuracy and 97.07% macro-F1, while 1-NN inference on the 128-D projected embeddings reaches 98.07% and 97.48%, improving on the trained head by 0.41 macro-F1 points without parameters of its own. The softmax macro-F1 of 0.9707 sits 0.0039 (1.56 standard deviations) above the 5-fold mean of 0.9668±0.0025; since the test partition is not a fold, we report this gap as an observation only.

Figure 7 localises the 168 softmax errors of this evaluation, as a post-selection description rather than a further selection step. *Populus* has the lowest recall (85/99, 85.9%), whose uncertainty is quantified below. The largest directional error cells are *Chenopodium*→NoPollen (18), Urticaceae→*Morus* (16), *Morus*→Urticaceae (13), and NoPollen→*Chenopodium* (11). Of the 24 softmax errors within NoPollen, 11 are assigned to *Chenopodium*; conversely, only 23 of 6,247 pollen images are assigned to NoPollen, including 18 *Chenopodium* images.

3.8. Where the errors are: rare taxa and difficult negatives

The aggregate macro-F1 conceals how precisely each class is measured (Figure 8a). Recall is therefore accompanied by Wilson 95% intervals (Wilson, 1927);

Journal Pre-proof

Table 3: **Phase III: post hoc date-disjoint sensitivity analysis.** No acquisition date is shared between training and evaluation (Table B.1); 17 classes, test $n = 5{,}377$, hyperparameters carried over from Phase II-B. The 13-class diagnostic excludes the classes with fewer than 50 grouped-test images (NoPollen, *Populus*, *Castanea*, Brassicaceae). Values are not comparable with Table 1, which uses a different partition, class count and sample size.

		Softmax head		1-NN (cosine)				
Model	Objective	Acc.	$F1_{macro}$	Acc.	$F1_{macro}$	Softmax $F1_{macro}$ (13 cls.)	Sil. ↑	Val. best
DINOv2-ViT Base	hybrid, seed 42	0.9639	0.8854	0.9647	0.8946	**0.9670**	0.753	0.9457
DINOv2-ViT Base	hybrid, seed 1	0.9578	0.8781	0.9624	0.8927	0.9584	0.711	0.9224
DINOv2-ViT Base	hybrid, seed 2	0.9613	**0.8941**	0.9635	0.8901	0.9647	0.751	0.9422
ConvNeXt-Base	hybrid	0.9613	0.8803	0.9634	0.8925	0.9591	**0.799**	0.9263
DaViT-Small	hybrid	0.9533	0.8857	0.9526	0.8747	0.9513	0.787	0.9154
Control: cross-entropy with the same projection head, read the same two ways								
DINOv2-ViT Base	CE + proj. (a)	0.9645	0.8913	**0.9678**	0.8838	0.9659	0.780	0.9402
DINOv2-ViT Base	CE + proj. (b)	**0.9647**	0.8888	0.9663	**0.8958**	0.9661	0.769	0.9369

Note: Val. best = best validation macro-F1. Bold marks the best value per column. The control runs (a) and (b) share seed 42 and differ only in kernel-level nondeterminism. Seed and architecture dispersion and the paired tests are reported in Section 3.6; DaViT-Small is significantly less accurate than the hybrid DINOv2 model and the control (McNemar $p \leq 5 \times 10^{-3}$) despite a similar macro-F1, because paired accuracy and macro-F1 weight the class-support imbalance differently.

Journal Pre-proof

Journal Pre-proof

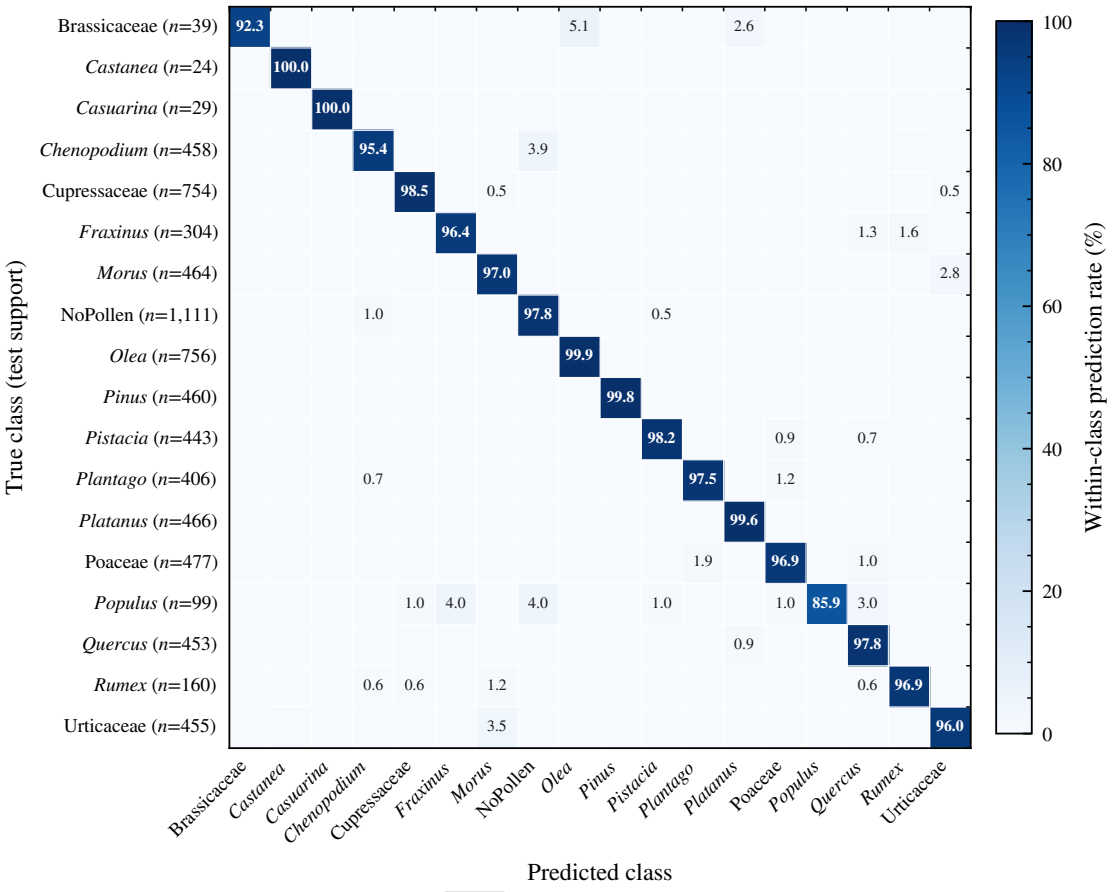


Figure 7: **Held-out softmax error structure of PollenFormerUCO.** Row-normalised confusion matrix on the image-level test partition (n = 7,358); cells of at least 0.5% of their row are annotated in percent, and row labels give test support. The matrix is descriptive and was not used for selection.

for the rarest taxa the interval, not the point estimate, is the informative quantity: Brassicaceae is measured on 39 images, so its recall of 0.923 carries an interval 17.7 percentage points wide, and *Castanea* (24) and *Casuarina* (29) both record recall 1.000 with intervals reaching down to 0.862 and 0.883, against 0.7 points for *Olea* on 756. Per-class scores for the rare taxa are individually uninformative and should

Journal Pre-proof

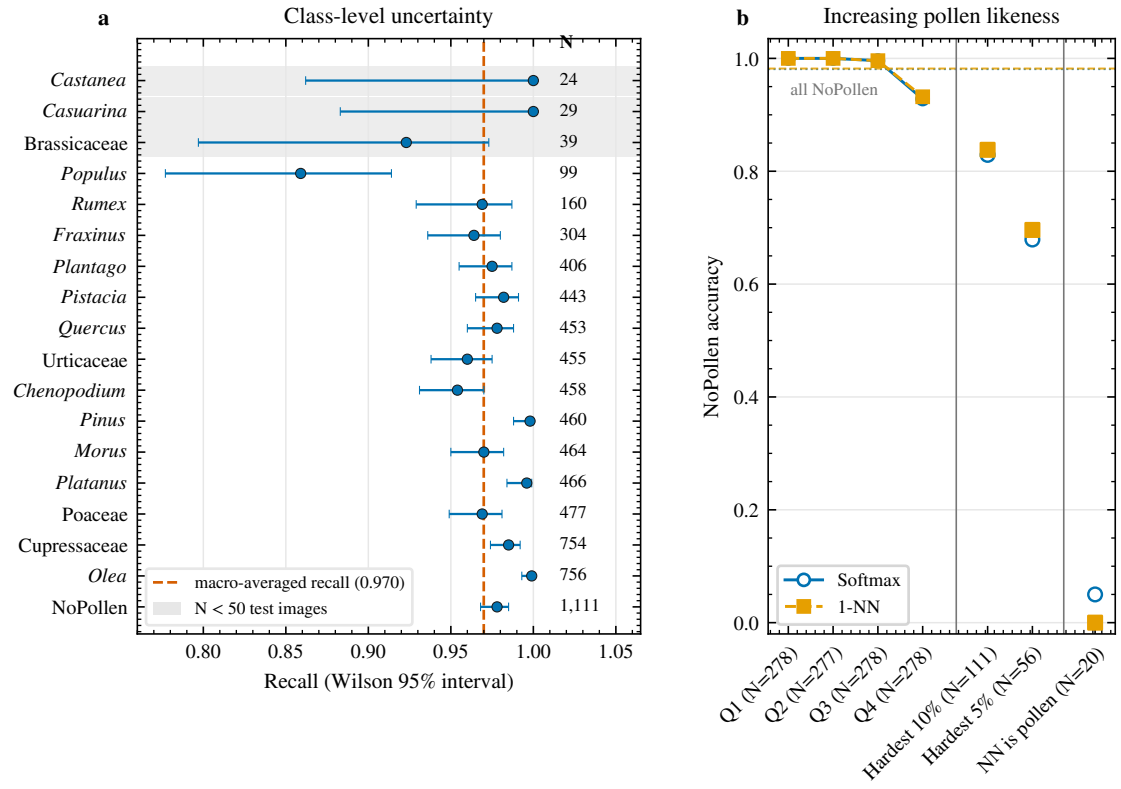


Figure 8: **Rare-class uncertainty and difficult negatives for PollenFormerUCO.** (a) Recall of each class with its Wilson 95% interval, ordered by test support. Points further right are better; a longer interval means a less precise estimate, as for the rare taxa in the shaded rows (fewer than 50 test images). The dashed line marks macro-averaged recall. (b) Accuracy within NoPollen as objects become more pollen-like (higher is better): four disjoint quartiles (joined), the nested hardest-10% and hardest-5% tails, and objects whose nearest neighbour is pollen.

not be compared without their intervals.

Class imbalance was nevertheless not the binding constraint here: the four least-represented taxa reach F1 between 0.9239 and 1.0000, against 0.9707 macro-F1 overall, so they are classified about as well as the abundant ones. All experiments therefore use unweighted cross-entropy, by design. The rarest taxa are ambient samples

Journal Pre-proof

(Section 2.1), so their scarcity reflects how seldom they occur in the monitored air, which is also the class prior a deployed classifier meets. Loss reweighting, resampling and targeted augmentation all over-represent these taxa artificially — in the objective or in the data — and shift the decision rule towards classes that are genuinely rare, inflating their false positives in the counts a monitoring network reports. Resampling and augmentation also multiply images from very few sessions — *Casuarina* comes from a single date — rather than adding independent information (Dablain and Chawla, 2024). With 24 to 39 test images for these taxa, their Wilson intervals are moreover wider than the difference any such remedy could plausibly produce.

Difficult negatives within NoPollen. NoPollen mixes trivial background with particles closely resembling grains, so its aggregate accuracy cannot show by itself whether performance reflects genuine discrimination. We therefore stratified its 1,111 test objects by s_{pollen}, the maximum cosine similarity to any pollen training embedding, which ranks difficulty in the representation used at inference and needs no subtype annotation (Figure 8b). The two easiest quartiles are classified without error, while accuracy falls to 92.8% in the most pollen-like quartile and to 67.9% over the hardest 5%. For the 20 objects whose nearest training neighbour is a pollen grain, the softmax head recovers one and 1-NN none, so the aggregate 97.8% overstates performance on genuinely ambiguous negatives.

The difficulty is, however, concentrated — only 1.8% of negatives fall into the failing regime — and systematic, with *Chenopodium* accounting for 11 of the 24 misclassified images. In the opposite direction only 23 of 6,247 pollen test images (0.37%) were assigned to NoPollen. The residual risk therefore sits in a small, identifiable population of pollen-like debris, and s_{pollen} is a ready criterion for routing those objects to expert review.

Journal Pre-proof

3.9. Exploratory gallery extension with unseen taxa

Operational use requires handling taxa absent from the training set. We examine this exploratorily with additional images from the same department and BAA502 system, annotated by expert aerobiologists but excluded from UCOPollen, so they appear in none of the preceding splits; they test unseen taxa *within the same acquisition domain*, not cross-device, cross-region or cross-year transfer. Both experiments use cosine 1-NN over $L2$-normalised 128-D PollenFormerUCO embeddings, with no retraining. The base gallery is the 34,257 training embeddings; for each unseen taxon, $\min(n, \max(5, \lfloor n/2 \rfloor))$ images were added to the gallery and the rest held out as queries, under a fixed seed.

3.9.1. Experiment 1: gallery extension with novel taxa

Seven taxa absent from training — Cyperaceae, Ericaceae, Juncaceae, *Ligustrum*, Myrtaceae, *Salix* and *Ulmus* — each contributed a reference gallery of 5–23 images, and 59 disjoint queries were classified against the augmented gallery (Figure 9a). Overall 1-NN accuracy reached 52.5% and the true class was among the five nearest neighbours for 61.0% of queries, without a single gradient step.

Performance splits along morphology rather than gallery size. Ericaceae was recognised perfectly (8/8) and *Ulmus* and Myrtaceae reached 56.5% and 57.1%, whereas Juncaceae (0/4, confused with *Quercus*), *Salix* (0/3) and *Ligustrum* (0/2, mapped entirely to *Olea*) were never recognised. The last is biologically interpretable — both genera are Oleaceae with tricolpate grains — and marks the boundary of the method: distinctive taxa are captured by the existing embedding, whereas taxa close to a training class need dedicated data.

A majority vote among five neighbours lowers accuracy to 37.3%, as expected when a new taxon contributes at most 23 embeddings against 34,257. Several taxa

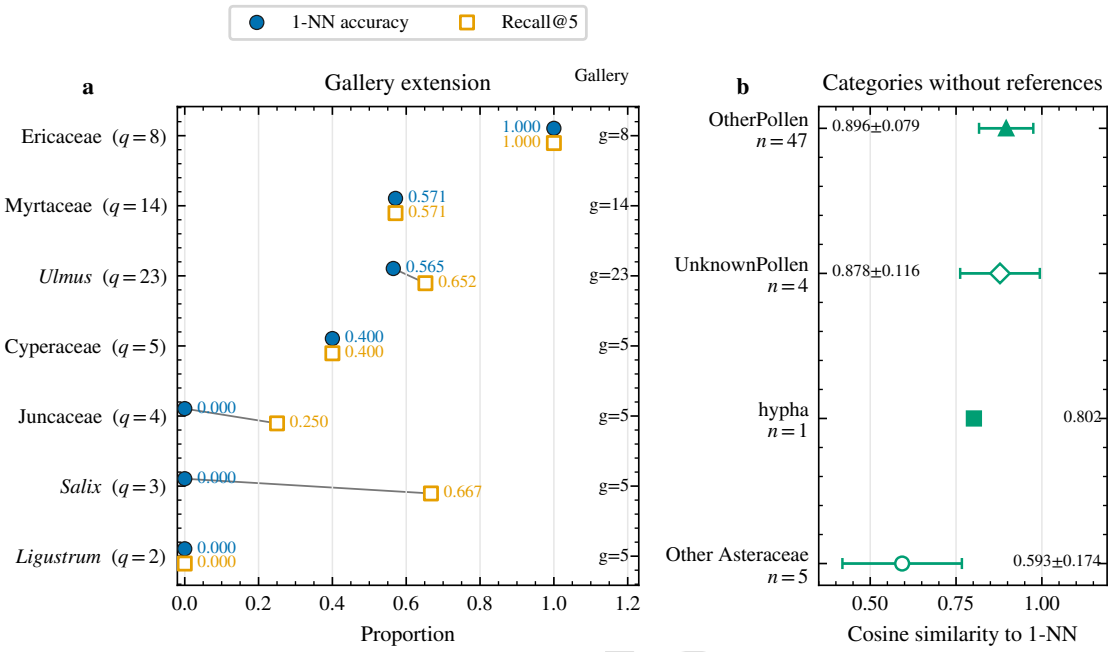


Figure 9: **Exploratory open-set behaviour.** Cosine 1-NN over 128-D PollenFormerUCO embeddings, without retraining. (a) Accuracy and Recall@5 for seven unseen taxa; q is query support and g gallery size. (b) Mean nearest-neighbour similarity ($\pm$ SD) for categories without references; the single hypha has no dispersion.

contribute only a handful of queries, so the aggregate figures are the reliable ones.

3.9.2. Experiment 2: behaviour on out-of-distribution samples

Fifty-seven further external images belonging to neither the training classes nor the added gallery were queried against the 25-class gallery of Experiment 1: OtherPollen (n=47), Other Asteraceae (n=5), UnknownPollen (n=4) and a single hypha (Figure 9b).

Two regimes emerge. The 32 OtherPollen samples labelled *Taxus* by the device software gave uniformly high similarity (0.907 ± 0.073), and 31 were assigned to Cupressaceae — biologically coherent, since both are conifers with spherical inaper-

Journal Pre-proof

turate grains, but *confident*, so a distance threshold alone would not flag them as unknown. Other Asteraceae, in contrast, gave low similarity (0.593 ± 0.174) and no consensus across four classes, as did UnknownPollen; that combination is precisely the signal a rejection mechanism exploits. The hypha was assigned to NoPollen (sim $= 0.802$), the correct operational response for a non-pollen artefact.

3.9.3. Implications for open-set deployment

Together the two experiments show that the projected embedding can be queried beyond the closed training taxonomy and that its failure mode is predictable from morphology. Distant unknowns give low similarity and no consensus, which a rejection rule can exploit; close ones give confident, taxonomically sensible but nominally wrong assignments, which require gallery extension or dedicated rejection, such as per-class distance thresholds or Mahalanobis outlier detection. This is same-device, preliminary evidence, not a demonstration of operational readiness.

4. Discussion

Automatic monitoring answers the limitations of Hirst-type counting — expert identification, delayed data and inter-analyst variability (Tummon et al., 2021; Buters et al., 2024; Smith et al., 2019) — and reviews of the field identify classification algorithms and comparable datasets as the binding constraint (Farooq et al., 2025), the gap UCOPollen and this benchmark address. Deployment must also account for the variability among automatic systems across devices, taxa and algorithms (Maya-Manzano et al., 2023) and for environmental conditions (González-Alonso et al., 2024).

The training regime, not the architecture, governs performance. Two questions usually conflated in backbone benchmarks — which architecture is better, and how

Journal Pre-proof

Journal Pre-proof

the backbone is adapted — have answers of very different magnitudes here. Under frozen probing the eleven checkpoints span 0.1641 macro-F1 and DINOv2-ViT Base leads; under an identical full fine-tuning budget the spread collapses to 0.0043, the ranking inverts, and no pairwise difference survives the corrected paired test. For practitioners, the choice among a modern CNN, a hybrid-attention Transformer and a self-supervised ViT does not determine the outcome, and since DaViT-Small matches the best model within the cross-validation dispersion at 56% of the parameters, parameter efficiency becomes the operative criterion. For the literature, a linear-probe ranking is not an architectural ranking: stopping at Phase I would have selected a different backbone and attributed to architecture a property of frozen-feature separability. Under the hybrid objective, by contrast, DINOv2-ViT Base is significantly more accurate than both supervised backbones (Section 3.5), which is why it underpins PollenFormerUCO.

The adaptation schedule decides whether a self-supervised backbone delivers. A representation learned from the 142 million images of LVD-142M (Oquab et al., 2024) is already close to what the task requires, and driving the whole backbone at the rate suited to a randomly initialised head discards it. DINOv2-ViT Base selects a backbone rate one decade below that of both supervised backbones and two below its own head (Section 2.4), and reaches 0.9748 macro-F1. When a self-supervised checkpoint appears to underperform after fine-tuning, the schedule should be examined before the architecture.

The projection head is what makes the embedding deployable. The learned projection yields a 128-D space markedly better separated than the 768-D token it is computed from (test silhouette 0.617 $\rightarrow$ 0.809), in which a parameter-free nearest-neighbour rule matches the trained head. That geometry lets unseen taxa be recognised from

Journal Pre-proof

a handful of reference images without retraining (Section 3.9), which an operational system needs because real samples contain taxa absent from any fixed label set. Under the post hoc date-disjoint analysis a cross-entropy model carrying the same projection head matches the hybrid (McNemar $p = 0.88$), so the capability follows from the projected space rather than from the triplet term. Metric learning is one route to that space, and on the image-level partition the best one; it is not a precondition.

Errors concentrate where morphology is genuinely close. Minority taxa such as *Castanea* and Brassicaceae are drawn towards similar or more abundant classes, and *Morus* against Urticaceae accounts for 32 misclassifications, while distinctive and well-represented taxa such as *Pinus*, *Olea* and *Platanus* are near-perfect. These patterns align with the taxa human experts also find difficult (González-Alonso et al., 2023).

4.1. Limitations

All data originate from a single institution and a single imaging device, so cross-device and cross-regional validation remains necessary before deployment in heterogeneous networks.

The partition is at the image level. No image appears in more than one subset, but 191 of the 257 acquisition dates are represented in all three, most markedly for the three taxa enriched through controlled feeding: 94.3% of *Pistacia*, 43.1% of *Morus* and 24.3% of *Chenopodium* images come from a few feeding dates. Phase III separates dates only within the limits of the surviving metadata — 65.9% of the corpus, *Casuarina* excluded, NoPollen reduced to 81 images — and is post hoc; grouping by calendar date also does not guarantee independent biological material, since a feeding campaign can span consecutive days. Group-aware evaluation of this

Journal Pre-proof

kind of data therefore needs the acquisition context to be recorded at capture rather than reconstructed afterwards.

The two dispersion estimates each hold one factor fixed: cross-validation varies the partition at a single seed, and the Phase III repetitions vary the seed at a single partition, so neither describes their combination.

No class-imbalance remedy was applied, for the reasons given in Section 3.8, so its effect on macro-F1 for this dataset was not measured.

The composition of NoPollen could not be quantified (Section 2.1), so Section 3.8 characterises difficulty through embedding geometry rather than particle subtype. The gallery experiments used reference sets of 5 to 23 images and query sets as small as two to five, so per-class rates are indicative, and no calibrated unknown-class rejection was evaluated; per-class distance thresholds, entropy gating or Mahalanobis detection are the main requirement for unsupervised operation.

5. Conclusions

We introduced UCOPollen, a Mediterranean dataset of 48,950 expert-annotated microscopic images, and benchmarked eleven pretrained vision backbones on it under controlled training regimes. Under frozen probing the checkpoints spanned a wide range and a self-supervised ViT led; once the three finalists were fine-tuned under an identical budget, that spread collapsed by more than an order of magnitude, the ranking inverted and no pairwise difference proved distinguishable.

PollenFormerUCO, a fully fine-tuned DINOv2-ViT Base with a learned 128-D projection, attains the highest macro-F1 of the study through a parameter-free cosine 1-NN readout (98.07% accuracy, 97.48% macro-F1), and its softmax head is more accurate than both supervised backbones trained under the same objective. More importantly for monitoring, its compact embedding lets new taxa be recognised from

Journal Pre-proof

a few reference images without retraining, the property needed to scale recognition to the open set of taxa a sampler encounters. That capability derives from the projection head rather than from the triplet term, since a cross-entropy control carrying the same head matches it under date-disjoint evaluation. Gallery extension succeeds for morphologically distinctive taxa and fails for those close to a training class, so calibrated rejection and external validation remain prerequisites for operational use.

Two constraints should inform future benchmarks on this kind of data. A date-disjoint partition of UCOPollen is constructible for only 17 of the 18 classes and for the images carrying a genuine acquisition timestamp, so group-aware evaluation requires acquisition context recorded at capture rather than reconstructed afterwards. And because an anchor contributes a triplet only when its class appears at least twice in a batch, the in-batch sampler forms few triplets for the rarest taxa, which limits what the metric term can contribute to a macro-averaged score.

Two directions would most strengthen this line of work. On the data side, material from further automatic monitoring devices, sites and seasons, with the acquisition context of every image (device, date and time, and sampling session) recorded at capture, would make group-aware partitions possible by construction and allow the cross-device, cross-region and cross-year validation that operational use requires. On the modelling side, the natural next steps are calibrated unknown-class rejection for gallery-based recognition, a triplet sampler that raises coverage of the rarest taxa without altering the class frequencies seen by the classifier, and lightweight backbones with approximate nearest-neighbour search for real-time inference.

CRediT authorship contribution statement

Ricardo Espantaleón-Pérez: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writ-

Journal Pre-proof

Journal Pre-proof

ing – review & editing, Visualization. **Qasim Farooq:** Resources, Data curation, Validation, Writing – review & editing. **Rocío López-Orozco:** Resources, Data curation, Validation, Writing – review & editing. **Mónica González-Alonso:** Resources, Data curation, Validation, Writing – review & editing. **Samer Alashhab:** Resources, Data curation, Software, Validation, Writing – review & editing. **Jose Oteros:** Resources, Validation, Supervision, Writing – review & editing. **Jeroen Buters:** Resources, Validation, Writing – review & editing. **Carmen Galán:** Resources, Validation, Writing – review & editing. **Manuel J. Marín-Jiménez:** Conceptualization, Methodology, Validation, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The source code, trained model weights, and UCOPollen dataset will be publicly available in the project repository[1]. A live interactive demo of PollenFormerUCO is available at https://richardesp-pollenformer-uco-demo.hf.space/ for direct evaluation on new images.

Funding

This study was supported by the European Union (NextGenerationEU/PRTR) and by the project *A SYstem for reaL-time obserVation of Aeroallergens* (SYLVA),

[1] https://github.com/AVAuco/pollenformerUCO

Journal Pre-proof

Journal Pre-proof

grant agreement no. 101086109, under call HORIZON-CL6-2022-GOVERNANCE-01, 2023–2026.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author Ricardo Espantaleón-Pérez used ChatGPT (OpenAI), Google Gemini (Google), and Claude Code (Anthropic) in order to improve the language and readability of the manuscript and to assist with drafting and editing. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Appendix A. Frozen-backbone benchmark: checkpoints and results

Table A.1 resolves the exact pretrained weights and preprocessing configuration of every backbone in the Phase I benchmark, and Table A.2 reports their Phase I results.

Appendix B. Construction of the date-disjoint partition

Only the 65.9% of images carrying a recovered acquisition timestamp (32,255 of 48,950) are eligible for grouping; the remainder carry a day-level identifier derived from the database object id, which records *insertion* rather than acquisition. Whole dates are allocated by a greedy rule that places each date where it least disturbs the per-class target proportions, measured as squared *relative* deviation so that a taxon contributing 81 images weighs as heavily as one contributing 4,559; without the relative weighting the rule sweeps scarce classes into the training subset. Table B.1 gives the resulting composition. Every date appears in exactly one partition and all 17 retained classes are present in all three, both asserted when the split is built.

Journal Pre-proof

Table A.1: **Exact pretrained checkpoints used in the frozen-backbone benchmark.** Identifiers resolve the weights and preprocessing configuration used by `torch.hub` or `timm` v1.0.16.

Backbone	Exact identifier	Pretraining	Input
DINOv2-ViT Base	`facebookresearch/dinov2:dinov2_vitb14`	LVD-142M, self-supervised	224
DINOv2-ViT Small†	`facebookresearch/dinov2:dinov2_vits14`	LVD-142M, self-supervised	224
DaViT-Small	`timm/davit_small.msft_in1k`	ImageNet-1K, supervised	224
DeiT-Small†	`timm/deit_small_patch16_224.fb_in1k`	ImageNet-1K, supervised	224
ConViT-Small†	`timm/convit_small.fb_in1k`	ImageNet-1K, supervised	224
MobileViT-S†	`timm/mobilevit_s.cvnets_in1k`	ImageNet-1K, supervised	256
ConvNeXt-Base	`timm/convnext_base.fb_in1k`	ImageNet-1K, supervised	224
ConvNeXt-Small†	`timm/convnext_small.fb_in1k`	ImageNet-1K, supervised	224
EfficientNetV2-S	`timm/tf_efficientnetv2_s.in1k`	ImageNet-1K, supervised	224
EfficientNetV2-M	`timm/tf_efficientnetv2_m.in1k`	ImageNet-1K, supervised	224
ResNet18†	`timm/resnet18.a1_in1k`	ImageNet-1K, supervised	224

Note: † The run export recorded only the architecture string for these backbones. The weight tag shown is the `timm` or `torch.hub` default resolved for that architecture at the version stated in the caption.

Table A.2: **Frozen-backbone linear-probe benchmark (Phase I).** A linear head trained on each frozen backbone with Cross-Entropy, LR 10^{-3} and early stopping on validation macro-F1. CV is five-fold on the training partition, and the test partition has $n = 7{,}358$. SSL = self-supervised learning. Rows are ordered by test macro-F1; bold marks the best value per column.

Backbone	Family	Params (M)	Epoch	CV $F1_{macro}$	Test acc.	Test $F1_{macro}$
DINOv2-ViT Base	SSL ViT	86.59	31	$\mathbf{0.8168 \pm 0.0107}$	**0.8722**	**0.8323**
DaViT-Small	Hybrid attention	48.99	44	0.8043 ± 0.0044	0.8576	0.8225
ConvNeXt-Base	CNN	87.58	49	0.7943 ± 0.0058	0.8471	0.8118
ConViT-Small	Hybrid attention	27.35	26	0.7815 ± 0.0054	0.8407	0.7981
EfficientNetV2-S	CNN	20.20	30	0.7916 ± 0.0063	0.8464	0.7955
DeiT-Small	Distilled ViT	21.67	34	0.7755 ± 0.0070	0.8292	0.7913
DINOv2-ViT Small	SSL ViT	22.06	28	0.7681 ± 0.0060	0.8372	0.7885
ConvNeXt-Small	CNN	49.47	49	0.7703 ± 0.0037	0.8326	0.7883
EfficientNetV2-M	CNN	52.88	44	0.7409 ± 0.0071	0.8149	0.7594
ResNet18	CNN	11.19	49	0.7222 ± 0.0053	0.8047	0.7316
MobileViT-S	Mobile hybrid	4.95	45	0.6604 ± 0.0084	0.7605	0.6682

Journal Pre-proof

Table B.1: **Phase III: the date-disjoint partition.** Whole acquisition dates are assigned to a single partition. Only images with a recovered acquisition date are eligible (65.9% of UCOPollen), and *Casuarina* is excluded because all its images share one date.

Class	Train	Validation	Test	Total
Brassicaceae	160	41	42	243
Castanea	92	26	23	141
Chenopodium	513	139	115	767
Cupressaceae	2,952	673	899	4,524
Fraxinus	1,249	281	282	1,812
Morus	946	363	427	1,736
NoPollen	52	6	23	81
Olea	3,006	884	669	4,559
Pinus	1,846	535	481	2,862
Pistacia	366	182	110	658
Plantago	1,624	369	369	2,362
Platanus	1,672	657	398	2,727
Poaceae	1,941	505	449	2,895
Populus	429	78	21	528
Quercus	1,876	377	475	2,728
Rumex	587	120	121	828
Urticaceae	1,754	430	473	2,657
Total	21,065	5,666	5,377	32,108
Acquisition dates	193	36	27	256

Note: The proportions (65.6% / 17.6% / 16.7%) depart from 70/15/15 because dates are indivisible. Four classes have fewer than 50 images in some partition (NoPollen, *Populus*, *Castanea*, Brassicaceae).

Appendix C. Training dynamics of the final fits

Figure C.1 shows the per-epoch validation macro-F1 and cross-entropy loss of the six final fits of Phases II and II-B. Every run stops ten epochs after its selected checkpoint, before the 50-epoch ceiling. Validation loss reaches its minimum earlier than macro-F1 (epochs 4–18) and rises by at most 0.023 by the selected epoch while macro-F1 still improves, the usual signature of increasingly confident predictions rather than of degraded classification; checkpoints are therefore selected on macro-F1, the metric reported throughout.

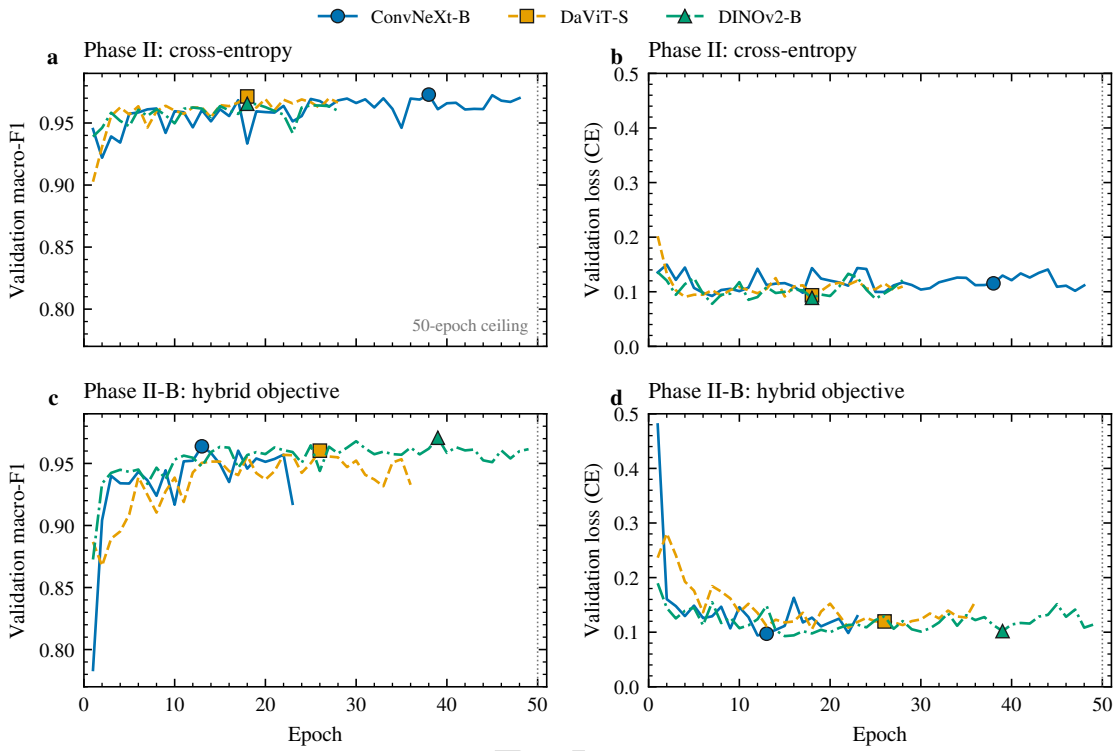


Figure C.1: **Validation trajectories of the final fits.** Validation macro-F1 (a, c) and validation loss (b, d) per epoch for Phase II (a, b) and Phase II-B (c, d); the loss is plain cross-entropy for both objectives, so the rows are comparable. Filled markers mark the selected checkpoint (maximum validation macro-F1), each curve ends where early stopping (patience 10) halted the run, and the dotted line marks the 50-epoch ceiling.

References

Aggarwal, C.C., Hinneburg, A., Keim, D.A., 2001. On the surprising behavior of distance metrics in high dimensional space. Springer Berl. Heidelb. , 420–434doi:10.1007/3-540-4450.

Boullayali, A., Hassoun, M., Bouziane, H., 2025. Variation of pollen season trends

Journal Pre-proof

under mediterranean climate: a systematic review. Aerobiologia 41, 469–488. doi:10.1007/s10453-025-09862-0.

Buters, J., Clot, B., Galán, C., Gehrig, R., Gilge, S., Hentges, F., O'Connor, D., Sikoparija, B., Skjoth, C., Tummon, F., et al., 2024. Automatic detection of airborne pollen: an overview. Aerobiologia 40, 13–37. doi:10.1007/s10453-022-09750-x.

Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Commun. Stat. 3, 1–27. doi:10.1080/03610927408827101.

Dablain, D.A., Chawla, N.V., 2024. Data augmentation's effect on machine learning models when learning with imbalanced data, in: 2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10. doi:10.1109/DSAA61799.2024.10722815.

Davies, D.L., Bouldin, D.W., 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-1, 224–227. doi:10.1109/TPAMI.1979.4766909.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

Ding, M., Xiao, B., Codella, N., Luo, P., Wang, J., Yuan, L., 2022. Davit: Dual attention vision transformers, in: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (Eds.), Computer Vision – ECCV 2022, Springer Nature Switzerland, Cham. pp. 74–92.

d'Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G., Sagun, L., 2022.

Journal Pre-proof

Convit: improving vision transformers with soft convolutional inductive biases*. J. Stat. Mech.: Theory Exp. 2022, 114005. doi:10.1088/1742-5468/ac9830.

Farooq, Q., López-Orozco, R., Martínez-Bracero, M., Lucas, R.W., Galán, C., Oteros, J., 2026. Automatic monitoring and imaging-based identification of airborne olea pollen. Environ. Res. , 124268doi:10.1016/j.envres.2026.124268.

Farooq, Q., Oteros, J., Galán, C., 2025. Advancing in the pollen frontier: a comprehensive evaluation and meta-analysis of automatic pollen monitoring systems. Aerobiologia 41, 527–546. doi:10.1007/s10453-025-09865-x.

González-Alonso, M., Oteros, J., Widmann, M., Maya-Manzano, J.M., Skjøth, C., Grewling, Ł., O'Connor, D., Sofiev, M., Tummon, F., Crouzy, B., Clot, B., Buters, J., et al., 2024. Influence of meteorological variables and air pollutants on measurements from automatic pollen sampling devices. Sci. Total Environ. 931, 172913. doi:10.1016/j.scitotenv.2024.172913.

González-Alonso, M., Boldeanu, M., Koritnik, T., Gonçalves, J., Belzner, L., Stemmler, T., Gebauer, R., Łukasz Grewling, Tummon, F., Maya-Manzano, J.M., Ariño, A.H., Schmidt-Weber, C., Buters, J., 2023. Alternaria spore exposure in bavaria, germany, measured using artificial intelligence algorithms in a network of baa500 automatic pollen monitors. Sci. Total Environ. 861, 160180. doi:10.1016/j.scitotenv.2022.160180.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. doi:10.1109/CVPR.2016.90.

Journal Pre-proof

Hirst, J.M., 1952. An automatic volumetric spore trap. Ann. Appl. Biol. 39, 257–265. doi:10.1111/j.1744-7348.1952.tb00904.x.

Konijn, T., Bijl, I., Cao, L., Verbeek, F., 2025. Analysis of 3d urticaceae pollen classification using deep learning models. Proc. 18th Int. Jt. Conf. Biomed. Eng. Syst. Technol. 18, 288–295. doi:10.5220/0013102700003911.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986. doi:10.1109/CVPR52688.2022.01167.

López-Orozco, R., García-Mozo, H., Oteros, J., Galán, C., 2023. Long-term trends and influence of climate and land-use changes on pollen profiles of a mediterranean oak forest. Sci. Total Environ. 897, 165400. doi:10.1016/j.scitotenv.2023.165400.

Maya-Manzano, J.M., Tummon, F., Abt, R., Allan, N., Bunderson, L., Clot, B., Crouzy, B., Daunys, G., Erb, S., González-Alonso, M., et al., 2023. Towards european automatic bioaerosol monitoring: comparison of 9 automatic pollen observational instruments with classic hirst-type traps. Sci. Total Environ. 866, 161220. doi:10.1016/j.scitotenv.2022.161220.

McInnes, L., Healy, J., Saul, N., Großberger, L., 2018. Umap: Uniform manifold approximation and projection. J. Open Source Softw. 3, 861. doi:10.21105/joss.00861.

McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153–157. doi:10.1007/BF02295996.

Journal Pre-proof

Mehta, S., Rastegari, M., 2022. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178 doi:10.48550/arXiv.2110.02178.

Nadeau, C., Bengio, Y., 2003. Inference for the generalization error. Machine Learning 52, 239–281. doi:10.1023/A:1024068626366.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2024. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 doi:10.48550/arXiv.2304.07193.

Oteros, J., Pusch, G., Weichenmeier, I., Heimann, U., Möller, R., Röseler, S., Traidl-Hoffmann, C., Schmidt-Weber, C., Buters, J., 2015. Automatic and online pollen monitoring. Int. Arch. Allergy Immunol. 167, 158–166. doi:10.1159/000436968.

Rostami, M.A., Balmaki, B., Dyer, L.A., Allen, J.M., Sallam, M.F., Frontalini, F., 2023. Efficient pollen grain classification using pre-trained convolutional neural networks: a comprehensive study. J. Big Data 10, 151. doi:10.1186/s40537-023-00815-3.

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. doi:10.1016/0377-0427(87)90125-7.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, . pp. 815–823. doi:10.1109/CVPR.2015.7298682.

Journal Pre-proof

Journal Pre-proof

Shamrat, F.M.J.M., Idna Idris, M.Y., Zhou, X., Khalid, M., Sharmin, S., Sharmin, Z., Ahmed, K., Moni, M.A., 2024. Pollennet: A novel architecture for high precision pollen grain classification through deep learning and explainable ai. Heliyon 10. doi:10.1016/j.heliyon.2024.e38596.

Smith, M., Oteros, J., Schmidt-Weber, C., Buters, J.T.M., 2019. An abbreviated method for the quality control of pollen counters. Grana 58, 185–190. doi:10.1080/00173134.2019.1570327.

Tan, M., Le, Q.V., 2021. EfficientNetV2: Smaller models and faster training, in: Proceedings of the 38th International Conference on Machine Learning (ICML), pp. 10096–10106.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H., 2021. Training data-efficient image transformers & distillation through attention, in: Meila, M., Zhang, T. (Eds.), Proceedings of the 38th International Conference on Machine Learning, PMLR. pp. 10347–10357.

Tummon, F., Arboledas, L.A., Bonini, M., Bütler, F., Celenk, S., Clot, B., Crouzy, B., Gehrig, R., Gilge, S., Hajkova, L., Magyar, D., Menzel, A., O'Connor, D., Sikoparija, B., Skjøth, C.A., Sofiev, M., Sozinova, O., Vasilatou, K., de Weger, L.A., 2021. The need for pan-european automatic pollen and fungal spore monitoring: a stakeholder workshop position paper. Clin. Transl. Allergy 11, e12015. doi:10.1002/clt2.12015.

Wightman, R., 2019. Pytorch image models. GitHub repository. URL: https://github.com/huggingface/pytorch-image-models. version 1.0.16.

Wilson, E.B., 1927. Probable inference, the law of succession, and statistical infer-

Journal Pre-proof

ence. Journal of the American Statistical Association 22, 209–212. doi:10.1080/01621459.1927.10502953.

Zhang, X., Zhao, L., Tong, D.Q., Wu, G., Dan, M., Teng, B., 2016. A systematic review of global desert dust and associated human health effects. Atmosphere 7, 158. doi:10.3390/atmos7120158.

Šaulienė, I., Šukienė, L., Daunys, G., Valiulis, G., Vaitkevičius, L., Matavulj, P., Brdar, S., Panic, M., Sikoparija, B., Clot, B., Crouzy, B., Sofiev, M., 2019. Automatic pollen recognition with the rapid-e particle counter: the first-level procedure, experience and next steps. Atmos. Meas. Tech. 12, 3435–3452. doi:10.5194/amt-12-3435-2019.

Journal Pre-proof

Journal Pre-proof

Highlights

Vision Transformers for Pollen Grains Classification: A Systematic Backbone Benchmark on UCOPollen

- UCOPollen: 48,950 expert-annotated images of 17 Mediterranean pollen taxa
- Eleven CNN, ViT and self-supervised backbones in four phases, one on held-out dates
- Hybrid cross-entropy and triplet metric learning benchmarked on all three finalists
- PollenFormerUCO (DINOv2) gives the best macro-F1, 97.5%, with cosine 1-NN
- Unseen taxa can be added to a reference gallery without retraining the model

Journal Pre-proof

Journal Pre-proof

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof